社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-10

MULTI-MODAL SOUND LOCALIZATION FROM A MOBILE PLATFORM

Jani Even, Nagasrikanth Kallakuri, Yoichi Morales, Carlos Ishi, Norihiro Hagita

ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan
even@atr.jp

ABSTRACT

This paper presents a multi-modal sound source localization method for mobile platforms. The sound source localization is performed while the robot is autonomously navigating through the environment by combining the power and bearing estimation given by a steered response power (SRP) algorithm with the range estimation obtained from the on-board laser range finders (LRF). First the positions of the sound sources in the environment are determined by taking into account the robot pose which is estimated with a particle filter and the estimated power is accumulated in the cells of a grid map covering the environment. Finally, a local maxima search is performed on this grid map to find the area with higher estimated sound power that correspond to the sound source locations.

1. INTRODUCTION

Sound source localization has been a topic of interest in the audio processing community for a long time (see [1]). The most effective techniques that emerged are either based on the estimation of the time delay of arrival at microphone pairs, or on the estimation of a steered response power (SRP) or on spectral decomposition techniques like the MUSIC algorithm. All these approaches rely on the use of microphone arrays. Using a robot, it is possible to explore the space and effectively extend the operational range of sound localization. Thus a natural framework for sound source localization from a robot is to use a conventional sound localization algorithm at different locations and combine the results from all these different locations [2, 3, 4, 5]. Since these localization results are obtained for different times, it is important to distinguish between fixed sound sources and moving sound sources. In this paper, we are interested in the localization of the environmental noises that are fixed sources.

The authors in [5] rely on triangulation to estimate the positions of the sound sources using audio scans taken by an autonomous mobile robot. One of the very interesting approaches in this area is the use of evidence grids in [3, 6]. The space to be explored is partitioned into grid cells of fixed size. Then the probabilities of having a sound source in each of the cells are estimated during the exploration. To achieve this, at a given location, an SRP with phase transform (SRP-PHAT [7]) is estimated for a grid centered on the robot and these estimated powers are used to update the evidence grid. In that method, the robot is tele-operated to gather sound data in the vicinity of the sound sources [6].

In this paper, we present a framework for localizing sound sources using an autonomous mobile robot equipped with a microphone array. The novelty of the present work is in obtaining the audio information about the environment using a multi-modal approach. In particular, the laser range finder (LRF) data are explicitly used during sound source localization. Most autonomous mobile robots are equipped with LRFs and odometry (obtained from the encoders on the wheels) to localize themselves in the environment. This ability to estimate precisely the range of the objects around the robot is exploited in this paper to solve the problem of poor range estimation from audio localization techniques. In the proposed approach, the audio modality is used to estimate the bearing of the sound sources whereas their ranges are obtained using laser range finders. Consequently, our approach assumes that the geometric coordinates of sound sources are detectable by the LRFs on the robot. While this assumption is a bit restrictive when a two dimensional horizontal plane is scanned, it will be reasonable when extended to a three dimensional scan. In the proposed framework, sound source localization is performed while the robot is autonomously navigating through the environment. During navigation, the two on-board LRFs (front and back) provide range scans and a steered response power (SRP) algorithm generates audio scans. The SRP gives the bearing of the candidate sound sources and an estimate of the received audio power. Combining the bearing, the received power and the range information, an estimate of the emitted power from candidate sound sources is computed. The audio and LRF scans are acquired at regular intervals and for each of the audio scan, the power of the most powerful emitting candidate sound source is accumulated on a grid map that covers the environment. The cells from that grid map contains an average of the estimated emitted power and a count of the number of visits (the number of time a cell has been selected). This procedure requires to transform the sound source positions from the

THIS RESEARCH WAS FUNDED BY THE MINISTRY OF INTERNAL AFFAIRS AND COMMUNICATIONS OF JAPAN UNDER THE STRATEGIC INFORMATION AND COMMUNICATIONS R&D PROMOTION PROGRAMME (SCOPE).

robot referential to the room referential. This is performed by taking into account the robot pose which is estimated with a particle filter. Finally, a local maxima search on this grid map finds the locations of the sound sources in the environment by selecting the cells with higher power.

