Graph model-based feature point matching

Wang, Chen

2013

Wang, C. (2013). Graph model-based feature point matching. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54912>

<https://doi.org/10.32657/10356/54912>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

GRAPH MODEL-BASED FEATURE POINT MATCHING

WANG CHEN

WANG CHEN

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

2013

2013

GRAPH MODEL-BASED FEATURE POINT MATCHING

WANG CHEN

WANG CHEN

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2013

Acknowledgments

Pursuing my PhD degree might be the most interesting but challenging journey that I've experienced in my first thirty years. The moments of joy and tears in this journey will remain dear to me. It is my pleasure to spend four years in the School of Electrical and Electronic Engineering, Nanyang Technological University. Here I met so many great people that gave my research work constructive suggestions, unflagging encouragement, and successive support. Therefore, I would like to give them my sincere thanks.

First and foremost I would like to express my upmost appreciation to my supervisor, Professor Kai-Kuang Ma. During my doctoral program, he gave me much freedom in doing research so that I was allowed to select the research topic in which I was interested, to brainstorm new ideas and to build up the prototype for the experiments and demonstrations. He respected any idea from me and spent much of his valuable time on discussing with me under what circumstance the proposed method could work well or totally fail and why. Based on these discussions, he provided me with his insightful guidance and constructive suggestions. More importantly, he also patiently trained me how to effectively communicate my research ideas to the other researchers, which is proven to be a precious asset for my future career and life.

I would like to extend my gratitude to our project advisors from DSO National Laboratories, Dr. Ng Teck Khim and Mr. Chia Kar Wee for their constant support,

guidance and helpful suggestions. Through the research discussions with them, more technical challenges in the practical situations were discovered and properly treated.

I would also like to give my sincerest thanks to the members of Temasek Lab at NTU, Dr. Tian Jing, Dr. Wei Zhe, Mr. Zeng Huanqiang and Ms. Ong Huiting. Without their invaluable advices and friendly assistances on my research work, I could never accomplish this thesis. Their friendships mean more than what I could ever express. Meanwhile my friends in China and in Singapore also deserve my special thanks, as you were the sources of laughter, joy and support.

Last but not least, I wish to thank all my family members for their unselfish support up to now. It is your love that motivated and inspired me all the time. I hope that this work could make you proud of me.

Contents

Acknowledgments	i
List of Figures	vi
List of Tables	xii
Summary	xiii
1 Introduction	1
1.1 Feature Point Matching	1
1.2 Motivations and Objective	3
1.3 Main Contributions and Highlights	5
1.4 Organization	8
2 Background	9
2.1 Feature Point Detection	9
2.2 Local Feature Descriptor Computation	13
3 Bipartite Graph Model-Based Mismatch Removal for Wide-baseline Image Matching	16
3.1 Literature Review	17
3.2 Proposed Region-to-region Matching Method	20

3.2.1	SIFT-based Initial Feature Point Matching Pairs	20
3.2.2	Proposed Algorithm for Merging Image Segments	21
3.3	Proposed Region-to-region Similarity Measurement Metric	25
3.3.1	Photometric Consistency	26
3.3.2	Neighborhood Consistency	27
3.3.3	Point-to-point Matching Pair Density	30
3.4	Generation of Coherent Region Pairs via the Bipartite Graph	31
3.4.1	Model Representation	31
3.4.2	Bipartite Graph Matching	32
3.4.3	Proposed Process for Selecting the Best Matching Pairs	33
3.5	Experimental Results	34
3.6	Summary	39
4	Common Visual Pattern Discovery via Directed Graph Model	41
4.1	Overview of Common Visual Pattern Discovery	42
4.2	Background	44
4.3	Undirected Graph Model: Conventional Approach	46
4.3.1	Framework Overview	46
4.3.2	Undirected Graph Model Approach	47
4.3.3	Graph Matching Method: An Equivalent Approach	50
4.4	Directed Graph Model: Proposed Approach	52
4.4.1	Model Setup and Motivation	52
4.4.2	The Gaussian Link-Weight Mapping Function	56
4.4.3	Matrix Representation of Directed Graph Model	59
4.5	Strongly-Associated Subgraph Identification	60
4.5.1	Problem Formulation for Identifying Strongly-Associated Subgraph	60

4.5.2	Proposed Numerical Method: An Evolutionary Approach . .	62
4.5.3	Topological Splitting	64
4.6	Experimental Results and Discussions	67
4.6.1	Objective Evaluation: Experimental Setup and the Ground Truth	67
4.6.2	Objective Evaluation: Experimental Results	70
4.6.3	Performance Evaluation with Natural Images	74
4.6.4	Near-duplicate Image Retrieval	82
4.7	Summary	84
5	Feature Histogram Equalization: A Feature Contrast Enhance- ment Technique	86
5.1	Background and Literature Review	87
5.2	What is Feature Contrast?	91
5.3	Feature Histogram	94
5.4	Feature Histogram Equalization (FHE)	98
5.5	Experimental Results	102
5.5.1	Applying the Proposed FHE to Feature-Point Correspondence Application	102
5.5.2	Simulation Setup and the Ground Truth Generation	104
5.5.3	Simulation Results and Discussions	106
5.6	Summary	110
6	Conclusion and Future Works	112
6.1	Conclusion	112
6.2	Future Works	116
6.2.1	Fast Feature Point Matching	116
6.2.2	Feature Enhancement	116

List of Figures

1.1	A typical framework of the feature point matching system.	2
1.2	The proposed framework of feature point matching system.	5
2.1	An illustration of the blob structure that commonly encountered in the images: three black vertical lines stand for the three imaging planes placed at the different distances to the object in the scene; the three red vertical lines stand for the blob structures.	10
3.1	The flowchart of our proposed bipartite graph model-based mismatch removal method.	21
3.2	An example of one initial region matching pair (blue regions). A pair of yellow crosses denotes an established SIFT-based point-to-point matching pair.	22
3.3	An example of one-to-many initial region matching pairs. Each pair of yellow crosses denotes an established SIFT-based point-to-point matching pair, and the blue segmented region in the left image actually corresponds to the blue, red, gray and green image segments in the right image.	23

