IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA ANALYTICS FOR LARGE-SCALE COMPLEX DATA ENVIRONMENTS

Received September 13, 2017, accepted October 12, 2017, date of publication October 30, 2017, date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2768078

Robust Image Feature Matching via Progressive Sparse Spatial Consensus

YONG MA[1,2], JIAHAO WANG[1], HUIHUI XU[1], SHUAIBIN ZHANG[1], XIAOGUANG MEI[1], AND JIAYI MA[1,2,3]

[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 100081, China
[3]Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China

Corresponding author: Jiayi Ma (jyma2010@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773295 and Grant 61503288, in part by the Hubei Provincial Key Laboratory of Intelligent Robot under Grant HBIR 201702, in part by the Joint Advanced Research Foundation of Departments of Equipment and Education under Grant 6141A02022303, and in part by the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant 2016IRS15.

ABSTRACT In this paper, we propose an efficient algorithm, termed as progressive sparse spatial consensus, for mismatch removal from a set of putative feature correspondences involving large number of outliers. Our goal is to estimate the underlying spatial consensus between the feature correspondences and then remove mismatches accordingly. This is formulated as a maximum likelihood estimation problem, and solved by an iterative expectation-maximization algorithm. To handle large number of outliers, we introduce a progressive framework, which uses matching results on a small putative set with high inlier ratio to guide the matching on a large putative set. The spatial consensus is modeled by a non-parametric thin-plate spline kernel; this enables our method to handle image pairs with both rigid and non-rigid motions. Moreover, we also introduce a sparse approximation to accelerate the optimization, which can largely reduce the computational complexity without degenerating the accuracy. The quantitative results on various experimental data demonstrate that our method can achieve better matching accuracy and can generate more good matches compared to several state-of-the-art methods.

INDEX TERMS Feature matching, spatial consensus, sparse approximation, progressive, outlier.

I. INTRODUCTION

This paper focuses on the establishment of accurate feature correspondences between two images of the same scene, which is a fundamental problem in computer vision, image analysis and pattern recognition [1]–[3]. The features are often salient points with associated descriptors extracted by some detectors, such as Scale Invariant Feature Transform (SIFT) [4] or Shape Context [5], and the goal is to find point correspondences according to their positions and descriptors.

The matching problem is an ill-posed problem and typically regularized by first imposing a similarity constraint requiring that two points can be matched if they have similar feature descriptors, and then imposing a geometric constraint requiring that the correspondences should satisfy some global geometrical relationship [6]. In general, the similarity constraint is able to generate a putative correspondence set containing major correct matches, or inliers, which largely reduces the scale of possible matches. However, due to viewpoint changes, repeated patterns, as well as occlusions, the correspondences established by only local feature descriptors become unreliable. Thus the geometric constraint is further adopted to remove the false matches, or outliers. For example, fit a spatial transformation between two feature sets and discard those matches which do not obey the transformation. In this paper, we focus on mismatch removal from a set of putative matches.

To efficiently remove mismatches, one of the most widely used methods is the RANdom SAmple Consensus (RANSAC) [7]. It tries to get a minimum outlier-free subset to estimate a given parametric model by resampling. RANSAC needs to know the image transformation model in advance, which cannot work well if the spatial transformation is non-parametric, or non-rigid. To address this issue, recently some new non-parametric model-based methods are proposed, including Identifying Correspondence Function (ICF) [8], Graph Shift (GS) [9], Vector Field Consensus (VFC) [6], Coherent Spatial Relations (CSR) [10] and

Locally Linear Transforming (LLT) [11]. These methods usually interpolate the underlying image transformation with kernel functions using unsupervised learning, and are able to handle complex non-rigid deformation.

Although various robust estimators have been proposed to distinguish inliers from outliers, it is still a challenging task to customize a practical algorithm when dealing with real-world matching problems. On the one hand, the existing matching methods typically require that the putative set should not contain large proportion of outliers. To achieve this goal, Lowe [4] suggested using a distance ratio threshold to filter out mismatches. It calculates the ratio of the Euclidean distance of the closest neighbor and the second-closest neighbor, and preserves only those matches with ratios below a predefined threshold. Based on this strategy, Pele and Michael [12] replaced the Euclidean distance with earth movers' distance, which further enhanced the matching accuracy in the putative set. However, by using only descriptor similarity to suppress mismatches, these strategies at the same time inevitably discard a lot of correct matches [13], leading to the performance degradation in the subsequent applications such as image retrieval [14], visual homing [15], object recognition [16], etc. Therefore, it is desirable to ensure that the putative set covers the whole true matches; this requires the matching algorithm to be robust even in case of extremely high proportion of outliers. On the other hand, the existing matching methods often suffer from low efficiency due to their large computational complexities, which is problematic in handing large scale problems. For example, for methods based on non-parametric models, the number of model parameters increases with respect to the scale of the putative set; the total time complexity can then achieve $O(N^3)$ or even larger. Simultaneously, if the outlier percentage reaches up to 80% in the putative set, RANSAC will need more than 20 million times of sampling to generate a satisfying result [8]. Therefore, it is desirable to seek a fast implementation to reduce the computational complexity, especially for real-time tasks.

In view of the above problems, we propose an efficient algorithm for robust feature matching even in case of extremely large number of outliers. To this end, we first introduce a mixture model composed of a Gaussian distribution and a uniform distribution, respectively indicating the inliers and outliers, to estimate the underlying spatial relationship (or spatial consensus) between feature points. The mismatch removal problem then can be formulated as a maximum likelihood estimation problem and solved by an iterative Expectation-Maximization (EM) algorithm. To enable our method to address extremely large number of outliers, we also introduce a progressive matching strategy. The key idea is to use the matching results on a small putative set with high inlier ratio to guide the matching on a larger putative set with lower inlier ratio but covering more true matches. We model the spatial transformation in a reproducing kernel Hilbert space (RKHS) using the thin-plate spline (TPS) kernel [17], which is efficient to handle both rigid and non-rigid motions. In addition, to reduce the computational complexity of our proposed method, we introduce a sparse approximation based on the idea of the subset of regressors method [18]. This can significantly accelerate our method without sacrifice the matching accuracy. Experimental results on various image data demonstrate the superior performance of our method compared to several other state-of-the-art matching methods.

Our contribution in this paper includes the following two aspects. Firstly, we introduce a progressive matching strategy combined with a maximum likelihood spatial consensus estimation, which can not only find more feature correspondences, but also can successfully distinguish the inliers even in case of extremely large outlier ratio. Secondly, a sparse approximation is applied to the estimation of spatial consensus, which greatly reduces the computational complexity and hence enables our method to be applicable to large scale matching problems.

The rest of the paper is organized as follows. Section II describes some background material and related work. In Section III, we present the proposed matching method, including the maximum likelihood formulation, sparse approximation solution, and the progressive matching strategy. Section IV illustrates the experimental performance of our method with comparison to other state-of-the-art methods on real image data, followed by the concluding remarks in Section V.

II. RELATED WORK

Image registration has been widely used in many fields including computer vision [19], [20], pattern recognition [21]–[23], medical image analysis [24], [25], and remote sensing [26], [27]. Exhaustive reviews on the image registration methods can be found in the literature [28], [29]. The registration methods can be broadly classified into area-based and feature-based methods. The area-based methods usually contain three types of methods, such as correlation-like methods [30], Fourier methods [31] and mutual information methods [32]. These methods deal with the original image intensity values directly and hence are preferable in case of few prominent details. However, they typically suffer from illumination changes, image distortions, and heavy computational complexities. Alternatively, feature-based methods are more robust and potentially faster, if implemented in the right way. They work by first extracting salient features from the image pair and then establishing correspondences and estimating spatial transformation between them, which are further used to align the image pair together. In this procedure, the image registration reduces to a feature matching problem, where the goal is to determine the correspondences between two feature point sets. Next, we briefly overview some works that are most relevant to our approach.

A popular strategy for establish feature correspondences is a two-step matching strategy [6]. In the first step, a set of putative correspondences is constructed by pruning all possible feature correspondences based on a similarity constraint, which discards those correspondences with sufficient

dissimilar descriptors. In the second step, a certain robust estimator is adopted to detect and remove the false matches in the putative set according to some geometrical constraint such as homography and epipolar geometry [10]. To address the mismatch removal in the second step, a plenty of methods have been proposed over the last decades, which can be roughly divided into four categories, including statistical regression methods, resampling methods, non-parametric interpolation methods, and graph matching methods.

