An Accurate and Efficient Voting Scheme for a Maximally All-Inlier 3D Correspondence Set

Hamdi Sahloul, Shouhei Shirafuji, and Jun Ota

Abstract—We present a highly accurate and efficient, yet simple, two-stage voting scheme for distinguishing inlier 3D correspondences by densely assessing and ranking their local and global geometric consistencies. The strength of the proposed method stems from both the novel idea of post-validated voting set, as well as single-point superimposition transforms, which are computationally cheap and avoid rotational ambiguities. Using a well-known dataset consisting of various 3D models and numerous scenes that include different occlusion rates, the proposed scheme is evaluated against state-of-the-art 3D voting schemes, in terms of both the correspondence PR (precision–recall) AUC (area under curve), and the execution time. A total of 374 experiments were conducted for each method, which involved a combination of four models, 50 scenes, and two down-samplings. The proposed scheme outperforms the state-of-the-art 3D voting schemes in terms of both accuracy and speed. Quantitatively, the proposed scheme scores $97.0\% \pm 12.9\%$ on the PR AUC metric, averaged over all of the experiments, while the two state-of-the-art schemes score $74.2\% \pm 22.2\%$ and $78.3\% \pm 26.4\%$. Furthermore, the proposed scheme requires only $24.1\% \pm 6.0\%$ of the time consumed by the fastest state-of-the-art scheme. The proposed voting scheme also demonstrates high robustness against occlusions and scarce inliers.

Index Terms—Outlier rejection, post-validated voting scheme, all-inlier correspondence set, local rigidity constraint, single-point superimposition transforms

1 Introduction

Finding maximal plausible correspondences within a contaminated set is a generic problem at the core of numerous state-of-the-art computer vision techniques, such as localization [1], [2], [3], motion estimation [4], pose estimation [5], [6], [7], [8], [9], recognition [10], [11], [12], reconstruction [13], [14], [15], registration [16], [17], [18], [19], and tracking [20], [21], [22]. A crucial step of these applications is to recover at least one geometric hypothesis that receives sufficient support from the initial or putative correspondence set [23], [24].

Although correspondence estimation has undergone large advancements in the last few decades [25], contaminated correspondence sets are still unavoidable. Approaches for correspondence estimation include spectral embeddings [26], [27], [28] and feature matching, where features are either local-image features [29], [30], [31], [32], RGB-D features [33], [34], [35], 3D features [36], [37], [38], or combinations of the above. Nevertheless, uncertainties may arise due to the locality of measurements, similarities in geometry and texture, or ambiguities originating from clutter and occlusions. All of these issues increase the correspondence outlier rate, causing a combinatorial search explosion while seeking a hypothesis which is maximally supported by consistent correspondence inliers.

Excluding spurious matches from high-outlier-rate correspondences is still an open problem featuring high-dimensionality issues [39]; yet, little research has been done in this area. One of the early popular proposals was random sample consensus (RANSAC) [40], which is based on random sampling, as the name implies, and thus, suffers from repeatability issues. Notably, Optimal RANSAC [41] is an extension of the random sampling algorithm, which addresses the repeatability issue to a high extent. Notwithstanding, sampling methods in general are still sensitive to high outlier rates and require a substantially large number of samples for robust estimation, which is time-consuming.

Recently, more robust schemes have been proposed, including guided sampling [42], [43], mathematical optimization [39], [44], [45], and convex/graph matching [25], [46], [47], [48], [49], [50]. However, all of these approaches are either complex and slow, or sacrifice the correspondence quantity for the sake of quality [51]. Although a few applications are aimed at either ultra speed or extreme accuracy, most applications seek a balance between these two aspects. Simultaneously recovering high quality and abundant correspondences is indispensable in obtaining a plausible hypothesis set for proper model fitting [43]. Accordingly, voting-based schemes have gained momentum recently, due to their balanced performance.

More to the point, several studies have incorporated correspondence consistency voting [23], [24], [51], [52], [53] to increase the correspondence inlier rate, either by truncating inconsistent correspondences or utilizing the voting ranks to weight the model parameters. Some voting schemes [23],

- *H. Sahloul is with Department of Precision Engineering, Graduate School of Engineering, University of Tokyo7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. E-mail: sahloul@race.t.u-tokyo.ac.jp.*
- *S. Shirafuji and J. Ota are with the Research into Artifacts, Center for Engineering (RACE), School of Engineering, University of Tokyo7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.*
  E-mail: {shirafuji, ota}@race.t.u-tokyo.ac.jp.

Manuscript received 7 Apr. 2019; revised 8 Dec. 2019; accepted 31 Dec. 2019. Date of publication 10 Jan. 2020; date of current version 3 June 2021.
(Corresponding author: Hamdi Sahloul.)
Recommended for acceptance by V. Lepetit.
Digital Object Identifier no. 10.1109/TPAMI.2020.2963980

[24] have adapted the nearest-neighbor similarity ratio (NNSR) [30], which was one of the earliest techniques to detect spurious correspondences formed by indistinct features. However, the NNSR was originally proposed for high-dimensional intensity-based local-image features, and thus its quality is questionable for low-dimensional geometric features. Similarly, the local rigidity constraint (LRC) has been employed in some voting schemes [23], [54] to ensure compatible euclidean distances in the surrounding neighborhoods of the two corresponding points. Per contra, a rigidity constraint is not sufficient to ensure the rotational compatibility of neighboring correspondences. It is worth noting, however, that neighborhood measurements are unavoidable in voting schemes, and it is believed that Ref. [24] made an error in claiming that k-nearest neighbor (k-NN) queries are avoided in their method, while in reality they employed a local reference frame (LRF) estimation method [55], which internally and unavoidably relied on k-NN queries. Although LRFs [55], [56], [57] have been recently utilized in voting schemes [23], [24] to assess the global consistency of correspondences, LRFs suffer from noise and eigenvector sign ambiguities. Even after resolving these ambiguities by following certain conventions [58], using LRFs for global verification is still debatable, as they were originally intended for local feature description. To the best of our knowledge, both the accuracy and robustness of voting schemes remain challenging problems at present.

Accordingly, the problem of rejecting outliers to find a maximal plausible correspondence set still persists, which has been set as the objective of this manuscript. Similar to the state-of-the-art 3D correspondence voting schemes [23], [24], we follow a two-stage scheme concept: in the first stage, a voting set is elected based on the top coarsely-estimated likelihoods; in the later stage, the correspondences are validated against the voting set and their fine-tuned likelihoods are estimated accordingly. Nonetheless, it is challenging to come up with criteria for each stage that maximize the accuracy without affecting the efficiency. Our approach for the first voting stage involves utilizing the LRC, similar to Ref. [23], to obtain coarse inlier ranking scores. However, unlike Ref. [23], the resulting LRC scores are not subject to a hard-threshold and are not combined with external scores. Instead, they are utilized to guide the global scoring stage, rather than constraining it and limiting the overall performance.

Importantly, the strength of our proposed method stems from the second voting stage, in which contamination is minimized in both the voting set and its elementwise hypotheses. While previous methods [23], [24] utilized the putative correspondences in composing the global-stage hypotheses (from both the correspondence source and destination sides, in the case of Ref. [23], and from the source correspondence side in the case of Ref. [24]), we opted to compose our hypotheses solely from the voting set, to minimize outlier effects. Furthermore, the voting set is post-validated in the second stage of our scheme, which was not carried out in previous methods [23], [24]. Our proposed method also differs from both previous methods [23], [24], as it does not rely on their underlying incompetent criteria of NNSR or LRFs, which lacks local-rigidity checks or suffers from rotational ambiguity, respectively. Instead, our proposed single-point superimposition transforms (1PSTs) are rotational ambiguity-free and computationally cheap, compared to the noisy and usually ambiguous LRFs.

Briefly, we propose a voting scheme that is:

- highly accurate and extremely efficient,
- deterministic and rigorously repeatable, and
- simple to implement.

The remainder of this manuscript is organized as follows. Section 2 describes the proposed scheme, while Section 3 explains the experimental setup, the dataset, and the performance metrics that were utilized. The results are discussed in Section 4, and Section 5 presents the conclusions and future work.

2 Methodology

Let $\mathcal{P}, \acute{\mathcal{P}} \subset \mathbb{R}^3$ be the model and scene point clouds (multiple rigid objects), and let $\mathcal{C} = \{(\boldsymbol{p}, \acute{\boldsymbol{p}}) : \boldsymbol{p} \in \mathcal{P}, \acute{\boldsymbol{p}} \in \acute{\mathcal{P}}\} \subset \mathcal{P} \times \acute{\mathcal{P}}$ be the initially-provided correspondence set. The goal is to compute a likelihood set, $\mathcal{S} \subset [0, 1]$, where each individual element $s(\boldsymbol{c}_i) \in \mathcal{S}$ represents the likelihood that a correspondence $\boldsymbol{c}_i \in \mathcal{C}$ is valid (i.e., an inlier).

In the first voting phase, a voting set $\mathcal{C}^l \subset \mathcal{C}$ is elected, based on the top-ranked elements of a local coarse ranking set, $\mathcal{L}$, that is estimated through verification of the LRC (local rigidity constraint, Section 2.1). In the second voting phase, the voting set $\mathcal{C}^l$ and the putative correspondence set $\mathcal{C}$ are assessed against each other. The targeted likelihood set $\mathcal{S}$ is estimated by calculating the elementwise covariance of the putative set with single-point superimposition transforms, which are derived from the voting set (Section 2.2). A schematic representation of our proposed method is shown in Fig. 1.

