Acoustic Simultaneous Localization and Mapping Using a Sensor-Rich Smartphone

Xi Yu Song[1(✉)], Mei Wang[2], Hong-Bing Qiu[1], and Xueming Wei[1]

[1] Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, Guilin University of Electronic Technology, Guangxi Zhuang Autonomous Region, Guilin 541004, China
{songxiyu,qiuhb}@guet.edu.cn, 31696712@qq.com
[2] College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China

Abstract. The problem of simultaneous localization and mapping (SLAM) has been extensively studied by using a variety of specialized sensors. In this paper, we show that the SLAM could be realized using a sensor-rich smartphone. We assume that an indoor pedestrian always carries a sounding smartphone and the pedestrian moves autonomously inside a room. At every step, the loudspeaker of the smartphone produces a chirp pulse (frequency band is in the upper of human hearing area), the microphone of this smartphone registers the echoes, and the inertial sensors record the accelerometer and gyroscope readings, then the position of the moving pedestrian and the geometry map of the room are done simultaneously. However, when in a rectangular room of regular shape, reconstructing the room geometry at each sound source position is quite redundant. To avoid this redundancy and improve the sound source localization performance, we address SLAM by Matrix Analysis-based geometry estimation, and then this information is applied to the real-time positioning requirements taking the advantage of multi-source information fusion concept. Finally, we show the effectiveness of the proposed SLAM method by experiments with real measured acoustic events, the result fully demonstrates that the proposed SLAM method could be easily implemented using the smartphone carried by the pedestrian.

Keywords: Room geometry reconstruction · Smartphone-based self-positioning · Simultaneous localization and mapping

1 Introduction

The goal of simultaneous localization and mapping (SLAM) is to reconstruct both the trajectory of the moving target and the map of the environment. Studies on SLAM are presented a lot for example in wifi [1], ranging [2], visual [3] and their fusion [4, 5]. However, SLAM solution based on echoes are minimal [6]. Three problems in this field have recently received considerable attention [7]: the first one is room geometry reconstruction, the second one is self-localization of the moving target, and the third one is the simultaneous localization and mapping.

J. L. Afonso et al. (Eds.): SESC 2019, LNICST 315, pp. 246–253, 2020.
https://doi.org/10.1007/978-3-030-45694-8_19

Room geometry reconstruction is a popular and useful topic in acoustic SLAM. Many estimation works of the room geometry, or the reflective surfaces, have been proposed. However, to easily distinguish the echo orders used in room geometry reconstruction, most of these works rely on microphone arrays, which is high cost and not easy to carry with [8, 9]. Inspired by works introduced in [10–12], which show that room geometry reconstruction can be achieved by one single microphone and only one sound source, we realize that the room geometry reconstruction work could also be conducted using a sensor-rich smartphone, which could be regarded as a setup with colocated source and receiver since the distance between them (source and receiver in a smartphone) is very close.

With the reconstruction results, one can enhance source localization performance, i.e., contraction positioning range from the room level [13] to the sub-meter level [14]. In this work, we proposed a method of combining infrastructure-free indoor acoustic self-positioning with pedestrian dead reckoning (PDR) self-positioning, which verifies the rationality of PDR results through the acoustic constraint between a sound source and its image sources.

The solution for the acoustic SLAM seems to always jointly localizing target in an unknown room and estimating the room geometry from echoes. It should be necessary for a robot in the rooms of complex construction, irregular shape and severe non-line of sight situation. However, when in a rectangular room of regular shape, reconstructing the room geometry at each sound source position is quite redundant. How to avoid this redundancy and improve the sound source localization performance is what we aim to. Instead of estimating the room geometry on every single step of the moving target, we address SLAM by Matrix Analysis-based geometry estimation, and then room geometry information is applied to the real-time positioning requirements. Moreover, the multi-source information fusion concept is applied to simpler the single channel acoustic SLAM problem. Finally, we show that those acoustic SLAM problems can be solved by taking advantage of a sensor-rich smartphone.

2 Theory

We define a room to be a K-faced rectangular room, the pedestrian holding a smartphone in this room is modeled as a sound source (the loudspeaker of the user smartphone), for ease of explanation, pedestrian and sound source are hereafter equivalently used in this paper. We worked in two-dimensional (2D) space, ignoring the floor and the ceiling, given $K = 4$. Image source model proposed by Allen and Berkley [15] is adopted to model room acoustics, as shown in Fig. 1.

