IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, October 2001

JOINT AUDIO-VIDEO OBJECT TRACKING

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 object localization and tracking algorithm integrating audio and video based object localization results. A face tracking algorithm and a microphone array are used to compute two single-modality speaker position estimates. These position estimates are then combined into a global position estimate using a decentralized Kalman filter. Experiments with a model railway show that such an approach yields more robust results for audio-visual object tracking than either modality by itself.

1. INTRODUCTION

Object localization and tracking is a well studied subject, which has put forth a large number of implemented systems for various applications such as robotics, scene analysis, person recognition, etc. Most of these systems fall into two categories: systems based on the analysis of video sequences or systems processing microphone array signals. Since either of these modalities (visual and acoustical) has its specific strengths and weaknesses, it is desirable to integrate the information of both modalities. This way, one can obtain more robust position estimates.
This paper describes a system for joint audio-visual object tracking. Although the general methodology is valid for any kind of objects which can be seen and heard at the same time, the specific implementation discussed here aims at tracking human speakers. The visual object localizer combines skin color based face detection and eye localization by principal component analysis. The acoustical localizer is an effective implementation of a steered beamformer for a microphone array. A decentralized Kalman filter is used at the fusion center to integrate such diverse signals as color video sequences and audio tracks.

2. OBJECT LOCALIZATION FROM VIDEO SEQUENCES

A robust face tracker for real-time operation was presented in [1]. This system combines feature invariant (skin color) and appearance based methods (eye detection) for face detection. We now use this face tracking algorithm for joint audio-video object localization.
Figure 1 shows a block diagram of the complete face tracking algorithm. It is implemented on an SGI O2 workstation, supporting real-time operation with 25 frames per second. Our implementation is based on the IRIX '*Video Library*' directly interfacing with the O2 standard video hardware. Once a new frame has been captured, its color information is subsampled to reduce the data that has to be processed. A subsampling factor of four turned out to be sufficient for real-time operation. At the next step, a foreground segmentation is carried out. For this purpose, a background image is captured at the beginning of the tracking session. To cope with changes in the background during the tracking session, the adaptive background scheme from [2] was added. Skin color segmentation is performed on the detected foreground pixels. Based on the skin color segmentation mask, the position and size of the dominant skin color region in the input frame is computed. Simultaneously, the user's eyes are located and tracked. In the next sections, we give a short overview illustrating the main components of the algorithm.

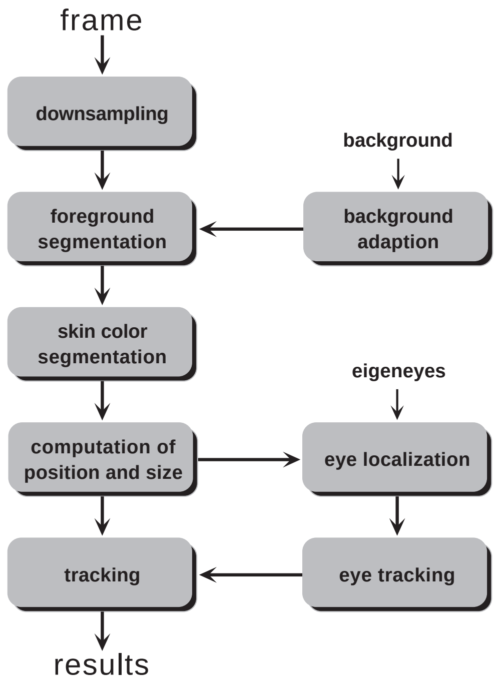


Fig. 1. Block diagram of the face tracking algorithm

2.1. Skin color based face detection

Face localization is performed using the statistical properties of human skin color. Many recent publications confirmed human skin color as a powerful feature for face detection. To improve the robustness of color segmentation, a foreground/background segmentation step is introduced before color segmentation is carried out. Skin color modeling and segmentation is performed using the YCrCb color space. It provides separation between the luminance (Y) and the chrominance (Cr,Cb) components.

2.1.1. *Skin color segmentation*

Statistical models are used to model the characteristics of human skin color. Among several models, the histogram based color model described in [3] was chosen for our algorithm. The model is trained using data from a set of hand labeled training images. In our case two classes of pixels were considered: skin and non-skin pixels. Given skin and non-skin histograms, the histogram counts are converted into estimates for the discrete probability distributions $\hat{P}(CrCb|\text{skin})$ and $\hat{P}(CrCb|\text{non-skin})$ in the usual manner:

$$\hat{P}(CrCb|\text{skin}) = \frac{c_s[CrCb]}{T_s}, \tag{1a}$$

$$\hat{P}(CrCb|\text{non-skin}) = \frac{c_n[CrCb]}{T_n} \tag{1b}$$

where $c_s[CrCb]$, $c_n[CrCb]$ denote the pixel counts for a certain $CrCb$ color pair in the skin and non-skin histograms and T_s, T_n are the total pixel counts contained in the skin and non-skin histograms, respectively. Studies have shown, that the human skin colors cluster in a small region of the color space and that there is a significant degree of separation between the skin and non-skin image classes.