2.1 First Voting Stage: Voting Set Election

This section is aimed at electing a voting set $\mathcal{C}^l \subset \mathcal{C}$. Its cardinality, $|\mathcal{C}^l| = k_l$, is the first free parameter of our proposed method. The voting set elected in this stage is utilized to assess the putative correspondences $\mathcal{C}$ in the next voting stage. To elect the voting set, the local rigidity constraint is utilized, which asserts the mutual compatibility of two correspondences $\boldsymbol{c}_i \overset{\text{def}}{=} (\boldsymbol{p}_i, \acute{\boldsymbol{p}}_i) \in \mathcal{C}$ and $\boldsymbol{c}_j \overset{\text{def}}{=} (\boldsymbol{p}_j, \acute{\boldsymbol{p}}_j) \in \mathcal{C}$. These two correspondences are said to have a high compatibility likelihood, $\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j) \approx 1$, when their corresponding domain and co-domain euclidean distances are approximately equal, $\|\acute{\boldsymbol{p}}_j - \acute{\boldsymbol{p}}_i\|_2 \approx \|\boldsymbol{p}_j - \boldsymbol{p}_i\|_2$, thus complying with the rigidity constraint. While several studies have formulated the likelihood of the pairwise rigidity constraint as a minimum between two ratios (e.g., see Ref. [23]), we opt to formulate it as a Gaussian function, in order to capture the physical aspects of the acquisition sensor:

$$\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j) = \exp\left(-\frac{(\|\acute{\boldsymbol{p}}_j - \acute{\boldsymbol{p}}_i\|_2 - \|\boldsymbol{p}_j - \boldsymbol{p}_i\|_2)^2}{2{\sigma_a}^2}\right), \tag{1}$$

where σ_a is the standard deviation of the acquisition accuracy, which constitutes the second free parameter of the proposed scheme.

Due to the pairwise nature of this constraint, one needs to pair each correspondence with several others to accumulate

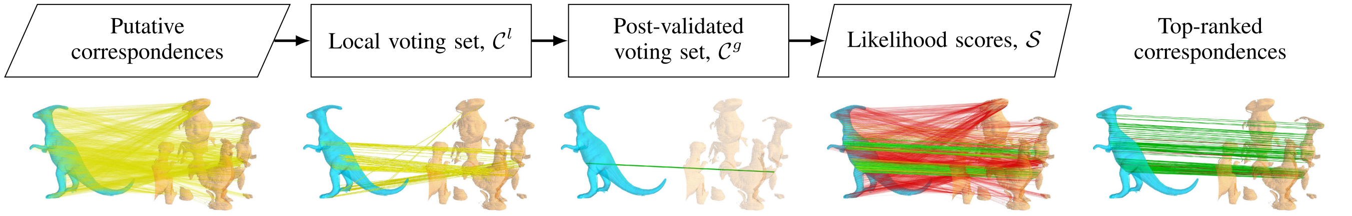
Putative correspondences
Local voting set, $\mathcal{C}^l$
Post-validated voting set, $\mathcal{C}^g$
Likelihood scores, $\mathcal{S}$
Top-ranked correspondences

Fig. 1. An overview of the proposed voting scheme. Our proposed scheme takes a set of putative correspondences (the yellow-colored dense lines between the teal-colored source model and the tan-colored destination scene) as input and processes them in two stages. In the first stage, Section 2.1, a local voting set, $\mathcal{C}^l$ (shown as yellow-colored sparse lines), is elected based on the local rigidity constraint. Each element in the voting set represents a global hypothesis, and their elementwise support by the putative correspondence is utilized to post-validate the voting set at the second voting stage, Section 2.2. The top supported hypotheses forms the global voting set, $\mathcal{C}^g$. In the case shown above, only a single element is selected (shown as the single green-colored line). Using the post-validated voting set, the likelihood scores, $\mathcal{S}$, for the putative correspondences being inliers are computed, according to their covariance with the post-validated voting set, where green-colored lines correspond to the inlier ones, and magenta-colored lines correspond to the outliers. It is up to the high-level application to decide whether to truncate the scores, based on some threshold, or to utilize them all in a weighted model.

a decent disjoint probability estimation. Nonetheless, excessive verification might lead to a combinatorial explosion. Without loss of generality, we exploit the fact that inlier correspondences tend to appear in groups [23], [59], in order to improve the probability estimation while avoiding computational issues. Although this assertion might reduce the inlier recall, it is compensated for in the global voting stage (Section 2.2).

Accordingly, for every matching keypoint of the model $\mathcal{P}_c = \{\boldsymbol{p} : (\boldsymbol{p}, \acute{\boldsymbol{p}}) \in \mathcal{C}\}$ the k-NN originating from the same keypoints set, $\mathcal{N}_k(\boldsymbol{p}) \subset \mathcal{P}_c$, is utilized, where k defines the size of the neighborhood (i.e., $|\mathcal{N}_k|$). If applicable, reusing such neighborhood information from the feature computation phase will save some computational power; otherwise, an approximate k-NN method [60] can be utilized for fast estimation. Upon the availability of neighborhood information, the local coarse ranking set $\mathcal{L} = \{\Lambda_l(\boldsymbol{c}_i) : \boldsymbol{c}_i \in \mathcal{C}\} \subset \mathbb{R}$ is estimated as the summation of the neighborhood pairwise likelihoods:

$$\Lambda_l(\boldsymbol{c}_i) = \sum_{\boldsymbol{p}_j \in \mathcal{N}_k(\boldsymbol{p}_i), (\boldsymbol{p}_j, \acute{\boldsymbol{p}}_j) \in \mathcal{C}} \lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j). \tag{2}$$

Notably, we arrived at the same conclusion as Ref. [23]; the most efficient spatial neighborhood size $|\mathcal{N}_k| = k$ is actually given by the cardinality of the voting set, $|\mathcal{C}^l| = k_l$. That is, the accuracy of the final likelihood scores decreases linearly with neighborhood sizes smaller than the cardinality, and begins to saturate for larger ones.

Subsequently, the putative set $\mathcal{C}$ is sorted in descending order, according to its local coarse ranking scores $\mathcal{L}$, and is denoted as $\mathcal{C}^{\mathcal{L}}$, from which the top k_l-elements constitute the voting set, $\mathcal{C}^l$

$$\begin{aligned} \mathcal{C}^{\mathcal{L}} &= \{\mathcal{C}_i : i \in \arg\,\mathrm{sort}(-\mathcal{L})\}, \\ \mathcal{C}^l &= \{\mathcal{C}_i^{\mathcal{L}}\}_{i=1}^{k_l}. \end{aligned} \tag{3}$$

Thus, the first voting stage is concluded by the election of the voting set $\mathcal{C}^l$, which is utilized in the next section for the assessment of the putative correspondences $\mathcal{C}$.

2.2 Second Voting Stage: Post-Validation and Scoring

In the first voting stage, we elected a voting set $\mathcal{C}^l$ from the top-ranked elements of a local coarse ranking set $\mathcal{L}$ based on local neighborhood measurements and support. However, inliers surrounded by contaminated neighborhoods would not have received enough support at that stage. In this stage, we address this particular issue by assessing both the voting set $\mathcal{C}^l$ and the putative correspondence set $\mathcal{C}$ against each other, and measure the covariance to global single-point superimposition transforms derived from the voting set. Unlike previous work [23], [24] which utilize LRFs, 1PSTs are computationally cheap and, more importantly, rotational ambiguity-free. Moreover, we construct the hypotheses solely from the voting set and evaluate both the putative and voting sets against each other, while previous work [23], [24] involved a contaminated putative set in forming their hypotheses and only evaluated the putative set without post-validating the voting set.

Initially, each voting element $\boldsymbol{c}_i \in \mathcal{C}^l$ is assigned a rotational ambiguity-free and computationally cheap single-point superimposition transform, $\mathbf{T}(\boldsymbol{c}_i) = [\,\mathbf{R}(\boldsymbol{c}_i) \quad \boldsymbol{t}(\boldsymbol{c}_i)\,] \in \mathbb{SE}_3$, as a candidate hypothesis, where $\mathbf{R}(\boldsymbol{c}_i) \in \mathbb{SO}_3$ is a rotation matrix and $\boldsymbol{t}(\boldsymbol{c}_i) \in \mathbb{R}^3$ is a translation vector. While the translation vector can be given by $\boldsymbol{t}(\boldsymbol{c}_i) = \acute{\boldsymbol{p}}_i - \mathbf{R}(\boldsymbol{c}_i)\boldsymbol{p}_i$, the rotation matrix is a little bit more involved. For the purpose of computing $\mathbf{R}(\boldsymbol{c}_i)$, we utilize a method for superimposition transform estimation [61] and borrow some concepts from another method [38] originally proposed for LRF estimation. Briefly, a covariance matrix between the corresponding model and scene points is computed and then decomposed to estimate a superimposition transform, as per Ref. [61]. Furthermore, the covariance weights and centroids are computed in a similar manner to Ref. [38]. The covariance matrix $\mathbf{C}(\boldsymbol{c}_i) \in \mathbb{R}^{3\times3}$ is given by:

$$\mathbf{C}(\boldsymbol{c}_i) = \sum_{\boldsymbol{p}_j \in \mathcal{N}_k(\boldsymbol{p}_i), \boldsymbol{c}_j \in \mathcal{C}} \omega(\boldsymbol{c}_i, \boldsymbol{c}_j)\,(\acute{\boldsymbol{p}}_j - \acute{\boldsymbol{p}}_i)\,(\boldsymbol{p}_j - \boldsymbol{p}_i)^\intercal, \tag{4}$$

where $\omega(\boldsymbol{c}_i, \boldsymbol{c}_j) = \exp(-\|\boldsymbol{p}_j - \boldsymbol{p}_i\|_2^2 / 2\sigma_r^2)\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j)^p \in \mathbb{R}$ is a weighting term, similar to Ref. [38], to provide robustness against both clutter and occlusions. Unlike Ref. [38], the weighting term $\omega(\boldsymbol{c}_i, \boldsymbol{c}_j)$ is a bilateral kernel, which provides robustness against within-neighborhood inconsistencies, which partially depends on the local rigidity pairwise consistency $\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j)$ defined in Eq. (1). Additionally, our weighting-term formulation includes σ_r as the standard deviation of the neighborhood radius, as well as the power term $p \in \mathbb{R}$ to adjust the standard accuracy deviation σ_a without recomputing $\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j)$. Furthermore, as per Ref. [38], the covariance centroids are approximated by the neighborhood center points, in order to speed up the computation. However, it is worth noting that the covariance matrix in our formulation is not normalized, as this has no effect on the underlying rotation. Additionally, the difference vectors, their euclidean distances, and the neighborhoods $\mathcal{N}_k(\boldsymbol{p})$, previously computed in Eq. (2), are reutilized in Eq. (4), which contributes to the computational efficiency of our approach.