Early in the statistics literature, it has been shown that maximum likelihood estimation of model parameters using L_1 norm is more robust and capable of resisting a larger number of outliers compared with quadratic L_2 norm [33], [34]. Based on adaptive boosting learning, Liu *et al.* [35] proposed a regression method for 3D rigid registration. In addition, a guided matching scheme is introduced based on statistical optical flow which has achieved promising results [36]. The most popular resampling method is RANSAC [7] as well as its variants including MLESAC [37] and PROSAC [38]. These methods aim to obtain a smallest possible outlier-free subset to estimate a given parametric model through resampling. The statistical regression and resampling methods need to define a parametric model for the image transformation in advance, which cannot work well in case of non-rigid transformation. Moreover, if the outlier ratio in the putative set is large, they also tend to severely degrade [8].

To address these issues, several non-parametric model-based methods [6], [8], [11] have recently been introduced, which commonly interpolate a non-parametric function by applying a slow-and-smooth prior. The ICF method learns a correspondence function pair mapping feature points across two images by using support vector regression, and then removes the false matches according to the estimated correspondence functions [8]. The VFC method interpolates a global smoothness motion field associated with the image pairs by using regularized kernel methods which can resist quite a large number of outliers [6]. In contrast, the LLT method recovers a smoothness image transformation by preserving local neighborhood structure of feature points. These methods often have computational complexities larger than $O(N^3)$, limiting their uses in real-time applications including object tracking, visual odometry, SLAM, etc. Graph matching provides another strategy for solving the matching problem [9], [13]; it provides considerable flexibility for building models and delivers robust matching and recognition. Graph matching problems usually incorporate pair-wise constraints, and they can be cast as a quadratic assignment problem. These methods however suffer from similar drawbacks of their NP-hard nature.

Except for mismatch removal, some efforts focus on generating better putative correspondences have also been made. For example, Guo and Cao [39] proposed a triangle constraint as a preprocess for pruning false matches, while Hu *et al.* [40] introduced to select an appropriate descriptor rather than a global descriptor for each feature point during matching. Ma *et al.* [41] provided an effective way called Locality Preserving Matching (LPM) to filter out false matches by preserving the local neighborhood structures of those potential true matches. In addition, a cascade scheme has been used to alleviate the loss of true correspondences [13]. In this work, we propose an effective method for boosting true matches while avoiding false matches. It is general and has low complexity which can be applied to handle various matching tasks.

III. METHOD

This section describes the proposed feature matching method. We start by briefly introducing the general regularization technique, and then present the formulation of our method and derive its EM solution together with a sparse approximation by using regularized kernel method. We subsequently lay out our progressive matching strategy for handling extreme outliers, followed by the analysis of computational complexity and some implementation details. Finally, we discuss the relation to existing work. Throughout the paper we use the following notations:

- $(\mathbf{u}_n, \mathbf{v}_n)$ - a feature correspondence,
- $\mathbf{f}$ - the spatial transformation,
- $\boldsymbol{\theta}$ - an unknown parameter set,
- $\mathbf{A}_{3\times3}$ - the TPS affine matrix,
- $\mathbf{W}_{N\times3}$ - the TPS non-affine coefficient matrix,
- K - the TPS kernel
- $\mathbf{P} = \mathrm{diag}(p_1, \cdots, p_N)$ - the match posterior probabilities.

A. TIKHONOV REGULARIZATION

Given a set of input-output pairs $S = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, the primary goal with learning is to interpolate a function $\mathbf{f}$ from them that can give good predictions for new inputs rather than precisely fit the given samples. Typically, we have limited number of samples of data in a much higher-dimensional space, and hence we cannot expect to obtain satisfying performance by blindly choosing a model. For example, a highly-parameterized model will probably overfit the data, and a too simple model may not adequately describe the data. Regularization in this context provides us with one way to strike the appropriate balance in creating the model. The goal of regularization is to solve the empirical error minimization problem by controlling the complexity of the function space, for example, by introducing a penalty term into the empirical error

$$\mathrm{ERR}(\mathbf{f}) + \lambda \mathrm{PEN}(\mathbf{f}), \tag{1}$$

where the first term is the empirical error measuring the fitting degree of the function and the samples, the second term is a penalty item which requires the function to be not too complex, and λ is used as a regularization parameter to make a trade-off between the two items. By using the L_2 loss on the data fitting and L_2 functional norm on the model complexity, the Tikhonov regularization minimizes the following

regularized risk functional [42]:

$$\mathcal{E}(f) = \min\{\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2 + \lambda\|\mathbf{f}\|^2\}. \quad (2)$$

By choosing a specifical kernel in a reproducing kernel Hilbert space (RKHS) to define $\mathbf{f}$, the minimization problem in Eq. (2) is equivalent to solving a linear system [42].

B. A MAXIMUM LIKELIHOOD FORMULATION

The Tikhonov regularization in Eq. (2) does not consider the outlier issue in the given samples. In other words, if the data samples contain outliers, the fitting result will be badly biased. Suppose we obtained a set of putative feature correspondences $S = \{(\mathbf{u}_n, \mathbf{v}_n)\}_{n=1}^{N}$ which may contaminated by some unknown mismatches, where $\mathbf{u}_n$ and $\mathbf{v}_n$ are the spatial positions of two feature points in the original two images. We use homogeneous coordinates, e.g., $\mathbf{u} = (\mathbf{u}^x, \mathbf{u}^y, 1)$, and the underlying spatial transformation between the feature correspondence is denoted as $\mathbf{f}$, e.g., $\mathbf{v}_n = \mathbf{f}(\mathbf{u}_n)$ if $(\mathbf{u}_n, \mathbf{v}_n)$ is an inlier. Due to the existence of outliers, it is desirable to have a robust estimation of $\mathbf{f}$. To this end, we make the assumption that the noise on inliers is Gaussian on each component with a zero mean and a uniform standard deviation σ, and the outlier distribution is uniform $1/a$, where a is the area of input image. We assume the uniform distribution on the outliers is based on the observation that the false matches can occur anywhere in an image pair, and this assumption has also been widely used in the matching problem [6], [37], [43]. Let γ be the percentage of inliers which we do not know in advance. Thus, the likelihood is a mixture model

$$\begin{aligned} p(S|\boldsymbol{\theta}) &= \prod_{n=1}^{N} \sum_{z_n} p(\mathbf{u}_n, \mathbf{v}_n, z_n|\boldsymbol{\theta}) \\ &= \prod_{n=1}^{N} \Big(\frac{\gamma}{2\pi\sigma^2} e^{-\frac{\|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2}{2\sigma^2}} + \frac{1-\gamma}{a} \Big), \end{aligned} \quad (3)$$

where $\boldsymbol{\theta} = \{\mathbf{f}, \sigma^2, \lambda\}$ include a set of unknown parameters to be solved, $z_n \in \{0, 1\}$ is a latent variable associated the n-th correspondence with $z_n = 1$ indicating a Gaussian distribution and $z_n = 0$ denoting a uniform distribution.

Generally, the true parameter set $\boldsymbol{\theta}$ maximizes the likelihood. Next, we seek a maximum likelihood estimation of $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(S|\boldsymbol{\theta})$, which is equivalent to solving the minimal energy

$$E(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \ln \sum_{z_n} p(\mathbf{u}_n, \mathbf{v}_n, z_n|\boldsymbol{\theta}). \quad (4)$$

Thus we can obtain the spatial transformation $\mathbf{f}$ from the optimal solution $\boldsymbol{\theta}^*$.

