A SENSOR NETWORK FOR EVENT RETRIEVAL IN A HOME LIKE UBIQUITOUS ENVIRONMENT

Gamhewage C. de Silva, Steve Anavi, Toshihiko Yamasaki, Kiyoharu Aizawa

Department of Frontier Informatics
The University of Tokyo

ABSTRACT

We present the current status of a system based on a network of sensors for event retrieval from a home like environment. A large number of cameras, microphones and pressure based floor sensors are used for continuous data capture. The data are analyzed independently and the results recorded in a central database, where they are combined for efficient retrieval and summarization of the video and audio data. We describe the detection of basic actions, events, and faces using image analysis. The users can query the system interactively to retrieve video, audio, and key frames corresponding to events. We report results of performance evaluations of the algorithms and discuss issues related to sensing, analysis and retrieval of data.

Index Terms— Multimedia systems, Multisensor systems, image sensors, ubiquitous environments

1. INTRODUCTION

With recent advances in electronics and communication technologies and the availability of high computing power and storage at relatively low costs, the use of image sensor networks has become more feasible and affordable. As a result, there has been a recent growth in research related to applying sensor networks to home like environments, serving two main objectives. One aims at providing services to the people in the environment by detecting and recognizing their actions [1]. The other aims at retrieval of multimedia recorded in the environment [2].

In this paper, we present a distributed sensor network for event retrieval from continuously recorded multimedia data in a home-like environment. The objective is to design a system that can help the residents to *recall* things that were forgotten, *discover* things that were unknown to them, and *identify* their behavioral patterns.

2. ENVIRONMENT AND SENSOR NETWORK

This research is based on *NICT Ubiquitous Home* [3], a ubiquitous environment simulating a two room house (Figure 1). Ceiling-mounted stationary cameras acquire images at 5 frames per second and store them as JPEG files with a resolution of 320×240 pixels. Omni-directional microphones are installed in the middle of each room. The other microphones have cardioid response, and are directed towards the middle of the room/corridor that they are installed. Monophonic audio is sampled at 44.1 kHz from each microphone and recorded in *mp3* file format. The floor sensors are point-based pressure sensors with a resolution of 180×180mm. A sensor is considered to be in state '1' when the pressure on that sensor is above a preset threshold, and state '0' otherwise. The pressure on each sensor is sampled at 6 Hz and state transitions are recorded.

One day of continuous capture in ubiquitous home results in 408 hours of video and 600 hours of audio data. Automated retrieval is essential for efficient experience retrieval from such a large quantity of data.

A series of "real-life experiments" was conducted for data collection. In each experiment, a family lived in ubiquitous home for 1-2 weeks. The families lead their normal lives during this stay, leaving the house for work and other activities when necessary. The images, audio and sensor data were stored with timestamps for synchronization.

3. SYSTEM DESCRIPTION

Figure 2 shows an outline of the proposed system. Data from floor sensors, microphones and cameras are analyzed separately for retrieving different types of events. The data analysis is therefore distributed on the basis of the type of sensors. The results are written to a central relational database, which the users access through a graphical user interface by submitting interactive queries. The following subsections describe the system in detail.

3.1. Retrieval using floor sensor data

A footstep placed on and removed from a floor sensor results in a pair of consecutive state transitions, hereafter referred to as a *sensor activation*. An Agglomerative Hierarchical Clustering (AHC) algorithm [4] is used to segment the sensor activations into footstep sequences of different persons. Figure 3a is a visualization of this process. The grid corresponds to the floor sensors. Sensor activations that occurred later are indicated with a lighter shade of gray.

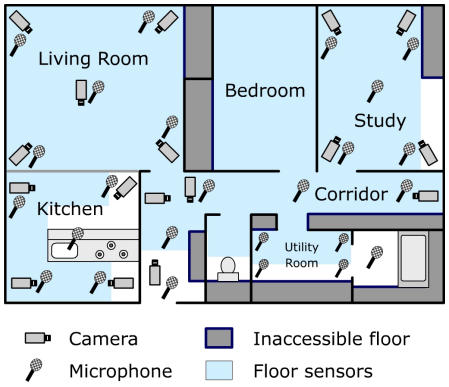

Figure 1. Ubiquitous home sensor layout.

