# Fusion of Sound Source Localization and Face Detection for Supporting Human Behavior Analysis

Markus Niiranen Janne Vehkaperä Satu-Marja Mäkelä Johannes Peltola Tomi Räty

VTT Technical Research Centre of Finland
Kaitoväylä 1, 90571, Oulu, Finland

firstname.surname@vtt.fi

## ABSTRACT

This paper describes a demonstrated concept implementation that combines sound source localization and face detection from video stream for supporting human behavior analysis. System monitors space containing multiple persons using microphone array and video camera. The aim is to detect which person in the scene is producing the sound that is received by the microphones. For this task the microphone array localizes the sound in the environment. Simultaneously face detection is performed to the video signal produced by the monitoring video camera. If face is detected from the bearing of the sound the system may decide that the sound is produced by the person who's face is detected. Preliminary results indicate that the fusion may give useful information for human behavior analysis for space containing multiple persons.

## Categories and Subject Descriptors

H5.5.[**Inforamtion Interfaces And Presentation (e.g. HCI)**]: Sound and Music Computing – *Signal analysis, synthesis, and processing;* I4.6[**Image Processing and Computer Vision**] Scene Analysis - *Object Recognition, Tracing.*

## General Terms

Algorithms.

## Keywords

Audio localization, audio detection, microphone arrays, face detection.

## 1. INTRODUCTION

Automatic monitoring of people activities is a widely researched topic for different applications such as surveillance systems and user interfaces that utilize human behavior modeling. Such user interfaces include intelligent meeting rooms, robot interfaces, security applications and artistic installations [4], [5],[6],[7]. This paper describes a demonstrated system that localises the sound source and combines the sound bearing information with face localization in order to reveal additional information for human behavioural analysis in space containing multiple persons. The aim is to detect which person in the scene is producing sound, such as speech.

Both audio source and face localization are well studied areas. Audio source localization is often performed in security applications. Júlian & al presents a custom designed acoustic enclosure for the microphone array [1]. Four microphones are positioned in the enclosure so that they can detect sounds from all directions. In 2D plane one pair is positioned to vertical axis and one pair to horizontal axis, which is an initial starting point for the experiments. In a more advanced sound localization methods the microphones are positioned side by side in single or multiple lines with equal spacing as in [2].

Detecting faces has been originally performed for images. Face detection can be done in multiple ways using for example facial features, skin color or face template matching. Good overview of different techniques is presented in [3]. The same methods can be applied for video sequence but in real time requirements one must consider the computational load for detecting face from each frame.

When we know the position of the face in video and bearing of the sound, we can couple the information from audio and video sensor and detect who is speaking or creating other kind of sound, such as clapping or yelling.

Audio source localization would also benefit from audio analysis that can distinguish between different of audio classes [8] and detect when speech or other, i.e. human made, sounds are present. This would also give valuable information while detecting cues from human behavior.

In chapter 2 the audio localization algorithm and in chapter 3 the face localization is explained. In chapter 4 the demonstrated system is described briefly.

## 2. AUDIO LOCALIZATION ALGORITHM

The audio source localization and monitoring platform consists of a microphone array containing four microphones, an audio interface and an audio sensor server. The audio sensor server controls the microphone array and communicates with the

Mobimedia 2008 July 7-9, 2008, Oulu, Finland.
 ISBN 978-963-9799-25-7
DOI 10.4108/ICST.MOBIMEDIA2008.4071

session server. One of the microphones in the array is marked as a master microphone that can be listened remotely. Audio data is primarily transmitted and sound intensity is monitored using the master microphone

The state diagram for the Audio Localization Platform is shown in the Figure 3. Here is a short description of the audio sensor's life span. The reference to a corresponding state is included in brackets.

At initial point the audio sensor will wait for the connect message from the session server. When the connection is done the audio sensor will be initialized (Initialization). On successful initialization the audio sensor starts running the main loop. The audio sensor will wait for the control messages transmitted from the session server, or that the audio buffer is filled in the audio recorder (Handle messages). The control messages (Shutdown, Set detection sensitivity, Data query) are processed accordingly. The audio data in the filled audio buffer (Record audio) is transmitted to the session server using streaming services (Transmit audio). Then the audio data is processed with the sound detection (Sound detection) and bearing algorithms (Sound bearing calculation). This sensor data is transmitted to the session server with constant intervals, whenever a sound event has been detected or on request from the session server (Transmit data). Then the audio sensor returns to wait for the incoming messages. The audio sensor finishes data processing when a shutdown message is received from the session server (Shutdown).

