ISPRS Journal of Photogrammetry and Remote Sensing 196 (2023) 502–513

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Robust feature matching via progressive smoothness consensus

Yifan Xia [a,1], Jie Jiang [b,1], Yifan Lu [a], Wei Liu [b], Jiayi Ma [a,*]

[a] *Electronic Information School, Wuhan University, Wuhan 430072, China*
[b] *Tencent, Shenzhen 518057, China*

ARTICLE INFO

Keywords:
Feature matching
Smooth function estimation
Progressive optimization
Image registration

ABSTRACT

Feature matching is a long-standing fundamental and critical problem in computer vision and photogrammetry. The indirect matching strategy has become a popular choice because of its high precision and generality, but it finds only a limited number of correct matches, and the mismatch removal phase does not utilize the critical feature descriptors. To this end, this paper proposes a novel and effective feature matching method, named *Progressive Smoothness Consensus* (PSC). Our PSC designs an objective function to directly construct correct matches from two feature point sets. To optimize the objective, we introduce a stepwise strategy, where a small but reliable match set with the smooth function is used as initialization, and then the correct match set is iteratively enlarged and optimized by match expansion and smooth function estimation, respectively. In addition, the local geometric constraint is added to the compact representation with a Fourier basis, thus improving the estimation precision. We perform the match expansion as a Bayesian formulation to exploit both the spatial distribution and feature description information, thus finding feasible matches to expand the match set. Extensive experiments on feature matching, homography & fundamental matrix estimation, and image registration are conducted, which demonstrate the advantages of our PSC against state-of-the-art methods in terms of generality and effectiveness. Our code is publicly available at https://github.com/XiaYifan1999/Robust-feature-matching-via-Progressive-Smoothness-Consensus.

1. Introduction

In computer vision and photogrammetry, constructing reliable correspondences between two images with similar contents or objects is a fundamental problem, and it acts as a key prerequisite for a wide spectrum of tasks including image registration and fusion, 3D reconstruction, structure-from-motion, and panoramic stitching (Ma et al., 2021; Jiang et al., 2022a). A common and efficient strategy is to first extract feature points with salient structure in two images and then establish point-to-point matches, which is known as feature matching.

The local feature descriptors provide a significant aid in solving the feature matching problem, such as classic Scale-Invariant Feature Transform (SIFT) (Lowe, 2004). These feature descriptors capture distinctive visual properties around the feature points, and correspondences can be then easily found by the similarities of feature descriptor vectors in two images. However, the inherent ambiguity of local feature representation invariably results in the presence of numerous mismatches (outliers). Therefore, a popular indirect matching strategy incorporating the mismatch removal stage is proposed, which additionally utilizes the spatial distribution constraint to remove the mismatches from the putative matches. In particular, putative matches are generally determined by the threshold of Nearest Neighbor Distance Ratios (NNDR) (Lowe, 2004). Nevertheless, for image pairs with large illumination or viewing-direction changes, too high a threshold would miss many correct matches (inliers) whose ratios are not high enough. While for simple transformations, too low a threshold would produce too many false matches, which may disturb the performance of subsequent mismatch removal. In addition, in the cases of duplicate structures, descriptors of the same world point in two images may not satisfy the nearest neighbor relationship.

As indicated above, the core issue of the mismatch removal stage is to exploit the geometric topology of putative matches to preserve correct matches and remove mismatches. When there is a global geometric transformation (*e.g.,* homography or epipolar) between two images, estimating a parametric model with finite Degrees of Freedom (DOF) based on resampling is an evident choice, such as the long-established RAndom SAmple Consensus (RANSAC) method (Fischler and Bolles, 1981) and its multiple variants (Raguram et al., 2012; Barath et al., 2019, 2020). Even though this framework has been a

[*] Corresponding author.
E-mail addresses: xiayifan@whu.edu.cn (Y. Xia), zeus@tencent.com (J. Jiang), lyf048@whu.edu.cn (Y. Lu), wl2223@columbia.edu (W. Liu), jyma2010@gmail.com (J. Ma).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.isprsjprs.2023.01.016
Received 5 September 2022; Received in revised form 16 November 2022; Accepted 21 January 2023
Available online 31 January 2023

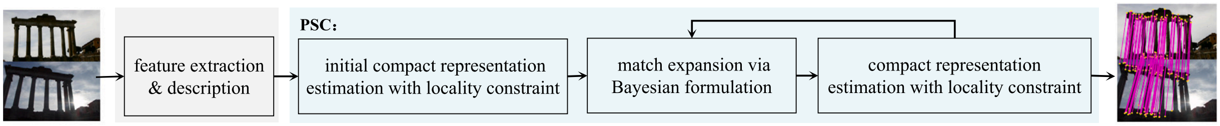
PSC:
feature extraction & description
initial compact representation estimation with locality constraint
match expansion via Bayesian formulation
compact representation estimation with locality constraint

Fig. 1. The overview of PSC, which utilizes the coordinates and descriptors of extracted feature points to establish sufficient and accurate correspondences, thus solving the feature matching problem. In the last match result, the yellow dots represent feature points and the magenta lines represent the correspondences constructed by PSC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

success, it has some fundamental flaws. On the one hand, the required runtime increases exponentially with the rate of outliers, which makes these methods unsuitable for heavily contaminated data. On the other hand, the parametric model restricts the generality of such methods and makes it problematic to solve challenging scenarios, such as drastic distortion.

To address the above issues, the pioneering work Vector Field Consensus (VFC) (Ma et al., 2014) is published and has driven a series of studies based on non-parametric models. Such methods are based on the well-established regularization theory (Girosi et al., 1995), which impose smoothness constraints to create flexible mapping functions that accommodate high DOF. The main drawback of VFC is the high computational complexity (*i.e.,* $O(N^3)$ in time and $O(N^2)$ in space). On this basis, SparseVFC (Ma et al., 2013) and Compact Representation Consensus (CRC) (Fan et al., 2021) have been proposed to reduce the time complexity to $O(N)$. In particular, the latter uses the Fourier basis to construct a coarse-to-fine representation, efficiently maintaining smoothness while having high numerical stability. However, it is likely that the global distribution of a feature set is discontinuous, or that there exists at least one obvious repetitive structure, where a simple smoothness-driven global function would retain many false matches. At this point, observed that local structure is significantly preserved during image transformations, Ma et al. (2015) added local consensus constraint to further distinguish difficult outliers.

However, indirect matching, which simply decomposes feature matching into two steps, still has obvious limitations. Firstly, the correct matches in the final match set are only from the putative matches, which means that the potential correct matches outside the putative matches cannot be found, which may sometimes even neglect almost whole correct matches. Secondly, the mismatch removal stage only considers the spatial distribution but ignores the critical feature descriptor information that may be significant to refine the matching confidence. Therefore, a direct matching method that utilizes both feature descriptors and spatial distributions is prone to solving the feature matching better. In addition, by observing the application of non-parametric models to feature matching, we propose an idea: a small but reliable putative match set can strongly contribute to a valid smooth function, while a reliable global transformation function can in turn contribute to finding correct matches that are missed by the descriptor threshold, and then added matches are conducive to updating a more encompassing non-parametric model. Further, this process can be iterated multiple times to obtain a mapping function that accurately and comprehensively represents the image transformation.

As a consequence, we propose a novel feature matching method, *i.e., Progressive Smoothness Consensus* (PSC). We formulate the direct feature matching problem as a probability-based model and derive a solution by match expansion and smooth function estimation. Specifically, a small but reliable match set with smooth function is obtained by NNDR and smooth function estimation and then used as initialization, and then the correct match set is enlarged by match expansion in description space and optimized by smooth function estimation in spatial distribution space. To efficiently find high-quality matches, the match expansion is formulated in a Bayesian manner and exploits both spatial geometry and feature description information. In addition, to improve the efficiency and accuracy, we introduce the compact Fourier-basis representation with the local geometric constraint to estimate the global transformation. The overall process of our PSC is shown in Fig. 1. Extensive experiments on a wide variety of datasets have demonstrated the advantages of our PSC over the commonly used and state-of-the-art methods, as reflected in terms of generality to various descriptors, effectiveness to feature matching, and promising applicability for high-level vision tasks, such as homography and fundamental matrix estimation and image registration.

The major contributions of this paper are as follows:

- We design a probability-based objective function for direct feature matching. The solution is freed from the limitations of putative matches and thus able to construct more correct matches compared with indirect matching while retaining generality to various descriptors.
- We combine the smooth function estimation with match expansion in a step-wise strategy, where the compact Fourier-basis representation with the locality constraint guarantees the accuracy of PSC while match expansion makes use of both the spatial geometry and feature descriptors to improve the matching effectiveness.

The rest of the paper is organized as follows. Section 2 describes background material and related work. Section 3 introduces the details of our method in terms of problem formulation, smooth function estimation, Bayesian formulation for match expansion, and smooth function progression. Section 4 presents the experimental results on the tasks of feature matching, relative pose estimation, and image registration to demonstrate the superiority of our method. The concluding remarks are drawn in Section 5.

2. Related work

This section briefly reviews the research works relevant to our method and is presented below in three parts, *i.e.,* local feature matching, direct matching, and indirect matching.

2.1. Local feature matching

Feature matching based on local areas of images is usually performed by (i) feature extraction and description, (ii) feature matching between images, and (iii) application-based post-processing. To enhance the performance of the first step in this pipeline, many local feature detection algorithms such as BRISK (Leutenegger et al., 2011), ORB (Rublee et al., 2011), and KAZE (Alcantarilla et al., 2012) are developed after SIFT (Lowe, 2004) and SURF (Bay et al., 2006). Recently, deep learning-based feature description methods have gained significant improvements in repeatability and accuracy, such as local affine invariant-based AffNet (Mishkin et al., 2018), SIFT-based compact descriptor HardNet (Mishchuk et al., 2017), self-supervised framework SuperPoint (DeTone et al., 2018), and Second Order Similarity Network (SOSNet) (Tian et al., 2019). Simple matching methods include Lowe's ratio test (*i.e.,* NNDR) (Lowe, 2004) and the Mutual Nearest Neighbors (MNN), *etc.* In many real cases of large deformations or wide baseline imaging, feature descriptors based on the local structure would have a difficulty in maintaining the same descriptor vector of a real point, so simple filters inevitably produce a large number of outliers or lose many inliers. In such cases, relationships between feature points or global transformations require consideration.

2.2. Direct matching

This type of methods deals directly with two sets of feature points extracted from two images to establish one-to-one point correspondences. Zaragoza et al. (2013) proposed as-projective-as-possible warps based on an estimation technique called Moving Direct Linear Transformation (Moving DLT), and this method is mainly for the image stitching application. Developed for global modeling with smoothness constraint, Lin et al. (2014) used the bilateral domain to reformulate a piece smooth constraint as continuous global modeling. Their subsequent work (Lin et al., 2016) extends their matching method to better suit structure-from-motion by exploiting epipolar constraint. Another existing type of direct matching idea is the use of pair-wise information, for example, graph matching. Conventional graph matching methods (Leordeanu and Hebert, 2005; Cho et al., 2010) usually formulate the problem as an integer quadratic programming, and find an indicator vector that maximizes or minimizes the objective function. The technical difficulties of existing optimization-based methods lie mainly in the high search complexity that grows quadratically with the number of feature points.

