Combined Visual-Acoustic Grasping for Humanoid Robots

G. Milighetti, T. Emter, H.-B. Kuntze
Fraunhofer Institute IITB
Systems for Measurement, Control, and Diagnosis (MRD)
Fraunhoferstraße 1, 76131 Karlsruhe, Germany
{milighetti,kuntze}@iitb.fraunhofer.de

D. Bechler, K. Kroschel
Universität Karlsruhe (TH)
Institut für Nachrichtentechnik
Kaiserstr. 12, 76128 Karlsruhe, Germany
{bechler,kroschel}@int.uni-karlsruhe.de

Abstract – This paper presents an efficient approach for the combined audio-visual localization of a fallen object. The localization algorithm is incorporated into a multi-sensor robotic platform to supervise and to adapt dynamically the discrete plan of a robot performing a pick-and-place task. Initial audio position estimates of the fallen object are delivered by means of time-delay estimates in two microphone pairs. For a successful operation even in acoustically adverse environments, reliability criteria for the time-delay estimates are introduced. These audio position estimates are improved by visual data using the disparity information in stereo images of a stereo-camera system. The performance of the audio-visual localization algorithm is evaluated for real data in a common office environment. The proposed real-time system shows robustness and accuracy localizing with ease a fallen wooden brick.

Index Terms – audio-visual localization, time-delay estimates, multi-sensor platform, pick-and-place task.

I. INTRODUCTION

Nowadays, the audio-visual scene analysis becomes more and more important in many areas. A typical application regards teleconferencing systems to steer electronically the camera to the active speaker without a human operator [1]. But also in surveillance systems the audio-visual perception is of great interest. For instance, in security domains undesired intruders shall be detected or public scenes have to be observed, e.g. in a tramway, to detect and prevent hooliganism [2]. A further example affects the observation of a driver. By means of acoustic and visual data the vigilance of the driver can be detected to avoid accidents due to fatigue or inattention.

In the last few years a new application has gained increasing interest: the audio-visual scene analysis for autonomous humanoid robots. In order to build collaborative human-friendly robots which are able to respond appropriately to their users’ needs, the robots have to be equipped with acoustic and visual sensors. Up to now, the localization and tracking of the users and the detection of their communicative cues such as gestures, gaze direction and speech recognition were the principal fields of research. Another very important aspect, which was somehow neglected in the past, is to guarantee the security of the users. One main goal of humanoid robots is to support handicapped or elderly people in household environments. For these persons it is fundamental that the robot is able to compensate for their deficits to hear and to see. A characteristic example regards the detection of a fallen object which might be a potential obstacle for the user. In this paper, a method for the robust audio-visual localization of such an event is presented. The proposed algorithm is incorporated into a real-time demonstrator of a multi-sensor robotic platform performing a pick and place task, which is interrupted by the event of a fallen wooden brick.

The paper is organized as follows: first an overview about the experimental platform is given in the next section. The algorithm developed for the audio localization and the one used for the visual localization are then presented in section III. After a brief description of the control architecture in section IV some experimental results are shown in section V.

II. EXPERIMENTAL PLATFORM AND TASK DESCRIPTION

The experimental platform assembled at Fraunhofer Institute IITB (Figure 1) for the development and investigation of the presented algorithm consists of two modular 7DoF AMTEC robot arms (A) and a 2DoF pan-tilt sensor head (B). For the communication and interaction with its environment the robot is equipped with various visual, acoustic and force sensors. The head is provided with a stereo camera able to localize predefined objects and an acoustic sensor (microphone array) able to determine the position of sound sources (Figure 2). Moreover an optical 3D laser stripes sensor (C) for a more accurate localization of objects in the close range is integrated in one gripper and two force-torque sensors are mounted on the wrists of the two arms (D). Finally a slip sensor (E) can be installed in one of the grippers.
In order to demonstrate the audio-visual localization algorithm a case study typical for a household environment will be considered within this paper. While the robot is performing a “pick and place” task to transport an object between two points (e.g. taking different ingredients and putting them into a pan), two different events can occur:
- the carried object is lost → the robot has to pick it up and go on with the primary task;
- another object is dropped down in the workspace of the robot → the robot has to remove it before going on with the primary task.