3.4	A graphical illustration of the proposed segment merging operation: (a) established point-to-point matching pairs based on the SIFT feature descriptors and the image segments generated by [49], (b) <i>initial</i> region matching pairs, (c) the final <i>candidate</i> region matching pairs.	24
3.5	An example of one-to-many region correspondence case after applying the proposed segment merging operation.	24
3.6	An example of an one-to-one coherent region pair by applying our proposed mismatch removal method.	26
3.7	A simple illustration showing why a checking of the neighborhood of each support region might be useful on the establishment of more reliable region-to-region correspondences. The smaller ellipses indicate the support regions centered at the respective two feature points, and the larger ellipses indicate the <i>extended</i> support regions which are obtained by extending the area coverage of the corresponding support regions by a factor of two. Note that all the support regions and the extended support regions are normalized into unit circles, followed by cutting a square image patch with the size of 100×100 for demonstration as shown in (c)-(f).	28
3.8	A demonstration of six wide-baseline image pairs downloaded from [20]. Each image pair has two images containing the same scene but captured under a large viewpoint variation.	35
3.9	An illustration of point-to-point matching pairs by applying the nearest neighbor matching criterion and our proposed mismatch removal method on image pair <i>mex</i> shown in (a)-(b) and image pair <i>church</i> shown in (c)-(d).	37

4.1	A demonstration on how difficult to determine whether there is any visual object commonly presented in both images and, if any, how many of them?	42
4.2	Examples of two graph models: (a) conventional <i>undirected</i> graph model, in which only one link weight is associated with each link (without direction). Here, all the link weight values are presumably computed based on Eq. (4.2); (b) <i>directed</i> graph model with <i>two</i> link weights on each link (with the pointing direction as indicated by the arrow). Here, the two link weight values are computed by applying Eqs. (4.3) and (4.4) developed in this chapter, respectively.	54
4.3	A demonstration of a limitation on discriminating two visual objects (i.e., compact disks) when they are undergoing the same scale change. (a) Two visual objects having the same scale-change factor $\alpha = 0.625$; (b) The obtained common visual pattern discovery result by using our proposed digraph approach and being viewed as <i>one</i> visual object. Note that the circles are the supported regions generated by the SIFT algorithm [17]; (c) <i>Two</i> visual objects are identified by further applying our developed topological splitting.	66
4.4	An illustration of two point sets A and B that are generated to simulate the SIFT feature points detected on the square-shaped visual object commonly presented in Image 1 and Image 2, respectively. Note that, the two point sets contain two categories of points: matched and mismatched. The former ones are represented by the red dots, while the latter ones are represented by the blue stars.	68
4.5	(Continue)	71

4.5	Performance comparison among Leordeanu <i>et al.</i> [12], Liu <i>et al.</i> [14], Cour <i>et al.</i> [63] and our proposed method. In all our experiments conducted, the number of correct correspondences is 15 (i.e., $P = 15$) as the ground truth. Multiple pairs of point sets A and B will be generated through simulation to incorporate various combinations of the geometric deformation and outlier that are often encountered in the natural images.	72
4.6	(Continue)	75
4.6	(Continue)	76
4.6	(Continue)	77
4.6	(Continue)	78
4.6	(Continue)	79
4.6	Common visual pattern discovery results based on six thoughtfully-chosen image pairs. In general, (a)-(f) tend to yield more <i>outliers</i> due to a severely cluttered background, while (g)-(l) contain more <i>geometric deformations</i> due to large view point variations.	80
4.7	Performance comparisons among Liu <i>et al.</i> 's method [12], Wu <i>et al.</i> 's method [90], and Zhu <i>et al.</i> 's method [91] on Columbia dataset for near-duplicate image retrieval.	83
5.1	A gray-scale image (534×800) and its histogram are shown in (a) and (b), respectively. After applying the <i>image histogram equalization</i> (IHE) process to (a), the equalized image and its histogram are shown in (c) and (d), respectively. Note that the IHE stretches the original image intensity values over its full dynamic range, from 0 to 255, for an 8-bit gray-scale image.	90

5.2	A demonstration of <i>low</i> feature contrast through the image pair “Wall.” A large number of repeated patterns (i.e., bricks) are presented in the image pair (a) and (b); this shows that the image content of “Wall” is inherited with strong <i>self-similarity</i> . To further appreciate this aspect, a set of zoom-in pictures of several local regions centered on various feature points, including the point “A” in (a), are shown in (c) and (d), respectively. Note that all the regions are re-sized to 85×85 for a better visualization.	92
5.3	A demonstration of <i>high</i> feature contrast through the image pair “Graf.” One can observe that the “Graf” is full of distinctive shapes and regions and therefore presents much less degree of self-similarity, compared with that shown in Figure 5.2. Similarly, a set of zoom-in pictures of several local regions centered on various feature points, including the point “B” in (a), are shown in (c) and (d), respectively. Note that all the regions are re-sized to 85×85 for a better visualization.	93
5.4	(Continue)	107
5.4	(Continue)	108

- 5.4 The *recall* (horizontal axis) versus *precision* (vertical axis) curves resulted by exploiting the initial feature point matching on the generated SIFT feature descriptors *with* and *without* incorporating the proposed FHE for various test cases: (a)-(c) viewpoint change, (d) zoom and rotation, (e) illumination change, (f)-(g) blurring, and (h) JPEG compression distortion. The test image pairs I_1 and I_6 are constantly used for testing the largest content variations in each scene category. Note that the range of precision in (a) is fairly small—in the range of $[0, 0.02]$ (i.e., showing nearly the same performance), while the remaining plots are constantly in the range of $[0, 1]$, showing appreciable or even significant precision improvement. 109