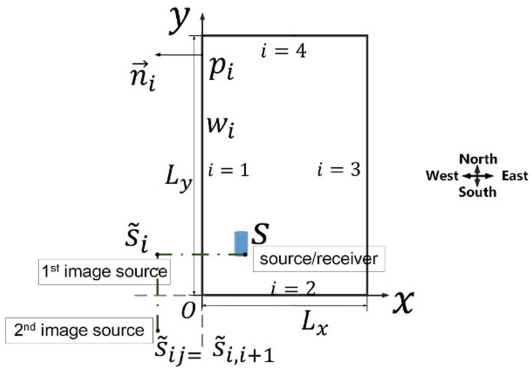

Fig. 1. Illustration of the image source model for 1st and 2nd -order images. One corner of the room denoted by O is taken as the original point. Sound source denoted by S is placed at an unknown initial position, wall index i is marked in anti-clockwise order. $p_i, i = 1, \ldots, K$, is an arbitrary point on wall w_i. n_i is the outward pointing normal vector of wall w_i. $\tilde{s}_i$ is the 1st -order image source of S corresponding to wall w_i. $\tilde{s}_{ij}$ is the 1st -order image source of $\tilde{s}_i$ corresponding to the $(i + 1)$ st wall. $[L_x, L_y]$ is the unknown room size, i.e., the required room geometry information.

2.1 Matrix Analysis-Based Room Geometry Reconstruction

When the loudspeaker S of the smartphone chirps in an indoor environment, the smartphone microphone M records both the direct path of the sound and its reflections from the walls. We set up the link between the 1st and 2nd -order images as descripted in [10] and use the same notation:

$$Q = A^2E + EA^2 - 2AN^TNA. \tag{1}$$

Where $N \overset{\text{def}}{=\!=} [n_1, \ldots, n_K]$ denotes the normal matrix, and is easy to be determined by the geometry directions, $A = diag(a_1, \ldots, a_K)$, $a_i = \|\tilde{s}_i\|$. It is well known all the early reflections should be within 0.05 s after the source stops sounding, to simplify the processing of the 1st and 2nd -order echo measurements for A and Q, we adopt the generalized cross correlation (GCC) given by the phase transform (PHAT) in the time domain [16] to separate arrivals that were close in time.

We assume that the pedestrian sounds at the beginning point, $R(\tau)_{s,r}$ represents the GCC-PHAT, the subscript s and r represents the emitting and receiving signals, respectively. To reduce the number of measurements, the first $M = 0.2f_s$ peaks of $R(\tau)_{s,r}$ are chosen to form the peak combination cells $\mathcal{C}_{1st}$ and $\mathcal{C}_{2st}$, where f_s is the sample frequency:

$$\mathcal{C}_{1st} = C_M^K, \quad \mathcal{C}_{2st} = C_M^{K(K-1)/2}. \tag{2}$$

Each cell element in $\mathcal{C}_{1st}$ is used to compose A, and each cell element in $\mathcal{C}_{2st}$ is used to compose Q with a form of $Q = \left(\|\tilde{s}_{ij}\|^2\right)$, which means Q has the 2nd -order delays as its elements. Match the A and Q using Eq. (1), and hold the A if this equation succeeds. Thus, we have A and Q as:

$$A \leftarrow diag[\left\| c * \left(\frac{\mathcal{C}_{1st}(m)}{f_s}\right)\right\|, \quad m = 1, \ldots, M. \tag{3}$$

$$Q \leftarrow \left(\|\mathcal{C}_{2st}(m)\|^2\right), \quad m = 1, \ldots, M. \tag{4}$$

Where c is the sound speed. Once the A is confirmed, the required room geometry information $\left[L_x, L_y\right]$ and the position of the initial position $S(x, y)$ of S could be obtained:

$$L_x = (\|\tilde{s}_1\| + \|\tilde{s}_3\|)/2, L_y = (\|\tilde{s}_2\| + \|\tilde{s}_4\|)/2. \tag{5}$$

$$S(x, y) = S(\|\tilde{s}_1\|/2, \|\tilde{s}_2\|/2). \tag{6}$$

2.2 Self-localization of the Moving Target

Even though the reconstruction could generate the every position estimation of the moving pedestrian, the matching process of A and Q from $\mathcal{C}_{1st}$ and $\mathcal{C}_{2st}$ is very time-consuming. Therefore, we only carry out such analysis at the very beginning of the moving pedestrian self- localization, then the results of one-time geometry estimation and the initial position of sound source will be used as the known priori for moving pedestrian self-localization. Actually, these results are the necessary condition of our prior work (about moving pedestrian self-localization) introduced in [14]. Here, we give a brief review of the moving pedestrian self-localization process. According to pedestrian dead reckoning (PDR) model, the position of S at time $t + 1$ could be inferred by the position of S at time t with the step length L_{tra} and the heading angular θ as