C. TRANSFORMATION MODELLING WITH SPARSE APPROXIMATION

Before we solve the optimization problem in Eq. (4), we first consider the problem of transformation modelling. For image pairs of statical scenes, the spatial transformations can be characterized by epipolar geometry, i.e., $\mathbf{vFu}^T = 0$ with $\mathbf{F}$ being a 3×3 fundamental matrix with 8 degrees of freedom. Furthermore, if the image pairs are of planar scenes or taken by camera in a fixed position during acquisition, then the spatial transformations will degrade to homography or even affine model, i.e., $\mathbf{v} = \mathbf{uH}$ with $\mathbf{H}$ being the corresponding 3×3 coefficient matrix. However, these parametric model cannot work well if the image pairs involve dynamical scenes or non-rigid motions. Moreover, in many practical tasks such as image retrieval and object matching and tracking, the transformation models are often unknown in advance, which further limits the application of parametric model.

In this paper, to make our method more general, we consider the non-parametric model and require the transformation to lie within an RKHS. More specifically, the TPS kernel is chosen to parameterize the transformation. The TPS is a general purpose spline tool which produces a smooth functional mapping for supervised learning [10]. Specifically, it also has been applied to the dimensionality reduction problem and shown promising results [44]. TPS has no free parameters that need manual tuning and also has a closed-form solution which can be decomposed into a global linear affine motion and a local non-affine warping component controlled by coefficients $\mathbf{A}$ and $\mathbf{W}$, respectively:

$$\mathbf{f}(\mathbf{u}) = \mathbf{u} \cdot \mathbf{A} + \widetilde{K}(\mathbf{u}) \cdot \mathbf{W}, \quad (5)$$

where $\mathbf{A}$ is a matrix of size 3×3, $\mathbf{W}$ is a matrix of size $N \times 3$, and $\widetilde{K}(\mathbf{u})$ is a $1 \times N$ vector defined by the TPS kernel, i.e., $K(r) = r^2 \log r$, with each entry $\widetilde{K}_n(\mathbf{u}) = K(|\mathbf{u} - \mathbf{u}_n|)$.

Feature matching methods based on non-parametric model often lead to computational complexity as least $O(N^3)$, as the number of parameters in a non-parametric model is proportional to the number of putative correspondences (see the coefficient matrix $\mathbf{W}$ for example), i.e., their models involve $O(N)$ numbers of parameters to be determined [18]. This will be problematic if the putative set contains a large number of correspondences. In fact, the spatial transformation between two images should not depend on the number of putative correspondences, as it would not change with respect to the change of putative correspondences. Therefore, if the scale of the putative correspondences is large, the standard non-parametric model will contain a lot of redundant parameters. To address this issue, we adopt a sparse approximation of the non-parametric model, and choose a fix set of M ($M \ll N$) bases $\{\widetilde{\mathbf{u}}_1, \widetilde{\mathbf{u}}_2, \cdots, \widetilde{\mathbf{u}}_M\}$ to construct the transformation:

$$\mathbf{f}(\mathbf{u}) = \mathbf{u} \cdot \mathbf{A} + \widetilde{K}^s(\mathbf{u}) \cdot \mathbf{W}, \quad (6)$$

where $\widetilde{K}(\mathbf{u})$ is a $1 \times M$ vector with each entry $\widetilde{K}_m(\mathbf{u}) = K(|\mathbf{u} - \widetilde{\mathbf{u}}_m|)$, and the coefficient matrix $\mathbf{W}$ in this case is of size $M \times 3$. The choice of bases $\{\widetilde{\mathbf{u}}_1, \widetilde{\mathbf{u}}_2, \cdots, \widetilde{\mathbf{u}}_M\}$ could be arbitrary, which could be an arbitrary subset of the original feature points $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}$. This follows [45] who found that this sparse approximation works well and simply selecting a random subset of the feature points performs no worse than those more sophisticated and time-consuming methods.

D. AN EM SOLUTION

Next, we consider the solution of the minimal energy problem in Eq. (4). There are several techniques can be used to solve this problem, where the well-known EM algorithm provides a natural framework for achieving the goal. The E-step basically estimates the responsibility indicating to what degree a sample belonging to inlier under the given spatial transformation $\mathbf{f}$, while the M-step updates $\mathbf{f}$ based on the current estimate of the responsibility. Following standard approach and omitting the terms independent of $\boldsymbol{\theta}$, we obtain the complete-data log likelihood as follows:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} p_n \|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2 - \ln\sigma^2 \sum_{n=1}^{N} p_n + \ln\gamma \sum_{n=1}^{N} p_n + \ln(1-\gamma)\sum_{n=1}^{N}(1-p_n), \quad (7)$$

where $p_n = P(z_n = 1|\mathbf{u}_n, \mathbf{v}_n, \boldsymbol{\theta}^{\text{old}})$ is a posterior probability indicating to the degree the match $(\mathbf{u}_n, \mathbf{v}_n)$ belonging to an inlier under the current estimated spatial transformation $\mathbf{f}$.

E-Step: Denote $\mathbf{P} = \text{diag}(p_1, \cdots, p_N)$ a diagonal matrix, where the responsibility p_n can be computed by applying Bayes rule:

$$p_n = \frac{\gamma e^{-\frac{\|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2}{2\sigma^2}}}{\gamma e^{\frac{-\|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2}{2\sigma^2}} + \frac{2\pi\sigma^2(1-\gamma)}{a}}. \quad (8)$$

M-Step: We determine the revised parameter estimate $\boldsymbol{\theta}^{\text{new}}$ as follows: $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$. Taking the derivative of $\mathcal{Q}(\boldsymbol{\theta})$ with respect to the variance σ^2 and the mixing coefficient γ and setting them to zero, we obtain

$$\sigma^2 = \frac{\sum_{n=1}^{N} p_n \|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2}{2 \cdot \text{tr}(\mathbf{P})}, \quad (9)$$

$$\gamma = \text{tr}(\mathbf{P})/N, \quad (10)$$

where $\text{tr}(\cdot)$ denotes the matrix trace.

To complete the M-step, we have to estimate the spatial transformation $\mathbf{f}$, which is relatively complex and we leave it in the next section. Once the EM iteration converges, we get the transformation $\mathbf{f}$. The mismatches can then be removed by checking whether they are consistent with $\mathbf{f}$. With a predefined threshold τ, the inlier set $\mathcal{I}$ is determined by the following formula:

$$\mathcal{I} = \{n | p_n > \tau, n = 1, \cdots, N\}. \quad (11)$$

E. TRANSFORMATION ESTIMATION

According to complete-data log likelihood in Eq. (7), the spatial transformation $\mathbf{f}$ is estimated by minimizing a weighted empirical error function as follows:

$$\mathcal{Q}(\mathbf{f}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} p_n \|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2. \quad (12)$$

This is an ill-posed problem, as there are infinite solutions for the transformation $\mathbf{f}$. To make the problem well-posed, we consider the regularization technique, as mentioned in Section III-A. The TPS regularization, e.g. the L_2 functional norm in Eq. (2), is defined as [10]:

$$\|\mathbf{f}\|^2 = \text{tr}(\mathbf{W}^T \mathbf{K}^s \mathbf{W}), \quad (13)$$

where $\mathbf{K}^s \in \mathbb{R}^{M\times M}$ is the so-called Gram matrix with $\mathbf{K}^s_{ij} = K(|\widetilde{\mathbf{u}}_i - \widetilde{\mathbf{u}}_j|)$. Therefore, the weighted empirical error function in Eq. (12) becomes a weighted regularized risk functional

$$\mathcal{Q}(\mathbf{f}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} p_n \|\mathbf{v}_n - \mathbf{f}(\mathbf{u}_n)\|^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T \mathbf{K}^s \mathbf{W}). \quad (14)$$

We use a matrix form to rewrite Eq. (14) and obtain

$$\begin{aligned}\mathcal{Q}(\mathbf{A}, \mathbf{W}) &= \frac{1}{2\sigma^2}\|\mathbf{P}^{1/2}(\mathbf{V} - \mathbf{U}\mathbf{A} - \mathbf{K}^s_{\mathbf{u}}\mathbf{W})\|_F^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}) \\ &= \frac{1}{2\sigma^2}\|\widetilde{\mathbf{V}} - \widetilde{\mathbf{U}}\mathbf{A} - \mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\|_F^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}), \end{aligned} \quad (15)$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N)^T$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N)^T$ are matrices of both size $N \times 3$, $\widetilde{\mathbf{U}} = \mathbf{P}^{1/2}\mathbf{U}$, $\widetilde{\mathbf{V}} = \mathbf{P}^{1/2}\mathbf{V}$, $\mathbf{K}^s_{\mathbf{u}} \in \mathbb{R}^{N\times M}$ with $\mathbf{K}^s_{\mathbf{u}ij} = K(|\mathbf{u}_i - \widetilde{\mathbf{u}}_j|)$, and $\|\cdot\|_F$ denotes the Frobenius norm.

