Multimedia Systems
https://doi.org/10.1007/s00530-021-00792-8

SPECIAL ISSUE PAPER

Adaptively feature matching via joint transformational-spatial clustering

Linbo Wang[1] · Li Tan[1] · Xianyong Fang[1] · Yanwen Guo[2] · Shaohua Wan[3]

Received: 25 July 2020 / Accepted: 5 April 2021

Abstract
The transformational and spatial proximities are important cues for identifying inliers from an appearance based match set because correct matches generally stay close in input images and share similar local transformations. However, most existing approaches only check one type of them or both types consecutively with manually set thresholds, and thus their matching accuracy and flexibility in handling large-scale images are limited. In this paper, we present an efficient clustering based approach to identify match inliers with both proximities simultaneously. It first projects the putative matches into a joint transformational-spatial space, where mismatches tend to scatter all around while correct matches gather together. A mode-seeking process based on joint kernel density estimation is then proposed to obtain significant clusters in the joint space, where each cluster contains matches mapping the same object across images with high accuracy. Moreover, kernel bandwidths for measuring match proximities are adaptively set during density estimation, which enhances its applicability for matching different images. Experiments on three standard datasets show that the proposed approach delivers superior performance on a variety of feature matching tasks, including multi-object matching, duplicate object matching and object retrieval.

Keywords Clustering · Density estimation · Feature matching · Mode-seeking

1 Introduction

Establishing feature correspondences among images has received lots of attention due to its important role in various applications such as object recognition, tracking, and near-duplicate image retrieval. Despite tremendous previous efforts, descriptor based feature matching remains challenging in identifying inliers from a putative match set formed by comparing local feature descriptors [1, 2]. In this paper, we aim for pruning wrong matches from the putative match set by exploring transformational and spatial proximities among match inliers via a joint domain clustering process.

Most existing approaches seek to alleviate mismatches by incorporating transformational proximity, which assumes correct matches to share similar local transformations. Among them, geometric voting techniques, such as RANSAC [3] and Hough transform [4, 5], are employed to find feature matches with inter-image rigid object transformations. Graph-based approaches [6, 7] usually encode pair-wise geometric constraints in a graph, aiming to detect matches with locally consistent transformations using graph optimization. Clustering-based methods [8–11] project the putative matches into the transformation space and identify match clusters with close transformational proximity. Deep learning based approaches usually embed the transformational proximity in a deep network for detecting correct matches [12–17]. Meanwhile, spatial proximity which expects correct matches to have close keypoints in each image is also explored for wrong match pruning. This is typically achieved by distance based neighborhood inconsistency measuring [18], keypoint motion smoothness evaluation among nearby matches [19, 20].

The two lines of researches suggest that both proximities are helpful cues for rejecting wrong matches, and exploring them together ought to further enhance the matching

✉ Xianyong Fang
fangxianyong@ahu.edu.cn

[1] MOE Key Laboratory of Intelligent Computing and Signal Processing, School of Computer Science and Technology, Anhui University, Hefei, China

[2] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

[3] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China

Published online: 10 June 2021

performance. However, checking the two proximities consecutively [21] often fails to reject small groups of matches wrongly mapping similarly structured texture patterns. Meanwhile, over-segmenting input images and checking transformational coherence with regional constraints is another beneficial trial [5], but its performance depends on the quality of region partition. Therefore, how to combine both proximities checking in an effective manner remains open for study.

In addition, checking both the proximities generally involves manual thresholds, e.g., the threshold of the symmetric transfer error (STE) for measuring the transformational similarity [5, 22] and the minimum neighboring distance [18, 21] for spatial closeness. This kind of thresholding can limit the matching performance since the proximities among matches are likely to vary with images.

To this end, we revisit the two proximities of correct matches with two observations. Firstly, match inliers are often located on the instances of the same object and stay close in both input images. Secondly, when mapping the same object instances, they typically share similar transformations in their spatial local neighborhoods. This indicates correct matches tend to group together locally and form multiple dense distributions in the underlying joint transformational-spatial domain. Meanwhile, wrong matches are of random transformations and keypoint locations. They may accidentally share either transformational or spatial proximities but not both with each other or correct matches. Therefore, they usually scatter all around in the joint domain. This is illustrated by the Fig. 1, which plots the transformational space (Fig. 1c) of putative matches (Fig. 1b) extracted from an image pair (Fig. 1a). Obviously, the space contains 6 dense local distributions, corresponding to six groups of correct matches mapping keypoints gathered on different objects, as well as a large number of noisy mismatches. Therefore, obtaining reliable matches is then to locate significant clusters in the underlying spatial-transformation space.

We introduce here an effective clustering scheme to detect correspondence groups in the joint domain, a combination of the transformational domain formed by the affine transformations estimated from putative matches and the spatial domain formed by keypoint locations of matches. The key idea is to measure the density of each match point using joint kernel density estimation, and determine the cluster modes by a density based mode-seeking process, which starts from each match point and recursively shifts to its local neighbors with maximum density until a stationary point (local mode) is reached. Matches subject to the same mode are then grouped together, with resultant match set recovered from significant clusters. Figure 1d shows the clustering result of Fig. 1c using our approach. Six clusters are successfully discovered despite their arbitrary shape, and corresponding matches are shown in Fig. 1e. Note that the mismatches mixed in the clusters of correct matches in the transformational space (black points in Fig 1d) do not show in the Fig. 1e, which demonstrates the merits of integrating both proximities in the matching framework.

