In Vision Modelling and Visualization (VMV), Saarbruecken, Germany, 2000

Object Localization by Joint Audio-Video Signal Processing

G. Wang, R. Rabenstein, N. Strobel* and S. Spors

University of Erlangen-Nürberg, Telecommunications Laboratory
Cauerstr. 7, 91058 Erlangen, Germany
Email: wang@LNT.de

Abstract

There are many different approaches either for the localization of sound sources or for tracking of visible objects in image sequences. However, most applications use only one modality, that is, they process only audio or video information for object localization. In this paper we introduce a method for the estimation of object positions based on joint audio-video information. The key technique is *a modified decentralized Kalman filter (MDKF)*, where the object localization problem is viewed as state estimation. First, the position of the object is estimated based on audio and video information separately. Then, the locally estimated results are further processed in a decentralized Kalman filter for data fusion. At the output we obtain the joint estimation results. Experiments have shown that the joint estimation provides more correct localization results than obtained by using audio or video information only.

1 Introduction

Object localization based on audio and video information is an important topic in various applications, e.g. analysis of dynamic scenes, video conferences, analysis of traffic situations, and others.

In the past, many methods have been reported for object localization based on audio or video information, respectively. Object localization using audio information requires a microphone array. To estimate the position of an object we record the sound signals generated by the object using a number of spatially distributed microphone sensors. Assuming that the amplitude gradient across the microphone array is negligible, the geometric information can be considered to be encoded in the time differences of arrival of the wavefront at the microphones. Therefore, the time delays between different microphone signals can be expressed by the unknown source location parameters. For this purpose there are two methods, namely direct and indirect acoustic source localization methods. The comparison of these two methods can be found in [1].

Object localization using video informations is usually based on the analysis of image sequences. There are basically two methods, on which the analysis of video information is based, namely the use of two-dimensional and of three-dimensional models. For two-dimensional models, a number of techniques are in current use. They include comparison between the foreground and background of a scene, search for characteristic features e.g. by template matching, segmentation according to color information, and others. The use of three-dimensional models is less advanced. An example can be found in [2].

All methods for object localization using audio or video informations suffer from errors caused by reflections, background noise and illumination changes. Many developments have

*now with Siemens AG.

been made to improve the localization for a single modality. Here, we show how to improve the accuracy of object localization by joint audio-video signal processing. The basic idea is to combine the different locally estimated results from the different sensors by data fusion techniques. The algorithm introduced in this paper is based on the decentralized Kalman filter [3,4]. To make the method more general and more practical, we propose here *a modified decentralized Kalman filter (MDKF).*

The paper is organized as follows. Following the introduction we review some theory about the decentralized Kalman filter and introduce a new modified decentralized Kalman filter. In section 3 and section 4 the two local estimators for audio and video information are described. The improvements of the localization are shown in section 5.

2 Joint Audio-Video Signal Processing

Viewing object localization as a state estimation problem has two advantages. At first, we can apply Kalman filtering techniques already developed for general estimation problems. This provides a solid mathematical basis for our data fusion problem. Secondly, state estimation is compatible with models for object movements derived from the physical laws of dynamic motion.

A decentralized Kalman filter structure suitable for the use with different kinds of sensors has been introduced in [4]. A decentralized Kalman filter consists of a fusion centre and two and more local Kalman filters, which are used to generate the local estimate based on the corresponding local measurements. The local measurements are provided by different sensors. Furthermore, the local Kalman filters can also have different state space models. For joint audio-video object localization the audio sensor is a microphone array and the video sensor is a camera. The fusion of audio and video position estimations is carried out in the fusion centre. The fusion centre yields the global a posteriori estimation for the object position.

2.1 Decentralized Kalman Filter with Single State Model

If a dynamic system can be described by a state-space model, we can apply a Kalman filter for estimating the system state. Suppose that a system can be described by a single state model

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}[k]\mathbf{x}[k] + \mathbf{b}[k]u[k] + \mathbf{v}[k] \quad (1) \\ \mathbf{y}[k] &= \mathbf{C}[k]\mathbf{x}[k] + \mathbf{n}[k], \quad (2) \end{aligned}$$

where $\mathbf{v}[k]$ and $\mathbf{n}[k]$ represent the process and measurement random noise, respectively. Usually, it is assumed that $\mathbf{v}[k]$ and $\mathbf{n}[k]$ are normally distributed with zero mean and covariance matrices $\mathbf{R}_{vv}[k]$ and $\mathbf{R}_{nn}[k]$.

