# Tracking a Moving Person Based on a Microphone Array and a Stereo Camera

Kristian Kroschel, Markus Schlosser, Michael Grimm

*Institut für Nachrichtentechnik, Universität Karlsruhe, Kaiserstr. 12, 76128 Karlsruhe, Germany*

*Email: kroschel@int.uni-karlsruhe.de*

## Introduction

The task to track a person is of interest if a human cooperates with a robot, e.g, since the human is a source of audio and video signals it is obvious that the tracking task is based on these signals instead of radar signals, laser scanners etc. since these approaches require a higher demand of technology on one hand. On the other hand audio and video signals might be used in a cooperative environment of a human and a robot to use the audio signal for speech recognition and the video signal to extract gestures, respectively.

In this contribution it is shown how the two data sources can be fused so that each of the source contributes at its best to the goal to determine the true location of the human.

## Audio Versus Video Sensor Data

Primarily, video signals have an advantage over audio signals when tracking a person since the video signal is permanently present whereas the audio signal in case that it is speech will not always be available. The argument that a moving person generates sounds permanently by walking on the ground demonstrates that audio signals may be helpful, too.

The following table summarizes the pros and cons of audio and video data:

| video data | audio data |
|---|---|
| + high accuracy | + omnidirectionality |
| - limited area of view | - environmental noise |
| - obstruction (objects) | - reverberation |
| - influence of illumination | - low accuracy |

This table shows that both sensor data can profit from each other when the task has to be solved to track a moving person, i.e. to calculate the time dependent Cartesian coordinates $x, y$ and $z$. It might be sufficient to measure the coordinates $x$ and $y$ when a person is moving on the floor but since the reference point of a person, the mouth or the center of the head, e.g., is moving also along the $z$-axis, the three coordinates are of interest. Furthermore, the moving person might start from a sitting position or will sit down at the end of motion so that generally all three coordinates are of interest.

To measure the position of the person, a microphone array together with a stereo camera is used. The microphone array shown in Fig. 1 is three dimensional and thus is able to measure the azimuth $\Theta$, the elevation $\Phi$ and the distance or range $r$, the latter with low accuracy due to sharp angles of the lines between the microphones and the sound source, i.e. the mouth.

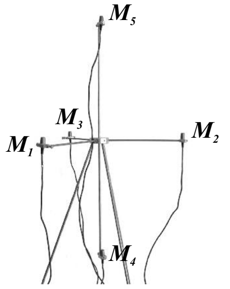

**Fig. 1:** Microphone array consisting of 5 microphones $\mathrm{M}_i$.

The angles $\Theta$ and $\Phi$ depend on the time difference of arrival of signals travelling to pairs of microphones and are extracted from the peak of the generalized cross correlation function [1].

The parameters $\Theta$, $\Phi$ and $r$ are extracted in the video channel from the face which is detected by its color and approximated by an ellipse with the reference point at the crossing of its diameters.

A nonlinear transformation is required to gain the coordinates $x, y, z$ from $\Theta, \Phi, r$. This could be done by an appropriate algorithm. But since there are the additional tasks,

- fusion of audio and video data,
- reduction of noise and reverberation

to be solved, an extended Kalman filter is used for the solution of all three tasks.

## Extended Kalman Filter

It is assumed that the person to be tracked is moving with constant velocity. Therefore the state vector of the Kalman filter is given by [2]

$$\mathbf{p}(k) = (x, y, z, \dot{x}, \dot{y}, \dot{z})^T, \quad 1 \leq k \leq K$$

whereas the vector containing the (corrupted) measurements is

$$\mathbf{p}_r(k) = (\Theta_V, \phi_V, r_V, \Theta_A, \phi_A, r_A)^T, \quad 1 \leq k \leq K$$

with standard deviations $\sigma_\Theta^{(A)} = \sigma_\Phi^{(A)} = 2°$, $\sigma_r^{(A)} = 0.2\,\mathrm{m}$ in the audio and $\sigma_\Theta^{(V)} = \sigma_\Phi^{(V)} = 1°$, $\sigma_r^{(V)} = 0.1\,\mathrm{m}$ in the video case.

The model on which the Kalman filter [3] is based can be written as

$$\begin{aligned}\mathbf{p}(k+1) &= \mathbf{A}\,\mathbf{p}(k) + \mathbf{B}\,u(k)\\ \mathbf{p}_r(k) &= \mathbf{C}\,\mathbf{p}(k) + \mathbf{n}(k)\end{aligned}$$

with the matrices [2]

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & T & 0 & 0\\ 0 & 1 & 0 & 0 & T & 0\\ 0 & 0 & 1 & 0 & 0 & T\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} T^2/2\\ T^2/2\\ T^2/2\\ T\\ T\\ T \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The standard deviation of the noise $\mathbf{n}(k)$ is given above and the standard deviation of the input process $u(k)$ is typically $\sigma_U = 0.3\,\mathrm{m/s}^2$.

