Common Visual Pattern Discovery via Nonlinear Mean Shift Clustering

Linbo Wang, Dong Tang, Yanwen Guo, and Minh N. Do, *Fellow, IEEE*

***Abstract*— Discovering common visual patterns (CVPs) from two images is a challenging task due to the geometric and photometric deformations as well as noises and clutters. The problem is generally boiled down to recovering correspondences of local invariant features, and the conventionally addressed by graph-based quadratic optimization approaches, which often suffer from high computational cost. In this paper, we propose an efficient approach by viewing the problem from a novel perspective. In particular, we consider each CVP as a common object in two images with a group of coherently deformed local regions. A geometric space with matrix Lie group structure is constructed by stacking up transformations estimated from initially appearance-matched local interest region pairs. This is followed by a mean shift clustering stage to group together those close transformations in the space. Joining regions associated with transformations of the same group together within each input image forms two large regions sharing similar geometric configuration, which naturally leads to a CVP. To account for the non-Euclidean nature of the matrix Lie group, mean shift vectors are derived in the corresponding Lie algebra vector space with a newly provided effective distance measure. Extensive experiments on single and multiple common object discovery tasks as well as near-duplicate image retrieval verify the robustness and efficiency of the proposed approach.**

***Index Terms*— Common pattern discovery, local affine region, mean-shift clustering, near-duplicate image retrieval.**

Manuscript received July 30, 2014; revised April 18, 2015; accepted September 10, 2015. Date of publication September 23, 2015; date of current version October 7, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61373059, Grant 61321491, and Grant 61502005, and in part by the Open Foundation through the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, under Grant KFKT2015B03. The work of L. Wang was supported by the High-Level Talents Program through Anhui University, China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Adrian G. Bors. *(Corresponding author: Yanwen Guo.)*

L. Wang is with the Key Laboratory of Intelligent Computing and Signal Processing, School of Computer Science and Technology, Ministry of Education, Anhui University, Hefei 230039, China, and also with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: wanglb.2005@gmail.com).

D. Tang is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: tangdong408.cn@gmail.com).

Y. Guo is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China, also with the University of Illinois at Urbana–Champaign, Urbana, IL 61820-5711 USA, and also with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: ywguo.nju@gmail.com).

M. N. Do is with the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: minhdo@illinois.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2481701

I. Introduction

With the rapid development of photographing devices, such as digital cameras, cellular phones, etc., a huge amount of images have been generated nowadays. It poses great challenges to the analysis of the underlying information for computer vision and multimedia communities. One of the challenges is to automatically discover common visual patterns in a given image set, which is very promising and beneficial for various applications, such as object recognition [1], near-duplicate image retrieval [2], image database browsing [3], video object co-segmentation [4] and image co-matting [5], etc.

Given a pair of images, a common visual pattern (CVP) depicts a group of local image regions with similar visual appearance as well as consistent spatial layout. Without other prior knowledge, identifying CVPs is generally infeasible since there is no appropriate way to represent and compare the CVPs in different images. Therefore, the task is often addressed by leveraging the discriminative power of local invariant features [6]–[8] to recover and group together feature correspondences, and to further derive the underlying CVPs.

Despite the noticeable efforts devoted to consistently improving the discriminative power of local features so far, robustly detecting CVPs in real-world images remains challenging. The challenges mainly come from several aspects. First, instances of a CVP on different images may exhibit different scales, orientations and viewpoints. Second, photometric and geometric deformations may exist across images. Third, apart from CVP regions, images usually contain significant clutters and noises, which often interfere with the discovery process. The first row of Fig. 1 shows an example. We can see that accurately identifying all the CVPs is not a trivial task.

Feature matching techniques are the potential solutions for the CVP discovery problem. They aim at identifying correct local feature correspondences between input images. Recently, many graph-based feature matching algorithms [9]–[12] have been proposed. The general routine is to first establish a feature correspondence candidate set, then construct a matching graph with nodes representing appearance similarity and edges encoding geometric compatibility among the candidate correspondence pairs, and finally derive solutions by maximizing the inherent energy function (or solving its dual form). Although promising performance has been demonstrated, such global quadratic optimization based approaches generally suffer from high computational cost, hence cannot handle time-intensive tasks, e.g. large scale near-duplicate

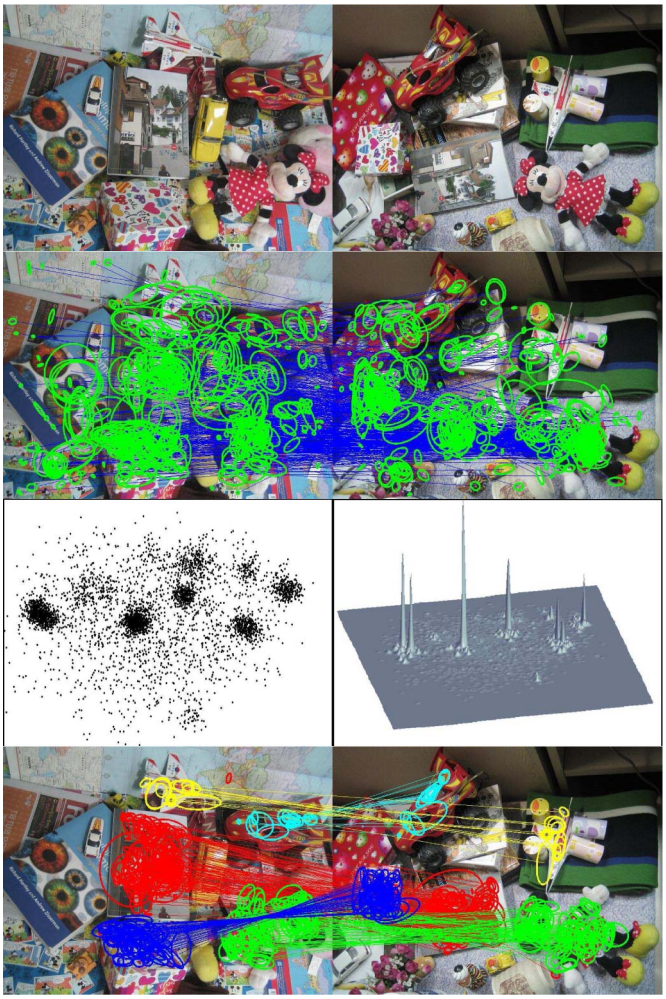

Fig. 1. Five CVPs between two images are automatically detected using our approach. Local affine regions in the input images (1st row) are extracted and matched based on SIFT features (2nd row). For better viewing experience, we only show 300 randomly selected pairs, which still contain significant distracting outliers. The 3rd row shows 2D projection and associated 3D density plot of the transformation space. Our approach identifies all the CVPs faithfully (4th row), despite the heavy noises as well as deformations on certain objects.

image retrieval. On the other hand, heuristic CVP discovery approaches [2], [13] proceed by gradually selecting good feature correspondences from the candidate set according to certain geometric criteria, however may fail to achieve optimum solutions. Besides, only a small fraction of existing approaches have the ability to handle a pair of images with multiple CVPs.

In this paper, we consider the task of CVP discovery from a new perspective. We treat each CVP as two groups of local interest regions, which are coherently projected between two images. Geometric transformations derived from different aligned local region pairs are not necessarily identical due to deformations, but are expected to be similar to each other. This means their transformation coordinates tend to be close enough, revealing a dense distribution in the underlying transformation space. Hence, identifying CVPs can be achieved via a transformation clustering procedure using mode finding in a properly constructed transformation space.

To this end, we propose a nonlinear mean shift algorithm for estimating the modes in the transformation domain, which is constructed by stacking up 2D similarity transformations derived from geometric parameters of matched SIFT features on the input images. Compared with the conventional mean shift which requires a vector space to function properly, nonlinear mean shift has proved to be effective for data points on Riemannian manifolds [14], [15], e.g. matrix Lie groups, where 2D similarity transformations lie. Note that as we show later, when local affine regions are described with SIFT descriptor, each of the similarity transformations computed corresponds to an affine projection between a local interest region pair. Since local regions are approximately planar and/or rigid, we assume a pair of them can be affinely mapped to each other, which does not deny the global deformation across the whole CVP body. Besides, this assumption is also admitted and experimentally proved to be effective in many existing literatures [11]–[13].