Kalman filters are basically classified as central Kalman filters and decentralized Kalman filters. A Kalman filter is said to be a central Kalman filter, if all measurements are processed by this Kalman filter. A decentralized Kalman filter is defined as a kind of Kalman filter, in which the measurements are first processed through different local Kalman filters. Their estimations are sent into a fusion centre for data fusion. If the measurements come from different kinds of sensors, it is favourable to use a decentralized Kalman filter to estimate the system state. The decentralized Kalman filter may not only reduce the calculation complexity but also improve the estimation accuracy.

Figure 1 shows the structure of the decentralized Kalman filter used for joint audio-video object localization, where we have two kinds of sensors, microphone array and video camera. The microphone array captures the sound wave and yields the digital signals through an A/D converter, while the video camera supplies an image sequence of the object. The measurements from audio and video sensors are first fed into the corresponding local central Kalman filters to generate the local estimations, respectively. The fusion centre then combines the local estimations to calculate a global estimation of the system state. To describe a decentralized Kalman filter it is necessary to assume that the process noise $\mathbf{v}_i[k]$ and the measurement noise $\mathbf{n}_i[k]$ are independent. Under this assumption we can express the global a posteriori state estimation in Figure 1 as [4]

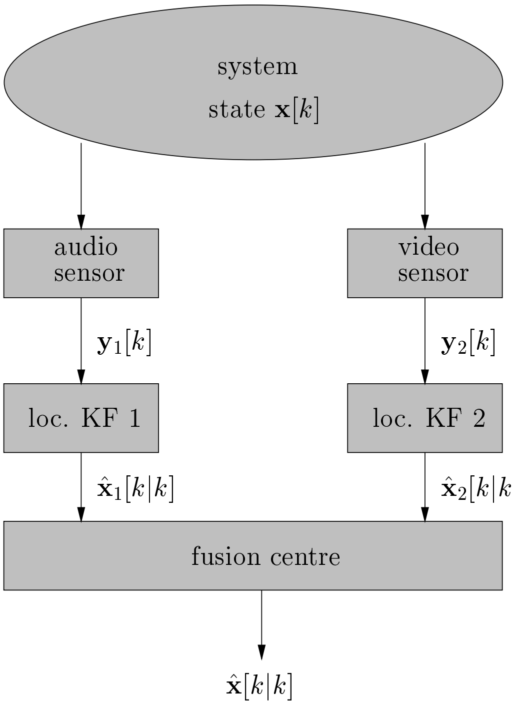

Figure 1: Decentralized Kalman filter for joint audio-video object localization.

$$\begin{aligned}
\hat{\mathbf{x}}[k|k] &= \mathbf{P}[k|k]\Big(\mathbf{P}^{-1}[k|k-1]\hat{\mathbf{x}}[k|k-1] \\
&+ \sum_{i=1}^{2}\{\mathbf{P}_i^{-1}[k|k]\hat{\mathbf{x}}_i[k|k] \\
&- \mathbf{P}_i^{-1}[k|k-1]\hat{\mathbf{x}}_i[k|k-1]\}\Big), \quad (3)
\end{aligned}$$

where the matrices $\mathbf{P}[k|k-1]$ and $\mathbf{P}[k|k]$ denote the global a priori and a posteriori estimation error covariance, respectively. $\mathbf{P}_i[k|k-1]$ and $\mathbf{P}_i[k|k]$ are the corresponding local error covariances.

2.2 Decentralized Kalman Filter with Multiple State Model

In subsection 2.1 we assumed that the system has only a single state model. However, in some situations it is difficult to use only one state model to describe a system. For example, if we want to describe a object that may perform different kinds of motion, we should use different motion models. Furthermore, for a multiple state system the applied Kalman filter should be adaptive according to the measurements to decide which model will be used. In this case, the optimal a posteriori state estimation can be written as [5]

$$\hat{\mathbf{x}}[k|k] = \sum_{i=1}^{L}\hat{\mathbf{x}}_{\alpha_i}[k|k]p(\alpha_i|\mathbf{Y}[k]), \quad (4)$$

where α_i denotes the i-th state-space model to be used and $p(\alpha_i|\mathbf{Y}[k])$ is the model probability. Figure 2 shows the block diagram of a central adaptive Kalman filter. Here, $\mathbf{Y}[k]$ denotes the measurement vector containing all measurements from different sensors, while $\mathbf{y}_1[k]$ and $\mathbf{y}_2[k]$ represent the measurement vectors from sensor 1 and sensor 2, respectively.

