Matching Images with Multiple Descriptors: An Unsupervised Approach for Locally Adaptive Descriptor Selection

Yuan-Ting Hu, Yen-Yu Lin, *Member, IEEE,* Hsin-Yi Chen, Kuang-Jui Hsu, and Bing-Yu Chen, *Senior Member, IEEE*

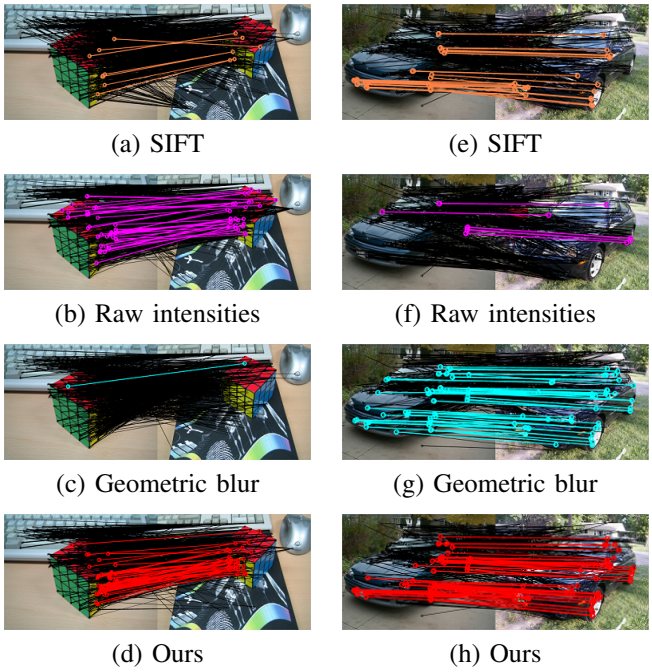
***Abstract*—With the aim to improve the performance of feature matching, we present an unsupervised approach for adaptive description selection in the space of homographies. Inspired by the observation that the homographies of correct feature correspondences vary smoothly along the spatial domain, our approach stands on the unsupervised nature of feature matching, and can choose a good descriptor locally for matching each feature point, instead of using one global descriptor. To this end, the *homography space* serves as the domain for selecting various heterogeneous descriptors. Correspondences obtained by any descriptors are considered as points in the space, and their geometric coherence and spatial continuity are measured via computing the *geodesic distances*. In this way, mutual verification across different descriptors is allowed, and correct correspondences will be highlighted with a high degree of consistency (i.e., short geodesic distances here). It follows that *one-class SVM* can be applied to identifying these correct correspondences, and achieves adaptive descriptor selection. The proposed approach is comprehensively compared with the state-of-the-art approaches, and evaluated on five benchmarks of image matching. The promising results manifest its effectiveness.**

***Index Terms*—Image feature matching, descriptor selection, geometric verification, homography space, geodesic distance, one-class SVM.**

I. Introduction

IMAGE matching aims to identify common regions across images. As a key component of image content analysis, image matching has attracted great attention for several years. It is one of the critical stages in widespread image processing and computer vision applications, such as panoramic stitching [1], object recognition [2], [3], image retrieval [4], and common pattern discovery [5].

Coupling interest points with local feature descriptors has been proven to be an effective way of image matching [6], [7]. Although the development of powerful descriptors [3], [8]–[13] has gained significant progress, there is in general no such a descriptor that is sufficient for dealing with all kinds of challenges in feature matching. Most descriptors are designed on the trade-off between *distinctiveness* and *invariance*. The more distinctive the descriptor is, the higher precision but the lower recall it may get. On the contrary, descriptors with high degrees of invariance often result in high recall but low precision. It implies that the goodness of a descriptor is usually *image-dependent*. Without any prior knowledge about images, using only one descriptor becomes insufficient and unreliable to conquer the wild image matching problems.



(a) SIFT (e) SIFT
(b) Raw intensities (f) Raw intensities
(c) Geometric blur (g) Geometric blur
(d) Ours (h) Ours

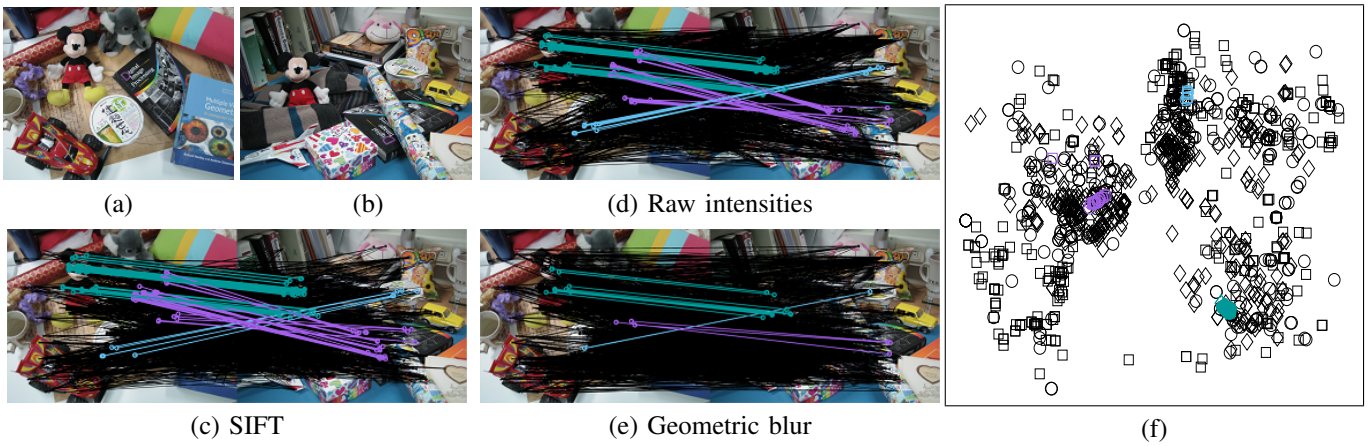
Fig. 1. Feature matching on two image pairs, (a) ∼ (d) `magic cube` and (e) ∼ (h) `car`. The matching results by using three different descriptors, including SIFT, raw intensities, and geometric blur, are shown in the first three rows, respectively. While correct correspondences are drawn in a specific color, wrong ones are in black. In `magic cube`, color/intensity information is important for matching owing to the high degree of color coherence. In contrast, shape and gradient features are more reliable in `car`. This example indicates that the performance of a descriptor varies from image to image. In addition, the deficiency of using a single descriptor is revealed. Our approach instead makes use of multiple, complementary descriptors, and can achieve superior matching results, as shown in (d) and (h).

Fig. 1 shows the matching results on two image pairs, `magic cube` and `car`, by using three descriptors, *SIFT* [3], *raw intensities*, and *geometric blur* [8], and our approach, respectively. The strong color coherence presents in the case of `magic cube`, so the color-based descriptor, raw intensities, gives good results. On the other hand, better performance is

This work was supported in part by Ministry of Science and Technology (MOST), Institute for Information Industry (III), National Taiwan University and Intel Corporation (NTU-ICRP) under Grants: MOST 103-2221-E-001-026-MY2, MOST 104-2628-E-001-001-MY2, MOST 103-2911-I-002-001, III 104-EC-17-A-24-1170 and NTU-ICRP-104R7501.

Y.-T. Hu, Y.-Y. Lin and K.-J. Hsu are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan. E-mail: r01922042@ntu.edu.tw, yylin@citi.sinica.edu.tw, kjhsu@citi.sinica.edu.tw

H.-Y. Chen and B.-Y Chen are with Nation Taiwan University, Taipei 106, Taiwan. E-mail: fensi@cmlab.csie.ntu.edu.tw, robin@ntu.edu.tw



(a) (b) (d) Raw intensities (c) SIFT (e) Geometric blur (f)

Fig. 2. (a) & (b) Two input images for feature matching. (c) Matching results by using SIFT. (d) Matching results by using raw intensities. (e) Matching results by using geometric blur. Each wrong correspondence is drawn in black, while each correct one is in a specifical color depending on the common object that it resides in. (f) The 2D visualization of correspondences in the homography space via classical multi-dimensional scaling (MDS). The circle, square, and diamond markers denote correspondences obtained by SIFT, raw intensities, and geometric blur, respectively. This figure demonstrates that not only geometric coherence but also spatial continuity are highly relevant to the correctness of correspondences.

achieved by the shape-based descriptors, SIFT and geometric blur, in the case of `car`. However, none of the three descriptors perform well in both the two cases. This example points out not only the performance fluctuation of a descriptor among images but also the deficiency of using a single descriptor.

In view of these issues, we aim at improving the quality of image matching with the aid of multiple, complementary descriptors. Two challenges arise in this scenario. First, features extracted by different descriptors are usually of different dimensions and with different scales of statistics. Also their adopted metrics for similarity measurement are diverse. How to effectively fuse information coming from heterogeneous descriptors becomes a challenging problem. Second, image matching in general is an unsupervised task. Without ground truth, one could hardly evaluate the goodness of descriptors. The other challenge is that when feature matchings by different descriptors present, how to identify correct ones from them. In this paper, we present an unsupervised approach that can effectively overcome the two problems, and generate accurate correspondences by leveraging complementary descriptors.

The idea of our approach is illustrated through Fig. 2. Fig. 2(a) and 2(b) show two input images for matching. The matching results by three descriptors, including SIFT, raw intensities, and geometric blur, are plotted in Fig. 2(c), 2(d), and 2(e), respectively. Wrong correspondences are drawn in black. Each correct correspondence is displayed with a specific color according to the common object that it resides in. Despite the varieties, a correspondence by any descriptor can be specified by its *geometric transformation* (or *homography*) in the same way. It implies that correspondences by all descriptors can be treated as points in the homography space, which can then serve as the common domain for fusing heterogeneous descriptors. We compute the pair-wise distances among points (correspondences), and show them in Fig. 2(f) via classical *multi-dimensional scaling* (MDS) [14], which summarizes high-dimensional data in a low-dimensional space by taking the pair-wise distances as input. The circle, square, and diamond markers denote correspondences obtained by using SIFT, raw intensities, and geometric blur, respectively. It can be observed in Fig. 2(f) that correct (colored) correspondences on the same object share similar homographies no matter by which descriptors they are established. They hence gather together in the homography space, while incorrect correspondences distribute irregularly. This observation suggests that geometric consistency among correspondences is highly relevant to their correctness. Moreover, Fig. 2(f) also indicates that correct correspondences are spatially correlated, since only correct correspondences within the same objects are geometrically consistent. The details of the adopted homography space and the similar measure between correspondences will be introduced later.