The transformational and spatial proximities are checked with adaptively set kernel bandwidths during density estimation. The spatial bandwidth is set as the upper bound of spatial distance that can keep most match inliers having correct match neighbors. Meanwhile, the transformational

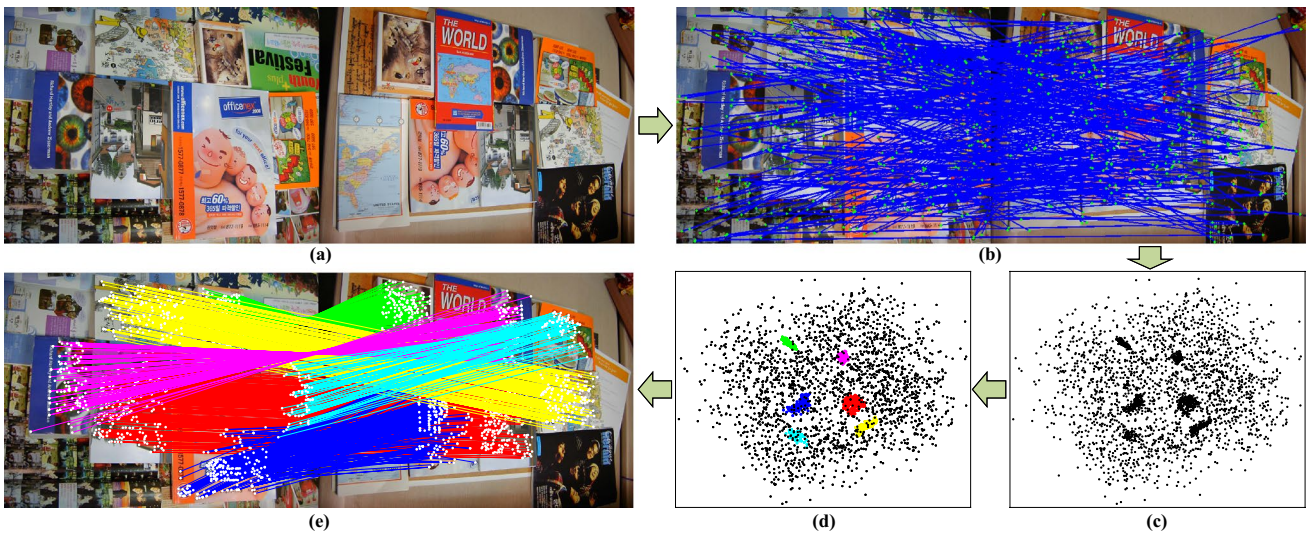

Fig. 1 Overview of our approach. Given two images (**a**), we obtain putative matches (**b**) based on distance of local features, construct their transformational domain (projected to 2D plane (**c**)), perform clustering in the joint transformational-spatial domain (**d**), and recover the final matches (**e**). For the clarity, we only show 300 pairs of putative matches in (**b**), and maximally 100 pairs of matches for each cluster in (**e**)

bandwidth is adaptively tuned to minimize the density entropy, which helps to obtain coherently varying local density distribution for the joint domain and facilitates the mode-seeking process. The resultant bandwidths are more effective for matching different images comparing with manually thresholding, where a large or small bandwidth leads to expansive or shrinkage matching object profiles.

To summarize, our approach has following advantages.

- We propose to identify correct matches with spatial and transformational proximities via density based clustering in the joint domain, which better explores both proximities for feature matching than existing approaches.
- We propose an adaptively tuning strategy for setting the spatial and transformational bandwidths, so that the clustering performance can be better guaranteed for different images.
- Extensive experiments and comparisons show that the proposed approach can effectively handle various matching tasks, including multiple object matching in single or multiple images, and duplicate image retrieval.

2 Related work

2.1 Descriptor based feature matching

The keypoint matching by quantizing the appearance of local regions with discriminative descriptors is a widely studied topic. Notable progress has been achieved by various feature descriptors proposed to date, ranging from the classical techniques in [1, 23], to more recently CNN-based ones in [2, 12, 24]. While they largely form the basis of modern techniques of object matching and extended applications [25–31], it is generally believed that the feature correspondences established by descriptors alone can be enhanced via mutual geometric consistency checking.

2.2 Transformational proximity based feature matching

This line of work identifies match inliers by assuming they share similar local transformations. They can be categorized into voting based, graph matching based, clustering and deep learning based approaches. Voting based methods continuously hypothesizes a global transformation and select the one with most supportive feature correspondences via voting. However, these methods, e.g., Hough transform [4, 5], RANSAC [3], are usually restricted to handle matching of a single rigid object.

By defining a graph encoding both appearance similarity and pairwise geometric coherence [6, 7, 17, 32, 33] or even higher order constraints [34], graph matching based approaches exhibit more flexibility in dealing with non-rigid object matching. Clustering based approaches, on the other hand, have so far sought to group together and identify correct matches in the underlying transformational space. Various clustering techniques has been exploited, including bottom-up agglomerative clustering [8, 35], graph-shift based clustering [10], PageRank based clustering [9] and mean-shift clustering [11], etc. Deep learning has so far been explored and achieved good progress in various computer vision tasks [36, 37]. Most deep learning based matching approaches [14, 15] take raw images as input and design a deep network to learn the image level geometric transformation and identify correct matches jointly, which are typically used to address the stereo problems. More recently, [12, 13] propose to train multi-layer perceptron integrating parametric transformation to classify putative matches as inliers or outliers. All these approaches rely on transformational proximity among matches to rule out false matches, and thus may fall short of robustness if transformations of false matches are close to those of true matches.

2.3 Integrating Spatial proximity for feature matching

Spatial proximity is usually used for feature matching by exploring the neighborhood structural consistency nearby a match. LPM [18] counts the mismatched neighboring keypoints to identify unreliable matches efficiently. Motion smoothness among spatially close matches is also explored in existing approaches. GMS [19] assumes match inliers are grid-wisely motion consistent and detects reliable matches by grid-based motion consistency consensus. These approaches usually run in an efficient manner, but the matching performance depends on the appearance variation in local regions as it dominates the spatial proximity of the underlying matches. LMR [38] performs neighborhood structural consensus with a two-class SVM classifier for mismatch removal. Besides, [21] prunes matches by checking their spatial and transformational proximities successively. [5] treats matches with keypoints on the same boundary preserving local regions and similar transformations as neighbors and identified matches with high neighborhood density as inliers.