To solve the TPS parameter pair $\mathbf{A}$ and $\mathbf{W}$, we use a QR decomposition [46]:

$$\widetilde{\mathbf{U}} = [\mathbf{Q}_1, \mathbf{Q}_2]\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}, \quad (16)$$

where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are orthonormal matrices of sizes respectively $N \times 3$ and $N \times (N-3)$, and $\mathbf{R}$ is an upper triangular matrix of size 3×3. With the QR decomposition in place, Eq. (15) becomes

$$\begin{aligned}\mathcal{Q}(\mathbf{A}, \mathbf{W}) &= \frac{1}{2\sigma^2}\|\widetilde{\mathbf{V}} - [\mathbf{Q}_1, \mathbf{Q}_2]\begin{bmatrix}\mathbf{R}\\ \mathbf{0}\end{bmatrix}\mathbf{A} - \mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\|^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}) \\ &= \frac{1}{2\sigma^2}\left\|\begin{bmatrix}\mathbf{Q}_1^T\widetilde{\mathbf{V}}\\ \mathbf{Q}_2^T\widetilde{\mathbf{V}}\end{bmatrix} - \begin{bmatrix}\mathbf{R}\mathbf{A}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{Q}_1^T\mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\\ \mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\end{bmatrix}\right\|^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}) \\ &= \frac{1}{2\sigma^2}\|\mathbf{Q}_2^T\widetilde{\mathbf{V}} - \mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\|^2 + \frac{1}{2\sigma^2}\|\mathbf{Q}_1^T\widetilde{\mathbf{V}} - \mathbf{R}\mathbf{A} - \mathbf{Q}_1^T\mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\|^2 + \frac{\lambda}{2}\text{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}). \end{aligned} \quad (17)$$

Since each term in the last equation is non-negative and only the second term involves the parameter $\mathbf{A}$, to minimize this equation, the second term should be required to be zero at the optimal solution:

$$\|\mathbf{Q}_1^T\widetilde{\mathbf{V}} - \mathbf{R}\mathbf{A} - \mathbf{Q}_1^T\mathbf{P}^{1/2}\mathbf{K}^s_{\mathbf{u}}\mathbf{W}\|^2 = 0, \quad (18)$$

Algorithm 1 Mismatch Removal by SSC

Input: Putative set $\{(\mathbf{u}_n, \mathbf{v}_n)\}_{n=1}^{N}$, parameters λ, τ, M
Output: Inlier set $\mathcal{I}$

Initialization: $0 < \gamma < 1$, $\mathbf{W} = \mathbf{0}$, $\mathbf{A} = \mathbf{0}$;
Compute a according to the area of the given image;
Compute Gram matrix $\mathbf{K}^s$ and $\mathbf{K}_{\mathbf{u}}^s$ using the TPS kernel;
repeat
  E-step:
    Update the responsibility p_n by equation (8);
  M-step:
    Update transformation $\mathbf{f}$ by using equations (22) and (19);
    Update σ^2 and γ by equations (9) and (10);
until *some stopping criterion is satisfied*;
The inlier set is determined by equation (11).

Therefore, the optimal solution of $\mathbf{A}$ should be as follows:

$$\mathbf{A} = \mathbf{R}^{-1}\mathbf{Q}_1^T(\widetilde{V} - \mathbf{P}^{1/2}\mathbf{K}_{\mathbf{u}}^s\mathbf{W}). \tag{19}$$

We submit the solution of $\mathbf{A}$ into Eq. (17) and obtain

$$\mathcal{Q}(\mathbf{W}) = \frac{1}{2\sigma^2}\|\mathbf{Q}_2^T\widetilde{\mathbf{V}} - \mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}_{\mathbf{u}}^s\mathbf{W}\|^2 + \frac{\lambda}{2}\mathrm{tr}(\mathbf{W}^T\mathbf{K}^s\mathbf{W}). \tag{20}$$

We take the derivative of $\mathcal{Q}(\mathbf{W})$ with respect to the variable $\mathbf{W}$ and set it to zero, and obtain

$$(\mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}_{\mathbf{u}}^s)^T(\mathbf{Q}_2^T\widetilde{\mathbf{V}} - \mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}_{\mathbf{u}}^s\mathbf{W}) - \lambda\sigma^2\mathbf{K}^s\mathbf{W} = \mathbf{0}. \tag{21}$$

By denoting $\mathbf{S} = \mathbf{Q}_2^T\mathbf{P}^{1/2}\mathbf{K}_{\mathbf{u}}^s$, we obtian the optimal solution of $\mathbf{W}$

$$\mathbf{W} = (\mathbf{S}^T\mathbf{S} + \lambda\sigma^2\mathbf{K}^s + \epsilon\mathbf{I})^{-1}\mathbf{S}^T\mathbf{Q}_2^T\widetilde{\mathbf{V}}, \tag{22}$$

where $\epsilon\mathbf{I}$ is used for numerical stability.

Until now, we have solved all the parameters in the M-step. As the spatial transformation is computed only using those underlying inliers with sparse approximation, we call this strategy Sparse Spatial Consensus (SSC) and summarize it in Algorithm 1.

F. THE PROGRESSIVE MATCHING STRATEGY

Note that for most feature matching methods the proportion of outliers in the putative set in general should not be too high, and hence some sophisticated strategies [4], [12], [39] are used to filter out mismatches during the construction of putative correspondences. However, this process will also lead to discarding those unstable correct matches, and sometimes these discarded correct matches dominate the whole true matches which will degrade the subsequential applications, and hence a putative set covers more true matches is desirable. To solve this dilemma, we introduce a progressive matching strategy in this section, as shown in Fig. 1. First, we construct a putative set $S_0 = \{(\mathbf{u}_i, \mathbf{v}_i)\}_{i=1}^{N_0}$ by using a small distance ratio threshold t_0 of SIFT matches[1] [4], as shown in Fig. 1a. In this putative set, the inlier percentage is typically high, and hence using our SSC proposed in Algorithm 1 works well, as shown in Fig. 1b. Then we construct a larger putative set $S_1 = \{(\mathbf{u}_i, \mathbf{v}_i)\}_{i=1}^{N_1}$ using a larger distance ratio threshold t_1, as shown in Fig. 1c. Clearly, S_1 contains S_0 and it is much larger than S_0 with much more mismatches and expected to cover more true correspondences. To enable our SSC to work well on S_1, we introduce a strategy that using the result on S_0 to guide the matching on S_1, which can generate the matching results in Fig. 1d.

Algorithm 2 Progressive Sparse Spatial Consensus

Input: Putative sets S_0, S_1, parameters λ, τ, M
Output: Inlier set $\mathcal{I}_1$

Perform SSC on S_0 using Algorithm 1 and obtain $\mathcal{I}_0$;
Using $\mathcal{I}_0$ to initialize $\mathbf{P}$ with Eq. (23);
Perform SSC on S_1 using Algorithm 1 and obtain $\mathcal{I}_1$.

The major reason why our SSC cannot work well on S_1 is because that the EM iteration is easy to get trapped into local extrema. However, if we give a good initialization to the EM iteration, then it is definitely possible to reach a satisfying solution. To this end, we use the matching result on S_0 to initialize the EM iteration of SSC on S_1. For example, after we obtain the matching result on S_0, the transformation $\mathbf{f}$ on S_0 could be used to initialize the transformation $\mathbf{f}$ on S_1, as the true transformation should not change in case of different putative sets. This is equivalent to using the responsibility p_i of correspondence $(\mathbf{u}_i, \mathbf{v}_i)$ on S_0 to initialize the responsibility on S_1. More specifically,

$$p_i = \begin{cases} 1, & \text{if } i|_{(\mathbf{u}_i,\mathbf{v}_i)} \in \mathcal{I}_0 \\ \epsilon, & \text{otherwise,} \end{cases} \tag{23}$$

where ϵ is a small number used for numerical stability, and $\mathcal{I}_0$ is obtained according to Eq. (11). By using the preserved correspondences in S_0, we could recover the transformation $\mathbf{f}$ from S_1 in the first EM iteration. In this way, the EM iteration is able to avoid many of the local extrema inherent in the SSC formulation and obtain a good estimate very quickly.

