A HIERARCHICAL APPROACH TO UNSUPERVISED SHAPE CALIBRATION OF MICROPHONE ARRAY NETWORKS

Marius Hennecke[1], *Thomas Plötz*[2], *Gernot A. Fink*[1], *Jörg Schmalenströer*[3], *Reinhold Häb-Umbach*[3]

[1]Department of Computer Science, TU Dortmund University, Dortmund, Germany
[2]Robotics Research Institute, TU Dortmund University, Dortmund, Germany
[3]Department of Communications Engineering, University of Paderborn, Paderborn, Germany

ABSTRACT

Microphone arrays represent the basis for many challenging acoustic sensing tasks. The accuracy of techniques like beamforming directly depends on a precise knowledge of the relative positions of the sensors used. Unfortunately, for certain use cases manually measuring the geometry of an array is not feasible due to practical constraints.

In this paper we present an approach to unsupervised shape calibration of microphone array networks. We developed a hierarchical procedure that first performs local shape calibration based on coherence analysis and then employs SRP-PHAT in a network calibration method. Practical experiments demonstrate the effectiveness of our approach especially for highly reverberant acoustic environments.

Index Terms: microphone array, unsupervised calibration, shape estimation, acoustic localization, SRP-PHAT

1. INTRODUCTION

Microphone arrays are spatial configurations of multiple sensors that are used simultaneously for recording multi-channel acoustic data [1]. As such arrays provide richer sensing capabilities than an isolated acoustic sensor they can be applied to a variety of challenging sensing tasks. The most important ones are beamforming, source localization, and blind source separation (BSS). In beamforming one tries to focus the sensitivity of an array on a specific direction of interest while at the same time suppressing interfering sounds from others. Similar techniques are also used for localizing sound sources (cf. [2]). BSS aims at isolating the true signal of a desired source from interfering ones and thus goes beyond simple beamforming.

As there is no geometric aspect in the interpretation of BSS no knowledge about the sensor positions is necessary. However, for both beamforming and source localization a precise knowledge of the shape of the microphone array, i.e. the relative positions of the sensors w.r.t. each other, is mandatory.

Microphone arrays are frequently built in certain shapes such as linear arrays with equal or logarithmic sensor spacing, as T-arrays with three linearly arranged microphones and a fourth offset from the linear base, or as circular arrays. Given that these arrays are sufficiently small their shape can easily be measured by hand. However, as soon as either a single array becomes large (cf. [3]) or a combination of multiple arrays within a microphone array network is used, automatic array shape calibration becomes an issue.

Ideally, such an automatic calibration procedure would not require special calibration signals or even special calibration hardware. It would rather work in a completely unsupervised manner relying only on acoustic signals picked up naturally by the array. In this paper we propose a method for unsupervised microphone array shape calibration that extends techniques proposed recently for solving this challenging problem. The key idea is to exploit the intrinsic hierarchy found in larger microphone arrays during the calibration process. Consequently the method consists of a local part that works for groups of sensors that are in near vicinity and form a small local array. Given the local calibration these arrays can be used for source localization and the measurements obtained can be exploited for the calibration of the relative position of microphone array pairs in the network using a robust matching procedure of localized acoustic sources. In this paper we focus on planar arrays assuming constant heights of the acoustic sources relative to the array network plane. However, this restriction does not limit the proposed idea of hierarchical calibration of microphone arrays in general.

2. BACKGROUND

In principle, microphone array networks are special cases of general sensor networks as they exist, for example, for distributed wireless sensing. Therefore, the methods for microphone array shape calibration are inspired by those used for node localization in general sensor networks (cf. [4]).

Basically, they can be distinguished either as being supervised, i.e. using known acoustic targets for calibration, or as being unsupervised. A supervised method was proposed in [3] addressing shape calibration for the Huge Microphone-Array. It relies on a complex acoustic apparatus with five tweeters arranged in a pyramidal shape. Shape calibration is based on time-delay estimation of short chirp signals emitted from known positions. Due to the well-defined setup the technique is also used for automatic gain calibration.