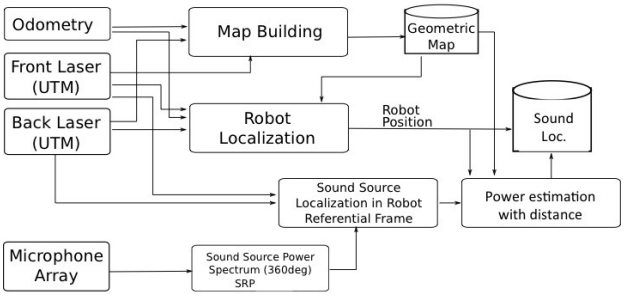


Fig. 1. Block diagram of the system.

2. PROPOSED APPROACH

A block diagram describing the proposed approach is shown in Fig. 1. The main processing blocks will be described in the following sections.

2.1. Map Building

The map building is performed in advance. The aim is to create a map that describes the environment in which the robot is due to navigate autonomously. The environment is represented by an occupancy grid, namely a grid which cells are either empty (open space) or occupied (walls and structures). The occupancy grid map is referred to as the *geometric map* in the remainder.

In this work we use them for building grid maps. To build the map, we controlled the robot with a joystick through the environment gathering odometry and laser sensor information. Then we used iterative closest point based SLAM to correct the trajectory of the robot and to align the laser sensor scans [8] using the $3DToolkit$ library framework [9], [10]. With the resulting aligned scans the occupancy grid map was created [11], [12]. The map obtained for one of the test environments is shown in Fig. 2.

2.2. Robot Localization

The goal of the robot localization task is to precisely estimate the pose (location and orientation) of the robot in the geometric map representing the environment. We used a particle filter approach to estimate the robot position with a weighted set of M particles. Each particle has a pose given by the state vector $\{x_m(k), y_m(k), \theta_m(k), w_m(k)\}$ containing a candidate position and orientation of the robot and the associated weight. While the robot moves, each particle also moves based on the odometry readings and the probabilistic motion model, which describes the uncertainty in the robot motion (prediction step). In the correction step, the particle filter estimates the posterior density by considering measurement likelihood. This likelihood is estimated from the LRF scans using the ray casting approach likelihood model in [13]. The map update depends on the state of the particle dispersion and the matching of the laser scans. The particles, which are more likely to be correct after ray casting, have a higher likelihood score, and therefore, more weight. Particle re-sampling is performed regularly and the robot pose $\{x(k), y(k), \theta(k)\}$ is given by the average weight of the particles.

2.3. Audio scanning

In order to present this framework in greater detail, let us first briefly describe the SRP approach to sound source localization (see [7] and references herein). The goal of sound source localization is to estimate the position of sound sources in a search space using the audio observation. At the sampled time k, the observed signals from the Q microphones of the array are $v_1(k), \cdots, v_Q(k)$. Because the geometry of the microphone array is precisely known, it is possible to *focus* the array using spatial beam forming to estimate the sound from a spatial location. The beam forming output is denoted by

$$s(k, [x, y, z]) = \mathcal{F}(v_1(k), \cdots, v_Q(k), [x, y, z]), \quad (1)$$

where $[x, y, z]$ are the coordinates of the focus point in the array referential. The SRP is obtained by computing the power of this output over T samples

$$J(k, [x_n, y_n, z_n]) = \frac{1}{T} \sum_{\tau=0}^{T-1} s^2(k - \tau, [x_n, y_n, z_n]) \quad (2)$$

for a set of N locations $[x_n, y_n, z_n]_{n \in [1,N]}$ in the search space. The locations corresponding to the peaks of the SRP gives the sound sources' positions. There are several ways to obtain the beam forming output, compute the power and select the set of locations. In the remainder, the SRP obtained at the time k that contains the power from the N locations is referred to as the kth audio scan.

In this paper, the SRP processing is done in the frequency domain after applying a short time Fourier transform (STFT) to the observed signals sampled at 48kHz (the analysis window is 25 ms long and the shift of the window is 10 ms). Then the SRP is computed for the frequency band $[1000, 6000]$ Hz using 10 STFT frames for averaging the power. Thus a new audio scan is available every 100 ms. A delay and sum beamformer is used to *focus* the observations.

