A Multi-Sensor Object Localization System

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

Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
E-mail: {spors, rabe, strobel}@LNT.de

Abstract

This paper presents a localization and tracking system integrating multiple sensors. Object localization results from local sensor systems are fused using a decentralized Kalman filter. An audiovisual speaker tracking system is evaluated, which is based upon a video based face tracker and a microphone array. A quantitative analysis shows that the presented bimodal tracking system can deliver more robust and reliable results than either of the two single modalities.

1 Introduction

Modern object oriented coding algorithms, like the emerging MPEG-4 standard, have strong requirements on machine based scene analysis. During the recording of natural scenes, scene analysis can be supplemented by object localization algorithms. Many single-sensor techniques already exist for this purpose. They are, e.g. based on microphone arrays, video cameras, or range sensors. Since all of these sensors have their specific strength and weaknesses, it is often advantageous to combine information from various sensor modalities to arrive at more robust position estimates. This paper presents a multimodal object localization framework which is based on data fusion by decentralized state estimation. For this purpose the decentralized Kalman filter is utilized.
We proceed as follows: The theory of multi-sensor object localization is reviewed in Section 2. Based on the derived algorithm an audiovisual speaker tracking system is presented in Section 3 followed by an quantitative analysis of the tracking accuracy in Section 4.

*now with Siemens Medical Solutions, Erlangen, Germany

2 Multi-Sensor Object Localization

Object tracking and data fusion can be seen in the context of parameter estimation. The data we intend to fuse in our multimodal object localization framework are position estimates delivered by monomodal object localization algorithms based on a single sensor type. In general object localization algorithms deliver noisy position measurements $\mathbf{y}_i[k]$ based on the raw sensor data of the local sensors. Parameter estimation tries to estimate the true value or state of the object based on the measurements and appropriate system models. In our case the system state $\mathbf{x}[k]$ consists of the object position and other state variables. By integrating parameter estimation with data fusion we build a multimodal object localization algorithm based on the Kalman filter as shown in the next sections.

2.1 Monomodal State Estimation

In situations where the systems dynamics can be described by a state-space model, the Kalman filter (KF) algorithm provides an efficient computational solution for estimating the state of a system. The Kalman filter can be characterized as a model-based predictor followed by an observation-dependent corrector. The linear discrete Kalman filter [3] is based upon a linear state-space model for system characterization,

$$\mathbf{x}_i[k+1] = \mathbf{A}[k]\mathbf{x}_i[k] + \mathbf{b}[k]u[k] + \mathbf{v}_i[k] \quad (1a)$$

$$\mathbf{y}_i[k] = \mathbf{C}_i[k]\mathbf{x}_i[k] + \mathbf{n}_i[k] \quad (1b)$$

where $\mathbf{x}_i$ denotes the system state and $u[k]$ a control input. The random variables $\mathbf{v}_i$ and $\mathbf{n}_i$ model the additive process and measurement noise. They are assumed to be independent from each other and from the system state $\mathbf{x}_i[k]$. Furthermore it is assumed that they are normally distributed with zero

mean and covariance matrixes $\mathbf{R}_{vv}^{(i)}[k]$ and $\mathbf{R}_{nn}^{(i)}[k]$. The input of the linear discrete Kalman filter is the position estimate of the i-th sensor given through the measurement vector $\mathbf{y}_i[k]$, where $\mathbf{C}_i[k]$ denotes the observation matrix. Note that we assume identical state models (1a) for the local sensor nodes, the measurement channels (1b), however, can differ from each other. The Kalman filter algorithm consists of a set of equations that can be found e.g. in [3, 12].

2.2 Multimodal State Estimation

In the previous section we introduced the Kalman filter for monomodal object localization. The position estimates computed by the local Kalman filters for each sensor system are only based on their sensor observations. We now combine these local estimates to arrive at a more robust global position estimate. At first glance sensor data fusion can be performed by mapping all local measurement vectors $\mathbf{y}_i[k]$ into one global measurement vector

$$\mathbf{y}[k] = \left[\begin{array}{cccc} \mathbf{y}_1[k] & \mathbf{y}_2[k] & \dots & \mathbf{y}_M[k] \end{array} \right]^T \quad (2)$$

and using the Kalman filter for state estimation. This scheme is referred as measurement fusion and has some benefits compared to state-vector fusion as presented in [4]. The only drawback is the centralized structure of using one centralized Kalman filter for multimodal state estimation. However, a decentralized structure of the tracking system is more useful in practical applications. This section shows how to arrive at a joint position estimate using a decentralized Kalman filter, which is a decentralized implementation of the standard Kalman filter [5, 16, 17].