Inspired by the observation in Fig.2, this work carries out feature matching with multiple descriptors, and can distinguish itself with the following two main contributions. First, we present an approach for descriptor selection that stands on the unsupervised nature of image matching. It can determine the correctness of correspondences, and choose an appropriate descriptor for matching each feature point without any training data. Specifically, we estimate the geometric and spatial consistency among correspondences via computing *geodesic distances* on a designed graph to smoothly transfer the carried information in the homography space. Through this process, correct correspondences are highlighted with strong coherence with each other. It follows that *one-class SVM* [15] can be applied to picking these correct correspondences. Second, our approach is comprehensively evaluated on five benchmarks of image matching, and jointly takes five descriptors into account, including *SIFT* [3], *LIOP* [12], *DAISY* [11], *geometric blur* (GB) [8] and *raw intensities* (RI). Our approach is compared with four image matching baselines and four baselines of descriptor fusion. It achieves significantly better results than those by the best descriptors and baselines in most cases.

The rest of this manuscript is organized as follows. A review of the related works is given in Section II. The problem we address is stated in Section III. We described the adopted homography space and the used similarity measure between correspondences in Section IV. The proposed approach is specified in Section V. The experimental setup and results are

given in Section VI and Section VII, respectively. Finally, we conclude this work in Section VIII.

II. Related Work

The literature on image feature matching is quite extensive. Our review focuses on those crucial to the development of the proposed approach.

A. Local Feature Descriptors

Local feature descriptors [7] have been extensively studied, especially since the seminal works by Schmid and Mohr [16] and Lowe [3]. Various descriptors have been designed to be robust to noises while invariant to particular types of deformations in matching. For example, *SIFT (scale-invariant feature transform)* [3] describes image regions in the gradient domain, constructing a 128-dimensional histogram, and is known to be robust to scale and orientation changes. *LIOP (local intensity order pattern)* [12] encodes both the local and global ordinal information, and can alleviate the unfavorable variations caused by the changes of lighting conditions. *DAISY* [11] is featured with fast feature extraction, while keeps robust to viewpoint changes. In addition, diverse visual cues have been explored in descriptor construction, such as color characteristics [17], shapes [8], [18], internal self-similarities [19], topological information [20], and local symmetries [13]. These descriptors are designed on the trade-off between distinctiveness and invariance. Thus, there does not exist an optimal descriptor in a wide range of test images. By contrast, we introduce our approach into image matching by employing multiple descriptors to complement one another, and thus solve this problem.

B. Correspondence Verification

Identifying correct feature correspondences from candidates is an important step in image matching. Geometric layout checking is one of the most effective ways, because the geometric layout of feature correspondences often reveals their correctness. *RANSAC* [21] is a classic method for removing outliers through geometric verification. It estimates a global transformation and rejects outliers simultaneously. A correspondence is considered as an outlier and deleted if it is inconsistent with the transformation that the majority agree. One advantage of RANSAC is its easy implementation to fulfill geometric verification. However, RANSAC is not able to deal with non-rigid transformations, and would be computationally expensive when the number of outliers becomes large.

The methods in [22]–[26] relax the geometric assumption of correspondences from obeying a global transformation to a smooth feature mapping function. The feature points are linked to their corresponding points through a smooth mapping function, which makes a non-rigid transformation expressible. Ma et al. [23], [24] and Pang et al. [25] have demonstrated an effective way to determine the parameter values of the mapping function with the *vector field*. However, owing to the smoothness assumption of the mapping function, these methods are not designed for matching multiple objects, and have to be combined with additional models for handling it.

Instead of deriving the transformations involved in matching, non-rigid deformations can be dealt with via measuring the similarities between correspondences, since correct correspondences tend to be consistent with each other. Both the photometric information given by descriptors and the geometric relationship of correspondences can be used in similarity computation. Some examples of similarity measures can be found in [5], [27]–[29]. With the similarities between correspondences, a branch of research efforts, e.g., [28]–[33] cast the task of correct correspondence identification as an optimization problem over the matching score. *Graph matching* [30], [32], [34] is one of the representative techniques in optimizing the matching score. Although it is an NP hard problem, various graph matching algorithms have been proposed to get the approximate solutions. Nonetheless, these approaches are sensitive to outliers, and are less robust in multiple object matching as pointed out in [35].

Clustering based techniques for grouping correspondences with geometric constraints have been explored. Cho et al. [36] established a linkage model of correspondence clusters, and iteratively merged the clusters based on their geometric consistency. Zhang et al. [37] refined and reformulated the linkage model as a directed graph to further eliminate ambiguousness. The computational efficiency might be an issue due to the iterative algorithms. Liu and Yan [5] found local maximizers on the matching score and merged them if any two of them are similar enough. The clustering framework can handle multiple transformations, but the parameter values of the developed models are difficult to determine, such as the thresholds of identifying the correct clusters, the criteria of merging clusters and the scale factor in [5].

Voting schemes can be used for measuring the consistency between correspondences. The correspondences are checked by the pair-wise geometric consistency via mutual voting among correspondences. In Avrithis and Tolias' work [38], correspondences are transformed into Hough space and the voting results are collected efficiently with a pyramid structure. Chen et al. [39] cast the voting process as a kernel density estimation problem in the transformation space. These approaches can identify multiple objects effectively. However, voting methods would become less powerful to find correct correspondences when the number of correct matchings is so less that votes from them are not dominant during voting. Our proposed approach in spirit follows the idea of project correspondences into transformation space, but we use the geodesic distance to include the spatial layout for improving the geometric consistency measurement. The most important feature of our approach is that multiple descriptors are considered. The proposed approach increases the number of correct matchings, and can avoid the situation mentioned above.

C. Matching with Multiple Descriptors

Since different descriptors can catch diverse visual cues, using multiple descriptors has been a feasible way for improving performance. A number of approaches, such as [40]–[47], have been developed to fuse diverse descriptors for improving image matching, retrieval, classification and alignment.

Mortensen et al. [41] proposed to *concatenate* SIFT [3] and shape context [9], and reported good results in matching. However, simple feature concatenation ignores the possible variations of feature dimensions and feature value scales among descriptors. It may lead to suboptimal performance, especially when less powerful descriptors have dominant feature dimensions or values. To address this problem, Bosch et al. [40] represented images under each descriptor as a *kernel matrix*. The works by Brox and Malik [44] and Weinzaepfel et al. [46] integrated descriptor matching for handling large displacement in optical flow, and combined multiple visual evidences represented in form of *energy functions*. Although kernel matrices and energy functions can serve as the unified domains for descriptor fusion, these approaches tune or learn fixed weights for descriptor combination. It may not be suitable for image matching, because the optimal descriptors change from image to image. Furthermore, using brute force search to determine descriptor weights may become infeasible, when there are a large number of descriptors to be considered. Besides, image matching is an unsupervised task, and no training or validation data are available for determining descriptor weights in general.

Our approach tackles these issues, and has the following two advantages: 1) Multiple descriptors are represented by their homographies in matching so that we can work with complementary descriptors without worrying about their diversities of feature dimensions or feature value scales; 2) Our approach allows geometric checking across descriptors, and consensus correspondences will be revealed through the process. It means that the plausible correspondences by various descriptors can be jointly identified in a fully unsupervised manner.

III. Problem Statement

We aim to match two given images I^P and I^Q, which come with the sets of detected feature points, $U^P = \{u_i^P\}_{i=1}^{N^P}$ and $U^Q = \{u_i^Q\}_{i=1}^{N^Q}$, respectively. The support region of each feature $u_i \in U^P \cup U^Q$ is assumed to an ellipse in this work. These feature points can be obtained by using off-the-shelf detectors, such as *Harris-Affine* [48], *Hessian-Affine* [48], the salient region detector [49], or their combinations. We use Hessian-Affine detector for its efficiency and high repeatability. Multiple descriptors are employed to characterize each feature point. The center and the described appearances of feature u_i are respectively denoted by $\mathbf{x}_i$ and $\{\mathbf{f}_{i,m}\}_{m=1}^{M}$, where M is the number of the employed descriptors. The set $\tilde{\mathcal{C}} = U^P \times U^Q$ covers all possible feature *correspondences* (or *matchings*). Our goal is to detect correct correspondences in $\tilde{\mathcal{C}}$ as many as possible.

The number of the detected feature points in an image, i.e., N^P or N^Q, is often in the order of 10^3. Directly searching correct correspondences in $\tilde{\mathcal{C}}$ may be inefficient. Hence we start from compiling a reduced set $\mathcal{C}$ of $\tilde{\mathcal{C}}$ by filtering out correspondences that are unlikely to be correct. For each feature $u_i^P \in I^P$, we find the set of the most similar r matchings, $\mathcal{C}_{i,m} = \{(u_i^P, u_{i_k,m}^Q \in I^Q)\}_{k=1}^{r}$, with descriptor m and the yielded distance $\|\mathbf{f}_{i,m}^P - \mathbf{f}_{i_k,m}^Q\|$, by searching the whole U^Q. Since total M descriptors are adopted, at most $r \times M$ matchings of u_i^P are kept in $\mathcal{C}$ after removing the duplicates. Namely,

$$\mathcal{C} = \bigcup_{i=1}^{N^P} \mathcal{C}_i, \text{ where } \mathcal{C}_i = \bigcup_{m=1}^{M} \mathcal{C}_{i,m}. \tag{1}$$

We will work on the reduced candidate set $\mathcal{C}$. The value of r controls the trade-off between efficiency and accuracy. We will analyze the effect of r on our approach in the experiments.

IV. Homography Space: The Domain for Descriptor Selection

In this section, we first introduce how to compute the geometric transformations of correspondences and how to measure their geometric dissimilarity. Then, we show how to fuse information grabbed by diverse descriptors in the homography space, where correspondences obtained by different descriptors are treated in a unified manner.

The elliptical region of feature u_i can be specified by mapping a circular region centered on the origin via the affine transformation:

$$T(u_i) = \begin{bmatrix} A(u_i) & \mathbf{x}_i \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^{2\times 1}$ is the feature center, and $A(u_i) \in \mathbb{R}^{2\times 2}$ is a non-singular matrix accounting for the scale, the shape, and the orientation of u_i. After normalization with transformation $T(u_i)^{-1}$, all the adopted descriptors can be applied to u_i, and generate $\{\mathbf{f}_{i,m}\}_{m=1}^{M}$. Refer to [48] for the details.