Our work is inspired by existing studies that the transformational and spatial proximities are both helpful for wrong match pruning, but explores them together in the joint transformation-spatial space to identify match inliers with density based clustering. This utilizes the complementary nature of two proximities and enhances performance.

3 Algorithm

Given two input images I and I', we first detect local interest regions in each image using popular region detectors [1, 39]. Each region is further affinely adapted [39] to extract SIFT features. Initial correspondence set $M = \{M_i\}_{i=1}^{N}$ are obtained by nearest neighbor searching in the feature space. We then estimate an affine transformation X_i mapping each region pair of M_i as in [11] and compute the mutual distance of two transformations using STE [22]. Transformations and keypoint locations of all putative matches are then collected to form a joint transformational-spatial space. Correct matches are then obtained via a joint domain clustering process, which is illustrated in Fig. 2. Details are discussed below.

3.1 Joint density evaluation

3.1.1 Kernel density estimation

The density of a match M_i is evaluated with joint kernel density estimation in the transformational-spatial space as

$$\rho_i = C \sum_{i=1}^{N} f\left(\left\|\frac{d_{ij}^t}{h_t}\right\|^2\right) f\left(\left\|\frac{d_{ij}^s}{h_s}\right\|^2\right), \tag{1}$$

where d_{ij}^t measures the transformational dissimilarity of M_i and M_j using STE; d_{ij}^s computes the spatial distance by $d_{ij}^s = 0.5 \times (d_{ij}^{s_I} + d_{ij}^{s_{I'}})$, with $d_{ij}^{s_I}$ encoding the Euclidean distance between the keypoints of M_i and M_j on I; h_t and h_s denote the bandwidths in the transformational and spatial domain respectively. C is a constant ensuring the integral value of ρ is 1.

The kernel profile f is applied to both the transformational and the spatial domains in Eq. 1. The density of each match is estimated by the product of the two kernel terms, since the two domains can presumably be independent with each other. Moreover, the choice of kernels rarely makes significant difference in the estimates because the difference using two kernel density estimates can be eliminated by setting up proper bandwidths [40]. Therefore, we choose the uniform kernel profile for f for simplicity. Formally, we define

$$f(x) = \begin{cases} 1 & 0 \le x \le 1, \\ 0 & x > 1. \end{cases} \tag{2}$$

The density ρ_i depicts the crowdedness of correspondences surrounding M_i in its local neighborhood. In general, wrong matches should be excluded from the neighbor set of a true match M_i when computing ρ_i, which is controlled by the kernel bandwidths with Eq. 1. Only the M_j satisfying $d_{ij}^t \le h_t$ and $d_{ij}^s \le h_s$ is considered for evaluating ρ_i with the uniform kernel profile, i.e., simultaneously checking both the transformational and spatial proximities is required. A transformation X_j generated by a wrong match M_j may be similar to the transformation X_i of M_i, however the distance between the keypoints of M_i and M_j is unlikely to be close at the same time. Consequently, it is more effective to estimate the match density in the joint domain rather than the individual transformational or spatial space if the bandwidths are properly set.

3.1.2 Adaptive bandwidth setting

Setting h_s is to set a threshold for spatial distance so that the density evaluation for most match inliers involves only correct neighbors. Equivalently, it is to determine the upper bound of d_{ij}^s ensuring the correctness of M_j nearby M_i.

The spatial distance between two matches can be impacted by the geometric variation in local area across images caused by viewpoint change or object deformation, etc. Assuming M_i and M_j situate a deformed object parts, then they are likely to recover two matches $M_{\bar{i}}$ and $M_{\bar{j}}$ sharing similar local transformations if local geometric deformation across images are relieved. This changes the spatial match distance from $d_{\bar{i}\bar{j}}^s$ to d_{ij}^s (Fig. 3). Moreover,

$$d_{ij}^s \le d_{\bar{i}\bar{j}}^s + d_{i\bar{i}}^t + d_{j\bar{j}}^t, \tag{3}$$

with $d_{\bar{i}\bar{j}}^s = \frac{1}{2}(|c_{\bar{i}}c_{\bar{j}}| + |c'_{\bar{i}}c'_{\bar{j}}|)$, $d_{i\bar{i}}^t = \frac{1}{2}(|c_i c_{\bar{i}} + c'_i c'_{\bar{i}}|)$, and $d_{j\bar{j}}^t = \frac{1}{2}(|c_j c_{\bar{j}} + c'_j c'_{\bar{j}}|)$. The latter two terms define the transformational distance between M_i and $M_{\bar{i}}$ as well as M_j and $M_{\bar{j}}$ respectively. Equation 3 holds because $|c_i c_j| \le |c_{\bar{i}}c_{\bar{j}}| + |c_i c_{\bar{i}}| + |c_{\bar{j}}c_j|$ and $|c'_i c'_j| \le |c'_{\bar{i}}c'_{\bar{j}}| + |c'_i c'_{\bar{i}}| + |c'_{\bar{j}}c'_j|$.