Moreover, to further speed the computations up, the power term $p = 1/0.16^2 \approx 39$ is empirically set to adjust the standard accuracy deviation σ_a to 16 percent of its original value without recomputing $\lambda_l(\boldsymbol{c}_i, \boldsymbol{c}_j)$. Moreover, it is sufficient to only consider the first $k_r = 18$ neighbors of the self-including neighborhoods, k-NN, while setting the standard deviation radius to just one half of the point cloud resolution, $\sigma_r = \frac{1}{2}\,\text{voxel}$. These parameters are believed suitable for datasets beyond our experimentations.

According to Ref. [61], the rotation matrix $\mathbf{R}(\boldsymbol{c}_i)$ is then formed by multiplying the left and right singular value decomposition (SVD) matrices, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{3\times 3}$, of the covariance matrix $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal = \mathbf{C}(\boldsymbol{c}_i)$

$$\mathbf{R}(\boldsymbol{c}_i) = \mathbf{U}\begin{bmatrix} 1 & & \\ & 1 & \\ & & \det(\mathbf{U}\mathbf{V}^\intercal) \end{bmatrix}\mathbf{V}^\intercal, \tag{5}$$

where the determinant $\det(\cdot)$ in the middle diagonal matrix is utilized to negate reflection cases. Notably, state-of-the-art methods [23], [24] have employed LRF algorithms for hypothesis estimation, which only depend on $\boldsymbol{p}$ and $\mathbf{V}$, while our hypothesis depends on $\acute{\boldsymbol{p}}$ and $\mathbf{U}$, as well. Although they eventually compose the final hypotheses for some $\mathcal{P} \times \acute{\mathcal{P}}$ LRFs set, their hypotheses are contaminated by the inclusion of non-voting set LRFs. Moreover, the LRFs suffer from sign ambiguities of the three eigenvectors $[\boldsymbol{v}_1 \;\; \boldsymbol{v}_3 \;\; \boldsymbol{v}_2] = \mathbf{V}$ which requires $\mathcal{O}(n^2)$ time complexity to resolve [58]. Even with such resolution, there remains no guarantee of correctness of such convention. See Fig. 2 for a graphical demonstration.

In spite of the estimated single-point superimposition transforms $\{\mathbf{T}(\boldsymbol{c}) : \boldsymbol{c} \in \mathcal{C}^l\}$ it is essential to exclude invalid transformations of the voting sets before imposing their dubious assessment on the putative correspondence set, $\mathcal{C}$. To address this chicken-and-egg problem, we consider the global compatibility likelihood of both the putative and voting sets, in an almost identical manner to Eqs. (1), (2), and (3). Consequently, the global pairwise likelihoods $\lambda_g(\boldsymbol{c}_i, \boldsymbol{c}_j)$ between the putative correspondence set $\boldsymbol{c}_i \in \mathcal{C}$ and the voting set $\boldsymbol{c}_j \in \mathcal{C}^l$ are formulated as

$$\lambda_g(\boldsymbol{c}_i, \boldsymbol{c}_j) = \exp\left(-\frac{\left\|\mathbf{R}(\boldsymbol{c}_j)\boldsymbol{p}_i + \boldsymbol{t}(\boldsymbol{c}_j) - \acute{\boldsymbol{p}}_i\right\|_2^2}{2{\sigma_e}^2}\right), \tag{6}$$

where σ_e is the standard deviation of the error tolerance, which led to compelling discriminative likelihoods in our experiments when set to four times the acquisition accuracy (i.e., $\sigma_e = 4\sigma_a$). From an efficiency point of view, the global pairwise likelihoods $\lambda_g \in \mathbb{R}^{|C|\times|\mathcal{C}^l|}$ should be computed once, and then reused in the following steps.

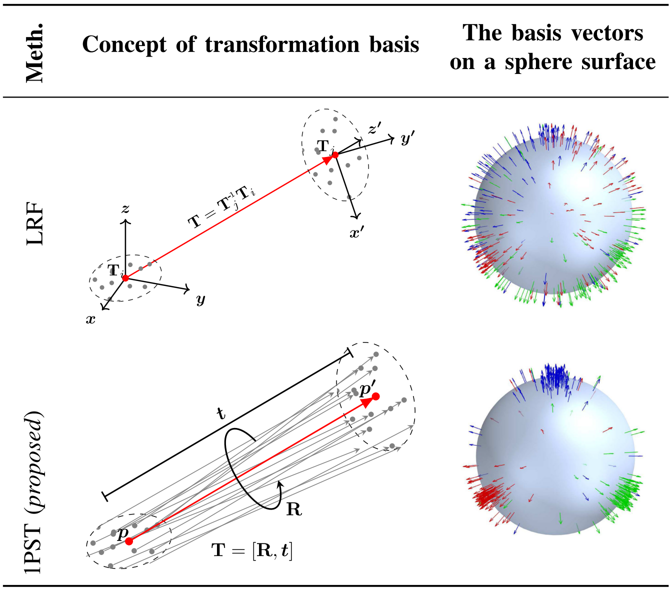
Meth.
Concept of transformation basis
The basis vectors on a sphere surface
LRF
z'
y'
$\mathbf{T}_j$
$\mathbf{T} = \mathbf{T}_j^{-1}\mathbf{T}_i$
x'
z
$\mathbf{T}_i$
y
x
1PST (proposed)
$\boldsymbol{t}$
$\boldsymbol{p}'$
$\boldsymbol{p}$
$\mathbf{R}$
$\mathbf{T} = [\mathbf{R}, \boldsymbol{t}]$

Fig. 2. The concept of both global-geometric consistency techniques: the existing LRF (local reference frame) and the proposed 1PST (single-point superimposition transform). The first column depicts a single correspondence, shown as a red line, between both the source and destination points, $\boldsymbol{p}$ and $\acute{\boldsymbol{p}}$, for which the elliptical dotted spheres surrounding them depicts their neighborhoods. The second column shows the basis vectors formed using the corresponding technique over a sphere surface for all correspondences within a voting set. Since LRF, as the name implies, computes the basis transformation of each source and destination point separately, $\mathbf{T}_i \in \mathbb{SO}_3$ and $\mathbf{T}_j \in \mathbb{SO}_3$, the transformation basis for the correspondence is obtained by composing the inverse of its destination point transform and the source point transform, $\mathbf{T} = \mathbf{T}_j^{-1}\mathbf{T}_i$. However, there are two issues with the LRF technique: (1) due to eigenvectors' sign ambiguity of each point transformation, there is no guarantee that their composed transform is rotational ambiguity-free, even after following certain conventions to resolve them. (2) the impurities within the neighborhoods of each point are not accounted for. As a result, the transforms of the voting set of previous methods [23], [24] that utilizes LRF are very chaotic, as shown over the sphere surface on the first row. On the other hand, 1PST takes into consideration the neighboring correspondences (shown as gray lines) to filter out impurities, as well as computing the correspondence transform, $\mathbf{T} \in \mathbb{SE}_3$, from both the translation, $\boldsymbol{t} \in \mathbb{R}^3$, and rotation, $\mathbf{R} \in \mathbb{SO}_3$, in a single pass to avoid sign ambiguities. As a result, the voting set transformation is accurate, as shown over the sphere surface in the second row.