A homography in this work refers to the geometric transformation of a feature correspondence. For a correspondence between $u_i^P \in U^P$ and $u_j^Q \in U^Q$, the transformation from the support region of u_i^P to that of u_j^Q can be derived as

$$H_{ij} = T(u_j^Q) * T(u_i^P)^{-1} \in \mathbb{R}^{3\times 3}. \tag{3}$$

H_{ij} is a 6-dof (degrees of freedom) affine transformation (or affine homography). Thus, it can be viewed as a point in the 6-dimensional affine homography space $\mathcal{H}$. Note that an affine matrix characterizes the transformation instead of a general perspective matrix, because it had a better match with the adopted Hessian-Affine detector, which is affine invariant.

Consider two correspondences $c = (u_i^P, u_j^Q) \in \mathcal{C}$ and $c' = (u_{i'}^P, u_{j'}^Q) \in \mathcal{C}$. We use the *reprojection error* to measure their geometric dissimilarity. Specifically, the homography matrices, H_{ij} and $H_{i'j'}$, of the two correspondences are firstly computed by Eq. (3). The *projection error* of (u_i^P, u_j^Q) with respect to $H_{i'j'}$ is then calculated by

$$d_{err}(u_i^P, u_j^Q, H_{i'j'}) = \|\mathbf{x}_j^Q - \rho(H_{i'j'} \begin{bmatrix} \mathbf{x}_i^P \\ 1 \end{bmatrix})\|, \tag{4}$$

where function $\rho : \mathbb{R}^3 \to \mathbb{R}^2$ is defined as

$$\rho(\begin{bmatrix} a \\ b \\ c \end{bmatrix}) = \begin{bmatrix} \frac{a}{c} \\ \frac{b}{c} \end{bmatrix}. \tag{5}$$

The projection error $d_{err}(u_i^P, u_j^Q, H_{i'j'})$ is the induced error when changing the homography from H_{ij} to $H_{i'j'}$ on correspondence (u_i^P, u_j^Q). The reprojection error between correspondences c and c' is then defined as

$$d(c,c') = \frac{1}{4}\big(d_{err}(u_i^P, u_j^Q, H_{i'j'}) + d_{err}(u_j^Q, u_i^P, H_{i'j'}^{-1}) + d_{err}(u_{i'}^P, u_{j'}^Q, H_{ij}) + d_{err}(u_{j'}^Q, u_{i'}^P, H_{ij}^{-1})\big). \quad (6)$$

We will use the reprojection error to measure the geometric dissimilarity between correspondences in $\mathcal{C}$.

We consider characterizing each feature point u_i with total M kinds of different descriptors, i.e., $\{\mathbf{f}_{i,m} \in \mathcal{F}_m\}_{m=1}^M$. The resulting representations by these descriptors are typically of various dimensions and even in diverse forms, such as vectors, histograms, and pyramids. To avoid the difficulties caused by these varieties, we use the homography in Eq. (3) as the representation for a feature correspondence. As the homography contains geometric relationship, it allows us to select good descriptors based on the geometric information. In the following sections, each correspondence in $\mathcal{C}$ in Eq. (1) is represented as the corresponding homography, and hence can be treated as a point in the homography space.

V. The Proposed Approach

We formulate the task of image matching as finding an appropriate matching for each feature point u_i^P in image I^P, i.e., picking the most plausible correspondence from $\mathcal{C}_i$ in Eq. (1). In this work, multiple descriptors collaborate in the sense that they jointly determine candidate correspondences with diverse region characteristics and different kinds of invariance, so the probability that the correct correspondence resides in $\mathcal{C}_i$ largely increases. The goal at this stage is to determine the correct correspondence for each u_i^P, if it exists. The unsupervised nature of image matching makes this task very challenging, because no prior knowledge or relevant training data are available.

We tackle this issue based on the observation that the homographies of correct correspondences vary smoothly along the spatial locations in the image. Specifically, we firstly employ a graph to encode the spatial arrangement among correspondences, and compute the *geodesic distances* on the graph for geometric coherence estimation. Then, we utilize *one-class SVM* [15] to identify correct correspondences, since it can effectively separates alike (both geometrically and spatially coherent here) data from the outliers. The two steps are respectively described in the following.

A. Geometric and Spatial Coherence Estimation

To jointly consider the geometric and spatial relationships among correspondences, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which each correspondence $c_i \in \mathcal{C}$ is associated with a vertex $v_i \in \mathcal{V}$, while an undirected edge $e_{ij} \in \mathcal{E}$ is added to connect vertices v_i and v_j if the end points in image I^P of c_i and c_j are near enough. That is, the number of vertices in $\mathcal{G}$, $|\mathcal{V}|$, is the same as the number of correspondences in $\mathcal{C}$. In our implementation, we consider two feature points in I^P are nearby if one point belongs to the k spatial nearest neighbors of the other point. The resulting edge set $\mathcal{E}$ hence encodes the spatial connection among correspondences. The value of k controls the neighborhood size. Its effect on the performance will be investigated in the experiments. Weight w_{ij} assigned to edge e_{ij} is defined as

$$w_{ij} = \begin{cases} d(c_i, c_j), & \text{if } e_{ij} \in \mathcal{E}, \\ \infty, & \text{otherwise}, \end{cases} \quad (7)$$

where the geometric dissimilarity $d(c_i, c_j)$ between correspondences c_i and c_j is given in Eq. (6). With the weighted graph, we compute the geodesic distance between each pair of vertices (i.e., correspondences). We denote the geodesic distance between correspondences c_i and c_j by $d_{geo}(c_i, c_j)$ hereafter. It can be seen that graph $\mathcal{G}$ integrates the spatial continuity into the estimation of geometric coherence. The use of geodesic distance catches the phenomenon that the homographies of correct correspondences on the same object may change smoothly along their spatial locations. Therefore, the resulting dissimilarity measure can deal with possible deformations in matching.

Suppose there are N correspondence candidates, i.e., $\mathcal{C} = \{c_1, ..., c_N\}$. We compute the pair-wise geodesic distances $\{d_{geo}(c_i, c_j)\}_{i,j=1}^N$. The correct correspondences will be highlighted with strong geometric and spatial coherence (short geodesic distances) with other correct correspondences. It is worth pointing out that compared with $d(c_i, c_j)$ in Eq. (6), the geodesic distance $d_{geo}(c_i, c_j)$ can more faithfully measure the dissimilarity between c_i and c_j, since the spatial information is taken into account to remove the noises, i.e., incorrect correspondences whose homographies happen to be consistent with those of the correct ones. The performance gain of using geodesic distances over reprojection errors will also be evaluated in each of the used dataset in the experiments.

B. Correct Matching Identification by One-class SVM

At this stage, we apply one-class SVM to distinguishing the correct correspondences from candidate set $\mathcal{C}$. One-class SVM is one of the state-of-the-art methodologies for unsupervised classification. It separates positive and negative data in an asymmetrical scenario: it assumes that positive data are similar to each other, while negative data are different in their own ways. The reason why we use one-class SVM for correct correspondence identification is justified in the following.

In our case, the correct correspondences are usually geometrically and spatially consistent with each other, i.e., short geodesic distances among them. On the other hand, the wrong matchings are caused by various factors, so their homographies often irregularly distribute. It results in that the homography of a wrong correspondence tends to be dissimilar to most homographies of all the other correspondences. The correct matchings meet the definition of positive data in one-class SVM in the sense that they are both geometrically and spatially coherent, and have short geodesic distances between them. On the other hand, the wrong matchings are wrong due to various causes. They correspond to negative data, since they are inconsistent with each other. Thus, our case closely meets

the scenario of one-class SVM. For the set of correspondence candidates, $\mathcal{C} = \{c_1, ..., c_N\}$, one-class SVM predicts their labels by solving the following constrained optimization problem

$$\begin{aligned} \min_{\mathbf{w},\{\epsilon_i\}} \ & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{Co \cdot N}\sum_{i=1}^{N}\epsilon_i - \nu \\ \text{subject to } & \mathbf{w}^\top \phi(c_i) \geq \nu - \epsilon_i, \text{ for } 1 \leq i \leq N, \\ & \epsilon_i \geq 0, \text{ for } 1 \leq i \leq N, \end{aligned} \tag{8}$$

where Co and ν are the two parameters of one-class SVM. We set $Co = 1$ and $\nu = 0.5$ in all the experiments. As a kernel machine, one-class SVM can work on nonlinearly mapped data. The function ϕ maps the data, correspondences here, to a *Reproduced Kernel Hilbert Space* (RKHS), which is implicitly defined by the adopted kernel. The optimization of one-class SVM can be accomplished by referencing only the inner products of pairs of the mapped data, and the inner product can be efficiently computed via the *kernel trick*, i.e., $k(c_i, c_j) = \langle \phi(c_i), \phi(c_j) \rangle$.

Note that we don't explicitly define the feature representation of correspondences in $\mathcal{C}$, but their pair-wise dissimilarity through d_{geo}. A kernel function is used to encode the similarity among data. In this work, the kernel matrix $K \in \mathbb{R}^{N \times N}$ and the kernel function $k(\cdot, \cdot)$ are defined as follows:

$$\begin{aligned} K(i,j) &= k(c_i, c_j) \\ &= \exp\left(-\frac{d_{geo}(c_i, c_j)}{\sigma}\right), \text{ for } 1 \leq i, j \leq N, \end{aligned} \tag{9}$$

where σ is the hyperparameter. We set σ as the average geodesic distance from each correspondence to its nearest neighbor. It follows that each correspondence c_i is predicted via $\text{sign}(f(c_i))$, where score $f(c_i)$ is in form of:

$$f(c_i) = \mathbf{w}^\top \phi(c_i) - \nu = \sum_{j=1}^{N} \alpha_j k(c_i, c_j) - \nu, \tag{10}$$

and $\{\alpha_j\}_{j=1}^{N}$ are the optimized coefficients of the support vectors. Note that the results of image matching are often jointly measured by *precision* and *recall*. For each feature point u_i^P in image I^P, we pick its correspondence as the one that has the highest prediction score in $\mathcal{C}_i$ (cf. Eq. (1)). All picked correspondences are further sorted according to their prediction scores. Precision-recall analysis can then be carried out with the sorted list and a set of thresholds.