The method proposed by [5] also relies on the use of known calibration signals (short chirps). Though source and microphone positions need not to be known in advance, the method requires prior knowledge regarding the number of sources and microphones. Nonlinear Maximum Likelihood estimation using time-of-flight (TOF) data is performed for shape calibration.

Recently, some methods were proposed that try to solve the shape calibration problem in an unsupervised mode. The method described in [6] makes no assumptions about source or microphone positions. However, the far-field assumption needs to be valid for all sources and only a single array can be considered. Affine structure from sound is derived using Singular Value Decomposition (SVD) and then the shape is recovered using non-linear optimization.

An energy-based method that relies on received signal strength (RSS) was proposed in [7]. It is capable of jointly estimating relative positions of speakers and microphones. However, due to the use of RSS only, the method is rather inaccurate.

In [8] an alternative approach to unsupervised shape calibration for planar array networks was presented. Evaluating time-of-arrival

978-1-4244-2710-9/09/$25.00 © 2009 IEEE

and angle-of-arrival data the positions of acoustic sources are estimated. Additionally, the positions and rotations of the arrays used are derived. Incorporating prior knowledge regarding sensor and source locations maximum a-posteriori optimization is applied. Furthermore, the Cramér-Rao bound is computed in order to give a reliability measure for the estimation.

The most promising unsupervised and furthermore least constrained method for microphone array shape calibration was proposed by McCowan and colleagues [9]. Though it is only applicable to rather small arrays, the technique does not require any calibration signal at all because it solely relies on the coherence function of a diffuse noise field. According to the authors the latter can be found in typical reverberant environments such as offices (reverberation time $t_{60} > 400\,\mathrm{ms}$) or cars.

Noise fields can be characterized by the complex coherence between two measurement positions i, j

$$\Gamma_{ij}(f) = \frac{\phi_{ij}(f)}{\sqrt{\phi_{ii}(f)\phi_{jj}(f)}} \tag{1}$$

where ϕ_{ij} and ϕ_{ii} denote cross- and auto-spectral densities, respectively. These two quantities must be estimated in practice, e. g. with a one-pole recursive time smoothing of Fourier transformed signal blocks $X_i(f)$ and $X_j(f)$

$$\phi_{ij}^{(k)}(f) = \alpha\phi_{ij}^{(k-1)}(f) + (1-\alpha)X_i\,(f)X_j^*(f) \tag{2}$$

with block index k, smoothing factor α, and $(\cdot)^*$ denoting the complex conjugate. For the sake of brevity in the following the block indices are omitted.

Assuming a diffuse noise field and omnidirectional microphones, equation 1 results in (cf. [1, chap. 4])

$$\Gamma_{ij}^{\text{diffuse}}(f) = \operatorname{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) = \frac{\sin\left(2\pi f d_{ij} c^{-1}\right)}{2\pi f d_{ij} c^{-1}} \tag{3}$$

with frequency f, microphone distance d_{ij}, and speed of sound c. The latter is assumed to be constant. Note that the noise field model can also be derived for directional microphones.

Obtaining distances d_{ij} from coherence measurements is formulated as a non-linear least-squares model fitting problem:

$$d_{ij} = \arg\min_{d} \sum_{f=0}^{f_s/2} \left| \operatorname{sinc}\left(\frac{2\pi f d}{c}\right) - \Gamma_{ij}(f) \right|^2 \tag{4}$$

where $f_s/2$ is the Nyquist frequency. This optimization problem is solved for all pairs and time indices k via the well known Levenberg-Marquardt algorithm. Before proceeding, a representative out of all distance estimates for a pair must be calculated. In contrast to [9], we found the median to be a robust estimator for this purpose.

The last step in recovering the local array shape utilizes all pairwise distance estimates in order to find the shape, which best explains the measured distances. This problem can be solved analytically using classic multidimensional scaling (CMDS) (cf. [10]).