Replacing each central Kalman filter in Figure 2 by a decentralized Kalman filter according to Figure 1 gives the structure of the decentralized adaptive Kalman filter with multiple state models for different kinds of motions. An example for $L = 2$ is shown in Figure 3.

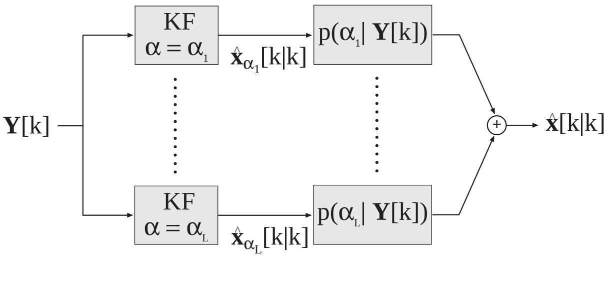
KF
$\alpha = \alpha_1$
$\hat{\mathbf{x}}_{\alpha_1}[k|k]$
$p(\alpha_1|\,\mathbf{Y}[k])$
$\mathbf{Y}[k]$
$\hat{\mathbf{x}}[k|k]$
KF
$\alpha = \alpha_L$
$\hat{\mathbf{x}}_{\alpha_L}[k|k]$
$p(\alpha_L|\,\mathbf{Y}[k])$

Figure 2: Block diagram of the adaptive Kalman filter.

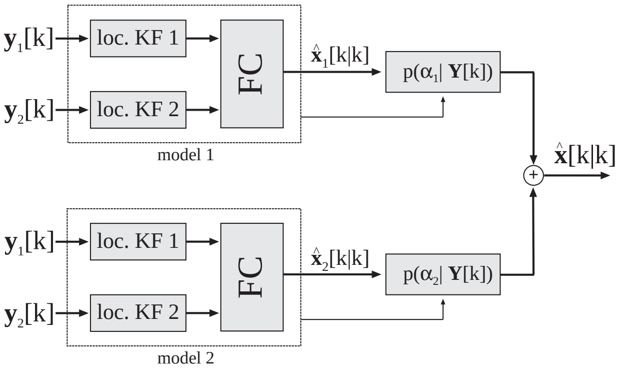

Figure 3: Block diagram of the decentralized adaptive Kalman filter.

2.3 Modified Decentralized Kalman Filter (MDKF)

From equations (1) and (2) we know that the difficulty to use the Kalman filter for state estimation is to find an expression between measurement $\mathbf{y}[k]$ and the system state $\mathbf{x}[k]$. In many applications it is not possible to describe the measurement channel by a mathematical method according to (2), because the relation between the system state $\mathbf{x}[k]$ and the measurement values $\mathbf{y}[k]$ is very involved. Also the use of an extended Kalman filter cannot solve this problem.

For example, in joint audio-video object localization the measurements from the audio sensors are digital signals representing the sound pressure and the measurements from the video sensors are images of the object. Suppose that we choose the system state as

$$\mathbf{x}[k] = \begin{bmatrix} x_x[k] \\ v_x[k] \\ x_y[k] \\ v_y[k] \end{bmatrix}, \qquad (5)$$

where $x_x[k]$, $x_y[k]$, $v_x[k]$, $v_y[k]$, are the horizontal and vertical components of the object position and velocity at k. Then, it is very difficult to find a function to express the relationship between $\mathbf{x}[k]$ and $\mathbf{y}[k]$. However, without a measurement channel expression as in equation (2), it is not possible to use the decentralized Kalman filter.

To solve this problem, we propose a modified decentralized Kalman filter. Its structure is shown in Figure 4. In addition to the decentralized Kalman filter, local estimators are inserted between sensors and local central Kalman filters. They estimate the object position from the sensor signals $\mathbf{y}_1[k]$ and $\mathbf{y}_2[k]$, respectively. The local estimators perform no recursive estimation and can provide only estimates of the current object position

$$\mathbf{z}_i[k] = \begin{bmatrix} z_{x_i}[k] \\ z_{y_i}[k] \end{bmatrix}, i = 1, 2. \qquad (6)$$

On the other hand, they do not need a motion model of the object.

