Co-Segmentation Guided Hough Transform for Robust Feature Matching

Hsin-Yi Chen, Yen-Yu Lin, *Member, IEEE,* and Bing-Yu Chen, *Senior Member, IEEE*

Abstract—We present an algorithm that integrates image co-segmentation into feature matching, and can robustly yield accurate and dense feature correspondences. Inspired by the fact that correct feature correspondences on the same object typically have coherent transformations, we cast the task of feature matching as a density estimation problem in the homography space. Specifically, we project the homographies of correspondence candidates into the parametric Hough space, in which geometric verification of correspondences can be activated by voting. The precision of matching is then boosted. On the other hand, we leverage image co-segmentation, which discovers object boundaries, to determine relevant voters and speed up Hough voting. In addition, correspondence enrichment can be achieved by inferring the concerted homographies that are propagated between the features within the same segments. The recall is hence increased. In our approach, feature matching and image co-segmentation are tightly coupled. Through an iterative optimization process, more and more correct correspondences are detected owing to object boundaries revealed by co-segmentation. The proposed approach is comprehensively evaluated. Promising experimental results on four datasets manifest its effectiveness.

Index Terms—Image feature matching, correspondence problems, Hough transform, co-segmentation, energy minimization.

1 Introduction

Establishing correspondences among two or more images has attracted great attention in the field of computer vision. Being a key component for image analysis, they are essential for a wide range of applications, such as object recognition [1], image retrieval [2], 3D reconstruction [3], image enhancement [4] and patch-based synthesis [5]. Despite the great applicability, two main difficulties hinder the advance in establishing correspondences of high quality. The predominant paradigm starts from analyzing local features to yield the candidates of correspondences. Although much progress has been made on the local feature descriptors, methods of this category often suffer from perspective changes, illumination changes, or cluttered backgrounds in the images. Thus the corrupted correspondences lead to *low precision* in feature matching. Many advanced methods, such as [6], [7], [8], [9], tackle this problem by ensuring the geometric consistency, which typically do not scale well due to high-order geometric checking. Hence, they often work on preselected, small subset of correspondence candidates, and result in *low recall.*

In this paper, we aim to address the aforementioned problems simultaneously. The proposed approach is developed upon the insight that nearby features on the same object typically share similar homographies if they are matched correctly. It follows that their homographies tend to gather together in the *transformation space*. On the other hand, each wrong matching is usually wrong in its own way. It implies that the *density* of each correspondence in the transformation space can verify its correctness. We leverage this property and cast the task of feature matching as a density estimation problem. Specifically, we identify correct correspondences by comparing the densities among mutually exclusive correspondences, those violating one-to-one constraints. That is, *Hough voting* for geometric checking is realized by computing the densities of homographies from correspondence candidates. Besides, it is also allowed to compile plausible correspondence candidates by investigating the concerted homographies of locally grouped features. We consider it as an inverted process of Hough voting, and use it to dynamically enrich potential correspondences for each feature. Furthermore, we show that both Hough voting and its inverted variant can be improved by integrating with image co-segmentation. The hypotheses of object boundaries, discovered by co-segmentation, facilitate the identification of relevant voters and locally grouped features.

The proposed approach carries out *Hough transform* and *inverted Hough transform* alternately to establish robust feature correspondences. It can distinguish itself with the following main contributions. First, every correspondence candidate is projected into a Hough space spanned by transformations. With the aid of image co-segmentation, correspondences associated to features within the same segments are considered for Hough voting. In this way, geometric verification boosts the precision of matching. The process of verification is also significantly speeded up, since only a small, relevant subset of correspondence candidates is taken into account in density estimation. Second, an inverted Hough transform is developed, which can recommend each feature point additional correspondences by investigating high-density homographies from features covered by the

- *H.-Y. Chen and B.-Y. Chen are with National Taiwan University.*
  E-mail: fensi@cmlab.csie.ntu.edu.tw; robin@ntu.edu.tw
- *Y.-Y. Lin is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan.*
  E-mail: yylin@citi.sinica.edu.tw

same segments. It considerably increases the recall of feature matching. Third, the proposed approach couples Hough voting and its inverted variant with image co-segmentation. Through an iterative optimization process, more and more correct correspondences are progressively detected owing to object boundaries revealed by co-segmentation. Finally, our approach is comprehensively evaluated and compared to the state-of-the-art systems on several benchmark datasets. The superior performance demonstrates its effectiveness.

2 Related Work

In this section, related works are briefly reviewed.

2.1 Matching via feature descriptor

Point-to-point matching with local feature descriptors is a principal way for correspondence problems. Some of notable researches, such as [10], [11], [12], have brought about significant progress in this area. Although these local descriptors are distinctive and powerful, no descriptor in general is sufficient for handling variation caused by complex combinations of nonrigid deformations, illumination and pose changes, in nowadays vision applications.

2.2 Matching via graph partition

One way to address matching ambiguity with additional geometric checking is to cast feature correspondence as a graph matching problem. Promising results via graph matching have been demonstrated [13], [9]. However, these methods typically work well on one common object with simple backgrounds, and do not deal with the cases where multiple sets of common features appear. As mentioned in [14], graph matching is sensitive to corrupt correspondences and outliers. In addition, high computational cost may restrict its applicability, especially when solving a generalized eigenvalue problem is required.

2.3 Matching via clustering

Research efforts on clustering-based mechanisms have been made to handle unconstrained matching cases. Bottom-up clustering can adopt locally adaptive constraints to aggregate coherent bundles of matches. Cho et al. [15] carried out object-based image matching via hierarchical agglomerative clustering. Yacov et al. [4] adopted a coarse-to-fine scheme and the coherence property of images to achieve dense matching. Liu and Yan [14] proposed a top-down clustering approach to detect dense neighborhoods on an affinity graph, and found common visual patterns. Despite the effectiveness, one major weakness of these methods lies in the high computational cost of clustering. Moreover, the optimal cluster numbers, criteria of cluster merging, and similarity thresholds typically vary from image to image.

2.4 Matching via voting

RANSAC [16], a geometric verification model, and its variants, such as [17], can be incorporated with local descriptors to enhance the performance. Yuan et al. [18] treated each correspondence as a voter, and maintained an affinity matrix to encode how these correspondences vote each other according to their compatibilities. Tolias and Avrithis [19] offered a variant of Hough transform for multi-object matching. They ranked the correspondences by adopting the mechanism of *pyramid match* [20]. Their method evenly quantizes the transformation space for fast matching. However, the transformations of correct correspondences often distribute irregularly. It may result in accuracy degradation. Our prior work [21] is a voting-based system. It has the advantage of voter selection for speeding up voting. It also supports correspondence enrichment.

In this work, we further integrate image co-segmentation into feature matching. The object hypotheses by co-segmentation facilitate not only relevant voter identification but also plausible correspondence recommendation. In our approach, feature matching and image co-segmentation are nicely coupled, and jointly lead to better performance.

2.5 Correspondence enrichment

Most feature correspondence methods work with a small, pre-selected subset of correspondences. Correspondence enrichment hence becomes an important task. Match-growing methods, [22], [23], propagate individual matches to nearby regions based on local appearance, but their performances heavily depend on the quality of initial matching. On the other hand, Čech et al. [24] developed a region-growing algorithm to distinguish correct and incorrect correspondences. Instead of using fixed shapes of measurement regions, they progressively grew regions by co-segmentation until reliable correctness identification can be reached. Cho and Lee [25] instead described a progressive graph matching framework to enrich initial matching. However, the yielded correspondences by their approach are biased to the density of features, and may be noisy due to diverse feature distributions in the two matched images. In contrast, our method works on feature bundles guided by co-segmentation, so the concerted transformations with high probability are transferred through mutually relevant features. It turns out that the information can be propagated more efficiently and the resulting candidates of correspondences are much more targeted.

2.6 Image co-segmentation

One line to address image co-segmentation is based on *Markov random field* (MRF). Methods of this category, e.g., [26], [27], often consist of an MRF model over each image, and a global consistency term among the foreground histograms. Another class of co-segmentation methods, such as [28], [29], uses graph-partitioning to obtain foreground/backgound labels. Recently, more strategies have been introduced to address image co-segmentation. For example, Faktor and Irani [30] considered a co-segment good if it can be easily composed by other co-segments, but is difficult to compose from the remaining parts. Wang et al. [31] extracted the *consistent functional maps* between image pairs to transfer segmentation across images. Sun and Ponce [32] carried out image co-segmentation with the aid of the detected discriminative parts of the object category. Despite these research efforts and positive results shown in [33], [34], image co-segmentation, in which common objects may exhibit high appearance variations, remains very challenging. As mentioned previously, an object often covers correspondences with coherent homographies. It also means that two adjacent regions with coherent homographies tend to belong to the same object. Our approach uses this property to improve co-segmentation with the matched correspondences.

3 Problem Definition

Given two images I^P and I^Q, two sets of feature points, $V^P = \{v_i^P\}_{i=1}^{N^P}$ and $V^Q = \{v_i^Q\}_{i=1}^{N^Q}$, are respectively detected. The region and the center of feature $v_i \in V^P \cup V^Q$ are denoted by S_i and $\mathbf{x}_i$, respectively. The appearance of v_i is described by descriptor $\mathbf{u_i}$, and its orientation θ_i is estimated by the dominant orientation in the gradient histogram [35]. The product $\mathcal{C} = V^P \times V^Q$ represents all the possible correspondences. Our goal is to find the correct correspondences in $\mathcal{C}$ as many as possible.