In both cases a small wooden brick will be used as object in order to provide some experimental results.

The robot solves the unexpected problem behaving like a human: first it detects by means of the audio array the impact and estimates its rough direction, then it turns the head in that direction looking for the object. Depending on the identified situation the correct strategy in order to reach the primary goal is then performed.

The algorithms which realize this behavior will be presented in the next section, while the architecture on which the control strategy is based will be presented in section IV.

III. MULTI-SENSOR DETECTION AND LOCALIZATION ALGORITHM

This section proposes an efficient method to reliably localize fallen objects with the help of a microphone array and a camera system while the robot is performing several tasks, i.e. the arms of the robot are moving. Therefore the assets and drawbacks of acoustic and visual localization have to be considered to be able to combine both modalities to a robust complete system. Visual systems offer high accuracy, but they have to cope with a restricted view angle, with changing illumination and with a relatively high computational load. In contrast, acoustic localization systems allow an omnidirectional view angle with low computational costs so that it is possible to sense the full scene without scanning. On the other hand acoustic systems suffer from relatively low accuracy and have severe problems with background noise and room reverberation. The fusion of acoustic and visual perception by using the acoustically located direction to steer the field of view of the camera to the fallen object combines the advantages of both methods.

For a visual localization in 3D of an object, a compact stereo-camera system can be adopted. The requirements for the microphone array are a limited size so that it can be easily integrated into the robot, as well as a small number of microphones for computational efficiency. The sensor head as shown in Figure 2 consisting of a stereo-camera and 2 microphone pairs M1M2 and M3M4 in a cross-configuration satisfies these specifications. The two cameras are 9 cm away from each other and the distance of two microphones in a sensor pair is 30 cm resulting in a compact size of the complete sensor head.

A. Acoustic Sound Source Localization

One major requirement for the localization algorithm is its real-time constraint demanding computational efficiency. In addition, the chosen method has to be able to localize short sound events which necessitates the analysis of the incoming signals in short frames. A last request regards the robustness of the algorithm against background noise, room reverberation as well as noise produced by the mechanics of the robot.

In order to determine the direction of arrival of a sound source with a microphone array several approaches are available [4]. To fulfill the above mentioned requirements the acoustic localization based on the Time Difference Of Arrival (TDOA) estimation [5] in microphone pairs seems to be the best choice as it needs only few microphones, has a relatively high spatial resolution and is computationally efficient.

A.1. TDOA-based Localization

The TDOA-based source localization is performed in two steps. First, the TDOAs of different microphone pairs are estimated. In a second step these TDOAs are used together with the known sensor positions to determine the direction of the sound source.

Signal Model

For a given pair of spatially separated microphones \( M_i \) and \( M_j \) the recorded sensor signals \( x_i(t) \) and \( x_j(t) \) coming from a remote sound source \( s(t) \) can be modeled mathematically as

\[
\begin{align*}
x_i(t) &= h_i(t) * s(t) + n_i(t) \\
x_j(t) &= h_j(t) * s(t - \tau_{ij}) + n_j(t),
\end{align*}
\]

where \( \tau_{ij} \) denotes the relative time delay of interest, * signifies the convolution operator, \( h_i(t) \) represents the acoustic impulse response between the source and the \( i^{th} \) sensor and the additive term \( n_i(t) \) summarizes the channel noise of the microphone system as well as the environmental noise recorded by the \( i^{th} \) sensor.