2.2.1 Single State Model

The decentralized Kalman filter (DKF) as used for the fusion of different modalities is a multi-sensor Kalman filter that has been divided up into modules associated with the local sensor systems. Each node computes a local a posteriori estimate $\hat{\mathbf{x}}_i[k|k]$ of the object position based on the position measurements $\mathbf{y}_i$ of the local sensor i. These partial estimates are finally assimilated to provide a global a posteriori estimate $\hat{\mathbf{x}}[k|k]$ in the fusion center. Figure 1 illustrates the structure of the decentralized Kalman filter. The time-update equations and measurement-

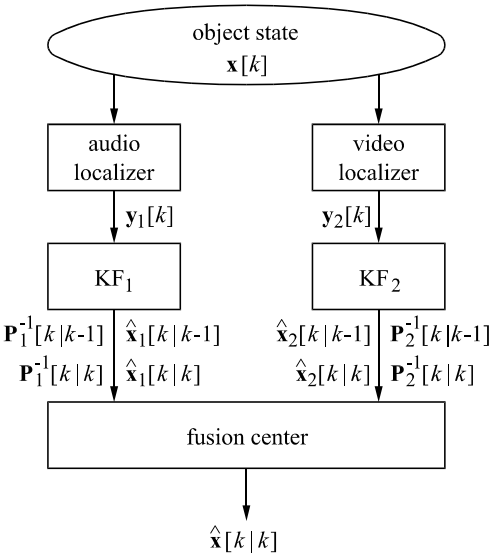


Figure 1: Structure of a multimodal object localization algorithm using a decentralized Kalman filter. Two modalities are fused here: audio based and video based object localization results. The estimated states $\hat{\mathbf{x}}_i[k|k]$ of the local Kalman filters (KF) are fused by the fusion center to a global state estimate $\hat{\mathbf{x}}[k|k]$.

update equations of a DKF with M sensors can, e.g. be found in [3]. If the measurement noise components $\mathbf{n}_i[k]$ of the local position measurements are independent, the centralized state estimate can be separated. It is assumed that the global state equation, describing the object dynamics, is equal for all local Kalman filters. Then the global state equation can be described in the same way as the local system dynamics

$$\mathbf{x}[k+1] = \mathbf{A}[k]\mathbf{x}[k] + \mathbf{b}[k]u[k] + \mathbf{v}[k]. \quad (3)$$

The global a posteriori state estimate can be expressed as

$$\hat{\mathbf{x}}[k|k] = \mathbf{P}[k|k]\Big(\mathbf{P}^{-1}[k|k-1]\hat{\mathbf{x}}[k|k-1] + \sum_{i=1}^{M}\{\mathbf{P}_i^{-1}[k|k]\hat{\mathbf{x}}_i[k|k] - \mathbf{P}_i^{-1}[k|k-1]\hat{\mathbf{x}}_i[k|k-1]\}\Big), \quad (4)$$

where $\mathbf{P}[k|k-1]$ and $\mathbf{P}[k|k]$ denote the global a priori and a posteriori error estimate covariances,

respectively, while $\mathbf{P}_i[k|k-1]$ and $\mathbf{P}_i[k|k]$ are their local counterparts at the two local processors. The vector $\hat{\mathbf{x}}[k|k-1]$ is the global a priori state estimate, and $\hat{\mathbf{x}}_i[k|k-1]$ together with $\hat{\mathbf{x}}_i[k|k]$ denote the local a priori and local a posteriori state estimates, respectively.
The global a posteriori error covariance matrix is given by

$$\mathbf{P}^{-1}[k|k] = \mathbf{P}^{-1}[k|k-1] + \sum_{i=1}^{M} \{\mathbf{P}_i^{-1}[k|k] - \mathbf{P}_i^{-1}[k|k-1]\}. \quad (5)$$