The key challenges on applying nonlinear mean shift lie in determining a specific form of mean shift vectors and a proper distance measure for updating the mean shift vectors. For the similarity transformation group, we show that a light-weighted form of mean shift vector can be derived in its corresponding vector space, Lie algebra. It considerably reduces the computational overhead by avoiding the expensive matrix exponential and logarithm operations, which are usually indispensable to achieve transformation between elements in the matrix Lie group and Lie algebra. On the other hand, we provide an effective distance measure with meaningful justification, which better guarantees the convergence property of the mean shift procedure under challenging conditions. As illustrated by Fig. 1, despite heavy noises introduced by incorrect initial correspondences (2nd row), multiple CVPs are successfully detected (4th row) by capturing the dense distribution attached to different modes in the transformation domain (3rd row).

To summarize, we present a new solution for the task of CVP discovery and the contributions are as follows:

1) Unlike existing optimization-based approaches which detect CVPs by feature matching with a discretized quadratic optimization formulation encoding geometric compatibility among candidate feature matches, we identify and group together correct feature correspondences of different CVPs by identifying their geometric consistency with a mean shift clustering procedure in the underlying geometric transformation space, bringing better convergence property compared with existing heuristic approaches. Therefore, it leads to higher efficiency and better performance in the task of CVP discovery under various conditions, e.g., heavy feature correspondence noise and object deformation, etc., as well as an extended application, near-duplicate image retrieval.
2) A similarity transformation space is constructed by stacking up geometric transformations estimated from SIFT-matched local interest regions. The matrix Lie group structure of the space is discussed for the designation of the mean shift clustering algorithm, and an efficient way for evaluating mean shift vectors in the matrix Lie algebra is presented.

3) Measures for evaluating the distance between two similarity matrices during mean shift iteration are investigated and a new distance measure is provided for better performance.

The rest of paper is organized as follows. We first elaborate on how to construct the transformation space based on the initial correspondences of local affine regions in Section II. This is followed by a description of the general nonlinear mean shift clustering algorithm on Riemannian manifold and matrix Lie groups in Section III, which is then extended to the CVP discovery task in Section IV. Extensive experiments are presented in Section V. Finally, we conclude the paper in Section VI.

II. Transformation Space Construction

This section presents the method of generating candidate transformations, each of which depicts the geometric relation between a pair of local features on the two input images. The transformations are later fed into the proposed clustering algorithm and CVPs are then identified by the similar transformations grouped together. To start off with, we extract local interest regions and describe them with the SIFT descriptor. A matching process is followed to generate candidate correspondence sets. The second step computes the geometric transformation between each matched pair.

A. Local Interest Region Extraction and Matching

Local interest region detection and description have been studied for many years and significant progress has been made so far. Here, we adopt multiple popular scale and affine invariant region detectors, including DoG [8], MSER [7], Harris and Hessian-affine [6]. Output of these detectors is a set of geometric invariant parameter tuples, with each tuple corresponding to a circular region (for scale invariant region detectors) or an elliptical region (for affine invariant region detectors). We then generate feature descriptions for these regions with the widely-used SIFT descriptor, and further establish initial region correspondences between two images by comparing feature distance. Despite the significant amount of false matches, correct region correspondences are generally preserved in the correspondence set thanks to the discriminative power of SIFT.

B. Transformation Computation

Once the initial correspondences are established, we turn to estimate the geometric transformation between each pair of the corresponding regions. As aforementioned, we assume here local region projections can be approximated with 2D affine transformations. Traditionally they are computed from all triple pairs of matched region centers, which is time-consuming. Instead, we exploit the geometric invariant parameters of SIFT features to estimate a 2D similarity transformation between each feature pair. The transformation can then be easily converted to the affine transformation between the two associated local affine regions as we show below.

Recall that a SIFT feature is essentially a gradient histogram of a circular local region with three parameters, including the centroid coordinates, radius and dominant angle. They depict the position, scale and gradient orientation of an object part respectively. While the feature itself is appearance-dependent, it fully relies on the three parameters to achieve geometric invariance. Henceforth, matching of a pair of SIFT features intrinsically admits the geometric correspondence between the underlying circular region pair, whereby a similarity transformation between the two local regions can be obtained.

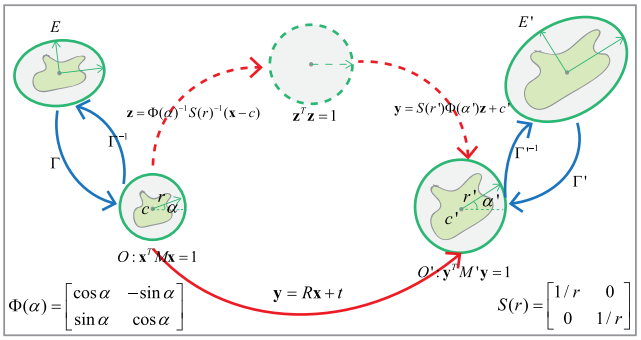

Fig. 2. Illustration of the transformation computation. E and E' are two ellipses generated by a local affine region detector. They are adapted to circles O and O' with Γ and Γ', respectively. c, r, α, M denote the centroid, scale, gradient orientation and quadratic form matrix of O. c', r', α', M' is similarly defined for O'. The transformation from O to O' is equivalent to converting the α-oriented O to a zero-oriented unit circle and further to the α'-oriented O' as illustrated above, thereby deriving the corresponding rotation and scaling matrix R and and translation t. Note that the subtitle i in the text is ignored for the sake of simplicity.

Let O_i, O_i' be two circles associated with a pair of matched local features. O_i is represented by three geometric parameters c_i, r_i, α_i, corresponding to the centroid coordinates, radius and orientation respectively. Note that here α_i is a crucial parameter for establishing the geometric correspondence since it records the gradient orientation of an object part. Similarly, parameters c_i', r_i', α_i' are defined for O_i'. As illustrated by Fig. 2, the transformation mapping each point on O_i to its counterpart on O_i' can then be expressed as

$$X_i = \begin{bmatrix} R_i & t_i \\ \mathbf{0} & 1 \end{bmatrix}, \tag{1}$$

where

$$R_i = \begin{bmatrix} s_i & 0 \\ 0 & s_i \end{bmatrix} \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix}, \quad s_i = \frac{r_i}{r_i'}, \quad \theta_i = \alpha_i' - \alpha_i, \tag{2}$$

and

$$t_i = c_i' - R_i c_i. \tag{3}$$

Basically, in the similarity transformation matrix X_i, R_i encodes the scaling and rotation transformations determined by the radius ratio s_i and orientation difference θ_i, while t_i denotes translation. The entire transformation process can be understood as converting an oriented circle to a unit one with zero orientation, and further to a new one with different geometric parameters.

Estimation of X_i above is solely built upon the circular geometric parameters of O_i and O_i', which are exactly the basic requirements for computing the corresponding

SIFT descriptors. Moreover, O_i is obtained either directly from the outputs of a scale invariant region detector, e.g., DoG [8], or by shape adaption from an elliptical region extracted with an affine invariant region detector, e.g. Harris-affine [6]. In the latter case, assuming E_i is the ellipse O_i adapted from, then E_i can be converted to O_i via an affine projection Γ_i, which is illustrated by Fig. 2. Likewise, for O_i' and its elliptical counterpart E_i', we can define an affine transformation Γ_i' mapping E_i' to O_i'. Consequently, the affine projection mapping E_i to E_i' is computed as $T_i = {\Gamma_i}'^{-1} X_i \Gamma_i$.

The above deduction suggests that the affine transformation T_i between two local affine regions can be estimated from the similarity transformation X_i. However, the actual computation of T_i is not necessary here considering threefold. First, the validity of T_i is guaranteed by the correctness of X_i. Second, the computation of X_i is much easier than T_i. Third, the relationships of X_i and s_i, θ_i greatly facilitate the conversion between the space of X_i and its algebra as shown in Section IV-A, which improves the clustering efficiency considerably.

For each region pair in the initial correspondence set, we compute its transformation X_i as described above. All the transformations are then stacked together, forming a similarity transformation space. Note that this is not a vector space endowed with the Euclidean distance metric. Given two arbitrary transformations X_1 and X_2, the matrix addition $X_1 + X_2$ is not really meaningful as the matrix multiplication $X_1 X_2$, which means to apply X_2 and X_1 successively.