Next, we study first two cases to set a specific value for h_s. (1) The local region surrounding M_i is heavily deformed. Then a reliable neighbor M_j ought to be very close to a correct match M_i. In extreme cases, $d_{\bar{i}\bar{j}}^s \approx 0$ and $d_{ij}^s \le d_{i\bar{i}}^t + d_{j\bar{j}}^t$ actually encodes the deformation-determined spatial distance between M_i and its neighbor M_j. Consequently, we set $h_s = 2h_t$ as h_t is the upper bound of $d_{i\bar{i}}^t$ and $d_{j\bar{j}}^t$ according to Eqs. 1 and 2. (2) The local region surrounding M_i is not

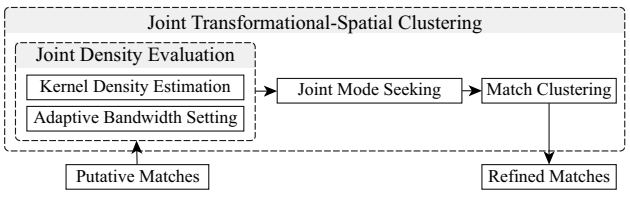

Fig. 2 The pipeline of match clustering in the joint domain. Please refer to the text for details

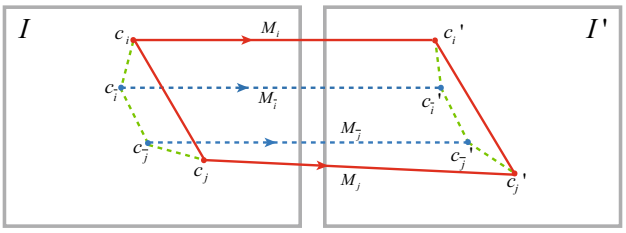

Fig. 3 Matches M_i and M_j deformed from $M_{\bar{i}}$ and $M_{\bar{j}}$, which map nearby object parts and share similar local transformations

deformed. In this case, the transformational constraint $d_{ij}^t \le h_t$ can effectively reject wrong neighbors inside the matching object, and $d_{ij}^s \le h_s$ is expected to identify the mismatch M_j sharing similar transformation with M_i but staying outside the object. Thus h_s is very flexible in value setting because the target wrong matches can assume to stay much more distant to M_j than those correct ones. Here, we set $d_{ij}^s \le h_s = 2h_t$ for picking out correct neighbors of M_i and keeping consistent with the former case. All the remaining cases can be seen as intermediate ones between the two extreme cases discussed. Therefore, we set $h_s = 2h_t$ for all cases.

Setting h_t To set h_t, we observe that the density of all matches may take few uniform values if h_t gets too small or too large, e.g., $\rho_i = \frac{1}{N}$ for an arbitrary match M_i when $h_t = 0$ or $h_t \ge h_{t\_\max}$ with $h_{t\_\max}$ being the maximum distance between two arbitrary matches. In such cases, the density distribution of the joint space is not varying smoothly, which obscures the boundaries and modes of different clusters. By contrast, a proper value of h_t would allow the neighborhood densities of matches in a cluster to vary coherently from its mode to verge area. We achieve this by introducing density entropy and minimizing it in the joint space to determine h_t as

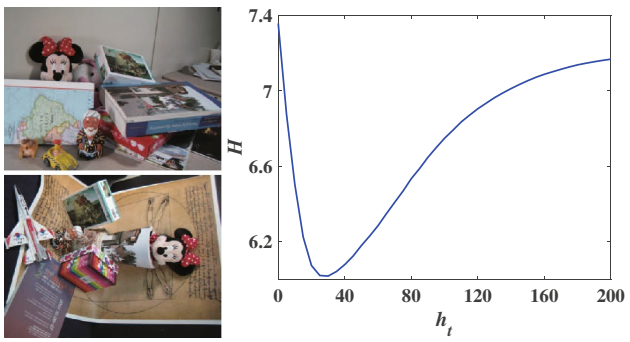

Fig. 4 The changing of entropy with varying h_t on an image pair

$$\arg\min_{h_t} \; H = -\sum_{i=1}^{N} \rho_i \log \rho_i, \tag{4}$$

where ρ_i is defined in Eq. 1. Basically, a small entropy H is obtained when the densities $\rho_i, i = 1...N$ are assigned with diverse values for different matches, which results in a more smoothly varying density distribution. We show an example (Fig. 4) illustrating the change of H with varying h_t after setting $h_s = 2h_t$. It can be observed that the value of H decreases rapidly with the increasing of h_t, until it reaches the local minimum. Thereafter, H increases steadily as h_t becomes large. Specifically, the largest entropy is reached when $h_t = 0$ with each match forming a single cluster, or $h_t \ge h_{t\_\max}$ with all matches assigned into the same cluster. Finally, h_t producing the minimum H is obtained in the interval $[0, h_{t\_\max}]$ by gradient descent. The density evaluation in Eq. 1 is settled after the optimization of h_t.

3.2 Density based joint mode seeking

A mode of the joint transformational-spatial domain is the local maxima of underlying probability density distribution according to the conventional paradigm. It can be defined alternatively as the point with density larger than all its neighbors considering the discrete nature of the joint domain. Formally, the modes are

$$\mathcal{M} = \{M_i \mid \rho_i > \rho_j, \; \forall M_j \in \mathcal{N}_i, \; i = 1, \dots, N\}, \tag{5}$$

where $\mathcal{N}_i$ denotes the neighbors of M_i,

$$\mathcal{N}_i = \{M_j \mid d_{ij}^t \le h_t, \; d_{ij}^s \le h_s, \; j = 1, \dots, N\}. \tag{6}$$

$\mathcal{N}_i$ is consistent with the density estimator in Eq. 1 since all points in $\mathcal{N}_i$ are involved in evaluating the density of M_i.

Seeking the modes is done by starting from each point and recursively shifting to the one with maximum density in the neighborhood until convergence, where each stationary point corresponds to a local mode. The process can also be regarded as to find the most reliable matches, whereby their confidences are supported by respective neighboring points.

3.2.1 Convergence analysis

The movement M_i making for mode seeking can be equivalently expressed as

$$M_{\zeta_i} = \arg\max_{M_j \in \mathcal{N}_i} \; p_{ij}(\rho_j - \rho_i), \tag{7}$$

with $p_{ij} = \frac{1}{|\mathcal{N}_i|}$ denoting the possibility of M_i shifting to one of its neighbors M_j. Consequently, $p_{i\zeta_i}(\rho_{\zeta_i} - \rho_i)$ defines the highest expectable density increment, and moving to M_{ζ_i}

witnesses the maximum increase in the density. In addition, the density keeps increasing if successive shifts are conducted along the steepest ascent direction. The move step can never be too large to jump over the stationary point considering the discrete structure of the joint domain. Thus the mode seeking process is bound to converge to local modes after finite steps.

