Three Dimensional Object Tracking Based on Audiovisual Fusion Using Particle Swarm Optimization

Fakheredine Keyrouz, Ulrich Kirchmaier, Klaus Diepold
Technische Universität München
Arcistr. 21, 80333
Munich, Germany

Abstract—In this paper a new method of detecting and tracking a human person in three dimensional space using audio and video data is proposed. A simple tracking system with two microphones and stereo vision is introduced. The audio information is resulting from the Generalized Cross Correlation (GCC) algorithm, and the video information is extracted by the Continuously Adaptive Mean shift (CAMshift) method. The localization estimates delivered by these two systems are then combined using a novel Particle Swarm Optimization (PSO) fusion technique. In our approach the particles move in the 3D space and iteratively evaluate their current position with regard to the localization estimates of the audio and video module. This facilitates the direct determination of the object’s three dimensional position. Compared to existing methods, this novel technique achieves faster tracking performance while being independent of any kind of model, statistic, or assumption.

I. INTRODUCTION

Object tracking research is of increased importance today mainly due to growing security requirements. Applications such as video-conferencing, surveillance, smart automobiles, and automatic scene analysis are few examples in the field of autonomous systems heavily relying on audiovisual tracking. A variety of single-sensor techniques based solely on sound, e.g. microphone arrays, or vision, e.g. stereo camera systems, already exists for that purpose. As all of those have their specific weaknesses when they are deployed as stand-alone systems, it is obviously advantageous to combine the information obtained by two or more audiovisual sensors.

Different approaches for audiovisual fusion have been established. Mostly recognized are techniques like Kalman filtering fusion [1], particle filtering [2], [3], as well as approaches using probabilistic graphical models [4], such as Bayesian networks [2], [5], [6], or hidden Markov models (HMM) [7]. The above mentioned fusion methods have a common drawback. They rely on a certain model. The Gaussian distribution for example is a prerequisite for an efficient Kalman tracking. In a real-world scenario, especially in cluttered or noisy scenes, however, measurements tend to have non-Gaussian, multi-modal distributions. Furthermore, the Kalman filter hypothesizes a pre-defined and constrained model for the movement and speed of the object to be tracked. Therefore its linear state transition system fails with sudden and abrupt movements.

With the methods becoming more complex, like in the case of particle filtering or Bayesian networks, the drawback may be partly overcome on the expense of exponentially increasing complexity and processing time. Furthermore, those methods still rely on different assumptions, models, training sets, statistics, and transition probabilities that have to be postulated.

In addition, many object tracking techniques merely aim on detecting and tracking objects on the screen, without determining their positions in the real-world reference frame. This is mainly due to the fact that the calculation of the three dimensional coordinates represents an additional complex task for conventional tracking which includes error prone disparity calculation and triangulation.

In this paper, a new method of detecting and tracking a human person directly in three dimensional space using audio and video information is proposed. The audio and vision information are fused in a novel manner using the Particle Swarm Optimization algorithm (PSO). This tracking technique is faster than existing approaches while being independent of any kind of model, statistic or assumption.

In the following sections, a description of the tracking system will be presented. First, the PSO concept is introduced in section II. In section III and IV we describe the audio and the video tracking systems independently. In section V we present a novel method for audiovisual fusion based on PSO. Experimental results and comparison with current tracking techniques are discussed in section VI. Section VII concludes the current study and introduces venues for future work.

II. PARTICLE SWARM OPTIMIZATION

The PSO is a stochastic, population-based evolutionary optimization method. Descending from the research area of artificial intelligence, it was first used in [8] e.g. to simulate the social behavior of swarms like bird flocking or fish swarms.

In recent years, researchers proved the PSO to be a highly efficient optimization method and a search algorithm with high performance capabilities as well as outstanding flexibility.
The algorithm has already been applied to solve general optimization problems, as well as to provide solutions in computer vision, ad-hoc networking, search engines, and object detection or tracking [9], [10], [11].

The basic idea is to minimize a cost or 'fitness' function \( F(x) \), with \( F : \mathbb{R}^n \rightarrow \mathbb{R} \), using a swarm of dynamic particles. 'Flying' through the parameter space \( x \in \Omega \subset \mathbb{R}^n \), all particles search for the minima. In every iteration step, each of them evaluates the function along its trajectory \( x(t) \), while keeping track of the best personal solution it has found so far (\( pbest \)), where \( F(x) \) is minimized along \( x(t) \). The current best (or optimal) solution among all the points is also tracked (\( gbest \)). At any given time \( t \), the velocity \( v_i(t) \) of a particle \( i \) is then updated to point towards \( pbest \) and \( gbest \), up to a random factor defined by system parameters.