This process could be performed progressively to boost more true correspondences. For example, construct an even larger putative set S_2 with an even larger distance ratio threshold t_2, and then use the matching result on S_1 to guide the matching on S_2. In our evaluation, we found two iterations can already achieve satisfying results, for example, we construct S_0 using a small t_0 and directly construct S_1 at $t_1 = 1$. As this progressive strategy is base on the SSC algorithm, we call it Progressive Sparse Spatial Consensus (PSSC) and summarize it in Algorithm 2.

[1]The value of distance ratio threshold t ranges from 0 to 1, where smaller t indicates larger inlier ratio in the putative set with less true matches, and $t = 1$ equals to the nearest neighbor matching strategy.

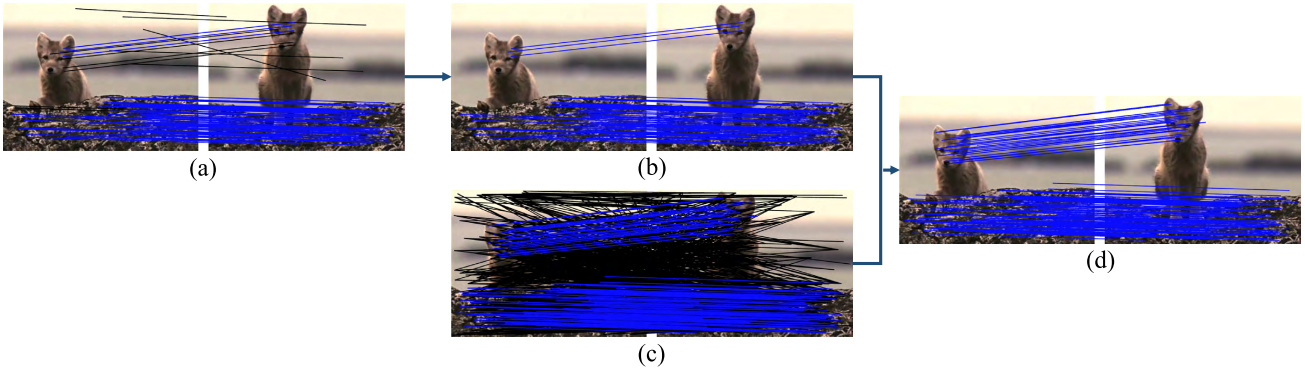


FIGURE 1. Schematic illustration of our progressive matching strategy. (a) A small putative set with high inlier ratio. (b) The matching result of our SSC on (a). (c) A small putative set with low inlier ratio but covers more true matches. (d) The matching result of our PSSC on (c) by using (b) for guided matching. The blue lines denote true matches and the black lines denote false matches.

G. COMPUTATIONAL COMPLEXITY

To compute the transformation $\mathbf{f}$, the most time-consuming step is the calculation of $\mathbf{W}$ in Eq. (22), which has time complexity $O(MN^2)$ due to the matrix inversion and the matrix multiplication operations. Besides, for a matrix of size $M \times N$, its time complexity of QR decomposition is about $2MN^2$, and hence the time complexity of Eq. (16) is about $O(N)$. Therefore, the total time complexity for each EM iteration of our method is about $O(MN^2)$. The space complexity of our method scales like $O(MN)$ due to the memory requirements for storing the matrix $\mathbf{K}_{\mathbf{u}}^{s}$. Since M is a constant and $M \ll N$, the time and space complexities can be written as $O(N^2)$ and $O(N)$, respectively. Without using the sparse approximation, the time and space complexities will increase to $O(N^3)$ and $O(N^2)$, respectively [10].

H. IMPLEMENTATION DETAILS

There are mainly three parameters in our algorithm: the regularization parameter λ, inlier threshold τ and the bases number M, where λ controls the trade-off between the closeness to the data and the smoothness of the transformation, τ is used to decide the correctness of a correspondence, and M is the number of bases used for sparse approximation to the TPS kernel. In our experimental evaluations, we found that our method is not very sensitive to these three parameters. Throughout this paper, we set $\lambda = 500$, $\tau = 0.5$ and $M = 30$.

In addition, the matching performance typically depends on the coordinate system in which the feature points are expressed. To alleviate the influence, we use data normalization, where a linear re-scaling of the matches is performed so that the two sets of points both have zero mean and unit variance. In addition, the constant a in Eq. (3) is set as the normalized area of input image after data normalization.

I. RELATION TO EXISTING METHOD

Our PSSC is related to the VFC algorithm [6]. On the one hand, both the two algorithms use the maximum likelihood spatial consensus estimation to formulate the matching problem, and the EM approach is adopted for optimization. On the other hand, our PSSC is different from VFC. We use the TPS kernel to parameterize the transformation model rather than the GRBF kernel in VFC. This can decompose the spatial transformation into explicit linear and nonlinear components, and the corresponding bending energy possesses a specific physical explanation, which is benefit to non-rigid matching. In addition, we generalize the formulation under a progressive matching framework to boost the number of true matches, and a sparse approximation is also applied to the TPS kernel to greatly reduce the computational complexity.

IV. EXPERIMENTAL RESULTS

In this section, we test the feature matching performance of our PSSC and compare with other six feature matching methods such as RANSAC [7], ICF [8], GS [9], VFC [6], CSR [10] and LPM [41]. We implement ICF and LPM and tune all their parameters accordingly to find optimal settings. For RANSAC, GS, VFC and CSR, we implement them based on the publicly available codes. For the six comparison methods, the putative correspondences are all established by using the nearest neighbor matching strategy. The experiments are performed on a desktop with 3.5 GHz Intel Core CPU, 64 GB memory and Matlab code. Throughout all the experiments, six algorithms' parameters are all fixed.

A. EXPERIMENTAL SETUP

We use the open source VLFeat toolbox [48] to extract the SIFT features, and then construct the putative correspondences according to the distance ratio threshold. Note that our method does not depend on any specific feature, some other features such as SURF [49] and ORB [50] can also be used to construct putative correspondences. The matching performance is characterized by precision and number of preserved correct matches, where the precision is defined as the ratio between the number of preserved correct matches and the number of whole preserved matches.

The match correctness for establish the ground truth is determined as follows. On the one hand, for image pairs related by parametric models such as homography, the ground

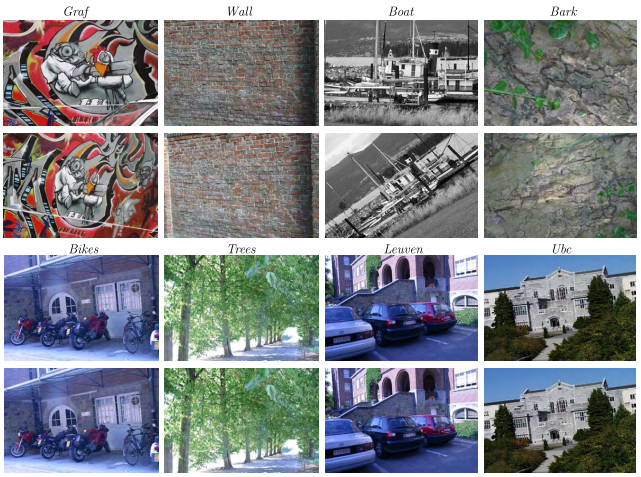


FIGURE 2. Some examples in the dataset of VGG [47].

TABLE 1. Performance comparison under different values of M.