3.3 Match clustering in the joint domain

The mode seeking process finds a path for each point shifting to a local mode in the joint transformational-spatial domain. The domain clustering can then be accomplished by grouping together the points reaching the same mode point. Moreover, two matches in the same cluster may associate with the same local feature in one image and thus become conflicted correspondences. In this case, the one with smaller density is pruned, whereby the one-to-one matching constraint is enforced within each cluster. This idea can better match repetitive patterns between images in comparison with existing approaches [21, 33] which apply a globally one-to-one matching constraint for wrong match pruning. Finally, the matches in significant clusters with more than a predefined number points (8 in our experiments) are kept and the noisy matches in the remaining clusters are discarded. The whole algorithm is summarized in Algorithm 1.

Algorithm 1: Joint Transformational-Spatial Clustering for Feature Matching

Input : Putative match set $\{M_i\}_{i=1}^{N}$; Transformational and spatial distance matrices of putative matches d_{ij}^{t} and d_{ij}^{s}; Minimum cluster member size $T = 8$.

Output: The clusters with member size larger than T.

1 Set $h_s = 2h_t$.;
2 Adaptively evaluate h_t (Eq. 4).;
3 Evaluate the density ρ_i of M_i (Eq. 1).;
4 Discover the mode of each M_i (Sec. III-B).;
5 Group $\{M_i\}_{i=1}^{N}$ based on their modes and prune conflicted matches (Sec. III-C).;

3.3.1 Implementation details and time complexity

The main time overhead lies in density estimation, which requires neighbor searching for each match. We partition input image planes into multiple grids and hash the features in each grid for speeding up. The spatial neighbors of each match are first searched in neighboring grids based on the spatial bandwidth and further verified using the transformational bandwidth.

The time complexity of density estimation is kN for N matches with k average search trials. Computing the bandwidths usually requires to run density estimation for t ($t < 10$ in our experiments) times. Besides, each match compares its density with its k_1 neighbors during mode seeking. Hence, the overall time complexity is $O((tk + k_1)N)$, which is lower than existing methods [5, 10] demanding for $O(N^2)$ to compute distances among initial matches, except for exclusive high computational steps.

4 Experiments

Three experiments are conducted in this section. First, multiple object matching is conducted on a benchmark dataset. Second, duplicate object matching in a single image is tested on a public dataset for verifying its robustness in matching repetitive patterns. Third, the performance of object retrieval is reported by comparison with existing methods.

All experiments are run on a laptop with Intel Core *i*5 2.67 GHz CPU and 4 GB memory, with the proposed approach implemented in Matlab. Besides, local features are extracted with VLfeat [41].

4.1 Multiple objects matching

The popular SNU dataset [42] is utilized as a test bed for this experiment. It consists of six pairs of images with each image pair containing at least 2 pairs of common objects. Finding object correspondences in the dataset is quite challenging because each pair of object instance undergoes random geometric distortions, background clutters, photometric variations and partially occlusions, etc.

The putative matches are established by checking the distance ratio of features with its nearest and second nearest features [1]. Correct matches are manually specified for quantitative performance evaluation. The number of putative and correct feature correspondences are shown in Table 1.

4.1.1 Competitors

We compare our approach against several algorithms, including the hough voting (HV) [5], common visual pattern discovery (CVP) [10], agglomerative correspondence clustering (ACC) [8], discrete tabu search for graph matching (DTS) [33], ensemble of weak geometric relation checking (EWGR) [21], locality preserving matching (LPM) [18], grid-based motion statistics (GMS) [19], neighborhood mining network (NMNet) [13] and learning for mismatch removal (LMR) [38]. HV and EWGR are implemented according to published papers, while the publicly available codes of the remaining methods shared by the respective authors are directly used. For fair comparison, the

transformational incompatibility between each correspondence pair is computed by the STE distance for each method if applicable.

4.1.2 Evaluation metrics

The 1-precision vs. recall curves is used to evaluate the performance of all algorithms. The controlling parameters of most methods are set with 10 varying values to obtain increasing recall rate. For ACC and our approach, we plot the kth points of the curve using the recall and precision of all the top k largest clusters. Besides, for NMNet and LMR, no proper control parameter is available, and thus only one point is shown to demonstrate the performance of each method.

4.1.3 Results

The quantitative results (Fig. 5) show that HV, EWGR and our approach slightly outperform other methods on image pairs with low transformational variations, including "Books", "Bulletins" and "Toys". This suggests integrating spatial proximity can complement the transformational proximity for better pruning wrong matches, even if the latter one alone performs well in the case object pairs in input images are approximately affinely-transformed.

For all the remaining image groups, "Jigsaws", "Mickeys" and "Minnies", they undergo more severe viewpoint changes and object distortions, which enlarge the transformational difference among neighboring correct matches. In these cases, our method identifies more correct matches than the others. On the contrary, HV evaluates the density of a match with neighbors inside the same BPLR, which may misidentify wrong matches if the shared BPLR crosses object boundaries or contains uniform background as in the case of "Jigsaws". Besides, the spatial proximity based approaches, namely, GMS and LPM, usually cannot guarantee to identify match inliers with high accuracy, but it is able to prune a considerable amount of wrong matches while saving most of the correct ones when the parameters are set properly. This enables them to be good preprocessing options for enhancing the matching performance of other methods. Finally, LMR and NMNet both incorporates neighborhood information in a learning framework for match reliability prediction. Among the two, NMNet only considers the transformational proximity for neighborhood feature extraction, which limits their performance when multiple object patterns with different transformations exist. LMR finds the spatial neighbors for each match and extracts geometric features for learning, which delivers better performance but also falls short of robustness when the object distortion turns severer. Overall, our approach performs consistently well in all cases, suggesting its effectiveness in identifying correct matches by integrating both spatial and transformational proximities in the manner of the proposed joint domain clustering.

