M-025

Selection from a Large Number of Audio and Video Sources for Personalized Video Retrieval in a Ubiquitous Environment

Gamhewage C. de Silva[†] Toshihiko Yamasaki[†] Kiyoharu Aizawa[†‡]

1. Introduction

Multimedia retrieval and summarization for ubiquitous environments is an active research area with several applications such as surveillance, study of human behavior [1], and taking care of the elderly [2]. The presence of a large number of sources and the lack of structure of the content makes this more difficult compared to retrieving broadcast media. Selection of the most appropriate sources is critical for accurate retrieval.

Most of the ubiquitous environments are equipped with various types of sensors that can provide supplementary and context data in addition to cameras and microphones. These data can be utilized to achieve more accurate and effective accurate retrieval [3].

This research is based on *Ubiquitous Home* [4], a two-bedroom house equipped with a large number of stationary cameras and microphones (Figure 1). The place is built as a test-bed for recording, retrieval and awareness of human behavior rather than surveillance, as evident from the positioning of cameras and microphones. Pressure-based sensors mounted on the floor are activated as people move inside the house. The amount of data recorded in a single day is approximately 500 GB.

Personalized video retrieval and summarization for this environment can be extremely tedious if performed manually. For example, if we want to see what Mr. Smith did in ubiquitous home during his visit in the morning of the 3rd of September 2004, it is necessary to watch the video from the camera showing the entrance to the house from early morning until the frames showing Mr. Smith entering the house are detected. Thereafter, it is necessary to pause and switch between several cameras and microphones to track him as he moves within the house, for the entire duration of his stay.

We intend to create a system where video for the above scenario can be retrieved and summarized as follows: first we enter the date and the time interval. This retrieves a set of key frames showing people who had been inside the house during this time interval. For the people who entered or left the house during this time interval, the key frames showing them entering or leaving the house will be displayed with timestamps. For those who remained inside, a key frame at the start of the time interval is displayed. By browsing only the key frames showing the persons entering the house, we can find the key frame showing Mr. Smith. By clicking on the frame, we can see a video clip or a set of key frames, showing what he did. The cameras and microphones are selected automatically as he moves, so that he can be seen and heard throughout the stay.

We propose to implement the above system by analyzing the floor sensor data and extracting video sequences and key frames for each person in the house. We design and conduct an experiment to evaluate the performance of the key frame extraction algorithm that we designed.

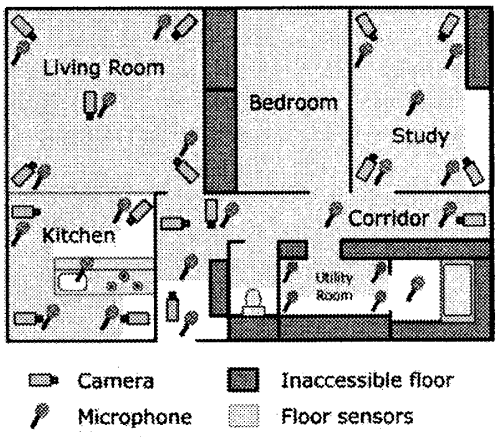

Figure 1. Ubiquitous home layout.

2. Video Retrieval

The pressure on floor sensors is sampled at 6 Hz. Footsteps of the people in the house result in state transitions on floor sensors under the feet. These transitions are recorded, with position and timestamp information.

A 3-stage Agglomerative Hierarchical Clustering (AHC) algorithm, described in our previous work [5], is used to segment floor sensor data into footstep sequences of different persons. This algorithm performs well in the presence of noise and activation delays, and despite the absence of floor sensors in some areas of the house.

We intend to create a video clip keeping a given person in view as he moves within the house. Since the cameras are stationary with fixed zoom, this seems trivial if footstep segmentation has been accurate. However, with more than one camera that can see a given position, it is necessary to select cameras in a way that a "good" summary can be acquired. We refer to this task as *video handover*. We used *position-based handover* [5], an algorithm developed in our previous work. This algorithm attempts to minimize the number of camera changes while keeping the relevant person the view.

The resulting video is sampled to extract key frames so that a more compact summary can be provided to the user. An *adaptive spatio-temporal sampling* algorithm [6] was used for key frame extraction. This algorithm attempts to sample key frames based on the time elapsed since last key frame, camera changes in the sequence and rate of footsteps the person makes.

3. Audio Retrieval

Audio acquisition capability is a new addition to the ubiquitous home; therefore the work on audio retrieval is still in initial stages. Our intention here is to 'dub' the video sequences

[†] Dept. of Frontier Informatics, The University of Tokyo

[‡] Dept. of Electronics Engineering, The University of Tokyo

created by video handover. Although there are a large number of microphones, it is not necessary to use all of them since a microphone can cover a larger range compared to a camera. Furthermore, frequent transitions of microphones can be annoying to listen.