List of Tables

3.1	The evaluation of the SIFT-based point-to-point matching pairs established on the six wide-baseline image pairs downloaded from [20]. The correct matches are visually identified by us and used as the ground truth.	36
3.2	A performance comparison of two mismatch removal methods: the topological filter [10] versus our proposed bipartite graph model-based method. The precision is computed by dividing the number of correct matches by the total targeted number, 120. These six wide-baseline image pairs are downloaded from [20].	38

Summary

Feature point matching aims to automatically establish point-to-point correspondences between two images acquired from the same scene but through two different viewpoints. This matching process is essential to many image processing and computer vision tasks, such as image registration, object detection and tracking, to name a few. However, the well-known SIFT feature descriptors together with the use of the nearest neighbor matching criterion oftentimes lead to many incorrect point-to-point correspondences. This would be incurred especially when the two images undergo a large viewpoint variation and/or contain a severely cluttered background. To improve the performance of feature point matching process, three contributions are made in this thesis: 1) mismatch removal; 2) common visual pattern discovery; and 3) *feature histogram equalization* (FHE). The first two are graph model-based approaches, and the last one is a novel technique on the enhancement of feature descriptors. More details are provided as follows.

For certain image processing applications such as matching two images captured from disparate views, or the so-called wide-baseline image matching as demonstrated in this thesis, a sufficiently large number of point-to-point correspondences are needed in the first place. However, these corresponding pairs often contain mismatched ones. Consequently, the accuracy of the spatial transformation estimation between the two images under matching will be reduced to a certain degree. Therefore, how to identify and remove the mismatched corresponding pairs from

the established ones is the main goal of our first contribution. For that, a bipartite graph model is exploited to the segmented image regions to establish all possible region-to-region correspondences. After that *one-to-one* region correspondences, called the *coherent region pairs* (CRPs), can be further identified by using the Hungarian method and the proposed region-to-region similarity measurement metric. The established CRPs will be utilized as the reference information in order to identify and remove those mismatched point-to-point correspondences. Extensive experimental results have demonstrated that our proposed mismatch removal method could reduce incorrect SIFT-based point-to-point correspondences for the application of wide-baseline image matching that normally requires a large targeted number of matching pairs.

For the second contribution on common visual pattern discovery, the entire work still begins with the SIFT feature descriptors. However, the main objective is to yield the *correct* point-to-point matching pairs in the first place, while the actual number of such established matching pairs tends to be small. The key novelty of our approach lies in the use of the *directed graph* (or *digraph*) model that has two link weights on each link, rather than one link weight inherited in the conventional graph model. The principle of *pairwise spatial consistency* is first exploited to generate the initial link weight for each link. Based on this link weight value, two *relative* link weights are then generated by considering the relativeness of neighboring vertices at each vertex of the link. For that, the *n-ranking* process and a novel link weight enhancement technique are proposed. Consequently, the resulted *relative* link weights are more robust to combat various adverse scenarios such as large viewpoint variations and indiscriminative feature descriptors. Based on the relative link weights generated at each assumed scale change factor, the strongly-associated subgraph can be extracted from the digraph by applying the non-cooperative game theory for handling non-symmetric adjacency matrix issue.

All the vertices (i.e., point-to-point correspondences) belonging to the strongly-associated subgraph extracted from the diagraph established at an assumed scale change factor are collectively treated as one common visual pattern; hopefully, this set of vertices corresponds to one visual object. If this is not the case, our proposed topological splitting algorithm might be able to further discriminate them. Extensive experiments have been conducted on the simulated SIFT feature points for standing out the technical challenges, followed by performing evaluations on six thoughtfully chosen natural image pairs and Columbia dataset to demonstrate the efficacy and robustness of the proposed method.

For the third contribution, *feature contrast* is first introduced as a measurement of the degree of self-similarity contained in an image pair. Note that, for an image pair with strong self-similarity, the established SIFT feature descriptors would be quite similar to each other, therefore the feature contrast is low. The FHE is then proposed to equalize SIFT feature descriptors by independently modifying the vector-component values of feature descriptors at each vector dimension; consequently, the feature contrast is effectively enhanced for the purpose of better discrimination of the feature descriptors. Extensive simulation results have clearly shown that our proposed FHE method could effectively improve the precision of SIFT-based point-to-point correspondences, especially to those image pairs containing a large amount of self-similar regions. As expected, if the image pairs contain only a few or even no self-similar regions, the performance gain would be marginal in this case.

Chapter 1

Introduction

1.1 Feature Point Matching

Given two sets of points, the process of identifying which point in one set corresponds to a point in the other set is a generic concern of the so-called *point matching* problem. This is an essential and challenging task that has been widely encountered in various domains, such as sonar processing [1], localization of DNA markers [2], to name a few. In the realm of image processing and computer vision, such point matching technique or the so-called *feature point matching* aims to automatically establish point-to-point correspondences between two images acquired from the same scene but through different viewpoints. Thus, this process plays a vital role in the following two categories of applications: 1) determining a certain geometric relationship or transformation between the given two images or two cameras, and 2) facilitating human beings or machines on identifying the same visual object appeared in the given two images. For the former category, some well-known examples are the spatial transformation estimation (e.g., affine transformation) of the two images under registration [3] [4] [5] or the epipolar geometry estimation in the camera calibration for 3D scene reconstruction [6] [7]. For the latter category,

Figure 1.1. A typical framework of the feature point matching system.

some typical applications are object detection [8] [9], object recognition [10] [11], common visual pattern discovery [12] [13] [14], and face matching [15] [16].

