LK_011

Sound Source Localization for Multimedia Retrieval in a Ubiquitous Environment

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

Abstract

We present a system for video retrieval based on analyzing audio data from a large number of microphones in a ubiquitous home. Silence elimination on individual microphones is followed by noise reduction based on regions consisting of multiple microphones, to identify audio segments. An algorithm based on the energy distribution of sounds in the house is used to localize sound sources, thereby removing sounds heard in regions other than they were generated. The algorithms were evaluated with 200 minutes of audio data from each of the microphones, gathered during an experiment during which a family lived in the ubiquitous home. It was possible to achieve an overall accuracy of above 80% from all algorithms.

1. Introduction

Audio analysis is a widely used approach for video retrieval, indexing and summarization due to a number of reasons. Audio has lower dimensionality, facilitating faster processing with less computing power. In some categories of video, such as news and sports video, there is a high degree of correlation between sounds and actual events. In case of studio-dubbed video, the high signal-to-noise ratio of audio results in more accurate results. The common approach for audio-based video retrieval is to divide the audio stream into a set of segments corresponding to different types of events, and use the segments as an index for video retrieval [1] [2].

Audio analysis can facilitate more efficient video retrieval in ubiquitous environments. Since most ubiquitous environments require multiple cameras for complete capture of the scene, the amount of video to be analyzed is quite large. Audio data can be successfully used to reduce the search space by selecting cameras according to the results of processing audio. Automated video surveillance [3], meeting video summarization [4] and tele-monitoring of patients [5] are some of the applications where audio analysis has been employed successfully.

However, audio analysis for a ubiquitous environment is more difficult than that for broadcast video. Multiple microphones are often required for capturing audio with adequate signal-to-noise ratio. Since the coverage of a microphone is much less restricted than that of a camera, sounds from a single source are picked up by several microphones, calling for the additional task of source localization. The most common approaches for this are based on time delay of arrival (TDOA), microphone arrays [6] and beam-forming techniques [7]. However, the conditions required by these approaches are difficult to satisfy in most ubiquitous environments.

The rest of this paper presents our work on sound source localization for video retrieval from a home-like ubiquitous environment with a large number of cameras and microphones. We introduce the ubiquitous environment and describe the algorithms used for sound source localization. We evaluate the performance of the algorithms and present the results. Finally we conclude the paper with suggestions for possible future directions.

2. Ubiquitous Home

This work is based on *Ubiquitous Home* [8], a two-bedroom house equipped with 17 cameras and 25 microphones for continuous capture of audio and video. Pressure based sensors, mounted on the floor, capture data corresponding to the footsteps. Cardioid and omni-directional microphones are mounted on the ceiling at locations shown in Figure 1a. The numbering of the microphones in Figure 1a will be used to refer to them in the coming sections of the paper. Figures 1b and 1c show the directional responses of omni directional and cardioid microphones respectively. Audio data from each microphone are captured at a sampling rate of 44.1 kHz.

Our goal in research related to ubiquitous home is to create a multimedia diary that the residents can browse in an efficient manner. At the current state, it is possible to retrieve personalized video and key frame sets using automated camera and microphone selection, and achieve basic activity classification by analyzing floor sensor data [9]. In this work, we extend the capability of the diary by incorporating audio analysis. Audio data can be used both to supplement video retrieval in the absence of floor sensor data, and to support queries related to events such as conversations.

3. System Description

The audio streams are partitioned into *segments* of one second each. Sets of audio segments that are captured during the same one-second interval (hereafter referred to as *segment sets*) are processed together. After eliminating silence, the segments are processed further to reduce noise. Source localization is applied on the resulting segments to identify the location(s) of the sound sources. The results are stored as indexes for video retrieval. The following sub-sections describe these steps in detail.

3.1 Silence elimination and noise reduction

An algorithm developed in our previous work [10] was used for silence elimination and noise reduction. We perform silence elimination for a single audio segment by comparing its RMS power against a threshold value. The threshold for each microphone was estimated by analyzing one hour of audio data for silence for that microphone, extracted from different times of day. Each audio segment is partitioned into *frames* having 300 samples. Adjacent frames had a 50% overlap. The root mean square (RMS) value of each frame is calculated. If the calculated RMS value is larger than the threshold, the frame is considered to contain sound. Sets of contiguous sound frames having duration less than 0.1s are removed. Sets of contiguous sound frames that are less than 0.5s apart are combined together to form single segments.

Data from multiple microphones in close proximity are used to reduce false positives resulting due to noise. For this purpose and for use in the following stages, the microphones are grouped into *regions* as specified in Table 1. A region-based voting algorithm based on the small duration and randomness of noise is used to reduce noise in sound segments.

