

Automatic Lifelog Media Annotation based on Heterogeneous Sensor Fusion

Ig-Jae Kim, Sang Chul Ahn, Heedong Ko and Hyung-Gon Kim

Abstract— Personal Lifelog media system involves capturing of great amount of personal experiences in the form of digital multimedia during an entire lifespan. However, the usefulness of those data is limited by lack of adequate methods for accessing and indexing such a large database. It is important to manage those data systematically so that user can efficiently retrieve useful experiences whenever they need. In this paper, we focus on presenting how to create metadata, which is the core of the systematical approach, by the fusion of sensor data from a set of heterogeneous sensors. With this metadata, we can support users to find their life history efficiently in our system.

I. INTRODUCTION

WRITING a diary and capturing photos, videos are natural activities, almost everyone has done at some point in their lives, that help to remember past experiences. It means that almost people want to log their everyday activities. However, it is not so easy to record everything occurred around us manually except capturing several photos or videos. Personalized Lifelog system can help those things more systematically.

It can record almost everything related user automatically occurred in daily living by wearing various sensors and wearable computer. Besides, it analyzes and classifies the captured data so that users can retrieve and replay their histories whenever they need. Like this, personalized Lifelog system aims to be a function of tracer of an individual's life by compiling a massive database of every activities of daily living. The captured data through Lifelog system include physical location recorded via wearable GPS sensor, motion information through triple axis accelerometers, audiovisual data and object information user contacts using camera/microphone and RFID sensor, respectively.

Recently, a number of works for Lifelog system have been proposed in the area of wearable computing, video retrieval. As one of earliest work for Lifelog system, Lamming and Flynn [1] record various personal activities such as personal location, file exchange, workstation activities. Microsoft's MyLifeBits project [2] tries to collect and store any digital information about a person such as documents, images, sounds and videos, but leaves the annotation to the user in early version. Since then, it has expanded in many directions, including new forms of capture using SenseCam [3] and more general use of typed links.

Aizawa et al. [4] found that watching Lifelog videos is a

critical problem and it would take another year to watch the entire Lifelog video for one year so that they suggested a new capturing system using GPS, motion and brain-wave analyzer with wearable camera. By using these sensor data, they try to estimate user's context so that they can respond to video retrieval queries of various forms correctly and flexibly. Mann [8] described EyeTap which facilitate the continuous archival and retrieval of personal experiences by way of lifelog video capture.

Recognizing general human activity or special motions is important key for automatic annotation. Randell et al. [5] have done early investigations of the problem using only single biaxial accelerometers and Kern et al. [6] summarized work on automatically annotating meeting recordings, extracting context from body-worn acceleration sensors alone and combining context from three different sensors for estimating the interruption of the user. Some researchers tried to infer the human activities by detecting human-object interaction [21].

Vemuri et al. [7] presented a method for audio-based memory retrieval. They developed a personal computer based memory retrieval tool allowing browsing, searching, and listening to audio and associated speech- recognizer generated transcripts.

In this paper, we propose to use wearable sensors in order to enhance the recorded data with contextual, personal information to facilitate user friendly retrieval. Sensors, such as accelerometers and biometric sensors, can enhance the recording with information on the user's context, activity and physical state. Such information can be used to annotate and structure the data stream for later associate access.

The remainder of this paper is organized as follows. Section 2 overviews our Lifelog system. Section 3 gives a detail description of creating metadata from capture sensor data. Section 4 presents the implementation of our proposed system and reports the experimental results, and Section 6 concludes the paper.

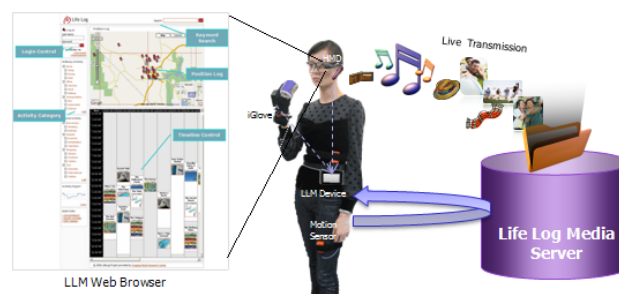


Fig. 1. Concept of Lifelog system

I. J. Kim, S. C. Ahn, H. Ko and H. G. Kim are with Korea Institute of Science and Technology, Imaging Media Research Center, 39-1 Hawolgok-dong, Seongbuk-gu, Seoul, Korea (e-mail: {kij, asc, ko, hgk}@imrc.kist.re.kr)