3. UNSUPERVISED ARRAY CALIBRATION: A HIERARCHICAL APPROACH

The new approach for array shape calibration (cf. Fig. 1) exploits the hierarchical structure of a typical microphone array network setup where sensors are distributed and grouped into local arrays. Assuming N microphones, the first step is to determine the number of local arrays A and the sensor-to-array mapping, i.e. $\mathcal{M}_k \subseteq \{1, 2, \ldots, N\}$ providing the set of all channel indices belonging to array k (first half of upper box). Subsequently, the local shape of an array is determined by diffuse noise model fitting (second half of upper box). Both steps can be performed in a completely unsupervised manner, provided that the diffuse noise model assumption holds.

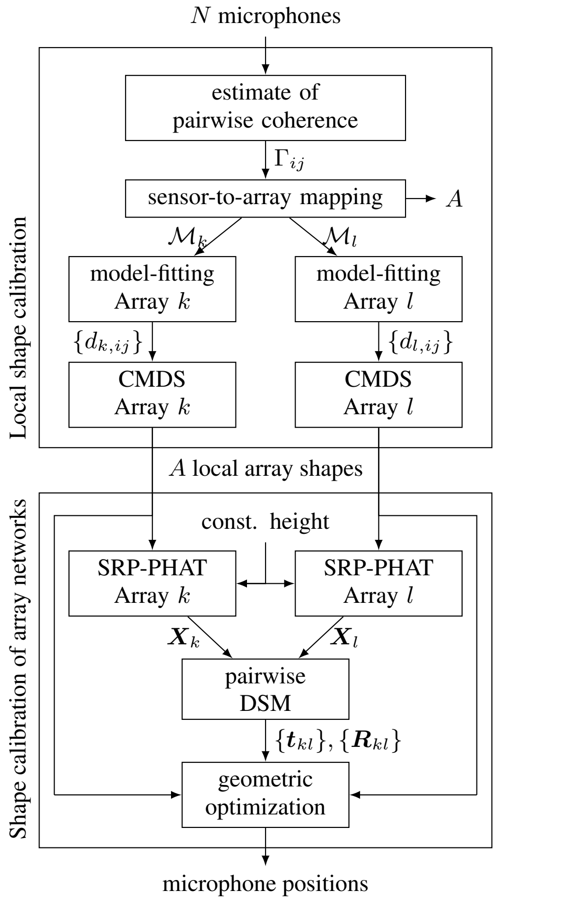

Fig. 1. Overview of the proposed hierarchical shape calibration

In order to determine the shape of the array network (lower box of Fig. 1), each local array is used for acoustic source localization. Matching of acoustic events that have been localized by two arrays provides an optimal transformation that consists of a translation $\boldsymbol{t}_{kl}$ and the relative rotation $\boldsymbol{R}_{kl}$. Incorporating all pairwise array translations and rotations a geometric optimization procedure gives the array positions $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A)$ and rotations $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A)$ w.r.t. an arbitrary reference $\boldsymbol{p}_1 \equiv 0$ and $\alpha_1 = 0$. Combining the latter with the local array shapes finally reveals the positions of all microphones. In the following, the overall calibration procedure is explained in detail.

3.1. Local Shape Calibration

The local shape calibration stage is divided into two distinct steps. First, the number of local arrays and the corresponding microphone-to-array mapping is determined. Afterwards, the local shape of each array that was found in step one is estimated according to [9]. For these steps, no user interaction is needed as long as the diffuse noise model assumption holds. In order to achieve a robust local shape calibration, only a small portion of noise is required.

