Integration of Multiple Sensors using a Generic Sensor Management Framework

Projektleiter
Prof. Dr. Burkhard Igel

Forschungsschwerpunkt
Process Improvement & CAQ

Zeitraum
2006 – 2008

Kooperation
Curtin University of Technology Perth, Australien

Förderung
DFG – Deutsche Forschungsgemeinschaft, Bonn

Kontakt
Prof. Dr. Burkhard Igel
Fachbereich Informations- und Elektrotechnik
Fachhochschule Dortmund
Sonnenstraße 96
44139 Dortmund
Tel.: (0231) 9112-357
E-Mail: igel @fh-dortmund.de

Abstract
Through the fact that security has become a major interest, intensive research has been done to develop surveillance systems to monitor public areas or areas with restricted access. To improve the performance and reliability of such systems the trend is to combine and integrate different sensor types to cover the weakness of one sensor by another type of sensor. This research proposes a generic sensor framework which supports different types of sensors. Our focus is the development of a joint-sensor calibration technique that uses audio and/or visual observations to improve the calibration process. One significant feature of this approach is the ability to check and update the calibration status of the sensor suite continuously, making it resilient to independent drift in the individual sensors and the natural environment.

1 Objectives

The aim of this research is to investigate a new approach for providing data of multiple sensors for tracking moving objects using a generic sensor management framework. It is postulated that this framework can handle different sensors without knowing their exact spatial location.
Key issues that this research intends to address are as follows:

- Learning the relationship between data of different sensors and using this knowledge for calibration issues.
- Designing and implementing a generic sensor management framework which is able to adapt itself to changes in the sensors alignment or environment parameters (i.e. when camera are slightly displaced or when temperature changes occur which affects the calibration of the audio sensor).
- Detection and tracking of objects using the joint sensor management framework.

2 Significance

The most significant aspect of this proposed research is the fact that it endeavours to create a generic sensor management framework that deals with multi sensor information of the same or different type of sensor. Such a framework is responsible to calibrate all sensors adaptively online through a feedback design between calibration and tracking. This is very important because sound source localisation is sensitive to environment parameters like air pressure or air temperature which cannot be assumed as constant. Another feature of this framework will be the ability to control the sensors like the pan and tilt angle of a PTZ camera. This also gives the system, once it has calibrated itself, the capability to steer an audio beam to a specific object based on the video information. To our knowledge no one has previously attempted to develop such an integrated sensor framework.

The stationary camera builds the base to which all additional sensors will be calibrated because the image of this camera represents the whole scene which is observed by the system. The novel idea in this framework is the feedback between calibration and tracking through the control unit. That is the key why the framework should calibrate itself and react dynamically to any change that influence the tracking performance. For instance, if the system is very confident that there is one object at the observed scene and this object is detected by the stationary camera and the audio sensors it can compute the special correlation between the sensors.

Work by M. J. Beal and N. Jojic [3] has covered the successful combination of a static camera and two microphones through a graphical model. Here the relative time delay between the microphone signal and the special shift of the object in the camera image is used to calibrate the sensors specially. But this proposed method is not able to interact with the sensors and doesn't use tracking information to verify the accuracy of the calibration.

Other work from Dmitry N. Zotkim et al. [4] used a sequential Monte-Carlo algorithm, also known as particle filter, to track objects as well as a way to calibrate the sensors relative positions. They proposed to include the intrinsic system parameter into the system space of the particle filter. The drawback of their proposed system is that the camera calibration is still performed offline by using a known calibration object.

3 Methodology

This section outlines the strategy of the investigation of the generic sensor management framework. The basic step is to develop acquisition components for audio and video sensors on a windows system. A promising technology for this task is DirectX which performs synchronised audio capturing for multi channels and parallel video capturing and rendering from multiple cameras with a minimal use of system performance.