II. OVERVIEW OF OUR LIFELOG SYSTEM

To capture and manage user's experiences in form of multimedia automatic annotation for the continuous media data with recognized metadata is required. For this personalized Lifelog system is divided mainly into three logical parts such as Lifelog media(LLM) client, user client and LLM server.

LLM client includes various wearable sensors to capture user experiences in form of digital media and manages its metadata that is used for the automatic annotation of user experiences. The collected user experiences and created metadata are sent to the LLM server in real-time. Most experience media data captured by LLM client are archived in the LLM server in raw format while other metadata is captured and organized to represent user experience in abstracted metadata database. LLM server also has web server to present the media and metadata to the User client.

User client is the unit where the user can review the Lifelog media data. This user client module physically can be any machine with web browsing and media handling capability. It even can be the same device with LLM client.

The implementation of LLM Client is done using ultra mobile computer with several wearable sensors. For the media transmission, we use VLC[9], an open source video utility, that captures and encodes in MPEG-4, and transmit it using RTP protocol directly to the LLM server in real-time.

Other sensor data are transmitted to the LLM server with windows socket communication. LLM server stores the media data as files and manages their metadata using database system. In LLM server, the video stream is received and saved as file while other sensor data are processed and saved to the database as metadata.

Metadata database system is implemented using these automatically generated metadata of location, time, action, and surrounding information including person, object, and environment. The database is located in LLM server and is indexed from the metadata generated by sensors in LLM client. The LLM data captured by the LLM client is organized in file system with timestamps.

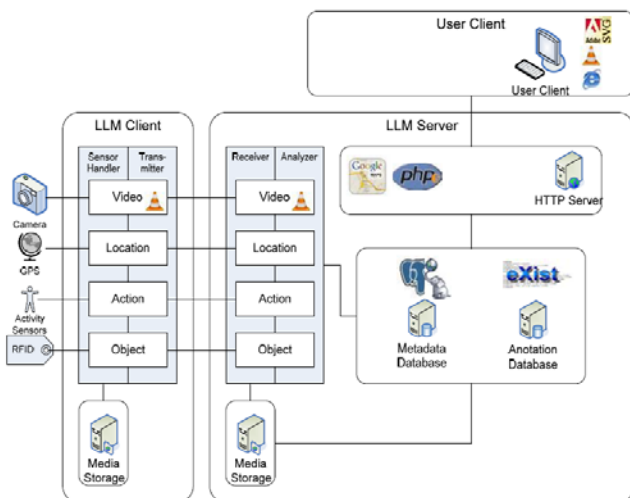


Fig. 2. Architecture of our Lifelog Media System

The user interface of the proposed personalized Lifelog media system is shown in Fig. 1. Using this interface, user can query using metadata in various query level. The search results may be displayed on map interface as well as under time line table arranged by location and time with thumbnail images[10]. User can select one of the candidate results to be played. While playing the video, the web interface can also display the video annotation taken from detailed metadata database. The media and metadata are presented using web browser.

The database in the Lifelog server manages media and metadata to enable efficient retrieval of Lifelog media. The usual method to manage metadata to describe multimedia is using time-based annotation that links metadata to some portion of the media file. XML is the most popular model used for this time-based annotation and several schemas like MPEG-7 [11] and Dublin Core [12] are already standardized to manage metadata of a video.

Various sensor fusion operations are performed on the captured sensor signals to enhance the analysis of the metadata. Fig. 3 shows how we can get the elements of metadata from several wearable sensors using sensor fusion technique.

Location information is provided by GPS receiver. Environmental information can be obtained by sensor fusion of RFID, audio information or by analyzing the video. Activity information is provided based on human motion and object identification.