To address the above issues, Cho and Lee (2012) proposed a progressive graph matching approach, which progressively updates the fixed size affinity matrix of feature correspondences by repeatedly using RRWM (Cho et al., 2010). Although it improves the efficiency of graph matching, the high computational loss still does not apply to large-scale feature matching. Another type of graph matching methods is based on a MaRkov Field (MRF), which builds graph structure on features of one image. Lee et al. (2020) progressively expanded the MRF model from a small set of seed matches to candidate matches, thus achieving high matching quality with less computation time.

Alternatively, the relaxed methods for direct matching have proven effective. Ma et al. (2018) proposed a guided matching strategy to preserve the neighborhood structure with high generality. Direct matching methods based on deep learning have received attention in the last two years, *i.e.,* using neural networks to match two sets of local features by jointly finding the correct matches and eliminating non-matchable points. Representatively, SuperGlue (Sarlin et al., 2020) proposes a flexible context aggregation mechanism based on attention and achieves a surprising performance. Chen et al. (2021) used a seeded graph matching network for learning to match features. For the reduction in runtime and memory consumption, ClusterGNN (Shi et al., 2022) uses a progressive clustering module to reduce redundant connectivity within subgraphs. However, such data-driven direct matching methods are not general, *i.e.,* they cannot be paired with untrained feature descriptors. For example, SuperGlue is inconvenient to use when paired with a non-public feature descriptor other than SIFT (Lowe, 2004) and SuperPoint (DeTone et al., 2018).

2.3. Indirect matching

Indirect matching has received a lot of attention in recent years as a concise and efficient matching strategy, and it focuses on detecting outliers of putative matches established by descriptor similarity. The resampling-based approaches, represented by RANSAC (Fischler and Bolles, 1981), have been a standard solution for decades. Such methods describe image transformations through parametric models, such as homography or fundamental matrix, *etc*. The improvements of RANSAC include model quality evaluation (Torr and Zisserman, 2000), fast verification (Chum and Matas, 2008) and local optimization (Chum et al., 2003), *etc.* In addition, USAC (Raguram et al., 2012) proposes a unified framework that integrates the various variants of RANSAC. Recently, MAGSAC (Barath et al., 2019) and its improvement (Barath et al., 2020) have moved away from inflexible requirements to set the threshold for inlier and outlier. However, the methods associated with RANSAC rely on the pre-specification of the parametric model, not applicable to real cases with non-rigid transformations.

VFC (Ma et al., 2014), as a practically powerful method, imposes the smoothness constraint in a reproducing kernel Hilbert space to estimate a non-parametric model. Based on this, SparseVFC (Ma et al., 2013) and CRC (Fan et al., 2021) improve the efficiency of matching. Local Linear Transformation (LLT) (Ma et al., 2015) adds local geometric constraint to improve matching accuracy. In addition, a number of relaxed methods have been developed. Grid-based Motion Statistics (GMS) (Bian et al., 2017) identify inliers by dividing the image into multiple grids and then examining the consensus of motion within the grid cells. Ma et al. (2019, 2022) constructed an objective function with a closed-form solution based on local topological consistency. Xia and Ma (2022) proposed locality-guided global-preserving optimization to reject mismatches. With the rise of deep learning techniques, learning-based approaches are gradually demonstrating their superiority in feature matching. For example, Yi et al. (2018) first attempted to use deep learning techniques for mismatch removal. Recently, Zhang et al. (2019) used an order-aware network to learn two-view correspondences and geometry, while Jiang et al. (2022b) proposed a graph attention network for mismatch removal.

However, such indirect matching methods are unable to find correct matches beyond the putative match set, and also ignore the critical information of local feature descriptions in the pruning process. Therefore, how to construct as many correct matches as possible while maintaining the generality and accuracy of indirect matching is meaningful. To address this issue, this paper proposes *Progressive Smoothness Consensus*, which is developed from non-parametric model-based mismatch removal methods such as CRC (Fan et al., 2021).

3. Methodology

This section describes our method for establishing reliable point-to-point matches between two feature point sets, which are extracted from two relevant images and described by feature descriptors (*e.g.,* SIFT (Lowe, 2004)). In the following, the process of our method would be elaborated on in detail.

3.1. Problem formulation

Suppose that we are given two images $\mathbf{A}$ and $\mathbf{B}$, with extracted keypoint sets $\mathcal{K}_{\mathbf{A}}$ and $\mathcal{K}_{\mathbf{B}}$, respectively, where any keypoint $\boldsymbol{\mathcal{K}}^i_{\mathbf{G}} = (\mathbf{p}^i_{\mathbf{G}}, \mathbf{d}^i_{\mathbf{G}})$, with $\mathbf{G} \in \{\mathbf{A}, \mathbf{B}\}$, $\mathbf{p}^i_{\mathbf{G}}$ being the coordinate vector of keypoint i in image $\mathbf{G}$, and $\mathbf{d}^i_{\mathbf{G}}$ being the associated visual descriptor. Our objective is to establish adequate and reliable one-to-one feature point matches across the two images. Denoting $\mathcal{M} = (\mathcal{K}^{\mathcal{M}}_{\mathbf{A}}, \mathcal{K}^{\mathcal{M}}_{\mathbf{B}})$ as the unknown correct match set, the optimal solution of the feature matching problem is:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \mathcal{Q}(\mathcal{M}; \mathcal{K}_{\mathbf{A}}, \mathcal{K}_{\mathbf{B}}), \tag{1}$$

with the objective function $\mathcal{Q}$ defined as:

$$\mathcal{Q}(\mathcal{M}; \mathcal{K}_{\mathbf{A}}, \mathcal{K}_{\mathbf{B}}) = -p(\mathcal{M}' | \mathbf{f}_0, \mathcal{M}_0) p(\mathbf{f}, \mathcal{M} | \mathcal{M}'), \tag{2}$$

where $\mathbf{f}_0$ is the initial smoothness-driven mapping function, $\mathcal{M}_0$ is the initial match set whose geometric relations satisfy $\mathbf{f}_0$, $\mathcal{M}'$ is a set of matches expanded from $\mathcal{M}_0$, and $\mathbf{f}$ is the smooth function describing the final match set $\mathcal{M}$. In particular, $\mathbf{f}_0$ and $\mathcal{M}_0$ are obtained from $\mathcal{K}_{\mathbf{A}}$ and $\mathcal{K}_{\mathbf{B}}$ by NNDR (Lowe, 2004) and a smooth function estimation. The two terms of objective function (2) are sequential in computational time, where the first term concerns the match expansion from $\mathcal{M}_0$ to $\mathcal{M}'$ and the second term refers to the smooth function $\mathbf{f}$ estimation with filtering $\mathcal{M}'$ to $\mathcal{M}$.

3.2. Smooth function estimation

For two feature point sets $\mathcal{K}_{\mathbf{A}}$ and $\mathcal{K}_{\mathbf{B}}$, NNDR (Lowe, 2004) can use the feature descriptor information with the ratio threshold as r_0 to construct a putative match set, in this case smooth function estimation can deal with this set to find $\mathbf{f}_0$ and $\mathcal{M}_0$. Alternatively, after the match expansion, $\mathcal{M}'$ can also be used as a putative match set, in this case smooth function estimation can process this set to find $\mathbf{f}$ and $\mathcal{M}$. In general, smooth function estimation aims to remove the outliers of the putative match set. Suppose a set of putative matches $\mathcal{S} = \{(\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathcal{Y} \ : \ n \in \mathbb{N}_N\}$, where $\mathbf{x}_n = \mathbf{p}^{\mathcal{S}}_{\mathbf{A},n} \in \mathbb{R}^D$ and $\mathbf{y}_n = \mathbf{p}^{\mathcal{S}}_{\mathbf{B},n} \in \mathbb{R}^D$, which denote D-dimensional spatial homogeneous coordinate vectors of feature points. The goal of this subsection is to recover the underlying transformation $\mathbf{y}_n = \mathbf{f}(\mathbf{x}_n)$ from $\mathcal{S}$, and the inlier set $\mathcal{I} \subseteq \mathcal{S}$ to be identified should be consistent with the transformation.

3.2.1. Compact representation with Fourier basis

To efficiently express the transformation, we denote $\mathbf{f}$ by a compact representation inspired by CRC (Fan et al., 2021):

$$\mathbf{f}(\mathbf{x}) = \sum_{t=1}^{T} \mathbf{a}_t \phi_t(\mathbf{x}), \tag{3}$$

where

$$\phi_t(\mathbf{x}) = \prod_{d=1}^{D} \cos\left(x_d \pi j_d^t\right), \quad \boldsymbol{j}^t \in \mathbb{N}^D, \tag{4}$$

x_d denotes the dth component of $\mathbf{x}$, j_d^t denotes the dth component of $\boldsymbol{j}^t$, and $\mathbf{a}_t \in \mathbb{R}^D$ is the function coefficient vector. Considering the steep distribution of feature correspondences near the boundary, the compact representation is expected to satisfy the Neumann condition with low complexity and please refer to Fan et al. (2021) for details. Hence, each basis function $\phi_t(\mathbf{x})$ is computed by the cosine elements of the Fourier basis (Grebenkov and Nguyen, 2013), with the corresponding eigenvalue $\mu_t = \pi^2 \|\boldsymbol{j}^t\|^2$. Specifically, for the smooth constraints of mapping functions, the Fourier basis is arranged in a set $\mathcal{B}_T$ in an ascending order of eigenvalues. The formula is expressed as follows:

$$\mathcal{B}_T := \{\phi_1, \phi_2, \dots, \phi_T\}, \quad 0 \le \mu_1 \le \mu_2 \le \cdots \nearrow \infty. \tag{5}$$

To further regularize the function (3), the coefficients $\mathbf{a} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T]^\top$ are interpreted as random variables with a normal distribution $\mathbf{a} \sim \mathcal{N}(0, \frac{1}{\lambda}\mathbf{R})$, and $\mathbf{R} := \mathrm{diag}(\omega_1, \omega_2, \dots, \omega_T)$. To promote damping of the high-frequency components and thus smoothness of the mapping function, the weights $\{\omega_t\}_{t=1}^T$ are constructed from eigenvalues, *e.g.,* $\omega_t = \mu_t^{-\frac{D}{2}}$, based on the Karhunen–Loeve expansion (Sullivan, 2015).