A particularity of sound source localization algorithm is that the estimation of a source range is imprecise whereas its bearing estimate is accurate. Thus spherical coordinates $[\rho, \theta, \phi]$ are often used to describe the search space. Contrary to [3], we assume the far field conditions hold (ρ large compared to array aperture) and a bearing only scan is performed. Namely the kth scan is a set of N angles $\theta_n(k) \in [0, 2\pi]$ with their associated power $J_n(k)$. In our approach the distances are obtained using the LRF scans as explained in Sect. 2.4.

Note that these scans are computed in search spaces in the array's frame of reference (as the position of the focus point have to be known in the array's frame of reference). Thus it is necessary to transform the poses of the robot at these locations to a global coordinate system to localize the sound sources in the global referential.

2.4. Emitted power estimation

The audio scans $\{\theta_n(k), J_n(k)\}$ are in the robot coordinate frame and the goal of the fusion procedure is to combine them with the range estimation from the LRFs and the knowledge about the robot pose in the global referential in order to estimate the position of the sound sources in the geometric map.

The main idea is to use the range estimation in the directions given by the SRP and combine it with the estimate of the received powers to estimate the powers that was emitted by the potential sound source candidates. For this purpose, the phase transform is not used as it discards the amplitude of the signals of interest.

For each of the directions $\theta_n(k)$, an estimated range $\rho_n(k)$ is given by the LRFs (the closest ray in the LRF scans is selected). Then, for estimated ranges in $[d_{\text{min}}, d_{\text{max}}]$, the estimated emitted power is

$$C_n(k) = J_n(k) \left(\frac{\rho_n(k)}{d_{\text{min}}} \right)^{\alpha} \tag{3}$$

where α controls the effect of the distance on the power (in free field $\alpha = 2$). Namely, the received power is corrected by the estimated distance to the sound source candidate in order to compensate for the power drop during propagation between the source and the array, see the circles representing the propagation in Fig. 3. A maximum distance d_{max} is set because the audio power decreases rapidly with the distance and sound sources are covered by the background noise for long distances.

For each of the audio scan, only the largest emitted power estimate $C_m(k) = \max_n C_n(k)$, obtained for $\{\theta_m(k), \rho_m(k)\}$, is considered. By combining the robot pose with the maximum power location $\{\theta_m(k), \rho_m(k)\}$ a position in the global referential is obtained. That position correspond to a cell $\{i, j\}$ of a grid map covering the room. This transform is illustrated in Fig. 4.

The average estimated power $P_{ij}(k)$ of that cell is obtained by taking

$$P_{ij}(k) = \frac{P_{ij}(k-1)K_{ij}(k-1) + C_m(k)}{K_{ij}(k-1) + 1} \tag{4}$$

$$K_{ij}(k) = K_{ij}(k-1) + 1 \tag{5}$$

where $K_{ij}(k)$ denotes the number of time for which the cell $\{i, j\}$ is visited ($K_{ij}(0) = 0$). This count is also used to remove cells that have been seen very few times. The grid map containing the average power is referred to as *power map* in the remainder.

Then the sound source localization is performed by finding the cells that have higher power in the power map. Namely, a local maxima search algorithm is used on the power map to find the candidate sound sources.

In practice, a small neighborhood of the cell $\{i, j\}$ is selected and the power $C_m(k)$ is distributed in that neighborhood. The size of the neighborhood is taken as $\Delta\theta\rho_n(k)$ where $\Delta\theta$ is the angular resolution of the audio scan. The N cells in this neighborhood receive the power $\frac{C_m(k)}{N}$ and are counted as visited one time. This smearing of the power is performed in order to take into account the larger uncertainty for longer ranges.

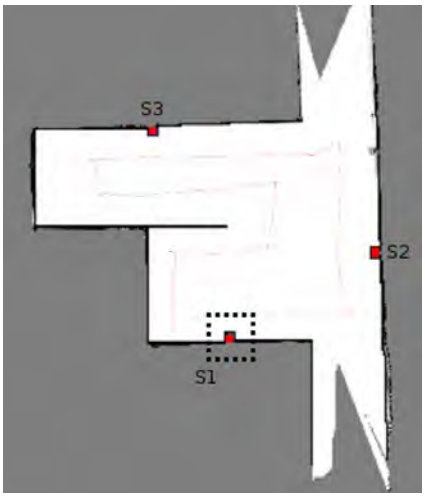


Fig. 2. Geometric map of the corridor with sound sources and robot trajectory.