$$S_{t+1}(x) = S_t(x) + L_{tra} cos\,\theta. \tag{7}$$

$$S_{t+1}(y) = S_t(y) + L_{tra} sin\,\theta. \tag{8}$$

Where $S_t(x)$ and $S_t(y)$ are the $x - axis$ and $y - axis$ position of S at time t, when $t = 0$, it means S is on the initial position $S(x, y)$. Thus, with the isosceles trapezoid model (ITM) introduced in [6], the 1st -order echo measurements can be concluded as

$$\begin{cases} r_{m_{t+1,i}} = r_{m_{t,1}} \pm 2L_{tra} cos\theta, i = 1, 3, S_{t+1}(x) > S_t(x) \\ r_{m_{t+1,i}} = r_{m_{t,1}} \pm 2L_{tra} sin\theta, i = 2, 4, S_{t+1}(y) > S_t(y) \\ r_{m_{t+1,i}} = r_{m_{t,1}} \mp 2L_{tra} cos\theta, i = 1, 3, S_{t+1}(x) < S_t(x) \\ r_{m_{t+1,i}} = r_{m_{t,1}} \mp 2L_{tra} sin\theta, i = 2, 4, S_{t+1}(y) < S_t(y) \end{cases}. \tag{9}$$

Where $r_{m_{t,i}}$ is the distance between S and its 1st -order image source for the ith wall at time t. Here, $i = 1$ refers specifically to the west wall, thus $r_{m_{t,1}} = 2S_t(x)$. $r_{m_{t+1,i}}$ is the distance between S and its 1st -order image source for the ith wall at time $t + 1$.

However, the widely used PDR method only provides a relative position estimate, with its accuracy degrading over time due to accumulative error. Thus, the adaptive step length algorithm presented by Shin et al. [17] and a heading correction method similar to the one presented by Deng et al. [18] are adopted to alleviate acumulative error impact. Furthermore, our proposed Sound Pressure Level Constraint with the help of room geometry information $[L_x, L_y]$ can verify the rationality of PDR results. Finally, pedestrian position value $\bar{S}_t$ is computed by the Levenberg–Marquardt algorithm-based weighted nonlinear least squares [19]:

$$\bar{S}_t = arg\min_{S_t} \varepsilon(S_t) \tag{10}$$

$$\varepsilon(S_t) = \left(d\left(r_{m_{t,k}}\right) - d\left(r_{e_{t,k}}\right)\right)^T * D^{-1} * \left(d\left(r_{m_{t,k}}\right) - d\left(r_{e_{t,k}}\right)\right) \tag{11}$$

where $(\cdot)^T$ is the transpose operation, $(\cdot)^{-1}$ is the inverse operation, and D is the noise covariance matrix. $D = \sigma^2 I_{K-1}$, where σ^2 is the noise covariance and I is the identity matrix. We refer readers to [14] for more details about $r_{e_{t,k}}$ and $r_{m_{t,k}}$.

3 Experiments

To verify the proposed approach, a field test was carried out on the cloister on Level 5 of the Library building, Jinji Campus, GUET, Guilin, Guangxi Zhuang Autonomous Region, China. The geometry of the cloister consists of labs, offices and classrooms, as shown in Fig. 2. The four sides of the library cloister are doors, glass windows, and walls; the ceiling is mainly glass and with steel stent supports; and the floor is covered with ordinary tile. The whole cloister is a rectangular ring. The cloister size was $[L_x, L_y]_{real} = [19, 35]$. The beginning point was set at $S(x, y) = [1.5, 9]$.

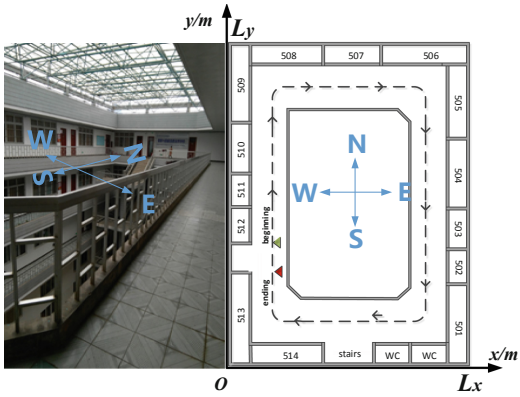

Fig. 2. Illustration of the fifth corridor of the Jinji Campus Library in GUET. GUET: Guilin University of Electronic Technology. The dashed lines are the reference walking lines. A small green triangle dot denotes the beginning point and a red one denotes the ending point. The dominant directions are denoted as E: East, S: South, W: West, N: North. (Color figure online)

To evaluate the performance of our approach when actual obstacles are present, especially in indoor situations, where pedestrians walk, all testing data collection took place during different days covering different times of the day. During the collection, students and staffs walked around normally as usual.

