論文内容の要旨

論文題目

Multimedia Experience Retrieval in a Ubiquitous Home
（ユビキタスホームにおける体験情報処理と検索）

氏名　デシルヴァ　ガムヘワゲ　チャミンダ

Automated capture and retrieval of multimedia experiences at home is interesting due to a number of reasons. A system with such capability can help the residents by *capturing* experiences that the residents do not want to be away from, for the sake of shooting photos or video. It can "entertain" the residents by allowing them to *recall* the happy moments and experiences, and *discover* things that were unknown to them. It can act as a "memory-aid" or a "healthcare assistant," thereby making life more comfortable for the elderly. If used over a long period of time, it can also help the residents to *identify* their behavioral patterns and take corrective action if necessary.

However, this is a difficult task with several challenges in different aspects. The number of sensors required for complete capture of experiences taking place in a home-like environment is quite large. Continuous capture is necessary to prevent missing experiences that residents are not prepared for, resulting in a large amount of multimedia content that is much less structured compared to those from any other environment. Recognition of actions, events and experiences using such data is extremely difficult. Queries for retrieval will be at high semantic level, and at different levels of granularity; a resident might just want to find out the number of visitors to the house on a certain day, or want to see the video of what he was

doing during the afternoon of a selected day. Different places of the home have different levels of privacy, restricting the ability to capture certain types of data in some locations.

In this research, we focus on capturing and retrieval of personal experiences in a ubiquitous environment that simulates a house, with the objective of creating an electronic chronicle that enables the residents to retrieve the captured video using simple and interactive queries. A large number of cameras and microphones are used to continuously record video and audio at desired areas of the house. Pressure based sensors, mounted on the house floor, record context data corresponding to the footsteps of residents. A given region of the house may contain none, some or all of these types of sensors, depending on the level of privacy in that region. One day of continuous capture in this house results in 408 hours of video and 600 hours of audio data, which amounts to about 500 GB of disk space, suggesting that manual retrieval is impossible.

Our approach in this work is to select sources that convey the most amount of information based on context data. Only the selected sources are queried to retrieve data, and these data are analyzed further for retrieval thereby minimizing the computational effort on content analysis. However, at the same time, the redundancy caused by the presence of a large number of sensors is utilized to improve the accuracy of retrieval.

Data from floor sensors are clustered using a hierarchical approach to segment footstep sequences of different persons. Algorithms for automatic video and audio handover are used for the creation of video clips using these sequences, while automatically changing cameras and microphones to keep the person in view and hear the sounds in his/her surroundings. Key frames are extracted from these videos the create summaries, allowing the users to get a quick preview of their content. An adaptive algorithm based on the time and location, and the rate of activity of the person is used to create complete and compact summaries.

Audio data from each microphone are segmented at two levels for retrieving audio events. First, data corresponding to silence and small noises are removed. Thereafter, sounds heard from regions other than where the microphone is located are removed using a sound source localization algorithm. The resulting audio segments are classified into different categories of sounds, to retrieve the sounds and video showing the locations where the sounds are heard.

Basic analysis of image data is used for the detection of selected events that take place inside the house. Floor sensor data are analyzed in combination with other sensory modalities, for recognition of some common actions inside the house. The results are written to a central a relational database, where they can be fused for accurate detection of activities.

The users, who are also residents, retrieve their experiences from the database through a graphical user interface by submitting interactive queries. This interface is designed based on the concepts of hierarchical media segmentation and Interactive retrieval, to facilitate effective retrieval with a minimal amount of manual data input using only a pointing device. Visualizations of different types of data at various levels of detail were included to help the user to retrieve required media and understand the results.

We evaluated the system using a two-pronged approach. Each functional component was evaluated individually, to ensure that it provides accurate results to the user and the other components using the results. We used standard accuracy measures and experiments where available, while designing experiments and defining new accuracy measures where necessary. We also conducted a user study for the purposes of gathering system requirements and evaluating the overall system. A set of "real-life experiments," in each of which a family actually lived in the house for a period of 7-14 days, were conducted for data collection. One of these families took part in a user study, where suggested system requirements, used the system for retrieving their experiences, and provided feedback.

Segmentation of floor sensor data followed by video handover enabled the creation of personalized video clips using a large number of cameras. It was possible to dub this video with reasonably good quality, using audio handover. Adaptive key frame extraction enabled retrieval of more than 80% of the key frames required for a complete summary of the video. Silence elimination and false positive removal from audio data produced results with a high accuracy of 98%. The scaled template matching algorithm we proposed is able to achieve localize sound sources with an accuracy of about 90%, despite the absence of microphone arrays or a beam-forming setup. The accuracy of audio classification using only time domain features is above 83%. Basic image analysis facilitated detection of events that are useful in understanding the activities that take place inside the house. Action detection using multiple sensory modalities yielded an average accuracy of approximately 78%.

The residents who evaluated the system found it useful, and enjoyed using it. They discovered events that they were not aware of before using the system. The residents wanted to keep some of the video they were able to retrieve, demonstrating the system's applicability. They found the system easy to learn and usable. The requirements they identified and the feedback they provided were valuable in improving the system.