The proposed approach is easy to implement. The geodesic distances are computed by finding shortest paths in $\mathcal{G}$, and a few packages of one-class SVM are available, such as *LibSVM* [50] which is adopted in this work. The time complexity of our method is between $O(N^2)$ to $O(N^3)$, i.e., the complexity of one-class SVM, where N is the number of correspondences. Note that in this work, N grows linearly with respect to the number of descriptors used for fusion. On the Co-reg dataset [51], each image has around 1,100 detected feature points. The average running time of our approach (calculating the geodesic distances and optimizing the one-class SVM) on each image pair is about 36 seconds on a modern PC with an Intel Core $i7$-4770 processor and 16GB memory.

VI. Experimental Setup

In this section, we introduce the details of our experimental setting, including the used feature detector, descriptors, baselines, datasets, and evaluation criteria.

A. Feature Detector and Feature Descriptors

The Hessian affine invariant detector [48] is used in our experiments to detect interest points and their surrounding elliptical support regions. Each detected region is normalized and rotated to the principal orientation. We apply all the feature descriptors to the normalized patches except for geometric blur. We will explain it later. The average number of detected interest points in an image is around 10^3.

In the experiments, we adopt five feature descriptors, including *SIFT* [3], *LIOP* [12], *DAISY* [11], *Raw intensity* (RI) and *geometric blur* (GB) [8]. The RI descriptor stores the pixel intensities in a raster scan order, and the normalized support region is used to construct a RI descriptor. The GB descriptor is designed to encode wide-range shape information. Thus, it is applied to support regions that are enlarged by three times in advance. Euclidean distance is used to measure the dissimilarity between two regions under each descriptor. The five descriptors capture diverse image characteristics and tend to complement each other.

B. Baselines

For performance comparison, we implemented eight baselines of image feature matching. Four of them are the state-of-the-art feature matching algorithms, including the graph matching method, *spectral matching* (SM) [27], the clustering-based approach, *agglomerative correspondence clustering* (ACC) [36], the voting-based method, *Hough voting* (HV) [39], and *vector field consensus* (VFC) [24]. The used affinity measure between correspondences in SM, ACC and HV is the reprojection error for fair comparison. Our approach can work with either a single or multiple descriptors. When a single descriptor is used, our approach is compared with the four matching algorithms. The parameter r for establishing initial correspondences is set to 5 as used in [39] when using a single descriptor.

The other four baselines are designed to fuse multiple descriptors, including *concatenation* (CAT), *concatenation with Hough voting* (CAT+HV), *Ranking*, and *Ratio*. The baseline CAT fuses multiple descriptors by concatenating all the feature descriptors. Nearest neighbor search is applied to finding the possible matching candidates. The baseline CAT+HV uses the initial candidates constructed by baseline CAT as the input of HV. Compared with CAT, CAT+HV additionally realizes geometric checking by Hough voting. In baseline Ranking, we find the first nearest neighbors of all feature points in image I^p with a specific descriptor, and rank the yielded correspondences according to the descriptor distances. The procedure is repeated for each descriptor. For each feature point in I^p, we determine its correspondence by using the descriptor that has the highest rank at this point. In baseline Ratio, we match each point in I^p to its first two nearest neighbors by using one specific descriptor, and compute the distance

ratio between the two matches. The smaller the ratio is, the more confident the descriptor is at this point. By comparing the ratios across descriptors, we find the correspondence of this point by using the descriptor with the smallest distance ratio. Matching methods, SM, ACC, HV and VFC, are designed to work with a single descriptor. We extend the four methods to use multiple descriptors by taking the union of correspondence candidates of all descriptors, i.e., those in Eq. (1), as input. Therefore, the proposed approach is compared with all the eight baselines when multiple descriptors are used.

For fair comparison, all the baselines and our approach in each comparison setup use the same detected interest regions, descriptors, and evaluation criteria in the experiments.

C. Datasets

The performance of our approach is evaluated on five benchmarks of image matching, including Object dataset [31], Co-recognition (Co-reg for short) dataset [51], Symfeat (SYM for short) dataset [13], VGG dataset [7] and VGG model house dataset available at http://www.robots.ox.ac.uk/~vgg/data1.html. Note that the VGG dataset is composed of eight image sets, and each set contains 6 images. We carry out image matching with two different degrees of variation in the experiment, i.e., the first image vs. the second one and the first one vs. the fourth one, which represent the slight and drastic variations in matching, respectively. Note that many image pairs in Co-reg dataset undergo transformations which are not purely affine. To get preeminent performance on all datasets is very challenging, because the five datasets exhibit a variety of variations, such as diverse kinds of deformation, various types of scenes, and different numbers of common objects. They jointly serve as a good test bed for performance evaluation.

D. Evaluation criteria

For performance measure, the evaluation metrics used in the experiments are introduced. For datasets VGG and SYM, we follow [7], and consider that a correspondence is correct if the overlap error is less than 50%. For datasets Object and Co-reg where manually annotated ground truth is available, a correspondence is correct if the distance between the matched feature point and the ground truth is within eight pixels.

After determining the correctness of correspondences, the performance of a matching algorithm can be measured by jointly taking *precision* and *recall* into account. While precision is the fraction of detected correspondences that are correct, recall is the fraction of correct correspondences that are detected. Specifically, the two terms are respectively defined as

$$\text{PRECISION} = \frac{n\text{TP}}{n\text{TP} + n\text{FP}}, \tag{11}$$

and

$$\text{RECALL} = \frac{n\text{TP}}{n\text{TP} + n\text{FN}}, \tag{12}$$

where $n\text{TP}$ and $n\text{FP}$ are the numbers of correctly and wrongly detected correspondences by a matching method, respectively. $n\text{FN}$ is the number of correct correspondences that are not detected.

For each matching approach, including ours and the eight adopted baselines, all the detected correspondences are sorted by its own criterion, such as the element values of the eigenvector in baseline SM, the ratio values in baseline Ratio, and the prediction score, Eq. (10), in our approach. By sampling on the sorted lists, the performance of each approach can then be represented by a *precision-recall curve* on an image pair or *mean average precision* (mAP) on a dataset. The mAP is the mean of the average precision, while the average precision on an image pair is calculated by averaging the precisions with different numbers of returned correspondences.

VII. Experimental Results

The performance of the proposed approach is evaluated and analyzed in this section, which is organized as follows: First, we investigate the effect of the two main parameters, r and k, on our approach. They control the sizes of the candidate set and the neighborhood, respectively. Second, the transformation space is visualized to verify that correct correspondences established by all the descriptors gather together in that space, while incorrect ones distribute irregularly. We also visualize and compare the homography spaces when the reprojection error and the proposed geodesic distance serve as the distance functions, respectively. Third, we colorize the established correspondences on Object dataset according to their associated descriptors, and analyze how our approach carries out locally adaptive descriptor selection. Fourth, our approach is compared with the eight baselines on three benchmarks of feature matching. The obtained quantitative results of all methods are presented in the forms of mAPs and precision-recall curves. Fifth, we show the matching results to demonstrate that our approach can leverage multiple, complementary descriptors to achieve remarkable performance gains in feature matching. Sixth, we present a set of experiments on the combinations of descriptors to show the performance of the proposed approach working with various types and numbers of descriptors. Finally, we investigate into the ability of handling perspective variations of the proposed matching algorithm.

A. Parameter Choosing

There are two important parameters, r and k, in our approach. Parameter r controls the size of $\mathcal{C}$ in Eq. (1). Parameter k decides the neighborhood structure in the constructed graph described in Section V-A. The effect of parameters r and k on our approach on Co-reg dataset is investigated in Fig. 3. It can be observed that our approach is not very sensitive to the two parameters, since the performance in mAP with various value combinations of r and k is still within 4 percent. Setting r to 3 gives better results, and there is no more performance improvement when k is set to larger than 80. Thus, we fix $r = 3$ and $k = 80$ for our approach in the following experiments.

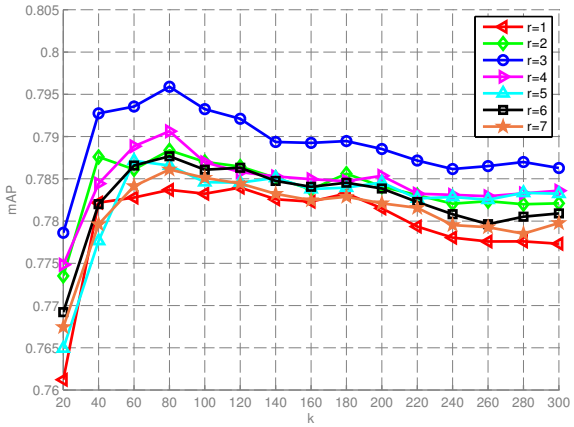


Fig. 3. The effect of parameters r and k on the performance (in mAP) of our approach on the Co-reg dataset.

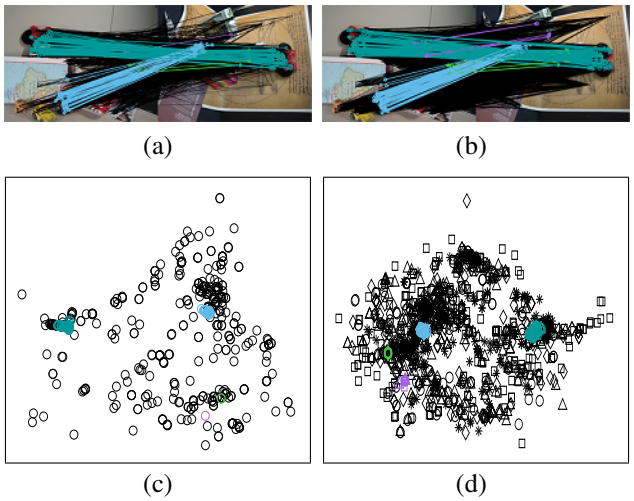


Fig. 4. (a) The matching candidates given by SIFT on image pair `Minnies` of the Co-reg dataset. (b) The matching candidates given by all the five descriptors on the same image pair. (c) 2D visualization of the homograph space of the matchings in (a). (d) 2D visualization of the homograph space of the matchings in (b).

B. Homography Space Visualization

To check whether the homography space is qualified to serve as a uniform domain for descriptor fusion, we visualize it by classical *multi-dimensional scaling* (MDS) [14], which can approximate the pair-wise distances between data in a 2D space. Because the number of the correspondences is too large to clearly show all correspondences in a figure, we use weighted sampling to only display a fraction of the correspondences for better visualization. The weights (or the probabilities of being sampled) are set to the densities in the homography space. The Co-reg dataset is used in the experiments. There are multiple common objects in every pair. This property allows us to observe the relationship between correct correspondences that belong to different objects.