3.2.2. Mixture model construction

For a robust estimation of $\mathbf{f}$ in the existence of outliers, we make reasonable assumptions about the putative feature correspondences: (i) feature correspondences are independently and identically distributed; (ii) for the inliers, the noise is Gaussian on each component with zero mean and uniform standard deviation σ; (iii) for the outliers, $\mathbf{y}_n$ lies randomly in a bounded region in $\mathbb{R}^D$ and its distribution is uniform $1/a$ with a denoting the region volume. A latent variable $z_n \in \{0, 1\}$ is associated with the nth correspondence $(\mathbf{x}_n, \mathbf{y}_n)$, where $z_n = 1$ indicates the correspondence being an inlier and otherwise for an outlier. In addition, each latent variable z_n fits a discrete distribution, *i.e.,* $p(z_n = 1) = \gamma$, and $p(z_n = 0) = 1 - \gamma$, where $\gamma \in [0, 1]$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N]^\top$ be the position data and the measurements, respectively. Under the *i.i.d.* assumption of data, the joint likelihood function about the mixture model takes the following form:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) &= \prod_{n=1}^{N} \sum_{z_n} p(\mathbf{y}_n, z_n | \mathbf{x}_n, \theta) \\ &= \prod_{n=1}^{N} \left(\frac{\gamma}{(2\pi\sigma^2)^{D/2}} e^{-\frac{\|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2}{2\sigma^2}} + \frac{1-\gamma}{a} \right), \end{aligned} \tag{6}$$

where $\theta = \{\mathbf{f}, \delta, \gamma\}$ denotes a set of unknown parameters. In Bayes rule, θ has a prior distribution $p(\theta)$, and the Maximum A Posteriori (MAP) estimation is formulated as:

$$\boldsymbol{\theta}^* = \arg\max_{\theta} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \arg\max_{\theta} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{7}$$

In the case of $\mathcal{M}'$ as the putative matches, maximizing $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})$ is equivalent to maximizing the second term $p(\mathbf{f}, \mathcal{M}|\mathcal{M}')$ of objective (2). Further, (7) is equivalent to seeking parameters that minimize the following energy:

$$E(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}) - \prod_{n=1}^{N} \ln \sum_{z_n} p(\mathbf{y}_n, z_n | \mathbf{x}_n, \theta). \tag{8}$$

3.2.3. Expectation–maximization solution

Considering the existence of latent variables, the EM algorithm (Dempster et al., 1977) can be used to solve this problem, which alternates with two steps: an expectation step (E-step) and a maximization step (M-step).

Following standard notations, we omit some terms that are independent of θ. Based on the negative log posterior function (8), the expectation of the complete-data log-likelihood is:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = &-\frac{1}{2\sigma^2} \sum_{n=1}^{N} p_n \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2 - \frac{D}{2} \ln \sigma^2 \sum_{n=1}^{N} p_n \\ &+ \ln(1-\gamma) \sum_{n=1}^{N} (1 - p_n) + \ln \gamma \sum_{n=1}^{N} p_n - \ln p(\boldsymbol{\theta}), \end{aligned} \tag{9}$$

where $p_n = P(z_n = 1 | \mathbf{x}_n, \mathbf{y}_n, \theta^{old})$ indicates the posterior probability of z_n. The E-step and M-step are presented below:

E-step: The current parameter values $\boldsymbol{\theta}^{old}$ are used to find the posterior distribution of the latent variables. Using the Bayes rule, it can be computed as follows:

$$p_n = \frac{\gamma e^{-\frac{\|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2}{2\sigma^2}}}{\gamma e^{-\frac{\|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2}{2\sigma^2}} + (1-\gamma)\frac{(2\pi\sigma^2)^{D/2}}{a}}. \tag{10}$$

M-step: The revised parameter estimation $\boldsymbol{\theta}^{new}$ is determined by this step using (9): $\boldsymbol{\theta}^{new} = \arg\max_{\theta} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$. The updating rules for σ and γ can be derived by taking the derivatives of (9) and setting them to 0. In the case of $\mathbf{P} = diag(p_1, p_2, \dots, p_N)$ as a diagonal matrix, we have:

$$\sigma^2 = \frac{tr\left((\mathbf{Y} - \mathbf{T})^\top \mathbf{P} (\mathbf{Y} - \mathbf{T})\right)}{D \cdot tr(\mathbf{P})}, \tag{11}$$

$$\gamma = \frac{tr(\mathbf{P})}{N}, \tag{12}$$

where $\mathbf{T} = \left(\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \dots, \mathbf{f}(\mathbf{x}_n)\right)^T$ and $tr(\cdot)$ is the trace.

3.2.4. Local geometrical constraint

The transformation function $\mathbf{f}$ represents the global geometric relationship between two images, which mainly comes to maintaining the overall smoothness and continuity within the feature domain. However, the global smoothness constraint can be maladaptive in the case of discrete feature distributions, discontinuous motions, and other complex non-rigid transformations, while the topology within the local region remains well maintained. Therefore, imposing constraints on the local structure can improve the accuracy and robustness of matching.

To preserve the local structure, we introduce an efficient technology similar to the LLE (Roweis and Saul, 2000), which characterizes the neighborhood structure in the low-dimensional manifold. Firstly, search for the K nearest neighbors for each feature point in $\mathbf{X}$. $\mathbf{W}$ denotes an $N \times N$ weight matrix, and $W_{ij} = 0$ when $\mathbf{x}_j$ is not in the K-neighborhood of $\mathbf{x}_i$. Secondly, minimize a reconstruction error expressed by the following cost function:

$$\begin{aligned} &\psi(\mathbf{W}) = \sum_{i=1}^{N} p_i \left\| \mathbf{x}_i - \sum_{j=1}^{N} \mathbf{W}_{ij} \mathbf{x}_j \right\|^2, \\ &\text{s.t.,} \quad \forall i, \sum_{j=1}^{N} \mathbf{W}_{ij} = 1. \end{aligned} \tag{13}$$

Optimal weight coefficients $\mathbf{W}_{ij}$ could be obtained by the least squares method. Considering the local conservation after transformation $\mathbf{f}$, a

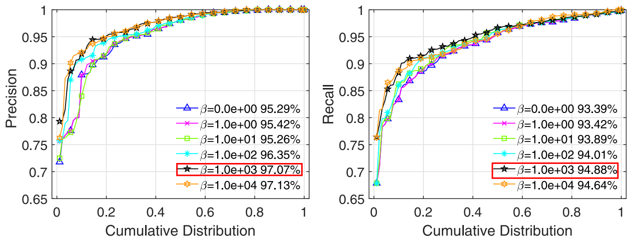

Fig. 2. Precision and recall with respect to the cumulative distribution by using different β to balance local geometric constraint term. When $\beta = 1000$, the smooth function estimation performs well for a wide range of images.

transforming cost term under a probability distribution requires to be minimized: $\sum_{i=1}^{N} p_i \|\mathbf{f}(\mathbf{x}_i) - \sum_{j=1}^{N} \mathbf{W}_{ij}\mathbf{f}(\mathbf{x}_j)\|^2$. Hence, after adding it to the likelihood function (9), the objective function of the M-step is updated as:

$$\hat{Q}(\theta, \theta^{old}) = Q(\theta, \theta^{old}) - \beta \sum_{i=1}^{N} p_i \left\| \mathbf{f}(\mathbf{x}_i) - \sum_{j=1}^{N} \mathbf{W}_{ij}\mathbf{f}(\mathbf{x}_j) \right\|^2, \tag{14}$$

where the parameter $\beta > 0$ balances these two terms. To solve the smooth function $\mathbf{f}$, we assume flat priors for σ and γ, and thus $p(\boldsymbol{\theta})$ degrades into $p(\mathbf{f})$. Abstracting the terms related to $\mathbf{f}$, the following functional is found:

$$\begin{aligned} \varepsilon(\mathbf{f}) =& \frac{1}{2\sigma^2} \sum_{n=1}^{N} p_n \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2 - \ln p(\mathbf{f}) \\ &+ \beta \sum_{i=1}^{N} p_i \left\| \mathbf{f}(\mathbf{x}_i) - \sum_{j=1}^{N} \mathbf{W}_{ij}\mathbf{f}(\mathbf{x}_j) \right\|^2. \end{aligned} \tag{15}$$

As aforementioned, the $\mathbf{f}$ function has $\mathbf{a}$ as coefficients with random distribution, and the term $-\ln p(\mathbf{f})$ can be denoted as $tr(\mathbf{a}^{\top}\mathbf{R}^{-1}\mathbf{a})$. Consequently, with $\lambda > 0$ as a predefined parameter, the following objective problem can be obtained:

$$\begin{aligned} \min_{\mathbf{f} \in span(\mathcal{B}_T)} \quad & \frac{1}{2\sigma^2} \|\mathbf{P}^{1/2}(\mathbf{Y} - \boldsymbol{\Gamma}\mathbf{a})\|^2 + \lambda tr(\mathbf{a}^T \mathbf{R}^{-1}\mathbf{a}) \\ & + \beta \sum_{j=1}^{N} \|\mathbf{P}^{1/2}(\mathbf{I} - \mathbf{W})\boldsymbol{\Gamma}\mathbf{a}\|^2. \end{aligned} \tag{16}$$

It can be seen that functional (16) is convex, and the optimal solution of the coefficients $\mathbf{a}$ of transformation functions can be obtained by solving the following linear system:

$$(\boldsymbol{\Gamma}^{\top}\mathbf{P}\boldsymbol{\Gamma} + 2\lambda\sigma^2\mathbf{R}^{-1} + 2\beta\sigma^2\boldsymbol{\Gamma}^{\top}\mathbf{Q}\boldsymbol{\Gamma})\mathbf{a} = \boldsymbol{\Gamma}^{\top}\mathbf{P}\mathbf{Y}, \tag{17}$$

where

$$\mathbf{Q} = (\mathbf{I} - \mathbf{W})^{\top}\mathbf{P}(\mathbf{I} - \mathbf{W}). \tag{18}$$

After the EM algorithm has converged, we can use a predefined threshold τ to filter matches with high poster probability p_n, thus constructing the inlier set,

$$\mathcal{I} = \{(\mathbf{x}_n, \mathbf{y}_n) : p_n > \tau, n \in \mathbb{N}_N\}. \tag{19}$$

To verify how the local geometrical constraint works, we collect in total 90 image pairs with various image types (30 pairs each of homography, wide-baseline, and nonrigid). Given the putative matches, the smooth function estimation with local geometric constraint aims to identify the correct matches from it. To evaluate the results, the precision ($\frac{\#inliers}{\#preserved\ matches}$) and recall ($\frac{\#inliers}{\#correct\ matches}$) are used as metrics, and the cumulative distributions with different values of parameter β are presented in Fig. 2.

So far, the parameters and inliers of this global transformation model have been derived, and the estimation procedure for the compact representation with locality constraints is summarized in Algorithm 1.

Referring to the objective (2), the inlier set $\mathcal{I}$ denotes the match set $\mathcal{M}_0$ in the case where the putative matches are determined by NNDR, or $\mathcal{I}$ denotes the match set $\mathcal{M}$ in the case of $\mathcal{M}'$ as the putative matches.