For each footstep sequence detected by footstep segmentation, we intend to create a video clip keeping the corresponding person in view as he moves inside the house. A position-based video handover algorithm [4] is used to accomplish this task. Figure 3b shows how the cameras are selected using this algorithm for the footstep sequence in Fig. 3a. The viewable region of each camera is shown by a lighter shade of the same color. The color of the path indicates the camera that has been selected. Audio handover [4] is performed in similar fashion to ensure that the person is heard throughout the video clip (Figure 3c).

The video sequence constructed using video handover is summarized by extracting key frames. An adaptive spatio-temporal sampling algorithm based on the elapsed time, camera transitions and the rate of footsteps was selected after designing and evaluating several algorithms [4].

3.2. Retrieval using audio data

We start by eliminating audio corresponding to silence in each *region* (non-overlapping regions are defined as labeled in Figure 1). The result is a set of audio segments for situations where a sound was *heard* at a given region. A scaled template matching algorithm based on the sound energy distribution among regions is applied to remove sounds heard from other regions, achieving sound source localization at region level [5]. After localization, the sound segments are classified using a multilayer perceptron network based on time domain audio features, into 8 classes such as voices, footsteps, vacuum cleaner, and environmental sounds. After classification, consecutive segments of the same class with a gap less than 3 s between them are merged together to form larger segments, so that a single video can be retrieved for those sounds.

3.3. Issues related to image analysis

The video capturing system in ubiquitous home is designed with emphasis on storage and viewing rather than image analysis. The frames have low resolution, and high JPEG compression. The cameras use automatic gain control to ensure that the images have appropriate brightness for a human observer. However, this makes image sequence analysis extremely difficult. We attempted image analysis at different levels of complexity for multimedia retrieval from ubiquitous home, as described in the following subsections.

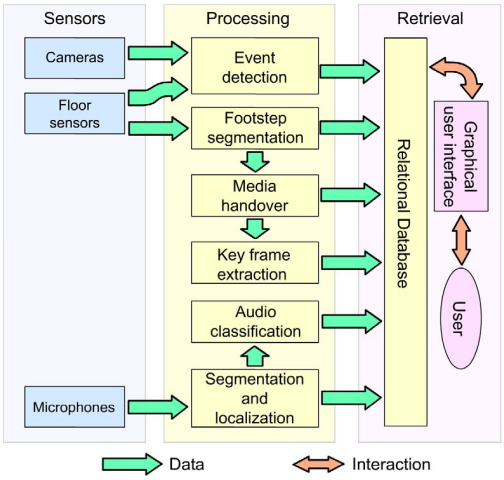

Figure 2. System overview.

3.4. Event detection using lighting changes

Lighting changes in a room, if combined with the scene context, can be used to identify significant events that take place in a house. Being very quick events, they can be used to create very short summaries of a day's events.

Lighting changes are relatively easy to detect as they are represented by sharp intensity level changes in image sequences. However, the problem is to find a single threshold level for this change that yields accurate event detection. For rooms or regions with windows, the amount of external light changes with the weather, curtains, and the time of day. For ubiquitous home, this task is made further complicated by automatic gain control in the cameras.

Our approach to solve this problem is to assign a rank of significance to each lighting change, based on its sharpness. The user selects the rank and browses the events through an interactive interface, thereby reducing the search space [6].

3.4. Detection of basic actions using image analysis

We attempt to detect a selected set of actions that take place in the living room, by analyzing images from multiple cameras. An algorithm based on image differencing in regions of interest is proposed for this task [7]. This approach is similar to the algorithm used by Jaimes et al. [8], but uses multiple cameras for robustness to occlusion.

Rectangular regions of interest are specified corresponding to the sofas and the wall-mounted television in the living room. The regions on different images corresponding to the same object are matched manually. Each region is modeled by the histogram of pixel intensities (hereafter referred to as the base histogram). Each object is modeled by a set of base histograms, corresponding to views from different cameras.