To verify the effectiveness of the bandwidth selection process, we show the average precision and recall rate with standard deviation for all 6 image pairs by taking h_t as different values in Table 2. As shown, small h_t leads to matching performance at a high precision but low recall rate, while large h_t results in a high recall but low precision rate. By contrast satisfactory performance is achieved when the adaptively tuned values are used. Figure 6a–c show the matching results of "Toys" using different h_t. Obviously, small h_t leads to shrinking match clusters, and each cluster maps part of a object pair, and large h_t results in expansive clusters with each cluster links a region pair overexpanding the true object area. By contrast, the adaptively tuned h_t presents visually more plausible matching results.

The average time overheads of different methods are also collected (Fig. 7). Feature extraction and putative correspondence constructed shared by all methods are excluded from the time statistics. HV, ACC and CVP generally requires relatively longer running time, while GMS, LMR, NMNet and our approach all can match two images within 1 s. Among the last four ones, the proposed method is slightly slower than the other three, with the largest margin less than 0.4 s. Given the final precision and recall rate reported, our approach outperforms 6 out of 9 competitors in terms of both effectiveness and efficiency, and achieves considerable performance gain over the remaining three with a slightly higher time overhead cost.

4.2 Matching duplicate objects in a single image

This experiment is done with the single image test set of identical object segmentation [43]. It contains 10 images, with 8 containing objects with more than 2 duplicates selected for testing. Initial matches are computed by finding 3 nearest neighbors for each feature in the descriptor space. Matches with two keypoints of distance less than 10 pixels are deemed unreliable and discarded. The odd and even rows of the first four columns in Fig. 8 show the original images

Table 1 The number of putative and correct matches of image groups in the SNU dataset

	Books	Bulletins	Jigsaws	Mickeys	Minnies	Toys
#correct	1024	419	182	118	200	681
#putative	3036	2202	2347	1873	1572	2960

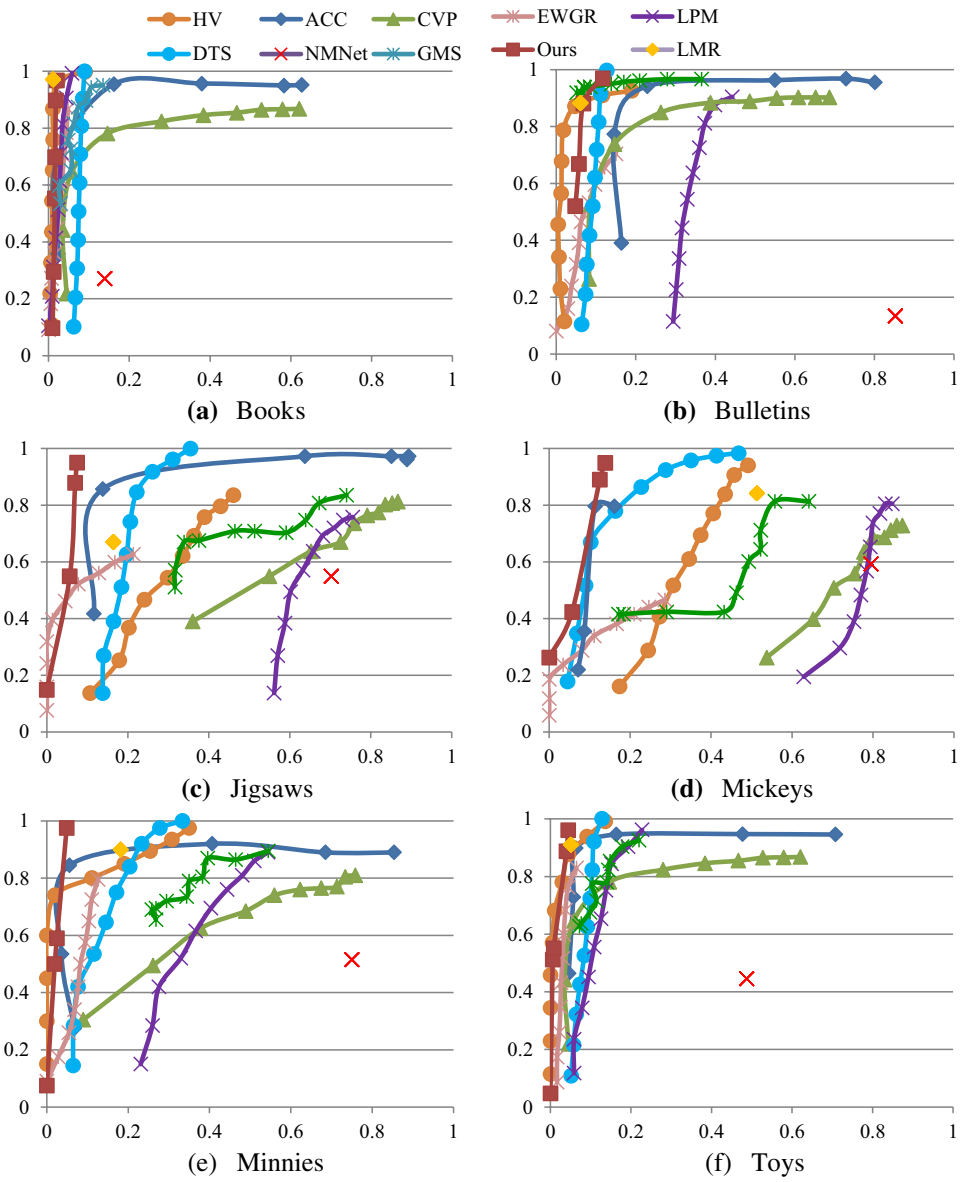

Fig. 5 Performance comparison of multiple object matching in terms of 1-precision (horizontal axes) and recall (vertical axes) on the SNU dataset

Table 2 Average precision and recall with standard deviation on the SNU dataset when setting h_t as adaptively tuned value Tuned as well as Tuned −10 and Tuned +10 respectively

	Tuned	Tuned −10	Tuned +10
Precision (%)	0.927 ± 0.042	0.961 ± 0.015	0.819 ± 0.072
Recall (%)	0.960 ± 0.008	0.876 ± 0.055	0.977 ± 0.009

and our results respectively. Two results of DTS [33] and ACC [8] are displayed in the last column.