Two wireless accelerometers are used for the classification of 5 human body states using decision tree, and detection of RFID tagged objects with hand movement provides additional object related hand motion information. Person information is calculated from analysis of video or audio. We combine the result from face recognition from input video and speech recognition from input audio. For object information user contact, we made RFID based glove system which use HF RFID system. Owing to range limitation of RFID system, we also have to analyze input video to find out the information of long-distance objects.

The following section will present a detailed explanation of how we calculate the metadata information from each sensor.

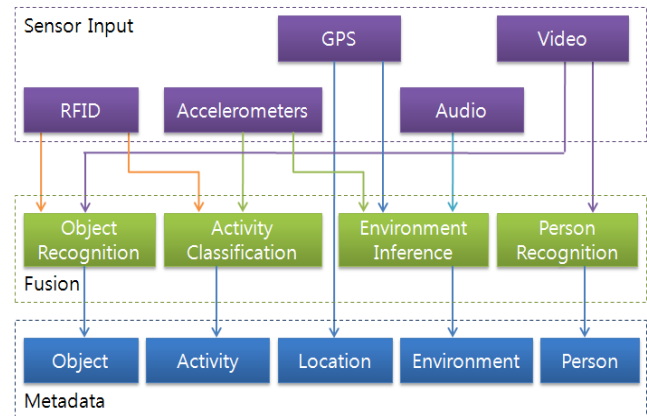


Fig. 3. Sensor fusion architecture for metadata creation

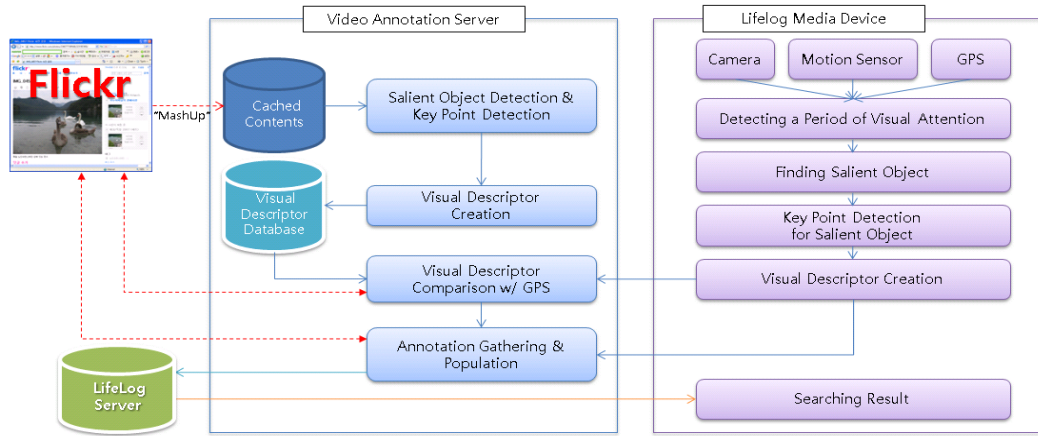


Fig. 4. Sensor assisted video annotation flow

III. LLM METADATA FROM WEARABLE SENSORS

A. Video

As mentioned in introducing previous works, it is hard to find a video we want precisely in an enormous amount of videos captured in our daily life if there's no appropriate explanation or some tags to the corresponding video. As part of strategies which can solve mentioned problem, video sharing website, such as Youtube [13], Pandorav [14] and so on, encourage users to write relevant tags to describe a piece of information for the uploading video. With this kind of information are possible to some extent. However, it is almost impossible for users to write a keyword or term manually to every video in case of capturing continuously, such like Lifelog system.

In this section, we introduce our method to make a relevant annotation for the captured video from our Lifelog system. As mentioned in previous section, our Lifelog system uses multiple sensors so that it can use those sensors to detect a period of recognizing object that users are concerned using the information from the sensors. Fig. 4 shows our method for video annotation using web service with some sensors. The detail explanation of the flow is as followings.

We assume that people are used to stop their walking and fix their eyes on something, if they find something to be concerned. We use three sensors to find such a state, such as GPS, accelerometer and image from camera. Using GPS, we can find a time when people stop their walking to see something by the following equation.

$$\|Lat^t - Lat^{t-1}\| + \|Lon^t - Lon^{t-1}\| < TH_{GPS} \quad (1)$$

Lat^t is a latitudinal value, Lon^t is a longitudinal value from GPS at time t and TH_{GPS} is a predefined threshold.