The color segmentation step classifies the pixels of an given input image into skin and non-skin pixels. Only the pixels that were identified as foreground pixels are processed further by skin color segmentation. The result is a binary mask, that marks the skin color areas in a given input image. A given pixel is classified as skin pixel, if the conditional probability $\hat{P}(\text{skin}|CrCb)$ is greater than a preselected threshold θ for the $CrCb$ color pair of this pixel:

$$\hat{P}(\text{skin}|CrCb) \geq \theta \tag{2}$$

Using the Bayes rule the conditional probability $\hat{P}(\text{skin}|CrCb)$ can be computed from the color histograms in the following way:

$$\hat{P}(\text{skin}|CrCb) = \frac{\hat{P}(CrCb|\text{skin})\,\hat{P}(\text{skin})}{\hat{P}(CrCb|\text{skin})\,\hat{P}(\text{skin}) + \hat{P}(CrCb|\text{non-skin})\,\hat{P}(\text{non-skin})} \tag{3}$$

where $\hat{P}(\text{skin})$ and $\hat{P}(\text{non-skin})$ are the prior probabilities for skin and non-skin.

2.1.2. *Face localization*

The face localization is implemented using a robust, statistics based method described in [4]. Starting point for the algorithm is the mask derived from the color segmentation performed on the input image as described in the previous sections. Based upon this mask two one-dimensional projected histograms along the x- and y-axis of the mask are computed. The center position and size of the dominant face in an input image is estimated based on the means and standard deviations of trimmed versions of the projected histograms.

2.2. Eye localization and tracking

Our method is based on the principle component analysis (PCA) which is better known as eigenface analysis. PCA has been mostly used for the localization and recognition of faces so far [5]. The research in [6] shows that PCA also provides a powerful framework for locating eyes. The aim of the PCA is to find the relevant characteristics of eyes from a set of training images. The basic idea is to use a unitary transform which transforms a given input image into a lower dimensional space. According to the eigenfaces used for face detection, the vectors of the transform matrix are called eigeneyes. The basic idea is, that the eigeneye basis provides the best reconstruction results for eye like regions and thus minimal reconstruction errors. The best match between reconstruction and the input image is an eye candidate.

Although the computational effort for the PCA detection scheme can be highly reduced by downsampling, it is still too high for a real time implementation with high frame rates on the given hardware. To reduce the computational complexity further, the eye detection and tracking task is divided into two steps: First the eye is detected using the algorithms described in the previous sections. Once the position of both eyes is known, they are tracked using a luminance-adapted block matching technique, as described in [2]. This provides robust eye localization through PCA and fast tracking using block matching.

3. OBJECT LOCALIZATION FROM AUDIO SIGNALS

Using audio signals, one can estimate the object position from time differences of arrival (TDOAs) of sound waves recorded at a microphone array. There are direct and indirect acoustic source localization methods. The direct approach is based on summing the systematically delayed microphone signals and observing the power of the overall output signal. This strategy is usually implemented using a steered filter-and-sum beamformer. Indirect techniques on the other hand, require two distinct processing steps. A set of time differences of arrival (TDOAs) of an acoustic wavefront recorded at separate microphone sensors is computed first. Then geometrical properties are used to infer the source position.

In this paper we use the audio localization technique described in [7]. In this algorithm the microphone array is operated as a steered filter-and-sum beamformer implemented as a summed correlator. A potential drawback to any steered beamformer approach is the fact that we have to focus at all potential speaker positions. Depending on the spatial accuracy desired, the search complexity may be considerable. To reduce the computational complexity, a hierarchical search strategy is used. Additionally a speech pause detector improves the robustness of the speaker localization algorithm by avoiding erroneous position estimates when no speech signal is present. Finally, the algorithm provides a source position estimate in terms of azimuth and range.