Fig. 4(a) and Fig. 4(b) show the sets of the initial matching candidates, via (1), by using SIFT and by using all the five descriptors, respectively. There are four common objects in this pair of images. The correct matchings are colored according to the objects that they lie on. By using the geodesic distance presented in Section V-A, the two sets of matchings

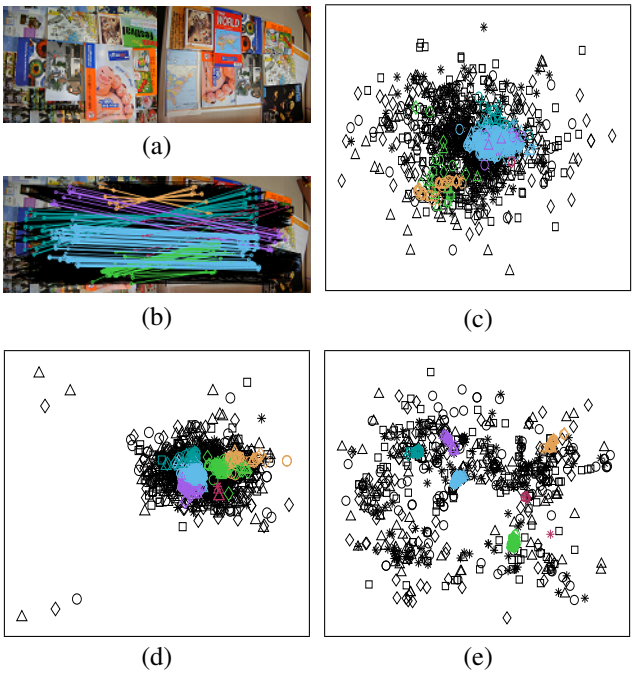
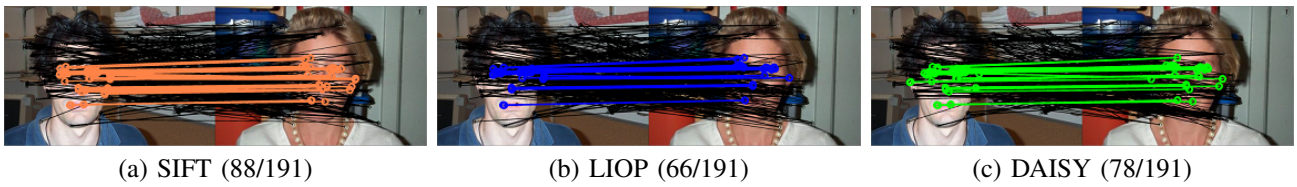


Fig. 5. (a) An image pair, `Books` of the Co-reg dataset. (b) The initial correspondences. (c) 2D visualization of these correspondences in the homography space when the intrinsic distance on the affine manifold is used. (d) 2D visualization of these correspondences in the homography space when the reprojection error is used. (e) 2D visualization of these correspondences in the homography space when the developed geodesic distance is used.

are visualized in Fig. 4(c) and Fig. 4(d), respectively. It turns out that no matter how many descriptors are used, the correct (colored) correspondences gather together, while wrong matchings distribute irregularly. This example also points out that using multiple descriptors helps to find out the correct correspondences that are sparsely detected with a single descriptor and may be neglected, such as the purple correspondences in Fig. 4(c). In Fig. 4(d), the purple correspondences given by diverse descriptors can mutually support in both geometric and spatial coherence estimation. It implies that they are more probably predicted as correct correspondences by one-class SVM.

As described in Section V-A, the developed geodesic distance computed over the designed graph takes both geometric consistency and spatial continuity into account. We compare the geodesic distance with the reprojection error as well as another alternative for measuring the dissimilarity between homographies by visualizing the homography spaces that they induce. The *geodesic* of two homographies on the affine manifold is considered in this alternative to the reprojection error. The *intrinsic distance* of the two homographies on the affine manifold is computed as the Euclidean distance of the corresponding Lie algebra of the two homography matrices as described in [52]. Refer to [52] for further details. In Fig. 5(a), two images to be matched are shown. The initial correspondence candidates are given in Fig. 5(b). We calculate the dissimilarity between these correspondences by using the geodesic on the affine manifold, the reprojection error and our proposed geodesic distance, and show their distributions in the homography space in Fig. 5(c), (d) and (e), respectively. It can be observed in Fig. 5(c) and (d) that the

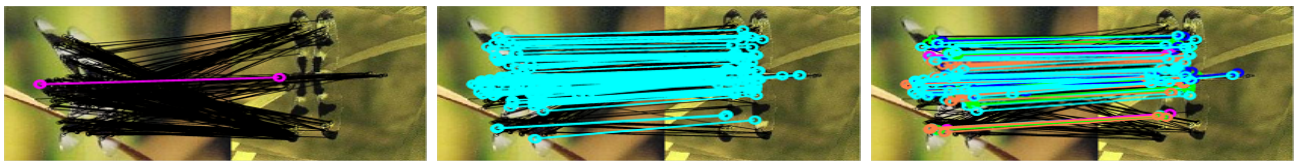


(a) SIFT (88/191) (b) LIOP (66/191) (c) DAISY (78/191)

(d) RI (45/191) (e) GB (38/191) (f) All (116/191)

Fig. 6. The matching results by our approach, on image `face` of the Object dataset, with (a) the SIFT descriptor, (b) the LIOP descriptor, (c) the DAISY descriptor, (d) the RI descriptor, (e) the GB descriptor, and (f) all the five descriptors. The recalls in Eq. (12), namely $n\text{TP}/n\text{TP}+n\text{FN}$, are shown in brackets. The correct correspondences are colored, and their colors indicate the descriptors by which they are established.

(a) SIFT (61/461) (b) LIOP (38/461) (c) DAISY (77/461)



(d) RI (3/461) (e) GB (132/461) (f) All (168/461)

Fig. 7. The matching results by our approach, on image `dragonfly` of the Object dataset, with (a) the SIFT descriptor, (b) the LIOP descriptor, (c) the DAISY descriptor, (d) the RI descriptor, (e) the GB descriptor, and (f) all the five descriptors. The recalls in Eq. (12), namely $n\text{TP}/n\text{TP}+n\text{FN}$, are shown in brackets. The correct correspondences are colored, and their colors indicate the descriptors by which they are established.

correct correspondences on an object may mix with the correct correspondences on the other objects because the intrinsic distance on the affine manifold and the reprojection error do not take spatial continuity into account. On the other hand, the correct correspondences on an object tightly assemble in Fig. 5(d), and are well separated from the rest. It implies that the developed geodesic distance can effectively consider both geometric and spatial consistency to better discover object-wise consistent homographies. This property facilitates correct correspondence identification by one-class SVM.

C. Locally Adaptive Descriptor Selection

Our approach can work with either a single descriptor or multiple descriptors. In this part of experiments, we show the matching results by applying our approach to the five adopted descriptors individually and jointly, and demonstrate the advantage of adaptive descriptor selection enabled by our algorithm. The Object dataset is used in the experiments. Each image pair in this dataset contains different object instances of the same category. The performance of a descriptor often varies from image pair to image pair, even from image region to image region, due to the diversity of intra-class variations.

Fig. 6 displays the matching results as well as the recalls by our approach, on image pair `face` of the Object dataset, with each of the five adopted descriptors and all of them. The correct correspondences are colored with their colors indicating the descriptors by which they are established, i.e., SIFT in orange, LIOP in blue, DAISY in green, RI in magenta, and GB in cyan. Descriptor SIFT finds the most correct matchings in this example. As shown in Fig. 6(f), our approach indeed selects most correspondences established by SIFT. Similar observation can be found in Fig. 7, in which the matching results on image pair `dragonfly` of the Object dataset are plotted. In this example, descriptor GB performs best, and our approach also finds most correspondences by GB. The results in Fig. 6 and Fig. 7 show why our approach can leverage multiple, complementary descriptors to boost the matching performance: It estimates both the geometric and spatial consensus in a feature point-specific manner, and hence can adaptively identify the correspondences established by the better descriptors.

D. Quantitative Results

We compare the performance of our approach by using the geodesic on the affine manifold, the reprojection error, and the proposed geodesic distance, and report the accuracies in mAP in TABLE I, in which the Co-reg, SYM and VGG datasets are used. The mAPs on the three datasets are improved when the

TABLE I
THE ACCURACY IN MAP OF USING THE REPROJECTION ERROR AND THE GEODESIC DISTANCE IN OUR APPROACH

mAP (%)	Co-reg	SYM	VGG
Geodesic on Affine Manifold	55.42	46.38	93.29
Reprojection Error	72.84	42.59	93.49
Proposed Geodesic Distance	**79.59**	**46.83**	**93.81**

geodesic distance is used. It indicates that the geodesic distance more faithfully grasps the intrinsic relationships between correspondences by exploring the graph which encodes both the spatial and geometric consistency. The performance gains over reprojection error, about 6% and 4%, on the first two datasets, i.e., Co-reg and SYM, and the performance gain over the geodesic on the manifold on Co-reg dataset is remarkable. The main reason is that the multiple common objects in Co-reg and the foregrounds and backgrounds in SYM typically have diverse transformations in matching, but each of these transformations tends to vary smoothly in the spatial domain. Hence, modeling spatial coherence is helpful. On the other hand, all interest points in each image of dataset VGG almost undergo the same transformation in matching. Giving additional spatial information does not help much in the cases. Using the geodesic on the affine manifold performs well on SYM and VGG datasets, but poorly on Co-reg dataset. The main reason is that dramatic viewpoint changes in Co-reg dataset make the estimated homograpies in Eq. (3) noisy. The transformations, $T(u_i^P)$ and $T(u_j^Q)$, of a homograpy are inferred from the detected regions whose stability often decreases in dramatic viewpoint changes. Noisy homograpies diminish the performance of using the geodesic on the affine manifold. The phenomenon can be observed in Fig. 5(c) where the correct correspondences within the same object less tightly gather together. On the contrary, the reprojection error is more stable in this case, because it considers the error of the reprojected center in Eq. (4), instead of the difference on the affine manifold.

We evaluate and compare our approach with the eight baselines. Five different descriptors, including SIFT, LIOP, DAISY, RI, and GB, are considered. TABLE II summarizes the performances in mAP of all the approaches on the three benchmarks. The first four approaches to image matching, i.e., SM [27], ACC [36], HV [39], VFC [24], consider a single descriptor at a time as in [24], [27], [36], [39]. Their performances with each of the five descriptors as well as the average performances are reported. Besides, we extend the four baselines to multiple-descriptor cases by applying them to the fused candidate sets constructed by multiple descriptors. Our approach is applied to both a single descriptor and multiple descriptors as well. The other four baselines jointly take multiple descriptors into account, so only the performances of descriptor fusion are reported. In TABLE II, the best performance on each benchmark is given in bold, while the best performance by using a single descriptor is given in italic and followed by a star sign.