3. EXPERIMENTS

For experimental validation we used a pioneer robot. This robot has a differential drive configuration and was equipped with two motor encoders and two laser range finder (UTM-30LX from Hokuyo, maximal range 30 m). The experimental platform can be observed in Fig. 5.

The microphone array is composed of 16 Sony ECM-C10 microphones mounted on a circular frame (diameter 31 cm). The audio capture interface is a Tokyo Electron Device

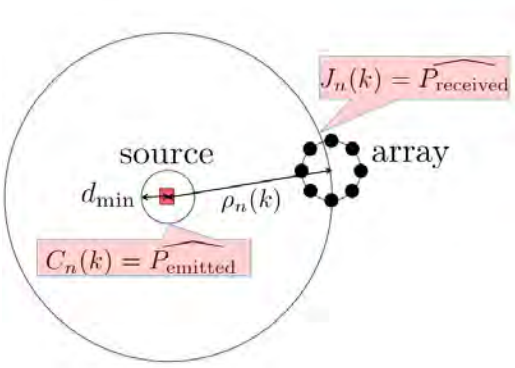


Fig. 3. Received and emitted powers.

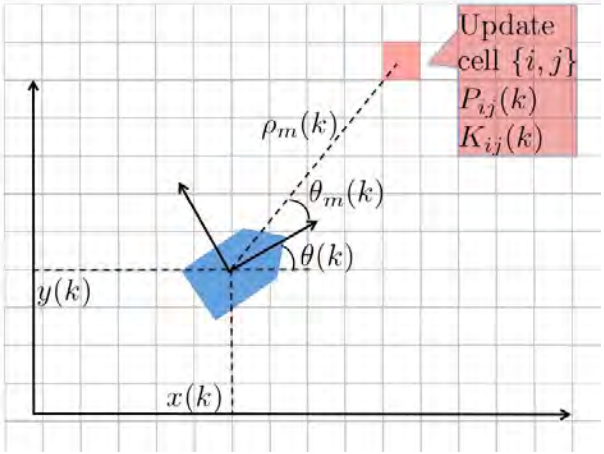


Fig. 4. Cell update using robot referential coordinates and robot pose.

Limited TDBD16AD-USB that samples the signals at 48kHz. The experimental evaluation of the approach was conducted in a corridor. Different sound sources of known intensities were setup in the environment.

Fig. 2 depicts the geometric map of the environment. The dimension of the cells in this map is 5 cm x 5 cm. The grid map used for localization has also 5 cm x 5 cm cells.

The robot navigates autonomously in the corridor using a set of way points that defined a loop covering all parts of the corridor In the remainder, a *run* corresponds to the robot performing one loop in the corridor. The sound source localization is performed during these runs.

Several (up to three) sound sources were placed in the environment (these locations are in the scan plane of the LRFs). These sound sources are loudspeakers playing recorded sounds. There was the sound of an air conditioning unit ($S1$ with a sound pressure of 78.5 dBA measured at 5 cm), the sound of a desktop computer fan ($S2$ at 77.5 dBA) and the sound of a server rack ($S3$ at 77 dBA). The sound pressure in the quiet corridor was around 42 dBA. The activation pattern

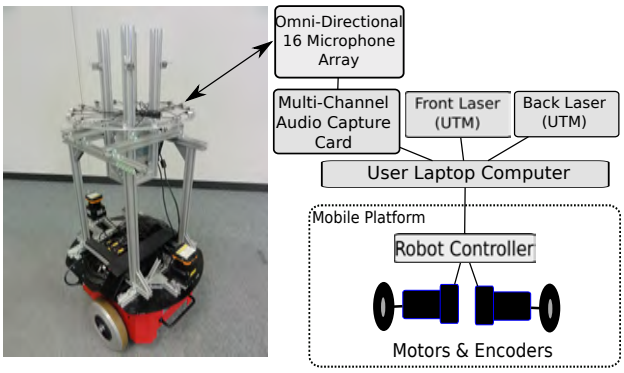


Fig. 5. Experimental robot platform with omni-directional microphone array and two laser range sensors.

Table 1. Parameters used during all runs.

$\Delta\theta$	α	$d_{\min}$	$d_{\max}$
3^o	0.5	0.3m	3m

of the sound sources for the runs can be observed in Table 2 and their positions are reference in Fig. 2. The parameters are given in Table 1. Note that α is set to a small value in order to avoid far estimate concentrating the power.