Silence elimination resulted in 0% false negatives (sound

Table 1. Assignment of microphones to regions

Region	Label	Microphones
Living room	LR	1-5
Study room	SR	6-10
Kitchen	KT	11-15
Corridor – region 1	C1	16,17
Corridor – region 2	C2	18-20
Utility room	UR	21-25

† Dept. of Frontier Informatics, the University of Tokyo

‡ Dept. of Electronics Engineering, The University of Tokyo

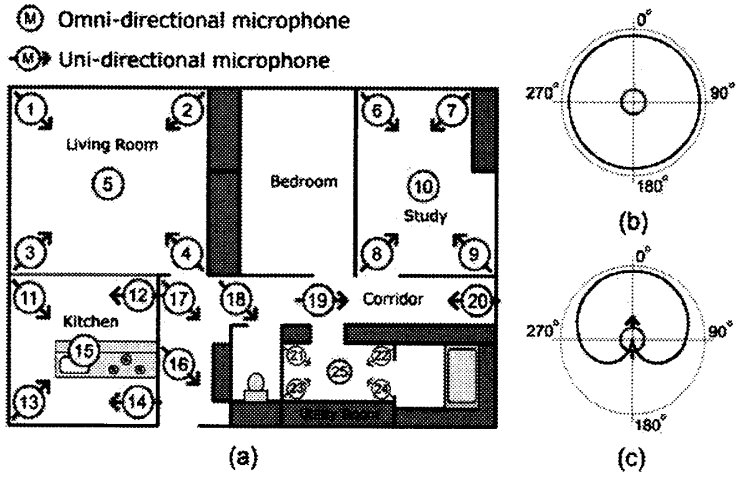

Figure 1. Microphone layout in Ubiquitous Home

misclassified as silence) and 2.2% false positives (silence misclassified as sound). The algorithm for noise reduction was able to remove 83% of the false positives that remained after silence elimination in individual audio streams.

3.2 Sound source localization

The sounds contained in the segments can be categorized into two types. One is *local* sounds, that is, sounds generated in the same region as the microphone belongs to. The other, *overheard* sounds, refers to the sounds that are generated in a region other than that the microphone belongs to. Each segment can contain either or both of these types. To prevent false retrievals, segments with only overheard sounds must be removed before further processing. We refer to this task as *sound source localization*, as it identifies the regions where one or more sound sources are present. With only a limited amount of directivity present in the setup of microphones, source localization for ubiquitous home is a difficult task.

3.3 Localization based on maximum energy

A simple approach for sound source localization is to select the region with the microphone that captures the sound with the highest volume [5]. Although this approach fails when multiple sound sources are active at the same time, we investigate its performance for comparison. We use the following algorithm for *localization based on maximum energy*.

For each segment set, the mean square value (which is proportional to short-term energy) of the samples in each segment is calculated. The average energy for each region is calculated by averaging mean square values for the segments from the microphones in that region. The region with the maximum average energy is selected as the location of the sound source.

4.4 Energy distribution template matching

A sound generated in one region of ubiquitous home can be heard in other regions, in different levels of intensity. Based on this fact, we model local sounds in each region with the variation of energy received by each microphone. For each region r, a number of segment sets are selected for instances where sounds were generated only in that particular region. The energy distribution $E(n)$ for each segment set is determined by calculating the energy for each segment in the set. The *Energy Distribution Template*, T_r of the region r is estimated by averaging all such energy distributions and normalizing to the range [0 1]. Figure 2 shows the energy distribution template for sounds originating in the living room.

Each audio segment set is a mixture of sounds from one or more regions of the house. We hypothesize that the energy distribution of a segment set is a linear combination of one or more energy distribution templates, and attempt to identify them. We use the following *scaled template matching algorithm* for this task. The main idea behind the algorithm is to repeatedly identify the loudest sound source available, and remove its contribution. This is repeated until there is no significant sound energy left to assume the presence of a sound source.

1. Calculate the energy distribution, $E(n)$ for the current segment set
2. Find the region r in the distribution with the maximum average value
3. Scale $E(n)$ by dividing by the max value, A_m in that region
4. Subtract the template T_r corresponding to that region, and obtain $G(n)$. Multiply by A_m to obtain $E'(n)$
5. If the average value of $E'(n) < 0.2$ then stop.
6. Repeat steps 2-5 on $E'(n)$, for k times where k is the number of regions where sound segments are detected after noise filtering.

4.4 Video Retrieval

After source localization, the sound segments from the same region are combined if they are less than 5 seconds apart. For the resulting segments, video is retrieved from all cameras capturing video from the corresponding region. Audio is retrieved from the microphone closest to the center of the region. All videos are played simultaneously, together with the audio segment. Each video segment starts 3 seconds before the starting time of the sound segment, as a sudden start of audio is annoying for the viewer in most cases.

4. Evaluation and Results

We used audio data captured from 7:45 a.m. to 11:05 a.m. on a day of the real-life experiment, for evaluation of sound source localization. The time interval was selected such that that all regions of the house were used for a considerable duration for ordinary household activities within this interval. The data were transcribed manually to find out the ground truth with regard to the sounds generated in each region.