M	10	20	30	40	50	N
Precision (%)	92.50	94.25	95.36	95.50	95.52	95.48
# Corr. Match	693.5	696.3	697.3	697.3	696.5	700.8
Time (s)	0.742	0.849	0.819	0.851	1.105	6.785

TABLE 2. Performance comparison under different values of t_1.

t_1	0.5	0.6	0.7	0.8	0.9	1.0
Inlier Ratio (%)	91.66	88.05	81.26	68.89	51.23	33.18
# Inlier	530.6	570.5	606.9	641.9	674.4	708.4
Precision (%)	94.31	94.02	94.58	96.32	95.95	95.36
# Corr. Match	519.6	558.5	591.2	625.1	658.0	697.3
Time (s)	0.094	0.100	0.233	0.265	0.518	0.819

truth model parameters can be obtained, and then we use an overlap error ϵ_S to determine the match correctness. Specifically, we first reduce the SIFT feature scales to one third of the original scales, and then a match is regarded as correct match if $\epsilon_S > 0$. This follows [6], [10] who found that this strategy is consistent with human perceptions. On the other hand, for image pairs related by non-rigid transformations, the ground truth model parameters usually cannot be obtained, and we determine the match correctness by manual checking. Although in this case the judgment of correct or false match seems arbitrary, to ensure objectivity we make the benchmark before conducting experiments.

B. RESULTS ON IMAGE PAIRS RELATED BY HOMOGRAPHY

We test our PSSC on the dataset of Visual Geometry Group (VGG) in the University of Oxford [47], which contains image transformations involving viewpoint change, scale change, rotation, image blur, JPEG compression, and illumination change. Some examples are shown in Fig. 2. The image pairs in this dataset are either of planar scenes or taken by camera in a fixed position during acquisition, and hence they always obey the homography. We use all the 40 image pairs for evaluation, where the ground truth homographies are supplied by the dataset. The reasons why we choose this dataset lie in two-fold: (i) it offers ground-truth for quantitative evaluation; (ii) it contains challenging situations in matching such as view point change, rotation, illumination variation, and so on. In the following, we first test the influences of different parameter settings, and then report the quantitative comparisons of different methods on the dataset.

We first investigate the influence of the choice of M, which is the number of basis functions used for sparse approximation. Five values of M including 10, 20, 30, 40 and 50 are chosen for test. In addition, we also test our algorithm without using the sparse approximation, i.e., $M = N$. The average precisions, numbers of preserved correct matches and run times on the whole dataset are summarized in Table 1. From the results, we see that by using or not using the sparse approximation, the average precisions and numbers of preserved correct matches are similar, but the average run times by using sparse approximation are much less. This demonstrates that the sparse approximation can largely speedup the matching procedure without much performance sacrifice. Moreover, $M = 30$ achieves the best trade-off between the accuracy and efficiency. Note that the average run time at $M = 30$ is even less than that at $M = 20$; this is due to that the EM iteration converges much faster at $M = 30$.

We then investigate the influence of the choice of different bases for sparse approximation. Except for simply selecting a random subset, we consider three other strategies: (i) using the $\sqrt{M} \times \sqrt{M}$ uniform grid points distributed evenly throughout the whole image; (ii) finding the M clustering centers of the feature points $\{\mathbf{u}_n\}_{n=1}^N$ by using a clustering algorithm such as k-means clustering; (iii) picking M bases sophisticatedly that minimize the residuals via sparse greedy matrix approximation [51]. By using these three strategies to select bases, we obtain the average correct match number and precision pairs (697.7, 95.38%), (696.1, 95.36%) and (697.5, 95.37%), respectively. We see that there is no significant difference among the different strategies. However, in the interests of computational efficiency, we implement the sparse approximation simply selecting random bases.

We subsequently investigate the influence of the choice of t_1, which is the distance ratio threshold of S_1, i.e., the larger putative set. Six values of t_1 including 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 are chosen for test. The average inlier ratios and inlier numbers in the putative sets are reported in the first three rows in Table 2. We see that as t_1 increases, the inlier ratio gets smaller while the inlier number becomes larger. The matching results are provided in the last three rows in Table 2, including the average precisions, numbers of preserved correct matches, as well as the average run times on the whole dataset. We see that the matching precision of our PSSC does not decrease with the decrease of inlier ratio. At $t_1 = 1.0$ which equals to the nearest neighbor

TABLE 3. Performance comparison under different values of λ.

λ	1	10	100	500	1000	5000
Precision (%)	94.02	94.61	94.94	95.36	95.72	95.73
# Corr. Match	697.7	698.5	697.7	697.3	696.6	695.8

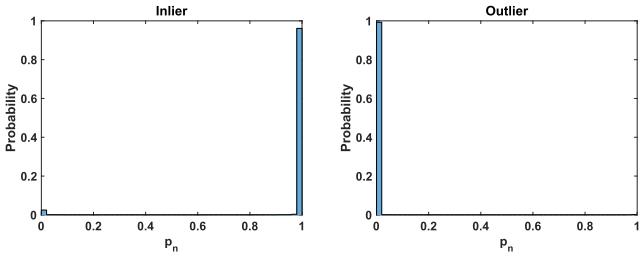


FIGURE 3. The probability distribution of p_n in Eq. (8) of each putative correspondence on the whole dataset of VGG [47].

matching strategy, the preserved correct matches achieve the largest number. In fact, we have also tried to construct an even larger S_1, for example, for each feature point in one set, we seek its K-nearest neighbors in the other set and construct K putative correspondences. We found that this procedure can only slightly increase the true match number but significantly increase the run time. Therefore, we just use the nearest neighbor matching strategy to construct S_1 in practice. Simultaneously, for the comparison methods in our experiments, we also construct the putative correspondences using the nearest neighbor matching strategy.

We next investigate the influence of the choice of λ, which controls the trade-off between the closeness to the data and the smoothness of the transformation. Six values of λ including 1, 10, 100, 500, 1000 and 5000 are chosen for test. The average precisions and numbers of preserved correct matches on the whole dataset are summarized in Table 3. We see that the value of λ does not influence the performance too much, where $\lambda = 500$ seems the best and we use it as the default value.

The influence of the parameter τ has also been investigated, which is used to decide the correctness of a correspondence. The probability distribution of p_n in Eq. (8) of each putative correspondence on the whole dataset is provided in Fig. 3, where the left figure is about the inliers and the right figure is about the outliers. We see that most of the putative correspondences have posterior probability either about 0 or about 1. Therefore, we can simply set the parameter τ to be 0.5.

Finally, we provide quantitative comparisons on the dataset with other six feature matching methods such as RANSAC [7], ICF [8], GS [9], VFC [6], CSR [10] and LPM [41]. The statistical results of precisions, numbers of preserved correct matches, as well as run times are reported in Fig. 4. We see that the precisions and correct match numbers of VFC and CSR are similar, both are ranked in the middle. The average precision of ICF is the lowest, and it cannot preserve too many correct matches either. LPM does not yield high precision, due to that the neighborhood structures among correct matches cannot be preserved well in case of severe false matches. RANSAC and GS preserve the least correct matches, although GS is able to achieve satisfying precisions. In contrast, by using the progressive matching strategy, our PSSC can achieve the best precisions and best correct match numbers, where the curves are consistently above those of the other methods. The average precision of our PSSC reaches up to 95.36%, which is far ahead of the second best (e.g., 88.27%). In fact, there are only four image pairs with precisions less than 95%, and we found that these image pairs typically have extremely small inlier ratios, e.g., lower than 5%. While on these image pairs, all the other comparison methods completely fail. Our PSSC generates the most number of correct matches on almost all the image pairs, and the average correct match number achieves about 768.1 on the dataset.

TABLE 4. The inlier ratios and inlier numbers in the putative sets on the six image pair in Fig. 5.

	DogCat	*Peacock*	*Fox*	*Mex*	*Tree*	*T-shirt*
Inlier Ratio (%)	25.98	28.88	33.16	10.82	5.75	12.60
# Inlier	132	409	128	130	218	201

We also report the comparison of run times of different methods on the right of Fig. 4. We see that by using the sparse approximation, our PSSC has comparable efficiency ranked in the middle tier. LPM is the most efficient as it merely requires to construct the neighborhood for each feature points. VFC also adopts a sparse approximation similar to our PSSC and hence it is quite efficient. Note that RANSAC also has similar average run time as our PSSC; this is due to that the maximum resampling time of RANSAC is set to a relatively small number (e.g., 200).