III. Mean Shift Clustering in the Transformation Space

Empirically, in the transformation space constructed above, transformations of false correspondence tend to stay far away from the others while transformations of true correspondences reside closely to each other, shaping multiple clusters with dense distributions. Moreover, each cluster encloses similar transformations linking pairs of local regions in the two input images. Joining these regions together in each image respectively leads to two pieces of large regions with similar geometric configuration, which naturally reveals a CVP. Thus the CVP discovery task can be achieved by performing a clustering procedure in the transformation space.

To this end, we choose the mean shift algorithm as our basic clustering scheme. The reason is that, parametric clustering approaches, e.g. k-means, seek to partition all the transformations into pre-specified number of clusters. This could be problematic considering that the number of clusters is not known a priori and the transformation set is likely to be very noisy due to false correspondences. By contrast, as a well-known non-parametric clustering algorithm, mean shift finds the modes of underlying distribution of the transformation space and performs clustering by grouping together all the transformations governed by the same mode. Hence it exhibits more flexibility as well as robustness to noise, and better suits our needs.

Unfortunately, the original mean shift algorithm is not directly applicable to this setting because of the non-Euclidean nature of the constructed transformation space. To be more specific, the transformation space of X_i is a standard geometric space, known as the Similarity group [14], denoted by Sim(2). It is one kind of matrix Lie groups, which are matrix groups equipped with the structure of analytic Riemannian manifold. This restricts the applicability of the conventional mean shift algorithm, which is essentially designed for clustering tasks in vector spaces. Nevertheless, a nonlinear mean shift technique has been proposed to cluster data points lying on matrix Lie groups in [14] and [15]. This forms the theoretical foundation of our CVP discovery algorithm. We now briefly review the nonlinear mean shift algorithm for the general Riemannian manifold and its extension for matrix Lie group, which is further customized to identify CVPs in the following section.

A. Mean Shift Over Riemannian Manifold

Data points of a matrix Lie group lie on a Riemannian manifold, which is a real differentiable manifold with its tangent space at each point being an Euclidean vector space. The nonlinear mean shift algorithm for matrix Lie groups is firstly proposed in [15]. It is later extended to all Riemannian manifolds, and successfully applied to various tasks, e.g. camera pose segmentation, diffusion tensor-MRI (Magnetic Resonance Imaging) filtering, etc. Next, we outline this general nonlinear mean shift algorithm which will lead to our customized version for the transformation space of X_i. More theoretical details can be found in [14] and [15].

Given a Riemannian manifold, the key idea of nonlinear mean shift algorithm is to seek the modes of the underlying distribution by deriving the mean shift vectors in its tangent space with a properly defined distance measure. Let $\mathbf{x}_i, i = 1, \ldots, n$ be n data points on a Riemannian manifold. Given a function $d(\cdot, \cdot)$ computing the distance between two points on the manifold surface, the density at the point $\mathbf{x}$ on the manifold can then be estimated with profile k and bandwidth h as

$$\hat{f}_{h,k}(\mathbf{x}) = \frac{c_{h,k}}{n} \sum_{i=1}^{n} k\left(\frac{d^2(\mathbf{x}, \mathbf{x}_i)}{h^2}\right), \tag{4}$$

where the constant $c_{h,k}$ is chosen to ensure the integral value of $\hat{f}$ equals 1. Taking the gradient of (4) and following the similar deduction routine as [16], the non-Euclidean mean shift vector can be expressed as

$$m_{h,g}(\mathbf{x}) = \frac{-\sum_{i=1}^{n} \nabla d^2(\mathbf{x}, \mathbf{x}_i) g\left(\frac{d^2(\mathbf{x}, \mathbf{x}_i)}{h^2}\right)}{\sum_{i=1}^{n} g\left(\frac{d^2(\mathbf{x}, \mathbf{x}_i)}{h^2}\right)}, \tag{5}$$

where $g(x) = -k'(x)$. The gradient term $\nabla d^2(\mathbf{x}, \mathbf{x}_i)$ lies in the tangent space to point $\mathbf{x}$ on the manifold. This makes the mean shift vector a weighted sum of the tangent vectors since the kernel terms $g\left(d^2(\mathbf{x}, \mathbf{x}_i)/h^2\right)$ are all scalars. Moreover, similar the standard mean shift algorithm, $m_{h,g}(\mathbf{x})$ is proportional to the normalized density gradient estimate of point $\mathbf{x}$, and it always points towards the direction of the maximum increase of the density, which guarantees that moving along the mean

shift vectors converges to the stationary points, i.e., local modes of the underlying density distribution.

The above deduction holds for the general Riemannian manifolds. Generalizing it to matrix Lie groups and further Sim(2) requires evaluation of a proper distance measure $d(\cdot,\cdot)$ as well as the specific form of mean shift vectors in the tangent space.

B. Mean Shift on Matrix Lie Groups

A matrix Lie group (denoted by G) is endowed with an Euclidean tangent space at each of its elements. In particular, the tangent space at the identity matrix e of G is termed Lie algebra (denoted by $\mathfrak{g}$). Elements in G and $\mathfrak{g}$ can be mapped to each other via matrix exponential and matrix logarithm [17]. Specifically, let $X \in G$ and $x \in \mathfrak{g}$ be a pair of corresponding matrices, then

$$X = \exp(x) = \sum_{j=0}^{\infty} \frac{1}{j!} x^j, \tag{6}$$

$$x = \log(X) = \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j} (X-e)^j. \tag{7}$$

We shall see later when G is Sim(2), $\mathfrak{g}$ is a vector space with its elements encoding the transformations of scaling, rotation, and translation.

The distance between two points on a manifold can be determined by the geodesic distance. A geodesic is a locally length-minimizing curve, and thus it measures the locally shortest path between points on the manifold. The length of geodesic is geodesic distance, and it can be mathematically defined by an integral over norms of tangent vectors on the geodesic [18]. Therefore, according to the mapping in (7), the distance between X and e can be evaluated by

$$d(X,e) = ||x||_H = ||\log(X)||_H. \tag{8}$$

Here, the norm is defined in the Lie algebra as

$$||x||_H^2 = \vec{x}^T H \vec{x}, \tag{9}$$

where $\vec{x}$ is the vector form of matrix x in $\mathfrak{g}$, and H is a positive definite matrix defining an inner product on the vector space of $\mathfrak{g}$. Note that the definition of H determines the distance measure and should be designed based on the characteristics of $\mathfrak{g}$. We will introduce a specific form of H specifically tailored to the Lie algebra of Sim(2) in Section IV.

To measure the distance between two arbitrary matrices X and Y in G, it is important to know that the mapping between G and $\mathfrak{g}$ is continuous and only one-to-one near the identity element of G, i.e., the exponential map is a local diffeomorphism between the neighborhood of $\mathbf{0} \in \mathfrak{g}$ and the neighborhood of $e \in G$. Moreover, the logarithm map can be defined only on the neighborhood of e, otherwise the series fails to converge. Therefore, to compute the distance between X and Y, an isomorphism mapping should be defined to project the point X to e and other points to neighbors of e. The mapping function can be right (or left) multiplication by the matrix inverse X^{-1}: $G \to G$. Suppose right multiplication is taken, then the distance is computed as

$$d(X,Y) = ||\log(YX^{-1})||_H. \tag{10}$$

Accordingly, the gradient term $\nabla d^2(\mathbf{x}, \mathbf{x}_i)$ in (5) can be approximated by the elements in the Lie algebra using

$$\nabla d^2(X,Y) = -\log(YX^{-1}). \tag{11}$$

With (10) and (11), the mean shift vector at the current center X can be estimated as

$$m_{h,g}(x) = \frac{\sum_{i=1}^{n} g\left(\frac{||x_i||_H^2}{h^2}\right) x_i}{\sum_{i=1}^{n} g\left(\frac{||x_i||_H^2}{h^2}\right)}, \tag{12}$$

where

$$x_i = \log(X_i X^{-1}). \tag{13}$$

is a vector in the Lie algebra encoding the transformed difference between X and X_i. Hence in principle $m_{h,g}(x)$ is completely evaluated in the Lie algebra.

With the computed mean shift vector, the center of local neighborhood is updated as

$$X^{(j+1)} = \exp\left(m_{h,g}(x^{(j)})\right) X^{(j)}, \tag{14}$$

where centroid of the current local window $X^{(j)}$ is moving along the geodesic defined by the mean shift vector and approaches the next estimate, $X^{(j+1)}$, until a local mode is reached, thereby determining the cluster assignment for each point.