Similar to Eq. (2), the global coarse ranking set $\mathcal{G} = \{\Lambda_g(\boldsymbol{c}_j) : \boldsymbol{c}_j \in \mathcal{C}^l\} \subset \mathbb{R}$ is estimated as the sum of the global pairwise-likelihoods over the entire putative correspondence set

$$\Lambda_g(\boldsymbol{c}_j) = \sum_{\boldsymbol{c}_i \in \mathcal{C}} \lambda_g(\boldsymbol{c}_i, \boldsymbol{c}_j). \tag{7}$$

Then, similar to Eq. (3), the voting set $\mathcal{C}^l$ is sorted in a descending order, according to the global coarse ranking scores $\mathcal{G}$, and is denoted as $\mathcal{C}^{\mathcal{G}}$. The top k_g-elements constitute the post-validated voting set $\mathcal{C}^g$

$$\begin{aligned} \mathcal{C}^{\mathcal{G}} &= \{\mathcal{C}^l_i : i \in \arg\,\text{sort}(-\mathcal{G})\}, \\ \mathcal{C}^g &= \{\mathcal{C}^{\mathcal{G}}_i\}_{i=1}^{k_g}, \end{aligned} \tag{8}$$

where k_g has some relation to the expected number of the multi-structures in the scene. In this work, we set $k_g = 1$, as the manuscript's scope is limited to single-structure rigid-

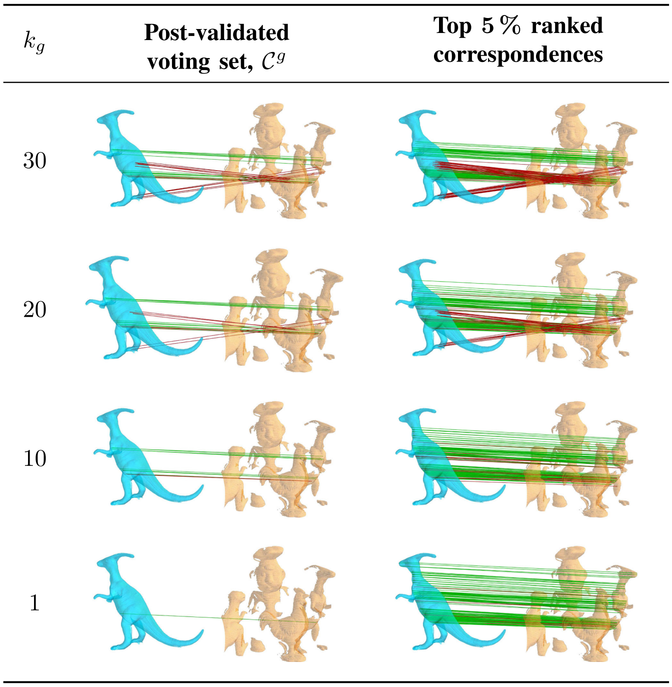
k_g
Post-validated voting set, $\mathcal{C}^g$
Top 5 % ranked correspondences
30
20
10
1

Fig. 3. The effect of the $k_g = |\mathcal{C}^g|$ parameter on the accuracy of the ranking processes, from which it is apparent that $k_g = 1$ is sufficient for our purposes. The green-colored lines indicate inlier correspondences, while the magenta-colored ones indicate outliers, while $\mathcal{C}^g$ is the post-validated voting set [Eq. (8)].

body correspondences. At any rate, $k_g \ll k_l$ should be maintained for proper likelihood estimation, as demonstrated in Fig. 3.

Finally, the fine-tuned likelihoods $\mathcal{S} = \{s(\boldsymbol{c}_i) : \boldsymbol{c}_i \in \mathcal{C}\}$ are estimated for each putative correspondence $\boldsymbol{c}_i \in \mathcal{C}$, by averaging their pairwise-likelihoods over the post-validated voting set

$$s(\boldsymbol{c}_i) = \frac{1}{k_g} \sum_{\boldsymbol{c}_j \in \mathcal{C}^g} \lambda_g(\boldsymbol{c}_i, \boldsymbol{c}_j). \tag{9}$$

This concludes the description of our proposed methodology.

For the sake of simplicity and ease of comparability with the state-of-the-art methods, we denote all methods as functions $f(\cdot,\cdot)$ of two arguments: the first denoting sorting/trimming stage technique and the second denoting the scoring stage technique. As the sorting stage in our proposed method was based on the LRC of the putative correspondences, and the scoring stage was based on the 1PST of such correspondences, we refer to our proposed method as $f(\mathrm{LRC}, 1\mathrm{PST})$.

3 Experimental Setup

In Section 2, we have proposed a two-stage voting scheme, in which a voting set is elected in the first stage and is filtered further in the second stage, before utilizing it to estimate the inlier likelihoods of the input putative correspondences. To demonstrate the accuracy and efficiency of this proposal, we performed experiments to compare our approach with the state-of-the-art. This section is dedicated to describing the experimental setup. First, the utilized dataset is introduced, and we explain how the putative correspondences were obtained from it. After that, the performance indices are discussed and, then, the compared methods and their parameters are briefly introduced.

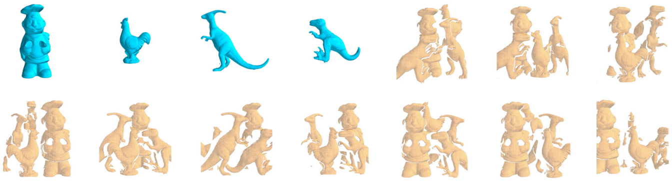

Fig. 4. Sample scans of the U3OR (UWA 3D object recognition) dataset [57], [66], which was utilized in our comparative evaluation. The dataset consists of four models (shown in teal color, namely: 'Chef', 'Chicken', 'Parasaurolophus', and 'T-Rex') and 50 scenes (the first ten are shown in tan color).

All compared methods were implemented as single-threaded Python$^{\mathrm{TM}}$ [62] scripts, and were evaluated using a laptop computer with a 2.7 GHz processor and 8 GB of available memory. Internally, the highly optimized NumPy package [63] was utilized for the linear algebra operations, while the 2D graphs and 3D graphics were generated using the Matplotlib [64] and the MayaVi [65] packages, respectively.

3.1 Dataset

The UWA 3D object recognition (U3OR) dataset [57], [66] features various real-world scanned objects, shown in numerous scenes with different occlusion and clutter rates (Fig. 4), and was utilized in our comparative experiments. The dataset consists of four models ('Chef', 'Chicken', 'Parasaurolophus', and 'T-Rex') and 50 scenes (RS1 to RS50). Each scene includes partial information about several models, with a total of 188 model–scene combinations and two different down-samplings; 374 of these combinations were utilized in this manuscript, as two pairs had no inlier correspondences after sampling their point clouds.

In order to generate the putative correspondence set, the dataset models and scenes were down-sampled to new resolutions, 2 and 5 mm, represented by 1 voxel hereafter. Additionally, since some related studies [23], [24] utilize LRF, which depends on surface normals, the datasets' surface normals were estimated using a principal component analysis (PCA)-based method [67]. Concisely, the normal vector is the eigenvector corresponding to the smallest eigenvalue of the covariance matrix constructed with the k-NN of a point. We limited the surface-normals neighborhood size to $k = 30$ with a radius of 2 voxel, to neutralize both over-sampled and distant points. After that, fast point feature histograms (FPFH) features [36] were computed for the down-sampled point clouds, which resulted in a 33-dimensional vector for each point, representing the point's spatial features. Finally, the feature-space vectors of the models and scenes were matched together to form the putative correspondences $\mathcal{C}$ using an approximate k-NN method [60] with $k = 1$. See Fig. 5 for a visualized conceptualization. The ground truth of a correspondence was constructed based on the ground-truth relative pose transformations of the model-scene pairs, which were part of the dataset as well. A correspondence was considered an inlier if it varied covariantly with the ground truth superimposition transform $\mathbf{T}^{gt}$ within an acceptable tolerance; that is, $\mathcal{C}^{gt} = \{\boldsymbol{c} : \left\|\mathbf{T}^{gt}\boldsymbol{p}_i - \acute{\boldsymbol{p}}_i\right\|_2 < \delta_e, \boldsymbol{c} \in \mathcal{C}\}$. The tolerance was set

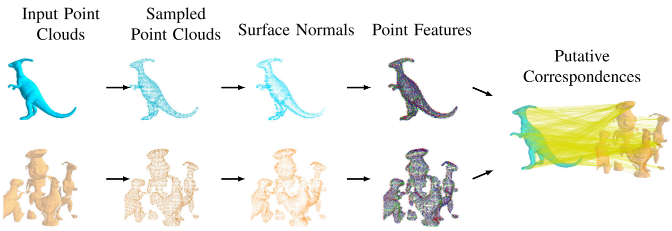

Fig. 5. Extraction of the putative correspondence set from the dataset. The dataset models (shown in teal color) and scenes (shown in tan color) are down-sampled, their surface normals and point features are computed, and finally, they are matched together to generate the putative correspondence set (the yellow-colored dense lines), which forms an input to the evaluated methods.

as twice the resolution of the point cloud, $\delta_e = 2$ voxel. Additionally, the dataset has an occlusion-rate ground truth, which we utilized along with the inlier fraction $|\mathcal{C}^{gt}|/|\mathcal{C}|$ to evaluate the robustness of the methods against these two challenges. For conciseness, only the means and standard deviations of the conducted experiments are reported.

3.2 Performance Metrics

As for the performance metrics, in order to evaluate both the accuracy and efficiency, we measured the precision–recall (PR) area under curve (AUC) and the execution time of each evaluated method. While it is the standard, in the context of retrieval and binary classification problems, to utilize the PR criteria, such a criteria represents only a single operating point of the voting scheme at a specific threshold. In other words, one must choose a score threshold $0 \leq s \leq 1$ to form a selected correspondence set $\mathcal{C}^s = \{\boldsymbol{c} : \mathcal{S}(\boldsymbol{c}) \geq s, \boldsymbol{c} \in \mathcal{C}\}$, and thus compute the precision $p(s) = |\mathcal{C}^{gt} \cap\ \mathcal{C}^s|/|\mathcal{C}^s|$ and the recall $r(s) = |\mathcal{C}^{gt} \cap\ \mathcal{C}^s|/|\mathcal{C}^{gt}|$. On the other hand, the aim is to evaluate the operating characteristics of the voting scheme for any threshold, hence PR AUC is the most appropriate single-number criterion capturing the parametric PR behavior throughout the entire range of s

$$PR_{AUC} = \int_0^1 p(s(r))\mathrm{d}r = \int_1^0 p(s)r'(s)\mathrm{d}s, \tag{10}$$

where $r'(\cdot)$ is the recall derivative.