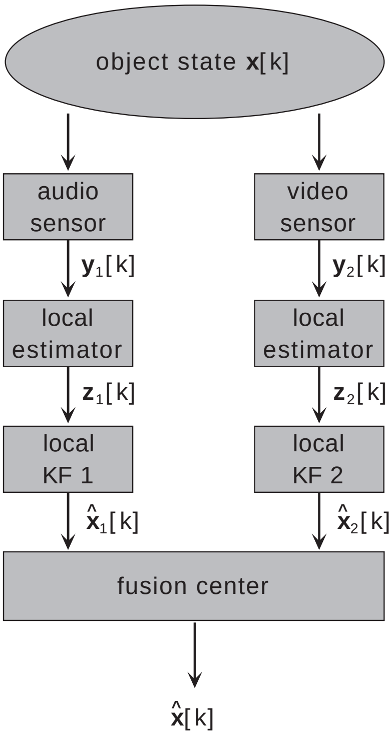


Fig. 2. Structure of the decentralized discrete Kalman filter

4. OBJECT LOCALIZATION USING AUDIO AND VIDEO INFORMATION

To increase the robustness of our tracking algorithm, we incorporate the video based position estimates as well as the audio based position estimates in the final position estimate. State estimation utilizes prior knowledge of the measurements and the system dynamics to obtain a more reliable estimate of the true system state. In the context of face localization, the state is identified as the center position or the size of the object tracked. Among various state estimation techniques, we decided to use the Kalman filter. It easily takes into account many important factors such as sequential time updates, measurement accuracy and target maneuver models. The parallelized or decentralized (linear) Kalman filter (DKF) provides a useful fusion framework for our application. The DKF is a multisensor Kalman filter that has been divided up into modules, each one associated with a particular sensor system. Figure 2 shows the structure of the DKF used. The local Kalman filter at the microphone array, the local Kalman filter at the video camera, and the global Kalman filter are the three main components needed to recursively calculate a joint object position estimate. The derivation of the DKF can be found in [8, 9]. All Kalman filters use the same dynamic model. To model the system dynamics, a motion model for the tracked object is needed. The linear motion model used here implies that the object moves with constant speed with respect to the Cartesian coordinate system used.

The video based face detection algorithm provides the Cartesian coordinates of the center position of the users head. The microphone array, on the other hand, observes the source position in terms of azimuth and range. The nonlinear relation between Cartesian and polar coordinates makes it necessary to combine a linear local Kalman filter and an extended Kalman filter when designing the overall fusion algorithm recursively computing the global position estimate in Cartesian coordinates.

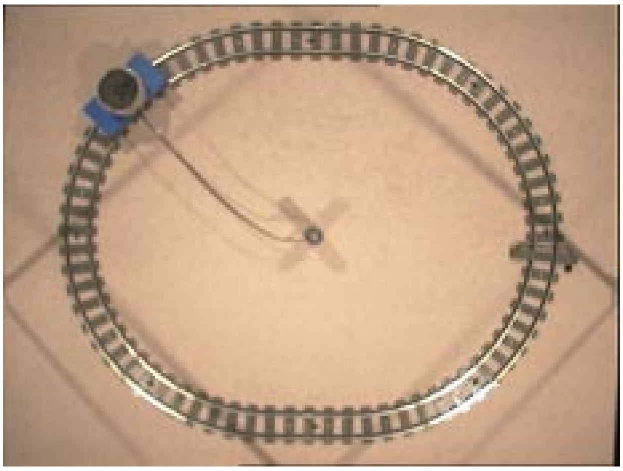


Fig. 3. View from the video camera on the model railway

5. RESULTS

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 our joint audio-video tracking algorithm we made experiments using a model railway with non-ambiguous color and a loudspeaker mounted on top of the engine. Figure 3 shows the view from the video camera on the model railway track. The skin color based localization scheme, trained on the color of the model railway, was used to localize and track the model railway. The loudspeaker played a voice signal, which was tracked by the audio localizer. The knowledge of the fixed railway track contour together with continuous measurements of the engine's exact position along the track provided the ground truth against which the audio-video tracking results could be compared. Figure 4 shows sample results from an tracking session with the model railway.

To demonstrate the increased robustness of joint audio-video processing against sensor failure, we assumed that both modalities suffer from poor localization conditions at different times. The audio localization results are shown on the top 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 KF1. The estimation result computed by the Kalman filter is depicted as a solid line. Furthermore, there are two instances in the sequence of position estimates where we dropped raw position estimates (observations) 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 got back on the track again. The situation is similar for video localization shown on the bottom. 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, we simulated two instances with missing video observations. As before at the audio localizer, the associated video position estimates were linearly extrapolated, since the associated video Kalman filter, KF2, uses the same motion model as the audio Kalman filter, KF1. The fusion result is shown on the top of Figure 4(b). We see that the joint estimation algorithm could successfully remove deviations due to unreliable audio or video observations. Finally, the plot on the bottom 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 can still rely on the more accurate single localizer estimate in these cases. This example shows, that joint audio-video object localization can provide more robust results than any of the two mono-modal methods.