Our approach can successfully reveal most of the correspondences among duplicated objects, with each linked by one or more matching clusters. ACC delivers clusters with wrong matches scattered in the image background due to no spatial proximity check. DTS on the other hand, presents sparse matches for some of the duplicated object pairs, partly because each feature can only be matched once under the global one-to-one matching constraint. Evidently, distinguishing different duplicated pairs for DTS requires further postprocessing.

4.3 Object retrieval

The dataset for this experiment is a subset of the Oxford dataset [44], which consists of 748 landmark images, including 55 query images depicting 11 different architectures.

(a) $h_t = 35$ (654 / 684) **(b)** $h_t = 20$ (415 / 432) **(c)** $h_t = 60$ (667 / 793)

Fig. 6 Matching results with different h_t on an image pair. The number of correct and returned matches are shown in the parentheses

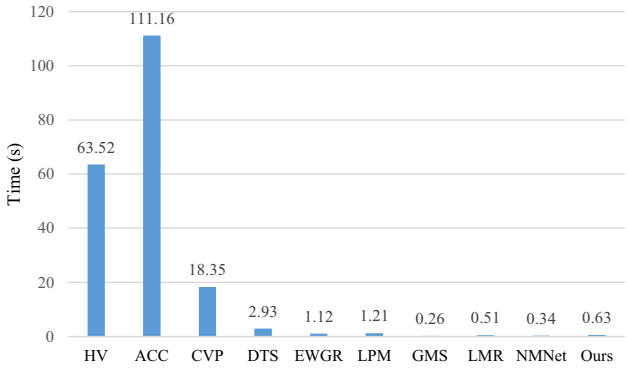

Fig. 7 Average running time (in seconds) on the SNU dataset

For each query image, a rectangle is defined to specify the retrieval region, and all the remaining images are classified as *good*, *ok*, *bad* and *junk*. The retrieval performance is measured over all the query images by mean average precision (mAP), as described in [45].

Several retrieval algorithms are taken for comparison, including bag-of-visual-word (BoVW) [45], spatial pyramid matching (SPM) [46], improved k-nearest neighbor searching (k-NNI) [47], feature matching with RANSAC verification (RANSAC). For BoVW and SPM, 10 K visual words are trained with k-means using 1 M features randomly selected from feature sets of all images. For the two spatial verification based approaches, RANSAC and ours, images are first ranked according to the number of features successfully matched to query image, and then the unmatched images are re-ranked using the k-NNI algorithm. Besides, the performance of our approach without re-ranking is also reported.

Table 3 shows the comparison result. Overall, RANSAC and our approach perform better than k-NNI, which indicates

Fig. 8 Matching multiple duplicate objects in a single image. The first four columns show the original images and our results with different color depicting different matching clusters. The last column shows the results of DTS (odd rows) [33] and ACC (even rows) [8]

Table 3 Retrieval performance in mAP on the subset of the Oxford dataset [45]

BoVW	SPM	*k*-NNI	RANSAC-Rerank	Ours	Ours-Rerank
0.501	0.515	0.800	0.826	0.836	0.852

the necessity for integrating geometric verification during feature matching. Finally, our approach further outperforms RANSAC even though most of the query images only contain one landmark, which suggests our approach can better handle object matching under complex conditions.

5 Conclusions

In this paper, we introduce a density-based clustering approach in the transformational-spatial joint domain for feature matching. Unlike existing methods incorporating either transformational or spatial proximity among matches or checking both proximities to refine the putative match set, we propose to identify match inliers by grouping together transformationally and spatially coherent matches through density estimation and mode-seeking based clustering in the underlying joint domain. To enhance the scalability of our approach, the bandwidths for measuring both proximities during density estimation is adaptively tuned. Experiments on multiple datasets show that the propose approach outperforms existing methods in the task of multiple object matching across images, duplicate object matching in a single image, as well as object retrieval.

Acknowledgements This work is co-supported by the National Natural Science Foundation of China (61502005), Anhui Science Foundation (1608085QF129).