We also find a time when people are fixing his eye on something interesting by wearing motion sensor which is attached to the camera on his head with a following equation.

$$\|Acc_x^t - Acc_x^{t-1}\| + \|Acc_y^t - Acc_y^{t-1}\| + \|Acc_z^t - Acc_z^{t-1}\| < TH_{motion} \quad (2)$$

$Acc_{x,y,z}^t$ is a value from each axis of accelerometer at time t .

After detecting a period of concern, we extract keyframe representing all frames corresponding the period.

Before we calculate the visual descriptors, we try to find a salient object in the keyframe. Because the human brain and visual system pay more attention to some part of an image, we call those area salient objects. It is natural that we focus salient objects and extract more feature vectors from them than others. It can enhance the efficiency of visual descriptor matching process. A number of techniques for finding salient object have been proposed since the work of Itti et al. [15]. Most existing approaches are based on the bottom up computational framework because visual attention is in general unconsciously driven by low-level stimulus in the scene such as intensity, contrast and motion [16]. We use an approach that is similar to the previous work [16]. We used a contrast map to extract accurate salient object. To make a contrast map, we generate color map and orientation map. Using these maps, we can find the area of visual attention.

Before generating color map, we apply Gaussian filter to the keyframe in order to remove Gaussian noise in advance.

We downsize the keyframe by a factor of 2, 3 and 4 and apply the different size of filters to the downsized keyframe. The filter calculates the difference of intensity between a center point and neighborhood points of an image. Because created feature map is sensitive the size of this contrast filter, we apply the filtering to the different level of image. We apply two different size of filter to each level of image and therefore we can get six feature maps for the extracted keyframe. By combining these feature maps, we can get a normalized color map (\bar{C}).

$$\bar{C} = \frac{1}{18} \left(\sum_{c=2}^4 D \left(\sum_{s=c+3}^{c+4} L(c,s) + \sum_{s=c+3}^{c+4} A(c,s) + \sum_{s=c+3}^{c+4} B(c,s) \right) \right) \quad (3)$$

We use CIELAB color model. $L(c,s)$ is a luminance difference between a center point and its neighborhood at image size c and filter size s and $A(c,s)$ and $B(c,s)$ are

results from filtering of color values in CIELAB model.

Since salient regions have special texture information, we also use Haar wavelet transform to get texture information of different level of detail. After one level Haar transform, we get the horizontal, vertical and diagonal texture information from the wavelet subbands. We then apply same process of generating color map to the created each subband and get the normalized orientation map(\bar{O}).

Finally, we can get the contrast map by linear combination of color and orientation map.

After finding salient object from the keyframe, we extract more visual descriptors from salient object than from others.

For any object there are many features, interesting points on the object that can be extracted to provide a "feature" description of the object. This description can then be used when attempting to locate the object in an image containing many other objects. There are many considerations when extracting these features and how to record them. SIFT image features provide a set of features of an object that are not affected by many of the complications experienced in other methods, such as object scaling and rotation [17]. While allowing for an object to be recognized in a larger image SIFT image features also allow for objects in multiple images of the same location, taken from different positions within the environment, to be recognized. SIFT features are also very resilient to the effects of "noise" in the image.

On this score, we use SIFT image features as visual descriptors in our system. When we search an image that has a high correlation to an image contains objects which user gives an attention, we extract more visual descriptors from the area of salient objects than from others so that we can enhance the searching efficiency.

These extracted visual descriptors of captured keyframe from Lifelog media device are sent to our video annotation server and compared to those of search results from photo sharing site, such as Flickr [18], considering GPS information (see Fig. 4 for the flow of this process). After comparing visual descriptors between them, we can read the annotation of web image that has a high correlation with captured keyframe using Mashup technique.

B. Audio

From audio input signal, we can get the information of environment where the audio is captured with motion sensor data. Gaussian Mixture Model (GMM) [19] is used as our classification tool. GMMs belong to the class of pattern recognition systems. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, it refines the weights of each distribution through expectation-maximization algorithms.