6. CONCLUSION

This paper presented a localization and tracking system integrating a video based face tracker and a microphone array for speaker tracking. A quantitative analysis has shown that the presented bimodal tracking system can deliver more robust and reliable results than either of the two single modalities.

7. REFERENCES

[1] S. Spors and R. Rabenstein, "A real-time facetracker for color video," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[2] L. P. Bala, K. Talmi, and J. Liu, "Automatic detection and tracking of faces and facial features in video sequences," *Picture Coding Symposium 1997, 10-12 September 1997, Berlin*, 1997.

[3] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1998, vol. 1, pp. 274–280.

[4] R. J. Quian, M. I. Sezan, and K. E. Matthews, "A robust real-time face tracking algorithm," in *Proceedings of the 1998 IEEE International Conference on Image Processing*, 1998, vol. 1, pp. 131–135.

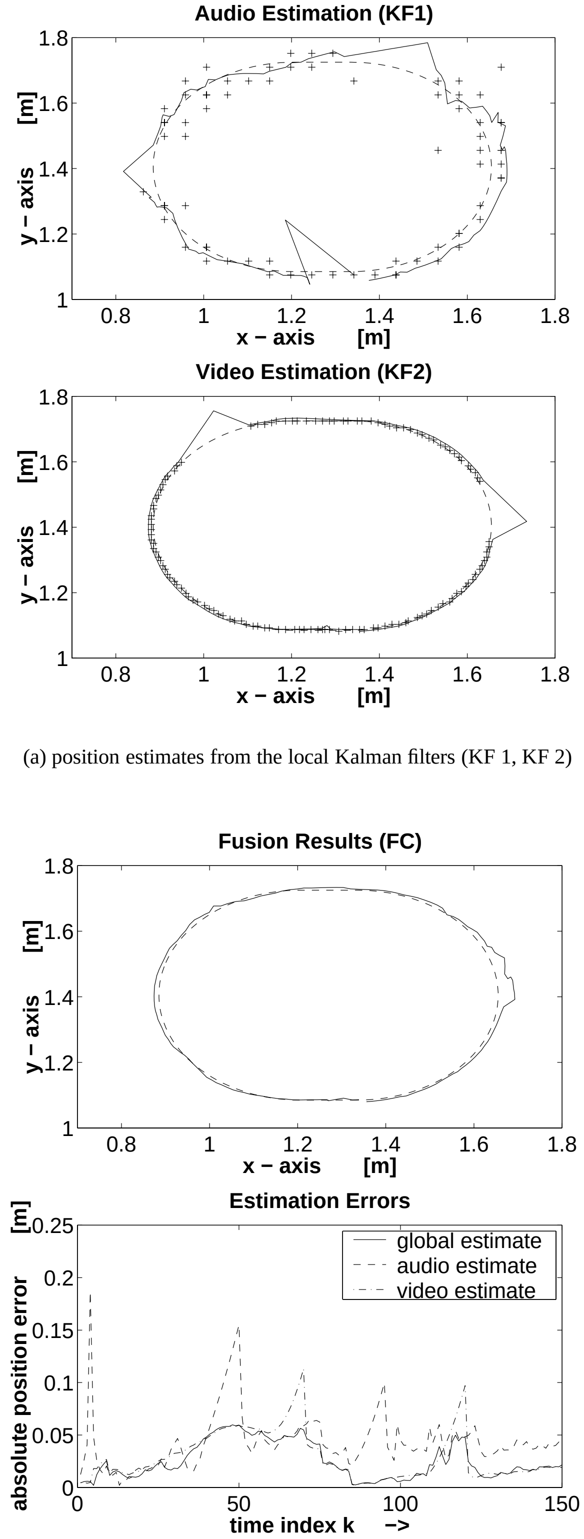
[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[6] K. Talmi and J. Liu, "Eye and gaze tracking for visually controlled interactive stereoscopic displays," *Image Communication*, vol. 14, no. 10, pp. 799–810, 1999.

[7] N. Strobel and R.Rabenstein, "Robust speaker localization using a microphone array," in *Proceedings of the X European Signal Processing Conference*, Tampere Finland, September 2000, vol. 3.

[8] G. Wang, R.Rabenstein, N.Strobel, and S.Spors, "Object localization by joint audio-video signal processing," *In Vision Modelling and Visualization*, pp. 97–104, 2000.

[9] N. Strobel, S.Spors, and R.Rabenstein, "Joint audio-video object localization using a recursive multi-state multi-sensor estimator," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 3781–3784.



(a) position estimates from the local Kalman filters (KF 1, KF 2)

(b) global position estimate and absolute position error

Fig. 4. Sample results from experiments with the model railway