Equations (4) and (5) summarize the decentralized Kalman filter algorithm. There is no need for communications from the fusion center to the local Kalman filters in this scenario. The fusion center only needs access to the a priori and posteriori state estimates $\hat{\mathbf{x}}_i[k|k-1]$ and $\hat{\mathbf{x}}_i[k|k]$ of the local Kalman filters and the appropriate error covariance matrixes $\mathbf{P}_i[k|k-1]$ and $\mathbf{P}_i[k|k]$. It is also possible to include reliability data provided by the local localization algorithms into the data fusion scheme. This can be done with the help of the measurement error covariance matrix $\mathbf{R}_{nn}^{(i)}[k]$. The respective values of the measurement error covariance matrix $\mathbf{R}_{nn}^{(i)}[k]$ can be adjusted according to the actual state of the local object localizer. If for example the object localizer has lost the object at the current timestep, high values for the diagonal elements of the measurement error covariance matrix of the respective local sensor would be chosen. This results in reduced trust of the fusion center in these measurements.
Theoretically there is no performance loss in the decentralized system, it delivers the same results as the centralized Kalman filter. Therefore, the DKF is a good choice for decentralized measurement fusion. The benefits of the DKF are the modular concept, allowing to add sensor systems on the fly, and the ease of parallel implementation.

2.2.2 Multiple State Models

In situations where objects can perform complex movements with different types of motion, it will be difficult to construct one state-space model that always fits. These problems can be overcome by using interacting multiple model (IMM) estimators [10]. Assuming that we can find appropriate state-space models for different parts of the object trajectory, an adaptive Kalman filter can learn from the measurements which of these models is the correct one. A derivation of the adaptive Kalman filter can be found, e.g. in [3].

2.3 Nonlinear State Estimation

In the previous sections, we assumed that the system dynamics and the measurement channel can be described by a linear state-space model (1). In some cases nonlinear state-space models are required. A nonlinear measurement equation is required for example when position estimates performed in different coordinate systems have to be fused or the local sensor systems have offsets and different orientations to the global coordinate system.
The state-space description of a nonlinear system is given as follows

$$\mathbf{x}[k+1] = \mathbf{f}(\mathbf{x}[k], u[k], k) + \mathbf{v}[k] \quad (6a)$$
$$\mathbf{y}[k] = \mathbf{h}(\mathbf{x}[k], k) + \mathbf{n}[k] \quad (6b)$$

where $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$ denote known nonlinear relationships which are in general dependent on the time index k. Following the construction of the decentralized linear Kalman filter, we assume identical nonlinear plant equations $\mathbf{f}(\cdot)$. The measurement models of the distributed sensors need, however, not to be identical. In the case of distributed sensors, we get

$$\mathbf{y}_i[k] = \mathbf{h}_i(\mathbf{x}_i[k], k) + \mathbf{n}_i[k] \quad (7)$$

where the measurement noise components $\mathbf{n}_i[k]$ are again assumed to be mutually independent.
Ideally, the final state estimate after fusing all individual nonlinear estimates should be identical to the centralized state estimate. Due to the nonlinear equations, a general solution to this problem appears to be difficult, and, to the knowledge of the authors, no general solution to has been presented so far.
Nevertheless a possible solution is to use the extended Kalman filter (EKF) to perform nonlinear state-estimation in a centralized fashion. The extended Kalman filter linearises the nonlinearities of the state-space equation (6) about the filter's estimated trajectory. For this purpose a linearized ver-

sion of the nonlinear state-space equation is used

$$\mathbf{x}[k+1] \approx \mathbf{f}(\hat{\mathbf{x}}[k|k], u[k]) + \mathcal{A}[k](\mathbf{x}[k] - \hat{\mathbf{x}}[k|k]) + \mathbf{v}[k] \quad (8a)$$

$$\mathbf{y}[k] \approx \mathbf{h}(\hat{\mathbf{x}}[k|k-1], k) + \mathcal{C}[k](\mathbf{x}[k] - \hat{\mathbf{x}}[k|k]) + \mathbf{n}[k] \quad (8b)$$

where $\mathcal{A}[k]$ and $\mathcal{C}[k]$ denote the Jacobian matrices of the partial derivatives of $\mathbf{f}(\cdot)$ and $\mathbf{h}(\cdot)$. The same derivation as for the centralized linear Kalman filter can then be used as illustrated in [3].

In cases where the system dynamics can be described by a linear state-space model (1a) and a nonlinear measurement equation (6b), the extended Kalman filter can be combined with the decentralized Kalman filter. Fortunately, the decentralized Kalman filter shown in Figure 1 is composed of autonomous components. This makes it possible to replace the local, initially linear Kalman filters with extended Kalman filters where required.

The straightforward derivation of the extended Kalman filter leads to problems in numerical stability during implementation. For practical implementation the unscented Kalman filter (UKF) [18, 8] provides a solution to overcome these problems.