Algorithm 1: Smooth function estimation: compact representation with locality constraint

```
Input: The correspondence set $\mathcal{S}$, basis functions $\mathcal{B}_T$, constants
       $\beta$ and $\lambda$, and inlier threshold $\tau$.
Output: Inlier set $\mathcal{I}$, global transformation $\mathbf{f}$
Initialize $\gamma$, $\mathbf{a}$, $\mathbf{T} = \mathbf{0}_{N\times D}$, $\mathbf{P} = \mathbf{I}_{N\times N}$;
Initialize $\sigma^2$ by (11);
Construct $\boldsymbol{\Gamma}$ using $\mathcal{B}_T$;
while $\mathcal{L}$ not converge do
    E-step:
        Update $\mathbf{P}$ by (10);
    M-step:
        Update $\mathbf{a}$ by (17) for $i = 1, 2, \ldots, D$;
        Update $\mathbf{f}$ and $\mathbf{T}$ by (3);
        Update $\sigma^2$ and $\gamma$ by (11) and (12);
end while
Determine the inlier set by (19).
```

3.3. Bayesian formulation for match expansion

The above section derives an initial smooth function $\mathbf{f}_0$ and an initial match set $\mathcal{M}_0$ from a putative match set, and these are estimates of overall correspondence between the two images within the geometry space. In particular, to ensure the high accuracy, the putative match set is determined by NNDR with a high threshold, so it has ignored many potentially correct matches in the feature description space, *e.g.*, feature correspondences with lower nearest neighbor distance ratios and non-nearest neighbor relationships.

To address this issue, we propose the match expansion based on a Bayesian formulation to construct a larger match set $\mathcal{M}'$. Since the match set $\mathcal{M}_0$ is consistent with the mapping function $\mathbf{f}_0$, the match set expansion can use $\mathbf{f}_0$ as a reference to find new feasible matches $\mathcal{M}^{add} = \mathcal{M}' \setminus \mathcal{M}$. In this case, maximizing $p(\mathcal{M}'|\mathbf{f}_0, \mathcal{M}_0)$ in the objective (2) is equivalent to maximizing $p(\mathcal{M}^{add}|\mathbf{f}_0, \mathcal{M}_0)$. To resolve this problem, we propose the following Bayesian formulation, which according to the chain rule can be decomposed into:

$$\begin{aligned} & p(\mathcal{M}^{add}|\mathbf{f}_0, \mathcal{M}_0) = p(\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}, \mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}|\mathbf{f}_0, \mathcal{M}_0) \\ & = p(\mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}|\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}, \mathbf{f}_0, \mathcal{M}_0) p(\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}|\mathbf{f}_0, \mathcal{M}_0), \end{aligned} \tag{20}$$

where $p(\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}|\mathbf{f}_0, \mathcal{M}_0)$ expresses a conditional prior for selecting viable keypoints $\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}$ from image $\mathbf{A}$, and $p(\mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}|\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}, \mathbf{f}_0, \mathcal{M}_0)$ denotes the probability of $\mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}$ relating to the keypoints $\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}$, the mapping function $\mathbf{f}_0$, and initial match set $\mathcal{M}_0$.

During the match expansion process, there is nothing wrong with assuming that match set $\mathcal{M}_0$ are inliers with the reliable reference value. The global transformation $\mathbf{f}_0$ represents the mapping relationship that the match set $\mathcal{M}_0$ conforms to. Naturally then, the greater the distribution density of a keypoint relative to the match set $\mathcal{M}_0$, the closer its potential mapping relationship is to the smooth function $\mathbf{f}_0$. To express this distribution density, we introduce the concept of $\digamma$*-Dist* with the definition as: $\digamma$*-Dist*$(\mathcal{K}_{\mathbf{G}}^{i}, \mathcal{K}_{\mathbf{G}}^{\mathcal{M}})$ is the Euclidean distance from the keypoint $\mathcal{K}_{\mathbf{G}}^{i}$ to the $\digamma$-th nearest neighbor within the keypoint set $\mathcal{K}_{\mathbf{G}}^{\mathcal{M}}$ in the same image $\mathbf{G}$.

Therefore, $p(\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}|\mathbf{f}_0, \mathcal{M}_0)$ can be formulated as:

$$p(\mathcal{K}_{\mathbf{A}}^{i} \in \mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}|\mathbf{f}_0, \mathcal{M}_0) = \begin{cases} 1, & \digamma - Dist(\mathcal{K}_{\mathbf{A}}^{i}, \mathcal{K}_{\mathbf{A}}^{\mathcal{M}_0}) < \epsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where $\mathcal{K}_{\mathbf{A}}^{i} \in \mathcal{K}_{\mathbf{A}} \setminus \mathcal{K}_{\mathbf{A}}^{\mathcal{M}_0}$ is any keypoint in image $\mathbf{A}$ that has not yet been matched, and ϵ is a predefined constant parameter.

The next step is to construct reliable feature matches for the selected keypoints $\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}$, *i.e.*, finding the corresponding keypoints $\mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}$ in image $\mathbf{B}$. This finding process follows two principles: descriptively, the constructed keypoints should have similar description vectors; geometrically, the matches obtained should be close to the global transformation $\mathbf{f}_0$. The detailed selection process is described below. Firstly, with respect to the exploitation of descriptor similarity, we consider that in practical imaging scenarios especially containing repetitive structures, the correct corresponding keypoints are not necessarily the nearest neighbors in the descriptor space. For a given keypoint $\mathcal{K}^i_{\mathbf{A}}$, we consider the top-$\kappa$ descriptor matches as the candidate matches, and the corresponding keypoints from image **B** form a set $\{\mathcal{K}^{i_1}_B, \mathcal{K}^{i_2}_B, \dots, \mathcal{K}^{i_\kappa}_B\}$, where the order is in an ascending order of descriptor distance, *e.g.,* $D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_1}_{\mathbf{B}}) < D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_2}_{\mathbf{B}})$. Here the distances are generally calculated using Euclidean distances. The current set of candidate corresponding keypoints is still redundant, so we further filter the set of candidate corresponding keypoints inspired by the Lowe's ratio (Lowe, 2004). This is done by calculating the relative values describing the descriptor distances and intercepting a portion of candidate keypoints by a threshold. In detail, if there exists an inequality relation with a predefined parameter φ in the set, *i.e.,*

$$\frac{D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_j}_{\mathbf{B}})}{D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_{j-1}}_{\mathbf{B}})} > \varphi, \tag{22}$$

we obtain the filtered candidate keypoint set as $\mathbf{C}_i = \{i_1, i_2, \dots, i_{j-1}\}$. In particular, if more than one inequality relation (22) exists in the top-κ descriptor matches, the first one is selected. If it does not exist, $\mathbf{C}_i = \emptyset$.

Finally, geometric constraints are imposed based on the smooth function $\mathbf{f}_0$ previously obtained, *i.e.,* the potential corresponding keypoints should be in the vicinity of mapping position $\mathbf{f}_0(\mathbf{p}^i_{\mathbf{A}})$. Therefore, the probability formula for the keypoints $\mathcal{K}^{\mathcal{M}^{add}}_{\mathbf{B}}$ is calculated as:

$$p(\mathcal{K}^j_{\mathbf{B}} \in \mathcal{K}^{\mathcal{M}^{add}}_{\mathbf{B}} | \mathcal{K}^i_{\mathbf{A}} \in \mathcal{K}^{\mathcal{M}^{add}}_{\mathbf{A}}, \mathbf{f}_0, \mathcal{M}_0) = \begin{cases} 1, & \mathcal{K}^j_{\mathbf{B}} \in \mathbf{C}_i \text{ and } D\left(\mathbf{p}^j_{\mathbf{B}}, \mathbf{f}_0(\mathbf{p}^i_{\mathbf{A}})\right) < \vartheta, \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

where $\vartheta \propto \frac{D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_j}_{\mathbf{B}})}{D(\mathbf{d}^i_{\mathbf{A}}, \mathbf{d}^{i_{j-1}}_{\mathbf{B}})} - 1$ is an adaptive parameter to balance the two aspects of descriptor similarity and geometric constraints. For example, the constraints imposed on the geometric aspects can be slightly relaxed for potential matches with high similarity in terms of local descriptors.

3.4. Smooth function progression

To increase the efficiency of our method, we design the passing strategy of key coefficients to speed up the convergence of the EM algorithm. Given that the previously constructed transformation $\mathbf{f}_0$ is estimated under relatively strict constraints, $\mathcal{M}_0$ has a higher degree of confidence relative to matches $\mathcal{M}^{add}$ obtained by the match expansion. Therefore, we initialize the posterior probability $p_n = P(z_n = 1|\mathbf{x}_n, \mathbf{y}_n, \theta^{old})$ and the mapping function coefficients **a** as follows:

$$p^{init}_n = \begin{cases} 1, & (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{M}, \\ \zeta, & (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{M}^{add}, \end{cases} \tag{24}$$

$$\mathbf{a}^{init} = \mathbf{a}^{old}, \tag{25}$$

where ζ is a predefined parameter significantly less than 1, and $\mathbf{a}^{old}$ is the coefficient matrix obtained in the previous transformation estimation. Afterward, a new representation function **f** and match set $\mathcal{M}$ can be calculated by the EM solution. In general, the updated match set $\mathcal{M}$ would contain more inliers than $\mathcal{M}_0$, and the smooth function **f** can better represent the true relationships of two images than $\mathbf{f}_0$.

A single match expansion may not be sufficient for feature matching tasks. Therefore, we further propose a stepwise strategy to construct as many correct matches as possible. Assuming M iterations, we have the following new objective function:

$$Q = -\sum_{m=1}^{M} P(\mathcal{M}'_m | \mathbf{f}_{m-1}, \mathcal{M}_{m-1}) P(\mathbf{f}_m, \mathcal{M}_m | \mathcal{M}'_m). \tag{26}$$

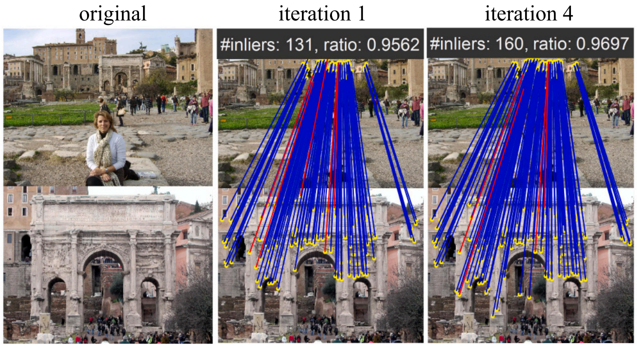

Fig. 3. The progressive smoothness function construction. As the iterative process progresses, each image pair shows the incrementally optimized correspondences. In the two image pairs on the right, the yellow dots represent feature points and the blue lines represent the correct correspondences with red lines denoting wrong correspondences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Algorithm 2: Progressive smoothness consensus

```
Input: Two keypoint sets $\mathcal{K}_{\mathbf{A}}$ and $\mathcal{K}_{\mathbf{B}}$
Output: Keypoint match set $\mathcal{M}^* = (\mathcal{K}^{\mathcal{M}^*}_{\mathbf{A}}, \mathcal{K}^{\mathcal{M}^*}_{\mathbf{B}})$
Obtain $\mathbf{f}_0$ and $\mathcal{M}_0$ by Alg. 1;
while $\mathcal{M}$ expands do
    Obtain $\mathcal{K}^{\mathcal{M}^{add}}_{\mathbf{A}}$ using (21);
    Obtain $\mathcal{M}^{add}_{m-1}$ using (23) and (20);
    Obtain $\mathbf{P}$ and $\mathbf{a}$ using (24) and (25)
    Obtain $\mathbf{f}_m$ and $\mathcal{M}_m$ using (7) by Alg. 1;
end while
Obtain the optimal match set $\mathcal{M}^* = \mathcal{M}_M$.
```

In this case, the cut-off condition for iterations is $|\mathcal{M}'_m| \leq |\mathcal{M}_{m-1}|$ or $|\mathcal{M}_m| \leq |\mathcal{M}_{m-1}|$. And $\mathcal{M}_M$ as the final match set is generally identified as the optimal solution $\mathcal{M}^*$ of the correct feature match set, and the number of iterations is typically $3 \sim 5$ in our experiments.