We firstly focus on the cases where a single descriptor is used. SIFT gives the best performance in dataset Co-reg, while LIOP performs best in dataset VGG. The experimental results show that the five descriptors complement each other and no single descriptor can get the best performance on all the datasets. The optimal descriptor for matching vary from image to image. It hence points out that fusing multiple descriptors can be a feasible way for improving performance. As for the performances of the image matching algorithms, baseline HV averagely gets the superior results, since it fully supports multiple object matching, and stably works with

TABLE II
THE PERFORMANCES IN MAP OF THE EIGHT BASELINES AND OUR APPROACH ON THE THREE DATASETS

method	descriptor	dataset		
		Co-reg	SYM	VGG
SM [27]	SIFT	55.30	18.92	70.73
	LIOP	38.74	16.79	87.71
	DAISY	46.71	22.72	78.62
	RI	34.57	7.99	59.55
	GB	12.13	32.57	62.96
	Average	37.49	19.80	71.92
ACC [36]	SIFT	60.28	26.74	82.53
	LIOP	29.83	19.49	91.76
	DAISY	36.49	29.97	86.33
	RI	15.10	10.70	67.98
	GB	8.88	29.28	64.36
	Average	30.12	23.57	78.59
HV [39]	SIFT	60.12	22.21	82.31
	LIOP	43.97	18.69	91.78
	DAISY	50.06	26.92	88.14
	RI	37.14	11.88	70.33
	GB	12.08	38.28	71.84
	Average	40.67	23.60	80.88
VFC [24]	SIFT	31.11	20.76	82.24
	LIOP	11.79	14.12	93.51
	DAISY	16.29	21.41	85.03
	RI	4.51	4.63	47.13
	GB	1.77	28.82	64.01
	Average	13.09	17.95	74.38
Ours	SIFT	*72.03**	25.11	85.24
	LIOP	51.17	20.41	***94.23****
	DAISY	61.30	30.28	90.70
	RI	35.89	11.38	68.57
	GB	10.80	*40.65**	71.50
	Average	46.24	25.57	82.05
CAT	All	19.62	7.90	61.27
CAT+HV	All	39.70	13.36	75.13
Ranking	All	48.48	27.99	85.68
Ratio	All	53.61	28.75	90.47
SM	All	55.45	37.36	92.11
ACC	All	56.00	36.65	85.73
HV	All	72.21	41.50	93.70
VFC	All	12.13	33.76	92.49
Ours	All	**79.59**	**46.83**	93.81

TABLE III
THE ACCURACY IN MAP OF FIVE MATCHING APPROACHES WITH RESPECT TO FIVE VARIATIONS ON VGG DATASET

Type of Variation	Blur	Viewpoint	Zoom+Rotation	Light	JPEG Compression
Image pair(s)	bikes, tree	graffiti, wall	bark, boat	leuven	ubc
SM	89.76	91.24	90.23	95.57	98.83
ACC	85.30	83.28	78.56	93.10	98.47
HV	93.80	92.36	91.00	96.36	98.90
VFC	95.44	90.37	86.58	96.54	98.62
Ours	94.60	93.01	90.16	96.25	98.65

various descriptors.

The four baselines, CAT, CAT+HV, Ranking and Ratio, perform diversely. Baselines CAT and CAT+HV give poor performance. Even their accuracies in mAP fall behind those by the first four image matching algorithms that work with a single descriptor. It reveals that concatenation is not a good strategy for descriptor fusion, because worse descriptors degrade the discriminative power of the concatenated descriptor. Baselines Ranking and Ratio, especially Ratio, lead to much better matching results. The two baselines averagely outperform the four image matching algorithms, but still fall behind them if the best descriptor in each dataset is chosen. For instance, baseline Ratio gives 53.61% in Co-reg and 28.75% in SYM, while baseline ACC with SIFT achieves 60.28% in Co-reg and baseline HV with GB achieves 38.28% in SYM. The results of applying the four baselines, SM, ACC, HV and VFC, to multiple descriptors are not always improved compared to their results using a single descriptor. For example, ACC with SIFT achieves 60.28% on Co-reg dataset while it gives 56% with multiple descriptors. In contrast, our performances are improved more consistently. Our approach allows mutual verification across different descriptors in an unsupervised manner, and correct correspondences will distinguish themselves with high coherence to each other. The quantitative results show that our approach can make the most of fusing various feature descriptors, and achieve significant performance gains over all the baselines on the three datasets. Note that the performance of our method with all the descriptors on VGG dataset decreases a little compared to the one with LIOP. It is because that LIOP has already got very high performance and thus there is not much space for the other descriptors to complement LIOP. This issue will be further analyzed in Section VII-F.

Our approach remarkably outperforms all approaches for comparison on Co-reg and SYM datasets. It shows that our approach can better tackle the variations in the two datasets, including the changes in viewpoints, rotations and scales combined with noises from the clutter backgrounds in Co-reg dataset and the dramatic variations ranging from lighting conditions (day/night), ages (old/nowadays scene) to rendering styles (photograph/drawing) in SYM dataset. In VGG dataset, there is only a single type of variation in each image pair, and totally five types of variations are involved, i.e., blur, viewpoint change, zoom and rotation, light change and JPEG compression. The performance in mAP regarding the five variations by our approach and the four matching methods, all using multiple descriptors, are reported in TABLE III. Our method surpasses the other four methods on dealing with viewpoint change and is comparable with them when handling other variations.

The mAP summarizes the performances of matching approaches on the whole dataset. To look inside how they work on individual images, precision-recall curves (precisely $1-$precision vs. recall curves here) are used. The Co-reg dataset consists of six image pairs. Large changes in viewpoints, rotations and scales combined with noises coming from the clutter backgrounds and occlusions make matching quite difficult on this dataset. The resulting precision-recall curves by all the approaches are shown in Fig. 8(a) $\sim$ 8(f). We show the results of SM, ACC, HV and VFC with a single descriptor and manually pick the best descriptor for them to draw their curves for the sake of clearness. Thus, their performances may be overestimated in this sense. Baselines ACC and HV can deal with multiple object matching, while baseline VFC and SM are less robust in the cases. Ranking and ratio can increase the recall with the aid of multiple descriptors in most cases, but their precision is unsatisfactory. Our method can effectively match multiple objects, and considerably boost both the recall and precision by leveraging multiple descriptors.

We also select three pairs of images from each of the other two datasets, and plot the corresponding precision-recall curves in Fig. 8(g) $\sim$ 8(l), respectively. These two datasets have different types of variations in matching, such as complex changes of illumination and redering styles in dataset SYM, and imaging condition changes in dataset VGG. Our approach can deal with these variations by adaptively picking appropriate descriptors in matching interest points, and result in the superior performance. An exception is shown in Fig. 8(j) leuven, an image pair from dataset VGG. As can be seen in TABLE II, descriptor LIOP achieves satisfactory results, and dominates the other descriptors on dataset VGG. Hence, the performance gain of our approach is not significant in the cases, especially on image pair leuven.

Our method tolerates a certain degree of perspective variations between images. Please see Section VII-F for sensitivity analysis on this issue. Our approach may fail to match repetitive patterns because of the geometric ambiguity, or fail to identify correct correspondences in a minority cluster in the homography space.