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
2. Schonberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M.: Comparative evaluation of hand-crafted and learned local features. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1482–1491 (2017)
3. Torr, P.H., Murray, D.W.: The development and comparison of robust methods for estimating the fundamental matrix. Int. J. Comput. Vis. **24**(3), 271–300 (1997)
4. Cho, M., Lee, K.M.: Progressive graph matching: making a move of graphs via probabilistic voting. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 398–405 (2012)
5. Chen, H.-Y., Lin, Y.-Y., Chen, B.-Y.: Robust feature matching with alternate Hough and inverted Hough transforms. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2762–2769 (2013)
6. Adamczewski, K., Suh, Y., Lee, K.M.: Discrete tabu search for graph matching. In: Proc. IEEE Int'l Conf. on Computer Vision, pp. 109–117 (2015)
7. Zanfir, A., Sminchisescu, C.: Deep learning of graph matching. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2684–2693 (2018)
8. Cho, M., Lee, J., Lee, K.M.: Feature correspondence and deformable object matching via agglomerative correspondence clustering. In: Proc. IEEE Int'l Conf. on Computer Vision, pp. 1280–1287 (2009)
9. Cho, M., Lee, K.M.: Mode-seeking on graphs via random walks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 606–613 (2012)
10. Liu, H., Latecki, L.J., Yan, S.: Fast detection of dense subgraphs with iterative shrinking and expansion. IEEE Trans. Pattern Anal. Mach. Intell. **35**(9), 2131–2142 (2013)
11. Wang, L., Tang, D., Guo, Y., Do, M.N.: Common visual pattern discovery via nonlinear mean shift clustering. IEEE Trans. Image Process. **24**(12), 5442–5454 (2015)
12. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2666–2674 (2018)
13. Zhao, C., Cao, Z., Li, C., Li, X., Yang, J.: Nm-net: mining reliable neighbors for robust feature correspondences. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 215–224 (2019)
14. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 6148–6157 (2017)
15. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: depth and motion network for learning monocular stereo. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5038–5047 (2017)
16. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: learning of structure and motion from video (2017). arXiv:1704.07804
17. Zhang, Z., Lee, W.S.: Deep graphical feature learning for the feature matching problem. In: Proc. IEEE Int'l Conf. on Computer Vision, pp. 5087–5096 (2019)
18. Ma, J., Zhao, J., Jiang, J., Zhou, H., Guo, X.: Locality preserving matching. Int. J. Comput. Vis. **127**(5), 512–531 (2019)
19. Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., Cheng, M.-M.: Gms: grid-based motion statistics for fast, ultra-robust feature correspondence. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4181–4190 (2017)
20. Jiang, X., Ma, J., Jiang, J., Guo, X.: Robust feature matching using spatial clustering with heavy outliers. IEEE Trans. Image Process. **29**, 736–746 (2019)
21. Wu, X., Kashino, K.: Robust spatial matching as ensemble of weak geometric relations. In: Proc. British Machine Vision Conf., pp. 25–1 (2015)
22. Ferrari, V., Tuytelaars, T., Van Gool, L.: Simultaneous object recognition and segmentation from single or multiple model views. Int. J. Comput. Vis. **67**(2), 159–188 (2006)
23. Leng, C., Zhang, H., Li, B., Cai, G., Pei, Z., He, L.: Local feature descriptor for image matching: a survey. IEEE Access **7**, 6424–6434 (2018)
24. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: Proc. European Conf. on Computer Vision, pp. 467–483 (2016)

25. Bai, X., Zhang, T., Wang, C., Abd El-Latif, A.A., Niu, X.: A fully automatic player detection method based on one-class svm. IEICE Trans. Inf. Syst. **96**(2), 387–391 (2013)
26. Gad, R., Abd El-Latif, A.A., Elseuofi, S., Ibrahim, H.M., Elmezain, M., Said, W.: Iot security based on iris verification using multi-algorithm feature level fusion scheme. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1–6. IEEE (2019)
27. Peng, J., Li, Q., Abd El-Latif, A.A., Niu, X.: Linear discriminant multi-set canonical correlations analysis (ldmcca): an efficient approach for feature fusion of finger biometrics. Multimed. Tools Appl. **74**(13), 4469–4486 (2015)
28. Peng, J., Wang, N., Abd El-Latif, A.A., Li, Q., Niu, X.: Finger-vein verification using Gabor filter and sift feature matching. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 45–48. IEEE (2012)
29. Wang, L., Chen, B., Xu, P., Ren, H., Fang, X., Wan, S.: Geometry consistency aware confidence evaluation for feature matching. Image Vis. Comput. **103**, 103984 (2020)
30. Zhang, T., Abd El-Latif, A.A., Wang, N., Li, Q., Niu, X.: A new image segmentation method via fusing ncut eigenvectors maps. In: Fourth International Conference on Digital Image Processing (ICDIP 2012), vol. 8334, p. 833430. International Society for Optics and Photonics (2012)
31. Jing, H., He, X., Han, Q., Abd El-Latif, A.A., Niu, X.: Saliency detection based on integrated features. Neurocomputing **129**, 114–121 (2014)
32. Wang, L., Zhen, H., Fang, X., Wan, S., Ding, W., Guo, Y.: A unified two-parallel-branch deep neural network for joint gland contour and segmentation learning. Future Gener. Comput. Syst. **100**, 316–324 (2019)
33. Suh, Y., Adamczewski, K., Lee, K.M.: Subgraph matching using compactness prior for robust feature correspondence. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5070–5078 (2015)
34. Lee, J., Cho, M., Lee, K.M.: Hyper-graph matching via reweighted random walks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1633–1640 (2011)
35. Wan, S., Xia, Y., Qi, L., Yang, Y.-H., Atiquzzaman, M.: Automated colorization of a grayscale image with seed points propagation. IEEE Trans. Multimed. (2020)
36. Gao, Z., Li, Y., Wan, S.: Exploring deep learning for view-based 3d model retrieval. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **16**(1), 1–21 (2020)
37. Gao, Z., Xuan, H.-Z., Zhang, H., Wan, S., Choo, K.-K.R.: Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet Things J. **6**(6), 9280–9293 (2019)
38. Ma, J., Jiang, X., Jiang, J., Zhao, J., Guo, X.: Lmr: learning a two-class classifier for mismatch removal. IEEE Trans. Image Process. **28**(8), 4045–4059 (2019)
39. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
40. Marron, J., Nolan, D.: Canonical kernels for density estimation. Stat. Probab. Lett. **7**(3), 195–199 (1988)
41. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
42. Cho, M., Shin, Y.M., Lee, K.M.: Co-recognition of image pairs by data-driven monte Carlo image exploration. In: Proc. European Conf. on Computer Vision, pp. 144–157 (2008)
43. Cho, M., Shin, Y.M., Lee, K.M.: Unsupervised detection and segmentation of identical objects. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1617–1624 (2010)
44. Lenc, K., Gulshan, V., Vedaldi, A.: Vlbenchmkars (2011). http://www.vlfeat.org/benchmarks/
45. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
46. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
47. Jégou, H., Douze, M., Schmid, C.: Exploiting descriptor distances for precise image search. Ph.D. dissertation. INRIA (2011)