TDOA Estimation

The most popular approach for determining the TDOAs is the Generalized Cross-Correlation (GCC) method [6]. The...
TDOA of interest \( \tau_{ij} \) is estimated at the time lag corresponding to the global maximum in the GCC function

\[
\hat{\tau}_{ij} = \arg\max R_{ij}^{GCC}(\tau).
\]  

(2)

The GCC function is defined as

\[
R_{ij}^{GCC}(\tau) = \int_{-\infty}^{+\infty} \Psi_{ij}(\omega)X_i(\omega)X_j^*(\omega)e^{i\tau\omega} d\omega,
\]  

(3)

where \( X_i(\omega) \) is the Fourier transform of \( x_i(t) \) and \( \Psi_{ij}(\omega) \) is a weighting function trying to decrease noise and reverberation influence. In real environments the Phase Transform (PHAT) weighting has shown best performance [4] and is defined as

\[
\Psi_{ij}^{PHAT}(\omega) = \frac{1}{\left| X_i(\omega)X_j^*(\omega) \right|}.
\]  

(4)

Determination of the Direction of Arrival

For a certain TDOA in a given pair of microphones all possible source positions are located on a hyperboloid which can be approximated by a cone. The estimated apex angle defining the opening of the cone is given by

\[
\hat{\alpha}_{ij} = \arccos \left( \frac{\hat{\tau}_{ij} \cdot c_s}{d} \right),
\]  

(5)

where \( d \) represents the distance between the two sensors and \( c_s \) signifies the velocity of sound.

For two pairs of microphones in a cross-configuration as depicted in Figure 2 the two corresponding cones intersect in two straight lines (Figure 3). The arising forward-backward ambiguity for the direction of arrival of the sound source is in the present case unproblematic, as the microphones are placed near a wall and thus the potential source position is restricted to the half-space in front of the array.

A.2. Signal Processing

The sampling frequency for the audio signals is \( f_s = 16 \, \text{kHz} \). The captured data are analyzed in frames of 8 ms to assure quasi-stationarity and to allow the detection of short acoustic events as e.g. the sound of a fallen object. For the segmentation a Hamming window with 75% overlap is used delivering every 2 ms a new TDOA estimate. For inter-sample precision quadratic interpolation is applied to the TDOA estimates by fitting a parabola through the maximum of the GCC function and its two direct neighbors.

A.3. Reliability Criteria and Data Clustering

Although the GCC approach seems to be very practical, its application in real acoustic environments is only of limited use. Even in mildly reverberant rooms the rate of the estimation error rises strongly delivering unreliable estimates. Therefore reliability criteria are required to evaluate the confidence of every single estimate.

As shown in [7] the value of the maximum peak in the GCC function can be used as an appropriate reliability indicator. Figure 4 shows the percentage of correct TDOA estimates of the two microphone pairs against the values of the maximum peak in the GCC function. As can be clearly seen the criterion of the maximum peak allows very convincingly a judgment about the current TDOA estimate: low criterion values mean low reliability whereas for increasing values of the criterion highly reliable estimates are delivered. Additionally, psycho-acoustical knowledge can be use to improve the TDOA estimation [3]. In closed reverberant rooms the human auditory system uses the so-called precedence effect [9] or law of the first wave-front. To help the listener in determining the position of a sound source, the human auditory system focuses on the direct sound wave, i.e. the first wave-front, ignoring the information contained in reflective, later arriving wave-fronts.

The typical length of a sound event coming from a fallen object is approximately 200 ms delivering for the chosen signal processing about 100 TDOA estimates. Since the sound of a falling object is a single acoustic event, only one final TDOA estimate is needed, suggesting the use of clustering methods of consecutive estimates. The proposed method to reduce all the estimates to a unique one takes the two reliability criteria, i.e. the maximum peak of the GCC function and the precedence effect, as well as averaging techniques into account. To consider the precedence effect, taking only the first 8 TDOA estimates of a sound event (corresponding to the
first 22 ms) has shown best performance. As the reliability criterion of the value of the GCC maximum is incorporated in the fitted parabolas of the quadratic interpolation, the 8 corresponding parabolas of the first 8 TDOA estimates are accumulated. The resulting maximum in the accumulated GCC function finally delivers a unique TDOA estimate in every microphone pair for the sound event of a fallen object.