The data collection tool used in this experiment was a Huawei Rongyao 7 smartphone installed with a chirp application developed by our team and already authorized by China National Intellectual Property Administration, which was used to emit and store the chirp sound signal. The chirp sample frequency was set as $f_s = 44.1$ kHz, the duration was $T = 0.006$ s, the lower frequency was $f_0 = 16$ kHz, the upper frequency was $f_1 = 22$ kHz, and the emitting interval was 0.3 s. The PDR sample frequency was set as $f_{pdr} = 20$ Hz.

With the Matrix Analysis-based Room Geometry Reconstruction method, we get the self-localization trajectory of the moving target as shown in red dot line in Fig. 3. The data in Table 1 shows the differences between the true value and the estimated value, and confirms the performance of the proposed method.

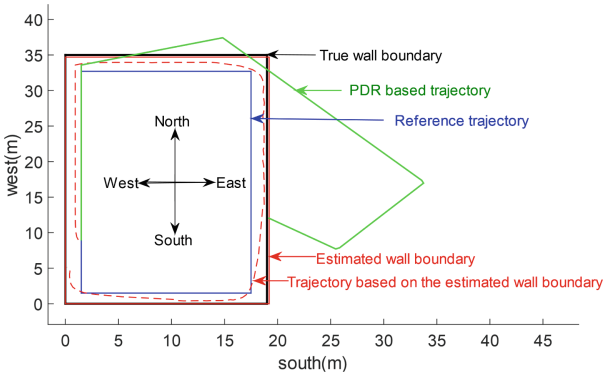

Fig. 3. Performance illustration of the proposed acoustic SLAM.

Table 1. Comparison of reconstruction results.

Parameters	True value	Estimated value	Errors
Room geometry information $[L_x, L_y]$ (m)	[19, 35]	[19.2, 34.7]	[0.2, 0.3]
Position of beginning point $S(x, y)$ (m)	[1.5, 9]	[1.2, 9]	[0.3, 0]

From the results shown in Fig. 3, we can infer the following conclusions:

① The output of the PDR trajectory (the green line) is continuous and has a similar shape to the reference trajectory (the blue line), but as time increased and the number of pedestrian steps increased, accumulative errors occurred in the accelerometer and gyroscope, resulting in positioning failure.

② Although the horizontal error of room reconstruction is 0.2 m and the vertical error is 0.3 m, the estimated wall boundary (the red line) is very close to the true one

(the black line); If you simply use the indoor map application, this effect can already meet the application needs.

③ When this estimated room geometry information $[L_x, L_y]_{estd}$ is applied in source positioning, the tracking trajectory (the red dot line) is always consistent with the trend of the reference trajectory.

④ Despite there is difference between the estimated wall boundary and the true wall boundary, the positioning trajectory (the red dot line) under the estimated wall boundary is still within the effective range of the true wall boundary. Moreover, error distribution depicted in Fig. 4 shows that the source positioning accuracy belongs to the decimeter level, which can meet the application needs of indoor SLAM robots and service robots.

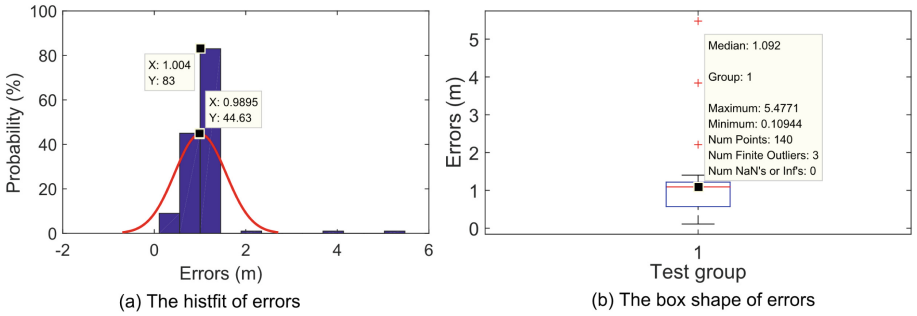

Fig. 4. Analysis of the proposed system errors. (a) is histfit of positioning errors, it shows the error probability distribution of every step of the moving pedestrian. 83% of the positioning error is around 1.0 m. Positioning errors far exceeding 1.0 m correspond to the three outliers in (b). They are related to the last two turns of the estimated trajectory (the red dot line) shown in Fig. 2. The cause of the outliers is the cumulative error of the smartphone's low cost gyroscope. However, the mean error of as low as 1.092 m verifies the effectiveness of the proposed smartphone based acoustic SLAM. (Color figure online)