3 Implementation of an Audio-Visual Object Localization System

The implementation of a joint audio-video processing system based on the theory discussed so far is illustrated below. There are many ways to implement the decentralized state estimator shown in the previous sections. These depend on the objects observed, the types of sensors available, and the requirements for localization and tracking. A system intended to track a single person in an audio-visual environment is presented here. It consists of a microphone array for audio localization and a video camera for tracking of human faces. Both sensors are combined using the recursive estimation scheme presented before. We only concentrate on 2 dimensional object localization in this context.

3.1 Audio Localization

In general microphone arrays are used for acoustic source localization. The existing acoustic source localization strategies can be divided loosely into three classes:

- maximizing the output power of a steered beamformer
- estimating the time delays of arrival (TDOAs) between microphone pairs for an acoustic wavefront
- high-resolution spectral estimation concepts

A detailed discussion of the different approaches is beyond the scope of this paper, but can be found in [2].

We are currently investigating two different algorithms for acoustic source localization: The first algorithm is based upon adaptive estimation of the room impulse responses between the source and the microphones as described in [1]. By computing the delay between the direct paths of a microphone pair the TDOA is obtained (second class). The second algorithm utilizes an efficient implementation of a steered filter and sum beamformer for evaluating the beamformer output associated with each hypothesized speaker position [14, 15]. The acoustic source position is found by maximizing the output power of the steered beamformer (first class). To inhibit erroneous estimates when no speech signal is present, a speech pause detector is employed for both algorithms.

The audio localization algorithms provide estimates of the azimuth $\theta[k]$ and the range $r[k]$. Due to the nonlinear relationship of the measurements $\theta[k], r[k]$ to the components of the state vector $\mathbf{x}[k]$

$$\theta[k] = \arctan\left(\frac{y[k]}{x[k]}\right) \quad (9a)$$

$$r[k] = \sqrt{x^2[k] + y^2[k]} \quad (9b)$$

the extended Kalman filter has to be used for local state estimation. In this case the Jacobian matrix $\mathcal{C}[k]$ takes the form

$$\mathcal{C}[k] = \begin{bmatrix} -\frac{y}{x^2+y^2} & 0 & \frac{x}{x^2+y^2} & 0 \\ \frac{x}{\sqrt{x^2+y^2}} & 0 & \frac{y}{\sqrt{x^2+y^2}} & 0 \end{bmatrix}. \quad (10)$$

3.2 Video Localization

The video localization system is a real-time face tracker with the following main elements: foreground-background segmentation, detection of

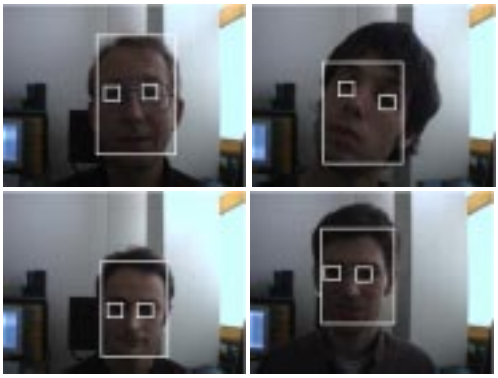


Figure 2: Sample results of the facetracking algorithm

skin-color regions, and detection of eye-like regions. Foreground-background segmentation is performed by comparing the actual captured frame with an pre captured background image at the beginning of the tracking process. Skin color segmentation is carried out on the detected foreground pixels. For this purpose a statistical skin color model [7] is utilized. Based on the results from the skin color detection task a robust statistics based algorithm estimates the center position and size of the face in the actual frame. To overcome with problems of skin color ambiguity additionally the eyes are searched in the detected facial area. This is performed by a principle component analysis (PCA) [9] based eye detection scheme. The relevant characteristics of human eyes are learned from a set of training images. These characteristics represented by a set of basis vectors, the eigeneyes, are then matched against the input frame. Some typical snapshots of tracking sessions can be seen in Figure 2. Details on the face tracker can be found in [13].

3.3 Fusion Center

The fusion center recursively combines the local a posteriori estimates $\hat{\mathbf{x}}_i[k|k]$ from the local Kalman filters into a global a posteriori estimate $\hat{\mathbf{x}}[k|k]$. It is based on the algorithm outlined in the above sections. To track a real object using a Kalman filter, a suitable motion model is needed. Since it is difficult to accurately describe complex maneuvers, we use a linear model as a first approximation instead. It is assumed that the object moves with piecewise constant velocity.