IV. Common Visual Pattern Discovery

As stated earlier, the CVP discovery is equivalent to executing a mean shift clustering process in the similarity transformation space. Although this could be achieved by the nonlinear mean shift clustering technique elaborated above, a key problem remaining unsolved is to design a distance measure that can faithfully quantify the difference between a similarity matrix X_i and the current center X during each iteration of mean shift. Specifically, it is to define H in (12) for deriving the norm of $x_i = \log(X_i X^{-1})$ in the Lie algebra of Sim(2). Since H measures the distance by combining all the components constituting x_i, determining H needs to firstly clarify the components of x_i. Through analyzing the matrix logarithm with respect to the similarity matrix, we show that x_i actually encodes three kinds of transformations, including rotation, scaling and translation. This fact can be further utilized to ease the procedure of mean shift clustering as well as to reduce the computational overhead.

Next, we first introduce the specific form of matrix x_i in the Lie algebra (denoted by $\mathfrak{sim}(2)$) and describe how to compute it in an efficient manner in order to improve the clustering efficiency. Thereafter, we elaborate on deriving the matrix H in (12), which measures the distance between the two similarity matrices X_i and X more precisely than existing approaches, providing better convergence performance during the mean shift procedure.

A. Computing Elements in $\mathfrak{sim}(2)$ for Mean Shift Clustering

Without loss of generality, assume X is the current center during the mean shift procedure while X_i is a similarity matrix nearby X in the extracted Sim(2) and $\mathbf{y}_i = X_i X^{-1}$ is its transformed version. As stated earlier, $\mathbf{y}_i$ and its corresponding element x_i in Lie algebra map to each other via matrix logarithm and exponential. The two elements are 3×3 matrices taking the form

$$\mathbf{y}_i = \begin{bmatrix} R & t \\ \mathbf{0} & 1 \end{bmatrix}, \quad \text{and} \quad x_i = \begin{bmatrix} \Omega & u \\ \mathbf{0} & 0 \end{bmatrix}. \tag{15}$$

Note that here y_i is also a similarity matrix and thus has the same form as (1) while x_i is the matrix form of (13) with Ω being a 2×2 matrix and u a 2D-vector. More specifically, if R is a transformation with rotation θ and scale s, it can be deduced that

$$R = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \text{and} \quad \Omega = \begin{bmatrix} \log(s) & -\theta \\ \theta & \log(s) \end{bmatrix}.$$

Now the vectors in $\mathfrak{sim}(2)$ can be simplified as $(\theta, \log(s), u^T)^T$, where θ and s represent rotation difference and scaling ratio between the two transformations X and X_i respectively.[1] Hence, it is a computationally more efficient form of (13), considering that the rotation and scaling encoded by X are easy to compute without resorting to matrix logarithm.

The specific form of u is more complex, however it is not necessarily to be made explicit. During mean shift clustering, the main procedure alternates between updating the mean shift vectors and center locations. Including u in mean shift vectors is to determine the translation component t of the new center matrix $\bar{X}$. Since the rotation and scaling components of mean shift vectors are independent of translation, they can be obtained in the first place and further used to compute matrix R of $\bar{X}$ (denoted by $R_{\bar{X}}$). Now recall that each transformation in the constructed Sim(2) is also associated with a pair of spatial coordinates, which perfectly map to each other by the corresponding transformation. Assume that the new center $\bar{X}$ is generated by a set of transformations in a local neighborhood with pairs of spatial coordinates, denoted by $\{(c_i(c_{i,x}, c_{i,y})^T, c_i'(c_{i,x}', c_{i,y}')^T)\}_{i=1,..,k}$. Matrix $\bar{X}$ ought to be compatible with all the transformations, meaning that $\bar{X}$ maps all pairs of spatial coordinates with the minimum projection error,

$$\min_{t_{\bar{X}}} \sum_{i=1}^{k} ||R_{\bar{X}} c_i - c_i' - t_{\bar{X}}||^2. \tag{16}$$

Here, $t_{\bar{X}}$ denotes translation component of $\bar{X}$. Thus it can be expressed as

$$t_{\bar{X}} = \sum_{i=1}^{k} (R_{\bar{X}} c_i - c_i')/k, \tag{17}$$

which is obtained by taking the gradient of (16). In this way, transformation matrix of the new center is determined without evaluating the value of term u.

B. Determining the Distance Measure

In general, the identity matrix is often taken as the value of H for evaluating the distance measure in (12) [14], [15]. It is equivalent to deducing the norm of the vector of the matrix x_i in $\mathfrak{sim}(2)$ with the standard inner product. However, as the elements constituting x_i are endowed with different meanings, e.g. θ for rotation difference between X and X_i, $\log(s)$ for scaling and u for translation, computing the distance simply as the sum of their absolute values is not well-justified. To relieve the problem, existing methods rely on a weighted sum to modulate the influence of individual components. For instance, the weights are set up in [19] in the way that a rotation by 180 degree corresponds to a scaling factor of 10 and a certain amount of displacement. The measure derived this way is tricky and error-prone due to lack of theoretical foundation.

To this end, we provide a specific form of H which yields a more effective distance measure integrating all kinds of transformation differences between two similarity matrices. The key idea here is to evaluate the dissimilarity of X and X_i by applying the two matrices to the same points and then computing the average distance between the projected locations. To begin with, let us re-visit the series of matrix logarithm of $\mathbf{y}_i = X_i X^{-1}$. As $\mathbf{y}_i$ is close to the identity element e, $x_i = \log(\mathbf{y}_i)$ is likely to be well approximated by the first-order Taylor expansion of (7), which results in the following distance measure according to (10),

$$d^2(X, X_i) \doteq ||\mathbf{v}||_H^2 = \vec{\mathbf{v}}^T H \vec{\mathbf{v}}, \text{ and } \mathbf{v} = X_i X^{-1} - e. \tag{18}$$

As in (9), $\vec{\mathbf{v}}$ is the vector form of matrix $\mathbf{v}$. Moreover, applying $\mathbf{v}$ to a homogeneous coordinate p', the norm of the resultant vector r can be obtained by

$$||r||^2 = ||\mathbf{v}p'||^2 = ||X_i X^{-1} p' - p'||^2. \tag{19}$$

Taking $p = X^{-1} p'$, we then have

$$p' = Xp, \text{ and } ||r||^2 = ||X_i p - p'||^2 = ||X_i p - Xp||^2. \tag{20}$$

Hence the norm faithfully reveals the projection difference when applying X and X_i to the same point p', forming a natural distance measure evaluating the proximity between X and X_i. This is illustrated by Fig. 3. Moreover, when n points are used for evaluation, the average projection difference is $\frac{1}{n}\sum_{k=1}^{n} ||r_k||^2$, which is equivalent to deriving the distance measure $d(X, X_i)$ in (18) by setting up H as

$$H = \begin{bmatrix} \mathcal{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{M} \end{bmatrix}, \text{ and } \mathcal{M} = \frac{1}{n}\sum_{k=1}^{n} p_k' p_k'^T. \tag{21}$$

Here, point p_k' can be determined under the condition that there exists a point $p_k = X^{-1} p_k'$. Since X originally maps between two circular regions, denoted by O and O' as shown in Fig. 3, all the points within O' are valid choices for p_k'.

[1] Generally speaking, when talking about the vector of x_i in $\mathfrak{sim}(2)$, it means the vector $(\theta, \log(s), u^T)^T$. To ease exposition, we use the equivalent vector $\vec{x}_i$ here by re-arranging all the elements of x_i in the form of a column vector.

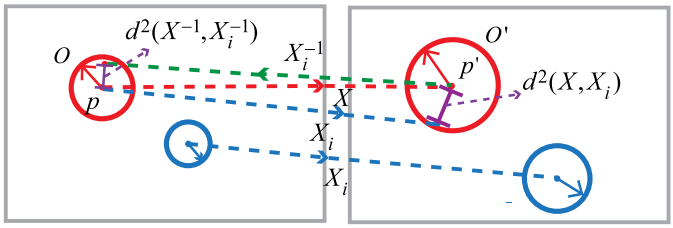

Fig. 3. Measuring the distance between center of the current local window X and a nearby transformation X_i in the similarity matrix Lie algebra with the new distance measure. X maps the circular region O to O', with p and p' being their centroids separately. $d^2(X, X_i)$ computes the distance between the projected locations of p using X and X_i. $d^2(X^{-1}, X_i^{-1})$ is similarly calculated by applying X^{-1} and X_i^{-1} to p'. The distance between X and X_i in the similarity matrix Lie algebra is evaluated by the average sum of $d^2(X, X_i)$ and $d^2(X^{-1}, X_i^{-1})$.