As shown in Fig. 1.1, a typical framework of the feature point matching system consists of the following two main stages to establish point-to-point correspondences across two images: 1) feature point detection and local descriptor computation; 2) local feature descriptor matching. In the first stage, variable sizes of circular- or elliptical-shaped regions of interest (or the so-called *support regions* [17]) are identified on each input image based on some distinctive features (e.g., corners, blobs) computed by using the Gaussian derivatives [17], image phase [18], or entropy [19]. The center point of each identified region is considered as a *feature point*, and a local *feature descriptor* (i.e., feature vector) of this feature point is established based on the support region. Such feature descriptor is oftentimes required to be invariant to certain geometric transformations and photometric changes. For example, the well-known *scale invariant feature transform* (SIFT) [17] feature descriptor, which is used in this thesis, is invariant to translation, rotation, scaling and up to a certain degree of illumination changes. This is achieved by computing the weighted gradient orientation histograms over a circular image region with the radius being equal to the estimated scale. Finally, at the end of the first stage, two sets of feature points are independently generated for the two input images, respectively, and each feature point is associated with a local feature descriptor to depict the support region that centered at this feature point.

In the second stage, point-to-point correspondences could be established by

applying the commonly used *nearest neighbor matching criterion* [17] to the local feature descriptors obtained in the previous stage. That is, each feature point in one set will be linked to a feature point in the other set provided that the Euclidean distance between their associated feature descriptors is the shortest among such measurements of all possible correspondences. Furthermore, this distance must be smaller than a pre-determined threshold. Otherwise, it will be declared that no point-to-point correspondence can be found for this feature point.

1.2 Motivations and Objective

Despite the huge success of the feature point matching algorithms [17] [20] [21] that are based on the framework discussed in Section 1.1, a large number of incorrect point-to-point correspondences would still be incurred especially when the two images under matching undergo a large viewpoint variation and/or contain a severely cluttered background [17]. This is mainly due to the following two reasons: 1) as a result of the large viewpoint variation incurred between the two images under matching, the feature descriptors used to represent the same physical surface of the scene in different images may appear dramatically different; 2) the severely cluttered background may introduce many irrelevant feature descriptors that will eventually disturb the local feature descriptor matching stage. This situation becomes even worse if the content of the background presents a strong degree of similarity with that of the object. In this case, the irrelevant feature descriptors established on the background are pretty much the same with the feature descriptors established on the object that commonly exists in the two images under matching. Therefore, it is extremely difficult to further distinguish one from the others. The above-mentioned two adverse factors will ultimately lead to a large percentage of incorrect matches

in the point-to-point correspondences generated by the feature point matching algorithms. Consequently, the applications based on such matching pairs will most probably produce unsatisfying results. For that, how to further refine these initially established point matching pairs to significantly increase the the precision (i.e., the percentage of correct matches) of the final point-to-point correspondences becomes the main focus of this thesis.

In the open literature, many efforts have been made to remove incorrect matching pairs while retain the correct ones in the initially established point-to-point correspondences or the so-called *initial feature point matching pairs*. This could be achieved by imposing additional geometric constraint on the established point-to-point correspondences. For example, in the well known *Random Sample Consensus* (RANSC)-based geometric verification methods [22] [23] [24], a linear spatial transformation (e.g., affine transformation) is assumed to be incurred between the two images under matching. The parameters of the transformation matrix can be therefore estimated by the RANSAC method. If the spatial coordinates of the two feature points in an established point-to-point matching pair don't conform with the assumed spatial transformation, this matching pair will be considered as a incorrect one. Although this kind of methods have been widely applied in various feature point matching-based applications, such as stereo vision and camera self-calibration, they are fundamentally suffered from the following two problems: 1) this kind of methods are not applicable if a non-linear spatial transformation is incurred between the two images under matching; 2) if the number of mismatches in the initial feature point matching pairs is high (e.g., the precision is lower than 50%), the performances of these methods would drop significantly.

To overcome these problems, Schmid *et al.* [25] use a large number of *local* linear spatial transformations to approximate the *global* non-linear transformation incurred between the two images under matching. Note that each local linear

Figure 1.2. The proposed framework of feature point matching system.

transformation here is estimated based on the fixed number of point-to-point correspondences that are spatially located near each other. More recently, the spatial relationship among multiple point-to-point matching pairs are commonly exploited for the refinement of the initial feature point matching pairs. For example, the pairwise spatial consistency principle [12] [26] applied to every two point matching pairs, and the topological relationship among every three point matching pairs [10], to name a few.

1.3 Main Contributions and Highlights

As discussed in Section 1.1, the applications of feature point matching can be broadly divided into two categories depending on the application's objective. Therefore, as shown in Fig. 1.2, the refinement of initial feature point matching pairs will be branched into two different processes in this thesis. That is, it will become either mismatch removal or correct match identification.

For the first category of applications, a *sufficiently large* number of point-to-point correspondences are needed for the RANSAC or the iterative least square [27] to reliably estimate the geometric relationship between the given two images or

two cameras. However, the initial feature point matching pairs often contain mismatches. Consequently, the accuracy of the estimation will be reduced. Therefore, how to identify and discard those incorrect matching pairs while maintaining enough number of point-to-point correspondences for the estimation is the main goal of the refinement process for this type of applications. In our work, such refinement process is denoted as *mismatch removal*.

For the second category of applications, a ‘large’ number of point-to-point correspondences are *not* necessarily needed. Instead, the main requirement for this category of applications is to yield the correct point-to-point correspondences with fairly high confidence in the first place. A typical example is the so-called *common visual pattern discovery* [14], which is exploited to facilitate human beings or machines to identify the same visual object in the given two images. In this case, one can imagine that even just one point-to-point correspondence that correctly links the same visual object in the two images is sufficient to achieve the application’s goal. Consequently, the number of final point-to-point correspondences after the refinement of feature point matching pairs could be fewer while the accuracy is of the highest concern. In our work, such refinement process is called *correct match identification*.