Based on initial distance estimates $d_{ij,0}$, which are also needed for solving eq. 4 iteratively, an undirected graph is built. In this case vertices represent microphones and edges reflect local adjacencies. The first zero crossing $f_{ij,0}$ of a coherence measurement Γ_{ij} (eq. 1) determines initial distance estimates $d_{ij,0} = c(2f_{ij,0})^{-1}$. An edge is inserted between channel i and j if $d_{ij,0} < d_{\max}$ with $d_{\max}$ being the threshold for the expected maximum distance. The connected components of this graph, found by a depth-first-search, provide the number of arrays plus the microphone-to-array mapping.

3.2. Shape Calibration of Array Networks

The calibrated local arrays are used for revealing the geometric shape of the array network by using them for acoustic source localization. Namely, a steered response power algorithm in combination with a phase transformation (SRP-PHAT) ([1, chap. 8]) is applied. It provides azimuth, elevation, and – to some extent – range information regarding an acoustic source relative to the center of the particular array. However, the latter can only be achieved if four or more sensors per array are available. If the source lies in the far-field of the array, range information is indeterminable. Note that SRP-PHAT does not constrain the kind of localizable signals. Advantageously, no special calibration signal is needed. For example unconstrained speech or ordinary hand claps are sufficient. In order to capture representative spatial variety, the acoustic source is required to be moving during the calibration procedure.

To circumvent the range ambiguity of the localization results, the acoustic source is constrained to lie on a plane parallel to the array network. The distance of the source from the network plane needs to be known. Note, that these constraints can often be easily fulfilled for typical use-cases, e.g., by estimating the aforementioned distance by exploiting a few range measurements of a near-field source. This allows to infer the Cartesian coordinates on the source plane from azimuth and elevation, respectively.

In order to estimate the distance between two arrays and their relative orientation, all sources localized by both arrays are analyzed. Hypotheses for the distance and the relative orientation of an examined pair of arrays are derived by error minimzation over localized acoustic events. This data set matching (DSM) is performed for every possible pair of arrays. Given n measurements, the matrix $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}, \boldsymbol{x}_{k,2}, \ldots, \boldsymbol{x}_{k,n})$ contains the coordinates of source positions estimated using array k. The optimal translation $\boldsymbol{t}_{kl}$ and rotation matrix $\boldsymbol{R}_{kl}$ for matching $\boldsymbol{X}_k$ and $\boldsymbol{X}_l$ is found by performing an SVD of their dispersion matrix (cf. [3])

$$\boldsymbol{D}_{kl} = \frac{1}{n}\boldsymbol{X}_k\boldsymbol{C}_n\boldsymbol{X}_l^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}, \qquad \boldsymbol{C}_n = \mathbf{I} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^{\mathrm{T}} \quad (5)$$

where $\boldsymbol{C}_n$ is the centering matrix of size n. The rotation and translation are given by $\boldsymbol{R}_{kl} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{t}_{kl} = \frac{1}{n}(\boldsymbol{X}_k - \boldsymbol{R}_{kl}\boldsymbol{X}_l)\,\mathbf{1}$.

Depending on SNR conditions and putative reverberations, acoustic source localization can generally contain a substantial number of erroneous detection results. Addressing increased robustness of the DSM procedure we apply the iterative, non-deterministic random sample consensus method [11]. The proposed procedure for matching two data sets $\boldsymbol{X}_k$ and $\boldsymbol{X}_l$ can then be summarized as follows:

1. Randomly choose the minimal number of points for calculating the model parameters. In our case two coordinates from each dataset that are assumed to be true positives, are sufficient. Initialize consensus set $\mathcal{C} = \emptyset$.
2. Calculate model parameters $\boldsymbol{t}_{kl}$ and $\boldsymbol{R}_{kl}$ through a DSM with the data points chosen in 1.
3. Determine the subset $\mathcal{X}_k^c$ of all data points from $\boldsymbol{X}_k$ which are in the vicinity of the transformed data set $\boldsymbol{X}_l'$. An Euclidean distance metric $\mathrm{d}(\cdot,\cdot)$ with a fixed distance threshold d_ε is used, i.e. $\mathcal{X}_k^c = \{\boldsymbol{x}_{k,i} |\, \mathrm{d}(\boldsymbol{x}_{k,i}, \boldsymbol{x}'_{l,i}) \leq d_\varepsilon, \forall i\}$.
4. If the subset $\mathcal{X}_k^c$ is larger than the consensus set $\mathcal{C}$, a DSM is performed using the whole subset. It becomes the new consensus set if its error is less than the error of the current one.
5. Repeat until $\mathcal{C}$ is large enough, i.e. $\mathrm{card}(\mathcal{X}_k^c) \geq \beta n$, where for example $\beta = 0.8$, or until the (predefined) maximum number of iterations $N_{\max}$ is reached.