However, in practice, we only use the center of O' considering that the resultant distance generally approaches the average projection difference using all the points, while computation of these points is time-consuming.

In addition, since each X_i in the extracted transformation space has an inverse transformation X_i^{-1}, a distance measure between X^{-1} and X_i^{-1} can be computed similarly as

$$d^2(X^{-1}, X_i^{-1}) \doteq ||\mathbf{v}'||_H = \vec{\mathbf{v}}'^T H' \vec{\mathbf{v}}', \text{ and } \mathbf{v}' = X_i^{-1}X - e \tag{22}$$

with

$$H' = \begin{bmatrix} \mathcal{M}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{M}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{M}' \end{bmatrix}, \text{ and } \mathcal{M}' = cc^T. \tag{23}$$

Here, c is the homogeneous coordinates of the center of the circle O. Note that $d^2(X^{-1}, X_i^{-1})$ also depicts the proximity between X and X_i because X and X^{-1} can be derived from each other directly. Moreover, $d^2(X, X_i)$ generally does not equal $d^2(X^{-1}, X_i^{-1})$ as illustrated by Fig. 3. Therefore, the final distance is determined by combining both distances as,

$$d^2([X; X^{-1}], [X_i; X_i^{-1}]) \doteq \frac{1}{2}\left(d^2(X, X_i) + d^2(X^{-1}, X_i^{-1})\right), \tag{24}$$

where $[X; X^{-1}]$ is the concatenation of X and X^{-1}.

A concern remained about the above distance measure is its validity when applying it to the mean shift algorithm under the condition that X_i is far way from X in the extracted transformation space, meaning $\mathbf{y}_i$ is not close to the identity matrix. In this case, the norm computed using (19) may not match the original distance produced by (10). However, it can be observed that the norm still accounts for the difference between X and X_i and (24) yields a large distance value. Recall that when running the mean shift iteration from X, it only relies on the transformations nearby X for estimating the shift vector and seeking the local mode. Remote transformations with distance larger than the kernel-bandwidth h are assigned zero weights by the kernel function $g(\cdot)$ in (12), and as a result are excluded from distracting the clustering process. This guarantees the effectiveness of the provided distance measure, which is also verified and compared with traditional distance measures in our experiments.

Algorithm 1 Mean Shift for CVP Discovery

Input: Similarity transformations $\{X_i\}_{i=1,\ldots,n}$, each associated with a vector $v_i = (c_{i,x}, c_{i,y}, c'_{i,x}, c'_{i,y}, \theta_i, \sigma_i)$, consisting of two centroid coordinates $c_i(c_{i,x}, c_{i,y})$ and $c'_i(c'_{i,x}, c'_{i,y})$, orientation difference θ_i, and $\sigma_i = \log(s_i)$ representing the logarithm of radius ratio. θ_i and s_i are defined in (2).
Parameters: $\tau_\theta = 20$, $\tau_s = 1$, $\xi = 4$.
for $i = 1$ to n **do**
  $X \leftarrow X_i$
  $v(c_x, c_y, c'_x, c'_y, \theta, \sigma) \leftarrow v_i$
  repeat
    for all $\{X_i\}_{i=1,\ldots,n}$ **do**
      Gather points nearby X and generate an index set $\{N_j\}_{j=1,\ldots,k}$ such that $|\theta - \theta_{N_j}| < \tau_\theta$ and $|\sigma - \sigma_{N_j}| < \tau_s$
    end for
    Compute H of (21) and H' of (23) using $(c'_x, c'_y, 1)^T$ and $(c_x, c_y, 1)^T$ respectively.
    for all $\{X_{N_j}\}_{j=1,\ldots,k}$ **do**
      $m_{h,g}(v) \leftarrow \dfrac{\sum_{j=1}^{k} g\left(\frac{d^2([X;X^{-1}],[X_{N_j};X_{N_j}^{-1}])}{h^2}\right)(v_{N_j} - v)}{\sum_{j=1}^{k} g\left(\frac{d^2([X;X^{-1}],[X_{N_j};X_{N_j}^{-1}])}{h^2}\right)}$
      $v(c_x, c_y, c'_x, c'_y, \theta, \sigma) \leftarrow v + m_{h,g}(v)$
      $R \leftarrow \begin{bmatrix} \exp(\sigma) & 0 \\ 0 & \exp(\sigma) \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
      $t \leftarrow \sum_{j=1}^{k} \left(R(c_{N_j,x}, c_{N_j,y})^T - (c'_{N_j,x}, c'_{N_j,y})^T\right)/k$
      $X_{pre} \leftarrow X$
      $X \leftarrow \begin{bmatrix} R & t \\ \mathbf{0} & 1 \end{bmatrix}$
    end for
  until $d(X, X_{pre}) < \xi$
  Store X as a local mode.
end for
Assign the cluster for data points based on their mode and derive the CVPs correspondingly.

C. Using Mean Shift for CVP Discovery

With the distance measure and gradient vectors specified above, the mean shift algorithm for the constructed transformation space is finalized with the radially symmetric Epanechnikov kernel [16]. After clustering, the transformations subjecting to the same mode are grouped together to form a cluster. Regions corresponding to the transformations of the same significant cluster are joined together on each input image separately and identified as a single CVP between the two input images. In addition, small clusters with the number of members less than 8 are automatically removed. The whole procedure proceeds as in Algorithm 1.

In principle, the complexity of the algorithm is $O(n^2 I)$, where n is the number of total similarity transformation matrices computed and I is the average number of iterations required for each transformation to reach its local mode.

The time bottleneck lies in calculating the distances between current center X and all the transformations. However, since the kernel function $g(\cdot)$ assigns zero value to the distance larger than the bandwidth h, we speed up the process by filtering the transformations beforehand and keeping only those nearby X in lines 6 to 8. In practice, it can be further accelerated using grid hashing, i.e., to partition the transformation space with multiple grids and only search the grids nearby the current center for its neighbors. This in general reduces the search trials to dozens of data points, and thus effectively improves clustering efficiency. Assuming that the average search trial is K, then complexity of the algorithm is reduced to $O(nKI)$.

V. Experiments

In this section, we evaluate the effectiveness of our proposed approach on three tasks. We first design a series of feature matching experiments to verify its robustness against various challenges, e.g., distracting outliers, deformation, etc. Thereafter, two other applications, multiple CVP discovery and near-duplicate image retrieval are further tested to demonstrate the performance. In addition, we also evaluate the choice for the kernel-bandwidth of mean shift h and the effectiveness of the provided distance measure in the first experiment.

A. Local Feature Matching on Synthetic Images

As explained previously, discovering CVPs essentially entails identifying correct feature correspondences between a pair of input images. In this experiment, we verify the robustness of the proposed method for matching image features under challenging conditions, including object deformation and noise disturbance. To present quantitative evaluations, sets of images are generated from an input image dataset with multiple Thin-Plate-Spline (TPS) transformations, followed by local interest region extraction and inliers and outliers determination of initial feature correspondences between each pair of deformed and original images. We then conduct similar experiments as described in SCC [2], whose performance is compared as well. Note that SCC is one of the state-of-the-art CVP discovery algorithms and it has demonstrated superior performance over several existing techniques, for instance [9].

We use the model images in ETHZ toys dataset [20] as our benchmark dataset. To apply TPS deformation to an image, 9×9 crossing points of a 10×10 meshgrid overlaid on the input image are selected as control points, each of which is then independently perturbed by the zero mean Gaussian noise $N(0, \sigma)$. A new image is generated by warping the control points on the original image to their new positions. To further simulate deformations of real-world images, we also rotate the new image with random rotation. We show a pair of the original and deformed images in Fig. 4.