There is little research on microphone selection for ubiquitous environments, at the time of writing. We implement a novel, simple algorithm for *audio handover*. Each camera is associated with one microphone for audio retrieval. For a camera in a room, audio is retrieved from the microphone that is located in the center of that room. For a camera in the corridor, the microphone closest to the center of the region seen by that camera is selected. This algorithm attempts to minimize transitions between microphones while maintaining a reasonable sound level.

4. Evaluation and Results

The system was tested on approximately 1500 hours of video and 700 hours of audio data. The data included those from a "real-life experiment", where a family of three members stayed in ubiquitous home for 10 days.

The video sequences retrieved by the system were evaluated subjectively. It was possible to view a person continuously in the sequences, and the viewers found the camera changes natural. Audio handover maintained an adequate sound level throughout the sequences. The dubbing was generally smooth, other than for the occasional instances where a person was moving from one room to another while talking. The user interface enabled fast retrieval of video. Typical search time for a query within a 3-hour session of data was found to be less than a minute.

To evaluate the performance of key frame extraction, we designed and conducted an experiment. Eight voluntary subjects took part in the experiment. The subjects extracted key frames from four video clips created by video handover, according to their own choice. The key frames selected by subjects were clustered to form an *average key frame set* for each sequence. It was observed that the subjects had a strong agreement on the actions and events to be selected as key frames. This suggests that the average key frame sets can be used as ground truth for performance evaluation of key frame extraction.

To evaluate the performance of key frame extraction quantitatively, we define the rank n performance, R_n as follows:

$$R_n = (K_n / N) \times 100\%$$

Here, K_n is the number of occasions a key frame is present within n frames from that of the average key frame set and N is the number of frames in the average key frame set. Figure 2 plots the cumulative performances against n. The results show that it is possible to extract key frames within a difference of 3 s, with an upper bound of around 80%, using only floor sensor data with this method.

5. Conclusion and Future Work

We have implemented personalized video summarization for a ubiquitous environment, by analyzing signals from pressure based floor sensors. Video and audio handover was used to select

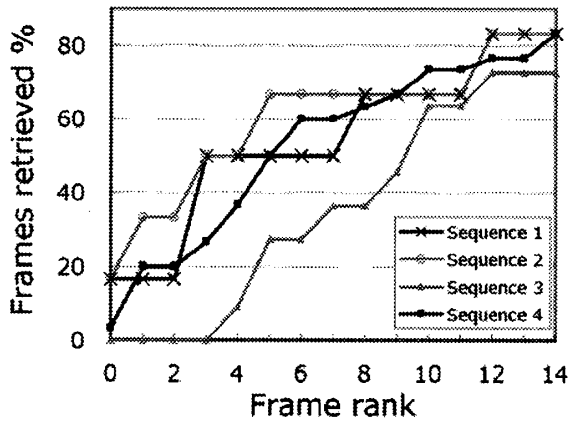

Figure 2. Performance of key frame extraction

cameras and microphones for synthesis of continuous video for a given person. An algorithm that is adaptive to the rate of the footsteps of the person was found to extract about 80% of the key frames that are required for a complete and compact summary of the video.

Future work will focus on extracting key frames for interaction among persons and between a person and an object. At the current state of the work, audio is retrieved using a simple algorithm merely for the purpose of video dubbing. Design and evaluation of better algorithms for audio handover is an interesting future direction. The possibility of using the audio data as a supplementary input for video retrieval is now under investigation.

Acknowledgments

We thank NICT for their cooperation. This work was partially supported by CREST and JST.

References

[1] T. Mori, H. Noguchi, A. Takada, T. Sato, "Sensing Room: Distributed Sensor Environment for Measurement of Human Daily Behavior", Proc. INSS2004, pp.40-43, 6 (2004).

[2] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata: Memory Cues for Meeting Video Retrieval, Proc. CARPE 2004, USA.

[3] M. Davis, S. King, N. Good, "From Context to Content: Leveraging Context to Infer Media Metadata", Proc. ACM Multimedia 2004. Pp. 188-195

[4] T. Yamazaki, "Ubiquitous Home: Real-life Testbed for Home Context-Aware Service", Proc. Tridentcom2005, pp.54-59, February 23, 2005.

[5] Gamhewage C. de Silva, T. yamasaki, T. Ishikawa, K. Aizawa, "Video Handover for Retrieval in a Ubiquitous Environment Using Floor Sensor Data", In proc. ICME 2005, July 2005, Netherlands.

[6] Gamhewage C. de Silva, T. Yamasaki, K. Aizawa, "Video Retrieval in a Ubiquitous Environment with Floor Sensors", to appear in Proc. ITE National Conference, August 2005, Japan.