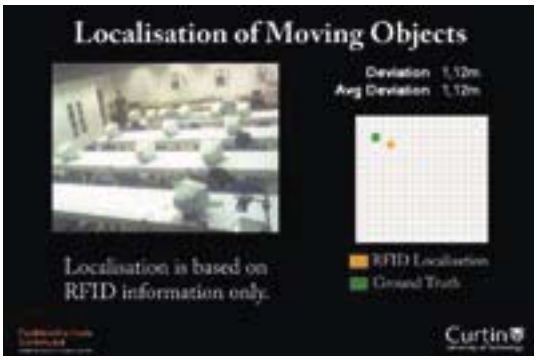
When the capturing is done, algorithms for object detection in a video stream will be investigated. In terms of the static camera a background subtraction algorithm will perform object detection. The information from the static camera is then used to detect the object into the PTZ camera by using colour information or template matching algorithms. The next component which is needed for the sensor framework is tracking. Therefore a particle filter will be developed. A variety of different types of particle filter algorithm has been proposed in literature. This research will use a Sequential Important Sampling (SIS) algorithm [1, 2] to perform the tracking task. At this stage it should deal only with objects extracted from a video stream.
After the system has the capability to detect and track object in the video stream, a sensor framework can be developed which tries to calibrate the cameras spatially. In this stage it is important that the framework is designed to cope with any arbitrary sensor suite.

Once the sensor framework can handle video data, algorithms will be developed for sound source localisation. This piece of work will be done

in conjunction with WATRI (Western Australian Telecommunications Research Institute). For the microphone arrangement a single linear array will be chosen at the beginning.

The final step will be to include the SSL into the sensor management framework to perform a calibration between all sensors. After the calibration is done, the framework should expand to steer an audio beam on a single object to record an enhanced audio signal based on the tracking results.

4 Ethical Issues
Due to the nature of this research, it will be necessary to record video and audio data for further processing. All subjects will be adult volunteers



who granted their permission that their image and sound samples will be recorded, stored and analysed via a signed statement subsequent to a full explanation by these researchers of:
- Exactly what recordings are required and how they are created
- The use of the recordings in the analysis
- Where the recordings will be stored and who will have access to them

Personal or confidential data about each subject is not required for the experiment, and so will not be requested, stored or used. The work will not involve the participation of minors, people with intellectual or mental impairment, persons highly dependent on medical care, people who are in a dependent or unequal relationship or collectives. The work will not involve medical trials of any nature, nor will it require the deception of its participants.

References:

[1] Arulampalam, M. S.; Maskell, S.; Gordon, N.; Clapp T. “A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking” IEEE TRANSACTIONS ON SIGNAL PROCESSING, FEBRUARY 2002

[2] Zia Khan; Balch, T.; Dellaert, F. “Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model“ Proceedings. 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 2003

[3] Beal, M.; Jojic. N. “A Graphical Model for Audiovisual Object Tracking” IEEE Transactions on Pattern Analysis and Machine Intelligence, July 2003, pp. 828- 836

[4] Zotkin, D.; Duraiswami, R.; Davis, L. “Joint Audio-Visual Tracking using Particle Filter” EURASIP J. Appl. Signal Processing, 2002, pp. 1154-1164

[5] Kuehnapfel, Th.; Tele, T.;Venkatesh, S.; Lehmann, E. “Calibration of Audio-Video Sensors for multi-modal event indexing”, MMSP-P1.5, ICASSP - IEEE Int. Conf. On Acoustics, Speech and Signal Processing, Hawai’i April 2007

[6] Kuehnapfel, Th.; Tele, T.; Venkatesh, S.; Nordholm, S.; Igel, B.; “Adaptive Speech Enhancement with Varying Noise Backgrounds”. 19th Intern. Conf. on Pattern Recognition, Dec 8-11, 2008, Tampa, FL, USA

recent results submitted to ICPR 2008
(International Conference on Pattern Recognition)

for further reading see ICPR 2008

Adaptive Speech Enhancement with Varying Noise Backgrounds

Thorsten Kühnapfel [1], Tele Tan [1], Svetha Venkatesh [1], Sven Erik Nordholm [2], Burkhard Igel [3]
[1] Dept. of Computing, Curtin University of Technology, Western Australia
[2] Western Australian Telecommunications Research Institute, Western Australia
[3] Dept. of Information Tech. and Electrical Eng., University of Applied Sciences Dortmund, Germany

Abstract

We present a new approach for speech enhancement in the presence of non-stationary and rapidly changing background noise. A distributed microphone system is used to capture the acoustic characteristics of the environment. The input of each microphone is then classified either as speech or one of the predetermined noise types. Further enhancement of speech in respective microphones is carried out using a modified spectral subtraction algorithm that incorporates multiple noise models to quickly adapt to rapid background noise changes. Tests on real world speech captured under diverse conditions demonstrate the effectiveness of this method.