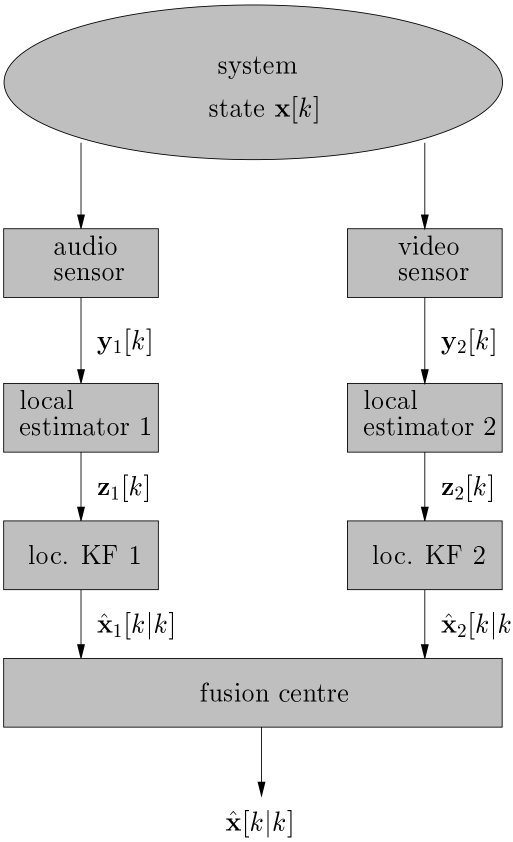

Figure 4: Modified decentralized Kalman filter (MDKF) for joint audio-video object localization.

The results $\mathbf{z}_1[k]$ and $\mathbf{z}_2[k]$ from the local estimators are now used as input for the decentralized Kalman filter. In contrast to Figure 1, there is a simple relationship between system state $\mathbf{x}[k]$ and the input signals of the local Kalman filter $\mathbf{z}_1[k]$ and $\mathbf{z}_2[k]$ (see (8)).

The remaining blocks in Figure 4, namely the local Kalman filter 1 and 2 and the fusion centre, represent a decentralized Kalman filter as in Figure 1. It is based on a linear motion model for the object dynamics and on a measurement model.

A simple motion model is obtained by assuming constant object velocity and a cartesian coordinate system. The resulting state-space equation can be expressed as

$$\underbrace{\begin{bmatrix} x_x[k+1] \\ v_x[k+1] \\ x_y[k+1] \\ v_y[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \underbrace{\begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}[k]} \underbrace{\begin{bmatrix} x_x[k] \\ v_x[k] \\ x_y[k] \\ v_y[k] \end{bmatrix}}_{\mathbf{x}[k]} + \mathbf{v}[k], \qquad (7)$$

where T is the time interval between subsequent estimates.

The corresponding measurement channel can be now simply described as

$$\underbrace{\begin{bmatrix} z_{xi}[k] \\ z_{yi}[k] \end{bmatrix}}_{\mathbf{z}_i[k]} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{C}[k]} \underbrace{\begin{bmatrix} x_x[k] \\ v_x[k] \\ x_y[k] \\ v_y[k] \end{bmatrix}}_{\mathbf{x}[k]} + \mathbf{n}_i[k], \quad i = 1,2 \qquad (8)$$

where $\mathbf{n}_i[k]$ denotes the local estimation errors.

3 Object Localization from Audio Information

From Fig. 4 we can see that the audio and video informations measured by microphone array and camera are first fed into two local estimators, respectively, in order to generate the local estimates. They act as inputs signals to the decentralized Kalman filter for data fusion. Therefore, the two local estimators play an important role in the whole system, since their accuracy can directly influence the final estimation results. In this section we describe the audio local estimator. The video local estimator is given in the next section.

The basic processing chain for object localization from audio information is shown in Figure 5. In order to localize the position of the acoustic source, we record sound signals, which is done with a microphone array. Then characteristic features of the recorded signals, in which the object position information is encoded, have to be extracted. At last, the extracted characteristics are used as inputs of the estimation algorithm to get the actual estimated source position.