Fig. 3 graphically illustrates the proposed progressive framework, where matching performance is gradually improved. Because our feature match method is based on a smoothness-driven mapping function that is continuously expanded and updated to obtain the correct matches between two images, we name this method as *Progressive Smoothness Consensus* (PSC). The overall algorithm flow is described in Algorithm 2.

3.5. Computational complexity

For two keypoint sets $\mathcal{K}_A$ and $\mathcal{K}_B$ with numbers as N_A and N_B, respectively, the main time complexity lies in computing the D-dimensional descriptor distances and is $O(N_A N_B D)$. And the space complexity is $O(N_A N_B)$ to store the distance matrix. In addition, for compact representation estimation as Algorithm 1, the main computational cost is to solve the linear system (17). For a putative match set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}^L_{i=1}$, the time complexity of the Fourier representation estimation is linear $O(L)$. And the space complexity is $O(L)$ to store Γ.

In particular, both initial putative matches for estimating $\mathbf{f}_0$ and candidate keypoint match set $\{\mathbf{C}_i\}$ can be established based on the same descriptor distance matrix $\mathbf{D}_{desc} \in \mathbb{R}^{N_A \times N_B}$, and therefore no additional computational loss is incurred. It is worth stating that the time and space complexities are consistent with CRC, therefore guaranteeing the efficiency of PSC.

3.6. Implementation details

There are some parameters of our PSC to be set, *i.e.,* r_0, T, γ, β, λ, K, τ, ϵ, $\digamma$, κ, φ, and ζ. Parameter r_0 is the threshold of the nearest neighbor distance ratio for obtaining the putative matches and set empirically to 1.5. Parameter T is the number of adopted basis functions, which is set empirically to 15. Parameter γ denotes the initial assumption on the inlier proportion of the correspondence set and is set empirically to 0.9. Parameter β controls the trade-off for the regularization of locality geometric constraints, and we empirically set it to 1000 as shown in Fig. 2. Parameter λ represents the balance factor for the regularization of $\mathbf{a}$ and is empirically set to 1. Local geometric constraint preserves K-neighborhood structure in the manifold, where parameter K is empirically set to 15. The threshold τ for filtering matches with posterior probabilities is empirically set to 0.75. Parameter ϵ is the threshold of $\digamma$-Dist for keypoints $\mathcal{K}_{\mathbf{A}}^{\mathcal{M}^{add}}$ and set to 0.005 with $\digamma = 10$. Parameter κ is the number of descriptor matches for initial candidate matches for finding $\mathcal{K}_{\mathbf{B}}^{\mathcal{M}^{add}}$ and empirically set to 10. Parameter φ is the threshold for constructing the candidate keypoint set $\mathbf{C}_i$, which is empirically set to 1.2. Parameter ζ is the passing value for the posterior probability p_n in smoothness representation progression. The data is normalized to $[0,1]^2$ before processing.

4. Experimental results

To comprehensively evaluate and analyze the performance of our PSC, we conduct extensive experiments. Firstly, we use PSC for generic feature matching and compare it with other state-of-the-art methods, as well as test the role of several technical modules of PSC and the generality to different descriptors. Secondly, we conduct the homography & fundamental matrix estimations on large datasets to reliably evaluate our method. Thirdly, we perform the image registration tasks to show the potential value for high-level vision tasks. The open source VLFeat (Vedaldi and Fulkerson, 2010) is used for SIFT, descriptor distance computation, and K-nearest neighbors with K-D tree (Vedaldi and Fulkerson, 2010). All experiments are performed on a desktop with a 2.90 GHz Intel Core CPU, 16 GB memory, and MATLAB R2022a code.

4.1. Feature matching

Feature matching aims to construct matching relationships between two images in terms of feature structure, *i.e.,* point-to-point matching in this paper. For a pair of images containing similar contents or objects, the most commonly used feature extraction and descriptor method SIFT (Lowe, 2004) is chosen to process each image and obtain a feature set for each. In this case, the task for feature matching methods is to find correctly corresponding pairs of feature points from two sets.

Datasets. The well-known *VGG* dataset[2] (Mikolajczyk et al., 2005) and its extended version *Hannover* dataset[3] (Cordes et al., 2013) are used, as they provide the ground-truth homography transformation for each image pair. The two datasets contain a total of thirteen sets (eight from *VGG* and five from *Hannover*), each of which has one source image and five target images totaling six medium resolution images. In each set, a source image and a target image form an image pair, so there are five image pairs with different levels of deformation in each set.

Evaluation metric. Heinly et al. (2012) proposed three evaluation metrics: *putative match ratio*, *precision*, and *matching score*. The *putative match ratio*, PMR $= \frac{\#matches}{\#features}$, which is the ratio of constructed matches to all detected features and represents the selectivity of the methods. A more restrictive method would yield a lower putative match ratio, otherwise the opposite. The *precision*, PC $= \frac{\#correct\ matches}{\#matches}$, indicates the ratio of the number of correct matches to the number of constructed matches. It denotes the purity of feature matching, *i.e.,* how many of the declared feature matches are correct. The *matching score*, MS $= \frac{\#correct\ matches}{\#features}$, represents the how many true matches are constructed among extracted features. It is worth noting that $\#features$ here refers to the number of feature points detected in the source image. In addition, running time, *i.e.,* is also used in the evaluation of feature matching methods.

Competitors. Nine comparison methods are chosen, which are the commonly used or state-of-the-art matchers: NNDR (Lowe, 2004), NNDR+Mutual Nearest Neighbor (MNN), GLPM (Ma et al., 2018), PFM (Lee et al., 2020), MAGSAC++ (Barath et al., 2020), GMS (Bian et al., 2017), OANet (Zhang et al., 2019), LOGO (Xia and Ma, 2022), and CRC (Fan et al., 2021). Briefly, NNDR and MNN are approaches that use only feature descriptors, where the latter would be more stringent than the former. GLPM proposes a guided strategy to preserve the local structure with a closed-form linearithmic solution. PFM formulates the feature matching problem as a Markov random field by progressive graph construction and optimization. These two methods combine the descriptor similarity and geometric structure and are direct matching methods. MAGSAC++ is a resampling-based method with an efficient model quality function. GMS detects the neighborhood consensus of feature points by gridding the image domain. OANet is a deep-learning based method which uses order-aware network and exploit both local and global spatial context to establish correspondences. LOGO uses the local geometric consensus as guide to construct matches using global structure. CRC constructs a smoothness-driven compact representation to filter the mismatches from putative matches. GMS, OANet, LOGO, and CRC are representative mismatch removal methods that deal with putative matches obtained by NNDR.

4.1.1. Qualitative comparison

To visualize the superiority of our PSC, three representative image pairs are selected as objects of feature matching to compare our method with the other three related and state-of-the-art methods. The qualitative results are shown in Fig. 4, where our method is able to construct a sufficient number of correct matches with the highest precision. Therefore, PSC performs satisfactorily even in the case of severe deformations.

4.1.2. Quantitative comparison

Table 1 shows the average performance of each feature matching method for all image pairs of the *VGG* and *Hannover* datasets, where PMR, MS, and PC are used as evaluation metrics. Since MS and PC can directly reflect the matching performance of methods, we highlight the best and second-best results in color. Methods that use only descriptors (NNDR and MNN) exhibit a linear trade-off between the MS and PC relative to the ratio threshold. GLPM and PFM as direct matching methods show significant improvement in MS compared to NN-based methods, *i.e.,* more correct matches between images are constructed. MAGSAC++, GMS, OANet, LOGO, and CRC are mismatch removal methods, which mainly improve matching performance in terms of PC, but still fall short relative to our PSC. In conclusion, our PSC not only demonstrates superiority in matching accuracy but also constructs sufficient matches for subsequent possible visual tasks.

4.1.3. Ablation study

To detail the role of the different technological innovations in our PSC, we construct ablation study on the VGG and Hannover datasets. NNDR, OANet, and CRC serve as comparison methods, where NNDR is the most commonly used, OANet is the representative deep-learning based, and CRC is the most relevant method to our PSC. MS, PC, and runtime (RT) are used as evaluation metrics to assess the matching efficiency and effectiveness of methods. The results are presented in Table 2.

Local geometrical constraint is based on the conservation of small regions during the image transformation, to facilitate the smooth function estimation which represents the global mapping relationship.

[2] https://www.robots.ox.ac.uk/∼vgg/research/affine/

[3] http://www.tnt.uni-hannover.de/project/feature_evaluation/

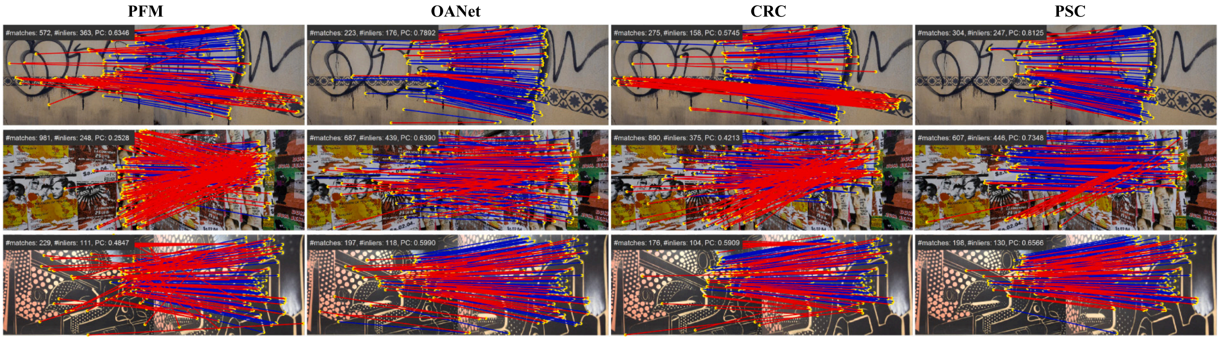
PFM
OANet
CRC
PSC
#matches: 572, #inliers: 363, PC: 0.6346
#matches: 223, #inliers: 176, PC: 0.7892
#matches: 275, #inliers: 158, PC: 0.5745
#matches: 304, #inliers: 247, PC: 0.8125
#matches: 981, #inliers: 248, PC: 0.2528
#matches: 687, #inliers: 439, PC: 0.6390
#matches: 890, #inliers: 375, PC: 0.4213
#matches: 607, #inliers: 446, PC: 0.7348
#matches: 229, #inliers: 111, PC: 0.4847
#matches: 197, #inliers: 118, PC: 0.5990
#matches: 176, #inliers: 104, PC: 0.5909
#matches: 198, #inliers: 130, PC: 0.6566

Fig. 4. Qualitative comparisons on representative image pairs from VGG and Hannovar datasets. In each pair of images, the yellow dots represent feature points, the blue lines are correctly identified matches, and the red lines are incorrect matches. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The quantitative results of feature matching on VGG and Hannover datasets, where our PSC is compared with the other nine methods including NNDR, MNN, GLPM, PFM, MAGSAC++, GMS, OANet, LOGO, and CRC. Putative match ratio (%), matching score (%), and precision (%) are used as evaluation metrics to assess the performance of methods, where the red indicates the best and the blue for the second.