1 Introduction

Audio sensors provide an effective and low cost way of measuring the sonic activity in and around places of interest. Whether used on its own, or with other sensors (CCTV, motion sensors), the quality of the sound captured in the presence of interfering noise signals is an important issue. To compensate for ambient noise, multiple microphones can be used to leverage on the detection and localisation of the speaker or speakers. One technique is to use beam steering or blind source separation to isolate the source of the speaker/s. However, the main drawback of this approach is the inability to work with sources that are not close to the microphone. One may consider distributing the microphone arrays in the area of interest, but this is costly. This paper explores the use of an alternative method of using several microphones to capture the noise and use the closest microphone to enhance speech signals.

For single microphone speech enhancement, spectral subtraction filters are commonly used, in which a single estimation of the background noise is made. There are different techniques to model the background; heuristically [5, 2, 9, 10, 6] by averaging out the recorded noise sequences or statistically by modelling each background coefficient as Gaussian random variables [4]. Recent work has focused on improving the filter coefficients [2] or investigating various smoothing techniques [9] to minimise speech distortion, known as "musical noise". The authors in [10] use a modified spectral subtraction approach with a low resolution gain function which is smoothed over time. The filter proposed in [6] divides the signal in different frequency bands and uses a weighting function to adjust the subtraction factor for each sub band. The limitation of using only one background model is that in a real world situation, the noise source can change rapidly and all the above methods require time to adapt.

To successfully use a spectral subtraction algorithm, voice activity detection (VAD) is essential. Martin [7] has reported a fast and effective algorithm for estimating the SNR based on short time power estimation. A disadvantage of this approach is that noise intensity estimation is sensitive to outliers and can lead to false detections. A more sophisticated VAD method is proposed in [8]. It uses a statistical approach to compare the second-order statistics of the test signal to the speech model. This is, however, complex and computationally expensive and has only been tested on stationary artificial or slightly non-stationary helicopter noise.

The aim of the proposed system is to provide sonic surveillance for an area of interest. As the size of this area increases it will be more economical to use a *network of distributed microphones*, instead of several microphone arrays. Within this network, the closest microphone to the source is used to enhance the speech signal, whereas the other microphones are used to capture and classify the noise source. The system first performs speech/non-speech classification using a new voice activation detection algorithm on all microphone inputs independently. For the non-speech segments, further classification into a predetermined list of background models is carried out. This classification result is used to provide the appropriate noise model to enhance the quality of the classified speech segments. The VAD is also used to set the parameters of the spectral filter. The experiments show the reliability of the noise classification and speech detection, in presence of real, non stationary background noise. The enhanced speech is then evaluated by a group of 11 people.

The novelty of the proposed system is that different noise models are constructed for the spectral subtraction algorithm, and thus the system rapidly adapts

to changes in typical noise sources in the environment. This is in contrast to other systems that need time to adapt to noise changes. Therefore, it can be used to increase the performance of any current spectral subtraction algorithms. Additionally, the noise classification is incorporated within the VAD to minimise false detection of speech.

2 Methodology

The proposed system consists of three sub modules: Noise classification, voice activity detection and speech enhancement. The noise classification is done by matching the features of the noise models with the input signal y. This correlation value is also used to enhance the reliability of the voice activity detection (see figure 1). Speech enhancement is then performed by combining the noise classification result of several microphones in conjunction with voice activity detection.