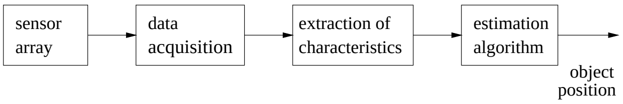

Figure 5: Processing chain for acoustic source localization.

The important topics in the processing chain shown in Figure 5 are the extraction of characteristic features and the subsequent estimation algorithm. In the proposed method we choose the time delays between the different microphones in the array as object characteristics. The first step for the extraction of the characteristic features is the calculation of cross correlation sequences between the different microphone signals. This is done in a block wise fashion. The block length and the audio sampling rate are chosen such that the time interval between subsequent position estimates corresponds to the frame rate of the video system. The calculated short time correlation functions are then used to estimate the source position with a steered beamformer [1]. If the object to be localized is a moving human speaker, a speech pause detector is used to prevent erroneous position estimation during speech pauses.

A crucial point in the implementation of the steered beamformer is the calculation complexity. Two measures for its reduction have been applied. At first, the steered beamformer is realized by a summed-correlator algorithm to avoid the manipulation of variable delays in the microphone signal lines.

Secondary, a hierarchical search structure is used. In a certain experimental setup, a full search of all passible discrete object localizations would have required to search for a total of 2496 possible positions. Application of a three stage algorithm in a cartesian grid similar to [6] reduced the search positions to 269. A further reduction is possible in a polar coordinate system by hierarchically searching for 1) the angle, 2) the distance and 3) the exact position. This resulted in a further decrease of the number of search position to 139 in the specific example.

A more detailed description of the audio local estimation based on a steered beamformer is given in [1,8].

4 Object Localization for Video Information

The localization estimator from video information is based on object color segmentation. The methods of object localization through object color are relatively new. Some results have been reported in [6,7]. The algorithm consists of two steps. First, the object is recognized through segmentation from the still background. Then, the color informations of pixels, which lie in the recognized area, are analysed to determinate the dominating region of the object.

The extraction of the foreground is based on the analysis of image differences . By comparison of two images at different times, the part in which there is small difference will be classified into background, and the part with large difference belongs to foreground. In the case that a stationary background is used, e.g. in a video conference, the difference is obtained through comparison with a reference image.

The difference between color images can be obtained in many ways. One of the possibilities is that only the illumination changes are taken into account. If the color information should be considered, a distance measure, e.g. the euclidian distance between the color channels can be applied. The foreground and background are segmented according to a threshold. Decision errors in the boundary can be eliminated through a median filter. Fig.6 shows an example for the segmentation of the foreground and background with and without median filter.

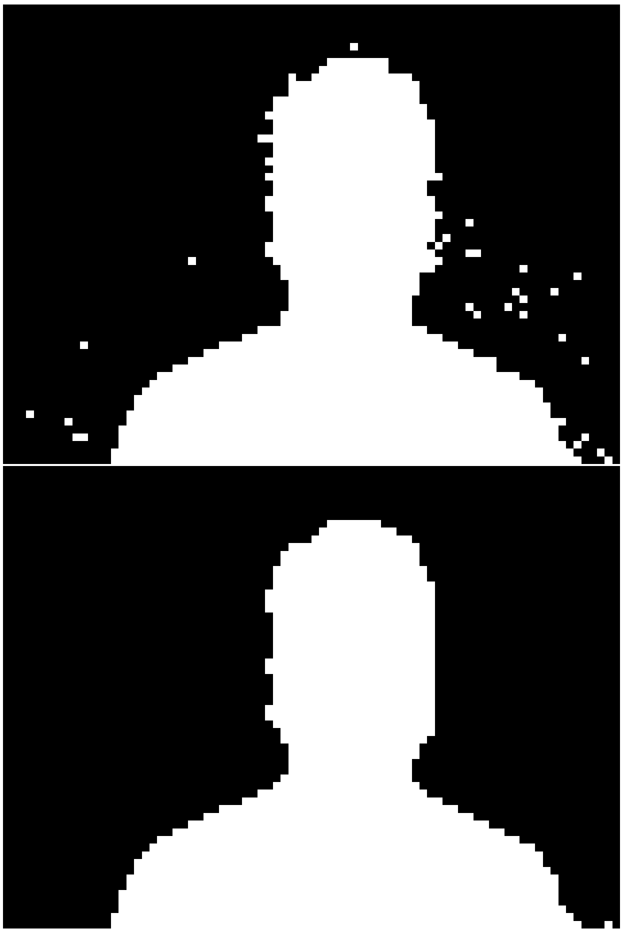

Figure 6: Segmentation of the foreground and background with (top) and without (bottom) median filter.