Given a deformed image I_Q and its original image I_P, we obtain their initial local feature sets with SIFT. With the known TPS and rotation parameters, we can then easily acquire n^i inliers in I_P and their corresponding n^i inliers in I_Q. Note that when multiple SIFT features are detected at the same location, inlier correspondences are confirmed by the the minimal SIFT features distance. Next we randomly collect n_p^o and n_Q^o

Fig. 4. A pair of the original (left) and deformed images (right). The TPS transformation is conducted with $\sigma = 10$.

features points as outliers from the remaining feature sets in I_P and I_Q, respectively. Thus the total number of feature points in I_P and I_Q are $n_P = n^i + n_P^o$ and $n_Q = n^i + n_Q^o$ and the matching problem is to derive correct correspondences from all the $n_P \times n_Q$ candidate sets. This experimental setting is very similar to the one in [2] in that the parameter σ controls the level of deformation between feature point sets, while n_p^o and n_Q^o determine the number of outliers. The difference, however, lies in that the point sets are gathered from extracted feature points and deformation is achieved by image level TPS transformation, instead of direct disturbance of each individual point. This not only better simulates the practical matching circumstances, but also preserves the underlying challenges caused by homogeneity of the data and the large search space during quantitative evaluation.

We use the publicly available code of SCC shared by the authors for comparison. The performance is measured based on the number of correct correspondences returned under the one-to-one matching constraint, which is enforced by exploiting a Spectral Matching [9] procedure as post-processing to filter the initially returned correspondences and only keep the top n^i pairs. Note the post-processing is only introduced in this experiment considering the matching noise is so high that each point has at least $\min(n_P, n_Q)$ candidate correspondences, among which only one is true.

For each combination of deformation level and numbers of inliers and outliers, we run both algorithms on all 40 pairs of the original and deformed images to produce mean performance scores and the standard deviations. The performance comparison is shown in Fig. 5. The 1st and 2nd rows plot the performance under varying deformations with fixed amount of inliers and outliers, meaning that n^i out of $n_P \times n_Q$ initial correspondences are groundtruths. It can be observed that both algorithms work well when the amount of noise is small. Otherwise, the performances of both algorithms degenerate with the increasing level of deformations. However, our method shows comparably superior ability in handling deformation. This could be explained by the fact that the variation among local geometric transformation caused by limited degree of deformation is generally continuous, and thus compatible transformations are identifiable through the clustering process as long as their underlying cluster structure is not destroyed. In the 3rd row, we keep the deformation parameter fixed with $\sigma = 5$ and change the number of outliers (left) as well as both inliers and outliers. Again, our method generates

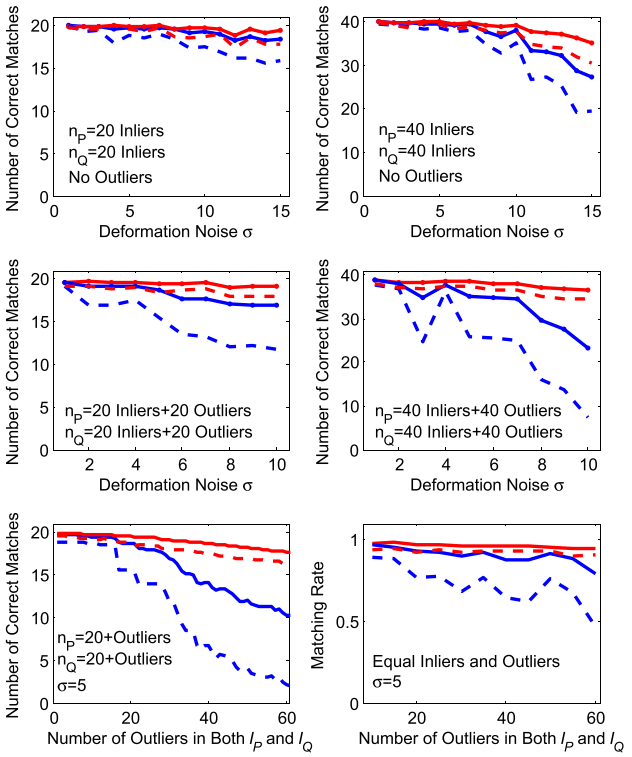

Fig. 5. Performance curves for our method (red lines) and SCC in [2] (blue lines). The mean performance is shown as a solid line, while one std below the mean is shown as a dotted line for each method. The number of correct matches is plotted against varying deformation noise with (2nd row) and without (1st row) outlier points. In the 3rd row, we keep the deformation parameter fixed with $\sigma = 5$ and change the number of outliers (left) as well as both inliers and outliers.

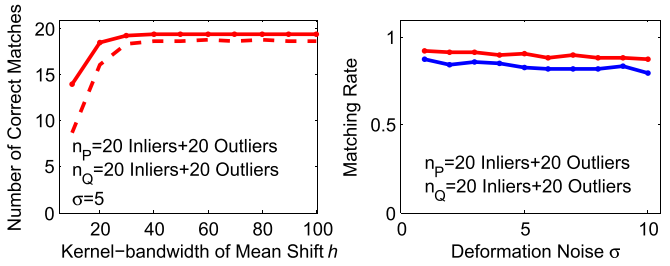

Fig. 6. Left: Performance curves using varying kernel-bandwidth for the mean shift algorithm. The mean performance is shown as a solid line, while one std below the mean is shown as a dotted line. Right: Curves of correct correspondence rate in the largest clusters obtained with the proposed distance measure (the red line) and the one using standard inner product function (the blue line).

higher average performance with low variance, which indicates it can better tolerate noise. Overall, the whole experiment suggests that correct correspondences can practically converge to the modes of the underlying density distribution of the affine transformation space through our proposed nonlinear mean shift procedure. Furthermore, this convergence process is robust under certain challenging conditions, including object deformation and distracting outliers.

In addition, we study how to specify a good value for the kernel-bandwidth h. As shown in the left sub-figure of Fig. 6, the performance of the algorithm continues to improve with the increasing of h until it reaches a stable state. This indicates that a large h is preferred for good performance.

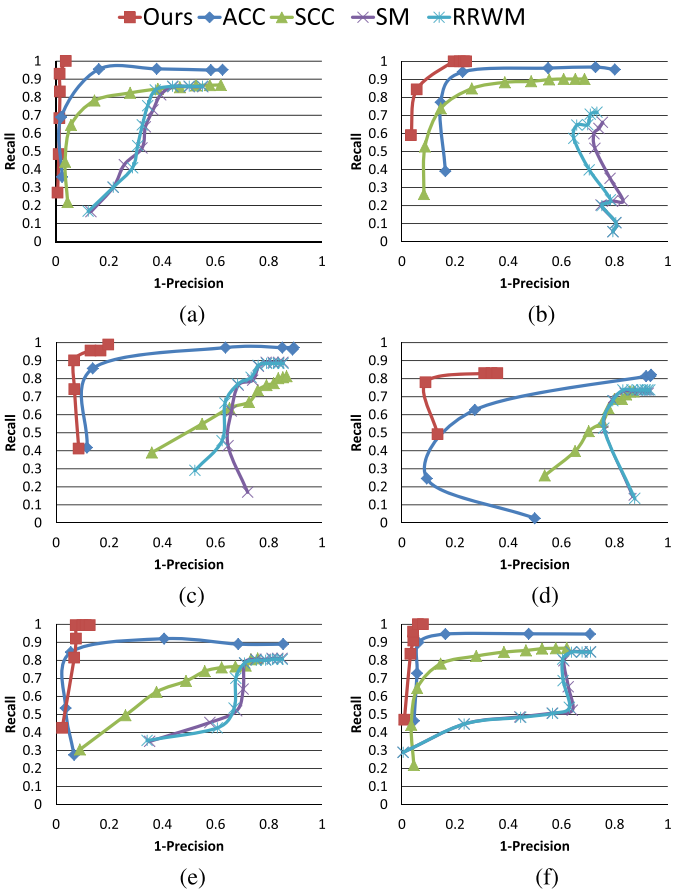

Fig. 7. Performance comparison of multiple common visual pattern discovery on each image of the SNU dataset. For each image, the number of groundtruth correspondences and initial correspondences, obtained by thresholding distance ratio of SIFT feature, are displayed in the parentheses. (a) Books (1024 / 3036). (b) Bulletins (379 / 2202). (c) Jigsaws (182 / 2347). (d) Mickeys (146 / 1873). (e) Minnies (200 / 1572). (f) Toys (681 / 2960).

However, a large h also incurs more computational overhead. Therefore, in all our experiments, we set $h = 40$ as it suggests a good tradeoff between accuracy and efficiency according to Fig. 6 (left). Moreover, in the right sub-figure of Fig. 6, we employ the same mean shift procedure to compare the newly provided distance measure with the one using standard inner product. The precision rate in the first cluster obtained is shown. Obviously, the proposed distance measure delivers better performance, which suggests that it can better facilitate convergence of the algorithm.