3.3 Compared Methods

In our comparative evaluation, besides our proposed method, three baseline methods—nearest-neighbor distance (NND), NNSR, and LRC—, two state-of-the-art methods [23], [24], and a robust randomized method (Optimal RANSAC [41]) were used for comparison.

3.3.1 The Baseline Methods

Nearest-neighbor distance scores putative correspondences according to their feature-space distances, while nearest-neighbor similarity ratio scores correspondence distinctiveness using the second-to-first feature-space distance ratio. These two methods are not expected to have significant scores, as they only depend on feature-based measurements. Local rigidity constraint, however, verifies the rigidity constraint of a correspondence in its local neighborhood, as described in Section 2.1, where its score is the normalized value of Eq. (2). This method is expected to perform quite well in comparison to the previous two, which is why it forms part of our proposed method.

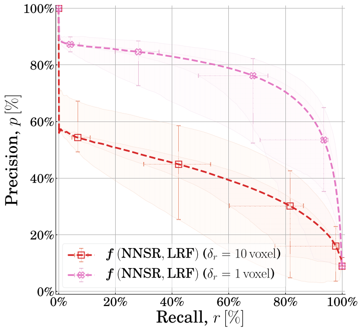

Fig. 6. Updating the parameter of a related study. Based on our experimental results, we updated the parameter δ_r of $f(\mathrm{NNSR}, \mathrm{LRF})$ related method [24] from 10 voxel to 1 voxel to achieve better performance. A higher the PR curve indicates better accuracy. Noting that NNSR (nearest-neighbor similarity ratio) and LRF (local reference frame) denote the underlying techniques utilized within the related method [24].

3.3.2 The State-of-the-Art Methods

The existing method of Ref. [23] initializes a voting set using both NNSR and LRC for inlier sorting, and both LRC and local reference frames are utilized for the final scoring. Accordingly, we denote the method of Ref. [23] by $f(\mathrm{NNSR} + LRC, \mathrm{LRC} + \mathrm{LRF})$. Similarly, the more recent existing method [24] is denoted by $f(\mathrm{NNSR}, \mathrm{LRF})$, as it sorts the voting set in the local phase using NNSR scores, and then utilizes LRFs to construct $\mathbb{SO}_3$ hypotheses for global verification and scoring. Refer to the paragraph just above Section 3 for the interpretation of $f(\cdot, \cdot)$.

3.3.3 Optimal RANSAC

Random sample consensus is a randomized method [40], in which tentative sample correspondences are iteratively drawn at random, and a superimposition hypothesis $\mathbf{T} \in \mathbb{SE}_3$ is elected if it receives sufficient support from the putative set. Optimal RANSAC [41], which we compare our approach against, improves upon the standard RANSAC algorithm's robustness by resampling the tentative correspondence set, ensuring repeatability. The scores are obtained according to the formula $s(\boldsymbol{c}_i) = \exp(-\|\mathbf{T}\boldsymbol{p}_i - \acute{\boldsymbol{p}}_i\|_2^2/2{\sigma_e}^2)$, which resembles Eqs. (6) and (9) in our proposed method.

3.4 Parameters

As for the parameters, we set the cardinality of the voting set $|\mathcal{C}^l|$ (and, thus, the local rigidity neighborhood size) to $k_l = 100$ in all of our experiments. The deviation of the acquisition accuracy, introduced in Eq. (1), was set to one fourth of the point cloud resolution, $\sigma_a = \frac{1}{4}$voxel. Thus, $\sigma_e = 4\sigma_a$ in Eq. (6) became 1 voxel. The LRC and Optimal RANSAC methods were assigned the same values of the σ_a and σ_e parameters. It is also worth noting that the reported parameter of Ref. [24], $\delta_r = 10$voxel, did not seem well-tuned and placed their method in a bad light. So here, we tuned it to 1 voxel for better performance, as shown in Fig. 6. The remaining parameters of both state-of-the-art methods [23], [24] were as suggested by their corresponding manuscripts.

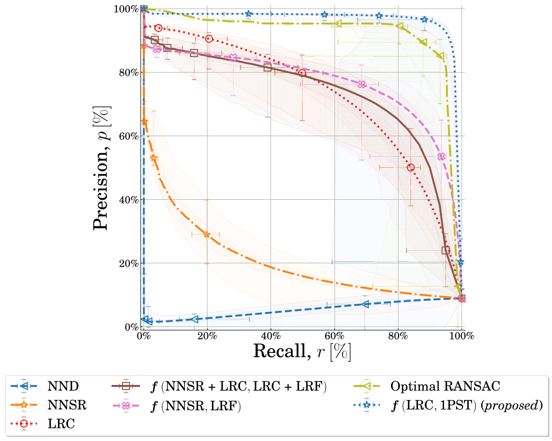

Fig. 7. Accuracy of the seven methods, in terms of their precision and recall means (lines) and standard deviations (shades) over 374 experiments. To enhance clarity, the nth means and deviations are denoted by the points and the error bars. The horizontal axis corresponds to the recall metric, while the vertical axis is the precision metric. In this parametric plot, the higher a curve from the horizontal axis, the better its results is. The proposed scheme $f(\mathrm{LRC}, 1\mathrm{PST})$ remains accurate over a large recall range and outperforms all the related methods, including the two state-of-the-art methods and a robust randomized method. Refer to Section 3.3 for the details of compared methods.

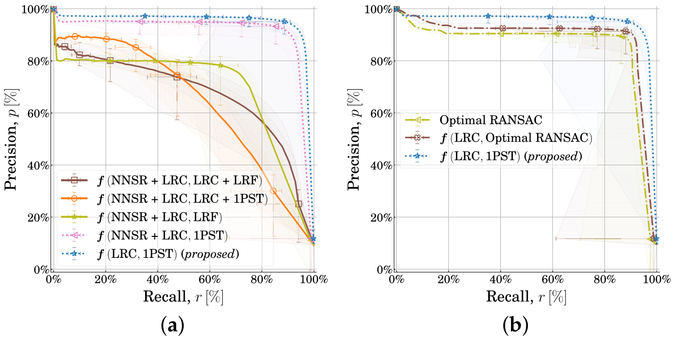

Fig. 8. Demonstration of the impact of the proposed stages, by comparing the proposed method to variants of competing methods combined with one of the proposed stages in terms of their precision and recall means (lines) and standard deviations (shades). To enhance clarity, the nth means and deviations are denoted by the points and the error bars. The horizontal axis corresponds to the recall metric, while the vertical axis is the precision metric. In this parametric plot, the higher a curve from the horizontal axis, the better its results. Note that only the data of 186 experiments (5 mm resolution) are utilized in these graphs, to avoid memory issues in the combined methods: (a) $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + LRF)$ method [23] with and without the proposed 1PST (single-point superimposition transform) and (b) Optimal RANSAC [41] with and without LRC (local rigidity constraint).

4 Results and Discussion

The proposed voting scheme in Section 2 was evaluated on the U3OR dataset [57], [66], after computing its putative correspondences (Section 3.1). The accuracy and efficiency of the results were interpreted in terms of the PR AUC and execution time criteria (Section 3.2) for the proposed method and several other methods, including the state-of-the-art methods [23], [24] (Section 3.3). This section initiates a discussion with the quantitative accuracy and efficiency results, while the qualitative results follows in later parts.

The proposed method $f(\mathrm{LRC}, 1\mathrm{PST})$ outperforms all compared methods, scoring $97.0\% \pm 12.9\%$ on the PR AUC metric PR_{AUC}, as shown in Fig. 7. This substantially high score in terms of precision and recall is attributed to the voting set post-validation (Eq. (8)) by collecting the putative correspondence support. Indeed, there is a similarity between the proposed method's operating characteristics and those of Optimal RANSAC, which utilizes the closely related concept of hypothesis support and scores the nearest to the proposed method: $95.2\% \pm 20.4\%$. On the other hand, it is worth noting that RANSAC is a randomized algorithm, thus its repeatability cannot match our deterministic voting scheme.

Additionally in Fig. 7, the baseline methods NND, NNSR, and LRC have PR AUC values of $5.7\% \pm 2.8\%$, $20.1\% \pm 10.4\%$, and $73.4\% \pm 22.0\%$, respectively. NND has the lowest operating characteristics, its precision remains below the correspondence inlier fraction $|\mathcal{C}^{gt}|/|\mathcal{C}|$. We believe that such worse-than-guessing performance is related to the indistinctness and low dimensionality of the FPFH features. These feature properties also affect the NNSR scores, to some extent, and might explain why its PR AUC curve exponentially decays and approaches the inlier fraction.

As for the state-of-the-art methods, $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + \mathrm{LRF})$ [23] scores $74.2\% \pm 22.2\%$, while $f(\mathrm{NNSR}, \mathrm{LRF})$ [24] scores $78.3\% \pm 26.4\%$. Remarkably, the LRC approach performs on par with the state-of-the-art methods, despite its simplicity, and proves to be a highly competent criterion for voting scheme initialization. We believe that the main reason $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + \mathrm{LRF})$ [23] does not score considerably higher than LRC, despite using LRC as part of the method, is two-fold. The major reason is that the local and global scores, LRC and LRF, are multiplied together at the scoring stage, instead of utilizing a weighted summation or relying solely on the global scores (see the following paragraph for further details). Another reason is that it uses hard thresholding, performed in several steps of the scheme, causing some loss of information. Overall, the related methods have some operating characteristic issues, and they are all outperformed by our proposed scheme, including the two state-of-the-art methods and the randomized one.