2.1 Noise Classification

For noise classification, the recorded signal $y_k(i)$ is transformed into the frequency domain $Y_k(i, f)$ via the fast Fourier transformation, where i is the time block index, f the frequency index and k is the index of the microphone. $|Y_k(i, f)|$ is then scaled by a Mel scale triangular filter [3] to obtain the final feature set $S_k(i, p)$, where p is the Mel scale filter index. In the initial learning process, a mean noise model $\overline{N}_k^q$ for microphone k and noise type q ($q : 1...Q$) is computed as

$$\overline{N}_k^q(i = 0, p) = \frac{1}{\tau + 1} \sum_{j=i}^{i+\tau} S_k(j, p) \tag{1}$$

where τ is the numbers of time blocks used. The classification decision η_k for microphone k is made by computing a normalised cross correlation coefficient c_k^q between each noise model $\overline{N}_k^q$ and the current feature set S_k and is computed as:

$$\eta_k(i) = \underset{q}{\operatorname{argmax}} \{c_k^q(i)\} \tag{2}$$

If the correlation coefficient c_k^q of the classified noise η_k is above a threshold (> 0.95), it indicates a single noise source and the corresponding noise model is updated as

$$\overline{N}_k^q(i, p) = (1 - \rho)\overline{N}_k^q(i - 1, p) + \rho S_k(i, p) \tag{3}$$

where ρ is an exponential updating factor.

Let l^q be the count of noise type q across the K microphones in which no active speech is detected. Then the final classification of the overall noise type is based on $\operatorname{argmax}_q \{l^q\}$.

2.2 Voice Activity Detection

The final VAD is based on two features, the signal power, P_{Y_k}, and the correlation coefficient c_k^q of the detected noise η_k of microphone k. Figure 1 shows the generic representation of the voice activity.

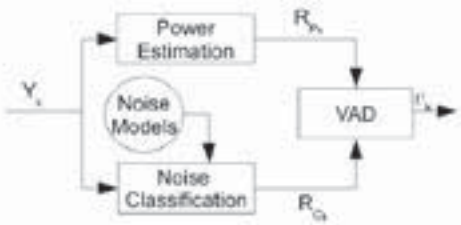


Figure 1. Generic model of voice activity detection

The signal power P_{Y_k} is fast to computed, but has the disadvantage of false detection if the noise intensity changes quickly. Therefore, the computed correlation coefficient c_k^q is used to reduce such errors as follows. Speech sequences are detected based on $a_k(i)$, computed as

$$a_k(i) = R_{P_k}(i) + R_{c_k}(i) \tag{4}$$

with

$$R_{P_k}(i) = \begin{cases} 1 & , \text{if } P_{Y_k}(i) \geq T_{P_k} \\ 0 & , \text{if } P_{Y_k}(i) < T_{P_k} \end{cases} \tag{5}$$

$$R_{c_k}(i) = \begin{cases} 1 & , \text{if } c_k^{q^*}(i) < T_c \\ 0 & , \text{if } c_k^{q^*}(i) \geq T_c \end{cases} \tag{6}$$

where T_c is the static threshold for classification and T_{P_k} is the dynamic threshold for the signal power. T_{P_k} is computed as

$$T_{P_k} = g\overline{P}_{Y_k} \tag{7}$$

where g is a scale factor to compensate for short-term variation and $\overline{P}_{Y_k}$ is the average energy of the signal, where no speech is detected. $\overline{P}_{Y_k}$ is estimated over time

References

[1] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, pages 208–221, 1979.

[2] C. Breithaupt, T. Gerkmann, and R. Martin. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *Sig. Proc. Letters*, 14:1036–1039, 2007.

[3] S. Davis and P. Mermelstein. Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. *IEEE Trans. on Acoust., Speech and Sig. Proc.*, ASSP-28:357–366, 1980.

[4] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, 32:1109–1121, December 1984.

[5] H. Gustafsson, S. Nordholm, and I. Claesson. Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Trans. on Speech and Audio Proc.*, 2001.

[6] S. D. Kamath and P. C. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, 4, 2002.

[7] R. Martin. An efficient algorithm to estimate the instantaneous snr of speech signals. *Euro. Conf. on Speech Communication and Tech.*, pages 1093–1096, 1993.

[8] B. McKinley and G. Whipple. Model based speech pause detection. *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc.*, 02:1179–1182, 1997.

[9] K. Wojcicki, B. Shannon, and K. Paliwal. Spectral subtraction with variance reduced noise spectrum estimates. *Austr. Int. Conf. on Speech Science & Techn.*, 1:76–81, 2006.

[10] L.-P. Yang and Q.-J. Fu. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The Journal of the Acoust. Society of America*, 117:1001–1004, 2005.