This is described by the equation

\[
\begin{align*}
    v_i(t+1) &= w \cdot v_i(t) + c_1 \cdot R_1 \cdot (pbest_i - x_i(t)) + c_2 \cdot R_2 \cdot (gbest - x_i(t)), \\
    x_i(t+1) &= x_i(t) + v_i(t+1),
\end{align*}
\]

where \( R_1 \) and \( R_2 \sim [0,1] \) are uniformly distributed random variables and \( w \), \( c_1 \) and \( c_2 \) are weighting parameters, \( w \) being a decay constant controlling the swarms convergence behavior. The parameters \( c_1 \) and \( c_2 \) represent the 'cognitive' and the 'social' component (the communication) of the swarm, respectively. The velocity \( v_i(t+1) \) describes the positional displacement between two time steps.

### III. Sound Source Localization based on GCC PHAT

The sound localization system, consisting of two horizontally spaced microphones, delivers an azimuth angle \( \nu \) which represents the relative angle between the origin of the system and the object being tracked.

In the Time Delay Of Arrival (TDOA)-based method of sound source localization, two or more sensors are used to estimate the traveling time, i.e. the time delay of a plane wave propagating across an array of sensors.

Assuming a single sound source with free-field planar wave propagation in low noise and low reverberant conditions. This wave is collected using two microphones separated by a distance \( b \). The microphone signals \( x_1(t) \) and \( x_2(t) \) can be modeled as follows [12]

\[
\begin{align*}
    x_1(t) &= s(t) + n_1(t), \\
    x_2(t) &= s(t - \tau_{\text{TDOA}}) + n_2(t),
\end{align*}
\]

with \( t \) being the time, \( s(t) \) being the signal of the source to be localized, \( \tau_{\text{TDOA}} \) being the Delay between the two sensors, and \( n_1(t) \), \( n_2(t) \) being noise signals assumed as mutually uncorrelated and wide-sense stationary processes. The Delay \( \tau \) can now be estimated by calculating the Generalized Cross Correlation (GCC) \( R_{x_1,x_2}(\tau) \), which is given by the inverse Fourier transform of the signals’ cross power spectral density, weighted with a filter function \( A(\omega) \)

\[
R_{x_1,x_2}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A(\omega)X_1(\omega)X_2^*(\omega)e^{-i\omega\tau}d\omega,
\]

assuming an infinite observation time, where \( \tau \) denotes the time delay between the frequency domain signals \( X_1(\omega) \) and \( X_2(\omega) \). The \( A(\omega) \) term describes the PHase-Transform (PHAT) weighting function, defined as

\[
A(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|}.
\]

This filter function tends to make the GCC suitable for narrow band signal and also more robust against reverberation [13].

The estimated delay \( \hat{\tau}_{\text{TDOA}} \) resulting from the signal \( s(t) \) is then obtained by a maximum search

\[
\hat{\tau}_{\text{TDOA}} = \arg \max_{\tau} R_{x_1,x_2}(\tau),
\]

which under consideration of the systems geography leads to the desired angle \( \nu \) using

\[
\nu = \arcsin \left( \frac{\hat{\tau}_{\text{TDOA}} \cdot c}{b} \right),
\]

where \( c \) denotes the velocity of sound and \( b \) represents the distance between the two microphones.

In the next section, the vision module will be explained.

### IV. Visual Object Localization via CAMShift

A convergent stereo camera system was used to track the person. The vision system delivers two correspondence or matching points, \( MP_l = (x_{MP_l}, y_{MP_l}) \) from the left frame and \( MP_r = (x_{MP_r}, y_{MP_r}) \) from the right frame. These points represent the 2D-projection of the object to be tracked.

Introduced in [14], the Continuously Adaptive Mean shift (CAMShift) is based on the mean shift algorithm [15]. The mean shift is a non-parametric approach to detect the mode of a probability distribution using a recursive procedure that converges to the closest stationarity point.

Basically, a search window is located on an initial position on a probability distribution. Then the mean location inside the window is calculated and the window is set to this mean location, which is repeated until convergence. This procedure causes the center of the window to migrate to the mode of the probability distribution.

In order to apply the mean shift method, which was implemented for probability distributions, to color tracking, a probability distribution image of the desired color, e.g. skin color in the case of face tracking, must be created. Therefore the image is in a first step transformed into the Hue Saturation Value (HSV) color space [16], which tends to be more robust against light changes in video sequences.

In a second step, a histogram has to be built over an image containing the desired color. This image could contain the front view of a face in case of skin color tracking.

The histogram bins can then be regarded as a look-up table implying the probability with which a pixel of certain color is
a skin pixel. That look-up operation realized for all pixels in the image is called the histogram back-projection.

The CAMshift algorithm expands the mean shift method by adjusting the size of the search window. By this extension, the method can be applied to image sequences which contain a changing shape of the tracked color distribution.