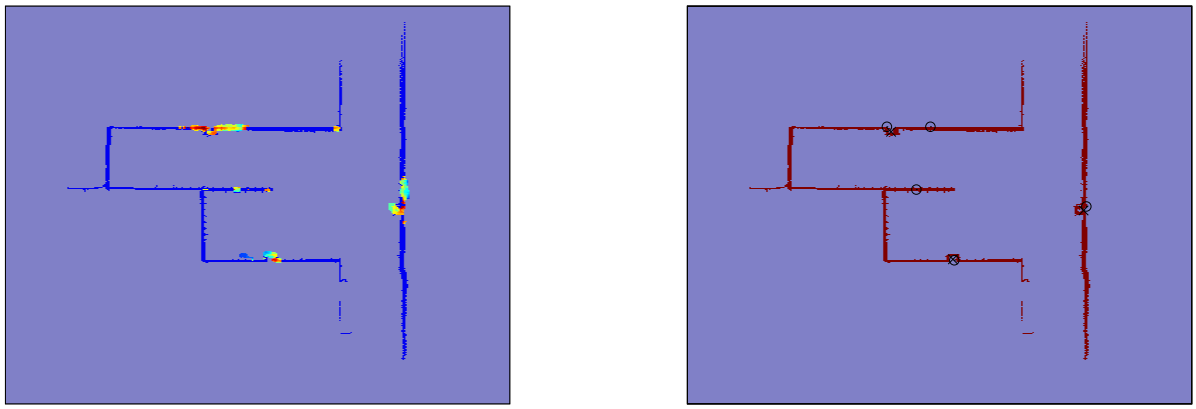
4. RESULTS

The power map is obtained by taking the accumulated power of the cells $P_{ij}(k)$ for which the number of visits $K_{ij}(k)$ is greater than 10% of the maximal number of visits. Thus an updated power map is available after each audio scan. Fig. 6-(a) shows the power map obtained at the end of the run 1 and Fig. 6-(b) shows the result of the local maxima search. The locations of the local maxima appear as black circles. The ground truth, i.e. the real positions of the sound sources are given as black crosses. For each of the sources the errors are given in Table 2. The results for run 2 are also in the table.

Figs. 7-(a) presents the power map difference (in dB) that is obtained by taking the difference of the power map for run 1 (three sound sources) and run 3 (no sound source). Figs. 7-(b) shows the same results for run 2 (two sound sources) and run 3 (no sound source).

5. DISCUSSION

The power map in Fig. 6-(a) illustrates the fact that large areas of higher power appear around the locations of the sources. Note that a few small areas with high power are also present in the power map. After local maxima search, the proposed approach successfully estimated the positions of the sound sources see Fig. 6-(b) (the three local maxima close to the true sources' location are the one with higher



(a) (b)

Fig. 6. (a) Power map for the run 1 , (b) local maxima search results (x) are the ground truth and (o) are the local maxima.

Table 2. Sound source localization results.

Run	Active Sources	Detected Sources	Error(m)
1	$S1$	$S1$	0.29
	$S2$	$S2$	0.22
	$S3$	$S3$	0.07
2	$S1$	$S1$	0.11
	$S3$	$S3$	0.18
3			

values). The average localization error was 0.17 m and the maximum error 0.29 m. Considering that the loudspeakers are not point sources but may span several cells and the localization was performed while moving, an error in the obtained range indicates a precise localization.

Another interesting point of the proposed approach is that the power maps contain estimates of the emitted power (these estimates are obtained by correcting the SRP without phase transform with a function of the estimated range, see Eq.3). Consequently, it makes sense to estimate sound source localization by using difference of power map in dB (equivalent to a power ratio). A *background power map* obtained when there is no sound source of interest (here the run 3) can be subtracted to a power map obtained when some sources of interest are present. In Figs. 7-(a) and (b), the difference of power maps clearly show the locations of the active sound sources. The local maxima search proved to be more easily conducted on the difference of power maps as they have larger dynamics and contains less false alarms (spurious local maxima). When it is not possible to obtain a good *background power map*, the local maxima search is to be applied on the power map.