To improve the performance of feature point matching, three main contributions are made in this thesis: 1) mismatch removal, 2) common visual pattern discovery, 3) feature histogram equalization (FHE). Highlight of each contribution is given in the following:

- Mismatch Removal

For the first contribution, a *bipartite* graph model-based mismatch removal method is proposed to identify the mismatched correspondences among the initial feature point matching pairs and then discard them. The key idea

of this mismatch removal method lies in the use of the proposed one-to-one *region-to-region* correspondences (or the *coherent region pairs* as coined in this thesis) as the reference information to re-evaluate each existing *point-to-point* correspondence to judge whether it is a correct match to retain or a mismatch to remove.

- Common Visual Pattern Discovery

For the second contribution, a *directed* graph (or *diagraph*) model-based correct match identification method is proposed with the main objective on yielding the correct point-to-point correspondences for the application of common visual pattern discovery. The novelty of our approach begins with the use of diagraph model that has two link weights on each link instead of one link weight only as inherited in the conventional (undirected) graph model. Consequently, the resulted two link weights are more robust to combat various adverse scenarios that are commonly encountered in the application of common visual pattern discovery, such as large viewpoint variations and indiscriminative feature descriptors.

- Feature Histogram Equalization (FHE)

For the third contribution, as shown in Fig. 1.2, feature histogram equalization is a novel local feature enhancement technique that can be ‘optionally’ applied to the computed feature descriptors to enhance the degree of discrimination among the feature descriptors. Thus, the precision of the initial feature point matching pairs (at a fixed recall rate) could be increased. The basic idea of FHE is to independently modify the vector-component values of feature descriptors at each vector dimension to effectively enhance the feature contrast. To our best knowledge, such an attempt has not been explored in any existing work.

1.4 Organization

The outline of this thesis is as follows. In Chapter 2, the background of feature point matching is briefly introduced to pave the way for the further discussions in the following chapters.

In Chapter 3, the proposed bipartite graph model-based mismatch removal method for the application of wide-baseline image matching is described in detail. Simulation results have reasonably demonstrated that our proposed mismatch removal method effectively identifies and removes a significant number of mismatched point-to-point correspondences for the wide-baseline image matching.

In Chapter 4, the proposed directed graph model-based correct match identification method for the application of common visual pattern discovery is elaborated at length. Extensive experiments have been conducted on the simulated SIFT feature points for standing out the technical challenges, followed by performing evaluations on six thoughtfully chosen natural image pairs and Columbia dataset to demonstrate the efficacy and robustness of the proposed method.

In Chapter 5, the proposed feature contrast enhancement technique—feature histogram equalization (FHE) is described in detail. Extensive simulation results have clearly shown that our proposed FHE method could effectively improve the precision of the initial feature point matching pairs, especially to those image pairs with low feature contrast, while the improvement could be marginal for the image pairs with high feature contrast as expected.

Finally, Chapter 6 summarizes this thesis and discusses two promising future research directions for improving the performance of feature point matching and local feature descriptors, respectively.

Chapter 2

Background

As mentioned in Chapter 1, all the research works in this thesis are actually starting from the local features. Therefore, in this chapter, some basic knowledge of the commonly used feature point detection and local feature descriptor computation methods will be briefly introduced first followed by a detailed explanation on why SIFT feature is chosen and extensively used in this thesis.

2.1 Feature Point Detection

The goal of feature point detection is to find the *support regions* [17] [28] that have distinctive and stable properties so that those regions could be easily found again when the image itself undergoes various geometric transformations, such as translation, rotation and scaling. The center point of each detected region is called a *feature point*. To achieve this goal, the support regions must satisfy the following two requirements: 1) it is a discriminating image region that significantly differs from other regions in the image; 2) it can be easily and reliably identified by the feature detector even the image undergoes various geometric transformations.

One of the earliest and most well-known feature point detection method is Harris

Figure 2.1. An illustration of the blob structure that commonly encountered in the images: three black vertical lines stand for the three imaging planes placed at the different distances to the object in the scene; the three red vertical lines stand for the blob structures.

corner detector [29], which tries to identify the image regions that contain corner points in the center. This is achieved by identifying the image pixel whose gradient is significantly large on two different orientations. After the corner point is detected, a *fixed* size of image region centered on this point is considered as the support region. Note that, this kind of support region could be accurately detected if the image itself only undergoes translation and rotation. Consequently, Harris corner detector is often said to be translation and rotation invariant in literature.

More recently, most research works in feature point detection is trying to find the support regions that can be reliably detected when the image undergoes more complex geometric transformations, such as scaling and affine transformation. For that, the approach based on the blob detection is commonly exploited in literature [30] [17] [31]. As shown in Fig. 2.1, blob structures are actually the reflections of the same physical surface of the object in the scene but at different imaging planes placed at different locations. If the imaging planes are parallel to the object but at different distances, which is the case illustrated in Fig. 2.1, each resultant blob structure is a circular image region with radius being proportional to the

distance between the imaging plane and the object. After normalize the circular blob structures into unit circles, the contents of the normalized regions should be almost the same. Therefore, circular blob structure is regarded as scale invariant [17]. On the other hand, if the imaging planes are placed at different viewing angles, the resultant blob structures can be best approximated by the elliptical image regions with different lengths and orientations of major and minor axes. Similar to the circular blob structures, the contents of the normalized elliptical blob structures should also be quite similar. Consequently, in literature, elliptical blob structure is regarded to be affine invariant [31]. Due to the nice geometric invariant properties of the blob structures, they are commonly used as the support regions in the feature point detection. Now the question boils down to how to identify the blob structures in an input image.

For the circular blob detection, the earliest detector can be traced back to the *Laplacian of Gaussian* (LoG) detector [30] which is based on the scale space theory [32] [33]. In order to detect the blob structure, N LoG band-pass filters with different variance values are convolved with the input image, respectively. Therefore, N filtered images are generated, and each pixel on the filtered image contains the following three information: 1) the pixel's spatial coordinates; 2) the variance value of the LoG filter applied to the input image; 3) the convolution result. Among all the pixels in the N filtered images, it is suggested in [32] that the spatial coordinates of the pixel whose convolution result is a local maximum or minimum is the center point of a circular blob structure. The radius of this blob structure can be determined by multiplying the variance value associated to this pixel with a constant. One of the main problem of the LoG detector is the high computational complexity caused by the convolutions between the LoG band-pass filters and the input image. To solve this problem, Lowe proposed the SIFT feature detector [17], where the LoG band-pass filter is approximated and replaced by the *Difference of*

Gaussian (DoG) filter to accelerate the speed in the blob detection.