Algorithm	NNDR (Lowe, 2004)	MNN	GLPM (Ma et al., 2018)	PFM (Lee et al., 2020)	MAGSAC++ (Barath et al., 2020)	GMS (Bian et al., 2017)	OANet (Zhang et al., 2019)	LOGO (Xia and Ma, 2022)	CRC (Fan et al., 2021)	PSC
PMR	27.90	22.65	25.85	30.21	24.12	22.29	25.80	23.47	23.63	26.16
MS	21.28	19.87	22.49	25.39	21.34	19.78	20.56	20.98	20.94	22.50
PC	58.33	65.80	67.04	65.30	67.72	66.35	67.93	68.04	67.53	68.15

Table 2
Results of ablation study. VGG and Hannover are the datasets, and MS (%), PC (%), and RT (ms) are regarded as the metrics to evaluate the matching results. w/o L.G.C stands for w/o local geometrical constraint. w/o M.E. stands for w/o match expansion, *i.e.,* only one mismatch removal. w/o S.R.P stands for no smoothness representation progression, *i.e.,* one mismatch removal and one match expansion. **Bold** indicates the best.

Method	MS	PC	RT
NNDR (Lowe, 2004)	21.28	58.33	657.9
OANet (Zhang et al., 2019)	21.46	67.62	822.6
CRC (Fan et al., 2021)	20.94	67.83	661.5
PSC w/o L.G.C	22.38	66.64	611.2
PSC w/o M.E.	20.90	**68.17**	**422.2**
PSC w/o S.R.P.	22.37	58.07	434.0
PSC full	**22.50**	68.15	703.2

For cases such as separated distributions or discontinuous motions, this design can help the algorithm to make the correct matching decision. In Table 2, PSC without local geometrical constraint has significantly lower precision than PSC full.

Match expansion aims to avoid the limitations of putative matches, thus finding as many feasible matches as possible in the feature description space. Therefore, the most obvious effect of this design is to increase the number of correct matches. As reflected in the ablation study, Table 2 indicates that full PSC has a significantly higher matching score compared to PSC without match expansion.

Smoothness representation progression represents the step-wise iterations of smooth function estimation and match expansion, which can both improve the accuracy of matching algorithm and increase the number of correct matches constructed. From Table 2, PSC full is clearly superior to PSC without smoothness representation progression in both matching score and precision.

It is worth noting that for the utilization of descriptor similarity, we calculate a generic descriptor distance matrix and then extract the valid information through matrix functions of MATLAB, which is more efficient than NNDR function of VLFeat in the mismatch removal strategy. Thus, the full PSC accomplishes the most advantages matching performance without much time consumption.

4.1.4. Descriptor generality testing

To test the effectiveness of our PSC with different descriptors, we perform generality testing experiments. In addition to the traditional handcrafted descriptors (SIFT (Lowe, 2004), SURF (Bay et al., 2006), and KAZE (Alcantarilla et al., 2012)), deep-learning based descriptors are also taken into account, including SuperPoint (DeTone et al., 2018), and HardNet (Mishchuk et al., 2017). For brevity, comparison methods include NNDR (Lowe, 2004), PFM (Lee et al., 2020), OANet (Zhang et al., 2019), and CRC (Fan et al., 2021). In particular, considering that the descriptor distance ratio threshold of 1.5 is too strict leading to failure for some descriptors (*e.g.,* HardNet), it is set to 1.3 here for all descriptors. In addition, as PFM requires the transformation orientation and scale information of feature points, it cannot be applied to deep-learning based feature descriptors, which only provide high-dimensional description vectors. MS and PC are used as evaluation metrics, and the results are presented in Table 3. As can be seen from it, our PSC has promising generality for various descriptors and is effective in constructing correct matches with satisfactory matching score and precision.

4.2. Homography & fundamental matrix estimation

Subsequently, we apply PSC to homography & fundamental matrix estimation tasks and compare it with several commonly used or state-of-the-art feature matching methods. In detail, after feature matching methods have obtained point-to-point correspondences between two images, from which a model estimator can derive a geometric model, such as homography or fundamental matrix. In particular, the most common and popular SIFT (Lowe, 2004) is chosen as the feature extraction and description method in estimation experiments.

Datasets. For homography matrix estimation, *HPatches* benchmark[4] (Balntas et al., 2017) contains 580 image pairs with ground-truth homography matrices. In detail, it contains a total of 116 scenes, where each scene consists of one source image and five target images thus five image pairs. And the maximum number of keypoints in each image

[4] https://github.com/hpatches/hpatches-dataset

Table 3
Results of ablation study. VGG and Hannover are the datasets, and MS (%), PC (%), and RT (ms) are regarded as the metrics to evaluate the matching results. w/o L.G.C stands for w/o local geometrical constraint. w/o M.E. stands for w/o match expansion, *i.e.,* only one mismatch removal. w/o S.R.P stands for no smoothness representation progression, *i.e.,* one mismatch removal and one match expansion. **Bold** indicates the best.

Method		NNDR	PFM	OANet	CRC	**PSC**
SIFT (Lowe, 2004)	MS	22.3	**25.4**	19.7	21.9	22.5
	PC	50.8	65.3	66.6	67.1	**68.2**
SURF (Bay et al., 2006)	MS	17.2	**24.0**	15.4	16.7	18.1
	PC	37.5	57.1	61.6	61.7	**62.3**
KAZE (Alcantarilla et al., 2012)	MS	23.2	**30.3**	22.4	23.0	25.5
	PC	49.6	57.3	62.2	**65.6**	64.8
SuPoint (DeTone et al., 2018)	MS	26.4	–	24.0	25.8	**27.8**
	PC	49.9	–	**61.5**	58.4	58.4
HardNet (Mishchuk et al., 2017)	MS	14.3	–	14.1	13.8	**15.4**
	PC	39.2	–	48.3	48.4	**49.0**
Average	MS	20.7	–	19.1	20.2	**22.0**
	PC	45.4	–	60.0	60.2	**60.3**

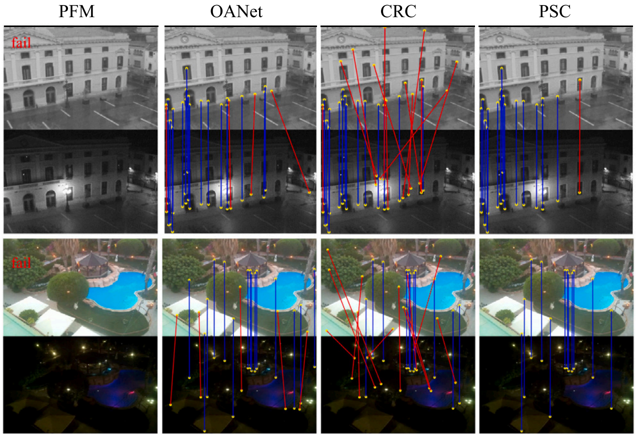

Fig. 5. Qualitative comparisons on the HPatches dataset for homography estimation. From left to right: PFM, OANet, CRC, and LOGO. There are two pairs of images from top to bottom. In each pair of images, the yellow dots represent feature points, the blue lines are correctly identified matches, and the red lines are incorrect matches. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is set to 4000. For fundamental matrix estimation, FM-Bench[5] (Bian et al., 2019) provides four large datasets, *i.e., TUM*, *KITTI*, Tanks and Temples (*T&T*), and Community Photo Collection (*CPC*). Specifically, (i) The *TUM* dataset (Sturm et al., 2012) is commonly used in SLAM and contains indoor scenes with short-baseline images in the resolution of 480 × 640. (ii) The *KITTI* dataset (Geiger et al., 2013) is captured by a camera on a moving vehicle and contains outdoor scenes with short-baseline images in the resolution of 370 × 1226. (iii) The *T&T* dataset (Knapitsch et al., 2017) comprises various scenes or objects and contains wide-baseline image pairs in the image resolution of 1080 × 1920 or 1080 × 2048. (iv) The *CPC* dataset (Wilson and Snavely, 2014) collects unstructured landmark images from Flicker and includes wide-baseline image pairs with different resolutions. In FM-Bench, each dataset contains 1000 pairs of images with ground-truth fundamental matrices. And the maximum number of feature points in each image is also set to 4000.

Evaluation metric. In addition to MS and PC as evaluation metrics, how to calculate the estimation error is the key to evaluating the results of homography & fundamental matrix estimation. After obtaining feature matches between an image pair, we adopt RANSAC (Fischler and Bolles, 1981) as the estimator to derive the geometric model (homography or fundamental matrix), which is compared with the ground-truth to calculate the error and thus assess the performance of methods. For homography estimation, *homography error* defined in SuperPoint (DeTone et al., 2018) is considered, which is less than 4 pixels that the corresponding estimated homography is identified as accurate. For fundamental metric estimation, following the FM-Bench (Bian et al., 2019), Normalized Symmetric Geometry Distance (NSGD) is adopted as the metric, which represents the ratio of SGD (in pixels) to the diagonal length of the source image. In this case, a fundamental matrix estimation is identified as accurate when its NSGD is less than 0.05. The accuracy (Acc.) is the ratio of accurate estimates to all estimates, which is used as the evaluation metric for geometric estimation.

Experimental setting. Due to the different image types, different RANSAC inlier-outlier thresholds for different datasets are set to achieve the best performance: 4 pixels for homography dataset (*i.e., HPatches*), 2 pixels for wide-baseline fundamental matrix datasets (*i.e., T&T* and *CPC*), and 0.5 pixels for short-baseline fundamental matrix datasets (*i.e., TUM* and *KITTI*). It is worth stating that the NNDR threshold involved in the feature matching methods during the experiments is set at 1.3 for homography and short-baseline fundamental matrix datasets, and 1.8 for wide-baseline fundamental matrix datasets.

4.2.1. *Qualitative comparison*

For homography estimation, qualitative comparisons regarding representative image pairs are presented in Fig. 5. Surprisingly, PFM (Lee et al., 2020) fails (cannot find the correct matches) when faced with such sparse sets of feature points, so its robustness is questionable. In contrast, our PSC yields significantly more accurate feature matches relative to two representative mismatch removal methods (OANet (Zhang et al., 2019) and CRC (Fan et al., 2021)), thus facilitating subsequent homography estimation.