In order to gain in-depth insights about the novelty and performance of the proposed method $f(\mathrm{LRC}, 1\mathrm{PST})$, especially in comparison to the $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + \mathrm{LRF})$ method [23] and Optimal RANSAC [41], several hybrid combinations between the existing methods and the proposed one were studied. The results are shown in Fig. 8. First, despite the fact that the LRC stage is utilized in the existing method [23], the originality and effectiveness of our formulation is apparent in Fig. 8a. The first curve corresponds to Ref. [23], while the second, $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + 1\mathrm{PST})$, is a hybrid method, replacing LRF with the proposed 1PST. In these two curves, LRC is utilized in the scoring phase, which drastically limits the performance, regardless of what other scoring technique (i.e., LRF or 1PST), it is combined with. This is the major issue with the method of Ref. [23] and can be alleviated by considering a weighted summation or just avoiding LRC in the scoring phase. However, scoring solely with LRF does not distinctively outperform the original approach [23], as demonstrated by the third curve, $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRF})$, due to the incompetencies of LRF. Indeed, no significant performance gain can be observed until 1PST is used solely in the

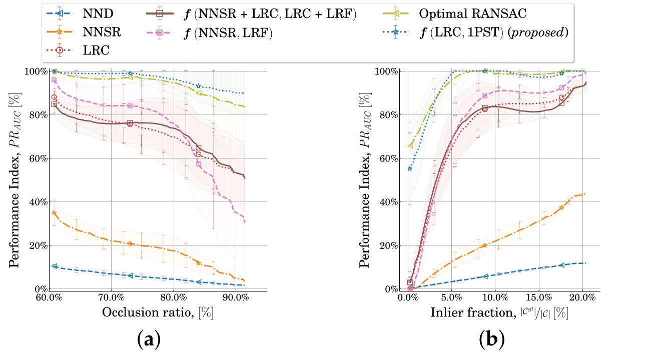

Fig. 9. Robustness analysis of competing methods against occlusions and scarce inliers in terms of the PR AUC means (lines) and standard deviations (shades), which corresponds to the vertical axis. To enhance clarity, the nth means and deviations are denoted by the points and the error bars. The vertical axes correspond to two: (a) the occlusions ratio and (b) the inlier fraction. The occlusion ratio denotes the ratio between observed surface points and the total surface points of the scene, which falls in the approximate range of 62%–93%, where the higher the occlusion ratio, the more challenging the problem. Similarly, the inlier fraction $|\mathcal{C}^{gt}|/|\mathcal{C}|$ denotes the ratio between the count of ground-truth inliers and the correspondences' set size, which falls in the approximate range of 1.5%–21%, where the lower the inlier fraction, the more challenging the problem. In both cases, the higher a curve is, the better its robustness. Refer to Section 3.3 for the details of compared methods.

scoring phase, as per the fourth curve, $f(\mathrm{NNSR} + \mathrm{LRC}, 1\mathrm{PST})$. Similarly, combining NNSR with LRC in the sorting phase harms the performance, rather than benefiting it, as shown by the proposed method's curve, $f(\mathrm{LRC}, 1\mathrm{PST})$, compared to the former ones.

Second, as shown in Fig. 8b, although coupling LRC with Optimal RANSAC [41] enables the truncation of some outliers and, thus, an improvement of the PR AUC score, it is still outperformed by $f(\mathrm{LRC}, 1\mathrm{PST})$, thanks to the proposed 1PST scoring stage. These in-depth experiments indicate the importance of both the LRC sorting and 1PST scoring stages, as no other combination outperforms our approach; thus, stressing the meticulousness and significance of the proposed method. It is worth noting that some hybrid combinations ran out of memory resources when utilizing the 2 mm resolution data; thus, only the 5 mm resolution data was utilized, for which the results are a bit different from, but consistent with, the rest of figures in this manuscript.

As shown in Fig. 9 the PR AUC score is inversely correlated with the occlusion ratio and directly with the inlier fraction; however, in both cases, the proposed method remains robust. Notably, the state-of-the-art method [24] denoted by $f(\mathrm{NNSR}, \mathrm{LRF})$ exhibits a lack of robustness against a high level of occlusions (Fig. 9a). This mostly originates from the fact this method does not perform local neighborhood rigidity checks, but rather uses methods that depend solely on feature-based measurements (i.e., NND and NNSR). Furthermore, LRC and $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + LRF)$ degrade, to some extent, and the most robust methods are the proposed method $f(\mathrm{LRC}, 1\mathrm{PST})$ and Optimal RANSAC. In Fig. 9b, only two methods demonstrate robustness against scarce inliers—namely, the proposed method and Optimal RANSAC—which is mostly due to their hypothesis-support strategies. Therefore, in summary, only the proposed method and Optimal RANSAC exhibit high robustness to scarce inliers and large occlusions.

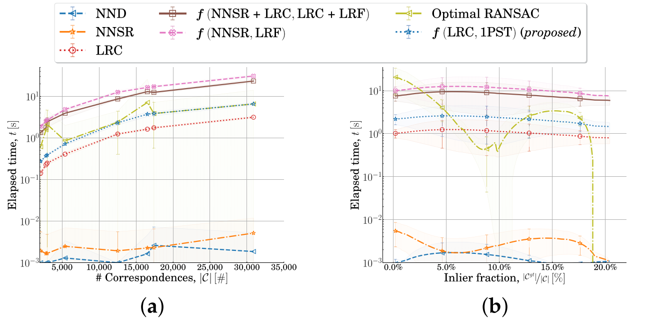

Fig. 10. Analysis of the computational efficiency of the seven methods, in terms of their execution time means (lines) and standard deviations (shades) over 374 experiments. To enhance clarity, the nth means and deviations are denoted by the points and the error bars. The vertical axis denotes the the elapsed time, while the horizontal axes corresponds to (a) correspondence set cardinality $|\mathcal{C}|$, and (b) the inlier fraction $|\mathcal{C}^{gt}|/|\mathcal{C}|$. Refer to Section 3.3 for the details of compared methods.

As for the computational efficiency analysis, aside from Optimal RANSAC, the time complexity of the remaining algorithms, including the proposed method, are $\mathcal{O}(|\mathcal{C}|)$, as the voting sets of all these algorithms are fixed in size. This can be also observed empirically from the linear relationships between problem size and execution time. Fig. 10 shows such relation, but in logarithmic scale to accommodate all results. In comparison to our proposal, as per Fig. 10a, both of the state-of-the-art methods [23], [24], $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + LRF)$ and $f(\mathrm{NNSR}, \mathrm{LRF})$, have higher slope coefficients, which indicates reduced efficiency. This is mostly because the hypothesis transforms are computed for the entire putative set $\mathcal{C}$ and due to the additional computational power being spent on resolving the LRF ambiguities. On the other hand, in the proposed method we compute the hypotheses only for the voting set $\mathcal{C}^{l} \subset \mathcal{C}$, and thus no further overhead ambiguity exists for the 1PSTs.

Importantly, as per Fig. 10b, despite the robustness of Optimal RANSAC, it takes a great deal of time when the inlier fraction is less than 10 percent, which limits its applicability to real-time scenarios. On the other hand, our method consumes a constant time, much less than the state-of-the-art methods, which can be tuned further based on the desired application. Remarkably, the proposed method's execution time is only $24.1\% \pm 6.0\%$ of the time taken by $f(\mathrm{NNSR} + \mathrm{LRC}, \mathrm{LRC} + \mathrm{LRF})$ [23], and $18.0\% \pm 4.1\%$ of the time taken by $f(\mathrm{NNSR}, \mathrm{LRF})$ [24].

Finally, Fig. 11 shows a qualitative evaluation on some of the dataset's more challenging scenes with respect to the four models. These qualitative results show similar tendencies to the conclusions from the PR AUC scores. It is apparent that only Optimal RANSAC and the proposed method $f(\mathrm{LRC}, 1\mathrm{PST})$ perform adequately. Nevertheless, Optimal RANSAC also has failure cases for all of the given qualitative examples, despite its high repeatability and robustness; which indicates the superior robustness and accuracy of the proposed method.