It can be read from the model equations that *late fusion* is implemented in this case which is realized by the matrix $\mathbf{C}$. No weighting was used for the two channels, i.e. audio and video have the same weights as far as the state vector $\mathbf{x}(k)$ is concerned, their reliability is expressed only by their noise standard deviations $\sigma^{(A)}$ and $\sigma^{(V)}$, respectively.

Since the predictive model of the Kalman filter has been used, prediction of the location is possible if one or both of the input channels are interrupted for a short period.

By the model given above, synchronization of both channels is assumed. The microphone array outputs a parameter vector $\Theta_A, \phi_A, r_A$ typically every 100 ms, whereas the stereo camera and the following signal processing unit operates on 15 frames/s, i.e. it generates a parameter vector $\Theta_V, \phi_V, r_V$ every 67 ms. Using time stamps on each data stream and interpolation, synchronous data are sent to the input of the Kalman filter.

## Application Example

It is assumed that a human walks with constant speed along an elliptic line marked on the floor. Along this path $K = 100$ locations are marked at which the location is sensed by the audio und video system. From the measured data these locations are determined by the audio and video data separately. Furthermore, the data are filtered by the Kalman filter and by this they are fused and the noise is reduced. The result is shown in Fig. 2.

To evaluate the result, the mean square error between the ground truth $\mathbf{p}^{(G)}(k)$ and the different estimators has been calculated

$$E^{(I)} = \frac{1}{K}\sum_{k=1}^{K} |\mathbf{p}^{(G)}(k) - \mathbf{p}^{(I)}(k)|^2, \quad I \in \{A, V, AV\}$$

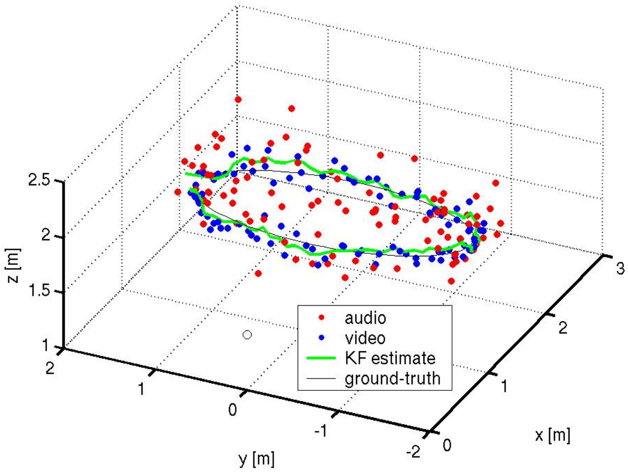

**Fig. 2:** Person moving with constant speed. Ground truth and estimates by audio and video data and fused data at $K = 100$ locations.

with $A$ for audio, $V$ for video and $AV$ for fused data. The numerical results are summarized in the following table:

| error | $E^{(A)}$ | $E^{(V)}$ | $E^{(AV)}$ |
|---|---|---|---|
| unfiltered | 0.0471 $\mathrm{m}^2$ | 0.0122 $\mathrm{m}^2$ | - |
| filtered | 0.0135 $\mathrm{m}^2$ | 0.0043 $\mathrm{m}^2$ | 0.0037 $\mathrm{m}^2$ |

From the results given in this table the following conclusions can be drawn:

- As expected from data precision, video outperforms audio for localization of moving persons.
- A significant improvement is gained if Kalman filtering is applied despite the simple model underlying the filter design.
- Fusion of audio and video data improves the result further. Even if this improvement is not very high, fusion is recommended for continuous tracking in case that one channel is interrupted.

## Acknowledgement

A part of this work has been supported by the Deutsche Forschungsgemeinschaft in the framework of the Sonderforschungsbereich 588 *Humanoide Roboter – lernende und kooperierende multimodale Roboter*.

## References

[1] C.H. Knapp, G.C. Carter, *The generalized correlation method for estimation of time delay.* IEEE Trans. on Acoustics, Speech and Signal Processing, **24 (4)**, 320-327, 1976.

[2] K. Kroschel et al, *Audio-Visual Scene Analysis for Humanoid Robots.* In: Colloque Annuel IAR Jahrestagung, Karlsruhe, 82-87, 2004

[3] K. Kroschel, *Statistische Informationstechnik*, 4th ed. Springer, Heidelberg, 2004.