For the elliptical blob detection, Mikolajczyk *et al.* first proposed the Harris affine feature detector and Hessian affine feature detector [31]. These two methods are in fact post-processing techniques to the existing SIFT feature detector. The main idea of these two methods is to iteratively adjust the detected circular image region to an elliptical one according to the dominant gradient orientations estimated in that region. The main difference between these two detectors lies in the way to estimate the dominant gradient orientations. For Harris affine feature detector, it uses the same method as that in the Harris corner detector. However, for Hessian affine feature detector, it uses the method based on the eigenvalues and eigenvectors of the auto-correlation matrix (or the so-called *Hessian matrix*) of the detected image region. Another well-known elliptical blob detection method is the *Maximally Stable Extremal Regions* (MSER) detector [28]. This method first binarizes the input images by using M different threshold to generate the M binary images. For each binary image, the connected component analysis is performed to further group the pixels whose intensity value is above the threshold into different candidate regions. If the candidate regions established on the M different binary images have some overlapped regions, these regions are considered as the blob structures and their bounding ellipses are outputted as the final detection results.

Mikolajczyk *et al.* [34] further provided a comprehensive evaluation of the most popularly used feature detectors and come to the conclusion that different feature detectors show their complementary properties and the so-called all-purpose detector doesn't exist. It is also pointed out in [34] that if the two images under matching undergoes a large scale change or/and viewpoint variation, Harris affine feature detector and Hessian affine feature detection perform the best among all the feature detectors being evaluated. This is why these two feature detectors are used in the following chapters of this thesis.

2.2 Local Feature Descriptor Computation

The main purpose for computing the feature descriptor is to find a discriminating representation for each support region that is identified in the feature point detection mentioned in Section 2.1. By performing the local feature descriptor computation, a multi-dimensional vector (or the so-called *feature descriptor*) will be established based on all the pixels in the identified support region. It is highly expected that each dimension of the feature descriptor will capture some unique characteristic of the support region, and this characteristic should remain invariant to the geometric variation or/and illumination change incurred to the image.

Recently, a large number of feature descriptors have been proposed in literature [17] [35] [21] [36]. According to their different motivations and objectives, these feature descriptors can be broadly classified into two categories: 1) feature descriptors with better performance; 2) feature descriptors that can be computed with faster speed or less memory requirement. More information about the feature descriptors in each category will be provided as follows.

For the feature descriptors that aim to provide better performance, different methods have been proposed to form the feature vector that could potentially capture more distinctive and robust characteristics of the support region. These characteristics include the gradient moment invariants [37], the shape context [38], the salient patterns obtained by applying steerable filters [39] or complex filters [40] and the gradient orientation distribution used to generate scale invariant feature transform (SIFT) [17] and gradient location-orientation histogram (GLOH) [35]. A comprehensive evaluation is performed by Mikolajczyk *et al.* [35], and SIFT is ranked as the best feature descriptor among all the frequently used ones due to its highly discriminative power and robustness to viewpoint and/or illumination changes. This is why in this thesis, SIFT feature descriptor is extensively used to

demonstrate the effectiveness of our proposed methods in the remaining chapters.

The success of SIFT feature descriptor lies in using gradient orientation histograms to represent the shape information in the support region. For that, the algorithm first equally divides each support region into 4 by 4 sub-regions. For each sub-region, the gradients of all the image pixels within the region are calculated first followed by quantizing the gradient orientations into 8 quantization steps: $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3}{4}\pi, \pi, \frac{5}{4}\pi, \frac{3}{2}\pi, \frac{7}{4}\pi$. Based on the quantized gradient orientation, an 8-bin gradient orientation histogram is computed to represent the shape information of this sub-region. Note that, for each support region, there are altogether 16 sub-regions, therefore, 16 8-bin gradient orientation histograms can be independently generated, and these histograms can be cascaded together to form a 128-dimension feature vector.

In fact, the gradient orientation histogram has been widely used to construct various local feature descriptors, for example, the gradient location-orientation histogram (GLOH). Compared with SIFT feature descriptor, GLOH divides the support regions into more unequal sized sub-regions and uses finer steps to quantize the computed gradient orientations. However, the main disadvantage of such histogram-based feature descriptor computation methods lies in the following two aspects: 1) high computational complexity; 2) large memory requirement to store the established high dimensional feature vectors. To tackle these two problems, many feature descriptors that can be computed with faster speed or less memory requirement has been proposed [21] [36]. For the former, the well-known *speeded up robust feature* (SURF) [21] utilizes integral images to speed up the computation of convolutions to establish SIFT-like local gradient histograms. The resulted 64-dimension SURF feature descriptors are proven to yield comparably performances as SIFT feature descriptors when there is only a small viewpoint variation incurred between the two images under matching [21]. For the latter, the *compressed histogram of gradients*

(CHoG) [36] exploits the Huffman and Gagic trees in the coding theory to further quantize and encode the gradient orientation histograms to reduce the number of bits to store the feature descriptors, such as SIFT and GLOH.