The recognition of object color is limited to the analysis of foreground. It is important to choose the correct color space to analyse the color information. If one takes the analysis in the RGB color space, then all three color channels have to be processed. Furthermore, the brightness can influence the color channels and impair the object color detection. So, it is useful to choose a color system that consists of an intensity channel and two color channels, for example the YUV color space or the YCrCb color space. For simplicity we can neglect the intensity channel Y, which makes the detection robust against illumination changes. Figure 7 shows a skin color distribution and a color distribution from clothing and background.

Because only a small area in the CrCb plane

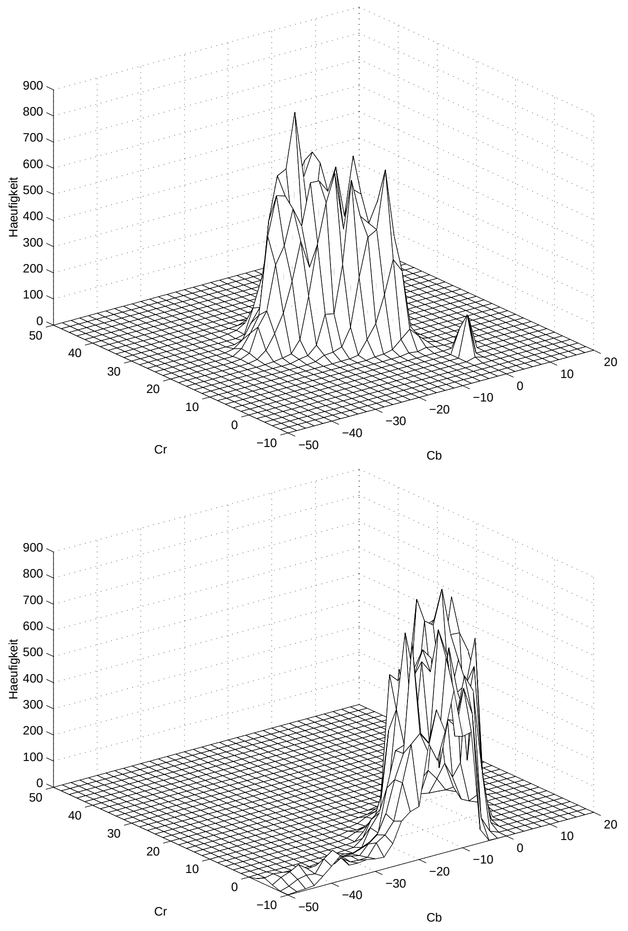

Figure 7: Skin color distribution (top) and the color distribution from clothing and background (bottom) in the CrCb plane.

is considered as object color, all pixels in the foreground should be compared to decide whether they belong to the object color area. An example for face localization is shown in Figure 8, where the white pixels indicate face color. It is clear to see that not only the face color area but also another skin color area is recognized. Therefore, to correctly localize the face, the position and size of the face have to be determined. First, the frequency distribution of the face color along the x and y direction is represented by two histograms. From the mean and variance values the position and size of the face can be evaluated. But the other color areas may misrepresent the face size. To avoid this error we use a method proposed in [6]. The result in Figure 9 shows that the hand in the lower right corner does not influence the determination of the face position and size.

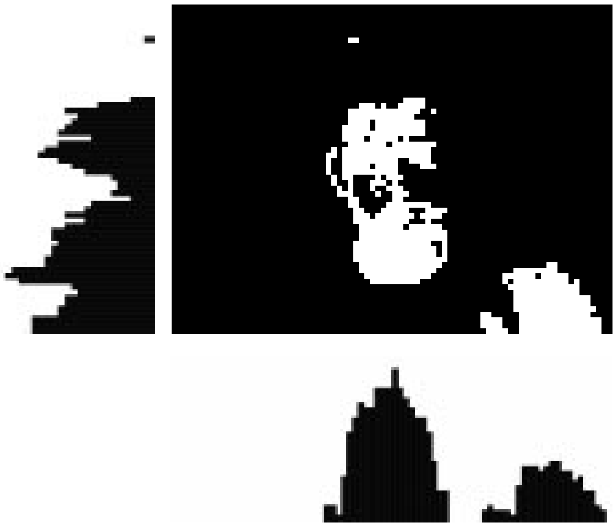

Figure 8: Recognition of the skin color areas in an image (white: skin color, black: no skin color)