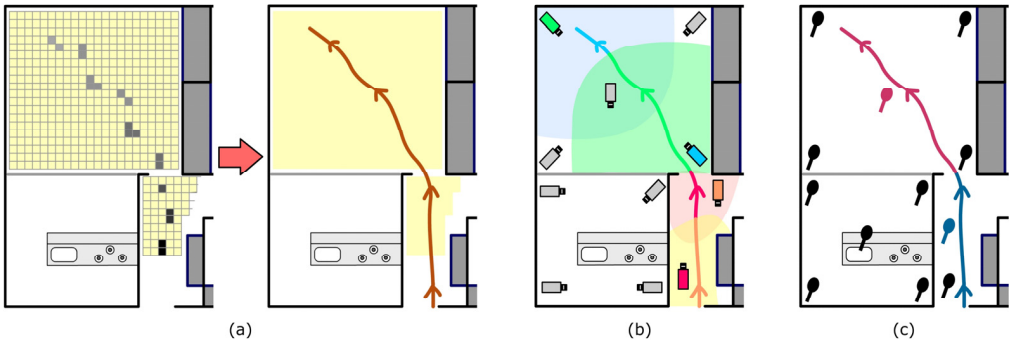

Figure 3. Personalized video retrieval using floor sensor data.

The following process is repeated for all frames during image analysis. The regions of interest are extracted and their intensity histograms calculated. Each histogram is compared with the corresponding base histogram. If the non-overlapping region of the two histograms exceeded a threshold of 50%, a candidate for an event for this particular region is registered. For example, if the region corresponds to a sofa the difference might have been caused by somebody sitting there. However, this event has to be verified by checking the corresponding regions of interest on images from the other cameras, to prevent false detections due to occlusion of objects. After verification, the interactions are identified using a set of heuristic rules.

Currently the proposed method is used to identify simple actions such as sitting, and turning the TV on and off. This can be expanded to identify more complex interactions by incorporating further processing on a sequence of images.

3.5. Face detection on image data

Face detection is prospective as a useful technique for video retrieval from ubiquitous home, owing to a number of reasons. The presence of the faces can be used as a cue to retrieve video or images for group activities. Better key frames can be extracted by ensuring that the face is visible in the selected key frames. Human-human interactions that are not captured by other sensors can be detected using the positions and orientations of multiple faces in an image. Face detection can act as a preliminary step for face recognition, making the system capable of person recognition.

However, due to the low resolution of images, the average size of faces in images from ubiquitous home is about 20x20 pixels. Poor quality of JPEG compression causes square blocks to appear on face regions (Fig. 4). Face detection on these images, therefore, is a difficult task.

Out of a number of face detectors, the Viola-Jones face detector [9] produced the best results on images from ubiquitous home. However, the accuracy of face detection was approximately 40%. There were two main reasons for low accuracy. The first was the presence of JPEG block on face regions, the edges around which tended to obscure the Haar-like features of a face used for detection. The second reason was that the average size of faces in the images was smaller than the size of those used for training the selected implementation of this face detector.

We made the following modifications to the face detection algorithm to obtain a higher accuracy. The images were enlarged in order to obtain larger faces and remove the edges around JPEG blocks. Three interpolation techniques, namely nearest neighbor, linear and cubic, were tried in order to select the best technique. The face detector was retrained using a data set consisting of the enlarged images, obtained from different cameras.

3.6. Integration and user interaction

The results of processing different types of data are stored in a relational database and indexed by date, time, location, person, and event types. These indices, decided based on a user study, facilitate integration of the results from processing that was hitherto distributed.

The user retrieves video, audio and key frames through a graphical user interface. The user interacts with the system using pointing gestures via a tablet monitor or a touch monitor. The interface is based on hierarchical timeline segmentation. The user starts by entering a day, upon which a summary of the day's activity is displayed along the time line. This is segmented in to one hour intervals, which are again partitioned to events detected using the algorithms described above. This allows the user to navigate easily within the large collection of data.

4. EVALUATION AND RESULTS

4.1 retrieval using sensor data

The individual algorithms were evaluated, defining accuracy measures where required. Accuracy of footstep segmentation was about 73%. The video clips created using video and audio handover were found to contain natural camera changes with adequate sound levels. Key frame extraction retrieved 80% of the most required key frames, according to a user study [4]. Sound source localization had an average

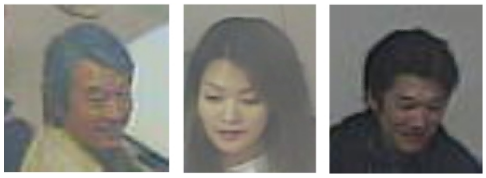

Figure 4. Faces extracted from ubiquitous home images.

precision of 92% and recall of 88%. Audio classification yielded an average accuracy of 83.2%.