4 Conclusions

In this paper, by taking advantage of the multi-source information extracted from the smartphone carried with a moving indoor pedestrian, we show the ways to solve the three significant SLAM problems and verify that the acoustic SLAM could be realized using a sensor-rich smartphone. Instead of redundantly reconstructing the room geometry at each sound source position in a rectangular room of regular shape, we consider about how to improve the sound source localization performance with the priori of room geometry. The proposed smartphone based acoustic SLAM shows us the validity of multi-source information fusion in single channel acoustic SLAM.

Acknowledgments. This work was supported by the Ministry of Education Key Laboratory of Cognitive Radio and Information Processing, the Wireless Broadband and Signal Processing Guangxi Key Laboratory.

Funding. This work was funded by the National Natural Science Foundation of China (Grant No. 61771151), by the GUET Excellent Graduate Thesis Program (Grant No. 16YJPYBS02), by the Guangxi Natural Science Foundation (Grant No. 2019GXNSFBA245103), and by the Guangxi Key Laboratory of Wireless Communication and Signal Processing Program (GXKL06180109).

References

1. Lawrence, N.D., Ferris, B., Fox, D.: WiFi-SLAM using gaussian process latent variable models. In: International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. (2007)
2. Djugash, J., Singh, S., Kantor, G., et al.: Range-only SLAM for robots operating cooperatively with sensor networks. In: IEEE International Conference on Robotics & Automation. IEEE (2006)
3. Zhou, H., Zou, D., Pei, L., et al.: StructSLAM: visual SLAM with building structure lines. IEEE Trans. Veh. Technol. **64**(4), 1364–1375 (2015)
4. Djugash, J., Singh, S.: Motion-aided network SLAM with range. Int. J. Robot. Res. **31**, 604–625 (2012)
5. Deibler, T., Thielecke, J.: Fusing odometry and sparse UWB radar measurements for indoor slam. In: Workshop on Sensor Data Fusion: Trends. IEEE (2014)
6. Krekovic, M., Dokmanic, I., Vetterli, M.: EchoSLAM: simultaneous localization and mapping with acoustic echoes. In: IEEE International Conference on Acoustics. IEEE (2016)
7. Dokmanic, I., Daudet, L., Vetterli, M.: From acoustic room reconstruction to SLAM. In: IEEE International Conference on Acoustics. IEEE (2016)
8. Dokmanić, I.: Acoustic echoes reveal room shape. Proc. Nat. Acad. Sci. U.S.A. **110**(30), 12186–12191 (2013)
9. Tervo, S., Korhonen, T.: Estimation of reflective surfaces from continuous signals. In: IEEE International Conference on Acoustics Speech & Signal Processing. IEEE (2010)
10. Dokmanić, I., Lu, Y.M., Vetterli, M.: Can one hear the shape of a room: the 2-D polygonal case. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE (2011)
11. Kreković, M., Dokmanić, I., Vetterli, M.: EchoSLAM: simultaneous localization and mapping with acoustic echoes. In: IEEE International Conference on Acoustics (2016)
12. Moore, A.H., Brookes, M., Naylor, P.A.: Room geometry estimation from a single channel acoustic impulse response. In: Signal Processing Conference. IEEE (2014)
13. Peters, N., Lei, H., Friedland, G.: Name that room: room identification using acoustic features in a recording. In: ACM International Conference on Multimedia (2012)
14. Song, X., Wang, M., Qiu, H., Luo, L.: Indoor pedestrian self-positioning based on image acoustic source impulse using a sensor-rich smartphone. Sensors **18**, 4143 (2018)
15. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(S1), 943–950 (1998)
16. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (2003)
17. Shin, S.H., Chan, G.P.: Adaptive step length estimation algorithm using optimal parameters and movement status awareness. Med. Eng. Phys. **33**, 1064–1071 (2011)
18. Deng, Z., Cao, Y., Wang, P., Wang, B.: An Improved heuristic drift elimination method for indoor pedestrian positioning. Sensors **2018**, 18 (1874)
19. Mensing, C., Plass, S.: Positioning algorithms for cellular networks using TDOA. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006