## 2.1 Sound event detection

The average sound pressure level is calculated for each sound frame using equations 1 and 2..

The sound intensity is defined as the average power of the signal ***x***.

$$P_{signal} = \frac{1}{N}\sum_{n=0}^{N-1}\left|x_n\right|^2 . \qquad (1)$$

The corresponding SNR of the signal is calculated as

$$SNR = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right).$$

10 consecutive values from average calculation are buffered and median filtered. The output of the filter is subtracted from the power sequence and the result is normalized (maximum value to 1) to emphasize relevant pulses. Frames are considered as sound events if a threshold value is exceeded. The threshold value is dependent on the detection sensitivity. Frames that are below the threshold value are considered as environmental sound.

The frames that are considered as environmental sound are used for calculating the running average of environmental noise power $P_{noise}$.

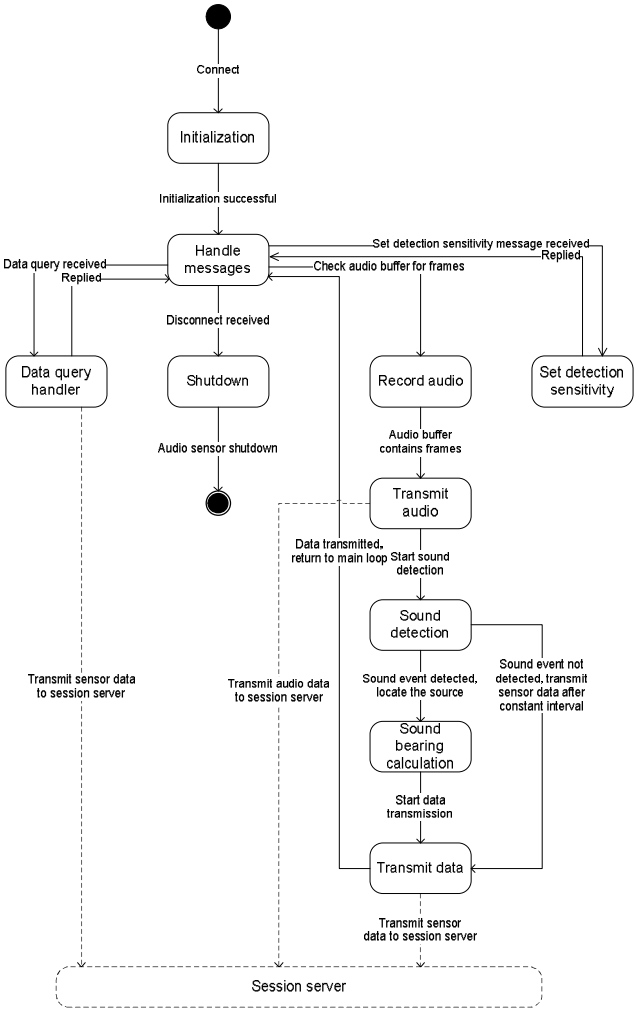


**Figure 1. A state diagram of the Audio Localization Platform.**

## 2.2 Sound bearing

The cross-correlation algorithm is used for estimating a time delay between microphone signals. However the sampling frequency of received audio signal is not enough for accurate delay estimation. Therefore an interpolation process is required to increase the accuracy before cross-correlation calculation. Signals are interpolated by the factor of 8. which increases the 44100 Hz sampling rate is to 352800 Hz. Interpolation is done by calculating FFT, zero padding the DFT and performing inverse FFT.

The cross-correlation between two signals is defined as:

$$r_{xy}(l) = \sum_n x(n)y(n-l) \qquad (3)$$

The acquired time delay is then used for calculating the bearing of the sound source. The sound source is assumed to be far away (sound source distance >> distance between microphones).

$$\cos(\beta) = \frac{T_{D\max}}{T_D} = \frac{d}{cT_D} \quad (4)$$

, where d = distance between microphones, c = speed of sound (345 m/s), $T_D$ = time delay between microphones

The time delay is calculated between microphones by cross-correlation. The time delay is the index of the maximum value of cross-correlated signals.