Chapter 3

Bipartite Graph Model-Based Mismatch Removal for Wide-baseline Image Matching

Wide-baseline image matching plays a vital role in many image processing and computer vision related applications. The matching operation is conducted over two separately acquired images of the same scene but from two widely disparate views. Most of the conventional approaches accomplish this task by performing a feature point matching process between the two images under matching. However, it is well-recognized that the conventional wide-baseline image matching method based on the SIFT descriptors alone usually yields a large number of mismatches, especially when the two images under matching undergo a large viewpoint variation or encounter a severely cluttered background [17]. Such mismatches will impair the system performance of the target application. Therefore, it is imperative to detect and remove those mismatched pairs from the initially established SIFT-based point-to-point correspondences. In this chapter, a bipartite graph model-based mismatch removal method is proposed to benefit those applications that require wide-baseline

image matching. The turnkey novelty lies in our proposed strategy on utilizing the developed *region-to-region* matching pairs as useful ‘reference information’ to help re-evaluate each initially established SIFT-based *point-to-point* matching pair so as to determine whether it is a correct match to retain or a mismatch to remove.

The remaining of this chapter is organized as follows. In Section 3.1, the relevant background and comparable state-of-the-art methods are succinctly reviewed. To realize our turnkey novelty as mentioned earlier, it is necessary to find out how to generate *region-to-region* matching pairs, or the so-called candidate region matching pairs, and this process is detailed in Section 3.2. In Section 3.3, a novel region similarity metric is proposed for the measurement of the degree of similarity between the two regions of each candidate region matching pair. In Section 3.4, a bipartite graph model is exploited to represent the established candidate region matching pairs. The Hungarian method [41] and the proposed region-to-region similarity measurement metric are then applied to the established bipartite graph model for searching the optimal one-to-one CRPs. Section 3.5 presents our simulation results to show the effectiveness of using the developed one-to-one CRPs to identify mismatched point-to-point matching pairs for removal. Conclusions are drawn in Section 3.6.

3.1 Literature Review

One intuitive and straightforward way for conducting mismatch removal is to first assume that a linear spatial transformation (e.g., affine or perspective transformation) is incurred between the two images under matching. A robust estimation of the transformation matrix can be obtained by using the *random sample consensus* (RANSC) method [42], for example. If the spatial coordinates of the two feature points in an established point-to-point matching pair don’t conform with

the assumed spatial transformation, this matching pair will be regarded as a ‘mismatch’ [43] [23]. However, a large number of mismatches still remain undetected through this approach. This is mainly due to the following two reasons. First, when the two images under matching are captured from two widely disparate views, a simple linear spatial transformation is oftentimes insufficient to depict the complex spatial transformation that is actually incurred between these two images. Second, when the precision (i.e., the percentage of correct matches) of the SIFT-based point-to-point matches falls below 50%, which is a scenario commonly encountered in the wide-baseline image matching, the spatial transformation matrix estimated by the RANSC-like methods becomes highly inaccurate [17].

Rather than making a certain assumption about the spatial transformation between two images, some mismatch removal approaches [22] [44] exploited a more universally applicable geometric constraint, called *epiploar geometry*, to improve the performance of mismatch removal. Note that, the epiploar geometry is used to depict the geometric constraint among the 3D point of the scene that the two images under matching aim to capture and the corresponding points in the two images. It is satisfied no matter how large the viewpoint variation is incurred between the two images. Consequently, those point-to-point matching pairs that violate the epiploar geometry are all considered as mismatches. However, the estimation of the fundamental matrix in the epiploar geometry is still obtained by the RANSC-like methods, which make the performances of this kind of approaches drop drastically when the precision of the initial feature point matching pairs is below 50%.

To tackle this problem, the geometric relationships among multiple feature points have been exploited for mismatch removal in literature. Tuytelaars and Van Gool [20] first proposed an iterative method to remove the incorrect point-to-point matching pairs based on the homographies that are locally estimated by the feature points near each other. Ferrari *et al.* [10] proposed a novel technique,

called the *topological filter*, by imposing their proposed *sidedness constraint* as a test condition; that is, for every three correct point-to-point matching pairs (thus, there are three feature points on each image), if one feature point lies on the right-hand side (or the left-hand side for the same argument) of the directed line passing through the other two feature points on the same image, this topological relationship should remain unchanged among the three corresponding feature points on the other image as well. By identifying and discarding those point-to-point matching pairs that violate this sidedness constraint, their method had demonstrated some promising results on mismatch removal, but at the expense of high computational complexity. Zhou *et al.* further proposed two geometric coding techniques [45] [46] to encode the relative spatial relationship between each two feature points in an image to some binary values (or the so-called *signatures*). It is expected that for correct feature point correspondences, the signatures of the feature points in one image should be equal to the signatures of the corresponding feature points in the other image. Based on this test condition, the incorrect feature point matching pairs can be identified and removed.

The goal of our work is to identify and eliminate mismatches from the established SIFT-based point-to-point matching pairs. The turnkey novelty lies in the strategy of constructing *one-to-one region-to-region* matching pairs, which are coined as the *coherent region pairs* (CRPs) in our work, and then utilizing them as the reliable reference information to verify whether each existing *point-to-point* correspondence is a mismatch or not. If a point-to-point matching pair is considered as a correct match, then the following conditions must be met: if one end point of a given correspondence pair (or a link) falls in a region of a CRP in one image, then the other end point of the same link must fall in the corresponding region of the same CRP presented in the other image. Otherwise, this point-to-point matching pair will be regarded as a *mismatch* and should be discarded.

Now, the question boils down to how to construct CRPs. For that, as shown in Fig. 3.1, we utilize the established SIFT-based point-to-point matching pairs to guide the merging of the segmented image regions that are spatially connected to form the *candidate* region matching pairs. Note that, each candidate region matching pair could be one-to-one, one-to-many, or many-to-one correspondence types. Therefore, such pairs can be best represented by a bipartite graph model that is widely used in image retrieval [47] [48]. Hungarian method [41] is then applied to this model for pruning those one-to-many and many-to-one matching pairs in order to identify the optimal *one-to-one* CRPs. The optimality is achieved by maximizing the total degree of region similarity over the entire bipartite graph model under the constraint that each region-to-region correspondence must be one-to-one. Note that for the two regions involved in each candidate region-to-region matching pair, their similarity is measured by the proposed region similarity measurement metric. Extensive simulation results have shown that our proposed bipartite graph model-based mismatch removal method could effectively identify a large number of mismatches for removal.