3.1 Transformation space

The local shape and the position of feature v_i can be described by a 3×3 matrix $T(v_i)$, which specifies an affine transform of v_i with regards to the *normalized patch* [35]:

$$T(v_i) = \left[\begin{array}{cc} A(v_i) & \mathbf{x}_i \\ \mathbf{0}^\top & 1 \end{array} \right] \in \mathbb{R}^{3\times 3}, \tag{1}$$

where $A(v_i) \in \mathbb{R}^{2\times 2}$ is a non-singular matrix.

Given a feature pair $v_i^P \in V^P$ and $v_{i'}^Q \in V^Q$, the *relative transformation* $H_{ii'}$ from v_i^P to $v_{i'}^Q$ can be derived as

$$H_{ii'} = T(v_{i'}^Q) * T(v_i^P)^{-1}. \tag{2}$$

In this work, we represent a feature *correspondence* as a triplet $m_{ii'} = (v_i^P, v_{i'}^Q, H_{ii'})$, i.e., two features in the opposite images and their relative transformation. The correspondence can be also symmetrically specified as $(v_{i'}^Q, v_i^P, H_{i'i}^{-1})$. As $H_{ii'}$ is a 6-dof affine homography, the geometry configuration of $m_{ii'}$ can be considered as a point in the 6-dimensional transformation space. It should be mentioned that the local transformation in Eq. (2) is not fully perspective but affine only. It has a better match to widely-used affine invariant detectors, such as Hessian-Affine detector [35]. Therefore, the computed transformations may not be reliable for objects with a significant 3D structure or under severe perspective distortion.

3.2 Distance metric in the transformation space

Given two correspondences $m_{ii'} = (v_i^P, v_{i'}^Q, H_{ii'})$ and $m_{jj'} = (v_j^P, v_{j'}^Q, H_{jj'})$, the *projection error* of $m_{jj'}$ with respect to $m_{ii'}$ can be defined as

$$d_{jj'|ii'} = ||\mathbf{x}_{j'}^Q - \rho(H_{ii'} \left[\begin{array}{c} \mathbf{x}_j^P \\ 1 \end{array} \right])||, \tag{3}$$

$$\text{where } \rho(\left[\begin{array}{ccc} a & b & c \end{array} \right]^\top) = \left[\begin{array}{cc} a/c & b/c \end{array} \right]^\top. \tag{4}$$

It checks if $H_{ii'}$ projects $\mathbf{x}_j^P$ around $\mathbf{x}_{j'}^Q$.

For a pair of correspondences $m_{ii'}$ and $m_{jj'}$, they are considered compatible if the corresponding homographies are similar. We hence adopt the *re-projection error* for dissimilarity measure, i.e.,

$$d(m_{ii'}, m_{jj'}) = \frac{1}{4}(d_{jj'|ii'} + d_{ii'|jj'} + d_{j'j|i'i} + d_{i'i|j'j}). \tag{5}$$

Note that it is symmetric and can serve as the distance function for correspondences in the transformation space.

4 The Proposed Approach

We investigate the geometric distribution of feature matchings to determine their correctness with the aid of *co-segmentation*. With the aim at identifying *accurate* and *dense* correspondences, the proposed approach carries out alternate Hough and inverted Hough voting. While the former discovers the consistent homographies by projecting correspondences into the transformation space, the latter recommends potential correspondences by referencing the concerted homographies within the same segments. The procedure is repeated iteratively until convergence.

To begin with, we first describe the construction of initial correspondences in Section 4.1. Next, co-segmentation is performed. With the obtained co-segments, the Hough transform for geometric verification and inverted Hough transform for correspondence enrichment are introduced in Sections 4.2 and 4.3, respectively. In Section 5, we will describe how the generated correspondences improve the image co-segmentation to lead to better co-segments, which then help feature matching.

4.1 Initial correspondence candidates

Our approach starts from the construction of initial correspondence candidates. For each feature $v_i^P \in V^P$, we find its r potential matchings $\{v_{i_k}^Q\}_{k=1}^r$ in V^Q according to the descriptor similarity, i.e., $||\mathbf{u}_i^P - \mathbf{u}_{i_k}^Q||$, and with the constraint that the r matchings do not highly overlap. The set of initial correspondences associated with v_i^P is

$$\mathcal{M}_i = \{m_{ii_k} = (v_i^P, v_{i_k}^Q, H_{ii_k})\}_{k=1}^r, \tag{6}$$

where H_{ii_k} is the relative transformation from v_i^P to $v_{i_k}^Q$. This process is repeated for each feature in I^P. Then the set of initial correspondences is constructed by

$$\mathcal{M} = \bigcup_{i=1}^{N^P} \mathcal{M}_i. \tag{7}$$

Set $\mathcal{M}$ is of size $|\mathcal{M}| = r \times N^P$. It contains many corrupted matchings, and may cover just a small subset of correct matchings in $\mathcal{C}$. We empirically set $r = 5$ in this work.

4.2 Hough transform for homography verification

The goal at this stage is to detect the correct correspondences in $\mathcal{M}$, which is either the initial correspondence set or the enriched set by the following stage. We investigate

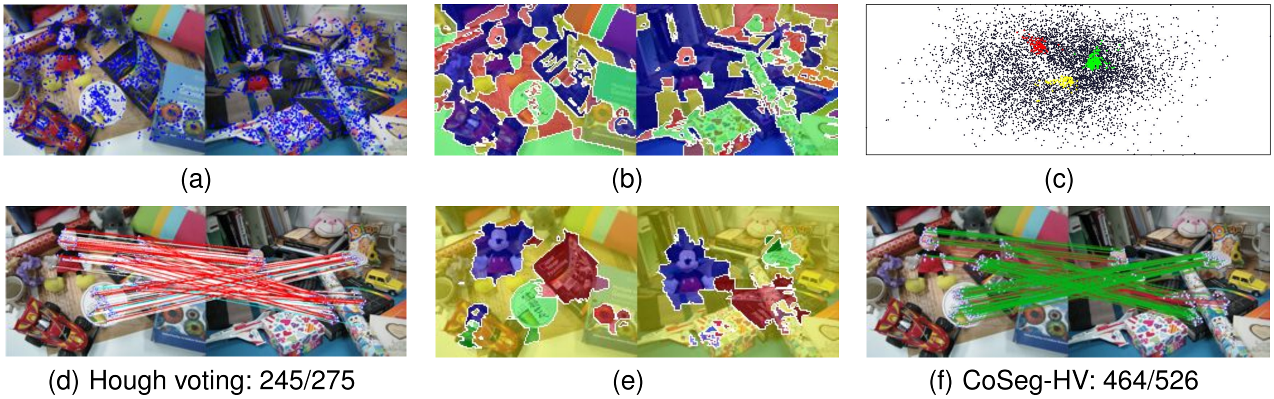


Fig. 1: Feature matching by our approach. (a) Input images I^P and I^Q together with all the detected feature points. (b) Initial co-segmentation results by [36]. (c) Correspondences in the Homography space. (d) Hough voting and its comparison with SIFT. 245 out of 275 correct correspondences in $\mathcal{M}$ are identified via Hough voting. White lines denote the correct correspondences detected by both approaches. Red and cyan lines are the correct correspondences by only Hough voting and the nearest SIFT searching, respectively. (e) The refined co-segmentation results by taking feature matchings into account. (f) Our approach CoSeg-HV. It recommends 251 $(= 526 - 275)$ correct candidates in $\mathcal{M}$ and leads to additional 219 $(= 464 - 245)$ correct correspondences (green lines) detected by the successive Hough voting.

the property that the transformations of correct correspondences are concerted while those of incorrect correspondences are different in their own ways. Hough voting for homography verification is employed since it can handle a high percentage of incorrect correspondences and detect correct correspondences via density estimation. Specifically, the relative transformation of each correspondence is treated as a point in *Hough space*, and it is considered as a hypothesis about the underlying homography of interest.

Despite its robustness, Hough transform is developed upon the assumption that the hypotheses are a sum of independent votes, and thereby neglects the *spatial* dependence among features. As pointed out in [37], choosing proper *voters* is critical in Hough transform, especially when voters are dependent. Inspired by the fact that nearby features on the same object are mutually dependent, we group relevant correspondences via co-segments detected by a co-segmentation algorithm, e.g., [36] or [27] in this work. It turns out that the performance of Hough voting is remarkably boosted. Furthermore, only relevant, small-size correspondences are involved in density estimation, instead of the whole $\mathcal{M}$. It significantly speeds up the process.

To formalize, let $\mathcal{B} = \{b_\ell\}$ be the set of the segments in image V^P obtained via co-segmentation. For each feature $v_i^P \in I^P$, we use $\pi(v_i^P) \in \mathcal{B}$ to denote the segment that covers the center of v_i^P. We then collect features relevant to v_i^P by checking if they reside in the same segment, i.e.,

$$G(v_i^P) = \{v_j^P | \pi(v_i^P) = \pi(v_j^P)\}. \tag{8}$$

We assume that the grouped features with high probability undergo similar transformations in matching. It follows that the correspondences relevant to v_i^P in Hough voting can be collected by

$$R(v_i^P) = \bigcup_{v_j^P \in G(v_i^P)} \mathcal{M}_j. \tag{9}$$

Given in Eq. (6), $\mathcal{M}_i$ consists of at most one correct correspondence. Hough voting as well as voters $R(v_i^P)$ are adopted to pick the most plausible correspondence associated with feature v_i^P. Specifically, it is accomplished by *normalized kernel density estimation* (KDE):

$$m_{ii'}^* = \underset{m_{ii'} \in \mathcal{M}_i}{argmax} \frac{1}{|R(v_i^P)|} \sum_{m \in R(v_i^P)} \exp\left(-\frac{d(m_{ii'}, m)}{\sigma}\right), \tag{10}$$

where σ is a positive constant whose value is set as the average distance among the existing correspondences. Note that the normalization term $1/|R(v_i^P)|$ does not affect the result in Eq. (10), but it is required in comparing densities across feature points.