C. RESULTS ON IMAGE PAIRS RELATED BY NON-RIGID DEFORMATION

In this section, we test our PSSC on image pairs involving non-rigid deformations. As in such test data the ground truth of matching correctness of each putative correspondence is manually determined, it is difficult to construct a large dataset with ground truth for quantitative evaluation. Thus we only choose several typical image pairs for evaluation such as *DogCat*, *Peacock*, *Fox*, *Mex*, *Tree* and *T-shirt*, as shown in Fig. 5. For the two image pairs of *DogCat* and *Peacock*, we first manually add a regular grid on it, and then warp it and take two views with different deformations. The image pair of *Fox* contains a moving fox leading to two different motion models in the scene. The two image pairs of *Mex* and *Tree* are two wide baseline image pairs, which is frequently encountered in epipolar geometry. The image pair of *T-shirt* involves natural non-rigid motion, which consists of scenes of a T-shirt undergoing two different deformations together with illumination changes.

The initial inlier ratios and inlier numbers in the putative sets on the six image pair are summarized in Table 4. We see

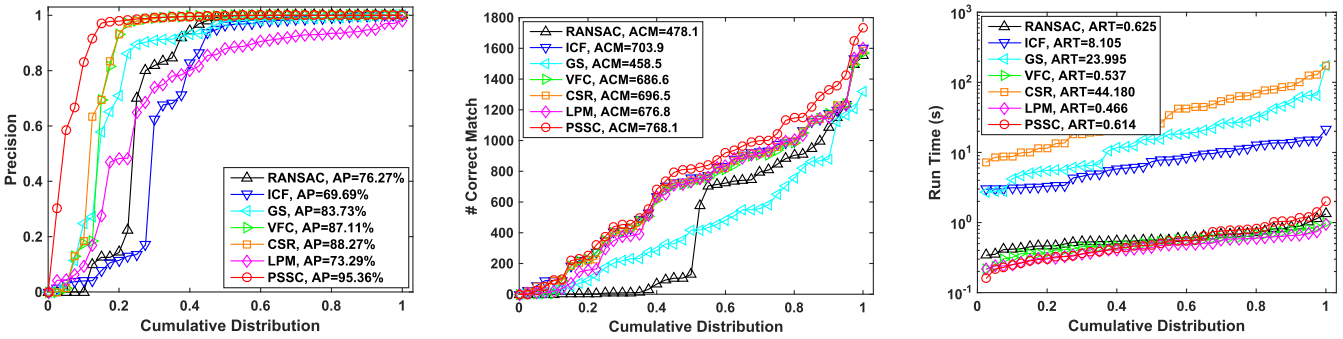


FIGURE 4. Precisions (left), numbers of preserved correct matches (middle) and run times (right) of six feature matching methods on the dataset of VGG [47]. The numbers in the three figures are average precisions (AP), average correct matches (ACM) and average run times (ATR), respectively. A point on the curve with coordinate (x, y) denotes that there are $100 * x$ percents of image pairs which have precisions, correct match numbers or run times no more than y.

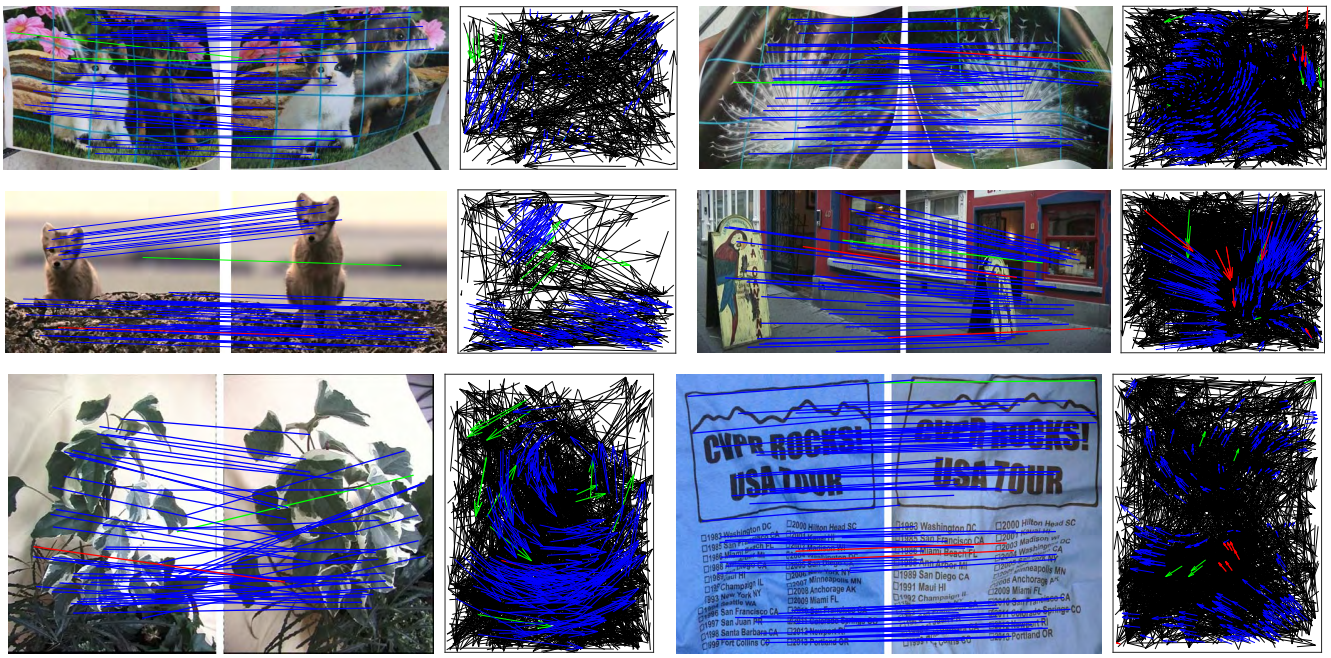


FIGURE 5. Matching results of our PSSC on six typical image pairs (from left to right and top to bottom: *DogCat*, *Peacock*, *Fox*, *Mex*, *Tree* and *T-shirt*) involve non-rigid deformations. The lines and arrows indicate the mismatch removal results. For each group of result, we use two types of representations to show the result, where the left figure provides a schematic illustration with at most 50 randomly chosen matches are shown for visibility, the right figure provides the complete mismatch removal result (blue = true positive, black = true negative, green = false negative, red = false positive). For each arrow, the head and tail correspond to the positions of two feature points in the left and right images, respectively. Best viewed in color.

that the inlier ratios are quite low, especially on the image pair of *Tree* which has only about 5.75% inliers. Therefore, the matching problem on these image pairs is quite challenging. The mismatch removal results of our PSSC are provided in Fig. 5. From the results, we see that our method is able to identify most of the inliers and outliers for all the test image pairs, even the inlier ratio is quite low.

We also provide the quantitative comparisons with other five feature matching methods on the six image pairs, where we do not use RANSAC for comparison due to that the image transformations on most image pairs cannot be modeled by a parametric model and hence RANSAC is not applicable. The precisions (%) and numbers of preserve correct matches of the five comparison methods are summarized in Table 5. Clearly, our method performs overall the best, especially in case of extremely low inlier ratio e.g., *Tree* and large non-rigid deformation (e.g., *T-shirt*). This demonstrate the generality of our PSSC for handling various feature matching problem. Note that ICF has the best results in terms of the preserved correct matches; this is because it completely fails and all the matches in the putative sets are taken as correct matches. In addition, VFC works better sometimes in terms of the matching precision, especially on the *Fox* pair. This is due to that it fits a simple smooth motion field on the ground which can precisely identify the correct matches on the ground, and hence leading to a higher precision. However, this simple

TABLE 5. Performance comparison of precision (%) and number of preserved correct matches for different feature matching methods on the six image pair in Fig. 5.

	DogCat	*Peacock*	*Fox*	*Mex*	*Tree*	*T-shirt*
ICF	(132, 25.98%)	(409, 28.88%)	(128, 33.16%)	(129, 10.74%)	(218, 5.75%)	(201, 12.60%)
GS	(106, 92.17%)	(327, 99.09%)	(142, 95.95%)	(85, 86.73%)	(101, 84.87%)	(109, 82.58%)
VFC	(125, 100.00%)	(359, 98.90%)	(97, 100.00%)	(92, 86.79%)	(146, 87.43%)	(187, 91.22%)
CSR	(120, 99.28%)	(399, 96.14%)	(120, 96.00%)	(113, 81.88%)	(156, 93.41%)	(177, 54.46%)
LPM	(125, 69.95%)	(408, 79.01%)	(127, 72.83%)	(84, 35.44%)	(163, 58.21%)	(171, 43.96%)
PSSC	(125, 100.00%)	(402, 98.77%)	(121, 99.18%)	(126, 95.45%)	(199, 91.28%)	(195, 96.53%)

motion field is not consistent with the motion of the other parts in the image pair, *e.g.* the fox, and then the matches on the fox will be all falsely removed, leading to an unsatisfying correct match number.