The CAMshift algorithm is applied to both left and right frames of the stereo vision system, yielding the two center points \((x_{cl}, y_{cl})\) and \((x_{cr}, y_{cr})\).

In the next step, the left center point \((x_{cl}, y_{cl})\) is searched in the right frame via a two dimensional block matching search. This search uses the two dimensional normalized cross correlation \(R(x, y)\).

The position with the maximum value of \(R(x, y)\) in the right frame represents the corresponding point \(M_{Pr} = (x_{MPr}, y_{MPr})\) to the CAMshift center point \((x_{cl}, y_{cl})\) in the left frame, which logically represents the left correspondence point, i.e. \(M_{Pl} = (x_{MPl}, y_{MPl})\) = \((x_{cl}, y_{cl})\). After correct detection of the matching points, they are forwarded to the PSO fusion block introduced in the following section.

V. AUDIOVISUAL INFORMATION FUSION AND TRACKING

The fusion module takes the information conveyed by the audio and vision algorithm and delivers an estimation of the 3D position corresponding to the tracked object, relative to the system origin. Before introducing our novel PSO-based audiovisual fusion approach, we will recapitulate the main steps of a well-known Kalman fusion technique which will serve as a reference system to compare the accuracy, performance, and execution time of the our new fusion algorithm.

A. Kalman-Based Fusion

As a reference tracking system, we implemented a three dimensional tracker using triangulation [17] and a Kalman filter [18].

Since our system deploys only one microphone pair and thereby only azimuthal localization is possible, we use an approach different from the decentralized system presented in [1]. In our approach, a basic linear Kalman filter, is applied to the tracking system. It uses a linear state transition matrix and adds the audio information by calculating an \(X_{audio}\) additionally to the \(X_{vis}\), using the relation \(X_{audio} = Z_{vis} \times \tan(\varphi)\) analog to the method explained in section V-B1, see also Figure 2. The \(X_{audio}\) is then added to the system by expanding the measurement vector \(z\) by one entry.

B. PSO-Based Fusion

The basic concept of our PSO-based audiovisual object tracker is illustrated in Figure 1. Every particle \(M\) represents a position in the three dimensional space, i.e. \(M \in \mathbb{R}^3\), in a coordinate system relative to the audiovisual system origin. With the position information obtained from the audio system \(\varphi\), and from the vision system \(MP_l\) and \(MP_r\), the particles moving in the three dimensional space can test a fitness function at their current position by calculating the angular and Euclidean distances to the positions favored by the audio and video system, respectively.

1) Audio: In order to evaluate the current position of a certain particle with reference to the azimuth angle obtained by the audio system as illustrated in section III, an audio distance \(D_{audio}\) is introduced. This variable represents the angular distance in radians between the audio azimuth angle \(\varphi\) and the angle \(\alpha\) which lies between the particle’s current position and the audio system origin. This is illustrated in Figure 2. The distance \(D_{audio}\) is normalized by \(\pi\), which is the greatest possible angular difference between \(\alpha\) and \(\varphi\).

\[
D_{audio} = \frac{|\varphi - \alpha|}{\pi},
\]

with

\[
\alpha = \arctan \left( \frac{Z_M}{X_M} \right),
\]

where \(Z_M\) and \(X_M\) are the \(Z\) and \(X\) coordinates of the particle’s position, respectively. Since the audio angle \(\varphi\) represents an azimuth angle, \(\frac{Z_M}{X_M}\) is equivalent to the tangent of the azimuth angle \(\alpha\).

2) Vision: In order to evaluate the particle’s current position with respect to the stereo vision system, the particle is projected on the left and right frame. This process leads to the left and right image points \(m_l = (x_{ml}, y_{ml})\) and \(m_r = (x_{mr}, y_{mr})\), respectively. Using a calibrated stereo camera system, the projection is obtained by

\[
m_l = P_l \cdot M, \tag{10}
\]

\[
m_r = P_r \cdot M, \tag{11}
\]

where \(P_l\) and \(P_r\) are the projection matrices for the left and right frame. These are defined as

\[
P_l = \frac{KK_l \cdot [I][0]}, \tag{12}
\]

\[
P_r = \frac{KK_r \cdot [R][t]}, \tag{13}
\]

under the assumption that the origin of the left image plane \(O_l\) is regarded as the system’s origin. The matrices \([I][0]\) and \([R][t]\) describe the homography between the left and right frame in a homogeneous coordinate system. \(R\) and \(t\) denote the rotation matrix and the translation vector, respectively. The part \(I\) represents the identity matrix. The terms \(KK_l\) and \(KK_r\) delineate the camera matrices.
The normalized values $D_{\text{left}}$ and $D_{\text{right}}$ represent the Euclidean distances between the projection of the current position of a particle $M$ on the left and right image frame, i.e. $m_l$ and $m_r$, and the corresponding localization points from the vision system, $MP_l$ and $MP_r$. This relation is defined as