The procedure of correspondence selection is repeated for each feature in image I^P. It results in N^P selected correspondences $\mathcal{M}^* = \{m_{ii'}^*\}_{i=1}^{N^P}$. We then sort them according to their associated densities in Eq. (10), and return the top correspondences by a proper threshold. In the experiments, our approach is evaluated by precision-recall curves, plotted with various thresholds.

An example of the verification results by Hough voting is shown in Fig. 1. The detected feature points are plotted in Fig. 1(a). The segments by co-segmentation algorithm [36] are displayed in Fig. 1(b). We use the re-projection error to measure the distance between two correspondences, and adopt *multi-dimensional scaling* (MDS) [38] to visualize points in the six-dimensional transformation space in a 2D plane. The homographies of the initial correspondences in the transformation space are shown in Fig. 1(c), in which incorrect correspondences are drawn in black while correct correspondences in the three common objects, including `cup noodles` (red), `book` (green), and `Mickey` (yellow), are drawn in the particular colors. It can be observed that correct correspondences on the same object often gather together. The detected correspondences by Hough voting and by the nearest neighbor search are compared in Fig. 1(d).

4.3 Inverted Hough transform for correspondence recommendation

Hough transform identifies correct correspondences $\mathcal{M}^* \subseteq \mathcal{M}$ and boosts the precision. The resulting correspondences can help image co-segmentation in finding more object-aware co-segments. This step for refining co-segmentation will be given in Section 5 for the sake of clarity. Assume that the refined co-segments are available. The goal of inverted Hough transform is to enrich $\mathcal{M}$ so that the recall can be increased. The grouped features by co-segments often have consensus transformations and can assist each other in finding plausible correspondences. We investigate this property and develop the inverted Hough transform, in which grouped features propagate their homographies to each other. It follows that concerted correspondences are recommended.

For each $v_i^P \in V^P$, we search the relevant features, $G(v_i^P)$ in Eq. (8). Each of these features delivers a hypothesis about the homography of v_i^P. These hypotheses are collected in

$$\tilde{M}_i = \{m^*_{jj'} | v_j^P \in G(v_i^P)\}, \tag{11}$$

where $m^*_{jj'}$ is the selected correspondence of v_j^P through Hough voting. The set $\tilde{M}_i$ may contain outliers caused by corrupted matchings. Hence, we pick the homography of the most plausible correspondence $\tilde{m}_{jj'} \in \tilde{M}_i$ for recommendation, where

$$\tilde{m}_{jj'} = \underset{m^*_{jj'} \in \tilde{M}_i}{argmax} \sum_{m \in \tilde{M}_i} \exp\left(-\frac{d(m^*_{jj'}, m)}{\sigma}\right). \tag{12}$$

Suppose that the relative transformation of $\tilde{m}_{jj'}$ is $H_{jj'}$. The projected region of v_j^P from I^P to I^Q through $H_{jj'}$ is denoted by S. The matching feature in I^Q is obtained by

$$v_k^Q = \underset{v_k^Q \in V^Q}{argmax} \frac{S \cap S_k^Q}{S \cup S_k^Q}. \tag{13}$$

It follows that correspondence $m_{ik} = (v_i^P, v_k^Q, H_{ik})$ is recommended, i.e., $\mathcal{M}_i \leftarrow \mathcal{M}_i \cup m_{ik}$. This process is done for each feature in I^P. The resulting correspondence set $\mathcal{M}$ in Eq. (7) is incrementally enriched.

Hough voting and its inverted variant are tightly coupled. While the former detects correct correspondences from the enriched candidates, the latter gives harmonic enrichment owing to better detection results. The alternate voting procedure is guaranteed to finish. It can be observed that the number of all the correspondences, i.e., $|\mathcal{C}|$, is finite and fixed. At each iteration, the number of correspondence candidates, i.e., $|\mathcal{M}|$, is monotonically increasing. Since $\mathcal{M}$ is a subset of $\mathcal{C}$, the iterative procedure must stop. In implementation, we terminate the procedure when $|\mathcal{M}|$ no longer increases or the maximum number of iterations is reached. Empirically, our approach rapidly converges after a few iterations, typically $2 \sim 4$, as shown in Section 6.5.4.

As our approach performs Hough voting by taking co-segmentation results into account, we term our approach as *Co-Segmentation guided Hough Voting* (or CoSeg-HV for short). Fig. 1(f) shows the matching results by CoSeg-HV. Compared with Hough voting, CoSeg-HV detects more accurate and dense correspondences, and hence improves the performance in terms of precision and recall. It is worth mentioning that the co-segmentation results can be further refined by considering the correspondences. We will describe it in the next section. Fig. 1(e) displays the refined co-segmentation results. We conclude this section by summarizing our approach in Algorithm 1.

Algorithm 1 The procedure of the proposed framework

```
Input: Feature sets $V^P$ and $V^Q$; Max iteration $T$
Output: Matched correspondences $\mathcal{M}^*$
Initialize correspondence sets $\{\mathcal{M}_i\}_{i=1}^{N^P}$ via (6);
Apply image co-segmentation to $I^P$ and $I^Q$;
while $t < T$ do
    $M^* \leftarrow \emptyset$
    for all $v_i^P \in V^P$ do
        Detect correspondence $m^*_{ii'} \in \mathcal{M}_i$ via (10);
        $\mathcal{M}^* \leftarrow \mathcal{M}^* \cup m^*_{ii'}$;
    end for
    Filter corrupt correspondences in $\mathcal{M}^*$ via (15);
    Cluster the remaining correspondences for co-segmentation initialization;
    Generate refined co-segments via (22) or (26);
    for all $v_i^P \in V^P$ do
        Identify recommended feature $v_k^Q$ via (13);
        Construct $m_{ik} = (v_i^P, v_k^Q, H_{ik})$;
        $\mathcal{M}_i \leftarrow \mathcal{M}_i \cup m_{ik}$;
    end for
end while
Sort elements in $\mathcal{M}^*$ with thresholding;
```

5 Enhanced Image Co-segmentation with Feature Matching

In this section, we show how the progressively accurate and dense correspondences by our approach improve co-segmentation. First, we use *one-class SVM* [39], [40] and *spectral clustering* [41] to filter corrupt correspondences and group the remaining ones, respectively. The results can serve as a good initialization of co-segments, which is essential to many existing co-segmentation algorithms. Second, based on feature matching, we introduce a *homography regularization* term and a *descriptor consistency* term into co-segmentation. While the former encourages intra-object geometric coherency, the latter encodes inter-image photometric consistency. We show the benefits from feature matching with two powerful co-segmentation algorithms, including the graph-partition model for multi-class co-segmentation [36] and the MRF-based two-class co-segmentation algorithm [27].

5.1 Corrupt correspondence filtering

Inspired by the observation that correspondences residing in the same object often have similar homographies, grouping correspondences based on their homographies helps in finding more object-aware co-segments. However, two issues arise if we directly cluster the correspondences found by our approach, i.e., those in $\mathcal{M}^*$. First, $\mathcal{M}^*$ consists of one correspondence for each feature point in image I^P, but only a fraction of feature points reside in common objects. Clustering correspondences in backgrounds does not make sense, because strong variation in their homographies presents. Second, spatially smooth co-segments are preferred in co-segmentation. However, the construction of $\mathcal{M}^*$ in Eq. (10) neglects the spatial positions of correspondences in the image. To address the two issues, we utilize one-class SVM with a designed distance to identify geometrically and spatially coherent correspondences from $\mathcal{M}^*$. The identified correspondences are then used to enhance co-segmentation in the forms of a good foreground initialization, the homography regularization term, and the descriptor consistency term.

One-class SVM is the state-of-the-art methodology for unsupervised classification. It works based on the assumption that positive data are similar to each other, while negative data are different in their own ways. We utilize one-class SVM for classifying matchings in $\mathcal{M}^*$, since it is analogous to feature matching where correct matchings are spatially and geometrically consistent with each other, while the incorrect matchings distribute irregularly. Specifically, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to model both the spatial and geometric relationships among correspondences. A node v_i is created for each correspondence $c_i \in \mathcal{M}^*$, while an edge $e_{ij} \in \mathcal{E}$ is added to connect v_i and v_j if their endpoints in I^P are within δ pixels. (δ is set as $\frac{\sqrt{img_h \times img_w}}{10}$ in this work, where img_h and img_w are the height and the width of I^P, respectively). We assign weight $w_{i,j}$ to edge $e_{i,j}$ as follows:

$$w_{ij} = \begin{cases} d(c_i, c_j), & \text{if } e_{ij} \in \mathcal{E}, \\ \infty, & \text{otherwise}, \end{cases} \tag{14}$$

where $d(c_i, c_j)$ is the geometric dissimilarity between correspondences c_i and c_j as given in Eq. (5). With the weights on the edges, the *geodesic distance* $d_{geo}(c_i, c_j)$ between each pair of correspondences c_i and c_j can be computed by Floyd-Warshall's algorithm. Compared with $d(c_i, c_j)$, $d_{geo}(c_i, c_j)$ further integrates spatial continuity into the estimation of geometric coherence. Thereby it more faithfully reflects the relationships between correspondences. With the kernel function $K(i,j) = k(c_i, c_j) = \exp(-\frac{d^2_{geo}(c_i,c_j)}{\sigma^2})$, one-class SVM predicts the correctness of each correspondence $c \in \mathcal{M}^*$ by sign($f(c)$), where

$$f(c) = \mathbf{w}^T \phi(c) - \nu = \sum_{j=1}^{|\mathcal{M}^*|} \alpha_j k(c, c_j) - \nu. \tag{15}$$

The value of σ is set as the average distance between all correspondences to their nearest neighbors. With training set $\mathcal{M}^*$, their coefficients, $\{\alpha_j\}$, in Eq. (15) are optimized by using package *LibSVM* [42]. We tune parameter ν by uniformly sampling from a large range, and set $\nu = 0.8$ in all the experiments. We denote the set of the correspondences selected by one-class SVM as $\tilde{\mathcal{M}}^*$.

5.2 Information transfer from feature matching to image co-segmentation

We describe our representations of information transferred from feature matching to image co-segmentation, including the initial seeds, the homography regularization term, and the descriptor consistency term.

For image co-segmentation, a set of images $\mathbf{I} = \{I^i\}_{i=1}^B$ containing instances of common objects is considered. The goal is to partition the pixels of all images into meaningful $K \geq 2$ classes, where K ($K-1$ common objects with one background class) is assumed to be known. For efficiency, the over-segmentation algorithm [43] is firstly applied to partition each I^i into S^i superpixels, i.e., $\{s^i_j\}_{j=1}^{S^i}$. Pixels of a superpixel share the same label in co-segmentation.

To have good initial co-segments, correspondences in $\tilde{\mathcal{M}}^*$ are clustered into $(K-1)$ classes via spectral clustering [41] with their pairwise geodesic distances as input. The clustered correspondences can then serve as the *initial seeds* in the generation of the initial co-segmentation results. For each superpixel that covers at least one correspondence in $\tilde{\mathcal{M}}^*$, its initial label is set as the majority cluster label of the correspondences falling into that superpixel. The remaining superpixels are initialized as background.

The *homography regularization term* is yielded by exploiting the geometric information carried by the homographies of the correspondences. Specifically, an affinity matrix $[\gamma^i_{g,h}]$ is used to represent how well two superpixels s^i_g and s^i_h in image I^i are grouped together according to the homography coherence:

$$\gamma^i_{g,h} = \sum_{m_{aa'} \in Q^i_g} \sum_{m_{bb'} \in Q^i_h} \exp(-\frac{d^2_{geo}(m_{aa'}, m_{bb'})}{\sigma^2_\gamma}), \tag{16}$$

where $d_{geo}(m_{aa'}, m_{bb'})$ is the geodesic distance between correspondences $m_{aa'}$ and $m_{bb'}$. Q^i_g and Q^i_h contain correspondences whose endpoints in image I^i fall in superpixels s^i_g and s^i_h, respectively. The value of σ_γ is set as the average geodesic distance between the correspondences in $\tilde{\mathcal{M}}^*$.

While the homography regularization term considers intra-image geometric coherence between superpixels, the descriptor consistency term encodes inter-image photometric consistency between superpixels by considering the similarity between matched feature points. Specifically, another affinity matrix $[\psi^{i,j}_{g,h}]$ is complied to measure the likelihood that superpixels s^i_g in image I^i and s^j_h in image I^j belong to the same class:

$$\psi^{i,j}_{g,h} = \sum_{m_{aa'} \in Q^i_g \cap Q^j_h} \exp(-\frac{\|\mathbf{u}^i_a - \mathbf{u}^j_{a'}\|^2}{\sigma^2_\psi}), \tag{17}$$

where σ_ψ is set as the average descriptor distance between matched feature points in matchings belonging to $\tilde{\mathcal{M}}^*$.

Most co-segmentation algorithms perform by taking the pair-wise similarities (or dissimilarities) between superpixels (pixels) into account. For example, the graph-based co-segmentation algorithm [36] formulates this information in form of *graph Laplacian*, and the MRF-based algorithm [27] model it in form of energy functions. As two kinds of similarity measure between superpixels, the homography regularization term and the descriptor consistency term can be easily incorporated into co-segmentation. Two examples are given in the next two subsections.

5.3 Graph-partition co-segmentation model

The model by Joulin et al. [36] partitions image set $\mathbf{I} = \{I^i\}_{i=1}^B$ of N pixels (or superpixels here) into K classes, and uses $y \in \{0,1\}^{N \times K}$ to represent the results, i.e.,

$$y_{nk} = \begin{cases} 1, & \text{if the } n\text{th superpixel is of class } k, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The co-segmentation model [36] jointly infers the segment label y and nonlinear separating surface (A, b) by minimizing the following energy function:

$$\min_{y, y 1_K = 1_N, A, b} E_U(y, A, b) + E_B(y) - H(y). \tag{19}$$

It includes a discriminative term E_U that maximizes class separability, a spatial consistency term E_B that encodes both the visual and spatial similarity, and a regularization term H that balances the cluster sizes. In this work, we follow the original definitions of E_U and H given in [36], and modify E_B to take the results of feature matching into account. Specifically, E_B in [36] is defined as

$$E_B(y) = \frac{\mu}{N} \sum_{I^i \in \mathbf{I}} \sum_{g,h \in \mathcal{S}^i} \sum_{k=1}^{K} y_{gk} y_{hk} L_{gh}, \tag{20}$$

where μ is a free parameter, and $\mathcal{S}^i$ is the index set of superpixels in I^i. $L \in \mathbb{R}^{N \times N}$ is a normalized graph Laplacian, and it is constructed with an affinity matrix, whose elements record the pair-wise similarities between superpixels in color and spatial position. Function $E_B(y)$ encourages clustering superpixels into homogeneous regions. It forces segmentation to preserve strong edges, which might not necessarily correspond to object boundaries especially when large intra-object color variability or cluttered backgrounds present. We leverage the information from feature matching to address this issue. With the affinity matrices $[\gamma_{g,h}^i]$ in Eq. (16) and $[\psi_{g,h}^{i,j}]$ in Eq. (17), two extra normalized graph Laplacian $L^\gamma \in \mathbb{R}^{N \times N}$ and $L^\psi \in \mathbb{R}^{N \times N}$ are constructed, respectively. The modified function $E_{B'}(y)$ is then defined as

$$E_{B'}(y) = \frac{\mu}{N} \sum_{I^i, I^j \in \mathbf{I}} \sum_{g \in \mathcal{S}^i, h \in \mathcal{S}^j} \sum_{k=1}^{K} y_{gk} y_{hk} (L_{gh} + \lambda^\gamma L_{gh}^\gamma + \lambda^\psi L_{gh}^\psi). \tag{21}$$

Via Eq. (21), feature matching is integrated into co-segmentation. Multi-type information, including color, spatial position, geometric, and descriptor similarities, are fused in the domain of graph Laplacian. With $E_{B'}(y)$ in Eq. (21), the new formulation for co-segmentation becomes

$$\min_{y, A, b} E_U(y, A, b) + E_{B'}(y) - H(y). \tag{22}$$

In the experiments, we tune and fix parameters λ^γ and λ^ψ in Eq. (21) for each adopted dataset. The same optimization procedure in [36] can be used to solve Eq. (22).

5.4 MRF-based co-segmentation model

The MRF-based co-segmentation algorithm [27] performs figure-ground separation over $\mathbf{I} = \{I^i\}_{i=1}^B$. The co-segmentation results on I^i is represented by binary vector $x^i \in \{0,1\}^{S^i}$. The algorithm by Chang et al. [27] optimizes $\{x^i\}_{i=1}^B$ by minimizing the following energy function:

$$F(x^i) = \sum_{i=1}^{B} L_i(x^i) + \lambda \sum_{i,j=1}^{B} G(x^i, x^j, I^i, I^j), \tag{23}$$

where $L_i(x^i)$ is the *intra-image* energy for the labeling x^i on I^i, and $G(x^i, x^j, I^i, I^j)$ is the *inter-image* energy measuring the inconsistency between I^i and I^j under the labeling x^i and x^j. Refer to [27] for the details of the two energy functions.

Except for the initialization, we further boost the performance of [27] by leveraging the results of feature matching. Note that the geometric regularization term in Eq. (16) and the descriptor consistency term in Eq. (17) respectively describe the intra-image and the inter-image interaction between superpixels. Thus, we integrate them into the intra-image energy and the inter-image energy, respectively. The modified intra-image energy $L_i'(x^i)$ becomes

$$L_i'(x^i) = L_i(x^i) + \lambda^\gamma \sum_{(g,h) \in \mathcal{E}^i} \gamma_{g,h}^i \delta[x_g^i \neq x_h^i], \tag{24}$$

where λ^γ is a nonnegative constant, and $\mathcal{E}^i$ contains the pairs of adjacent superpixels in I^i. The extra homography regularization term $\gamma_{g,h}^i$ in Eq. (24) encourages label coherence especially when superpixels s_g^i and s_h^i share similar homographies in feature matching. On the other hand, the modified inter-image energy $G'(x^i, x^j, I^i, I^j)$ is given by

$$\begin{aligned} G'(x^i, x^j, I^i, I^j) &= G(x^i, x^j, I^i, I^j) \\ &+ \lambda^\psi \sum_{g \in \mathcal{S}^i, h \in \mathcal{S}^j} \psi_{g,h}^{i,j} \delta[x_g^i \neq x_h^j], \end{aligned} \tag{25}$$

where λ^ψ is a nonnegative constant. $\mathcal{S}^i$ and $\mathcal{S}^j$ are the index sets of superpixels in images I^i and I^j, respectively. The second term in Eq. (25) penalizes label inconsistence in particular when there exist strong correspondences between superpixels. With $L_i'(x^i)$ in Eq. (24) and $G'(x^i, x^j, I^i, I^j)$ in Eq. (25), the new energy function becomes

$$F'(x^i) = \sum_{i=1}^{B} L_i'(x^i) + \lambda \sum_{i,j=1}^{B} G'(x^i, x^j, I^i, I^j). \tag{26}$$

Similarly, the parameters λ^γ and λ^ψ in Eq. (24) and Eq. (25) are tuned and fixed for each adopted dataset in the experiments. Like [27], graph-cut [44] is used to solve Eq. (26), which is still submodular.

6 Experimental Results

The performance evaluation and analysis of the proposed approach are reported in this section. Totally four sets of experiments are conducted. First, we visualize the transformation space to verify our assumption that correct correspondences gather together in that space, while incorrect

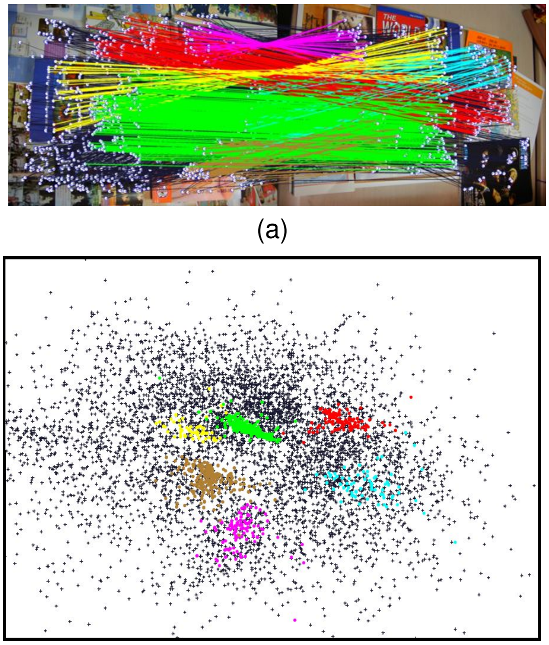


Fig. 2: (a) The matching results of our approach in an image pair. The incorrect correspondences are drawn in black. Each correct correspondence is shown in a particular color according to the object that it belongs to. (b) The corresponding 2D homography space generated by using multi-dimensional scaling.

ones distribute more dispersedly. Second, our approach is compared with some of the state-of-the-art approaches to matching multiple common objects in cluttered backgrounds. Third, we show that our approach can collaborate with different feature detectors and descriptors, and establish accurate correspondences across images where dramatic variations or deformations present. Finally, a few important issues regarding our approach are discussed, such as image co-segmentation initialization, the number of cosegments, the running time, and the convergence property.

6.1 Homography space visualization

We visualize the transformation (or homography) space, and check whether the assumption that correct correspondences tend to gather together in the space holds or not. As mentioned in Section 3, each correspondence is considered as a point in the homography space that is six-dimensional in this work, and the re-projection error measures the distance between two correspondences. We adopt *multi-dimensional scaling* (MDS) [38] to visualize the six-dimensional space in a 2D plane. MDS summarizes high-dimensional data in a low-dimensional space by approximating their pair-wise distances, i.e., re-projection error here. In Fig. 2(a), the matching results by approaches are shown, where the correspondence with the highest density in each detected feature of the left image is plotted. While wrong correspondences are drawn in black, each correct one is displayed in a specific color according to the common object that it resides in. The corresponding 2D plane generated by MDS is shown in Fig. 2(b). It can be observed that correct correspondences on the same object typically have consistent transformations, while the incorrect correspondences are irregularly distributed. It also indicates that geometric consistency among correspondences is highly relevant to their correctness.

6.2 Evaluation metrics

In this work, we analyze the performance of a matching algorithm by jointly considering *precision* and *recall*. While the former is the fraction of detected correspondences that are correct, the latter is the fraction of correct correspondences that are detected. In more detail, the two terms are respectively defined as

$$\text{PRECISION} = \frac{n\text{TP}}{n\text{TP} + n\text{FP}}, \tag{27}$$

and

$$\text{RECALL} = \frac{n\text{TP}}{n\text{TP} + n\text{FN}}, \tag{28}$$

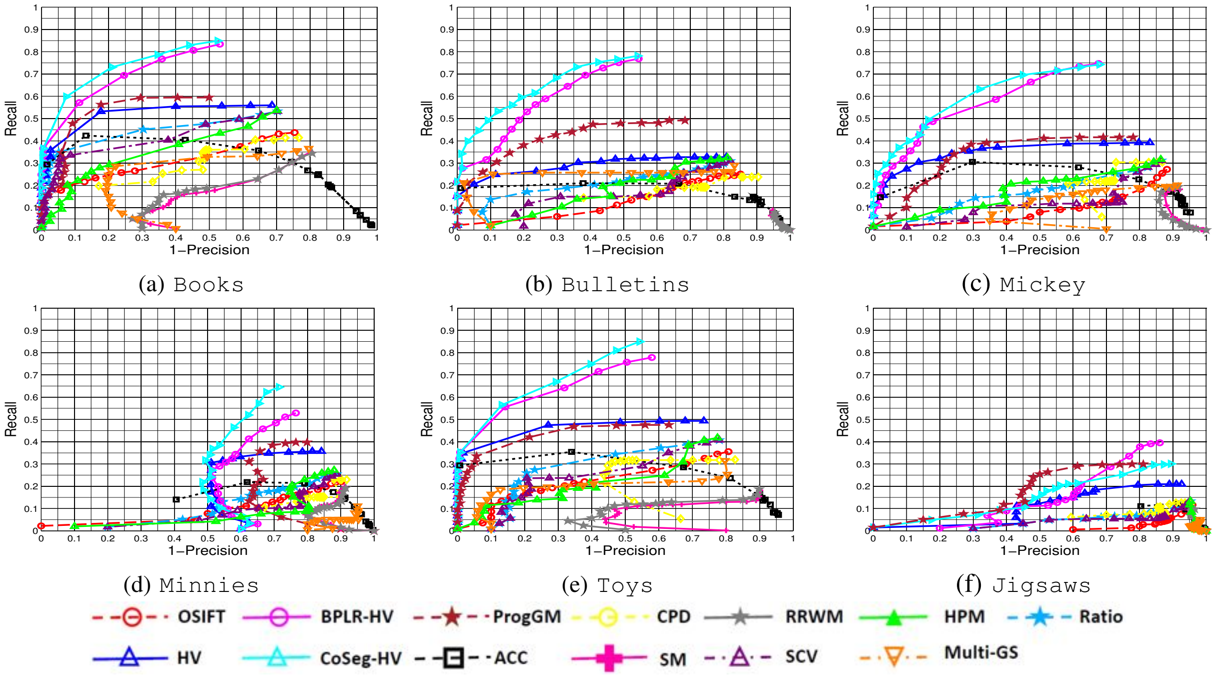
where $n\text{TP}$ and $n\text{FP}$ are the numbers of detected correspondences that are correct and incorrect, respectively. $n\text{FN}$ is the number of correct correspondences that are not detected. In other words, $(n\text{TP} + n\text{FP})$ is the number of correspondences returned by a specific algorithm. $(n\text{TP} + n\text{FN})$ is the number of the all correct correspondences in set $V^P \times V^Q$.

For a feature point in the first (left) image to be matched, its matching is considered correct if the distance between the other endpoint and the ground truth is within ϵ pixels. We set ϵ as 15 in this work. For each matching approach, including ours and the adopted baselines, all the detected correspondences are ranked by its own criterion, such as the element values of the eigenvector in spectral matching [13] or the estimated density, Eq. (10), in our approach. With a set of thresholds, the performance of each approach can then be presented by a precision-recall curve. We determine the thresholds by uniformly sampling the returned correspondences in the ranked list.

6.3 Matching with multiple common objects

The performances of our approach as well as the adopted baselines are evaluated on the *co-recognition* dataset [22]. This dataset contains six image pairs, and each pair has multiple common objects. The large appearance variations of common objects, partial occlusions, and cluttered backgrounds make matching on this dataset quite challenging. However, it provides a good test bed to manifest the importance of geometric verification and correspondence recommendation, since the initial correspondences are not reliable enough. Our approach, *Co-Segmentation guided Hough Voting* (CoSeg-HV), is compared with some of the state-of-the-art systems, each of which is briefly described and denoted in bold as follows:

- Descriptor-based approach: We adopt *opponent SIFT* (**OSIFT**) [45] for comparison. It is an important baseline, since we use OSIFT to compile the initial set of correspondences. We also adopt Lowe's *ratio test* [12] (**Ratio**), i.e., the ratio of distance from the closest neighbor to the distance of the second closest, for sorting the establishing correspondences.



(a) Books (b) Bulletins (c) Mickey

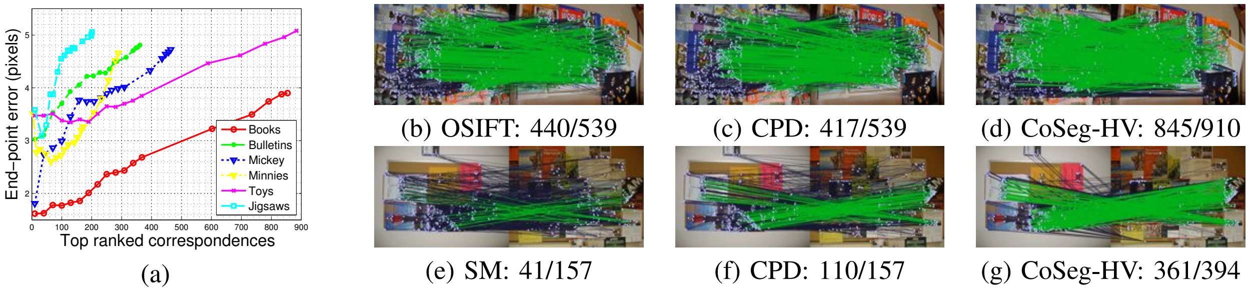
(d) Minnies (e) Toys (f) Jigsaws

Fig. 3: The performances of various approaches on the six image pairs of the co-recognition dataset.

- Clustering-based approach: Our approach is compared with the *common visual pattern discovery* (**CPD**) [14] and the *agglomerative correspondence clustering* (**ACC**) [15].
- Graph-based approach: Among the graph-matching algorithms, *spectral matching* (**SM**) [13] and *reweighted random walks* (**RRWM**) [46] are selected for their good performances.
- Voting-based approach: Among the voting-based algorithms, the variant of *Hough pyramid matching* (**HPM**) [47] and *multi-structure homography fitting* (**Multi-GS**) [17] are selected for their good performances. Note that the method in [47] is applied to correspondences established by matching the visual words. In our cases, there is no additional dataset for constructing visual words in advance. Thus, we implement its variant that is applied to the correspondence set $\mathcal{M}$ in Eq. (7).
- *Sequential Correspondence Selection* (**SCV**) [24] is also adopted for its support region selection.
- Feature matching and enrichment: We adopt the *progressive graph matching framework* (**ProgGM**) [25], which is one of the best approaches to correspondence enrichment.
- The variants of our approach: Two variants of our approach, including *Hough voting* (**HV**) and *BPLR guided Hough Voting* (**BPLR-HV**) [21]. HV carries out only the part of Hough transform in our approach. Comparing with HV can reveal the effect of correspondence enrichment especially in the aspect of RECALL. BPLR-HV adopts BPLR [48] for correspondence recommendation. Comparing with BPLR-HV, the advantage of integrating image matching with image co-segmentation can be explored.

In this set of experiments, our approach is collaborated with the co-segmentation model in [36], since it is designed for the cases where multiple common objects present. For the sake of fair comparison, all the approaches work on the same feature points detected by the Hessian affine detector [35] and depicted by the OSIFT descriptor [45]. The initial correspondence set of all the approaches is selected by the nearest search with the OSIFT descriptor. Note that we use the publicly available codes provided by the authors for all the compared methods except HPM, which is implemented by us and can work with multiple correspondence candidates in this experiment. We set the value of γ in Eq. (6) as 5 for all the methods, though the optimal value may vary from method to method.

The quantitative results in form of precision-recall curves are reported in Fig. 3. The baseline OSIFT achieves good scale invariance and robustness to a certain degree of viewpoint changes. However, it does not work well in this dataset, since the unary local features are not sufficient to handle cluttered backgrounds, complex transformations of common objects. Baseline Ratio gives better results than OSIFT. The performances of the approaches based on graph matching, i.e., SM [13] and RRWM [46], are not stable due to their sensitivity to outliers. In this dataset, wrong correspondences are much more than correct ones in initialization. Although Multi-GS is capable of fitting multiple homography structures, it still suffers from the large fractions of outliers here. HPM is marginally better, but still not satisfactory. It may result from the quantization errors in the pyramid of the transformation space. Although SCV performs additional process for support region selection, the performance gain over OSIFT is not significant in this dataset, because the Hessian-affine detector has already shown high repeatability in the highly textured images.



(a)

(b) OSIFT: 440/539

(c) CPD: 417/539

(d) CoSeg-HV: 845/910

(e) SM: 41/157

(f) CPD: 110/157

(g) CoSeg-HV: 361/394

Fig. 4: (a) Average endpoint errors of the top ranked correct correspondences by CoSeg-HV. (b) $\sim$ (g) The feature correspondences detected by different approaches on image pairs `Books` and `Bulletins`. In each figure, the adopted approach as well as its performance (correct detections/correct candidates in $\mathcal{M}$) are also shown. The correct correspondences are plotted in green, and the incorrect ones are in black. Note that the performance shown here is not the recall defined in Eq. (28). The denominator here is the number of correct correspondences in $\mathcal{M}$, while the denominator in recall is the number of correct correspondences in set $V^P \times V^Q$.

Instead, CPD [14], ACC [15], and HV show the advantage of investigating geometric consistency.

By progressively enriching true candidate matchings in $\mathcal{M}$, BPLR-HV, ProgGM and CoSeg-HV further improve the performance in terms of both precision and recall. Among them, the proposed CoSeg-HV and BPLR-HV consistently outperforms ProgGM in most cases. The objective functions of the correspondence verification and recommendation steps in our approach are both derived upon densities and hence coherent. The two steps complement each other to jointly lead to better results. In addition, CoSeg-HV is superior to BPLR-HV. This is because BPLR-HV investigates only the local segmentation or arrangement in a single image. In contrast, CoSeg-HV utilizes rich information across two images through co-segmentation. With the aid of the segment-level information discovered by co-segmentation, the transformations corresponding to real object hypotheses are more likely to be targeted. This results in correspondence enrichment of high quality.

We also report the average endpoint errors of the correct correspondences detected by CoSeg-HV in Fig. 4(a). Specifically, the correct correspondences are ranked by their densities in Eq. (10). The average endpoint errors in pixel are measured with different numbers of the top ranked correct correspondences. It can been observed that the average endpoint errors of the correct correspondences in most cases are less than 5 pixels. To gain insight into the quantitative results, we plot the correspondences by various approaches on a few image pairs in Fig. 4(b) $\sim$ 4(g). In each subfigure, the adopted approach as well as the performance (correct correspondences / correct candidates in $\mathcal{M}$) are shown. It can be seen that CoSeg-HV dramatically increases the number of correct candidates in $\mathcal{M}$, i.e., from 539 to 910 on image pair `Books` and from 157 to 394 on image pair `Bulletins`. Thus, CoSeg-HV detects more accurate and dense matchings as shown in Fig. 4.

6.4 Collaborating with other feature descriptors

Our approach can be considered as a geometric filter. It drops correspondences that are not consistent with others, and enhances the matching by propagating concerted transformations among dependent features. It can be applied to various types of feature descriptors, and improve their performances in the associated applications.

TABLE 1: The performances in mAP of our approach and the baselines on three datasets.

dataset	SYM-D [49]	HV	CoSeg-HV
SYM-BENCH	18.82%	41.71%	63.12%

dataset	LIOP [50]	HV	CoSeg-HV
Illumination	63.57%	80.25%	91.16%

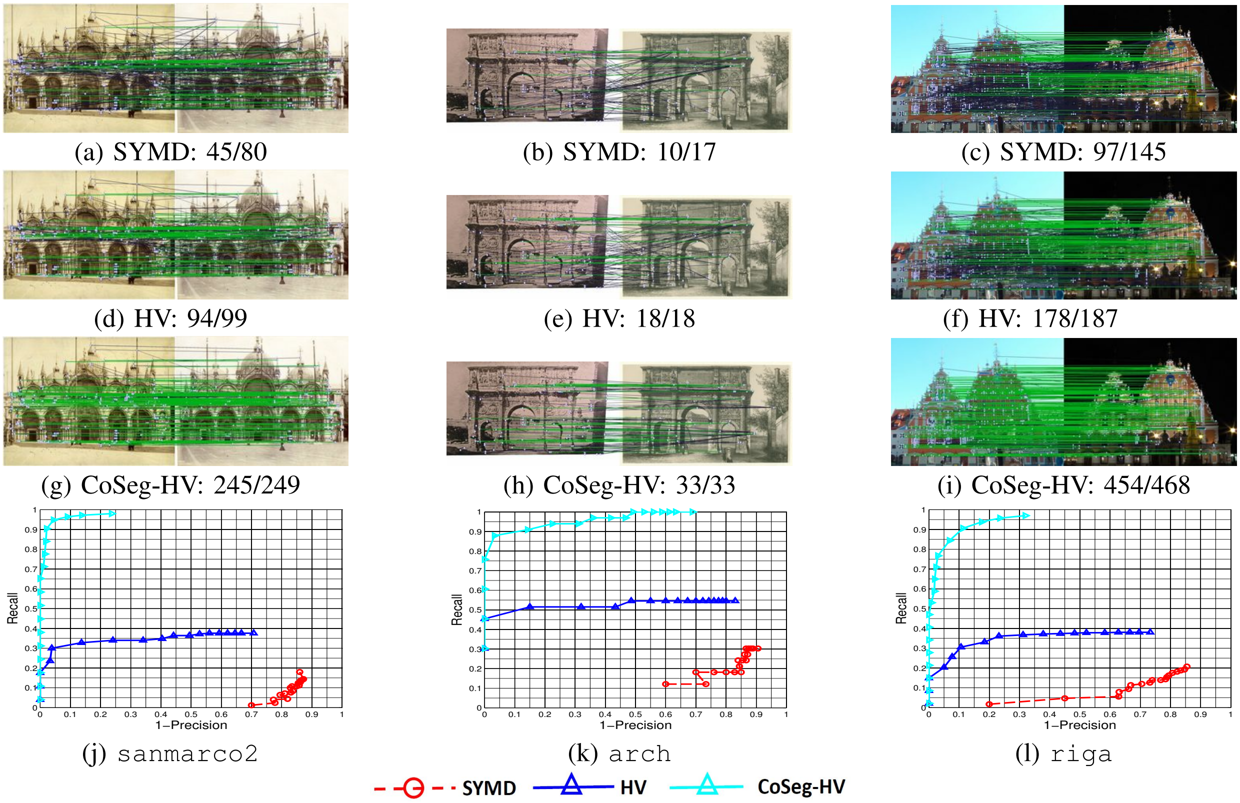
dataset	DAISY [51]	HV	CoSeg-HV
Daisy	52.28%	68.93%	71.50%

We conduct the experiments on the *SYM-BENCH* dataset [49], the *Illumination* dataset [50], and the *Daisy* dataset [51], where the *SYMD* descriptor [49], the LIOP descriptor [50], and the DAISY descriptor [51] are adopted, respectively. In this set of experiments, matching performances are evaluated by using *mean Average Precision* (mAP), which is the mean of the average precisions, while the average precision on each image pair is obtained by averaging the precisions computed with different numbers of the returned correspondences. In the first two experiments, our approach collaborates with the two-class co-segmentation algorithm by Chang et al. [27], because there is one single common object in each image pair of the two datasets. In the last experiment, our approach works with the multi-class co-segmentation algorithm by Joulin et al. [36] due to the presence of multiple common objects.

6.4.1 On working with the SYMD descriptor

We check whether our approach improves the SYMD descriptor on the challenging SYM-BENCH dataset [49], which contains 46 pairs of images exhibiting dramatic variations in lighting conditions, ages, and rendering styles. Some image pairs are pre-registered with a homography, while the others exhibit both geometric and photometric variations. A manually annotated homography for each pair is included in the ground truth. The *SYM-G* detector and the SYMD descriptor [49] are adopted to construct the set of initial correspondences. HV and CoSeg-HV are then applied to correspondence verification.

The overall performance in mAP is reported in TABLE 1. As highly similar patterns repeatedly appear in most images of this dataset, it causes *ambiguity* when establishing corre-



(a) SYMD: 45/80 (b) SYMD: 10/17 (c) SYMD: 97/145

(d) HV: 94/99 (e) HV: 18/18 (f) HV: 178/187

(g) CoSeg-HV: 245/249 (h) CoSeg-HV: 33/33 (i) CoSeg-HV: 454/468

(j) `sanmarco2` (k) `arch` (l) `riga`

Fig. 5: The matching results by three different approaches on three image pairs of the SYM-BENCH dataset [49], including image pair `sanmarco2` in the first column, image pair `arch` in the second column, and image pair `riga` in the third column. (a) $\sim$ (i) The used approaches as well as their performances (correct detections/correct candidates in $\mathcal{M}$) are attached below the figures. (j) $\sim$ (l) The corresponding precision-recall curves on the three image pairs.

spondences by considering only local descriptors. CoSeg-HV and HV overcome this issue by further enforcing geometric consistency in correspondence selection. They achieve mAPs of 63.12% and 41.71%, respectively, and considerably outperform the SYMD descriptor. Suffering from the same ambiguity problem, fewer correct correspondences are included in the initial set. CoSeg-HV tackles this issue by progressive correspondence enrichment, and is hence superior to HV. For the SYMD descriptor, we use the implementation provided by the authors [49][1] in this experiment. Note that the precision-recall curves are different from those reported in [49]. It is due to the different evaluation criteria and matching constraints.

The matching results and the precision-recall curves of three image pairs of the SYM-BENCH dataset are shown in Fig. 5. Compared with the SYMD descriptor, HV filters out many correspondences whose transformations are inconsistent. Compared with HV, the proposed CoSeg-HV generates more accurate and dense matchings. The advantages of using CoSeg-HV can be observed visually and quantitatively in Fig. 5.

6.4.2 On working with the LIOP descriptor

In the experiment, our approach collaborates with the LIOP descriptor [50], and establishes correspondences across images with drastic illumination changes. We perform the quantitative analysis on the complex illumination dataset used in [50]. It contains two image pairs `Desktop` and `Corridor`. Besides, the image pair `Leuven` with exposure change from Oxford dataset[2] is also adopted. For each image pair, the initial matching candidates are found by LIOP, which is robust to dramatic illumination changes via encoding local ordinary information of each pixel.

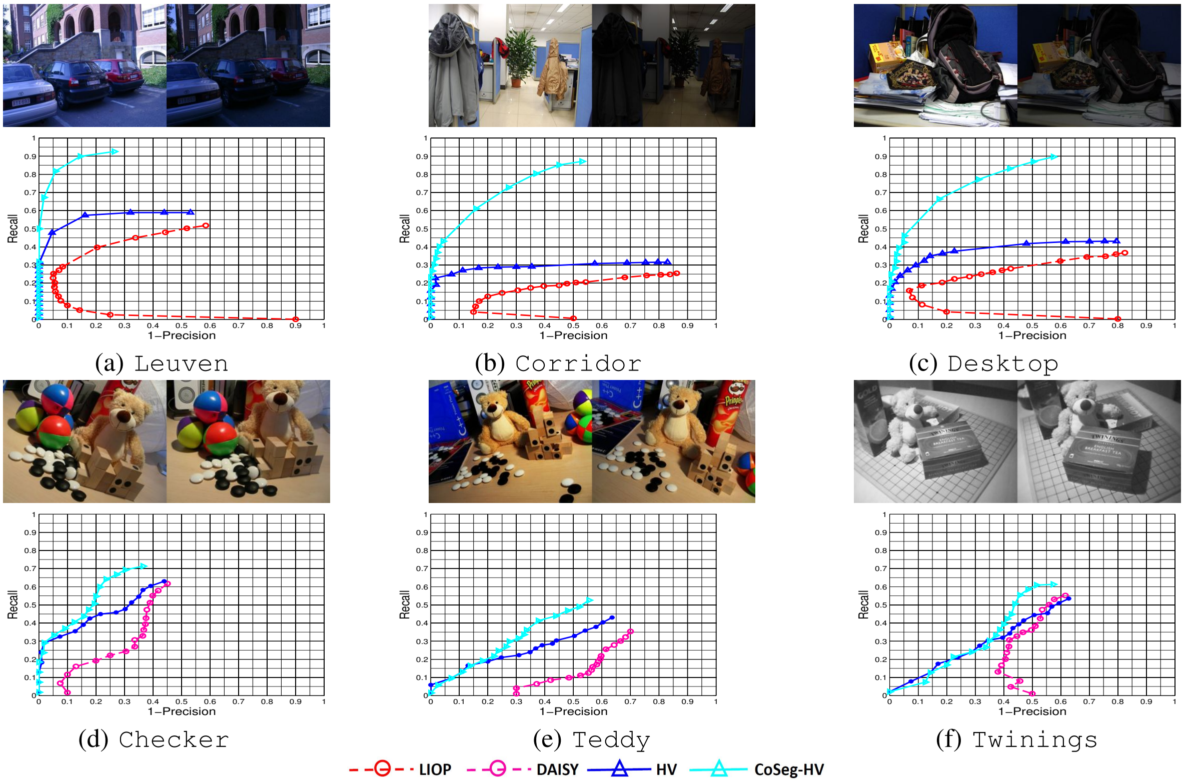
HV and CoSeg-HV are applied to the correspondences initially discovered by LIOP. The performances in form of precision-recall curves are shown in Fig. 6(a) $\sim$ 6(c). The obtained mean average precision (mAP) is reported in TABLE 1. It can be seen that despite the robustness to appearance change due to illumination variance, the performance of LIOP can still be enhanced by ensuring homography consistency. Besides, the correspondence candidates are enriched by leveraging the additional information grabbed through co-segmentation. It follows that CoSeg-HV leads to considerable performance gains, i.e., $10.91\%(= 91.16\% - 80.25\%)$ over HV and $27.59\%(= 91.16\% - 63.57\%)$ over LIOP. This is attributed to the fact that the co-segmentation results help to find the most plausible transformations across two images.

6.4.3 On working with the DAISY descriptor

In this experiment, our approach works with the DAISY descriptor [51], and establishes correspondences on six

[1] http://www.cs.cornell.edu/projects/symfeat/

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



(a) `Leuven` (b) `Corridor` (c) `Desktop`

(d) `Checker` (e) `Teddy` (f) `Twinings`

Fig. 6: Plug-in comparison with the LIOP and DAISY descriptors on six image pairs, including (a) `Leuven`, (b) `Corridor`, (c) `Desktop`, (d) `Checker`, (e) `Teddy`, and (f) `Twinings`.

TABLE 2: The effect of using OCSVM on our approach.

image pair	without OCSVM	with OCSVM
Books	0.4786	0.5035
Bulletins	0.5214	0.5451
Mickey	0.4679	0.5348
Minnies	0.2611	0.3531
Toys	0.4503	0.4824
Jigsaws	0.2530	0.3109
average	0.4054	0.4544

wide-baseline image pairs used in [51]. Image matching on these image pairs is challenging owing to the changes in contrast, scale, image quality, viewpoint and brightness. For each image pair, the initial correspondence candidates are found by the DAISY descriptor, which is robust to photometric and geometric variations. HV and CoSeg-HV are then applied to these initial candidates. The matching results and the precision-recall curves on three image pairs are shown in Fig. 6(d) ∼ 6(f), respectively. The overall performance in mAP is reported in TABLE 1. Compared with nearest neighbor search with the DAISY descriptor, HV utilizes Hough transform to enhance the precision especially when the recall is low. CoSeg-HV further boosts the recall by dynamic correspondence enrichment.

6.5 Comprehensive studies

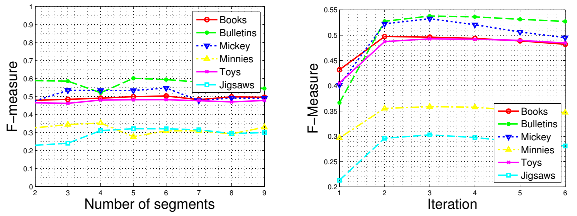
Here we discuss a few issues pertaining to our approach, including the initialization of image co-segmentation, the number of co-segments, the running time of each step in our approach, and the convergence property.

6.5.1 Co-segmentation initialization

One-class SVM (OCSVM) is utilized to select geometrically and spatially consistent correspondences. The selected correspondences act as the input to spectral clustering to initialize co-segments. This extra step, using OCSVM for correspondence selection, alleviates the problem that spectral clustering is sensitive to outliers. To single out its effect, we respectively evaluate the matching performances of our approach with and without this step. For a concise performance measure of feature matching, we compute *F-score*, the harmonic mean of precision and recall, with different thresholds, and take the average. TABLE 2 depicts the matching performances in F-score on the six image pairs of the co-recognition dataset. The results imply that better initialization of image co-segmentation with the aid of OCSVM is beneficial for image matching.

6.5.2 Number of co-segments

The number of segments K in the multi-class co-segmentation algorithm [36] is crucial. The true value of K is assumed known in our approach. Here we evaluate our approach with different numbers of co-segments. The numbers of common objects on the co-recognition dataset range from 3 to 6. Fig. 7(b) shows the performances, F-score, of our approach by setting K from 2 to 10 respectively. Compared with co-segmentation, the matching performance of our approach is less sensitive to the value of K. The main reason may be that our approach adopts the voted homography in Eq. (12) for enrichment, and is tolerant to moderate over-segmentation and under-segmentation.



(a) (b)

Fig. 7: (a) The performance of our approach with different numbers of co-segments. (b) The performance of our approach along the iterative procedure.

TABLE 3: Step-wise running time in second. [mean±std]

step	running time
feature detection and extraction	17.35 ± 5.60
Hough transform	6.06 ± 3.20
co-segmentation initialization	3.93 ± 1.49
co-segmentation	62.19 ± 4.08
Inverted Hough transform	3.37 ± 1.22

6.5.3 Running time

The step-wise running time of applying our approach to the co-recognition dataset is summarized in TABLE 3. Our approach is implemented in MATLAB, and executed on modern PC with an Intel Core $i7$ 3.4 GHz processor. The average number of interest points in an image is 1,466. It takes about 17.4 seconds to detect interest points, extract features, and find initial correspondences for an image pair. In the iterative procedure, two key components, Hough transform and inverted Hough transform, of our approach take about 6.1 and 3.4 seconds respectively. While the primary computational cost in the former component is on kernel density estimation, that in the latter component is on inferring the most plausible candidates for correspondence enrichment. Co-segmentation initialization by computing geodesic distance and running OCSVM spends around 3.9 seconds. The step of image co-segmentation is usually the computational bottleneck, and the running time is dependent on the adopted algorithm. It takes about one minute in our case where multi-class co-segmentation algorithm [36] is applied to the images of the co-recognition dataset.

6.5.4 Convergence analysis

We show the performance of feature matching along the iterative procedure in Fig. 7(b), in which each curve represents one of the six image pairs of the co-recognition dataset. It can be observed that the procedure rapidly converges after only a few iterations in all the six image pairs. The F-score averagely increases 10% in early iterations. The results demonstrate that our method provides fast convergence and high performance in feature matching.

7 Conclusion and Future Work

We have presented a simple but effective approach that carries out alternate Hough voting and its inverted variant to establish correspondences in complex matching tasks, and boosts the performances in both precision and recall. Our approach integrates image co-segmentation into the process of feature matching, and cast it as a density estimation problem in the homography space. Through iterative optimization, more correct correspondences are detected from the enriched candidates, while plausible enrichments are gradually revealed by taking the object segments into account. In the experiments, the proposed approach is comprehensively evaluated on four datasets coupled with different descriptors and co-segmentation algorithms. The promising results consolidate the effectiveness of our approach. For future work, we will apply the proposed approach to vision applications where accurate feature correspondences are appreciated.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. This work was supported in part by Ministry of Science and Technology, National Taiwan University and Intel Corporation under Grants 103-2221-E-001-026-MY2, 103-2911-I-002-001, and NTU-ICRP-104R7501.

References

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

[2] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2009, pp. 25–32.

[3] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. on Graphics*, vol. 25, no. 3, pp. 835–846, 2006.

[4] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Non-rigid dense correspondence with application for image enhancement," *ACM Trans. on Graphics*, vol. 30, no. 4, pp. 70:1–70:10, 2011.

[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. on Graphics*, vol. 28, no. 3, 2009.

[6] A. Albarelli, E. Rodolà, and A. Torsello, "Imposing semi-local geometric constraints for accurate correspondences selection in structure from motion: A game-theoretic perspective," *Int. J. Computer Vision*, vol. 97, no. 1, pp. 36–53, 2012.

[7] M. Leordeanu, M. Hebert, and R. Sukthankar, "An integer projected fixed point method for graph matching and map inference," in *Advances in Neural Information Processing Systems*, 2009, pp. 1114–1122.

[8] F. Zhou and F. De la Torre, "Factorized graph matching," in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 127–134.

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

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

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

[12] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int'l Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[13] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. Int'l Conf. Comput. Vis.*, 2005, pp. 1482–1489.

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

[15] M. Cho, J. Lee, and K. M. Lee, "Feature correspondence and deformable object matching via agglomerative correspondence clustering," in *Proc. Int'l Conf. Comput. Vis.*, 2009, pp. 144–157.

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

[17] T.-J. Chin, J. Yu, and D. Suter, “Accelerated hypothesis generation for multistructure data via preference analysis,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 34, no. 2, pp. 625–638, 2012.

[18] Y. Yuan, Y. Pang, K. Wang, and M. Shang, “Efficient image matching using weighted voting,” *Pattern Recognition Letters*, vol. 4, no. 33, pp. 471–475, 2012.

[19] G. Tolias and Y. Avrithis, “Speeded-up, relaxed spatial matching,” in *Proc. Int’l Conf. Comput. Vis.*, 2011, pp. 1653–1660.

[20] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image features,” in *Proc. Int’l Conf. Comput. Vis.*, 2005, pp. 1458–1465.

[21] H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, “Robust feature matching with alternate hough and inverted hough transforms,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2013, pp. 2762–2769.

[22] M. Cho, Y. M. Shin, and K. M. Lee, “Co-recognition of image pairs by data-driven monte carlo image exploration,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 2129–2136.

[23] V. Ferrari, T. Tuytelaars, and L. Van Gool, “Simultaneous object recognition and segmentation by image exploration,” in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 40–54.

[24] J. Čech, J. Matas, and M. Perďoch, “Efficient sequential correspondence selection by cosegmentation,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 32, no. 9, pp. 1568–1581, 2010.

[25] M. Cho and K. M. Lee, “Progressive graph matching: Making a move of graphs via probabilistic voting,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 492–505.

[26] C. Rother, T. P. Minka, A. Blake, and V. Kolmogorov, “Cosegmentation of image pairs by histogram matching - Incorporating a global constraint into MRFs,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2006, pp. 993–1000.

[27] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2011, pp. 2129–2136.

[28] A. Joulin, F. Bach, and J. Ponce, “Discriminative clustering for image co-segmentation,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2010, pp. 1943–1950.

[29] E. Kim, H. Li, and X. Huang, “A hierarchical image clustering cosegmentation framework,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 686–693.

[30] A. Faktor and M. Irani, “Co-segmentation by composition,” in *Proc. Int’l Conf. Comput. Vis.*, 2013, pp. 1297–1304.

[31] F. Wang, Q. Huang, and L. J. Guibas, “Image co-segmentation via consistent functional maps,” in *Proc. Int’l Conf. Comput. Vis.*, 2013, pp. 849–856.

[32] J. Sun and J. Ponce, “Learning discriminative part detectors for image classification and cosegmentation,” in *Proc. Int’l Conf. Comput. Vis.*, 2013, pp. 3400–3407.

[33] J. C. Rubio, J. Serrat, A. M. López, and N. Paragios, “Unsupervised co-segmentation through region matching,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 749–756.

[34] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, “Unsupervised joint object discovery and segmentation in internet images,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2013, pp. 1939–1946.

[35] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *Int. J. Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[36] A. Joulin, F. Bach, and J. Ponce, “Multi-class cosegmentation,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 542–549.

[37] P. Yarlagadda, A. Monroy, and B. Ommer, “Voting by grouping dependent parts,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 197–210.

[38] T. F. Cox and M. Cox, *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC, 2000.

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

[40] L. Manevitz and M. Yousef, “One-class SVMs for document classification,” *J. Machine Learning Research*, pp. 139–154, 2002.

[41] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[42] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[44] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[45] K. van de Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, 2010.

[46] M. Cho, J. Lee, and K. M. Lee, “Reweighted random walks for graph matching,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 144–157.

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

[48] J. Kim and K. Grauman, “Boundary preserving dense local regions,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2011, pp. 1553–1560.

[49] D. C. Hauagge and N. Snavely, “Image matching using local symmetry features,” in *Proc. Conf. Comput. Vis. and Pattern Recognit.*, 2012, pp. 206–213.

[50] Z. Wang, B. Fan, and F. Wu, “Local intensity order pattern for feature description,” in *Proc. Int’l Conf. Comput. Vis.*, 2011, pp. 603–610.

[51] E. Tola, V. Lepetit, and P. Fua, “DAISY: An efficient dense descriptor applied to wide baseline stereo,” *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

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

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

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