B. Multiple Common Visual Patterns Discovery

In this subsection, we test our approach for identifying multiple CVPs in real world image pairs. We use the SNU dataset [21], containing six image pairs that are particularly collected for the task. Each image pair contains at least 2 CVPs, each undergoing random geometric and photometric deformations, as well as clutter backgrounds and partially occlusion. This makes the dataset quite challenging.

Before testing, we extract local features of each image pair using SIFT in VLfeat [22], and obtain initial feature correspondences by thresholding distance ratio of the closest and second

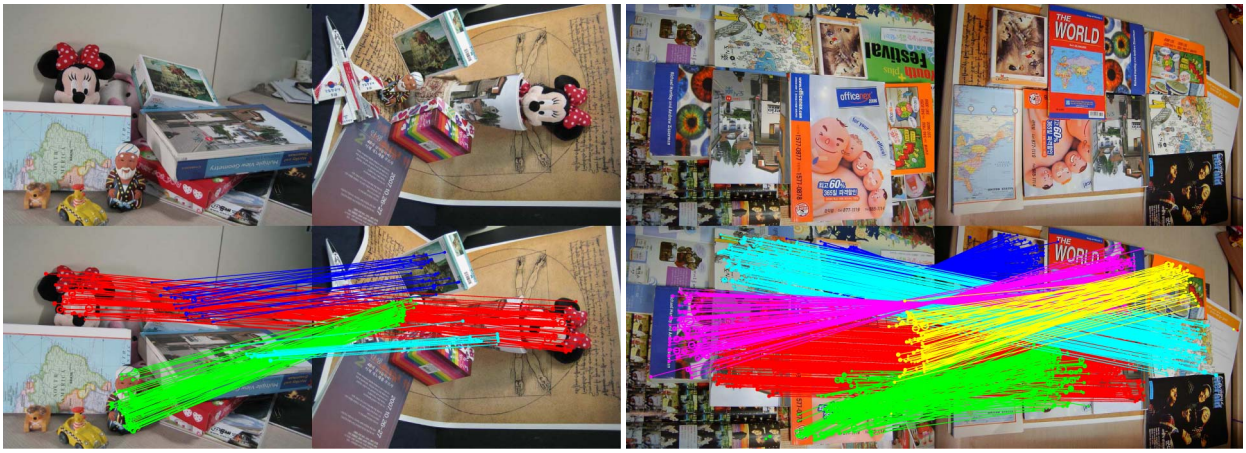

Fig. 8. Results of multiple common visual pattern discovery on the SNU dataset.

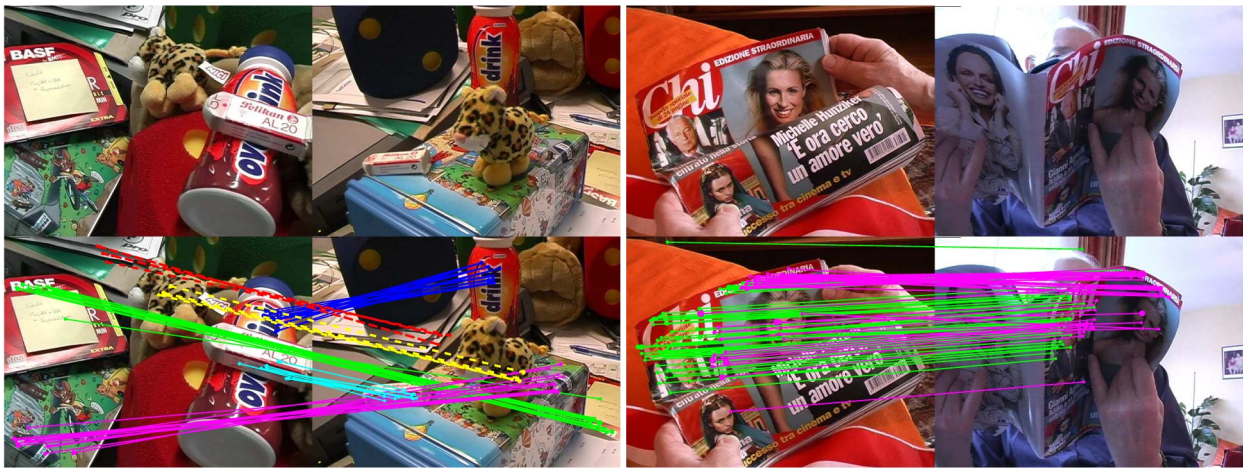

Fig. 9. Detection of multiple common patterns under different viewpoint changes. CVPs corresponding to insignificant clusters (number of feature correspondences < 8) are discarded after mean shift clustering (dotted lines). CVPs undergoing large viewpoint distortion may be missed due to insignificant clusters (yellow dotted lines).

closest features [8]. This preprocessing generally discards a large fraction of outliers while keeping most of the correct matches. However, a significant amount of distracting outliers are still retained since it fully relies on local feature descriptors to establish correspondences, without considering any geometric constraints. We then feed the initial correspondence set into our framework and several other approaches to verify the effectiveness. To quantitatively measure the performance, we manually pick up the groundtruth correspondences out of the initial match set.

We compare our approach with multiple state-of-the-art techniques, including ACC [13], SCC [2], SM [9] and RRWM [12]. The performances is measured by recall versus 1-precision curves, with

$$recall = \frac{CorrectMatchesReturned}{TotalCorrectMatches},$$
$$precision = \frac{CorrectMatchesReturned}{AllMatchesReturned}.$$

We plot the performance of each method by varying the value of its controlling parameter, such as the principal eigenvector in Spectral Matching [9], to gradually increase its recall. Since there is no such parameter in our case, we draw the kth point of our curve using the recall and precision rate of returned matches in all top k largest clusters.

We plot the quantitative results of all approaches in Fig. 7. As shown, SM and RRWM are more sensitive to outliers mainly because that they seek to separate correct matches from the initial match set in a global manner. Without considering local property, they are likely to handle single CVP cases well, but may fail for multiple CVPs since correspondences of one CVP may manifest itself as outliers to another. By contrast, ACC, SCC and our approach improve the performance by emphasizing more on the intra-correlation among correspondences constituting each individual CVP. Among them, our approach shows superior performance over the other two in most cases, which suggests that the proposed clustering strategy can faithfully recover the underlying modes of the extracted transformation set as well as true correspondences with respect to multiple CVPs.

Moreover, each CVP derived corresponds to one cluster in the transformation space, yielding clear boundaries among

different CVPs. Hence multiple CVPs can then be naturally separated from each other and recognized. We show two examples and our results on the SNU dataset in Fig. 8. It can be observed that a large number of true feature correspondences belonging to different CVPs are correctly identified and grouped together, verifying the effectiveness of the proposed approach.

Fig. 8 also indicates that our approach can detect CVPs under viewpoint changes given that some of the CVPs are imaged under different viewpoints (e.g. the old man toy in the left part of Fig. 8). Fig. 9 shows two challenging examples, where multiple CVPs in each example all undergo different degree of viewpoint changes and some suffered from severe deformation distortion. As shown, the proposed approach succeeds to find most CVPs in this case as well, which further verifies its capability in handling viewpoint changes. Note that we do not claim here that it achieves better viewpoint invariance than the SIFT descriptor. In fact, since the proposed algorithm focuses on identifying correct matches from initially SIFT-matched feature correspondences, its capability in handling viewpoint changes largely depends on what degree of viewpoint invariance SIFT can achieve. Our tests show that if $N(N \geq 8)$ true SIFT matches of a CVP survive under viewpoint changes, the mean shift procedure is generally able to identify most correct feature correspondences. Otherwise, the proposed approach may fail as illustrated by the tiger toy in Fig. 9 (left). A possible solution to relieve this issue is to replace the SIFT descriptor with Affine SIFT [23], which we plan to investigate in the future.

Finally, our approach can be naturally extended to find CVPs across multiple images. To start off with, SIFT-matched candidate feature matches are firstly collected from each image pair, and fed into the proposed CVP discovery algorithm, resulting in multiple matching clusters. For each two clusters, if they share more than fifty percent of the feature points in one image, they usually correlates with multiple instances of the same object in different images and hence are regarded as the same CVP. The visual result of one example is shown in Fig. 10(a), where the beer logo is shared across all four images. As shown, each group of blue lines between an image pair represents a CVP, and all the CVPs in the image group are the same. Similarly, we can also find CVPs within a single image and CVPs from both within an image and across multiple images simultaneously. Examples of the two cases are displayed in Fig. 10(b) and Fig. 10(c) respectively.