$$D_{\text{left}} = \frac{\sqrt{[x_{MPl} - x_m]^2 + [y_{MPl} - y_m]^2}}{\text{width} + \text{height}},$$

$$D_{\text{right}} = \frac{\sqrt{[x_{MP_r} - y_m]^2 + [y_{MP_r} - y_m]^2}}{\text{width} + \text{height}},$$

where $\text{width}$ and $\text{height}$ are the width and height of the left and right image frames. These parameters are illustrated in Figure 2.

3) **Fitness Function:** As explained in section II, each particle tests the quality of its current position in every iteration by calculating its fitness function $F$. This function must decrease when the particle’s position is close to the object to be tracked in the solution space. The distances $D_{\text{left}}$ and $D_{\text{right}}$ measure the position quality with respect to the video-based localization module. Likewise, $D_{\text{audio}}$ directly measures the quality with respect to the audio-based module. Consequently, the adequate fitness function $F$ is defined as a weighted sum of these three distance values

$$F = w_{\text{audio}} \cdot D_{\text{audio}} + w_{\text{left}} \cdot D_{\text{left}} + w_{\text{right}} \cdot D_{\text{right}},$$

where $w_{\text{audio}}$, $w_{\text{left}}$, and $w_{\text{right}}$ denote weighting factors of each component.

The PSO fusion algorithm delivers a 3D location estimate of the tracked object and stops iterating based on two criteria. This happens when a predefined maximum number of iterations has been executed, or when a minimum value $F_{\text{min}}$ for the fitness function $F$ is reached, i.e. the convergence criterion is fulfilled. The 3D position $\text{gbestPos}(X, Y, Z)$ of the global best solution $\text{gbest}$ that has been encountered so far represents the current position of the tracked object.

VI. COMPARISON AND RESULTS

To test our PSO based 3D tracker, audio and video recordings of a person moving and talking in an area facing the stereo camera and stereo microphone system were taken in the LDV lab of the TU Munich. Distorting noise was produced by the fans of several PC’s, distributed in the room. The hardware deployed consisted of two firewire cameras and two AKG omni-directional microphones. The videos were shot with 15 frames per second with a resolution of $640 \times 480$ pixels. The audio material was recorded using a sampling frequency of 44100 Hz. The algorithms were implemented in C++ using the OpenCV library [19] and tested using a 1.6 GHz AMD processor with 512 MB RAM.

Averaged over 400 frame pairs of video, our PSO based tracker requires a mean computation time of 100.42 ms, while keeping mean squared errors comparable to the Kalman based reference tracking system, i.e. an average of 10 cm deviation between both methods. Compared to this, 106.2 ms are needed by the Kalman tracker when applied to the same recording material. It should be mentioned that approximately 80 to 90 ms are required by the preprocessing steps of audio and visual localization. This shows that both Kalman and PSO based tracking modules are using a minor part of the overall execution time. It should be noted that, the linear triangulation method combined with a linear Kalman system are one of the fastest state-of-the-art tracking systems today. More complex systems achieve similar tracking performance, under the same experimental conditions, yet demanding higher computational costs.

With a mean execution time of about 100 milliseconds per frame, which was achieved without any code optimization, and without program parallelization or hardware accelerators, the 3D tracker implemented in this study can be regarded as a low delay tracker, which is dedicated for real time implementation.

VII. CONCLUSION AND FUTURE WORK

We present a novel 3D object tracking system based on fusing audiovisual information. Our PSO based fusion approach does not need any manual initialization and no models, statistical, or learning. It therefore overcomes the problems of other audiovisual fusion methods, that base on several assumptions regarding the distributions of variables, or that tend to become complex by reducing these limitations. Speed performance was shown to be slightly faster than the simplest of the existing systems, the Kalman tracking. It therefore outperforms more complex methods like particle filtering or Bayesian inference, which tend to become time computationally expensive. A further advantage of the algorithm is that it is robust to local minima, due to the iterative interaction between particles, and is therefore insusceptible to false localization.

In Future Work, the PSO object tracker is to be adapted for multi object tracking. Towards this end, the modules of the tracker must be appropriately parallelized. Another possible extension is to include the information of additional sensors, like e.g. range sensors. For this purpose, the PSO systems fitness function can easily be adopted by adding the weighted additional information to the fitness function and adapting the convergence criterion. Furthermore, the presented three dimensional particle search may be a replacement of existing triangulation functions and moreover a new and fast ability to create complex disparity maps out of point correspondences.
REFERENCES