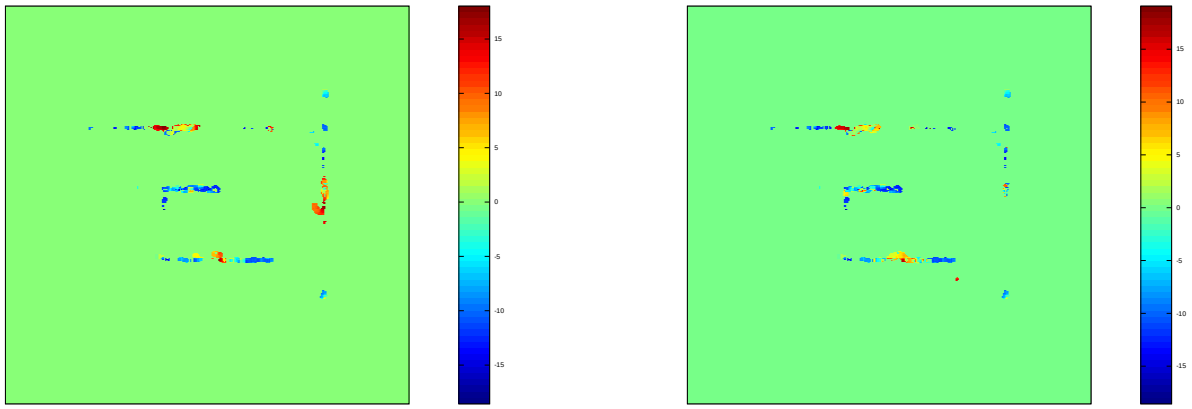
6. CONCLUSIONS

This paper presented a framework for localizing environmental sound sources using an autonomous mobile robot equipped with encoders, laser sensors and a 16 channel microphone array. The sound source localization results obtained in the experiments had an average distance error of 0.17 m using local maxima search, showing that the proposed framework is capable of localizing sound sources. Up to 3 sources within an 8m x 8m space were localized. The method is also robust towards false positive detections and noise effects produced by echoes. The novelty of the approach is in the combination of the audio scans with the LRFs scans. These kind of sound localization approach will aid the robot in attaining a better knowledge about environmental noise. It can be used for better speech recognition (suppressing the known environmental noise), effective human-robot interaction, and also for surveillance of environments. With the available framework, it is possible to extend the work to 3-Dimensional sound source localization that is more informative.

7. REFERENCES

[1] H. DiBiase, J. nad Silverman and M. Brandstein, *Microphone arrays : Signal Processing Techniques and Applications*, Springer-Verlag, 2007.

[2] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, sept.-2 oct. 2004, vol. 3, pp. 2123 – 2128 vol.3.



(a) (b)

Fig. 7. (a) Difference of power maps for run 1 and run 3 (scale in dB) , (b) Difference of power maps for run 2 and run 3 (scale in dB).

[3] Eric Martinson and Alan C. Schultz, "Auditory evidence grids.," in *IROS*. 2006, pp. 1139–1144, IEEE.

[4] K. Nakadai, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evalation," in *IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 561–566.

[5] Y. Sasaki, S. Thompson, M. Kaneyoshi, and S. Kagami, "Map-generation and identification of multiple sound sources from robot in motion," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2010*, 2010, pp. 437–443.

[6] Eric Martinson and Alan C. Schultz, "Robotic discovery of the auditory scene," in *ICRA*, 2007, pp. 435–440.

[7] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997*, 1997, pp. 375–378.

[8] P.J. Besl and H.D. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[9] D. Borrmann, J. Elseberg, K. Lingemann, Andreas Nüchter, and J. Hertzberg, "The Efficient Extension of Globally Consistent Scan Matching to 6 DoF," in *Proceedings of the 4th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT '08)*, Atlanta, USA, June 2008, pp. 29–36.

[10] slam6d, "Slam6d - simultaneous localization and mapping with 6 dof," Retrieved December May, 20 2011 from `http://www.openslam.org/slam6d.html`, 2011.

[11] Hans Moravec and A. E. Elfes, "High resolution maps from wide angle sonar," in *Proceedings of the 1985 IEEE International Conference on Robotics and Automation*, March 1985, pp. 116–121.

[12] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, June 1989.

[13] Sebastian Thrun, Wolfram Burgard, and Dieter Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, The MIT Press, 2005.