We strongly urge researchers to address the scarcity of unified benchmarks for correspondence evaluation, as well as the lack of consensus on appropriate evaluation criteria. We observed that the reported scores of any given method were often not directly comparable to another, as the proposed methods usually utilize different down-samplings

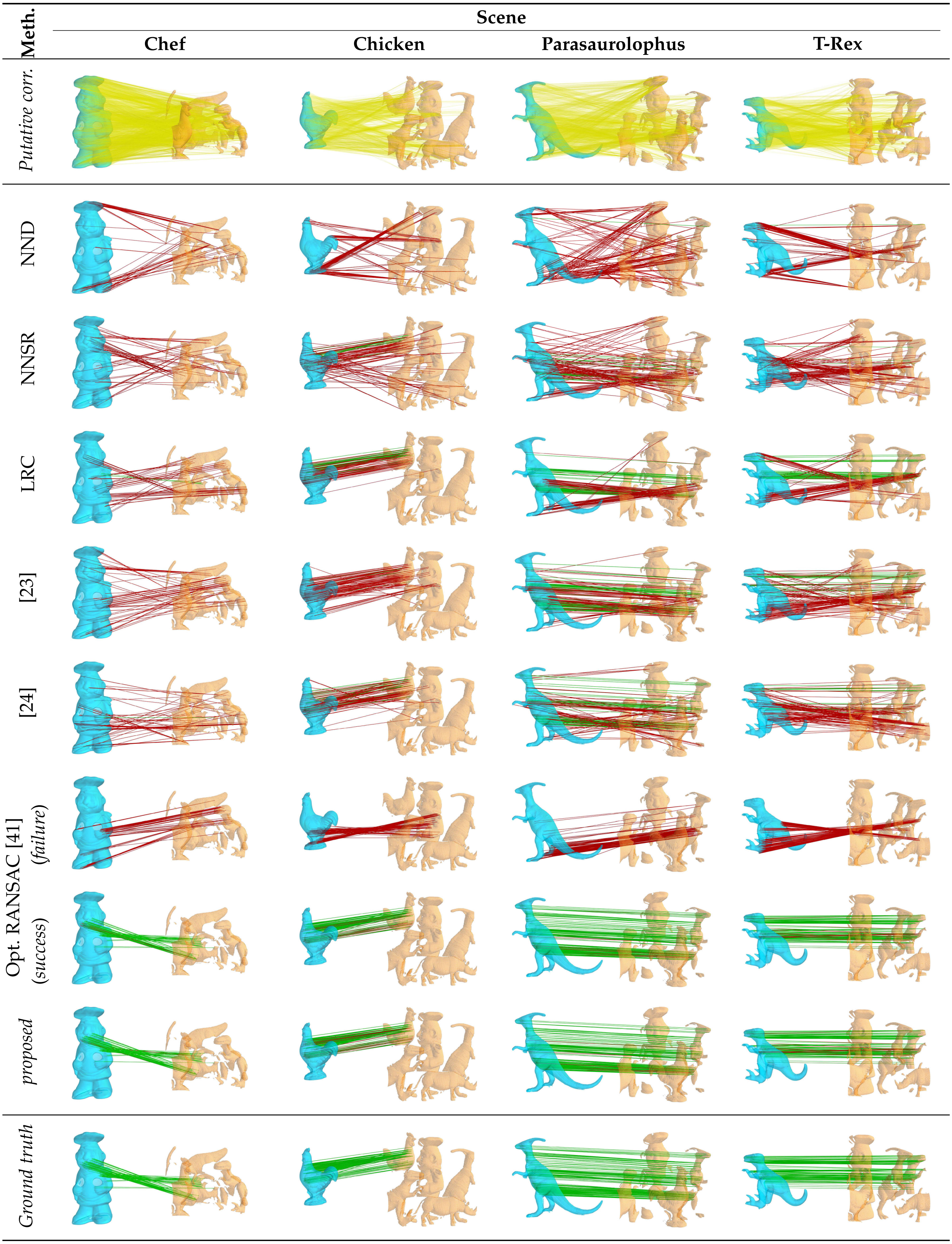

Fig. 11. Qualitative performance matching four models (shown in teal color) with some challenging sample scenes (shown in tan color), all from the U3OR (UWA 3D object recognition) dataset [57], [66] down-sampled to 5 mm. The input correspondences and their ground truths are shown in the first and last rows, respectively. The remaining rows show the qualitative results for each method, which consist of (from top to bottom) three baseline methods, two state-of-the-art voting schemes, a robust randomized method, and the proposed voting scheme. The green-colored lines indicate inlier correspondences, while the magenta-colored ones indicate outliers. Only top-ranked correspondences are shown for each method, with an equal count to their corresponding ground truth. Refer to Section 3.3 for the details of compared methods.

and different features for computing the correspondences, and also even follow different criteria for evaluation.

5 Conclusion

Rejection of spurious correspondences is an essential step for proper geometric modeling, high-level computer vision, and image processing tasks, such as motion estimation, recognition, and reconstruction, among others. Despite all the progress in the field during recent decades, the correspondence problem remains an open one, with few methods tackling it. Most of these methods are either too complex and slow, or lack adequate accuracy. For these reasons, we proposed an extremely efficient, robust, accurate, and simple voting scheme for correspondence scoring. Our proposed method consists of two stages, in which a voting set is elected in the first stage, based on local rigidity consistency. This voting set is post-validated, in the second stage, by enumerating its element-wise global support from the putative correspondence set. The post-validated voting set is then utilized to score the putative correspondence set, based on their pairwise covariances. While the proposal is simple, its novelty lays in the careful formulation for post-validation to solve this chicken-and-egg problem. It is worth noting that the method is very flexible with respect to the utilized correspondence estimation method, as it takes the raw correspondences as input without any additional dependency on the detector, descriptor, matching algorithm, or any additional information (such as correspondence quality scores).

The proposed scheme was evaluated on the U3OR dataset [57], [66]. It demonstrated a high level of accuracy, with an average of $97.0\% \pm 12.9\%$ for the PR AUC criterion over a total of 374 experiments. On the other hand, the state-of-the-art methods [23], [24] scored $74.2\% \pm 22.2\%$ and $78.3\% \pm 26.4\%$; thus, they were outperformed by our proposed method. The proposed method also demonstrated adequate robustness against occlusions and scarce inliers, and a high effectiveness. It seems as though the local rigidity constraint is the major player in occlusion robustness, while collecting hypothesis support improves performance when there are scarce inliers. The effectiveness of our proposed method comes from limiting the hypothesis computation to a relatively small voting set, and thus its execution time is only $24.1\% \pm 6.0\%$ of the time consumed by the fastest state-of-the-art method. Overall, the proposed method exceeded all the compared methods in all aspects.

However, we are not proposing an almighty scheme, as it is currently limited to single-structure rigid-body geometric fitting. Thus, addressing multi-structure geometric modeling would be an interesting extension of this proposal. Moreover, our proposed scheme picks the highest supported hypothesis without resampling. However, resampling seems to make the estimation more robust [41], [68], [69], and thus constitutes one of the future directions. Another future direction considers borrowing the concept of higher-than-minimal subset sampling [70], [71] into the voting schemes, perhaps by considering clustering correspondences [72] or poses [73]. Moreover, only down-sampled point clouds were employed, so involving complete point clouds in hypothesis election or fine-tuning is expected to bring about astonishing results. Finally, while the method is proposed for the 3D scenario, extending it to other scenarios (e.g., 2D), by adapting its scoring and transformation estimation techniques, seems feasible and has great implications.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and express their interest in further discussion and correspondence.

Abbreviations

k-NN k-nearest neighbor; 1PST single-point superimposition transform; AUC area under curve; FPFH fast point feature histograms; LRC local rigidity constraint; LRF local reference frame; NND nearest-neighbor distance; NNSR nearest-neighbor similarity ratio; PCA principal component analysis; PR precision recall; RANSAC random sample consensus; SVD singular value decomposition; U3OR UWA 3D object recognition.

References

[1] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1455–1461, Jul. 2017.

[2] J. Dong, E. Nelson, V. Indelman, N. Michael, and F. Dellaert, "Distributed real-time cooperative localization and mapping using an uncertainty-aware expectation maximization approach," in *Proc. Int. Conf. Robot. Autom.*, 2015, pp. 5807–5814.

[3] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6896–6906.

[4] A. D. Speers, B. Ma, W. R. Jarnagin, S. Himidan, A. L. Simpson, and R. P. Wildes, "Fast and accurate vision-based stereo reconstruction and motion estimation for image-guided liver surgery," *Healthcare Technol. Lett.*, vol. 5, no. 5, pp. 208–214, Oct. 2018.

[5] Y. Jiang, M.-C. Kang, M. Fan, S.-H. Chae, and S.-J. Ko, "A novel relative camera motion estimation algorithm with applications to visual odometry," in *Proc. Int. Symp. Multimedia*, 2018, pp. 215–216.

[6] S. H. Cen and P. Newman, "Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions," in *Proc. Int. Conf. Robot. Autom.*, 2018, pp. 1–8.

[7] M. Ariz, A. Villanueva, and R. Cabeza, "Robust and accurate 2D-tracking-based 3D positioning method: Application to head pose estimation," *Comput. Vis. Image Understanding*, vol. 180, pp. 13–22, Mar. 2019.

[8] N. Pitchandi and S. P. Subramanian, "Image uncertainty-based absolute camera pose estimation with Fibonacci outlier elimination," *J. Intell. Robot. Syst.*, vol. 96, pp. 65–81, Jan. 2019.

[9] V. Larsson, J. Fredriksson, C. Toft, and F. Kahl, "Outlier rejection for absolute pose estimation with known orientation," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 45.1–45.12.

[10] A. G. Buch, L. Kiforenko, and D. Kraft, "Rotational subgroup voting and pose clustering for robust 3D object recognition," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 4137–4145.

[11] M. Yuan, Z. Li, K. W. Wan, and W. Y. Yau, "Outlier detection using hierarchical spatial verification for visual place recognition," in *Proc. Int. Conf. Control Autom. Robot. Vis.*, 2018, pp. 1–6.

[12] Q. Zhu and Z. Mu, "Local and holistic feature fusion for occlusion-robust 3D ear recognition," *Symmetry*, vol. 10, no. 11, Nov. 2018, Art. no. 565.

[13] J.-K. Lee, J. Yea, M.-G. Park, and K.-J. Yoon, "Joint layout estimation and global multi-view registration for indoor reconstruction," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 162–171.

[14] X. Xie, T. Yang, D. Li, Z. Li, and Y. Zhang, "Hierarchical clustering-aligning framework based fast large-scale 3D reconstruction using aerial imagery," *Remote Sens.*, vol. 11, no. 3, Feb. 2019, Art. no. 315.