These results of microphone pairs are combined to get an estimate of the sound bearing.

## 2.3 Microphone array

In [1] is presented an equation for calculating the optimal sampling frequency where we can acquire the minimal distance between microphones as well. If the sampling frequency is increased, in this case multiplied by the factor of 8 to 352800 Hz, it is shown that the minimal distance between microphones shortens. This makes it possible with the presented microphone rack to increase the accuracy by increasing the microphone distances.

# 3. FACE DETECTION

The face detection component has been implemented using a similar state diagram as the audio localization platform, with an exception that the face detector is not capable to re-transmit video stream. The detection results are communicated using a socket based interface, thus audio localization and face detection can be run on different computers or it is possible to use multiple instances of face detectors or sound localization systems for monitoring same area. Fusion of the detection results can be collected on a single decision node that receives the results from different detectors.

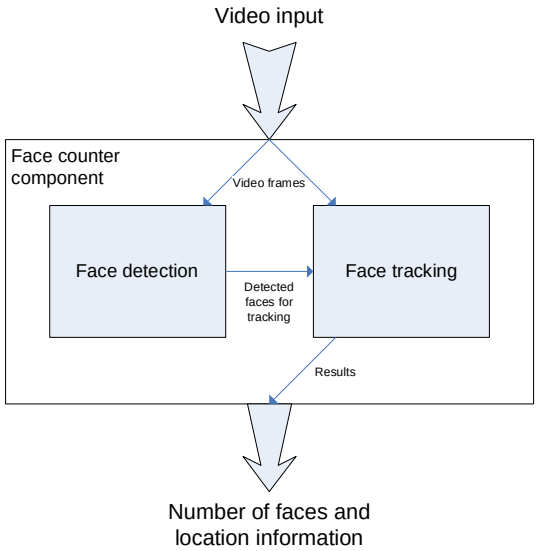


**Figure 2. Block diagram of face localization algorithm.**

The component evaluates the number of faces in video frame and tracks their movements in a live or recorded video stream providing the facial location information. The component utilises two algorithms for localization of faces. First detecting the faces is based on haar cascade classifier and then tracking the detected facial area movements from frame to frame. The component is designed for real-time applications. In general the detection requires more processing power than the tracking, thus tracking is performed more often than detection to reduce the processing load of the component. The ratio of detection and tracking is dependent on the available computational power, with a bias towards detection. Outline of the component is presented in figure 2.

Face Detection is based on the OpenCV library's (Open Source Computer Vision Library1) object detection and Face Tracking uses OpenCV's object tracking functionality. The Video Feature component returns the number faces as well as the estimated facial area (ellipse) along the tilt of the ellipse. In figure 3 there is a screenshot of the face localization algorithm.


**Figure 3. Example of face localization algorithm.**

# 4. FUSION OF THE ANALYSIS DATA

The fusion is performed on a separate application that listens results coming from analysis components using a socket based communication. The fusion algorithm knows the position of the microphones and the camera and based on all available information it can generate decision about which person is creating sound at the scene.

The microphones are positioned in a square form as shown in the Fig.4. The video sensor is positioned at the top of the microphone rack so the information of the image and audio can be adjusted to each other. The microphones and the camera in the figure are pointing towards the reader.

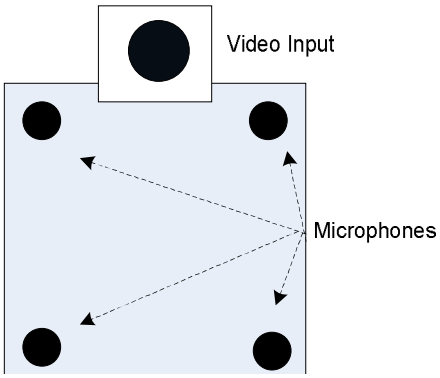


**Figure 4. Demonstration setup.**

In figure 5 is a simple example of using the sound and face localization together. The right hand image corresponds to the video frame in face detection component and the left hand image presents sound bearing so that the monitored area is divided into sectors representing the camera image regions. There are two faces detected in video frame and the sound bearing point to upper left quarter sector of the image plane. So we can assume that the sound source is the person on left.