V. CONCLUSION

Within this paper, we propose a new method named Progressive Sparse Spatial Consensus (PSSC) for robust feature matching. It formulates the mismatch removal as a maximum likelihood estimation problem and solves it by an iterative EM algorithm. The transformation between two images is characterized by a non-parametric model with TPS kernel, which enables our PSSC to be applicable in handling both rigid and non-rigid matching problems. We also adopt a sparse approximation to the TPS kernel so that it can work well on large scale matching problem. In addition, by using a progressive matching strategy, our PSSC is able to boost the number of true feature correspondences and can successfully remove outliers even the putative match set contains extremely number of mismatches. Experimental results on publicly available datasets with comparisons to other six state-of-the-art matching methods demonstrate that our PSSC algorithm can achieve much better results, especially when the putative set is badly degraded by the mismatches.

REFERENCES

[1] J. Ma, J. Zhao, J. Jiang, and H. Zhou, "Non-rigid point set registration with robust transformation estimation under manifold regularization," in *Proc. AAAI Conf. Artif. Intel.*, 2017, pp. 4218–4224.

[2] M. Baydoun and M. A. Al-Alaoui, "Enhancing stereo matching with classification," *IEEE Access*, vol. 2, pp. 485–499, 2014.

[3] J. Zhao, J. Ma, J. Tian, J. Ma, and D. Zhang, "A robust method for vector field learning with application to mismatch removing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2977–2984.

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

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[6] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

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

[8] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[9] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1609–1616.

[10] J. Chen, J. Ma, C. Yang, and J. Tian, "Mismatch removal via coherent spatial relations," *J. Electron. Imaging*, vol. 23, no. 4, p. 043012, 2014.

[11] 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," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[12] O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 495–508.

[13] C. Wang, L. Wang, and L. Liu, "Progressive mode-seeking on graphs for sparse feature matching," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 788–802.

[14] J. Chen, Y. Wang, L. Luo, J.-G. Yu, and J. Ma, "Image retrieval based on image-to-class similarity," *Pattern Recognit. Lett.*, vol. 83, pp. 379–387, Nov. 2016.

[15] J. Zhao and J. Ma, "Visual homing by robust interpolation for sparse motion flow," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 1282–1288.

[16] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.

[17] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.

[18] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognit.*, vol. 46, no. 12, pp. 3519–3532, 2013.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intel.*, 1981, pp. 674–679.

[20] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.

[21] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, 2016.

[22] Y. Yang, S. H. Ong, and K. W. C. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognit.*, vol. 48, no. 1, pp. 156–173, 2015.

[23] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.

[24] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[25] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–142, Nov. 2017.

[26] K. Yang, A. Pan, Y. Yang, S. Zhang, S. H. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, p. 581, 2017.

[27] Z. Wei *et al.*, "A small uav based multi-temporal image registration for dynamic agricultural terrace monitoring," *Remote Sens.*, vol. 9, no. 9, p. 904, 2017.

[28] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, Dec. 1992.

[29] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

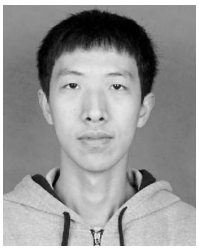
[30] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New York, NY, USA: Addison-Wesley, 2002.

[31] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, Aug. 1996.

[32] A. Rangarajan, H. Chui, and J. S. Duncan, "Rigid point feature registration using mutual information," *Med. Image Anal.*, vol. 3, no. 4, pp. 425–440, 1999.
[33] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.
[34] L. Yu, G. Zheng, and J.-P. Barbot, "Dynamical sparse recovery with finite-time convergence," *IEEE Trans. Signal Process.*, vol. 65, no. 23, pp. 6146–6157, Dec. 2017.
[35] Y. Liu, L. De Dominicis, B. Wei, L. Chen, and R. R. Martin, "Regularization based iterative point match weighting for accurate rigid transformation estimation," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 9, pp. 1058–1071, Sep. 2015.
[36] J. Maier, M. Humenberger, M. Murschitz, O. Zendel, and M. Vincze, "Guided matching based on statistical optical flow for fast and robust correspondence analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 101–117.
[37] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 138–156, 2000.
[38] O. Chum and J. Matas, "Matching with PROSAC—Progressive sample consensus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 220–226.
[39] X. Guo and X. Cao, "Good match exploration using triangle constraint," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 872–881, 2012.
[40] Y.-T. Hu, Y.-Y. Lin, H.-Y. Chen, K.-J. Hsu, and B.-Y. Chen, "Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5995–6010, 2015.
[41] J. Ma, J. Zhao, H. Guo, J. Jiang, H. Zhou, and Y. Gao, "Locality preserving matching," in *Proc. Int. Joint Conf. Art. Intell.*, 2017, pp. 4492–4498.
[42] C. A. Micchelli and M. A. Pontil, "On learning vector-valued functions," *Neural Comput.*, vol. 17, no. 1, pp. 177–204, 2005.
[43] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.
[44] X. Jiang, J. Gao, T. Wang, and D. Shi, "Tpslvm: A dimensionality reduction algorithm based on thin plate splines," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1795–1807, Oct. 2014.
[45] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. L. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2147–2154.
[46] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Comput. Vis. Image Understand.*, vol. 89, nos. 2–3, pp. 114–141, Feb. 2003.
[47] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.
[48] A. Vedaldi and B. Fulkerson, "VLFeat—An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.
[49] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.
[50] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.
[51] A. J. Smola and B. Schölkopf, "Sparse greedy matrix approximation for machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 911–918.

YONG MA received the degree from the Department of Automatic Control, Beijing Institute of Technology, Beijing, China, in 1997, and the Ph.D. degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003.

From 2004 to 2006, he was a Lecturer with the University of the West of England, Bristol, U.K. From 2006 to 2014, he was with the Wuhan National Laboratory for Optoelectronics, HUST, where he was a Professor of electronics. He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. His current research interests include signal and systems, remote sensing of the lidar and infrared, and infrared image processing, pattern recognition, and interface circuits to sensors and actuators.



JIAHAO WANG is currently pursuing the bachelor's degree majoring in electronic engineering with the Electronic Information School, Wuhan University. His research interests are in the areas of computer vision and pattern recognition.

HUIHUI XU is currently pursuing the bachelor's degree majoring in communication engineering with the Electronic Information School, Wuhan University. Her research interests are in the areas of computer vision and pattern recognition.

SHUAIBIN ZHANG is currently pursuing the bachelor's degree majoring in communication engineering with the Electronic Information School, Wuhan University. His research interests are in the areas of computer vision and pattern recognition.

XIAOGUANG MEI received the B.S. degree in communication engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007, the M.S. degree in communications and information systems from Huazhong Normal University, Wuhan, in 2011, and the Ph.D. degree in circuits and systems from HUST in 2016.

From 2010 to 2012, he was a Software Engineer with the 722 Research Institute, China Shipbuilding Industry Corporation, Wuhan. He is currently a Post-Doctoral Fellow with the Electronic Information School, Wuhan University, Wuhan. His current research interests include hyperspectral imagery, machine learning, and pattern recognition.

JIAYI MA received the B.S. degree from the Department of Mathematics Huazhong University of Science and Technology, Wuhan, China, in 2008, and the Ph.D. degree from the School of Automation, Huazhong University of Science and Technology, in 2014.

From 2012 to 2013, he was an Exchange Student with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA, USA. He is currently an Associate Professor with the Electronic Information School, Wuhan University, Wuhan, where he holds a post-doctoral position from 2014 to 2015. His current research interests include computer vision, machine learning, and pattern recognition.