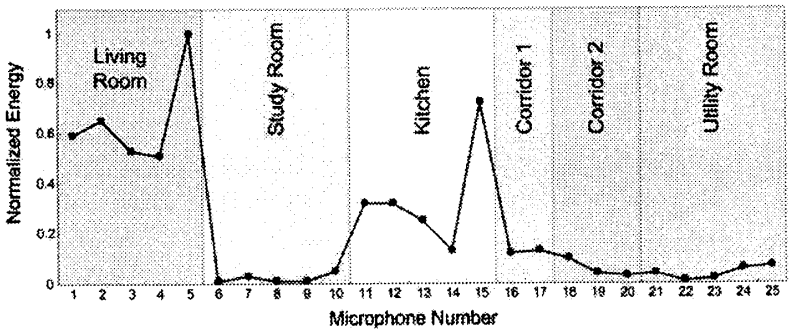

Figure 2. Energy distribution template for the living room

In source localization, it is important to minimize both retrieving overheard sounds as local sounds, and ignoring local sounds as overheard sounds. We define the Precision P, Recall R, and Balanced F-measure F of source localization for each region of the house, as:

$$P = N_L / N_T$$

$$R = N_L / N_A$$

$$F = 2PR/(P+R)$$

where

N_L = no. of local segments retrieved correctly

N_T = total no. of segments retrieved

N_A = actual no. of local segments

Table 2 compares the precision, recall, and F-measure for audio segments before source localization, results obtained by maximum energy-based selection, and results obtained by scaled template matching. It is evident that the scaled template matching has the best performance, although the accuracy is relatively low for some regions.

5. Conclusion

We have proposed a set of algorithms for analyzing audio data from a large number of microphones in a home-like ubiquitous environment for video retrieval. The scaled template matching algorithm is able to achieve generally accurate sound source localization, despite the absence of microphone arrays or a beam-forming setup.

Acknowledgments

We thank the staff at NICT Keihanna Info-Communications Research Laboratories, for their support.

Table 2. Performance of sound source localization

	Before Localization			Max. Energy			Temp. Matching		
Region	P	R	F	P	R	F	P	R	F
LR	0.56	1.00	0.72	0.97	0.93	0.95	0.95	0.99	0.97
SR	0.78	1.00	0.88	1.00	0.45	0.62	0.96	0.60	0.74
KT	0.72	1.00	0.84	0.96	0.78	0.86	0.97	0.79	0.87
C1	0.69	1.00	0.82	1.00	0.60	0.75	0.86	0.89	0.88
C2	0.46	1.00	0.63	1.00	0.80	0.89	0.84	0.97	0.90
UR	0.71	1.00	0.83	1.00	0.82	0.90	0.91	0.89	0.90

References

[1] Y. Wang, Z. Liu, J. Huang: Multimedia Content Analysis Using Both Audio and Visual Cues, IEEE Signal Processing Magazine, November 2000. 12-36

[2] R. Cai, L. Lu, A. Hanjalic: Unsupervised Content Discovery in Composite Audio: Proc. ACM Multimedia 2005, Singapore. (2005) Page 628-637

[3] A.F. Smeaton, M. McHugh: Towards Event Detection in an Audio-Based Sensor Network, Proc. ACM VSSN'05, Singapore. (2005) 87-94

[4] D. Gatica-Perez, D. Zhang, S. Bengio: Extracting Information for Multimedia Meeting Collections, Proc. ACM MIR'05, Singapore. (2005) 245-252

[5] M. Vacher, D. Istrate, L. Besacier, E. Castelli, J. F. Serignat: Smart audio sensor for telemedicine, Smart Objects Conference 2003, France. (2003)

[6] J. F. Chen, L. Shue, H. W. Sun, K. S. Phua: An Adaptive Microphone Array with Local Acoustic Sensitivity, in Proc. IEEE ICME 2005, The Netherlands. (2005)

[7] Hoshuyama, A. Sugiyama, A. Hirano: A robust adaptive beam former for microphone arrays with a blocking matrix using constrained adaptive filters, IEEE Trans. Signal Processing, vol. 47, no. 10, pp. 2677-2684, Oct. 1999.

[8] Ubiquitous Home: http://www.nict.go.jp/jt/a135/eng/research/ubiquitous_home.html, National Institute of Information and Communication Technology, Japan.

[9] G. C. de Silva, T. Yamasaki, and K. Aizawa: Experience Retrieval in a Ubiquitous Home, Proc. ACM CARPE 2005, Singapore. (2005) 35-44

[10] G. C. de Silva, T. Yamasaki, K. Aizawa: Creation of an Electronic Chronicle for a Ubiquitous Home: Sensing, Analysis and Evaluation, Proceedings of the IEEE eChronicles Workshop, USA. (2005) 35-44