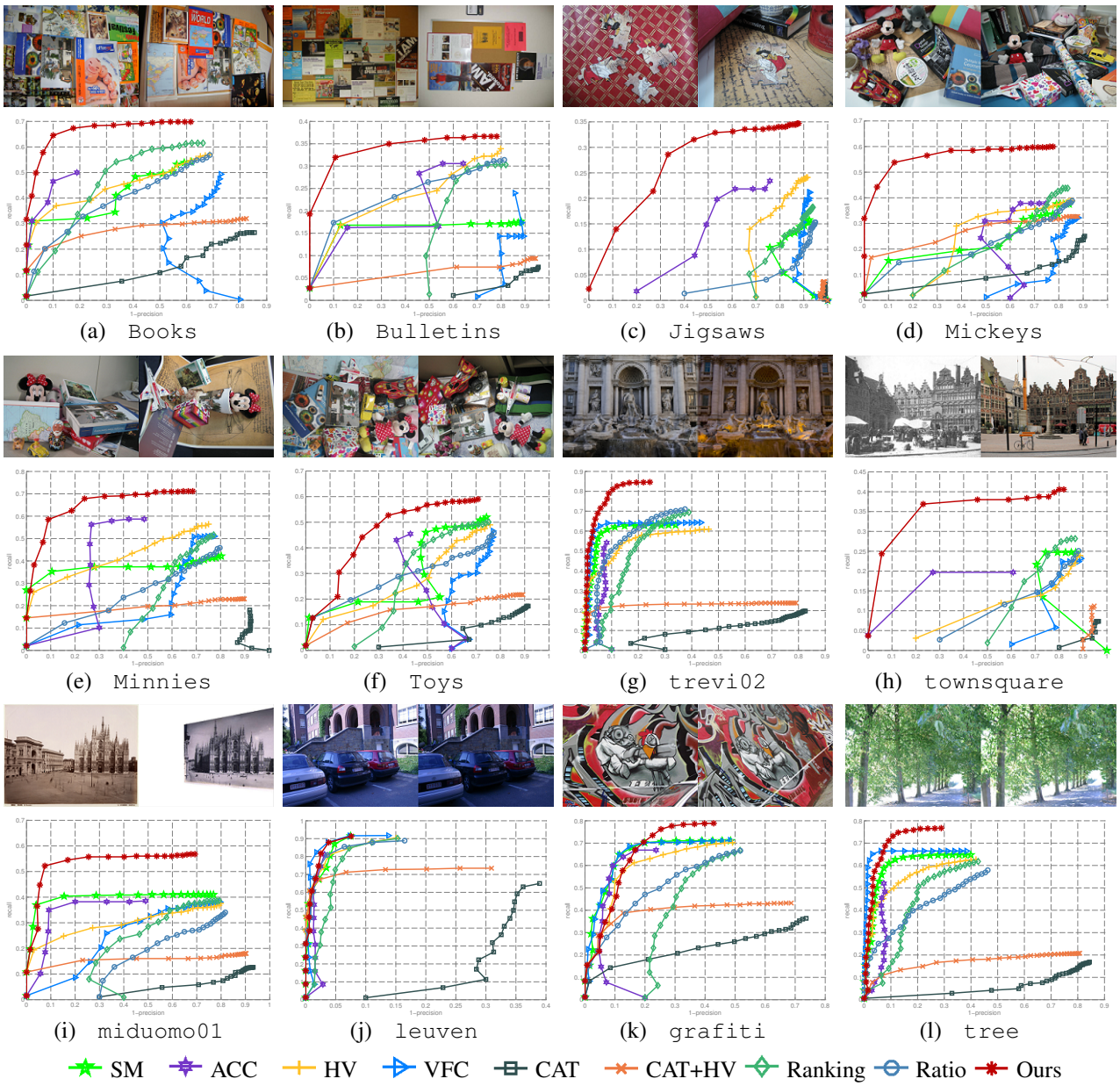


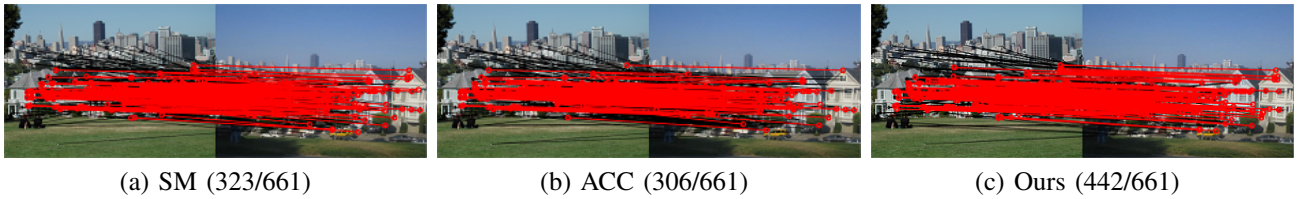
Fig. 8. The precision-recall curves on 12 image pairs. (a) ∼ (f) Six image pairs of the Co-reg dataset. (g) ∼ (i) Three image pairs of the SYM dataset. (j) ∼ (l) Three image pairs of the VGG dataset.

E. Visualization of Matching Results

To gain insight into the quantitative results, we display the matching results by our approach and the adopted baselines. We give two examples from each of the SYM and VGG datasets in Fig. 9 and Fig. 10, respectively.

In Fig. 9, the matching results by our approach and the four image matching algorithms on two image pairs, `paintedladies12` and `sanmarco2`, of the SYM dataset are shown. Our approach yields more dense and accurate matchings (red correspondences), and outperforms the four matching algorithms even if their respective best descriptors on this dataset have been manually chosen. In Fig. 10, our approach and the four baselines for descriptor fusion are compared on two image pairs, `wall` and `grafiti`, of the VGG dataset. Our approach in both cases carries out geometric verification, and effectively reduces the numbers of false positives (black correspondences) yielded by individual descriptors. Therefore, it achieves more satisfactory results.

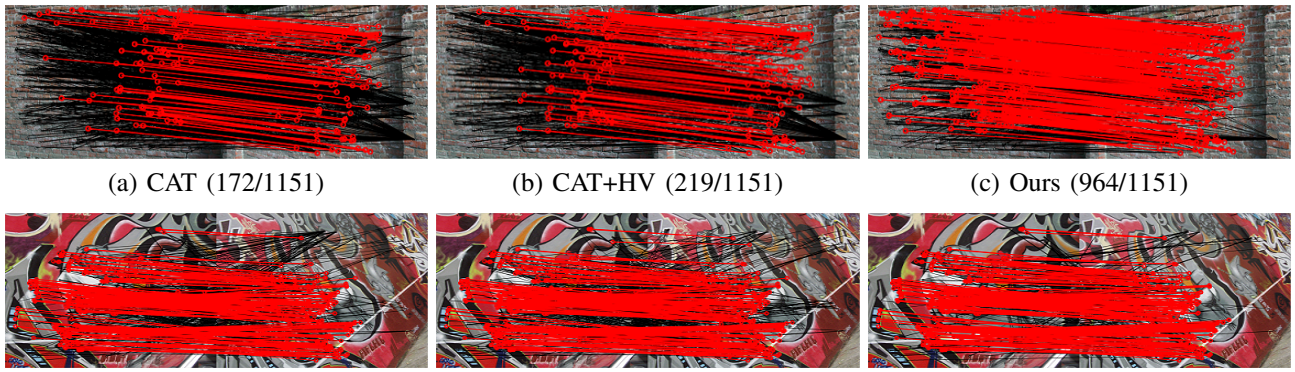
To summarize, the visualization of the matching results demonstrates that our approach can effectively leverage multiple descriptors: On the one hand, it allows geometric layout verification across heterogeneous descriptors, and hence results in higher precision. On the other hand, it increases the number of correct correspondence candidates with the aid of complementary descriptors, and leads to higher recall. These properties enable our approach to alleviate the unfavorable issues in image matching, such as the combined changes of lighting conditions and rendering styles in the SYM dataset, and imaging condition changes in the VGG dataset.



(a) SM (323/661) (b) ACC (306/661) (c) Ours (442/661)

(d) HV(652/1546) (e) VFC (616/1546) (f) Ours (795/1546)

Fig. 9. The matching results by our approach and the four image matching algorithms on two image pairs of the SYM dataset, including (a) $\sim$ (c) image pair `paintedladies12` and (d) $\sim$ (f) image pair `sanmarco2`. The recalls in Eq. (12), namely $n_{\text{TP}}/n_{\text{TP}} + n_{\text{FN}}$, are shown in brackets.



(a) CAT (172/1151) (b) CAT+HV (219/1151) (c) Ours (964/1151)

(d) Ranking (557/821) (e) Ratio (555/821) (f) Ours (657/821)

Fig. 10. The matching results by our approach and the four baselines for descriptor fusion on two image pairs of the VGG dataset, including (a) $\sim$ (c) image pair `wall` and (d) $\sim$ (f) image pair `grafiti`. The recalls in Eq. (12), namely $n_{\text{TP}}/n_{\text{TP}} + n_{\text{FN}}$, are shown in brackets.

F. Analytic Study on the Combinations of Descriptors

To better understand both the number and the type of feature descriptors that the proposed method needs to boost the performance, we present a set of experiments in which we apply our method to various combinations of the five adopted descriptors and analyze the performance with respect to the number and the type of feature descriptors used. Roughly speaking, SIFT, LIOP, and DAYSI are texture-based descriptors, while RI and GB are intensity-based and shape-based descriptors respectively. Texture-based descriptors are useful in Co-reg and VGG datasets owing to the highly textured images. In contrast, shape-based descriptors show high discriminative power on SYM dataset, since the shapes of buildings to be matched are coherent. Note that the ranking of descriptors on each dataset are basically the same with different geometric verification methods.

Due to the large number of possible descriptor combinations, we reduce the number of combinations in the following way. We first sort the five descriptors according to their performance by our method on each dataset as reported in TABLE II. We then add a descriptor at a time sequentially with respect to the sorted descriptor order. The performance of the descriptors is sorted both descendingly and ascendingly. Thus, the number of combinations is reduced to five for each order, and the total number is ten for each dataset.

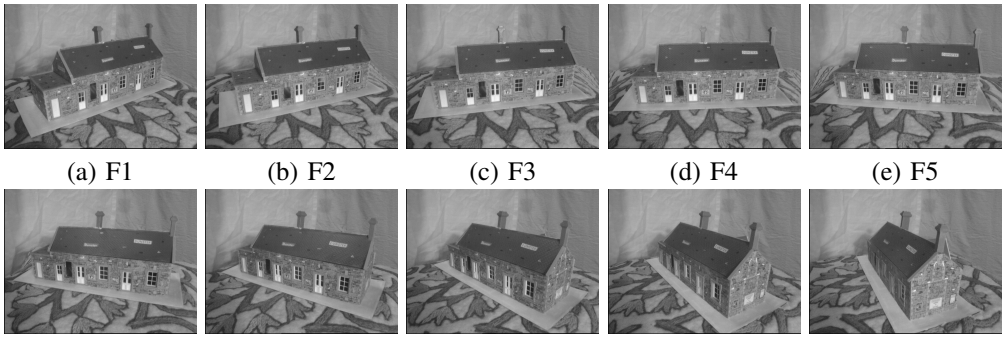
TABLE IV
THE ACCURACY IN MAP OF OUR APPROACH WITH DIFFERENT COMBINATIONS OF DESCRIPTORS ON CO-REG DATASET

Ascending	GB	+RI	+LIOP	+DAISY	+SIFT
mAP (%)	10.80	36.77	58.61	69.69	79.59
Gain (%)		+25.97	+21.84	+11.08	+9.90
Descending	SIFT	+DAISY	+LIOP	+RI	+GB
mAP (%)	72.03	78.32	79.89	80.25	79.59
Gain (%)		+6.31	+1.57	+0.36	-0.66

TABLE V
THE ACCURACY IN MAP OF OUR APPROACH WITH DIFFERENT COMBINATIONS OF DESCRIPTORS ON SYM DATASET

Ascending	RI	+LIOP	+SIFT	+DAISY	+GB
mAP (%)	11.38	22.92	31.16	36.20	46.83
Gain (%)		+11.44	+8.24	+5.04	+10.63
Descending	GB	+DAISY	+SFIT	+LIOP	+RI
mAP (%)	40.65	47.95	49.15	48.83	46.83
Gain (%)		+7.30	+1.25	-0.32	-2.00

For instance, the performance ranks of the five descriptors on Co-reg dataset from high to low are SIFT, DAISY,



(a) F1 (b) F2 (c) F3 (d) F4 (e) F5

(f) F6 (g) F7 (h) F8 (i) F9 (j) F10

Fig. 11. The ten frames in VGG model house dataset. (a) $\sim$ (j) frame 1 (F1) $\sim$ frame 10 (F10).

TABLE VI
THE ACCURACY IN MAP OF OUR APPROACH WITH DIFFERENT COMBINATIONS OF DESCRIPTORS ON VGG DATASET

Ascending	RI	+GB	+SIFT	+DAISY	+LIOP
mAP (%)	68.57	84.02	91.30	93.14	93.81
Gain (%)		+15.45	+6.28	+1.84	+0.67
Descending	**LIOP**	**+DAISY**	**+SFIT**	**+GB**	**+RI**
mAP (%)	94.23	94.13	93.71	93.67	93.81
Gain (%)		-0.10	-0.42	-0.04	+0.14

LIOP, RI and then GB, respectively. By adding in an descending direction, we test the five combinations, i.e., SIFT, SIFT+DAISY, SIFT+DAISY+LIOP, SIFT+DAISY+LIOP+RI and SIFT+DAISY+LIOP+RI+GB. The other five combinations in the ascending order are similarly given. The results of the combinations on Co-reg dataset, SYM dataset and VGG dataset are reported in TABLE IV, TABLE V and TABLE VI, respectively.