For classification of captured environment, we extract 26th MFCC feature vector from input audio signal. Input audio signal is converted to 13th order MFCC by frame (1 frame = 10 ms), which is more adequate for classification. Fig. 6 shows this process.

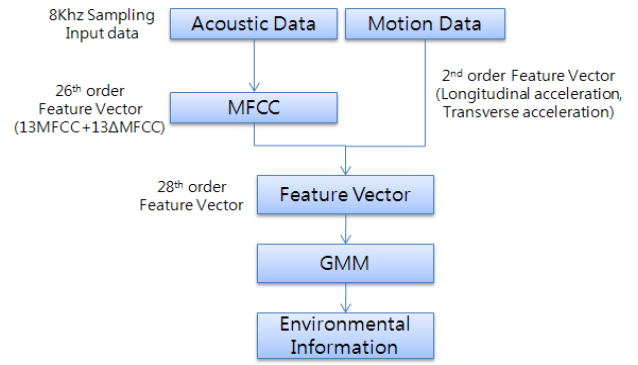


Fig. 5. Sensor fusion flow using audio signal and motion data for classification of captured environment

Besides, we extract 2nd order feature vector which is composed of longitudinal acceleration and transversal acceleration from motion sensor. We then compose a new feature vector which has 28th order by considering the frame correlation between these two feature vectors. The constructed feature vector is used to calculate likelihood using pre-captured learning data set. By comparing the result, we can classify the environment from pre-constructed environment model.

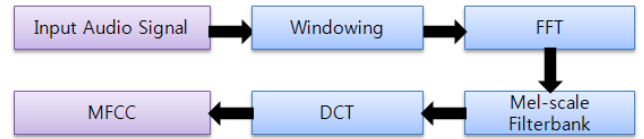


Fig. 6. Converting process from input audio signal to MFCC

C. Accelerometers and RFID

Activity of daily living(ADL) is a way to describe the functional status of person. It is, therefore, important to record such information for Lifelog system. To recognize ADL automatically, we used three triaxial accelerometers.

Accelerometers can provide quantitative measurements and respond to both acceleration for gravity and acceleration for body movement. This makes them suitable for measuring postural orientations as well as body movements. To detect the physical activities of the user, three triaxial accelerometers were worn on thigh, waist and wrist. Among them, two sensors which were attached to waist and thigh were used to classify the body state. The feature values of mean, energy, entropy and correlation in frequency-domain are calculated from acceleration data with 256 sample window.

According to the results that Bao et al. [20] showed a decision tree classifier which had the best performance in recognizing ADL, we also used decision tree classifier to recognize user's physical state, such as walking, running, lying, standing and sitting.

RFID systems have been widely applied over recent years for the identification and tracking applications [21]. Passive RFID tags do not need battery and are small enough to be attach on the small objects like screw driver and tooth brush.

The RFID tagged objects are detected by the glove which integrates the HF RFID reader of 13.56 Mhz. The reader is also small enough so that can be embedded in a wearable glove, iGrabber (see Fig. 7 for user wearing iGrabber). We can easily get the information of object that user contacts using iGrabber. By combining accelerometers and this iGrabber, we can extend our activity recognition module to the recognition of various instrumental ADL, such as drinking with cup, tooth brushing, ironing and so on. For instrumental activities, we examine the movement of hand and object ID to infer the more detailed intension of the user.

If the movement of hand is detected, then we try to detect the object ID taken by the hand using our RFID reader. If an object is detected, we can infer user's activity from the detected object ID in a classified body state.

IV. SYSTEM IMPLEMENTATION

A. Wearable Sensors for Lifelog Media Data

In Fig. 7, we show a Lifelog user wearing multiple sensors in our experiment. As explained in previous section, we record audiovisual data and detect some periods of visual attention using camera, microphone and accelerometer attached to the camera. By wearing iGrabber, we know the object information that user contacts. We also infer user's activity of daily living through triaxial accelerometers by combining captured object information.

All sensor data except video signal, which is connected to USB, are gathered to LLM device through Bluetooth. LLM data collected in LLM device are sent to LLM server through Wi-Fi in realtime. Fig. 1 shows this process conceptually. Fig. 9 shows an example of structure of metadata database from wearable sensors. One XML file describes one media file. Each XML annotation file contains tracks to describe modality and elements to describe metadata occurred in that video time.