Using all pairwise translations and rotation matrices the array positions $\boldsymbol{P}$ and their orientations $\boldsymbol{\alpha}$ can be determined by the following geometric minimization procedure

$$\boldsymbol{P} = \arg\min_{\boldsymbol{P}} \sum_{k=1}^{A} \sum_{l=k+1}^{A} \left\|\boldsymbol{t}_{kl} - (\boldsymbol{p}_l - \boldsymbol{p}_k)\right\|^2 \quad (6)$$

with $\boldsymbol{\alpha}$ defined analogously, whereas $\alpha_{kl} = \arccos((\boldsymbol{R}_{kl})_{1,1})$. Finally, all microphone positions are derived using local array shapes.

4. EXPERIMENTAL EVALUATION

In order to evaluate the effectiveness of the new approach we performed practical experiments within a challenging acoustic environment, namely a smart house. For a validation of the results we furthermore conducted local shape calibration experiments in a designated audio lab. We first describe the methodology, followed by a discussion of the achieved results.

4.1. Setup and Data Recording

The FINCA: The conference room of our smart house, the FINCA (http://finca.irf.de), is equipped with 16 Behringer ECM8000 omnidirectional microphones mounted in a coffered ceiling using fixing plates. Thus, almost arbitrary array layouts can be realized. All microphones are attached to two eight-port amplifiers (SM PRO Audio PR8E), which are connected to two M-Audio Delta 1010 sound cards. The FINCA has an approx. rectangular shape (3.7 m × 6.8 m).

The room exhibits highly reverberant acoustic conditions with $t_{60} \approx 500\,\mathrm{ms}$. Different recordings were made for two different array setups. For the local shape calibration step one minute of noise – produced mainly by computers above the ceiling – was recorded for each setup. In order to have a reproducible and spatially known moving acoustic source for the array network calibration step a loudspeaker (Behringer TRUTH B2030A) mounted on a mobile robot (Scitos G5) was chosen. Moving along an arbitrary path, white noise and speech samples were replayed and captured. Furthermore, recordings of a talking person wandering around within the FINCA in an unconstrained manner correspond to an additional data set. All recordings were made with a sampling rate of $f_s = 48\,\mathrm{kHz}$ and have an approximate length of one minute.

In order to determine the maximum distance for which the diffuse noise model assumption holds, the evaluated setups include a linear array with increasing inter-microphone distances (0.15 m up to 0.9 m). The different array network setups exhibit the following geometries: (F1) Two regular circular arrays are used, which are 2.5 m apart with eight microphones each and a diameter of 20 cm. (F2) One T-shaped array is mounted per corner consisting of four 14.1 cm spaced microphones each.

Audio-Lab Paderborn: The recordings are done in a lab of size 3.5 m× 7.3 m with a room reverberation time of approximately $t_{60} =$

Ground truth distance in mm	150	350	625	875
Distance error in mm(%)	1 (0.6)	4 (1.1)	37 (5.9)	129 (14.7)

Table 1. Error of estimates with increasing microphone distance.

Setup	F1		F2			
Mean error in mm	9	10	15	17	40	9

Table 2. Mean error per setup (F1: two circular arrays with a diameter of 20 cm, F2: four T-shaped arrays with 14.1 cm inter-microphone distances) and per array of Euclidean distances between ground truth and estimated local shapes.