Fig. 10. CVP discovery example in different input settings. (a) Finding CVP in a group of four images. The four images are put together for display clarity. (b) Detecting CVPs constructed by multiple instances of the same object in a single image. (c) Finding CVPs within an image and among multiple images simultaneously.

C. Near-Duplicate Image Retrieval

In this subsection, we evaluate our method on the task of near-duplicate image retrieval, which plays a critical role in many multimedia applications, such as news video search and copyright infringement detection, etc. Since near-duplicate images usually contain one or multiple CVPs, algorithms of CVP discovery are all expected to be applicable to the task.

We test the proposed approach on the Columbia dataset, containing 600 images (150 pairs of near-duplicate images and 300 non-duplicate images) extracted from TRECVID2003. We follow the same experimental settings as in [2] and target

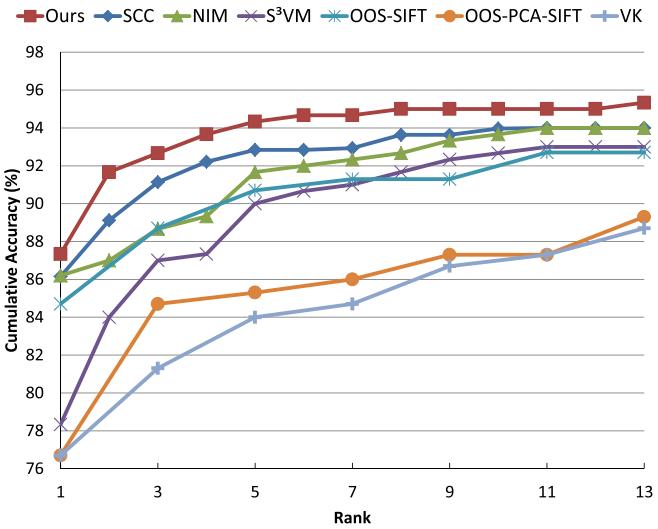

Fig. 11. Performance comparison of near-duplicate image retrieval on Columbia dataset.

on re-ranking top 50 correspondence candidates retrieved for each near-duplicate image using the method in [24]. Local invariant region detection and initial correspondence acquisition follow the same routine as the previous experiment.

In Fig. 11, we plot the cumulative accuracy of top 13 ranking results achieved by our and several

Fig. 12. Comparison of resultant correspondences obtained by our method (top row) and [2] (bottom row).

existing approaches, including SCC [2], NIM [24], OOS-SIFT [25], OOS-PCA-SIFT [26] and Visual Keywords (VK) [26]. As shown, our approach clearly outperforms the others. Fig. 12 demonstrates the representative results obtained by our method and SCC on three groups of images, verifying that our approach can robustly detect common visual patterns in real images. In addition, including feature and initial correspondence extraction, it takes about 0.7 hour for our approach to deal with all the 15000 pairs of images (300 near-duplicate images, each with 50 candidate correspondences). The algorithm can be further speeded up by other acceleration techniques, e.g. GPU acceleration, thus showing great potential in large-scale recognition applications.

VI. Conclusion

In this paper, we have presented a mean shift clustering based approach for the task of common visual pattern discovery between a pair of input images. A set of similarity transformation matrices is firstly computed from the geometric invariant parameters of SIFT feature correspondences. The transformations constitute a matrix Lie group, where a nonlinear mean shift algorithm can be utilized to find the modes of underlying density distribution. Considering the non-Euclidean nature of the similarity transformation space, mean shift vectors are derived in the corresponding Lie algebra vector space with a newly provided distance measure. Extensive experiments demonstrate the robustness and efficiency of the proposed approach.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful suggestions in improving this paper.

References

[1] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 26–33.

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

[3] G. Zhao and J. Yuan, "Mining and cropping common objects from images," in *Proc. ACM Multimedia*, 2010, pp. 975–978.

[4] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang, "Video object co-segmentation via subspace clustering and quadratic pseudo-Boolean optimization in an MRF framework," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 903–916, Jun. 2014.

[5] L. Wang, T. Xia, Y. Guo, L. Liu, and J. Wang, "Confidence-driven image co-matting," *Comput. Graph.*, vol. 38, pp. 131–139, Feb. 2014.

[6] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

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

[9] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Oct. 2005, pp. 1482–1489.

[10] L. Torresani, V. Kolmogorov, and C. Rother, "A dual decomposition approach to feature correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 259–271, Feb. 2013.

[11] H. Jiang, S. X. Yu, and D. R. Martin, "Linear scale and rotation invariant matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1339–1355, Jul. 2011.

[12] M. Cho, J. Lee, and K. M. Lee, "Reweighted random walks for graph matching," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 492–505.

[13] M. Cho, J. Lee, and K. M. Lee, "Feature correspondence and deformable object matching via agglomerative correspondence clustering," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1280–1287.

[14] R. Subbarao and P. Meer, "Nonlinear mean shift over Riemannian manifolds," *Int. J. Comput. Vis.*, vol. 84, no. 1, pp. 1–20, 2009.

[15] O. Tuzel, R. Subbarao, and P. Meer, "Simultaneous multiple 3D motion estimation via mode finding on Lie groups," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, vol. 1. Oct. 2005, pp. 18–25.

[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[17] W. Rossmann, *Lie Groups: An Introduction Through Linear Groups*, vol. 5. London, U.K.: Oxford Univ. Press, 2002.

[18] E. Begelfor and M. Werman, "How to put probabilities on homographies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1666–1670, Oct. 2005.

[19] N. J. Mitra, L. J. Guibas, and M. Pauly, "Partial and approximate symmetry detection for 3D geometry," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 560–568, 2006.

[20] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Simultaneous object recognition and segmentation from single or multiple model views," *Int. J. Comput. Vis.*, vol. 67, no. 2, pp. 159–188, 2006.

[21] M. Cho, Y. M. Shin, and K. M. Lee, "Co-recognition of image pairs by data-driven Monte Carlo image exploration," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 144–157.

[22] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. ACM Multimedia*, 2010, pp. 1469–1472.

[23] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.

[24] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *Proc. ACM Multimedia*, 2008, pp. 41–50.

[25] X. Wu, W.-L. Zhao, and C.-W. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," in *Proc. ACM Int. Conf. Image Video Retr.*, 2007, pp. 162–169.

[26] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 1037–1048, Aug. 2007.

Linbo Wang received the B.Eng. degree in computer science from Shandong University, Jinan, China, in 2005, and the Ph.D. degree in computer science from Nanjing University, China, in 2014. He is currently a Lecturer with the School of Computer Science and Technology, Anhui University, China. His research interests include computer vision, image processing, and computer graphics.

Dong Tang received the B.Eng. degree from Nanchang Hangkong University, China, in 2012, and the M.Eng. degree from Nanjing University, China, in 2015. His research interests include computer vision and image processing.

Yanwen Guo received the Ph.D. degree in applied mathematics from the State Key Lab of CAD&CG, Zhejiang University, China, in 2006. He is currently a Full Professor with the National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Jiangsu, China. He was a Visiting Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2006 and 2009, respectively, and the Department of Computer Science, The University of Hong Kong, in 2008, 2012, and 2013, respectively. He has been a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, since 2013. His research interests include image and video processing, vision, and computer graphics.

Minh N. Do (M'01–SM'07–F'14) was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Canberra, ACT, Australia, in 1997, and the D.Sc. degree in communication systems from the Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland, in 2001. He has been a Faculty Member with the University of Illinois at Urbana–Champaign, Champaign, IL, USA, since 2002, where he is currently a Professor with the Department of Electrical and Computer Engineering, and hold joint appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, and the Department of Bioengineering. His research interests include image and multidimensional signal processing, wavelets and multiscale geometric analysis, computational imaging, augmented reality, and visual information representation. He was a recipient of the Silver Medal from the 32nd International Mathematical Olympiad in 1991, the University Medal from the University of Canberra in 1997, the Doctorate Award from EPFL in 2001, the CAREER Award from the National Science Foundation in 2003, and the Young Author Best Paper Award from the IEEE in 2008. He was named a Beckman Fellow with the Center for Advanced Study, UIUC, in 2006, and received the Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007. He was a member of the IEEE Signal Processing Theory and Methods Technical Committee and the Image, Video, and Multidimensional Signal Processing Technical Committee, and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.