Fig. 7. Lifelog user with wearable sensors

B. Lifelog Web Client

The user interface in our Lifelog web client is shown in Fig. 8. Using this interface, a user can query using extracted metadata in various query level. The search results can be displayed on map interface as well as under time line table arranged by location and time with thumbnail images[9]. User can select one of the candidate results to be played. While playing the video, the web interface can also display

video annotation taken from metadata database.

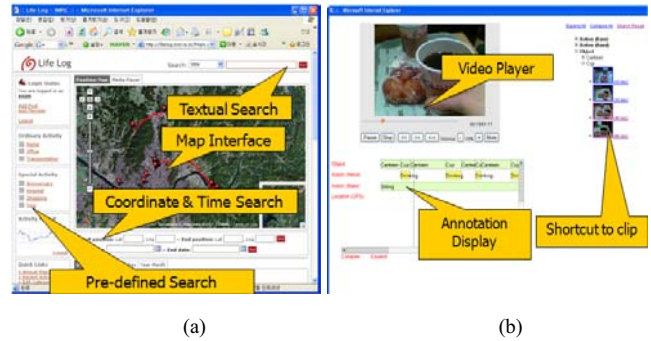


Fig. 8. Snapshot of our Lifelog web client. (a) Logged position can be displayed as red balloon on the map (b) web interface support video annotation according to video

```
<annotation>
<head>
<video src="1980.01.01_01.01.01.asf"/>
</head>
<body>
<track name="Location">
<el index="0" start="6" end="40">
<attribute >126.997 37.636397</attribute>
</el>
<el index="1" start="42" end="77">
<attribute >126.9999 37.677897</attribute>
</el>
</track>
<track name="Person">
<el index="0" start="4" end="51">
<attribute >Kim Hyoungnyoun</attribute>
</el>
</track>
<track name="Object">
<el index="0" start="1" end="5">
<attribute >Pen</attribute>
</el>
<el index="1" start="12" end="18">
<attribute >Pen</attribute>
</el>
</track>
<track name="Environment">
<el index="0" start="8" end="9">
<attribute >snow</attribute>
</el>
</track>
</body>
</annotation>
```

Fig. 9. Example of metadata for Lifelog media annotation

C. Experimental Result

Define Among the several sensor fusion results of our system, we show the result of audio and motion sensor fusion and the result of activity recognition.

TABLE I

RECOGNITION RESULT FROM AUDIO							
output Input	bus	theater	office	street	subway	down town	library
Bus	70	0	0	0	30	0	0
theater	0	100	0	0	0	0	0
office	0	2.5	95	2.5	0	0	0
street	0	0	2.5	97.5	0	0	0
subway	43	17	0	5	35	0	0
downtown	0	0	0	45	0	50	5
library	0	0	0	0	0	0	100

TABLE II

RECOGNITION RESULT FROM AUDIO AND MOTION DATA							
output Input	bus	theater	office	street	subway	down town	library
Bus	100	0	0	0	0	0	0
theater	0	100	0	0	0	0	0
office	0	0	92.5	2.5	0	5	0
street	0	0	2.5	97.5	0	0	0
subway	0	2.5	0	17.5	67.5	12.5	0
downtown	0	0	0	40	0	60	0
library	0	0	0	0	0	0	100

To evaluate the performance of environmental recognition using audio and motion data, we checked in every 3 seconds and calculate the average for one hour. The results (in Table I, II) show that the fusion with audio and motion data makes more robust than the case with audio only.

TABLE III
ACTIVITY RECOGNITION RESULT WITH MOTION DATA AND RFID

Activities	Accuracy (motion only)	Accuracy (motion + RFID)
Sitting	97.64	100.0
Standing	96.72	100.0
Walking	90.92	96.91
Lying	95.94	94.02
Running	94.68	97.91
Sitting + Drinking	94.57	94.38
Standing + Drinking	62.86	97.68
Walking + Drinking	88.19	94.86
Ironing + Standing	69.23	97.94
Cutting + Standing	56.35	98.67
Brush hair + Standing	84.16	98.34
Repairing + Standing	55.02	94.98
Overall	82.26	97.16

We've collected acceleration data from 18 subjects of 11 males and 7 females using 3 accelerometers. We used decision tree classifier to classify the motion data into predefined activities. The result (in Table III) shows that we can get more confident result from the fusion with motion data and RFID than from motion sensor only.