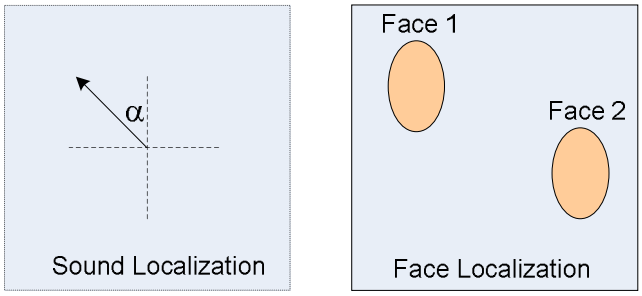


**Figure 5. Fusion of the audio visual information.**

## 5. CONCLUSION

The goal of the work was to obtain human behavior information from a monitored area so that we can automatically detect which person in the scene is producing sound at a given time.

The work consisted of defining a real-time distributed platform for analyzing the environment using different sensor information and implementing reasoning based on the sensor data.

The platform was used for implementing a sound source location using microphone arrays and face detection from video camera streams and for combining the received analysis results for obtaining information about human behavior in the monitored scene. The results obtained from this concept application indicate that fusion of sound localization and face detection can give information about the sound generating person in the monitored scene.

.In further work it should be investigated if the sound localization (or rather bearing detection) is sufficient in 2D space or should the localization be done in 3D space. On complexity aspect, 2D localization is recommended. This way audio sensor can provide bearing angle of the sound source in respect of a plane.

Different indoor spaces should also be studied to see how obstacles affect localization and how the methods presented in the state-of-the-art [4] could be used in these cases. Also the effect of background noise and case of multiple sound sources should be investigated.

## 6. ACKNOWLEDGMENTS

The acknowledgments are due to the European Commission funded CALLAS project (CALLAS IST-034800 Conveying Affectiveness in Leading-Edge Living Adaptive Systems) and Serket project. Serket is part of ITEA program, funded by TEKES (National Technology Agency of Finland).

## 7. REFERENCES

[1]Julián P., Andreou A.G., Riddle L., Shamma S., Goldberg D.H., Cauwenberghs G.: A Comparative Study of Sound Localization Algorithms for Energy Aware Sensor Network Nodes, IEEE Transactions on Circuit and Systems - I: Regular Papers, Vol. 51, No. 4, April 2004.

[2]Mungamuru B., Aarabi P.: Enhanced Sound Localization, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, Vol. 34, No. 3, June 2004.

[3] Yang M.-H., Kregman D. J, Ahuja N.: Detectecting Faces in Images: A survey, IEEE Transactions on Patterrn Analysis and Machine Intelligence, Bol. 24, No.1 Jan. 2004 pp 34 – 58

[4] Hyun-Don Kim; Komatani, K.; Ogata, T.; Okuno, H.G.: Auditory and visual integration based localization and tracking of humans in daily-life environments, Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference onOct. 29 2007-Nov. 2 2007 Page(s):2021 – 2027

[5] Gehrig, T.; Nickel, K.; Ekenel, H.K.; Klee, U.; McDonough, J.;Kalman filters for audio-video source localization ,Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on 16-19 Oct. 2005 Page(s):118 – 121

[6] A. Kushal; M. Rahurkar; Li Fei-Fei; J. Ponce; T. Huang: Audio-Visual Speaker Localization Using Graphical Models Pattern Recognition, 2006. ICPR 2006. 18th International Conference on Volume 1, 2006 Page(s):291 - 294

[7] Gilroy,S. W., Cavazza M., Chaignon, R., Mäkelä, S.-M., Niiranen M., André E., Vogt T., Billinghurst M., Seichter, H., and Benayoun M.: An Emotionally Responsive AR Art Installation Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality,ISMAR 2007, 13.-16.Nov, 2007

[8] Mäkelä S.-M., Peltola J., Myllyniemi M.: Mobile Video Capture Targeted Narrowband Audio Content Classification Proceedings of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006, May 15-19, 2006, Toulouse, France