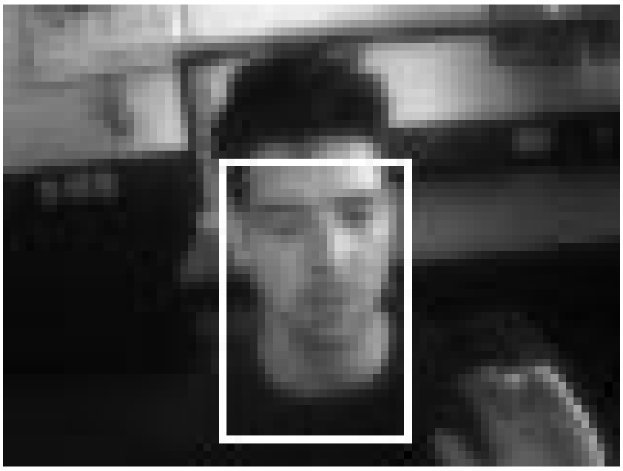

Figure 9: Determination of the position and size of the face in color regions of the other skin colors

5 Experiments

To test the proposed method we have used it for estimating the position of a whistling model railway moving along an oval track. The local estimates are obtained using the local estimators introduced in subsection 3 and

Table 1: Error variance of the audio, video and global estimation

audio estimation	$\sigma_{d_1}^2$	$1,8 \cdot 10^{-3}$	m^2
video estimation	$\sigma_{d_2}^2$	$2,4 \cdot 10^{-4}$	m^2
global estimation	σ_d^2	$1,5 \cdot 10^{-4}$	m^2

subsection 4. The results are given in Table 1, where $\sigma_{d_1}^2$, $\sigma_{d_2}^2$ and σ_d^2 denote the variance of the audio local estimation, video local estimation and global estimation of the true position $\mathbf{x}[k]$.

The results show that the error variance of the video estimation can be reduced considerably by data fusion with the audio estimation, although the audio estimation is less accurate than the video estimation.

6 Conclusion

In this paper we have introduced a method for object localization using joint audio-video signal processing. A new modified decentralized Kalman filter (MDKF) was introduced to estimate the global object position. As local estimators, a steered beamformer approach has been applied for acoustic source localization and skin color detection has been used to localize human faces in video sequences.

Experiments have shown that the modified decentralized Kalman filter (MDKF) is an effective approach for the system state estimation, especially, when the measure channel is difficult to describe. Furthermore, the MDKF can also provide more estimation accuracy than the local estimators.

References

[1] N. Strobel and Th. Meier and R. Rabenstein, “Speaker Localization Using a Steered Filter-And-Sum Beamformer”, *Workshop on Vision, Modeling, and Visualization*, Erlangen, November 1999.

[2] E. Steinbach and P. Eisert and B. Girod, “Motion-based Analysis and Segmentation of Image Sequences using 3-D Scene Models”, *Signal Processing*, 66 (2), pp. 233-247, April 1998.

[3] N. Strobel and S. Spors and R. Rabenstein, “Joint Audio-Video Object Localization Using a Recursive Multi-Stage Multi-Sensor Estimator”, *Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP)*, Istanbul, Juni 2000.

[4] H. Hashemipour and S. Roy and A. Laub, “Decentralized structures for parallel Kalman filtering”, *IEEE Tr. on Automatic Control*, vol. 33(1), pp. 88-93, 1988.

[5] R. Brown and P. Hwang, “Introduction to random signals and applied Kalman filtering”, Wiley, 1997.

[6] R. Quian and M. Sezan and K. Matthews, “A robust real-time face tracking algorithm”, *International Conference on Image Processing*, pp. 131-135, 1998.

[7] D. Chai and K. Ngan, “Face segmentation using skin-color map in videophone applications”, *IEEE Tr. on Circuits and Systems for Video technology*, vol. 9, No. 4, pp. 551-564, 1999.

[8] N. Strobel and R. Rabenstein, “Robust Speaker Localization Using a Microphone Array”, *European Signal Processing Conference (EUSIPCO)*, Tampere, Sept. 2000.