V. CONCLUSION

In this paper, we introduce our Lifelog system that enables user to capture, manage and retrieve experience media and especially, focus on explaining how to generate the metadata automatically using wearable sensors. Wearable sensors include audiovisual device, GPS, accelerometers and RFID sensor.

Our Lifelog system is good environment for testing multiple sensor fusion, because we have to use various sensors to capture different type of Lifelog media. We therefore can increase the confidence in finding metadata by combining different sensor data. For better Lifelog service, we need an annotation based on high level inference from low level sensor data. For this, we need more sophisticated sensor fusion technique.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MIC/IITA. [2006-S-032-03, Development of an Intelligent Service technology based on the Personal Life Log].

REFERENCES

[1] M. Lamming and M. Flynn, "Forget-me-not: intimate computing in human memory," in *Proceeding of International Symposium Next Generation Human Interface*, pp. 125-128, Feb. 1994.
 [2] J. Gemmell, G. Bell, R. Lueder, S. Drucker, C. Wong, "MyLifeBits: fulfilling the Memex vision," in *Proceeding of ACM Multimedia*, 2002, pp. 235-238.

[3] J. Gemmell, L. Williams, K. Wood, R. Lueder and G. Bell, "Passive Capture and Ensuing Issues for a Personal Lifetime Store," *ACM Workshop CARPE*, 2004, pp. 48-55.
 [4] K. Aizawa, D. Tancharoen, S. Kawasaki and T. Yamasaki, "Efficient Retrieval of Life Log Based on Context and Content," *ACM Workshop CARPE*, 2004, pp. 22-31.
 [5] C. Randell and H. Muller, "Context awareness by analyzing accelerometer data," in *Proceeding of Fourth International Symposium on Wearable Computers*, 2000, pp. 175-176
 [6] N. Kern, B. Schiele and A. Schmidt, "Multi-Sensor Activity Context Detection for Wearable Computing," in *European Symposium on Ambient Intelligence*, 2003.
 [7] S. Vemuri, C. Schmandt, W. Bender, S. Tellex and B. Lassey, "An audio based personal memory aid," in *Proceeding of Ubicomp*, 2004, pp. 400-417.
 [8] S. Mann, "Continuous Lifelong Capture of Personal Experience with EyeTap," *ACM Workshop CARPE*, 2004, pp. 1-21.
 [9] VLC Free and Open Source Video Player, Videolan, <http://www.videolan.org>.
 [10] I. J. Kim, S. C. Ahn and H. G. Kim, "Personalized Life Log Media System in Ubiquitous Environment," *Lecture Notes in Computer Science (LNCS)*, vol. 4412, pp. 20-29, Jan. 2007.
 [11] MPEG-7 Overview, International Organization for Standardization. <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>.
 [12] Dublin Core, Dublin Core Metadata Initiative(DCMI), <http://dublincore.org/index.shtml>.
 [13] <http://www.youtube.com>
 [14] <http://www.pandoratv.com>
 [15] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on Pattern Analysis Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.
 [16] S. Kwak, B. Ko and H. Byun, "Automatic Salient-Object Extraction Using the Contrast Map and Salient Points," *Lecture Notes in Computer Science*, Vol. 3332, pp. 138-145, 2005.
 [17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.
 [18] <http://www.flickr.com>
 [19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19-41, Jan. 2000.
 [20] L. Bao and S. S. Intille, "Activity Recognition from User-annotated Acceleration Data," in *Proceeding of Pervasive(LNCS 3001)*, pp. 1-17, 2004.
 [21] M. Philipose, K. P. Fishkin, M. Perkwitz, D. J. Patterson, D. Fox, H. Kautz and D. Hahnel, "Inferring Activities from Interactions with Objects," *IEEE Pervasive Computing*, pp. 50-57, 2004.