3.4 Real-Time Implementation

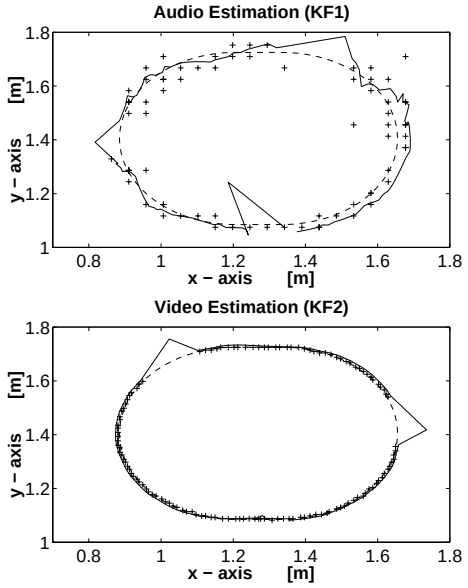
This section addresses real-time implementation issues of the described algorithms. The computational requirements of the decentralized state estimation and data fusion algorithms are quite low. The state estimates of the local Kalman filters have to be updated only each time new measurements are computed by the audio and video localization algorithms. The global position estimate is then computed by the fusion center each time the local state estimates arrive. However, the whole system has to be synchronized properly. The decentralized Kalman filter presented so far, assumes temporal alignment of the localization subsystems. This can be overcome by using an asynchronous formulation of the decentralized Kalman filter [11].
The main computational complexity of the multisensor object localization system lies in the localization algorithms itself. The current version of the facetracking algorithm was implemented on an SGI O2 workstation, providing real-time operation with 25 frames per second. Several optimizations had to be performed on the algorithm described above, details of the implementation can be found in [13]. Audio localization however, has even higher computational requirements because of the higher dimensionality of the input signals when using more than two microphones. A sufficiently high measurement update rate of the TDOA based localization algorithm using four microphones could only be achieved by downsampling the input signals and implementing the adaptive room impulse estimation in the frequency domain.

4 Quantitative Analysis

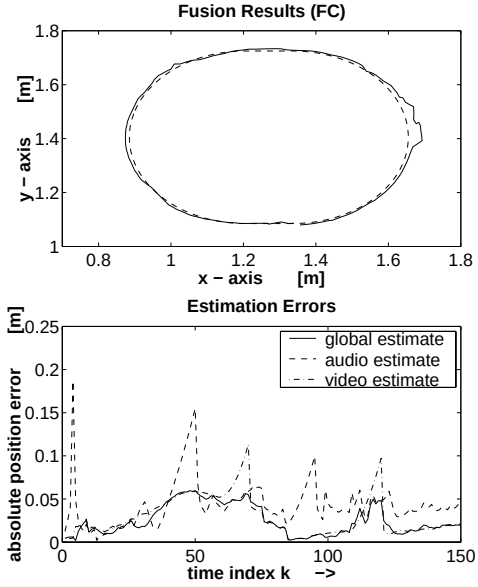
Tracking of a human speaker in an audio-visual environment is a very interesting application. Unfortunately, it does not easily facilitate a quantitative analysis since the true speaker position cannot be determined accurately by other means. To demonstrate the robustness and accuracy of joint audio-video tracking, this work will resort to an alternative setup: tracking of a model railway along an oval track in a plane. Knowledge of the fixed railway track contour together with continuous mea-

Figure 3: View from the video camera on the model railway

surements of the engine's exact position along the track provided the ground truth against which the audio-video tracking results can be compared. Figure 3 shows an example view from the video camera on the model railway. To demonstrate the increased robustness of joint audio-video processing against sensor failure, it has been assumed that both modalities suffer from poor localization conditions at different times. The audio localization results are shown in the upper plot of Figure 4(a). The dashed line is the railway track. The sequence of position estimates from the summed correlator beamformer is indicated by crosses (+). They represent the input data, $\mathbf{y}_1[k]$, to the local extended Kalman filter, KF_1. The estimation result computed by the Kalman filter is depicted as a solid line. Since the state estimation process started at the bottom part of the track, the initial error during the first steps of the Kalman recursion is clearly visible. Furthermore, there are two instances in the sequence of position estimates where the raw position estimates (observations) were dropped to mimic a silent acoustic source. In both cases, the Kalman filter extrapolated the position estimates based on the linear motion model of the local Kalman filter. When new input data became available, the position estimates resumed their proper course. The situation is similar for video localization shown in the lower plot of Figure 4(a). Since the camera usually has a much higher spatial resolution than the microphone array, the video position estimates are significantly more accurate in general. Again, two instances with missing video observations were simulated. As in the case of the audio localizer, the associated video po-