A.4. Results of the Acoustic Localization Algorithm
The presented algorithm has been tested using a wooden brick of the size 9 x 2.7 x 1.9 cm³, which has been dropped from a height of approximately 10 cm onto a wooden plate. Assuming the brick in the signal model as a point source the localization algorithm has to cope with an intrinsic error between ±2° and ±4° due to the size of the brick, depending on the distance from the array which ranges from 0.8 m to 1.5 m. The brick has been dropped at 12 defined positions 100 times, respectively. The multi-sensor platform is situated in a room measuring 4 x 5.5 x 2.8 m³. The 60 dB reverberation time $T_{60}$ of the room was experimentally determined by means of Schroeder’s backward integration method [8]. With a relatively high value of $T_{60} = 550$ ms previous acoustic localization systems based on the GCC function were not any longer able to deliver robust and reliable TDOA estimates. On the other hand, the Signal-to-Noise Ratio (SNR) was unproblematic with about 38 dB of the sound event of the falling brick while the robot is moving one arm in a circular motion above the table.

In Figure 5 the estimates of the localized positions on the wooden plate are shown. While the position estimates near the array are distributed closely around the true position, the estimates further away tend to be localized slightly off target but are still accurate enough to be able to steer the stereo-camera to the fallen object.

Considering that the view angle of the camera is about 40° the localization of the sound source has been deemed correct when the absolute value of the difference between the true cone angle $\alpha_{ij}$ of the horizontal and vertical microphone pair and their respective estimates is less than 20°

$$|\alpha_{ij} - \hat{\alpha}_{ij}| \leq 20°.$$ 

Using this evaluation method a percentage of correct localization estimates of 99.75% was achieved. In Figure 6 the error angles of all 1200 recordings are shown. The mean of the horizontal error angle has a bias of 1.0°, the vertical one a bias of -0.7°, respectively. Both errors are very low compared to the uncertainty due to the dimensions of the brick. The standard deviation is given by 1.9° for the horizontal sensor pair and 3.1° for the vertical one, respectively. These convincing results for the acoustical localization allow to direct the stereo-camera robustly and reliably to the direction of the fallen object.

B. Visual Localization
For the visual localization commercial algorithms are used. First a disparity image from the two black and white mono images is computed by means of the Small Vision System (SVS) [13]. Once the disparity image is computed, it can be further processed by the user.

Successively a pattern matching algorithm provided by the software Halcon [14] is performed looking in one of the two mono images for the pattern of the object to localize. Thus, an area of interest in the image is determined and the 3D position of every point in this area is extracted by the previously stored disparity image.

Building the center of gravity of the point cloud the 3D position of the object in camera coordinates can be delivered and finally transformed into the coordinate system of the robot.

The merging of the two acoustic and visual measurements in order to reach a higher reliability in the position estimation will be shown in the next section presenting some experimental results.
IV. CONTROL ARCHITECTURE

The basic structure of the control concept developed in order to supervise the robot throughout its task is shown in Figure 7. The state of the robot interacting during its task with a human operator and of the environment is supervised with the help of internal (e.g., encoder) and external sensors (e.g., camera, microphone, force-torque).

In the upper hierarchy level, a discrete control processes the measurements coming from the sensors and fuses them in order to generate diagnosis signals that contain quantitative information about the continuous state of the system (e.g., position of objects, sounds, forces). As a second step, this information is used for the identification of the discrete state or event providing the qualitative information about the present situation (e.g., a certain event has occurred). By interpreting the acquired knowledge about both the continuous and discrete state of the system, a decision unit generates a new appropriate task sequence or it adapts the present one in order to overcome the unexpected situation and to reach the initial desired goal.