[15] M. Cao *et al.*, "Fast and robust feature tracking for 3D reconstruction," *Opt. Laser Technol.*, vol. 110, pp. 120–128, Feb. 2019.

[16] Á. P. Bustos and T.-J. Chin, "Guaranteed outlier removal for point cloud registration with correspondences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2868–2882, Dec. 2018.

[17] V. T. Wang and M. P. Hayes, "Synthetic aperture sonar track registration using SIFT image correspondences," *J. Ocean. Eng.*, vol. 42, no. 4, pp. 901–913, Oct. 2017.

[18] N. Luo and Q. Wang, "Effective outlier matches pruning algorithm for rigid pairwise point cloud registration using distance disparity matrix," *IET Comput. Vis.*, vol. 12, no. 2, pp. 220–232, Mar. 2018.

[19] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[20] C. Fu, R. Duan, D. Kircali, and E. Kayacan, "Onboard robust visual tracking for UAVs using a reliable global-local object model," *Sensors*, vol. 16, no. 9, Sep. 2016, Art. no. 1406.

[21] G. Lu, L. Nie, S. Sorensen, and C. Kambhamettu, "Large-scale tracking for images with few textures," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2117–2128, Sep. 2017.

[22] G. Nebehay and R. Pflugfelder, "Clustering of static-adaptive correspondences for deformable object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2784–2791.

[23] A. G. Buch, Y. Yang, N. Kruger, and H. G. Petersen, "In search of inliers: 3D correspondence by local and global voting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2067–2074.

[24] J. Yang, Y. Xiao, Z. Cao, and W. Yang, "Ranking 3D feature correspondences via consistency voting," *Pattern Recognit. Lett.*, vol. 117, pp. 1–8, Jan. 2019.

[25] F. Zhou and F. D. la Torre, "Factorized graph matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 127–134.

[26] L. S. Shapiro and J. M. Brady, "Feature-based correspondence: An eigenvector approach," *Image Vis. Comput.*, vol. 10, no. 5, pp. 283–288, Jun. 1992.

[27] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 1482–1489.

[28] V. Jain, H. Zhang, and O. van Kaick, "Non-rigid spectral correspondence of triangle meshes," *Int. J. Shape Model.*, vol. 13, no. 01, pp. 101–124, Jun. 2007.

[29] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[31] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[33] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys, "3D model matching with viewpoint-invariant patches (VIP)," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[34] Y. Liu, H. Zhang, H. Guo, and N. Xiong, "A fast-brisk feature detector with depth information," *Sensors*, vol. 18, no. 11, Nov. 2018, Art. no. 3908.

[35] H. Sahloul, S. Shirafuji, and J. Ota, "3D affine: An embedding of local image features for viewpoint invariance using RGB-D sensor data," *Sensors*, vol. 19, no. 2, Jan. 2019, Art. no. 291.

[36] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 3212–3217.

[37] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3D object recognition," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2009, pp. 689–696.

[38] S. Salti, F. Tombari, and L. D. Stefano, "SHOT: Unique signatures of histograms for surface and texture description," *Comput. Vis. Image Understanding*, vol. 125, pp. 251–264, Aug. 2014.

[39] K. Jia *et al.*, "ROML: A robust feature correspondence approach for matching objects in a set of images," *Int. J. Comput. Vis.*, vol. 117, no. 2, pp. 173–197, Apr. 2016.

[40] 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, Jun. 1981.

[41] A. Hast, J. Nysjö, and A. Marchetti, "Optimal RANSAC—Towards a repeatable algorithm for finding the optimal set," *J. World Soc. Comput. Graph.*, vol. 21, no. 1, pp. 21–30, 2013.

[42] T.-J. Chin, J. Yu, and D. Suter, "Accelerated hypothesis generation for multi-structure robust fitting," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 533–546.

[43] H. S. Wong, T.-J. Chin, J. Yu, and D. Suter, "Dynamic and hierarchical multi-structure geometric model fitting," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1044–1051.

[44] E. Ask, O. Enqvist, and F. Kahl, "Optimal geometric fitting under the truncated L2-norm," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1722–1729.

[45] W.-Y. D. Lin, M.-M. Cheng, J. Lu, H. Yang, M. N. Do, and P. Torr, "Bilateral functions for global motion modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 341–356.

[46] C. Olsson, O. Enqvist, and F. Kahl, "A polynomial-time bound for matching and registration with outliers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[47] O. Enqvist, K. Josephson, and F. Kahl, "Optimal correspondences from pairwise constraints," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1295–1302.

[48] H. Wang, G. Xiao, Y. Yan, and D. Suter, "Mode-seeking on hypergraphs for robust geometric model fitting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2902–2910.

[49] J. Yan, C. Li, Y. Li, and G. Cao, "Adaptive discrete hypergraph matching," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 765–779, Feb. 2018.

[50] X. Lin, D. Niu, X. Zhao, B. Yang, and C. Zhang, "A novel method for graph matching based on belief propagation," *Neurocomputing*, vol. 325, pp. 131–141, Jan. 2019.

[51] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4181–4190.

[52] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2666–2674.

[53] Z. Zheng, Y. Ma, H. Zheng, J. Ju, and M. Lin, "UGC: Real-time, ultra-robust feature correspondence via unilateral grid-based clustering," *IEEE Access*, vol. 6, pp. 55 501–55 508, Oct. 2018.

[54] H. Chen and B. Bhanu, "3D free-form object recognition in range images using local surface patches," *Pattern Recognit. Lett.*, vol. 28, no. 10, pp. 1252–1262, Jul. 2007.

[55] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 356–369.

[56] J. Novatnack and K. Nishino, "Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 440–453.

[57] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes," *Int. J. Comput. Vis.*, vol. 89, no. 2/3, pp. 348–361, Sep. 2010.

[58] R. Bro, E. Acar, and T. G. Kolda, "Resolving the sign ambiguity in the singular value decomposition," *J. Chemometrics: A J. Chemometrics Soc.*, vol. 22, no. 2, pp. 135–140, Feb. 2008.

[59] P. Yarlagadda, A. Monroy, and B. Ommer, "Voting by grouping dependent parts," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 197–210.

[60] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2009, pp. 331–340.

[61] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, Sep. 1978.

[62] Python Software Foundation, "Python language reference, version 3.6," 2018. [Online]. Available: https://www.python.org/

[63] T. E. Oliphant, *Guide to NumPy*, 2nd ed. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 2015.

[64] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007.

[65] P. Ramachandran and G. Varoquaux, "MayaVI: 3D visualization of scientific data," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 40–51, Mar. 2011.

[66] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1584–1601, Oct. 2006.

[67] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, "Comparison of surface normal estimation methods for range sensing applications," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 3206–3211.

[68] O. Chum, J. Matas, and J. Kittler, "Locally optimized RANSAC," in *Proc. Joint Pattern Recognit. Symp.*, 2003, pp. 236–243.

[69] O. Chum, J. Matas, and S. Obdrzalek, "Enhancing RANSAC by generalized model optimization," in *Proc. Asian Conf. Comput. Vis.*, 2004, pp. 812–817.

[70] T. T. Pham, T.-J. Chin, J. Yu, and D. Suter, "The random cluster model for robust geometric fitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1658–1671, Aug. 2014.

[71] R. B. Tennakoon, A. Bab-Hadiashar, Z. Cao, R. Hoseinnezhad, and D. Suter, "Robust model fitting using higher than minimal subset sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 350–362, Feb. 2016.

[72] A. E. Johnson and M. Hebert, "Surface matching for object recognition in complex three-dimensional scenes," *Image Vis. Comput.*, vol. 16, no. 9/10, pp. 635–651, Jul. 1998.

[73] B. Drost and S. Ilic, "3D object detection and localization using multimodal point pair features," in *Proc. Int. Conf. 3D Imag. Model. Process. Vis. Transmiss.*, 2012, pp. 9–16.

Hamdi Sahloul received the BSc degree in computer and control engineering from Sana'a University, Sana'a, Yemen, in 2010, and the ME and PhD degrees in precision engineering from the University of Tokyo, Tokyo, Japan, in 2016 and 2019, respectively. By 2019, he started working for a robotics automation company, as a computer vision engineer. Between 2010 and 2014, he was working as a machine-to-machine systems designer. His research interests include robotics automation, computer vision, and machine learning.

Shouhei Shirafuji received the PhD degree in information science from Osaka University, Osaka, Japan, in 2014. He was a JSPS research fellow from 2014 to 2015. From 2015 to 2018, he was a postdoctoral researcher with the University of Tokyo, Japan. Since 2018, he has been an assistant professor with the Research into Artifacts, Center for Engineering, University of Tokyo, Japan. His main research interests include mechanical design, robotics, and bio-mechanics.

Jun Ota received the BE, ME, and PhD degrees from the Faculty of Engineering, University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1994, respectively. He is currently a professor of Research into Artifacts, Center for Engineering, (RACE), University of Tokyo, Japan. From 1989 to 1991, he worked with Nippon Steel Cooperation. In 1991, he was a research associate with the University of Tokyo. He became a lecturer and associate professor, in 1994 and 1996, respectively. In April 2009, he became a professor with the Graduate School of Engineering, University of Tokyo. In June 2009, he became a professor of RACE, University of Tokyo. From 2015, he has been a guest professor with the South China University of Technology. From 1996 to 1997, he was a visiting scholar with Stanford University. He received a Fellowship from the Robotics Society of Japan in 2016. His research interests include multi-agent robotic systems, embodied-brain systems science, design support for large-scale production/material handling systems, and human behavior analysis and support.