250 ms, representing typical audio-video communication conditions. We employed one T-shaped microphone array consisting of two linear arrays with 5 cm inter-microphone spacing and two linear arrays, consisting of two microphones at a distance of 5 cm. The arrays are mounted on the walls and an M-Audio Delta 1010 soundcard running at a sampling rate of 48 kHz is used for recordings. We decided to generate the diffuse noise by computer fans covered with sound absorbing foam parts, since experiments revealed that the ambient noise in the lab is too low for calibration purposes.

4.2. Evaluation Results

Local Shape Calibration: Preliminary experiments showed reasonable results with 10 s of noise, a block size of 43 ms and a smoothing factor $\alpha = 0.95$, which are used for the evaluation reported here. Table 1 shows the median of distance errors – over all block indices k – w.r.t. ground truth. It can be seen that the diffuse noise model assumption holds up to approx. 60 cm microphone distance leading to an error of 3.7 cm.

Results for the local shape calibration step are shown in table 2. Due to the translation and rotation ambiguity of the CMDS, a DSM is performed before determining the mean distance error. For F2 the third array leads to a higher error. Re-analyzing the setup we identified the position of the array, which is mounted very closely to a corner of the room, causing the accuracy drop. Apparently, the diffusivity assumption here does not hold.

Due to a low t_{60} time in the second setup only distance estimates for the linear arrays were determined. The errors are 8 mm(16 %), 9mm(18 %), 5 mm(10 %) and 14mm(28 %). The increase for the latter is again correlated with the array's mounting position in a corner.

Shape Calibration of Array Networks: The results reported in table 3 were obtained as the median of 100 Monte Carlo runs of the RANSAC ($N_{\max}$ = 200, β = 0.8, d_ε = 0.4 m) DSM procedure followed by the aforementioned geometric optimization. The sources' heights were assumed being constant and predetermined by range measurements of a near-field source using 20 s of speech.

The mean distance error is proportional to the localizing capability of the source signal. Hence, in each setup the best results can be achieved by using white noise. Even with speech as source signal the obtained calibration is quite accurate – relative to the respective array network dimensions.

FINCA setup	F1			F2		
Signal type	w	r	s	w	r	s
Mean error in mm	108	202	267	93	139	110

Table 3. Calibration results for different FINCA setups and three different signals: (w)hite noise, (r)eplayed speech and (s)peech.

5. CONCLUSION

Unsupervised shape calibration of microphone array networks is a demanding task, which is extremely relevant for numerous practical applications. In this paper we presented a new hierarchical approach that combines local shape calibration based on coherence analysis, and network calibration using SRP-PHAT. The focus of our work was on applications in highly reverberant acoustic environments.

By means of the proposed procedure we were able to successfully perform unsupervised array network calibration in a challenging setting. Thereby, the shapes of the particular microphone arrays together with their relative arrangements have been unveiled very precisely. The promising results were – to some extent – validated in an additional setting. However, reasoned by substantially differing acoustic conditions the calibration accuracy decreased for the latter.

Future work will consider methods for eliminating planar array constraints as well as improved independence of room acoustics.

6. REFERENCES

[1] M. S. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 8, pp. 157–180. Springer, 2001.

[3] J.M. Sachar, H.F. Silverman, and W.R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 42–52, 2005.

[4] N. Patwari, J. N. Ash, R. L. Moses, S. Kyperountas, A. O. Hero III, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, July 2005.

[5] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 70–83, 2005.

[6] S. Thrun, "Affine structure from sound," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1353, 2006.

[7] M. Chen, Z. Liu, L. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[8] R. L. Moses and R. Patterson, "Self-calibration of sensor networks," in *SPIE*, M. Carapezza, E. Ed., Aug. 2002, vol. 4743 of *SPIE*, pp. 108–119.

[9] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 666–670, March 2008.

[10] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, Number 88 in Monographs on statistics and applied probability. CRC Press, 2001.

[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.