(a) position estimates from the local Kalman filters



(b) global position estimate and absolute position error

Figure 4: Sample results from experiments with the model railway

sition estimates were linearly extrapolated since the associated video Kalman filter, KF_2, uses the same motion model as the audio Kalman filter, KF_1. The fusion result is shown in the upper plot of Figure 4(b). It may be seen that the joint estimation algorithm successfully removes deviations due to unreliable audio or video observations. Finally, the lower plot in Figure 4(b) shows how the audio, video, and joint audio-video position estimates differ from the true object positions. The absolute position errors of the audio and video position estimates peak at the startup of the audio estimator and when there are failures related to missing mono-modal sensor observations. Since these deviations do not coincide in time, the joint estimate relies on the more accurate single localizer estimate in these cases. This example shows that joint audio-video object localization provides more robust results than either of the two mono-modal methods employed independently.

5 Discussion and Conclusions

This paper presented a localization and tracking system integrating multiple sensor systems. The decentralized Kalman filter recursively combines local audio and video state estimates into a more reliable global state and, thus, position estimate. To this end, a common model of the system dynamics and a common coordinate system is needed. Although audio position estimates are often less accurate than the results obtained with a video localizer, they can still provide useful input for a joint audio-video object localization system. Nevertheless, by introducing a joint audio-video processor, a localizer that yields more reliable results than either one of the single-sensor systems is obtained.

The principles of multimodal object tracking are not limited to the use of decentralized Kalman filters. More advanced state estimation techniques like the CONDENSATION algorithm [6] were also proposed for state estimation in our context. These algorithms overcome the limitations of the Kalman filter, mainly the assumption of a unimodal density to model the system state. Due to our observations, the decentralized adaptive Kalman filter does not seem to limit the performance of our multimodal tracking algorithm. The underlying assumptions were shown to be valid in our scenarios.

References

[1] J. Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America*, 107(1):384–391, 2000.

[2] M. Brandstein and D. Ward, editors. *Microphone Arrays*. Springer, 2001.

[3] Robert Grover Brown and Patrick Y.C. Hwang. *Introduction to random signals and applied Kalman filtering*. Wiley, 1997.

[4] Q. Gan and C.J. Harris. Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 37(1):273–280, Jan 2001.

[5] H. R. Hashemipour, S. Roy, and A. J. Laub. Decentralized structures for parallel Kalman filtering. *IEEE Transactions on Automatic Control*, 33(1):88–93, 1988.

[6] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(2):5–28, 1998.

[7] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 274–280, 1998.

[8] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*, Orlando, Florida.

[9] M. Kirby and L. Sirovich. Application of the Karhuenen - Loève procedure for the characterization of human faces. In *IEEE Transactions on Pattern analysis and Machine intelligence*, volume 12, pages 103–108, 1990.

[10] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan. Interacting multiple model methods in target tracking: A survey. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):103–123, Jan 1998.

[11] B. S. Rao, H. F. Durrant-Whyte, and J. A. Sheen. A fully decentralized multi-sensor system for tracking and surveillance. *International Journal of Robotics Research*, 12(1):20–44, 1993.

[12] L. L. Scharf. *Statistical Signal Processing – Detection, Estimation, and Time Series Analysis*. Addison-Wesley, 1991.

[13] S.Spors and R.Rabenstein. A real-time face tracker for color video. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, USA, Salt Lake City, May 2001.

[14] N. Strobel, T. Meier, and R. Rabenstein. Speaker localization using steered filtered-and-sum beamformers. In B. Girod, H. Niemann, and H.-P. Seidel, editors, *Proceedings Vision, Modeling, and Visualization '99*, pages 195–202, Erlangen, 1999.

[15] N. Strobel and R. Rabenstein. Robust speaker localization using a microphone array. In *Proceedings of the X European Signal Processing Conference*, volume III, pages 1409–1412. EURASIP, 2000.

[16] N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video signal processing for object localization and tracking. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 10, pages 203–225. Spinger, 2001.

[17] N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, 18(1):22–31, Jan 2001.

[18] Eric A. Wan and Rudolph van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC)*, Lake Louise, Alberta, Canada, Oct. 2000.