Adding descriptors in the ascending order leads to monotonic performance increases on all the three datasets. It can be understood that adding good descriptors tends to increase the performance of our approach. Note that the performance gains after adding the best descriptors, SIFT and GB, on Co-reg dataset and SYM dataset respectively are still remarkable even if there are already four employed descriptors. While adding the best descriptor, LIOP, doesn't lead to such a notable improvement on VGG dataset as the previous two. The reason is that the correlation between LIOP and DAISY, the second best descriptor on VGG dataset, is high. The candidate sets established by two descriptors do not complement each other.

In contrast, adding descriptors in the descending order does not guarantee to increase the performance as the successively joined descriptors perform worse and worse in the single descriptor cases. However, the overall performance of fusing all the five descriptors is better than the best descriptor on Co-reg and SYM datasets, and only slightly falls behind LIOP on VGG dataset. It indicates that our approach can leverages multiple, complementary to boost the performance of feature matching.

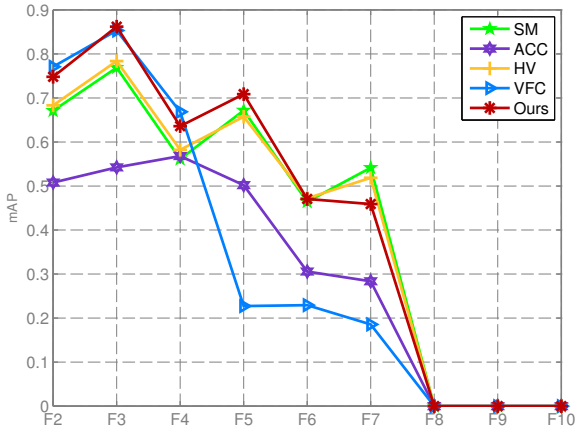


Fig. 12. The performance (in mAP) of matching frame 1 to the other 9 frames (F2 $\sim$ F10) of five matching algorithms on the VGG model house dataset.

G. Sensitivity Analysis of Perspective Variations

Affine homography is adopted in our work. We present a set of experiments to evaluate the proposed algorithm on matching images with perspective variations. VGG model house dataset is used in the experiments. The dataset as well as the detected interest points and the ground truth matchings are available at http://www.robots.ox.ac.uk/~vgg/data1.html. It contains 10 frames of a model house in a moving sequence. The images of the dataset are shown in Fig. 11. We match frame 1 to the other 9 frames to investigate the performance with different degrees of perspective change. Two detached frames in the sequence have a more dramatic perspective change than adjacent frames. The matching performance in mAP of our approach and the four matching algorithms, i.e., SM, ACC, HV and VFC, is reported in Fig. 12.

Though all the five matching methods use affine homography to characterize correspondences, they are tolerant of perspective variations to some degree. The performance of VFC and our algorithm on matching frame 1 to frames $2 \sim 4$ outperform the others, and our method can still match frame 1 to frames with larger perspective changes, such as frames 5 and 6, with higher performance compared to the others. All methods fail in matching frame 1 to frames $8 \sim 10$. The number of the corresponding points becomes less and less, because the detector cannot handle the drastic perspective variations.

VIII. Conclusion

We have presented an effective approach to matching images with multiple descriptors. The correspondences yielded by all descriptors are firstly projected into the homography space, in which both geometric and spatial consistence among them are measured by computing geodesic distances on a designed graph. One-class SVM is then employed to rank the correspondences according to their consensus with each other. The proposed approach is featured with high flexibility in the sense that it can work with any elliptical region detectors as well as heterogeneous descriptors. Besides, it selects good correspondences across descriptors in a fully unsupervised way. No prior knowledge about images to be matched is required. Our approach is comprehensively compared with the state-of-the-art algorithms and is evaluated on five benchmark datasets. The experimental results demonstrate that it can significantly boost the matching quality in both precision and recall. In the future, we will generalize and apply this work to computer vision applications where accurate and dense matchings are appreciated, such as image alignment, object recognition, and motion estimation.

References

[1] R. Szeliski and H.-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 1997.

[2] C. Olson and D. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *IEEE Trans. on Image Processing*, vol. 6, no. 1, pp. 103–113, Jan 1997.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideals, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, 2008.

[5] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.

[6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[8] A. Berg and J. Malik, "Geometric blur for template matching," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.

[9] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Euro. Conf. Computer Vision*, 2006.

[11] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2010.

[12] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," in *Proc. Int' Conf. Computer Vision*, 2011.

[13] D. Hauagge and N. Snavely, "Image matching using local symmetry features," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[14] T. F. Cox and M. A. A. Cox, *Multidimentional Scaling*. Chapman & Hall, London, 1994.

[15] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[16] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–534, 1997.

[17] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2006.

[18] X. Bai, C. Rao, and X. Wang, "Shape vocabulary: A robust and efficient shape representation for shape matching," *IEEE Trans. on Image Processing*, vol. 23, no. 9, pp. 3935–3949, Sept 2014.

[19] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[20] E. Lobaton, R. Vasudevan, R. Alterovitz, and R. Bajcsy, "Robust topological features for deformation invariant image matching," in *Proc. Int' Conf. Computer Vision*, 2011.

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

[22] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.

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

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

[25] S. Pang, J. Xue, Q. Tian, and N. Zheng, "Exploiting local linear geometric structure for identifying correct mtches," *Computer Vision and Image Understanding*, vol. 128, no. 0, pp. 51–64, 2014.

[26] J. Ma, W. Qiu, J. Zhao, Y. Ma, A. Yuille, and Z. Tu, "Robust l2e estimation of transformation for non-rigid registration," *IEEE Trans. on Signal Processing*, vol. 63, no. 5, pp. 1115–1129, 2015.

[27] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. Int' Conf. Computer Vision*, 2005.

[28] Y. Zheng and D. Doermann, "Robust point matching for nonrigid shapes by preserving local neighborhood structures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 643–649, 2006.

[29] O. Choi and I. Kweon, "Robust feature point matching by preserving local geometric consistency," *Computer Vision and Image Understanding*, vol. 113, no. 6, pp. 726–742, 2009.

[30] T. Cour, P. Srinivasan, and J. Shi, "Balanced graph matching," in *Proc. Advances in Neural Information Processing Systems*, 2006.

[31] M. Cho, J. Lee, and K.-M. Lee, "Reweighted random walks for graph matching," in *Proc. Euro. Conf. Computer Vision*, 2010.

[32] M. Leordeanu, R. Sukthankar, and M. Hebert, "Unsupervised learning for graph matching," *Int. J. Computer Vision*, vol. 96, no. 1, pp. 28–45, 2012.

[33] H. Li, X. Huang, and L. He, "Object matching using a locally affine invariant and linear programming techniques," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 411–424, 2013.

[34] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *Int. J. of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 03, pp. 265–298, 2004.

[35] C. Wang, L. Wang, and L. Liu, "Improving graph matching via density maximization," in *Proc. Int' Conf. Computer Vision*, 2013.

[36] M. Cho, J. Lee, and K.-M. Lee, "Feature correspondence and deformable object matching via agglomerative correspondence clustering," in *Proc. Int' Conf. Computer Vision*, 2009.

[37] W. Zhang, X. Wang, D. Zhao, and X. Tang, "Graph degree linkage: Agglomerative clustering on a directed graph," in *Proc. Euro. Conf. Computer Vision*, 2012.

[38] Y. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *Int. J. Computer Vision*, vol. 107, no. 1, pp. 1–19, March 2014.

[39] H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, "Robust feature matching with alternate Hough and inverted Hough transforms," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[40] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM Conf. Image and Video Retrieval*, 2007.

[41] E. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2005.

[42] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[43] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Improving local descriptors by embedding global and local spatial information," in *Proc. Euro. Conf. Computer Vision*, 2010.

[44] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2011.
[45] X. Bai, B. Wang, C. Yao, W. Liu, and Z. Tu, "Co-transduction for shape retrieval," *IEEE Trans. on Image Processing*, vol. 21, no. 5, pp. 2747–2757, 2012.
[46] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proc. Int' Conf. Computer Vision*, 2013.
[47] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Robust image alignment with multiple feature descriptors and matching-guided neighborhoods," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.
[48] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.
[49] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," in *Proc. Euro. Conf. Computer Vision*, 2004.
[50] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[51] M. Cho, Y.-M. Shin, and K.-M. Lee, "Co-recognition of image pairs by data-driven monte carlo image exploration," in *Proc. Euro. Conf. Computer Vision*, 2008.
[52] V. M. Govindu, "Lie-algebraic averaging for globally consistent motion estimation," in *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, 2004, pp. 684–691.

Yuan-Ting Hu received her bachelor's degree and master's degree in computer science and information engineering from National Taiwan University in 2012 and 2014, respectively. She was a research assistant in Academia Sinica, Taiwan from 2013 to 2015. Her research interests include computer vision, image processing and machine learning.

Yen-Yu Lin received the B.B.A. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His current research interests include computer vision, pattern recognition, and machine learning. He is a member of the IEEE.

Hsin-Yi Chen received her B.B.A. degree in business administration from National Taiwan University,and her M.S. degree in computer science and information engineering from National Taiwan University, where she is currently a Ph.D. candidate in the same program. She was a research assistant in Academia Sinica, Taiwan from 2012 to 2014. Her current research interests include computer vision, computer graphics and image processing.

Kuang-Jui Hsu received the B.S. in the Department of Electrical Engineering from National Sun Yat-sen University in 2011 and the M.S. in the Graduate Institute of Networking and Multimedia from National Taiwan University in 2013. He is currently a research assistant at the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. His research interests include computer vision, machine learning and image processing.

Bing-Yu Chen received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, in 1995 and 1997, respectively, and the Ph.D. degree in information science from The University of Tokyo, Japan, in 2003. He is currently a professor with National Taiwan University. He was a Visiting Researcher and Professor at The University of Tokyo during 2008 to 2012. His current research interests include computer graphics, image and video processing, and human-computer interaction. He is a senior member of ACM and a member of Eurographics.