Basic action detection had an accuracy of approximately 80% when evaluated on 3 hours of data with 3 residents. Table 1 presents the results of face detection for different interpolation techniques. Cubic interpolation yields best results, but the improvement compared to linear interpolation is marginal. Face detection was less accurate in the corridor than other regions of the house. This was partly due to the low level of illumination in the corridor. However, the main reason was the presence of motion blur in images, as the residents usually keep moving when in the corridor.

4.2 User study

A user study was conducted for evaluating the usability of the overall system. In this study, one of the families that took part in a real life experiment used the system to retrieve video and key frames for a 6 hour period from their stay. After using the system for about 45 minutes, the users were able to recall the events that took place, and discovered a few unknown events. They found the system helpful, and wished to keep some of the retrieved media.

5. DISCUSSION: DISTRIBUTED PROCESSING FOR REAL-TIME INTEGRATION

At the current state, the sensors in ubiquitous home capture data continuously and the data are analyzed offline. The reason is that our work has been carried out in a remote location from ubiquitous home. Furthermore, the data are used for a related project on continuous archival and retrieval of personal experiences.

As shown in Figure 2, the capture of different types of data is performed independently. There are opportunities for real-time distributed processing, applicable at two levels. The first level is parallel processing of sensor data, in which different sensor data streams are analyzed in parallel and indices are produced. The indices can later be integrated using the timeline. The second level is collaborative processing of sensor data, which can achieve not only real-time processing, but also a significant reduction of data storage. For example, the system can use the data from floor sensors to trigger cameras, thereby reducing the redundancy of the captured video data.

Table 1. Performance of face detection.

Method of interpolation	Precision	Recall	F-measure
Nearest Neighbor	0.77	0.81	0.79
Linear	0.92	0.87	0.89
Cubic	0.94	0.91	0.92

6. CONCLUSION

We have implemented a system for retrieval of events in a home using a network of cameras, microphones and pressure based floor sensors. The floor sensor data were analyzed using unsupervised data mining techniques for creation of personalized video clips, key frame extraction, and activity classification. Audio segmentation and classification enabled video retrieval for audio events of different categories. Image analysis at different levels of complexity facilitated detection of selected actions and events. Face detection yielded an accuracy of approximately 90%.

7. ACKNOWLEDGMENTS

This work was supported in part by JST of Japan. We thank Mr. Atsushi Omiya and family for their cooperation.

8. REFERENCES

[1] Philips Research. Ambient Intelligence: changing lives for the better. http://www.research.philips.com/technologies/syst_softw/ami/background.html, Philips Electronics N.V., 2005.

[2] Mori, T., Noguchi, H., Takada, A., Sato, T. "Sensing Room: Distributed Sensor Environment for Measurement of Human Daily Behavior", *Proc. INSS2004*, p.40-43, 2004.

[3] Yamazaki, T. Ubiquitous Home: Real-life Testbed for Home Context-Aware Service. *Proc. Tridentcom2005*, 2005, 54-59.

[4] Gamhewage C. de Silva, T. Yamasaki, K. Aizawa, "Evaluation of Video Summarization for a Large Number of Cameras in Ubiquitous Home", *Proc. ACM Multimedia 2005*, pp. 820-828, 2005.

[5] Gamhewage C. de Silva, T. Yamasaki, K. Aizawa, "Sound Source Localization Based on Energy Distribution Template Matching for a Ubiquitous Environment", submitted to *IEEE Transactions in Multimedia*.

[6] Gamhewage C. de Silva, T. Yamasaki, K. Aizawa, "Interactive Experience Retrieval for Ubiquitous Home", *Proc. ACM CARPE 2006.*

[7] Gamhewage C. de Silva, T. Yamasaki, K. Aizawa, "Experience Retrieval for Ubiquitous Home", *Proc. ACM CARPE 2005, p.35-44*.

[8] Jaimes, A. Omura, K., Nagamine, T., Hirata, K. Memory Cues for Meeting Video Retrieval. *CARPE 2004*, 74-85.

[9] P. Viola and M. J. Jones, "Robust real-time face detection," Int. J Computer Vision, vol. 57, no. 2, pp. 137–154, 2004.