For fundamental matrix datasets, the comparison results of feature matching are shown in Fig. 6. FM-Bench contains both short and wide baseline image pairs, as well as indoor and outdoor scenes, from which representative images are presented from top to bottom. As can be seen, the representative direct matching method PFM (Lee et al., 2020) has difficulty in guaranteeing clean matching results, while the advanced indirect matching methods (deep-learning-based OANet (Zhang et al., 2019) and handcrafted CRC (Fan et al., 2021)) are inferior to our PSC in terms of matching accuracy. Therefore, after examination of large datasets, our PSC is capable of constructing adequate and accurate feature matches.

4.2.2. *Quantitative comparison*

For homography estimation, Table 4 shows the quantitative results on HPatches dataset regarding eight comparison methods: NNDR (Lowe,

[5] https://github.com/JiawangBian/FM-Bench

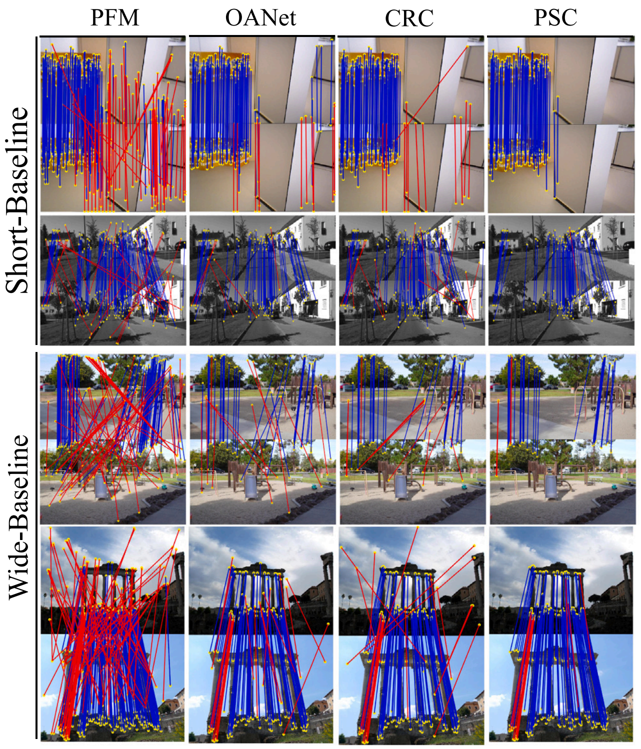

Fig. 6. Qualitative comparisons on short-baseline and wide-baseline fundamental matrix datasets. From left to right: PFM, OANet, CRC, and PSC. From top to bottom: matching results on *TUM*, *KITTI*, *T&T*, and *CPC*, respectively. The first two datasets are short-baseline and the last two are wide-baseline. In each pair of images, the yellow dots represent feature points, the blue lines are correctly identified matches, and the red lines are incorrect matches. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The quantitative comparison of homography estimation on the *HPatches* dataset. **Bold** indicates the best results.

Method	NNDR	MNN	GLPM	PFM	GMS	OANet	CRC	**PSC**
MS	22.6	20.5	22.4	21.2	21.4	21.3	22.2	**23.2**
PC	79.3	90.3	91.6	91.5	89.5	89.7	92.6	**93.3**
Acc.	76.2	78.1	79.5	78.6	79.0	78.1	78.5	**79.7**

Table 5
The quantitative comparison of fundamental matrix estimation on the short-baseline image datasets (*i.e., TUM* and *KITTI*). **Bold** indicates the best results.

Method	*TUM*			*KITTI*		
	Acc.	MS	PC	Acc.	MS	PC
NNDR (Lowe, 2004)	63.9	27.50	67.64	90.1	30.43	77.12
GLPM (Ma et al., 2018)	67.5	27.44	88.93	90.7	30.19	95.72
PFM (Lee et al., 2020)	66.0	32.53	88.76	90.5	33.48	96.08
MSC++ (Barath et al., 2020)	67.9	26.91	94.92	88.9	31.51	97.82
GMS (Bian et al., 2017)	66.2	24.36	92.39	90.9	27.42	96.51
OANet (Zhang et al., 2019)	65.9	24.09	94.61	90.3	25.04	96.44
LOGO (Xia and Ma, 2022)	64.0	24.93	92.70	**91.2**	28.76	96.80
CRC (Fan et al., 2021)	66.5	26.12	94.49	90.6	28.53	97.53
SG (Sarlin et al., 2020)	66.6	**41.03**	92.05	90.8	**43.96**	94.08
PSC	**72.8**	26.34	**94.96**	90.7	29.46	**98.27**

2004), MNN, GLPM (Ma et al., 2018), PFM (Lee et al., 2020), GMS (Bian et al., 2017), OANet (Fan et al., 2021), CRC (Fan et al., 2021), and our PSC. Our PSC has the best performance on three metrics (MS, PC, and Acc.), showing a clear superiority relative to other commonly used and state-of-the-art feature matching methods.

For fundamental matrix estimation, Table 5 presents the result statistics on the short-baseline datasets (*i.e., TUM* and *KITTI*), while Table 6 shows the results on the wide-baseline datasets (*i.e., T&T* and *CPC*). NNDR (Lowe, 2004) only makes use of feature descriptor information, which makes it difficult to estimate geometric transformation effectively. GLPM (Ma et al., 2018) and PFM (Lee et al., 2020) are handcrafted methods to directly match features, but they are usually not accurate enough. MSC++ (Barath et al., 2020), GMS (Bian et al., 2017), OANet (Zhang et al., 2019), and LOGO (Xia and Ma, 2022) are the state-of-the-art methods for mismatch removal, which have low matching scores. CRC (Fan et al., 2021) is worse than our method in all metrics, which indicates that the improvement of PSC over CRC is significant. SG (Sarlin et al., 2020), as an advanced directly feature matching method, is data-driven. However, it performs poorly on accuracy and precision for wide-baseline images, even with high match scores. Compared to the other nine feature matching methods, our PSC exhibits the most accurate feature matching performance with highest precision as a decent matching score. In addition, PSC shows the best geometric estimation performance on *TUM* and *CPC* datasets with highest accuracy. Thus, our PSC demonstrates superiority over other commonly used and state-of-the-art methods in the fundamental matrix estimation task.

Table 6
The quantitative comparison of fundamental matrix estimation on the wide-baseline image datasets (*i.e., T&T* and *CPC*). **Bold** indicates the best results.

Method	*T&T*			*CPC*		
	Acc.	MS	PC	Acc.	MS	PC
NNDR (Lowe, 2004)	70.9	3.22	52.11	38.1	1.91	52.91
GLPM (Ma et al., 2018)	**81.4**	3.47	71.84	47.0	2.19	77.76
PFM (Lee et al., 2020)	74.0	4.33	58.90	41.6	3.18	59.54
MSC++ (Barath et al., 2020)	79.6	3.15	75.29	46.2	1.85	78.50
GMS (Bian et al., 2017)	74.7	2.78	66.78	40.6	1.62	71.34
OANet (Zhang et al., 2019)	80.5	2.68	73.31	47.4	1.77	78.37
LOGO (Xia and Ma, 2022)	74.2	2.98	75.94	44.1	1.83	78.88
CRC (Fan et al., 2021)	79.4	3.06	69.44	43.2	1.79	74.99
SG (Sarlin et al., 2020)	59.7	**8.26**	51.11	27.8	**6.50**	71.25
PSC	79.9	3.37	**76.60**	**48.3**	1.93	**79.21**

4.3. Image registration task

To further exploit the practical value of PSC, we apply it to the image registration task, *i.e.,* maximizing the overlap between the source and target images. The detailed procedure is: firstly, our PSC obtains a set of feature point matches between two images; secondly, Thin Plate Spline (TPS) (Donato and Belongie, 2002) is chosen to estimate a transform function $\mathcal{F}$ because of its mapping generality and smoothness; thirdly, according to $\mathcal{F}$, the pixels in the source image are mapped to the corresponding positions in the target image, after which the pixel intensity of each coordinate of target image is calculated by a bilateral interpolation algorithm (Han et al., 2010).

Datasets. *RS* (Ma et al., 2019) and *720Yun* (Liang et al., 2020) are chosen as experimental datasets, which together contain 92 image pairs. These images are mainly remote-sensing images including color-infrared, SAR, and panchromatic photographs, and the scenes contain terrain, roads, buildings, terraces, *etc.* In particular, the image pairs of *720Yun* conforms to nonrigid transformation, which is challenging for image registration tasks.

Evaluation metric. To evaluate the registration performance of methods, 20 pairs of landmark pixel values $\{r_i, s_i\}_{i=1}^{M=20}$ are randomly chosen from each image pair, in which case Root Mean Square Error (RMSE), MAximum Error (MAE), and MEdian Error (MEE) are employed as quantitative metrics. The definitions of these metrics are as follows:

$$\mathrm{RMSE} = \sqrt{1/M \sum_{i=1}^{M} \left(r_i - \mathcal{F}\left(s_i\right)\right)^2}, \tag{27}$$

$$\mathrm{MAE} = \max \left\{ \sqrt{\left(r_i - \mathcal{F}\left(s_i\right)\right)^2} \right\}_{i=1}^{M}, \tag{28}$$

Table 7
The quantitative results of image registration. The average values and standard deviations of RMSE, MAE, and MEE are used for evaluation. **Bold** indicates the best.