In order to perform a flexible task sequence generation and to have the possibility of a fast on-line task adaptation, a discrete task structure has been developed which assures a transparent and efficient instrument for the on-line decision and is also easily accessible by the decision unit. The best way to obtain such a task structure is to base its architecture on a sequence of modules. Every task can thus be seen as the result of the execution of a chain of elementary actions called Primitive Skills (PS) each with its own sub-goal.

An example of the simple “pick and place” task implemented with a sequence of PS is given in Figure 8. Notice also the sub-task planned in case of loss of the load localizing it by means of the stereo camera and of the laser stripes sensor (LSS).

Once the decision unit has determined the PS sequence that has to be performed and the most appropriate controller in order to execute the currently active PS, in the lower hierarchy level the continuous control assures that the currently active PS is managed by the optimal specific controller.

More details about a Neuro-Fuzzy based diagnosis concept and about PS can be found in [11] and [12].

V. EXPERIMENTAL RESULTS

In order to validate the proposed concept, the scenario described in section II has been implemented on the test platform at IITB. As previously introduced, one arm performs a basic pick and place task moving continuously an object from position 1 to position 2 and back while the second arm is inactive.

The sound of an impact with the table lets the robot interrupt the sequence of the PS loop building the pick and place task. The new PS \texttt{LookByStereoCamera} is then dynamically initialized using the elevation and azimuth angles delivered by the microphone-array and introduced in the plan.

Thanks to the flexibility of the PS-based structure briefly presented in the previous section, such modifications of the actual plan can be made on-line without any problem, simply inserting in the existing PS sequence the needed new ones with commands like \texttt{InsertBehind}, \texttt{InsertBefore}.

Once the camera has been moved in the direction of the expected sound source, the software for the localization is activated (Figure 9) and the results of the localization algorithm are combined with the estimation of the audio array. Comparing so the positions delivered by the two different sensors and always taking into account the actual state of the robot arms an optimal control strategy able to overcome the unexpected situation can be planned.

The case shown in Figure 10-a leads for example to the conclusion that the sound occurred due to the loss of the carried object. The two measurements are in fact consistent.
and located close to the position of the Tool Center Point (TCP) of the robot. Of course information about the state of the gripper can be also used to make the decision more reliable. In Figure 10-b it can be observed how the two measurements are still consistent but located in a region far from the actual working area of the robot. Hence it can be concluded that the impact has been caused by a second object and not by the carried one. To remove this external object form the workspace the robot has to place the actual load and can then proceed with the removal action. Finally in Figure 10-c two inconsistent measurements are shown. In this case a more accurate investigation (maybe with the hand camera) is required before a decision can be made.

In all the considered situations, once the action needed to reestablish the initial conditions has been executed, the primary “pick and place” task can be resumed. The presented algorithm is implemented in real time confirming the convincing results shown in the simulations. The accuracy of the combined audio-visual localization is always sufficient for the execution of a safe grasping sequence.

The robustness of the proposed algorithm was proven by a large amount of successful experiments and also by replacing the wooden brick with different other objects as aluminum bricks or a spoon.

VI. CONCLUSION

In this paper an audio-visual algorithm to determine the correct position of a fallen object was presented. In a first step, the direction of the sound event was calculated acoustically based on Time Difference Of Arrival estimates in microphone pairs.

This initial estimation is improved by a visual localization system based on pattern matching algorithm, thus combining the advantages of both, the audio and video modality.

The proposed concept has been successfully implemented in a real time environment and tested on the multi-sensor robotic platform available at Fraunhofer Institute IITB. The accuracy and robustness of the algorithm were both very satisfactory.

ACKNOWLEDGEMENT

The provided results are part of the cooperation between the Fraunhofer Institute IITB and the Institut für Nachrichtentechnik of the University Karlsruhe within the framework of the research project Sonderforschungsbereich (SFB) No. 588 “Humanoid robots”. The SFB is supported by the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES