Computer Vision and Image Understanding 113 (2009) 726–742

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Robust feature point matching by preserving local geometric consistency

Ouk Choi *, In So Kweon

Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 24 March 2008
Accepted 5 December 2008
Available online 24 December 2008

Keywords:
Wide-baseline stereo
Object recognition
Affine regions
Pairwise constraints
Relaxation labeling

ABSTRACT

We present a method for matching feature points robustly across widely separated images. In general, it is difficult to match feature points correctly by using only the similarity between local descriptors. In our approach, the correspondence problem is formulated as an optimization problem with one-to-one correspondence constraints. A novel objective function is defined to preserve local image-to-image affine transformations across correspondences. This objective function enables our method to cope with significant viewpoint or scale changes between images, unlike previous methods that relied on the assumption that the distance or orientation between neighboring feature points are preserved across images. A relaxation algorithm is proposed for maximizing the objective function, which imposes one-to-one correspondence constraints, unlike conventional relaxation labeling algorithms that impose many-to-one correspondence constraints. Experimental evaluation shows that our method is robust with respect to significant viewpoint changes, scale changes, and nonrigid deformations between images, in the presence of repeated textures that make feature point matching more ambiguous. Our method is also applied to object recognition in cluttered environments, giving some promising results.

1. Introduction

Matching feature points across images is one of the fundamental problems in computer vision, with a variety of applications that includes 3D reconstruction [1,2], object recognition [3,4], categorization [5], and content-based image retrieval [6]. Although much research has been directed at reliable matching, there are many difficulties to overcome for practical use; the difficulties arise at the lowest level with feature point detection and description, and at the highest level with reducing ambiguity.

The main difficulty with feature point matching has been the fact that the images of the same 3D surface patch have quite different appearances if they are obtained from different viewpoints. There have been many approaches to solving this problem [3,7,8], which use the basic idea of detecting feature points together with local neighborhood regions that are covariant with the underlying viewpoint change. The detected regions are transformed into geometrically normalized regions by using local invariant transformations [8], so that the normalized regions and their descriptors will be invariant.

Although research directed at covariant region detection and invariant description has matured [8], there still remains the problem of ambiguity. The descriptors of the regions projected from different 3D surface patches can be similar because the regions are often too small to include sufficient distinctive textures. For this reason, it is often difficult to match feature points correctly using only the similarity between the local descriptors, without making other assumptions. Fig. 1 shows an example of ambiguous matching. Matching fails if we constrain one-to-one correspondence using only the similarity between the local descriptors, when the images are highly ambiguous. Various assumptions have been made to reduce the ambiguity arising from the local comparison, leading to useful constraints such as the epipolar constraint [1,9–11] and pairwise constraints [12,5,4,13].

The epipolar constraint is based on the assumption that the images are obtained concurrently or that the scene is static. The epipolar constraint is parameterized by a 3 × 3 fundamental matrix, and several wide-baseline stereo matching methods [9–11] have been proposed, together with improved algorithms based on Random Sample Consensus [15], aiming to estimate the fundamental matrix robustly. Although the fundamental matrix plays an important role not only in constraining the positions of the corresponding feature points but also in reconstructing the 3D structure of the scene, it hardly applies to object recognition, for which the images are not obtained concurrently and the objects may be near-planar and deformed nonrigidly.

Approaches that use pairwise constraints [12,5,4,13] have received a great deal of attention, especially for shape matching applications, because of their capability in matching feature points detected from deformable objects or shapes across images. In these approaches, it is often assumed that the distance or orientation between neighboring feature points are preserved across images,

* Corresponding author. Fax: +82 42 869 5465.
E-mail addresses: choi@rcv.kaist.ac.kr (O. Choi), iskweon@kaist.ac.kr (I.S. Kweon).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.12.002

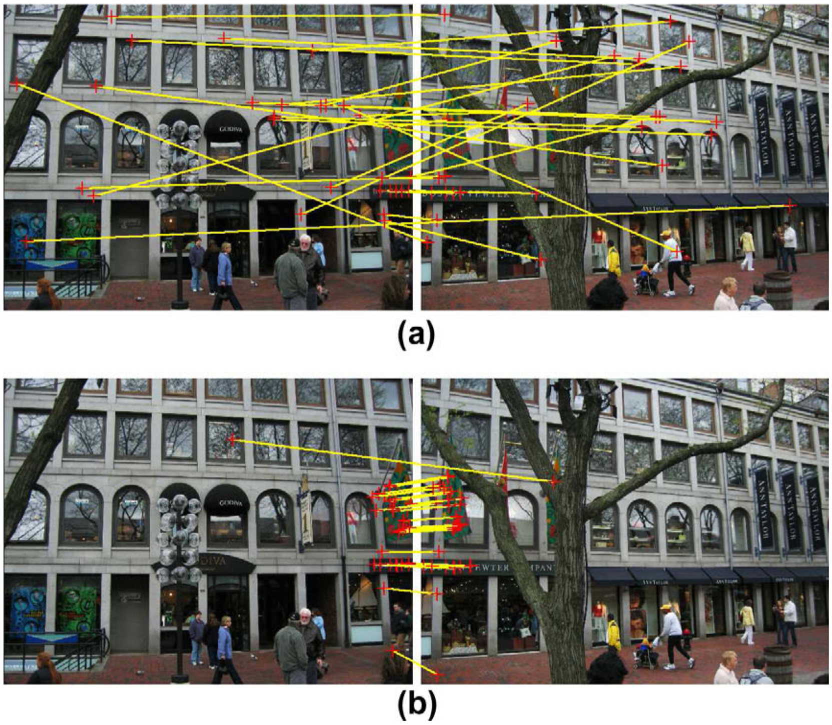


Fig. 1. A highly ambiguous image pair with small overlap and repeated textures [14]. The right part of the left image overlaps the left part of the right image. (a) Correspondences with the most similar local descriptors. (b) Correspondences detected by our method. Only the top 30 correspondences are displayed, to aid visibility. Our method detects correspondences from the actual overlapped regions.

resulting in pairwise constraints that aim to preserve them [12,5,4]. The constraints are usually adapted to an objective function instead of being applied as hard constraints, and the correspondence problem becomes an optimization problem with mapping constraints such as many-to-one or one-to-one correspondence constraints. The optimization problem is an NP-hard integer quadratic programming problem in which either 0 or 1 should be assigned to every candidate correspondence, so these approaches avoid combinatorial searching by approximating the objective function or by relaxing the mapping constraints.

To the best of our knowledge, Zhang et al. [12] first used pairwise constraints for feature point matching across images. In their approach, correspondences are initially detected by maximizing their objective function using a winner-take-all strategy, and the detected correspondences are used for estimating fundamental matrices. Berg et al. [5] used pairwise constraints for shape matching across images. They change the problem into simpler subproblems by approximating the quadratic objective function as linear functions for which integer linear programming finds the global optimum. Leordeanu and Hebert [4] applied pairwise constraints to both shape matching and feature point matching across images. They relax the integer constraints (assigning 0 or 1 to every correspondence) as well as the mapping constraints, and solve the relaxed problem with a spectral technique; they use a greedy algorithm as a postoptimization step to find a discretized solution. Zheng and Doermann [13] applied pairwise constraints to hand-drawn shape alignment problems. They relax the integer constraints into real-value constraints, and solve the problem using a relaxation labeling algorithm [16–18] that maximizes the objective function through simple iterative updates.

Although good optimization algorithms have been proposed in the approaches that use pairwise constraints [12,5,4,13], the underlying assumptions may not be suitable for solving the wide-baseline stereo matching problem that we are interested in, because the distance and orientation between feature points are not well preserved across widely separated images. In this paper, we assume that local image-to-image affine transformations, the so-called local feature transformations [19], are well preserved across correspondences. The images may not be obtained concurrently, and the scene objects may be deformed nonrigidly, provided any deformation is continuous. Fig. 2 shows our assumption.

A local feature transformation is a mapping from the neighborhood of a feature point in an image to the neighborhood of a potentially corresponding feature point in the other image; readers interested in estimating such transformations may refer to [8,20,19]. Local feature transformations have been used for the (quasi-)dense propagation of feature correspondences [21,22,20]. Ferrari et al. [21] matched regularly quantized regions in an image to possibly irregular regions in the other image by assuming that the regions in the neighborhood of a matched pair of features have a similar transformation to the local feature transformation of the matched pair. Based on a similar assumption, Vedaldi and Soatto [22] densely propagated a feature correspondence to the neighboring pixels, and Kannala and Brandt [20] extended the original match propagation [23], so that it can be applied to widely separated images. Although our work relies on a similar assumption, we focus on finding correct feature point correspondences rather than densely propagating feature correspondences that may be incorrect if the images are highly ambiguous. It is clear that the propagation approaches may benefit from the consistent feature point correspondences that we are aiming to provide.

1.1. Proposed approach

In this paper, we aim to reduce ambiguity in matching feature points across widely separated images. We formulate the correspondence problem as an optimization problem with one-to-one correspondence constraints. In our approach, a set of candidate

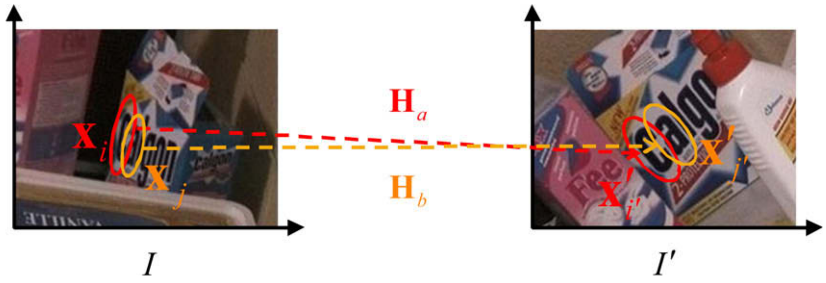


Fig. 2. The figure illustrates the assumption in this paper. The images I and I' were adapted from [7]. Feature points $\mathbf{x}_i$ and $\mathbf{x}_j$ in image I correspond to feature points $\mathbf{x}'_{i'}$ and $\mathbf{x}'_{j'}$ in image I', respectively. The local feature transformations $\mathbf{H}_a$ and $\mathbf{H}_b$ transform the neighborhoods of $\mathbf{x}_i$ and $\mathbf{x}_j$ to the neighborhoods of $\mathbf{x}'_{i'}$ and $\mathbf{x}'_{j'}$ respectively. Our assumption is that local feature transformations are well preserved across correspondences; $\mathbf{H}_a$ and $\mathbf{H}_b$ should be similar under this assumption.

correspondences is given as an input by using conventional local descriptor-based matching methods [7,3], without enforcing the mapping constraints. From the set of candidate correspondences, we aim to find a subset not only maximizing the objective function but also satisfying one-to-one correspondence constraints.

Our formulation is similar to those of previous approaches [12,5,4,13] that use pairwise constraints; however, affine viewpoint changes are taken into consideration. We define an objective function to preserve local feature transformations [19] across correspondences, unlike those of previous approaches [12,5,4] that were defined to preserve the distance or orientation between neighboring feature points across images. This newly defined objective function enables our feature point matching method to cope with significant viewpoint or scale changes between images.

We propose a relaxation algorithm for maximizing the objective function. Our algorithm is based on conventional relaxation labeling algorithms [16–18]. However, it differs from them in terms of the mapping constraints. Our algorithm imposes one-to-one correspondence constraints, unlike the conventional algorithms [16–18] that impose many-to-one correspondence constraints. We consider that our algorithm is more suitable for feature point matching problems, because information is more rapidly conveyed from unambiguous correspondences to ambiguous correspondences in our algorithm. This property may not be important for general point matching applications, for which the points may not have neighborhood patterns to be compared. In our problem, some feature points may have distinctive neighborhood patterns that result in unambiguous correspondences, and the property may be more effective. The property will be revisited in Section 3.

The remainder of this paper is organized as follows. In Section 2, we introduce our novel objective function and formulate feature point matching as a constrained optimization problem. In Section 3, our relaxation algorithm is described. In Section 4, the effectiveness of our method is demonstrated via experiments involving image pairs with significant viewpoint changes, scale changes, nonrigid deformations, and repeated textures. Our method is also applied to object recognition in Section 4. Finally, Section 5 concludes the paper.

2. Problem definition

In this section, we introduce our novel objective function, and formulate feature point matching as a constrained optimization problem. Our algorithm requires a set of candidate correspondences as an input. Although all possible pairs of feature points could be such an input, in practice we have to limit the number of candidate correspondences because computers have a limited amount of memory. We briefly explain how to find such a set of candidate correspondences, before formulating our problem.

2.1. Detection of candidate correspondences

We use affine invariant features, namely the Scale Invariant Feature Transform (SIFT) in a Maximally Stable Extremal Region (MSER) [7,3], although our method can be generalized to use other features, because significant viewpoint or scale changes are assumed between images $\mathscr{I}$ and $\mathscr{I}'$. Affine regions [7] are detected from images $\mathscr{I}$ and $\mathscr{I}'$, and we denote the regions as $\mathscr{R}_i$ and $\mathscr{R}'_{i'}$, respectively, where $i = 1, \ldots, N_{\mathscr{I}}$ and $i' = 1, \ldots, N'_{\mathscr{I}}$. A feature point $\mathbf{x}_i$ is defined as the centroid of an affine region $\mathscr{R}_i$, and $\mathbf{x}'_{i'}$ is defined as the centroid of $\mathscr{R}'_{i'}$. A feature point $\mathbf{x}_i$ is tentatively matched to a feature point $\mathbf{x}'_{i'}$ if the Euclidean distance d_a between the local descriptors (128-dimensional SIFT vectors [3]) $\mathbf{v}_i$ and $\mathbf{v}'_{i'}$ of the feature points $\mathbf{x}_i$ and $\mathbf{x}'_{i'}$ is smaller than a threshold value τ_d. We also limit the number N of candidate correspondences, for reasons of computational tractability, by taking N_{max} correspondences.

The set of candidate correspondences is denoted by:

$$\mathscr{M} = \{\mathbf{m}_a = (i, i') : a = 1, \ldots, N\}. \tag{1}$$

Note that the elements in $\mathscr{M}$ may not satisfy one-to-one correspondence constraints; $\mathscr{M}$ may contain correspondences of the form (i, j') or (k, i'), in conflict with (i, i') under one-to-one correspondence constraints.

For each correspondence $\mathbf{m}_a$, we consider not only the local-descriptor distance d_a, but also the local feature transformation $\mathbf{H}_a$ that maps the neighborhood of the feature point $\mathbf{x}_i$ to the neighborhood of the feature point $\mathbf{x}'_{i'}$ [19]. $\mathbf{H}_a$ is an affine transformation that can be parameterized by $\mathbf{x}_i$, $\mathbf{x}'_{i'}$ and $\mathbf{A}_a$, where $\mathbf{A}_a$ is a 2×2 matrix that can be decomposed as:

$$\mathbf{A}_a = \mathbf{\Sigma}'^{-\frac{1}{2}}_{i'} \mathbf{R} \mathbf{\Sigma}^{\frac{1}{2}}_{i}, \tag{2}$$

where $\mathbf{\Sigma}_i$ and $\mathbf{\Sigma}'_{i'}$ are the covariance matrices of the interior pixels of $\mathscr{R}_i$ and $\mathscr{R}'_{i'}$, respectively, and $\mathbf{R}$ is an orthogonal matrix that transforms the reference orientation of $\mathscr{R}_i$ into that of $\mathscr{R}'_{i'}$. A point $\mathbf{x}$ neighboring $\mathbf{x}_i$ is approximately transformed into a point $\mathbf{x}'$ neighboring $\mathbf{x}'_{i'}$ by the following equation:

$$\mathbf{x}' = \mathbf{A}_a(\mathbf{x} - \mathbf{x}_i) + \mathbf{x}'_{i'}. \tag{3}$$

The orthogonal matrix $\mathbf{R}$ may not be unique for a correspondence $\mathbf{m}_a$ because the reference orientation of a region is determined as a dominant image-gradient orientation that may not be unique [3]. Consequently, $\mathscr{M}$ may contain candidate correspondences that share both of the feature points but have different local feature transformations; these correspondences are also in conflict with each other. For this reason, if necessary, we sometimes write a correspondence as $\mathbf{m}_a = (\mathbf{x}_i, \mathbf{x}'_{i'})$ or $\mathbf{m}_a = (\mathbf{x}_i, \mathbf{x}'_{i'}, \mathbf{H}_a)$ to avoid notational ambiguity.

2.2. Objective function

For every pair of correspondences, $\mathbf{m}_a = (\mathbf{x}_i, \mathbf{x}'_{i'}, \mathbf{H}_a)$ and $\mathbf{m}_b = (\mathbf{x}_j, \mathbf{x}'_{j'}, \mathbf{H}_b)$, we can define a pairwise transformation error e_{ab} as follows:

$$\begin{aligned} e_{ab} &= e(b|a) + e(a|b), \\ e(b|a) &= \|\mathbf{x}'_{j'} - \mathbf{H}_a(\mathbf{x}_j)\| + \|\mathbf{x}_j - \mathbf{H}_a^{-1}(\mathbf{x}'_{j'})\|, \\ e(a|b) &= \|\mathbf{x}'_{i'} - \mathbf{H}_b(\mathbf{x}_i)\| + \|\mathbf{x}_i - \mathbf{H}_b^{-1}(\mathbf{x}'_{i'})\|, \end{aligned} \tag{4}$$

where $\mathbf{H}(\mathbf{x})$ denotes the point given by the transformation of $\mathbf{x}$ by $\mathbf{H}$. The error e_{ab} will be small if $\mathbf{H}_a$ and $\mathbf{H}_b$ are similar to each other, and will be equal to zero when the two transformations are identical, because $e(a|a) = e(b|b) = 0$ from (3). If two correspondences $\mathbf{m}_a$ and $\mathbf{m}_b$ are detected from the images of a smooth surface that can be locally approximated by planar patches, then we can expect the error e_{ab} to be small, although the converse is not always true.

A cost function $h(\mathscr{M}^*)$ may be defined as the sum of the error e_{ab} to find $\mathscr{M}^* \subset \mathscr{M}$ whose elements have similar local feature transformations:

$$h(\mathscr{M}^*) = \sum_{\mathbf{m}_a \in \mathscr{M}^*} \sum_{\mathbf{m}_b \in \mathscr{M}^*} e_{ab}. \tag{5}$$

However, the solution $\mathscr{M}^*$ minimizing the cost function (5) is always trivial, namely $\mathscr{M}^* = \phi$, and the trivial solution also satisfies one-to-one correspondence constraints. To avoid the trivial solution, we convert the minimization problem into a maximization problem by transforming the error e_{ab} into a binary compatibility weight w_{ab} that increases with decreasing error e_{ab} (*e.g.*, $w_{ab} = \exp(-e_{ab}^2/2\sigma^2)$):

$$h(\mathscr{M}^*) = \sum_{\mathbf{m}_a \in \mathscr{M}^*} \sum_{\mathbf{m}_b \in \mathscr{M}^*} w_{ab}. \tag{6}$$

The solution $\mathscr{M}^*$ maximizing the function (6) is $\mathscr{M}$; however, the elements in $\mathscr{M}$ may not satisfy one-to-one correspondence constraints as we discussed in the previous subsection.

To define our objective function more formally, we define a confidence value p_a for each candidate correspondence $\mathbf{m}_a$, simply referred to the confidence of $\mathbf{m}_a$. The confidence p_a takes a value in $\{0,1\}$: $p_a = 1$ if $\mathbf{m}_a \in \mathscr{M}^*$, and $p_a = 0$ otherwise. We denote the set of confidences as $\mathscr{P}$:

$$\mathscr{P} = \{p_a : a = 1, \ldots, N\}. \tag{7}$$

Under the definition of $\mathscr{P}$, the function $h(\mathscr{M}^*)$ is equivalent to a binary objective function $h(\mathscr{P})$ defined as:

$$h(\mathscr{P}) = \sum_{a=1}^{N} \sum_{b=1}^{N} w_{ab} p_a p_b. \tag{8}$$

Now, the problem becomes to find $\mathscr{P}$ that maximizes $h(\mathscr{P})$ while satisfying one-to-one correspondence constraints.

In previous work, the weight w_{ab} has been defined to encourage correspondences that preserve the distance or orientation between the feature points across images [12,5,4]. Unlike the previous definitions, we define the weight w_{ab} as a nonnegative decreasing function of e_{ab} so that pairs of correspondences with similar local feature transformations can be encouraged:

$$w_{ab} = \begin{cases} \exp(-e_{ab}^2/2\sigma^2), & \text{if } b \notin \mathscr{C}_a \text{ and } e_{ab} < 3\sigma, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\mathscr{C}_a$ is the set of the index a of a correspondence $\mathbf{m}_a = (i, i')$ and the indices of all the candidate correspondences in conflict with $\mathbf{m}_a$:

$$\mathscr{C}_a = \{b : \mathbf{m}_b = (j, j') \text{ such that } j = i \text{ or } j' = i'\}. \tag{10}$$

The weight w_{ab} does not construct a link, namely is equal to zero, if two correspondences $\mathbf{m}_a$ and $\mathbf{m}_b$ are identical or in conflict with each other; note that $a \in \mathscr{C}_a$. It is natural not to construct a link between conflicting correspondences because one of them must be incorrect under one-to-one correspondence constraints, and it is reasonable not to construct a link between a correspondence and itself. The weight w_{ab} also does not construct a link if e_{ab} is greater than 3σ. The truncation makes our method more efficient, because it is then not necessary to consider null links during the computation in Section 3.

The parameter σ can either be computed adaptively or chosen manually. An adaptively computed σ value is:

$$\sigma = \frac{1}{N} \sum_{a=1}^{N} (\min_{b \in \{1,\ldots,N\} - \mathscr{C}_a} (e_{ab})), \tag{11}$$

where *min* denotes the minimum value. The adaptive σ value (11) was determined on the assumption that a correct correspondence $\mathbf{m}_a$ would have a smaller value of the minimum error $(\min_{b \in \{1,\ldots,N\} - \mathscr{C}_a} (e_{ab}))$ than incorrect correspondences. We use this value unless otherwise specified.

In a similar manner of defining the binary objective function $h(\mathscr{P})$, we define a unary objective function $g(\mathscr{P})$ that encourages correspondences with small local-descriptor distances.

$$g(\mathscr{P}) = \sum_{a=1}^{N} w_a p_a, \tag{12}$$

where w_a is a unary compatibility weight that is a decreasing function of the local-descriptor distance d_a:

$$w_a = 1 - d_a. \tag{13}$$

Because we use normalized descriptor vectors $\mathbf{v}$ such that $\|\mathbf{v}\| = 1$ and all the elements of $\mathbf{v}$ are nonnegative, the distance d_a may be greater than 1. However, the threshold τ_d is usually set to 0.5, so the weight w_a is nonnegative. We tested various type of function that includes linear, exponential and Gaussian functions to define the weights w_{ab} and w_a; and we use Gaussian and linear weights for w_{ab} and w_a, respectively, because the combination gives the best results although the performance gap is not so significant.

Finally, our objective function is defined as the sum of the unary and binary objective functions:

$$f(\mathscr{P}) = g(\mathscr{P}) + h(\mathscr{P}). \tag{14}$$

There are some interesting issues such as balancing the unary and binary objective functions [24]; however, such issues are out of the scope of this paper.

The objective function $f(\mathscr{P})$ is maximized when $p_a = 1$ for every $a \in \{1, \ldots, N\}$, *i.e.*, $\mathscr{M}^* = \mathscr{M}$, without any mapping constraints. The condition for $\mathscr{P}$ to satisfy one-to-one correspondence constraints can be formally described as:

$$s_a = \sum_{b \in \mathscr{C}_a} p_a = 1, \quad \forall a \in \{a : p_a = 1\}. \tag{15}$$

Our problem is to find $\mathscr{P}$ that maximizes the objective function $f(\mathscr{P})$ and satisfies the condition (15). The proof for the equivalence between the condition (15) and one-to-one correspondence constraints can be found in Appendix A.

3. Relaxation algorithm

In this section, we present our optimization algorithm. Before presenting the algorithm, we address two basic properties that an algorithm for maximizing the objective function (14) should have. First, a good algorithm should find a solution satisfying the constraints; this property is denoted by P1. If any kind of relaxation is used (*e.g.*, $p_a \in [0,1], \forall a \in \{1,\ldots,N\}$ or $\|\mathbf{p}\| = 1$, where $\mathbf{p}$ is a vector containing the elements of $\mathscr{P}$), the final values of $\mathscr{P}$ should satisfy the condition (15) as closely as possible, because the integer constraint and one-to-one correspondence constraints are satisfied if the condition (15) is satisfied. Suppose that $\mathscr{P}^*$ is the optimal solution of the relaxed problem, whose elements do not satisfy the condition (15), and suppose that $\hat{\mathscr{P}}^*$ is a discretized solution computed from $\mathscr{P}^*$, whose elements satisfy the condition (15). If the final values of $\mathscr{P}^*$ are much different from those of $\hat{\mathscr{P}}^*$, then it is clear that $f(\mathscr{P}^*)$ does not approximate $f(\hat{\mathscr{P}}^*)$ well. From this point of view, the first property is important. The second basic property is that a good algorithm should find a solution that maximizes the objective function; this property is denoted by P2.

The approximate integer quadratic programming algorithm proposed by Berg et al. [5] has the second property (P2) approximately. The algorithm, however, does not have the first property (P1) because their formulation allows several features in image $\mathscr{I}$ to match the same feature in image $\mathscr{I}'$. The spectral technique proposed by Leordeanu and Hebert [4] has the second property (P2) approximately but strongly; the algorithm finds the global maximum of the relaxed problem efficiently. The algorithm, however, does not have the first property (P1) because of the strong relaxation, namely $\|\mathbf{p}\| = 1$. As a postoptimization step, they use a greedy algorithm to compute a discretized solution so that their final correspondences can satisfy the constraints. Relaxation labeling algorithms [16–18] have the second property (P2) weakly. The algorithms are based on simple update equations that can be regarded as a gradient ascent combined with a normalization; the algorithms find a local maximum. The algorithms [16–18], in their original form, are suitable only for correspondence problems with many-to-one correspondence constraints and therefore do not have the first property (P1).

3.1. Proposed algorithm

In our algorithm, the confidence p_a is relaxed to take on real values in $[0,1]$ for every $a \in \{1,\ldots,N\}$, to avoid combinatorial searching. Although the relaxation is the same as those in conventional relaxation labeling algorithms [16–18], our algorithm imposes the condition (15), unlike conventional algorithms.

Because we relax p_a to take on real values in $[0,1]$, a partial derivative q_a with respect to p_a can be computed as follows:

$$q_a = \frac{\partial f(\mathscr{P})}{\partial p_a} = w_a + 2\sum_{b=1}^{N} w_{ab} p_b, \quad \forall a \in \{1,\ldots,N\}. \tag{16}$$

We can see that the product of p_a and q_a is the contribution of $\mathbf{m}_a$ to $f(\mathscr{P})$:

$$f(\mathscr{P}) = f(\mathscr{P} - \{p_a\}) + p_a q_a. \tag{17}$$

A correspondence $\mathbf{m}_a$, therefore, can be considered good if $p_a q_a$ is large, because $\mathbf{m}_a$ contributes to $f(\mathscr{P})$ by a large amount. Motivated by this simple discussion, and by the fact that q_a is nonnegative, we propose an update equation replacing p_a by $p_a q_a$, followed by a normalization that imposes $p_a \in [0,1]$ and $s_a = \sum_{b\in\mathscr{C}_a} p_b = 1$.

$$p_a^{(t+1)} \leftarrow \frac{p_a^{(t)} q_a^{(t)}}{\sum\limits_{b\in\mathscr{C}_a} p_b^{(t)} q_b^{(t)}}, \quad \forall a \in \{1,\ldots,N\}, \tag{18}$$

where $p_a^{(t)}$ denotes p_a at time $t \in \{0,\ldots,T\}$, and $p_a^{(0)}$ is set to a constant p_0 for every $a \in \{1,\ldots,N\}$.

The relaxed solution $\mathscr{P}^*$ found by using our algorithm is defined as:

$$\mathscr{P}^* = \{p_a^* : p_a^* = p_a^{(T+1)}, \quad a = 1,\ldots,N\}, \tag{19}$$

and the discretized solution $\hat{\mathscr{P}}^*$ is defined as:

$$\hat{\mathscr{P}}^* = \{\hat{p}_a^* : a = 1,\ldots,N\}, \tag{20}$$

where $\hat{p}_a^* = 1$ if $p_a^* > p_b^*$ for every $b \in \mathscr{C}_a - \{a\}$, and $\hat{p}_a^* = 0$ otherwise. Finally, the solution set $\mathscr{M}^*$ of correspondences can be derived from $\hat{\mathscr{P}}^*$ as follows:

$$\mathscr{M}^* = \{\mathbf{m}_a : \hat{p}_a^* = 1\}. \tag{21}$$

From the discussion in Section 2, our algorithm can be considered to have the first property (P1) very closely, if $p_a^* \approx 1$ and $s_a^* = \sum_{b\in\mathscr{C}_a} p_b^* \approx 1$ for every $a \in \mathscr{S}^*$, where $\mathscr{S}^*$ is defined as a set of the correspondence indices with $\hat{p}_a^* = 1$:

$$\mathscr{S}^* = \{a : \hat{p}_a^* = 1\}. \tag{22}$$

Fig. 3 shows how closely the proposed algorithm satisfies the first property (P1) for the image pair in Fig. 1; one-to-one correspondence constraints are more closely satisfied with increasing T. Experimental evaluations similar to this one will be given at the end of Section 4 for other image pairs.

If we use a conventional relaxation labeling algorithm [18], the update equation becomes (23) by considering a feature point in image $\mathscr{I}'$ as a label.

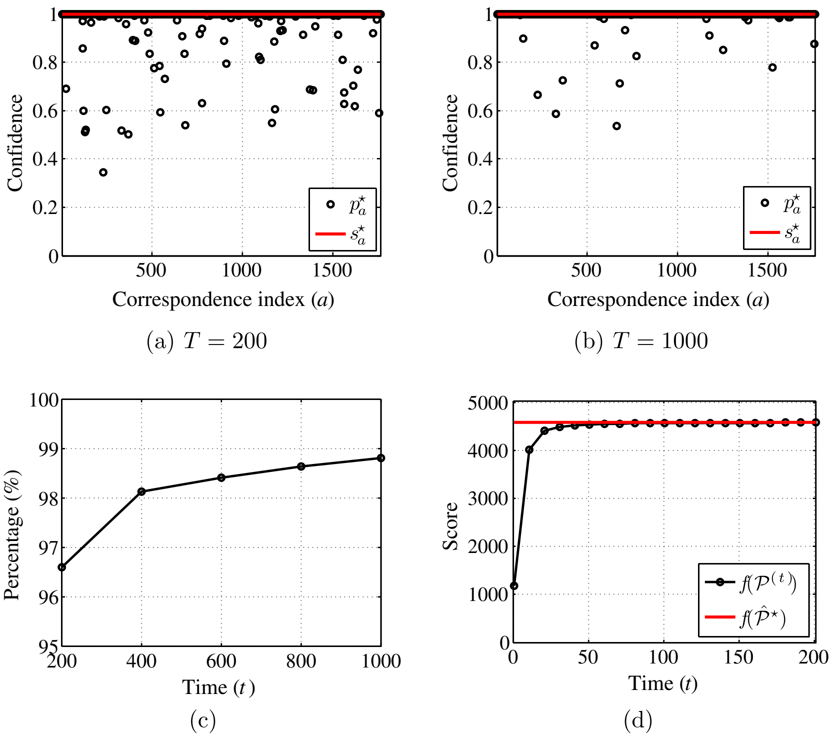
$$p_a^{(t+1)} \leftarrow \frac{p_a^{(t)} q_a^{(t)}}{\sum\limits_{b\in\mathscr{C}_a^{row}} p_b^{(t)} q_b^{(t)}}, \quad \forall a \in \{1,\ldots,N\}, \tag{23}$$

where $\mathscr{C}_a^{row} = \{b : \mathbf{m}_b = (j,j')\text{such that} j = i\}$. The update Eq. (23) performs normalization only in one direction (*i.e.*, $j = i$), so many-to-one correspondence is allowed.

Zheng and Doermann [13] suggested using alternated row and column normalizations [25] after each relaxation labeling update (23), if one-to-one matching is necessary. Because of outliers that do not have a corresponding feature point in the other image, the alternated normalization algorithm [13] uses dummy points that are matched to the outliers; the dummy points are also used in a soft assignment approach [25]. We applied the alternated normalization algorithm [13] to matching the images in Fig. 1, with 50 alternated row and column normalizations after each relaxation labeling update (23). Fig. 4 shows the results. We can see that one-to-one correspondence constraints are more closely satisfied by our algorithm with the same number of updates. Although Zheng and Doermann detect correspondences whose confidence is greater than p_{min} ($p_{min} = 0.95$) by matching the remaining feature points to the dummy points [26], we kept our definition of the discretized solution (20) for the comparison in Figs. 3 and 4 except that we harvest correspondences with a higher confidence than the confidence of matching either feature point to a dummy point. The percentage of correspondences satisfying the condition (15) in Fig. 4 grows from 55.8% into 95.4% at $T = 1000$, if the definition by Zheng and Doermann [26] is used. We checked that the converged confidence values are equivalent to a doubly stochastic matrix whose row and column sums are 1 except the last row and column [26]; however, we could not observe strong convergence of the confidence values to either 1 or 0 when using the alternated normalization algorithm [13].

Although it is clear from Fig. 3 that a large T value is better for satisfying the constraints more closely, we cannot set T to an arbitrarily large value because the running time of our algorithm is usually $O(NT)$. As a trade-off, T is determined adaptively by means of the following two conditions: T should be less than T_{max}, and the percentage of converged confidences, *e.g.*, $p_a^{(t)} < 0.01$ or $p_a^{(t)} > 0.99$, should be greater than τ_p. If either of these two conditions is satisfied, the iteration terminates.

It is not easy to show theoretically that the proposed algorithm has the second property (P2). One thing that we can do is to compare the values of $f(\hat{\mathscr{P}}^*)$ to those computed by state-of-the-art



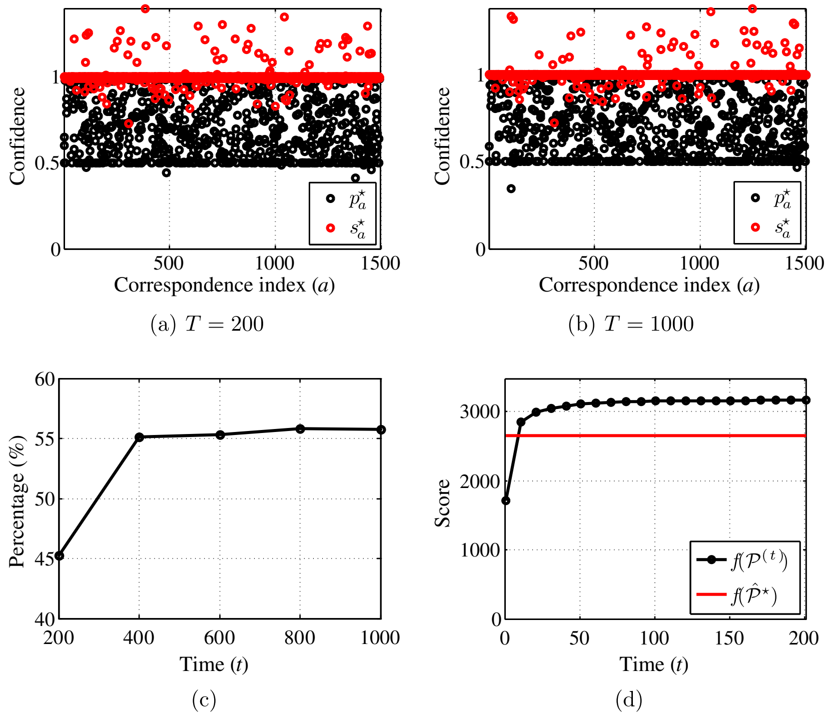
(a) $T = 200$

(b) $T = 1000$

(c)

(d)

Fig. 3. Evaluation of the property P1 for the image pair in Fig. 1. The proposed algorithm (18) has been used for the evaluation. (a) The graph displays p_a^* and s_a^* values for every $a \in \mathscr{S}^*$. For the case of one-to-one correspondence constraints being strictly satisfied, both values must be 1 (refer to Section 2 for details). p_a^* is greater than 0.99 (thus, $p_a^* \approx 1$) for 96.6% of the detected correspondences, such that $\hat{p}_a^* = 1$ after 200 updates. (b) The percentage increases to 98.8% after 1000 updates. For both of these cases, $s_a^* \approx 1$ for every $a \in \mathscr{S}^*$. (c) The graph displays the percentage as a function of T. One-to-one correspondence constraints are more closely satisfied with increasing T. (d) The graph displays the relaxed value of $f(\mathscr{P})$ as a function of t. $f(\mathscr{P}^{(t)})$, the relaxed value of $f(\mathscr{P})$, monotonically increases, and approaches the value of $f(\hat{\mathscr{P}}^*)$ with increasing t. We can see that $f(\hat{\mathscr{P}}^*)$ is well approximated by $f(\mathscr{P}^{(t)})$.



(a) $T = 200$

(b) $T = 1000$

(c)

(d)

Fig. 4. Evaluation of the property P1 for the image pair in Fig. 1. The alternated normalization algorithm [13] have been used for the evaluation. (a) The graph displays p_a^* and s_a^* values for every $a \in \mathscr{S}^*$. For the case of one-to-one correspondence constraints being strictly satisfied, both values must be 1 (refer to Section 2 for details). p_a^* is greater than 0.99 and s_a^* is less than 1.01 (thus, $p_a^* \approx 1$ and $s_a^* \approx 1$) for 45.2% of the detected correspondences, such that $\hat{p}_a^* = 1$ after 200 updates. (b) The percentage increases to 55.8% after 1000 updates. (c) The graph displays the percentage as a function of T. The percentage slowly grows with increasing T; however, it stops increasing at $T = 800$. (d) The graph displays the relaxed value of $f(\mathscr{P})$ as a function of t. $f(\mathscr{P}^{(t)})$, the relaxed value of $f(\mathscr{P})$, does not approximate the value of $f(\hat{\mathscr{P}}^*)$.

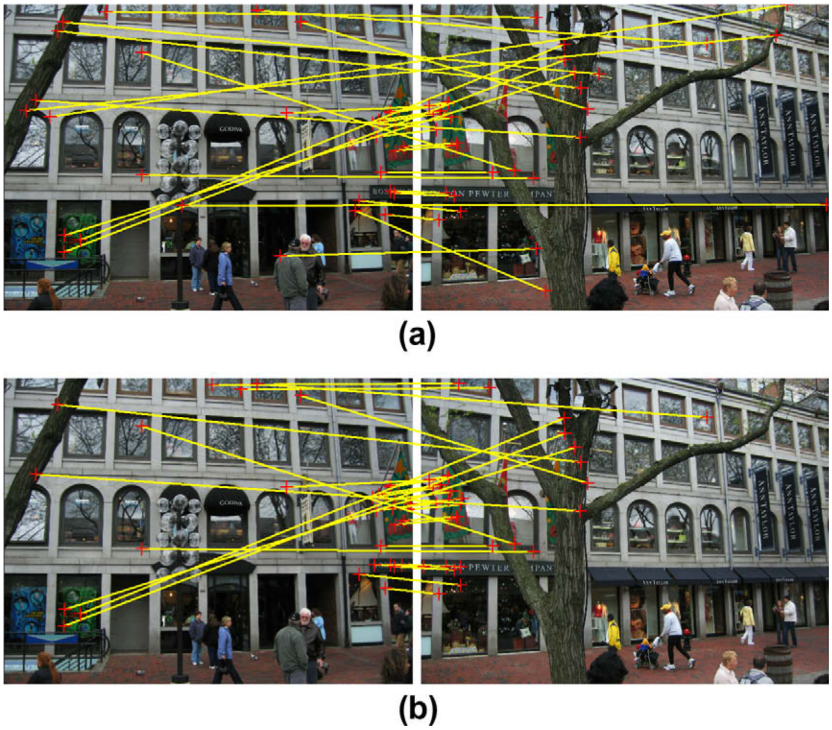


Fig. 5. (a) Unambiguous correspondences whose confidences converge to 1 after the first update. (b) Correspondences whose confidences converge to 1 (greater than 0.99) after the second update. Note that these correspondences also include the unambiguous correspondences in (a), although only the top 30 correspondences are displayed, to aid visibility. Additional correct correspondences linked to the unambiguous correct correspondences are detected after the second update.

algorithms such as the spectral technique [4]; the comparison will be given at the end of Section 4.

Our algorithm has a third important property (denoted by P3) for feature point matching. Conventional relaxation labeling algorithms [16–18] also have this property if one-to-one matching is not required. The set of candidate correspondences $\mathscr{M}$ may contain a correspondence $\mathbf{m}_a$ without any correspondences in conflict with it, *i.e.*, $\mathscr{C}_a = \{a\}$, because we limit the number of candidate corre-

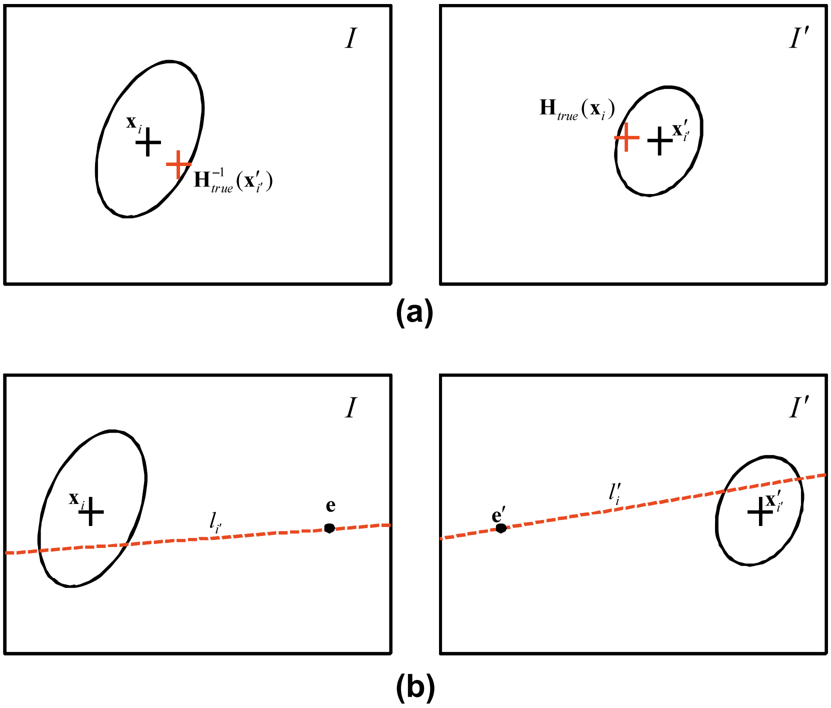


Fig. 6. (a) An example of an inlier with respect to a ground-truth homography. Both of the transformed feature points lie inside the fitted ellipses of the corresponding feature points. (b) An example of an inlier with respect to a ground-truth fundamental matrix. The epipolar line induced by each feature point intersects the fitted ellipse of the corresponding feature point.

spondences. For such a correspondence $\mathbf{m}_a$, the confidence $p_a^{(t)}$ is 1 for every t such that $t \geqslant 1$, irrespective of the choice of the objective function and the initial confidence value p_0, because $\sum_{b\in\mathscr{C}_a} p_b^{(t)} q_b^{(t)}$ is always equal to $p_a^{(t)} q_a^{(t)}$ for such a correspondence $\mathbf{m}_a$. Let these correspondences be called unambiguous correspondences, in that they already satisfy one-to-one correspondence constraints. The early-converged confidences can affect other confidences more when the $q_b^{(t)}$ of the geometrically linked correspondences $\mathbf{m}_b$ is increased. This means that the confidence flows from the unambiguous correspondences to ambiguous correspondences. Under the assumption that correct correspondences are more geometrically consistent, correct unambiguous correspondences will affect other correspondences more than incorrect unambiguous correspondences. Fig. 5 shows this effect on the image pair in Fig. 1. This property may not be important for general point matching applications, for which the points may not have neighborhood regions to be compared. In our problem, some feature points may have distinctive neighborhood regions resulting in unambiguous correspondences, so the property is more effective. We consider this property to be a crucial advantage in comparison with other algorithms such as integer linear programming [5] and the spectral technique [4].

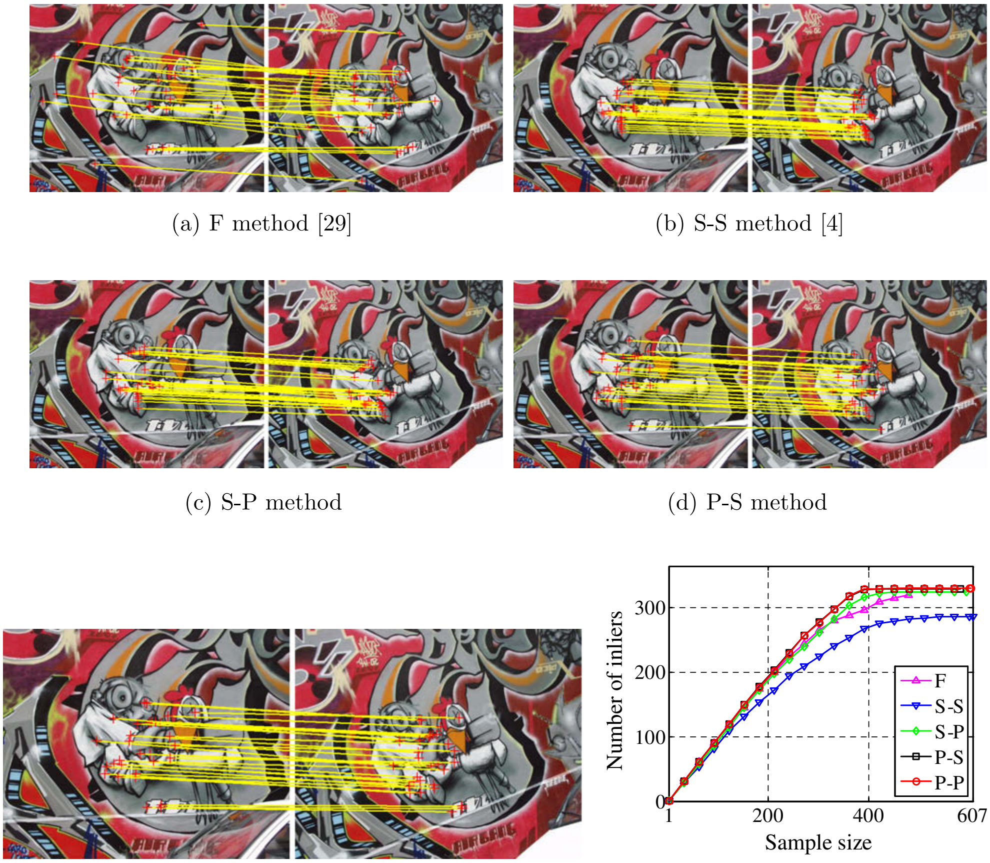
4. Experiments

In this section, we assess the proposed method via experiments with various types of image pairs. In the first two subsections, ground-truth global parametric models, such as homographies and fundamental matrices between the images, are given. We count correct correspondences using these models. The third subsection describes the application of the proposed method to matching images containing nonrigid objects. In the last subsection, we apply the proposed method to object recognition.

We implement four other methods for comparison. The first method is the texture-descriptor-based matching method proposed by Forssén and Lowe [27] (denoted by the F method), which uses local-descriptor distances only. We use the same feature detection and description methods as those used in this paper for fairness of comparison, although Forssén and Lowe [27] use multiscale features in their work. The detected correspondences are sorted in increasing order of the ratio between the best and the second-best dissimilarities [27], and one-to-one correspondence constraints are enforced by eliminating conflicting correspondences based on the sorting results. In addition, we reject correspondences with large local-descriptor distances ($d_a > \tau_d$) at the final stage, which we consider as bad correspondences with high dissimilarity.

The remaining three methods share the same unary objective function $g(\mathscr{P})$ as that of the proposed method. In fact, these methods are three different combinations of the objective functions and the optimization algorithms of the proposed method and the spectral method [4]. By decomposing the two methods component-by-component and recombining them in this way, we can identify which part of the method is the most effective.

The binary compatibility weight between $\mathbf{m}_a = (\mathbf{x}_i, \mathbf{x}'_{i'})$ and $\mathbf{m}_b = (\mathbf{x}_j, \mathbf{x}'_{j'})$, used by Leordeanu and Hebert [4] for matching SIFT features, can be described by:



(a) F method [29]

(b) S-S method [4]

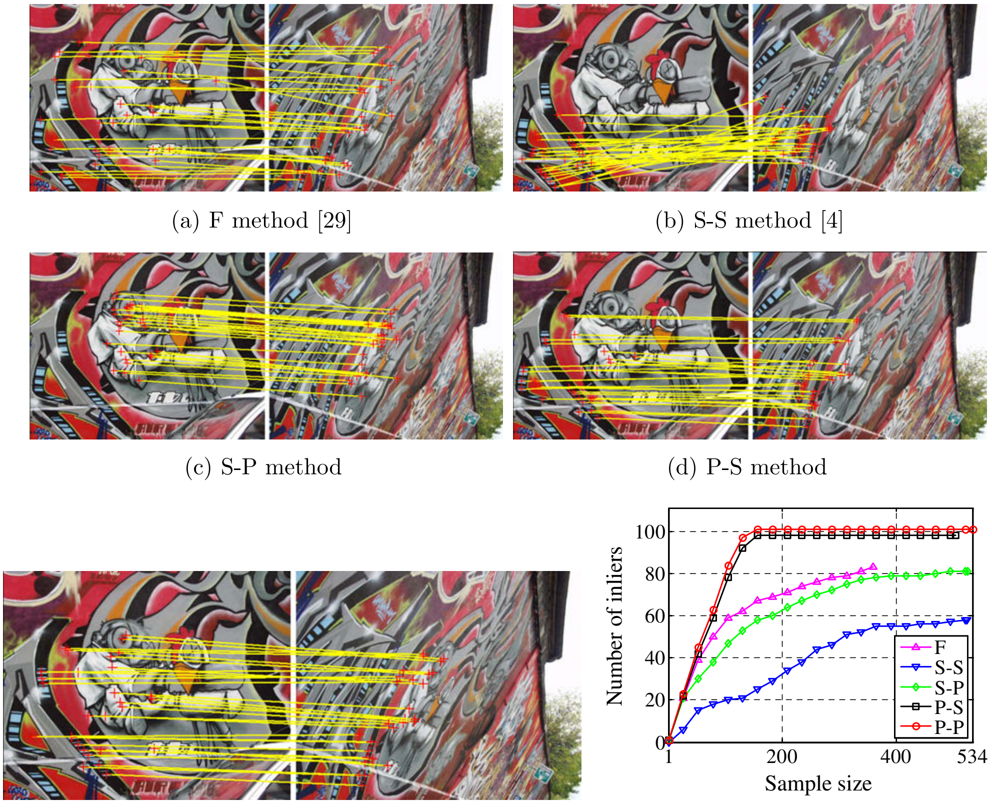
(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 7. A planar scene with a small viewpoint change [8]. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).



(a) F method [29]

(b) S-S method [4]

(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 8. A planar scene with a large viewpoint change [8]. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct ($k =$ sample size).

$$w_{ab} = \begin{cases} 1 - e_{ab}^2/\gamma^2, & \text{if } b \notin \mathscr{C}_a \text{ and } e_{ab} < \gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where:

$$e_{ab} = |\|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{x}'_{i'} - \mathbf{x}'_{j'}\||. \tag{25}$$

Additional constraints, for w_{ab} to have a nonzero value, are:

$$\begin{aligned} &\|\mathbf{x}_i - \mathbf{x}_j\| < r, \\ &\|\mathbf{x}'_{i'} - \mathbf{x}'_{j'}\| < r, \end{aligned} \tag{26}$$

where r is usually set to 200 pixels. The simple binary compatibility weight (24) encourages pairs of correspondences to preserve the distance between the feature points across images; it consequently tends to preserve the object scale. We set γ to be about one-tenth of the image size (usually 50 pixels), because the proposed adaptive parameter selection scheme was found not to work for this simple binary compatibility function.

The second method is the original spectral method [4] (denoted by the S-S method) that uses the simple binary compatibility weight (24) and the spectral technique. The ARPACK software [28] was used for the implementation of the spectral technique, and the greedy algorithm [4] was used to find discretized solutions. The third method (denoted by the S-P method) uses the simple binary compatibility weight (24) and our relaxation algorithm. The fourth method (denoted by the P-S method) uses our binary compatibility weight (9) and the spectral technique followed by the greedy algorithm. Finally, we refer to our proposed method as the P-P method.

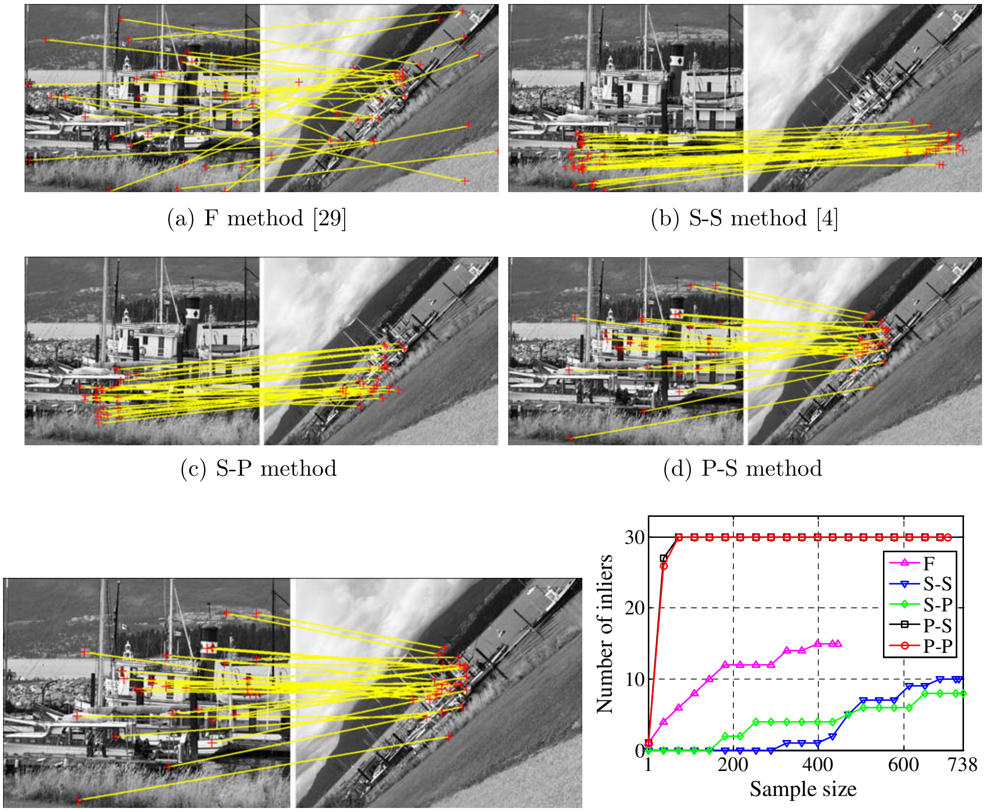
For the S-S and S-P methods that use the simple binary compatibility weight (24), the set $\mathscr{M}$ of candidate correspondences is reduced by deleting all correspondences $\mathbf{m}_b = (j, j')$ for which there is another correspondence $\mathbf{m}_a = (i, i')$ such that $i = j$, $i' = j'$, and $d_a < d_b$, because these correspondences are identical if we do not consider local feature transformations $\mathbf{H}_a$ and $\mathbf{H}_b$. Such identical correspondences result from multiple dominant image-gradient orientations used in invariant feature description [3]. For the P-S method, we use the original set $\mathscr{M}$.

The correspondences detected by the S-P and P-P methods, which use the proposed relaxation algorithm, are sorted in decreasing order of $\hat{p}_a^* \hat{q}_a^*$. Those detected by the S-S and P-S methods, which use the spectral technique [4], are sorted in decreasing order of p_a^* because p_a^* is the confidence of $\mathbf{m}_a$ in these methods. For all the image pairs in this paper, we use the set of parameter values $\tau_d = 0.5$, $N_{max} = 20000$, $p_0 = 0.5$, $T_{max} = 200$, $\tau_p = 99\%$, unless otherwise specified.

4.1. Planar and parallax-free scene results

In this subsection, we use image pairs of planar or parallax-free scenes. Some image pairs (Figs. 7–9) with viewpoint or scale changes were selected from the data set provided by Mikolajczyk et al. [8]. Ground-truth homographies between the images are provided in the data set, and we use them to evaluate the proposed method. For the other image pairs, handcrafted point-correspondences were used for estimating the ground-truth homographies.

When evaluating matching results using ground-truth models, the disparity between the centroid of an affine region and the homography-transformed centroid of the correctly corresponding affine region tends to be large if the two regions are large, because the affine assumption breaks more easily for large regions. For this reason, Mikolajczyk et al. [8] proposed an evaluation scheme based on region overlap, which encourages correspondences between



(a) F method [29]

(b) S-S method [4]

(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 9. A scene with a large-scale change [8]. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).

large regions to be counted as inliers. Here, we use a simpler scheme that also encourages correspondences between large regions.

Following [27,8], we fit an ellipse to each region of a detected correspondence. For example, the ellipse equation for a region $\mathscr{R}_i$ is:

$$(\mathbf{x}-\mathbf{x}_i)^T\Sigma_i^{-1}(\mathbf{x}-\mathbf{x}_i) = \kappa^2, \tag{27}$$

where $\mathbf{x}_i$ is the centroid of $\mathscr{R}_i$, Σ_i is the covariance matrix of the interior pixels of $\mathscr{R}_i$, and κ is a constant. A correspondence $\mathbf{m}_a = (\mathbf{x}_i, \mathbf{x}'_{i'})$ that satisfies both of the following conditions is counted as an inlier (a correct correspondence):

$$\begin{aligned}&(\mathbf{H}_{true}^{-1}(\mathbf{x}'_{i'})-\mathbf{x}_i)^T\Sigma_i^{-1}(\mathbf{H}_{true}^{-1}(\mathbf{x}'_{i'})-\mathbf{x}_i) < \kappa^2,\\&(\mathbf{H}_{true}(\mathbf{x}_i)-\mathbf{x}'_{i'})^T\Sigma_{i'}'^{-1}(\mathbf{H}_{true}(\mathbf{x}_i)-\mathbf{x}'_{i'}) < \kappa^2,\end{aligned} \tag{28}$$

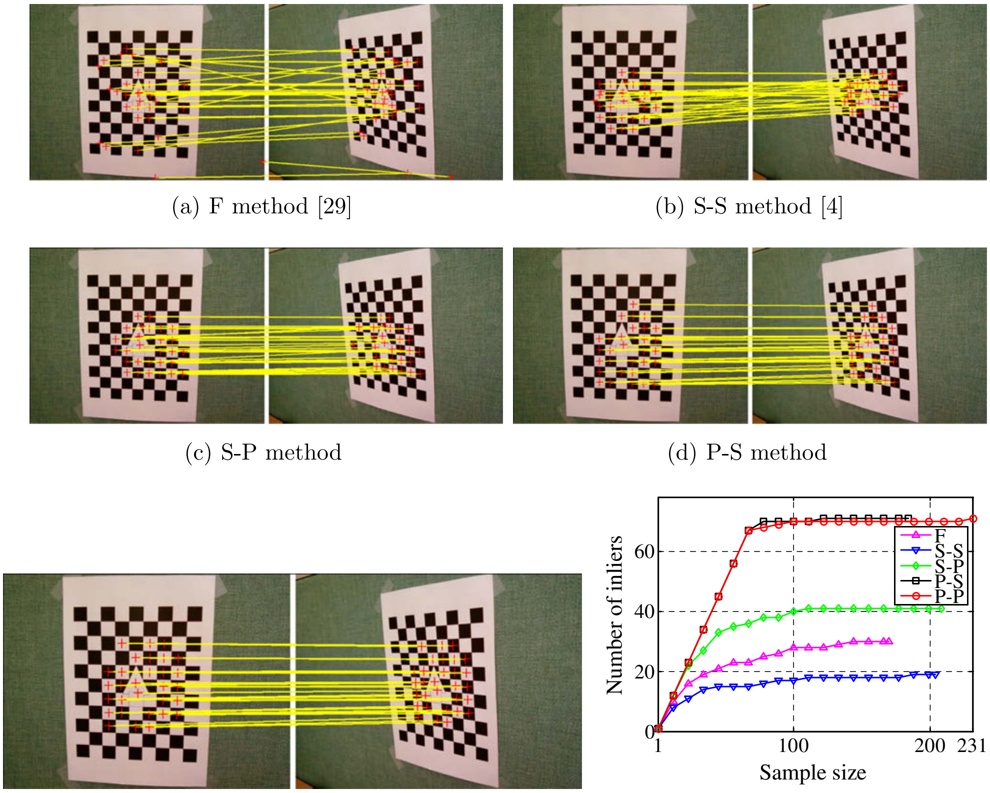
where $\mathbf{H}_{true}$ is the ground-truth homography. If the conditions are satisfied, each transformed feature point lies within the fitted ellipse of the corresponding feature point. This situation is illustrated in Fig. 6a. If κ is large, correct correspondences are not missed, at the expense of letting through some false correspondences. If κ is small, we can harvest correct correspondences with high accuracy at the expense of losing less-accurate correct correspondences. We set $\kappa = 1$ for the experiments, which is a small value that strongly discourages incorrect correspondences from being counted as inliers.

Figs. 7–12 show the image pairs of planar and parallax-free scenes. The image pair in Fig. 7 contains distinctive textures, with the viewpoint change being small. The image pairs in Figs. 8 and 10 contain a viewpoint-changed scene with a relatively small scale change. The image pairs in Figs. 9 and 11 contain a scale-changed scene with a relatively small viewpoint change. The image pairs in Figs. 10 and 11 contain a chessboard pattern, where the only distinctive part is the house-like pattern in the middle; from this pattern, human referees can find correct correspondences. We use these image pairs as simulative examples with high ambiguity. The image pair in Fig. 12 is especially hard to match because of the reflected tree pattern on the windows, in addition to the repeated textures. We can expect the methods using the proposed objective function (namely P-S and P-P) to work better than the other methods because our assumption holds best for the image pairs of planar and parallax-free scenes.

As expected, the P-P and P-S methods give the best results for most of the image pairs, which means that the proposed objective function is suitable for finding correct correspondences between the images robustly, with respect to viewpoint or scale changes. For the easiest image pair in Fig. 7, which has distinctive textures and a small viewpoint change, all the methods give good results. The F method gives good results for the image pairs with distinctive textures (Figs. 7–9), but does not give good results for the image pairs with highly repeated textures (Figs. 10–12). The S-P method gives meaningful results for the image pairs with small scale changes (Figs. 7, 8, and 10), but the method fails for the other image pairs. The S-S method gives the worst results for all the image pairs. It is interesting to note that the P-P method works better than the P-S method for the challenging image pair in Fig. 12. Indeed, our algorithm gives higher values of $f(\hat{\mathscr{P}}^*)$ than the spectral technique for this image pair, as will be revealed at the end of this section.

4.2. 3D scene results

In this subsection, we assess the proposed method for image pairs of static 3D scenes. The ground-truth fundamental matrices



(a) F method [29]

(b) S-S method [4]

(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 10. A viewpoint-changed planar scene with repeated textures. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).

for the image pairs were estimated from handcrafted point correspondences.

Moreels and Perona [29] have developed an evaluation scheme for 3D objects. The necessity for an additional image, however, discourages us from using this scheme. Forssén and Lowe [27] have recently developed an evaluation scheme that does not require an additional image, at the expense of letting through some false correspondences. The scheme uses epipolar tangents for the fitted ellipses. The tangents, however, do not exist when the epipole lies within the fitted ellipse. Fortunately, this kind of degeneracy did not affect their evaluation because the epipole was outside the images in their experiments.

Here, we use a simpler scheme. We regard a correspondence as correct if the epipolar line induced by a feature point intersects the fitted ellipse of the corresponding feature point (Fig. 6(b)). This scheme is not affected by the position of the epipole, although false correspondences are more likely to be let through.

Figs. 13–15 show the image pairs of static 3D scenes. Our assumption holds well for the image pair in Fig. 13 because the scene is mostly composed of planar objects. For the other image pairs, we cannot easily say that our assumption holds better, because the object scale is well preserved and the viewpoint change is not so severe.

The P-P and P-S methods give the best results for the image pairs. The F method gives good results for the image pairs in Figs. 13 and 14 with their distinctive textures, but does not give good results for the image pair in Fig. 15 with its repeated leaves. For the plant image pairs in Figs. 14 and 15 without severe scale changes, the S-P method gives good results, but it is less effective when the viewpoint is changed significantly (Fig. 13). The S-S method gives similar results to those of the S-P method, although the S-P method outperforms the S-S method in terms of the number of correct correspondences in the top k best correspondences.

4.3. Nonrigid objects

In this subsection, we assess the proposed method for image pairs of nonrigid objects. Ground-truth parametric models cannot be used in the evaluation because of the nonrigid deformation, so we manually count correct correspondences for the image pairs.

Figs. 16–18 show the image pairs of nonrigid objects. Figs. 16 and 17 show simulative examples with high ambiguity. The chessboard pattern in Fig. 16 is deformed without being occluded, but it is self-occluded in Fig. 17. For both image pairs, our assumption holds, so that we can expect the methods using the proposed objective function (P-S and P-P) to give good results in spite of the repeated textures. The image pair in Fig. 18 shows a person and a fountain in both images. Although the print on the T-shirt contains some repeated textures and is nonrigidly deformed, the surface is smooth and our assumption still holds. However, the continuity breaks at the shoulder, and the assumption does not hold between the arm and the torso. In fact, no algorithm in this paper finds correct correspondences on the arm and the face: the correspondences have neither the smallest local-descriptor distances nor geometric measurements consistent with neighboring correspondences.

The P-P method consistently gives the best results, for all the image pairs. Although the P-S method gives a good result for the image pair in Fig. 16, it gives poor results for the image pairs in Figs. 17 and 18, because it finds very strongly structured false correspondences for the image pair in Fig. 17, and finds structured false correspondences on the T-shirt in Fig. 18. In fact, the P-P method finds higher values of $f(\hat{\mathscr{P}}^{*})$ than the P-S method for the

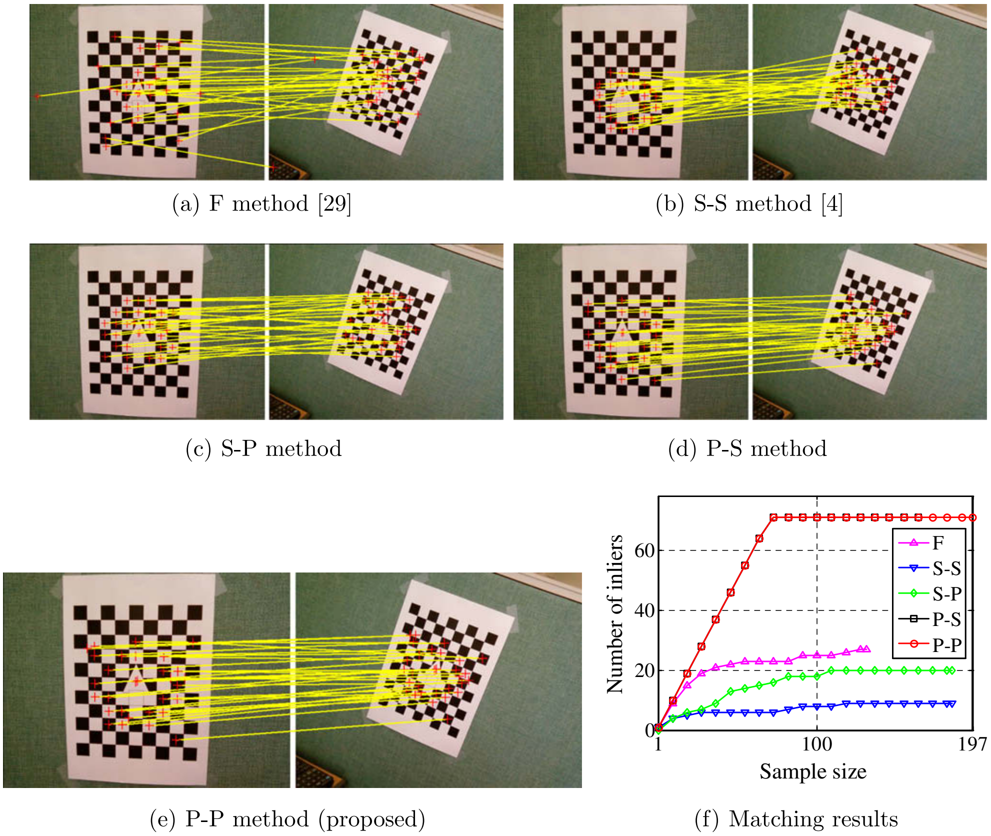

(a) F method [29]
(b) S-S method [4]
(c) S-P method
(d) P-S method
(e) P-P method (proposed)
(f) Matching results

Fig. 11. A scale-changed planar scene with repeated textures. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).

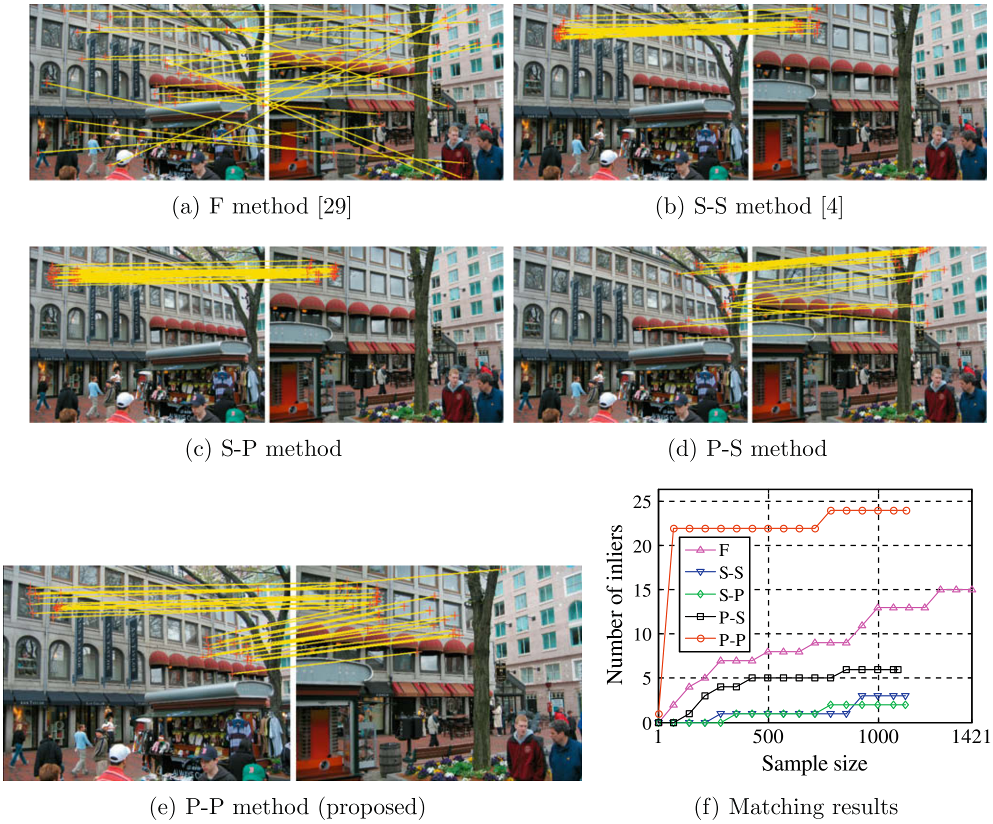
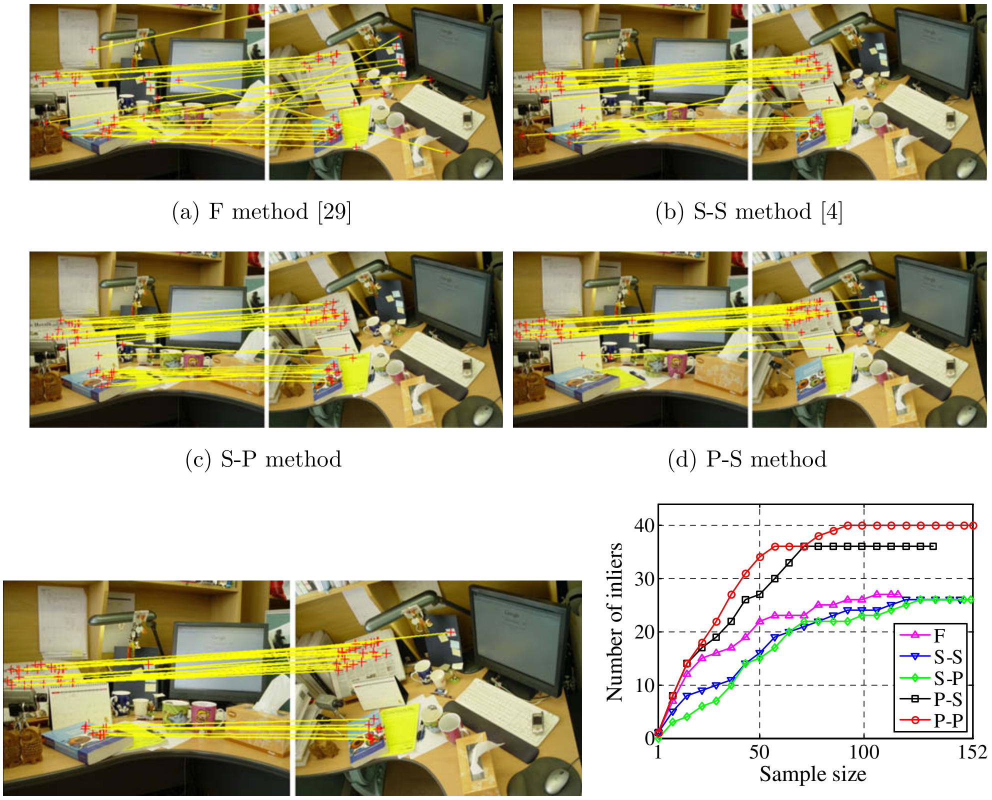

(a) F method [29]
(b) S-S method [4]
(c) S-P method
(d) P-S method
(e) P-P method (proposed)
(f) Matching results

Fig. 12. A viewpoint-changed near-planar scene with repeated textures [14]. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).



(a) F method [29]

(b) S-S method [4]

(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 13. A 3D scene with a viewpoint change. (a–e) Only the top 30 matches are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).

image pairs in Figs. 17 and 18. All the other methods perform poorly for all the image pairs of nonrigid objects with high ambiguity.

We evaluated the values of $f(\hat{\mathscr{P}}^*)$ to determine the effect of the optimization algorithms. Fig. 19a and b shows the values. As mentioned above, the P-P method gives higher values of $f(\hat{\mathscr{P}}^*)$ than the P-S method for the image pairs in Figs. 12, 17, and 18, resulting in the better performance of the P-P method. For the image pair in Fig. 17, the difference of the values is not so large (less than 20%) because the P-S method finds very strongly structured false correspondences. However, the difference is large (greater than 40%) for the practical image pairs for which the P-S method does not find very strongly structured correspondences (Figs. 12 and 18). It can also be observed that the P-P method consistently gives higher values of $f(\hat{\mathscr{P}}^*)$ than the P-S method, for the other image pairs.

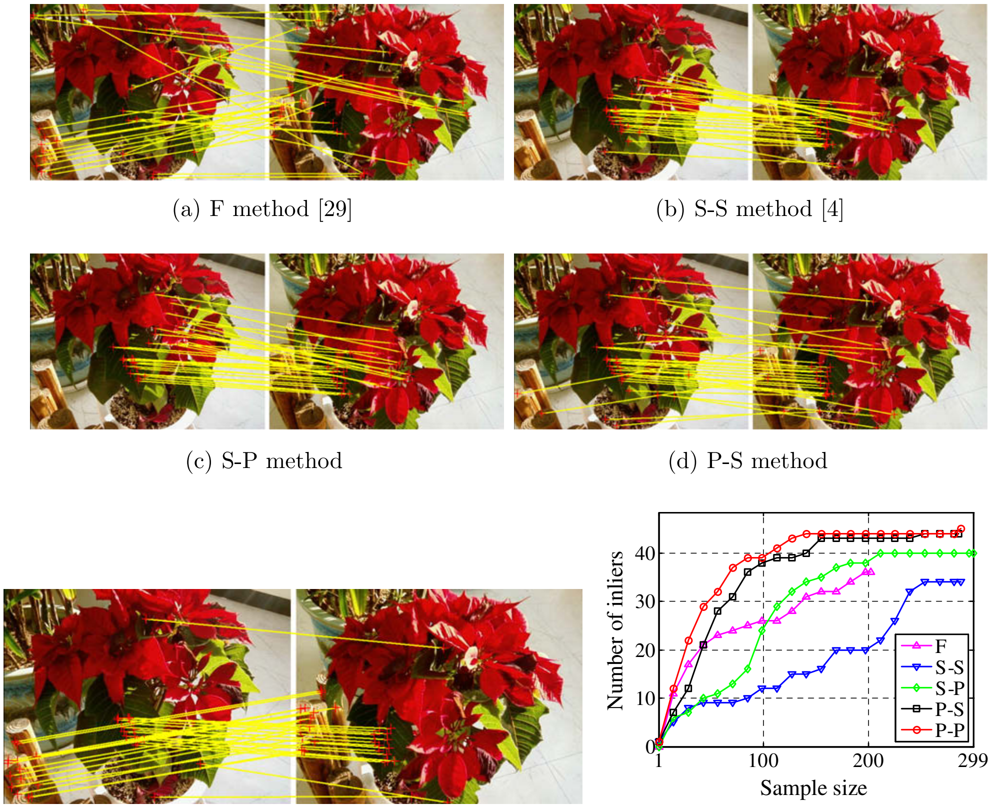
For those cases where the S-S and S-P methods are used (Fig. 19a), we cannot say that the S-P method finds higher values of $f(\hat{\mathscr{P}}^*)$, and we cannot see any correlation between the values and the number of correct correspondences either. It might be that they do not have a strong correlation because the assumption does not reflect the actual changes between the images. In spite of using an objective function that does not model the viewpoint or scale changes between the images, the S-P method outperforms the S-S method for most of the image pairs, in terms of the number of correct correspondences. We consider this better performance of our algorithm to be caused by the third property (P3) of our algorithm.

We evaluated the percentage of p_a^* greater than 0.99 ($p_a^* \approx 1$) for every $a \in \mathscr{S}^*$, to observe how well the proposed relaxation algorithm finds a feasible solution. Fig. 19c shows the results. Although it is hard to prove theoretically that the solution found by using our algorithm converges to a feasible solution, the graph shows that the percentage of converged confidences grows with increasing T: more than 95% of the confidences converge to 1 at $T = 1000$ on average. Our choice of $T_{max} = 200$ may not be suitable for finding more converged solutions. However, we did not increase T_{max}, for reasons of computational complexity.

4.4. Recognizing objects in cluttered environments

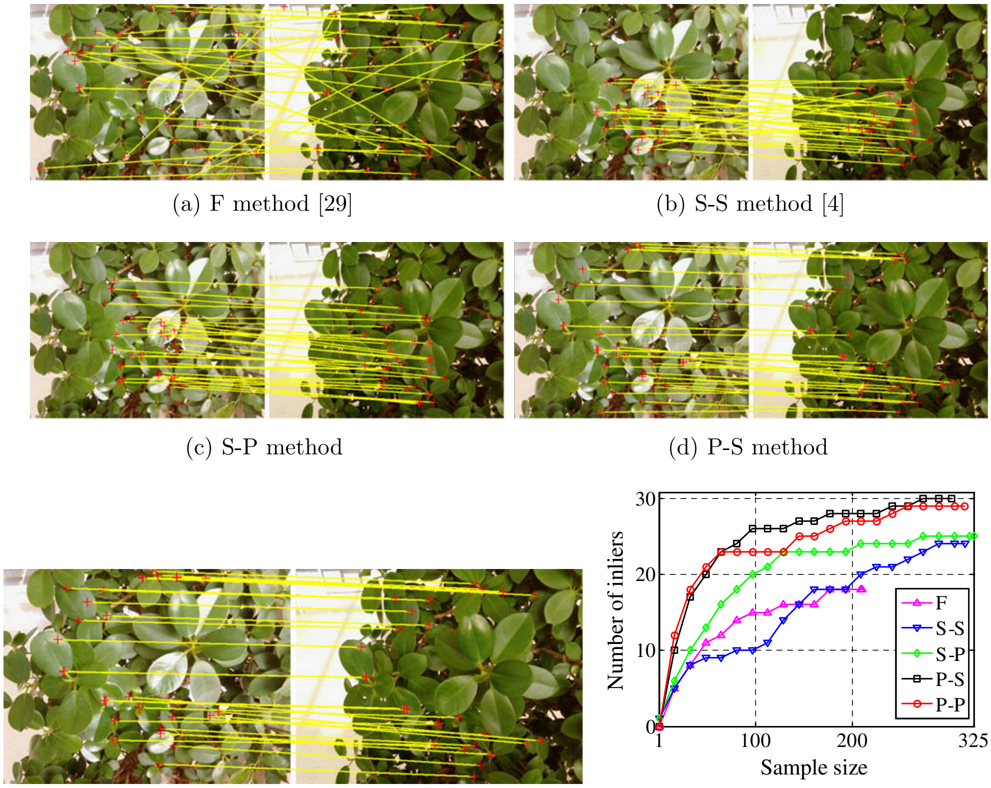
Background clutter should be considered when developing a robust recognition system, because false correspondences are frequently detected between the object in a data image and the cluttered background in a query image. The proposed method encourages geometrically consistent correspondences to be detected. This property motivates us to apply the proposed method to object recognition, under the assumption that false correspondences are not geometrically consistent.

The recognition scenario is as follows. A query image is given as an input, containing an object in a cluttered background. The proposed method is used to match the query image to every data image that contains an object in a simple black background (see Fig. 20). The detected correspondences are clustered to form Groups of Aggregated Matches (GAMs) [21] by grouping together correspondences with a small pairwise transformation error ($e_{ab} < 3\sigma$). The value of $f(\hat{\mathscr{P}}^*)$ is computed for the correspondences in the largest GAM between every query-and-data image pair, and is then divided by the number of feature points in the data image, to discourage an incorrect data image with a large number of feature points from getting a larger value of $f(\hat{\mathscr{P}}^*)$ than a correct data image with a small number of feature points. We finally classify the cluttered query image as an instance of



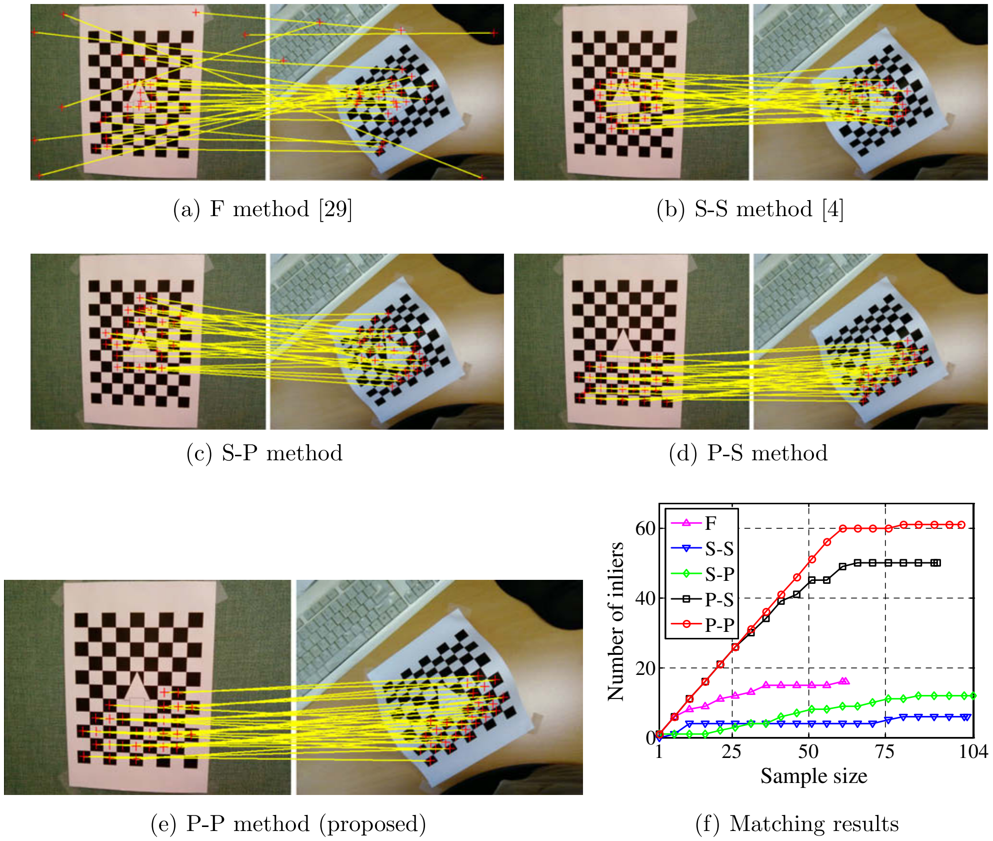
(a) F method [29]
(b) S-S method [4]
(c) S-P method
(d) P-S method
(e) P-P method (proposed)
(f) Matching results

Fig. 14. A plant scene. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct ($k =$ sample size).



(a) F method [29]
(b) S-S method [4]
(c) S-P method
(d) P-S method
(e) P-P method (proposed)
(f) Matching results

Fig. 15. A plant scene with repeated leaves. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct ($k =$ sample size).



(a) F method [29]

(b) S-S method [4]

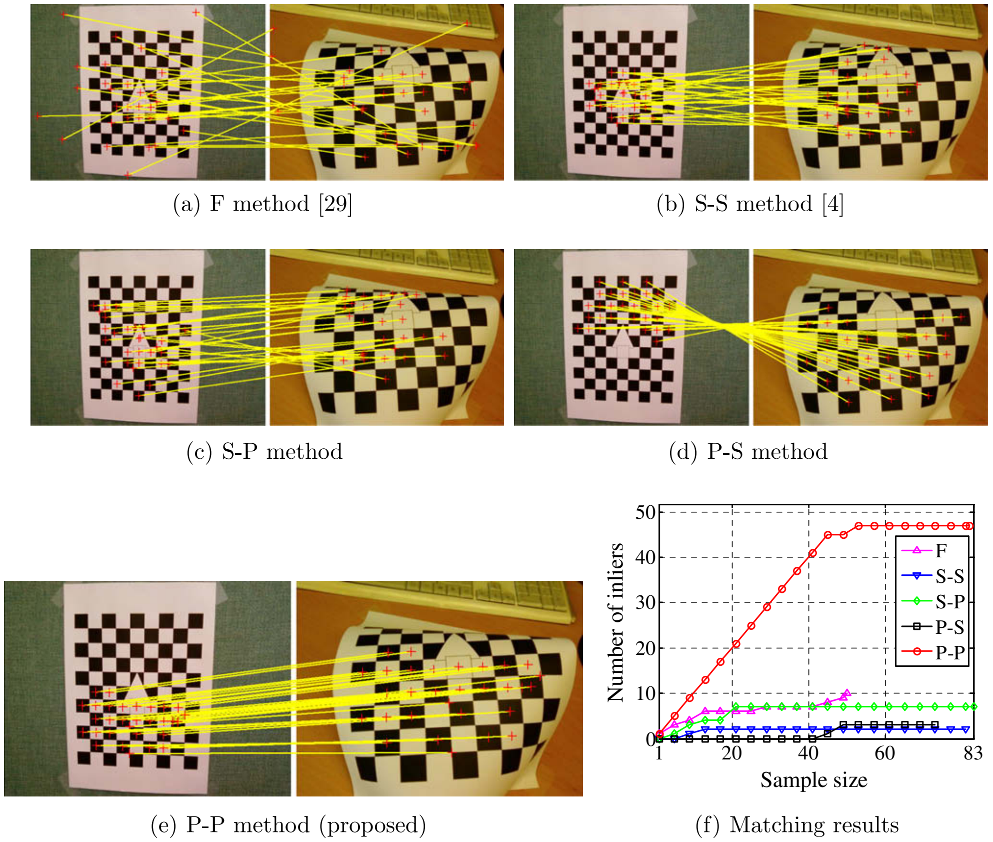
(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 16. A scale-changed nonrigid object with repeated textures. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k correct correspondences are correct (k = sample size).



(a) F method [29]

(b) S-S method [4]

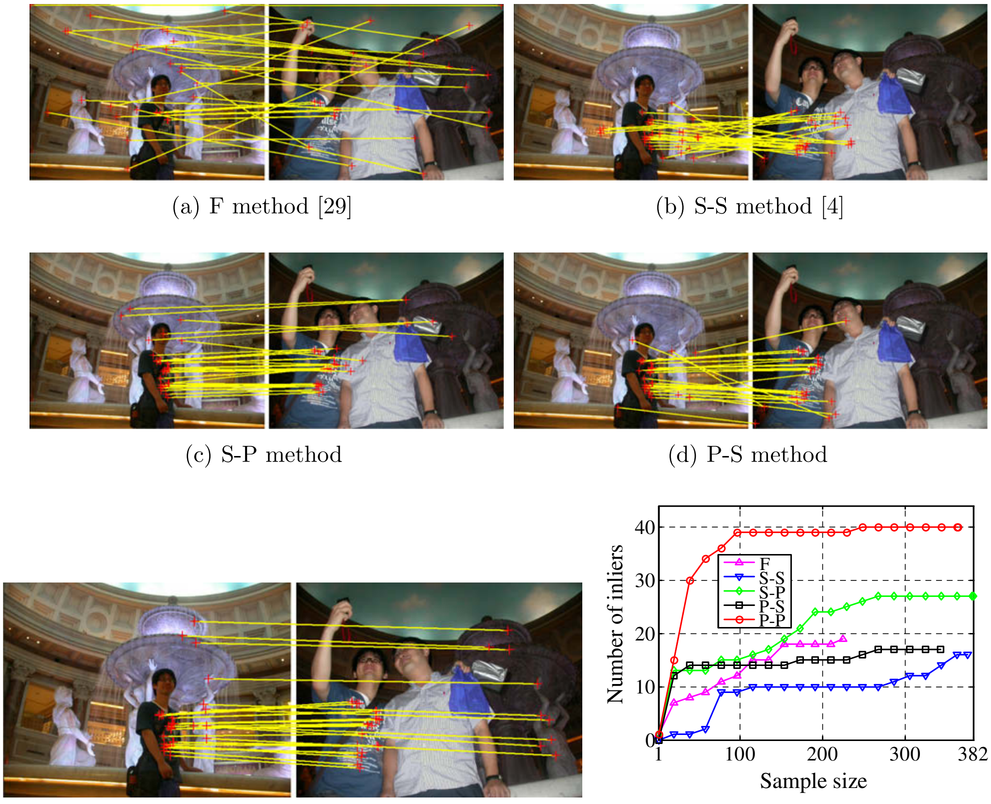
(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 17. A partially occluded nonrigid object with repeated textures. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k best correspondences are correct (k = sample size).



(a) F method [29]

(b) S-S method [4]

(c) S-P method

(d) P-S method

(e) P-P method (proposed)

(f) Matching results

Fig. 18. People in front of a fountain. (a–e) Only the top 30 correspondences are displayed, to aid visibility. (f) The graph displays how many of the top k correct correspondences are correct (k = sample size).

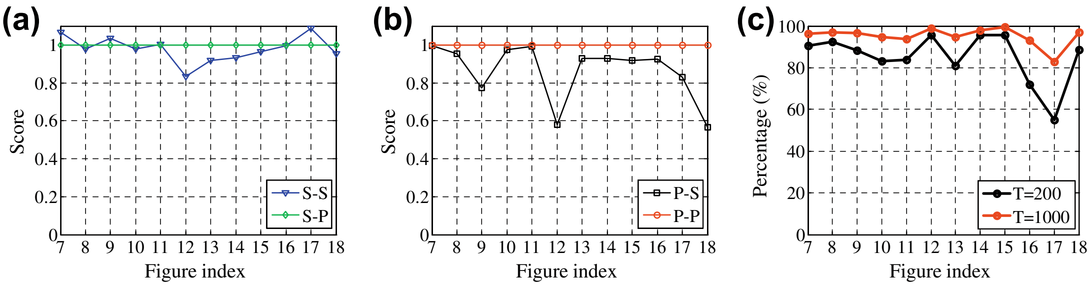


Fig. 19. (a,b) The values of $f(\hat{\mathscr{P}}^*)$ found by the proposed relaxation algorithm and the spectral technique [4]. In both graphs the values of $f(\hat{\mathscr{P}}^*)$ were divided by the value of $f(\hat{\mathscr{P}}^*)$ found by the proposed relaxation algorithm for easy comparison. (c) The graph displays the percentages of p_a^* greater than 0.99 for every $a \in \mathscr{S}^*$. The P-P method was used for the experiment. The percentage grows with increasing T.

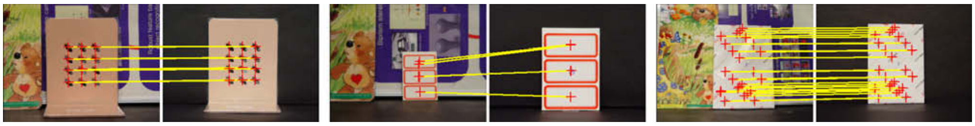


Fig. 20. Examples of query-and-data image pairs from the KAIST-104 DB [30]. The inliers of the largest GAM are displayed.

the data image that maximizes the normalized value of $f(\hat{\mathscr{P}}^*)$. For computing the recognition score $f(\hat{\mathscr{P}}^*)$ under fair conditions, we use a fixed parameter value ($\sigma = 10$) for clustering and score computation, but not for the matching.

For the experiment, we used images from the KAIST-104 DB [30]. Fig. 20 shows some examples. Because of the cluttered background, Kim et al.'s success rate was 71.15%, classifying 74 images correctly for 104 query images [30]. We could classify 85 images correctly, which is about a 10% improvement on state-of-the-art methods [30]. Fig. 21 shows a confusion matrix for the recognition results. The strong diagonal elements show that the proposed recognition score is effective.

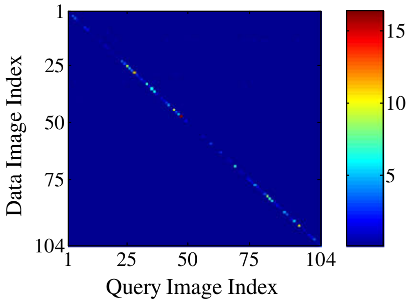


Fig. 21. Proposed recognition score. The graph shows the normalized values of $f(\mathscr{P}^*)$ for every query-and-data image pair used in the recognition experiment.

5. Conclusions and future work

In this paper, we have proposed a robust feature point matching method that benefits from a novel objective function and a relaxation algorithm. The proposed method showed consistently good matching results for challenging image pairs with viewpoint changes, scale changes, and nonrigid deformations, in the presence of repeated textures. The objective function was shown to be suitable for matching image pairs with (or without) significant viewpoint or scale changes, unlike previous functions that rely on the assumption that the distance between neighboring feature points is preserved across images. The proposed relaxation algorithm found good solutions that not only maximized the objective function but also maximized the number of correct correspondences. The proposed relaxation algorithm also showed better performance than state-of-the-art algorithms such as the spectral technique [4], for widely separated practical image pairs.

The major drawback of the proposed method is its memory requirements for computing and saving the pairwise links, which is approximately $O(KN)$, where K can be as large as N in the worst case. We are currently investigating an effective method for reducing it. The proposed relaxation algorithm can be applied to maximizing not only quadratic functions but also higher-order functions, provided they are differentiable. This advantage will lead us to consider higher-order relationships among three or more correspondences, and there may be other kinds of pairwise relationship that have yet to be discovered. We are investigating such pairwise and higher-order relationships for robust feature point matching.

Acknowledgement

Thank you to the reviewers for helpful suggestions. This research was supported by the Defence Acquisition Program Administration and Agency for Defense Development, Korea, through the Image Information Research Center at KAIST under the contract UD070007AD.

Appendix A

Here is a brief proof for the equivalence between the condition (15) and one-to-one correspondence constraints. If the condition (15) is satisfied, then $\sum_{b\in\mathscr{C}_a-\{a\}} p_b = 0$ for every a such that $p_a = 1$. In this case, $p_b = 0$ for every $b \in \mathscr{C}_a - \{a\}$ because $p_b \geqslant 0$ for every $b \in \{1,\ldots,N\}$, meaning that every $\mathbf{m}_b$ in conflict with $\mathbf{m}_a$ does not belong to the solution set $\mathscr{M}^* = \{\mathbf{m}_a : p_a = 1\}$. This proves that the condition (15) is a sufficient condition for one-to-one correspondence constraints. If one-to-one correspondence constraints are satisfied, then $p_b = 0$ for every pair of a and b such that $p_a = 1$ and $b \in \mathscr{C}_a - \{a\}$. It follows that $\sum_{b\in\mathscr{C}_a-\{a\}} p_b = 0$ because $p_b = 0$ for every $b \in \mathscr{C}_a - \{a\}$. Consequently, $s_a = p_a + \sum_{b\in\mathscr{C}_a-\{a\}} p_b = 1$ for every a such that $p_a = 1$. This proves that the condition (15) is also a necessary condition.

References

[1] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, second ed., Cambridge university, Cambridge, 2003.

[2] D. Martinec, T. Pajdla, 3D reconstruction by gluing pair-wise Eclidean reconstructions, or "how to achieve a good reconstruction from bad images, in: Interantional Symposium on 3D Data Processing, Visualization, and Transmission, 2006.

[3] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[4] M. Leordeanu, M. Hebert, A spectral technique for correspondence problems using pairwise constraints, in: International Conference on Computer Vision, 2005.

[5] A.C. Berg, T.L. Berg, J. Malik, Shape matching and object recognition using low distortion correspondences, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[6] T. Tuytelaars, L.V. Gool, Content-based image retrieval based on local affinely invariant regions, in: International Conference on Visual Information Systems, 1999, pp. 493–500.

[7] J. Matas, O. Chum, U. Martin, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: British Machine Vision Conference, 2002.

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L.V. Gool, A comparison of affine region detectors, International Journal of Computer Vision 65 (1–2) (2005) 43–72.

[9] P.H.S. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Computer Vision and Image Understanding 78 (2000) 138–156.

[10] P.H.S. Torr, C. Davidson, Impsac: Synthesis of importance sampling and random sample consensus, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (3) (2003) 354–364.

[11] O. Chum, J. Matas, Matching with prosac—progressive sample consensus, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[12] Z. Zhang, R. Deriche, O. Faugeras, Q.T. Luong, A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, Artificial Intelligence 78 (1–2) (1995) 87–119.

[13] Y. Zheng, D. Doermann, Robust point matching for nonrigid shapes by preserving local neighborhood structures, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 643–649.

[14] R. Szeliski, Iccv2005 computer vision contest, http://research.microsoft.com/iccv2005/contest/ (2005).

[15] M.A. Fischler, R.C. Bolles, Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[16] A. Rosenfield, R.A. Hummel, S.W. Zucker, Scene labeling by relaxation operations, IEEE Transactions on System, Man, and Cybernetics 6 (6) (1976) 420–433.

[17] R.A. Hummel, S.W. Zucker, On the foundations of relaxation labeling processes, IEEE Transactions on Pattern Analysis and Machine Intelligence 5 (3) (1983) 267–287.

[18] S.Z. Li, Markov Random Field Modeling in Image Analysis, Springer-Verlag, 2001.

[19] O. Choi, H.W. Kim, I.S. Kweon, Simultaneous plane extraction and 2d homography estimation using local feature transformations, in: Asian Conference on Computer Vision, 2007.

[20] J. Kannala, S.S. Brandt, Quasi-dense wide baseline matching using match propagation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

[21] V. Ferrari, T. Tuytelaars, L.V. Gool, Simultaneous object recognition and segmentation from single or multiple model views, International Journal of Computer Vision 67 (2) (2006) 159–188.

[22] A. Vedaldi, S. Soatto, Local features, all grown up, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[23] M. Lhuillier, L. Quan, A quasi-dense approach to surface reconstruction from uncalibrated images, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 418–433.

[24] T. Cour, P. Srinivasan, J. Shi, Balanced graph matching, in: Neural Information Processing Systems, 2006.

[25] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Computer Vision and Image Understanding 89 (2–3) (2003) 114–141.

[26] Y. Zheng, D. Doermann, Robust point matching for non-rigid shapes: a relaxation labeling based approach, Tech. Rep. LAMP-TR-117, University of Maryland, 2004.

[27] P.-E. Forssén, D.G. Lowe, Shape descriptors for maximally stable extremal regions, in: International Conference on Computer Vision, 2007.

[28] R. Lehoucq, K. Maschhoff, D. Sorensen, C. Yang, Arpack software, http://www.caam.rice.edu/software/arpack/, 1996.

[29] P. Moreels, P. Perona, Evaluation of features detectors and descriptors based on 3d objects, International Journal of Computer Vision 73 (3) (2007) 263–284.

[30] S. Kim, K.J. Yoon, I.S. Kweon, Object recognition using a generalized robust invariant feature and gestalt's law of proximity and similarity, Pattern Recognition 41 (2) (2008) 726–741.