Method	RMSE	MAE	MEE
NNDR (Fischler and Bolles, 1981)	256.3 (±141.8)	511.9 (±252.2)	344.3 (±203.2)
MNN	253.3 (±162.6)	498.3 (±284.6)	340.6 (±229.8)
GLPM (Barath et al., 2020)	164.1 (±147.8)	325.9 (±320.8)	220.9 (±198.3)
PFM (Ma et al., 2014)	110.6 (±158.0)	217.5 (±299.4)	146.0 (±211.7)
MSC++ (Barath et al., 2020)	171.2 (±135.6)	353.8 (±273.4)	227.0 (±185.8)
GMS (Bian et al., 2017)	232.5 (±169.2)	443.4 (±286.7)	314.3 (±239.0)
OANet (Zhang et al., 2019)	130.0 (±149.3)	257.1 (±279.9)	174.6 (±208.3)
LOGO (Xia and Ma, 2022)	238.3 (±431.9)	479.8 (±380.6)	316.1 (±430.7)
CRC (Fan et al., 2021)	147.5 (±155.9)	275.2 (±278.8)	202.8 (±218.9)
PSC	**63.48 (±107.3)**	**125.6 (±198.1)**	**83.98 (±147.1)**

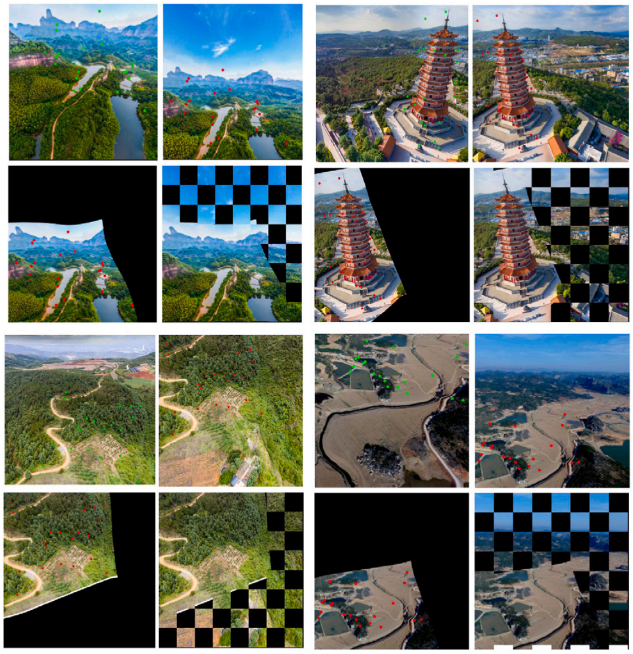

Fig. 7. Qualitative presentation of image registration results of our PSC. From top to bottom, left to right, there are four groups of images. In each group, the first row contains the source image and the target image, and the second row contains the warped target image and the stitched registration result, where green dots represent landmarks of the source image, and the red dots represent landmarks of the target image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\text{MEE} = \text{median}\left\{\sqrt{\left(r_i - \mathcal{F}\left(s_i\right)\right)^2}\right\}_{i=1}^{M}. \tag{29}$$

4.3.1. Qualitative results

The qualitative results of our PSC on image registration tasks are shown in Fig. 7, which has four groups of images. In each group, the first row contains the original source and target images, and the second row contains the warped target image and the registration result of two images. From the registration results, it can be seen that the warped target image is well stitched together with the source image, especially in the edge regions, which proves the effectiveness of our PSC in high-level vision tasks.

4.3.2. Quantitative comparison

The result statistics are presented in Table 7, where our PSC is compared with other nine feature matching methods. Clearly, due to the high-precision nature of our PSC, it shows a distinct advantage (lowest errors) on image registration task.

5. Conclusion and discussion

In response to the limitations of the current popular mismatch removal strategy: the segmented exploitation of geometric structure and descriptor similarity, this paper proposed an effective, robust, and general handcrafted feature matching method, named *Progressive Smoothness Consensus* (PSC). This method acts directly on two sets of feature points, both constructing an accurate feature matching relationship and guaranteeing an adequate number of matches, which improves the performance of computer vision tasks such as relative pose estimation and image registration.

Even though our method has shown significant advantages over commonly used or state-of-the-art methods in terms of the effectiveness of feature matching, generality to descriptors, and robustness to various images, it does not do quite well in computational time. Therefore, there is still room for further improvement in the efficiency of feature matching methods, which would be further investigated in the future.

CRediT authorship contribution statement

Yifan Xia: Methodology, Experiment, Writing – original draft. **Jie Jiang:** Methodology, Writing – review & editing. **Yifan Lu:** Methodology, Experiment. **Wei Liu:** Methodology, Writing – review & editing. **Jiayi Ma:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276192), and also by the Key Research and Development Program of Hubei Province, China (2020BAB113).

References

Alcantarilla, P.F., Bartoli, A., Davison, A.J., 2012. KAZE features. In: Proceedings of the European Conference on Computer Vision. pp. 214–227.

Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K., 2017. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5173–5182.

Barath, D., Matas, J., Noskova, J., 2019. MAGSAC: marginalizing sample consensus. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10197–10205.

Barath, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. Magsac++, a fast, reliable and accurate robust estimator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1304–1312.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. In: Proceedings of the European Conference on Computer Vision. pp. 404–417.

Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., Cheng, M.-M., 2017. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2828–2837.

Bian, J.-W., Wu, Y.-H., Zhao, J., Liu, Y., Zhang, L., Cheng, M.-M., Reid, I., 2019. An evaluation of feature matchers for fundamental matrix estimation. In: Proceedings of the British Machine Vision Conference. pp. 1–14.

Chen, H., Luo, Z., Zhang, J., Zhou, L., Bai, X., Hu, Z., Tai, C.-L., Quan, L., 2021. Learning to match features with seeded graph matching network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6301–6310.

Cho, M., Lee, K.M., 2012. Progressive graph matching: Making a move of graphs via probabilistic voting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 398–405.

Cho, M., Lee, J., Lee, K.M., 2010. Reweighted random walks for graph matching. In: Proceedings of the European Conference on Computer Vision. pp. 492–505.

Chum, O., Matas, J., 2008. Optimal randomized RANSAC. IEEE Trans. Pattern Anal. Mach. Intell. 30 (8), 1472–1482.

Chum, O., Matas, J., Kittler, J., 2003. Locally optimized RANSAC. In: Proceedings of the Joint Pattern Recognition Symposium. pp. 236–243.

Cordes, K., Rosenhahn, B., Ostermann, J., 2013. High-resolution feature evaluation benchmark. In: Proceedings of the International Conference on Computer Analysis of Images and Patterns. pp. 327–334.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–22.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 224–236.

Donato, G., Belongie, S., 2002. Approximate thin plate spline mappings. In: Proceedings of the European Conference on Computer Vision. pp. 21–31.

Fan, A., Jiang, X., Ma, Y., Mei, X., Ma, J., 2021. Smoothness-driven consensus based on compact representation for robust feature matching. IEEE Trans. Neural Netw. Learn. Syst..

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. Int. J. Robot. Res. 32 (11), 1231–1237.

Girosi, F., Jones, M., Poggio, T., 1995. Regularization theory and neural networks architectures. Neural Comput. 7 (2), 219–269.

Grebenkov, D.S., Nguyen, B.-T., 2013. Geometrical structure of Laplacian eigenfunctions. Siam Rebiew 55 (4), 601–667.

Han, J.-W., Kim, J.-H., Cheon, S.-H., Kim, J.-O., Ko, S.-J., 2010. A novel image interpolation method using the bilateral filter. IEEE Trans. Consum. Electron. 56 (1), 175–181.

Heinly, J., Dunn, E., Frahm, J.-M., 2012. Comparative evaluation of binary features. In: Proceedings of the European Conference on Computer Vision. pp. 759–773.

Jiang, S., Jiang, W., Guo, B., 2022a. Leveraging vocabulary tree for simultaneous match pair selection and guided feature matching of UAV images. ISPRS J. Photogramm. Remote Sens. 187, 273–293.

Jiang, X., Wang, Y., Fan, A., Ma, J., 2022b. Learning for mismatch removal via graph attention networks. ISPRS J. Photogramm. Remote Sens. 190, 181–195.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. Graph. 36 (4), 1–13.

Lee, S., Lim, J., Suh, I.H., 2020. Progressive feature matching: Incremental graph construction and optimization. IEEE Trans. Image Process. 29, 6992–7005.

Leordeanu, M., Hebert, M., 2005. A spectral technique for correspondence problems using pairwise constraints. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1482–1489.

Leutenegger, S., Chli, M., Siegwart, R.Y., 2011. BRISK: Binary robust invariant scalable keypoints. In: Proceedings of the International Conference on Computer Vision. pp. 2548–2555.

Liang, L., Zhao, W., Hao, X., Yang, Y., Yang, K., Liang, L., Yang, Q., 2020. Image registration using two-layer cascade reciprocal pipeline and context-aware dissimilarity measure. Neurocomputing 371, 1–14.

Lin, W.-Y.D., Cheng, M.-M., Lu, J., Yang, H., Do, M.N., Torr, P., 2014. Bilateral functions for global motion modeling. In: Proceedings of the European Conference on Computer Vision. pp. 341–356.

Lin, W.-Y., Liu, S., Jiang, N., Do, M., Tan, P., Lu, J., et al., 2016. Repmatch: Robust feature matching and pose for reconstructing modern cities. In: Proceedings of the European Conference on Computer Vision. pp. 562–579.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., 2021. Image matching from handcrafted to deep features: A survey. Int. J. Comput. Vis. 129 (1), 23–79.

Ma, J., Jiang, J., Zhou, H., Zhao, J., Guo, X., 2018. Guided locality preserving feature matching for remote sensing image registration. IEEE Trans. Geosci. Remote Sens. 56 (8), 4435–4447.

Ma, J., Li, Z., Zhang, K., Shao, Z., Xiao, G., 2022. Robust feature matching via neighborhood manifold representation consensus. ISPRS J. Photogramm. Remote Sens. 183, 196–209.

Ma, J., Zhao, J., Jiang, J., Zhou, H., Guo, X., 2019. Locality preserving matching. Int. J. Comput. Vis. 127 (5), 512–531.

Ma, J., Zhao, J., Tian, J., Bai, X., Tu, Z., 2013. Regularized vector field learning with sparse approximation for mismatch removal. Pattern Recognit. 46 (12), 3519–3532.

Ma, J., Zhao, J., Tian, J., Yuille, A.L., Tu, Z., 2014. Robust point matching via vector field consensus. IEEE Trans. Image Process. 23 (4), 1706–1721.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., Tian, J., 2015. Robust feature matching for remote sensing image registration via locally linear transforming. IEEE Trans. Geosci. Remote Sens. 53 (12), 6469–6481.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L., 2005. A comparison of affine region detectors. Int. J. Comput. Vis. 65 (1), 43–72.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. Adv. Neural Inf. Process. Syst. 30.

Mishkin, D., Radenovic, F., Matas, J., 2018. Repeatability is not enough: Learning affine regions via discriminability. In: Proceedings of the European Conference on Computer Vision. pp. 284–300.

Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.-M., 2012. USAC: A universal framework for random sample consensus. IEEE Trans. Pattern Anal. Mach. Intell. 35 (8), 2022–2038.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In: Proceedings of the International Conference on Computer Vision. pp. 2564–2571.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4938–4947.

Shi, Y., Cai, J.-X., Shavit, Y., Mu, T.-J., Feng, W., Zhang, K., 2022. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12517–12526.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of RGB-d SLAM systems. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems. pp. 573–580.

Sullivan, T.J., 2015. Introduction to Uncertainty Quantification. Vol. 63, Springer.

Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V., 2019. Sosnet: Second order similarity regularization for local descriptor learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11016–11025.

Torr, P.H., Zisserman, A., 2000. MLESAC: A new robust estimator with application to estimating image geometry. Comput. Vis. Image Underst. 78 (1), 138–156.

Vedaldi, A., Fulkerson, B., 2010. Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the ACM International Conference on Multimedia. pp. 1469–1472.

Wilson, K., Snavely, N., 2014. Robust global translations with 1dsfm. In: Proceedings of the European Conference on Computer Vision. pp. 61–75.

Xia, Y., Ma, J., 2022. Locality-guided global-preserving optimization for robust feature matching. IEEE Trans. Image Process. 31, 5093–5108.

Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., 2018. Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2666–2674.

Zaragoza, J., Chin, T.-J., Brown, M.S., Suter, D., 2013. As-projective-as-possible image stitching with moving DLT. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2339–2346.

Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H., 2019. Learning two-view correspondences and geometry using order-aware network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5845–5854.