3.2 Proposed Region-to-region Matching Method

3.2.1 SIFT-based Initial Feature Point Matching Pairs

Given a wide-baseline image pair, the SIFT feature descriptors [17] are formed on each image—one feature descriptor for each feature point detected by the *affine covariant region detector* [28] [31]. For each SIFT feature descriptor f given in one image, two most similar SIFT feature descriptors g and h are identified from the other image according to the l_2 -norm distance measurement, and the descriptor g is further defined as being ranked higher than the descriptor h on the similarity

Figure 3.1. The flowchart of our proposed bipartite graph model-based mismatch removal method.

measurement, both with respect to the descriptor f . If the following test condition [17] is met, then the descriptors f and g form a point-to-point matching pair (f, g) .

$$\frac{\|f - g\|_2}{\|f - h\|_2} < \tau, \quad (3.1)$$

where $\tau = 0.8$ is a pre-determined threshold as suggested in [17]. The established SIFT-based point-to-point matching pairs will be treated as the initial feature point matching pairs and further utilized to guide the merging of the segmented image regions to form the *candidate* region matching pairs as follows.

3.2.2 Proposed Algorithm for Merging Image Segments

To identify regions, an efficient graph-based image segmentation method proposed by Felzenszwalbis *et al.* [49] is applied to each image to obtain two sets of image regions or segments, $S = \{s_i | i = 1, 2, \dots, M\}$ and $T = \{t_j | j = 1, 2, \dots, N\}$, where M and N are the total numbers of image segments obtained from each image, respectively. Note that the parameters of this image segmentation method are set

Figure 3.2. An example of one initial region matching pair (blue regions). A pair of yellow crosses denotes an established SIFT-based point-to-point matching pair.

as $\sigma = 0.5$, $k = 100$, $min = 100$ to ensure that the two images under matching tend to be over-segmented. Based on the initially established SIFT-based point-to-point matching pairs, the initial region-to-region matching pairs can be quickly established as follows. If the two feature points of a given point-to-point matching pair respectively fall upon regions s_i and t_j , the pair (s_i, t_j) can be formed and denoted as an *initial* region matching pair. Meanwhile, region s_i or region t_j is called the *associate region* of the other. To demonstrate, a pair of yellow crosses as shown in Fig. 3.2 indicates an established point-to-point matching pair. A pair of blue-colored image segments, where the yellow crosses falling upon, forms an initial region-to-region matching pair.

Due to possible illumination, viewpoint and/or scale variations, a segmented region in one image might simultaneously correspond to multiple image segments presented in the other image; that is, ‘one-to-many’ and ‘many-to-one’ region matching pairs are quite likely encountered at this stage. A typical example of such case is shown in Fig. 3.3. Note that, the blue segmented region in the left image simultaneously corresponds to the four image segments in the right image. To achieve the final goal of establishing only *one-to-one* CRPs, a segment merging operation is proposed for reducing such cases as follows.

For a given region in Image 1 (e.g., s_1 in Fig. 3.4(a)), if there exist two or

Figure 3.3. An example of one-to-many initial region matching pairs. Each pair of yellow crosses denotes an established SIFT-based point-to-point matching pair, and the blue segmented region in the left image actually corresponds to the blue, red, gray and green image segments in the right image.

more associate regions in Image 2 (i.e., a ‘one-to-many’ region correspondence case) that are *spatially connected* (e.g., regions t_2 and t_3 in Fig. 3.4(b)), then these two adjacent associate regions are considered as being *over-segmented* and should be merged together to become one larger associate region instead (i.e., region $t_{2,3}$ in Fig. 3.4(c)). In general, the above-mentioned segment merging operation, as summarized in Algorithm 1, will be applied to ‘Image 1’ for each one-to-many region matching pair across the two images under matching, followed by applying the same operation to ‘Image 2’ for taking care of possible ‘many-to-one’ region matching pairs, likewise. After merging, all the region pairs are called the *candidate* region matching pairs (e.g., (s_1, t_1) and $(s_1, t_{2,3})$ in Fig. 3.4(c)).

One should note that not all the one-to-many and many-to-one initial region matching pairs were successfully merged in the above-mentioned stage, since the only segment merging criterion used was ‘spatially connected.’ That is, for a region in one image, whose associate regions appeared in the other image are spatially *disconnected*, these associated regions will not be merged and thus the one-to-many region correspondence case is still encountered. For example, as shown in Fig. 3.5, the blue region in the left image still corresponds to both the blue region and the green region in the right image after applying the proposed segment merging

Figure 3.4. A graphical illustration of the proposed segment merging operation: (a) established point-to-point matching pairs based on the SIFT feature descriptors and the image segments generated by [49], (b) *initial* region matching pairs, (c) the final *candidate* region matching pairs.

Figure 3.5. An example of one-to-many region correspondence case after applying the proposed segment merging operation.

operation.

Likewise, it is possible that many-to-one region correspondence cases can also happen, if the violation of the spatially connected criterion incurred. Since the established region-to-region correspondences will be used as the reference information to check whether each SIFT-based point-to-point matching pair is a correct match or a mismatch, these region-to-region correspondences must be *one-to-one* so that there won't be any ambiguity being created as the reference information. Therefore, it is necessary to further prune those one-to-many and many-to-one links to become an one-to-one correspondence individually (e.g., the one-to-many region correspondence case in Fig. 3.5 has become one-to-one region correspondence in

Algorithm 1 Proposed Segment Merging Algorithm

INPUT: The *initial* region matching pairs, and two sets of the image segments, S and T , generated by [49] for the two images under matching, respectively

OUTPUT: The *candidate* region matching pairs

```

for  $i = 1$  to  $M$  do
  For region  $s_i$ , find its associate region set  $O(s_i)$ .
  if regions  $t_a, \dots, t_b$  in region set  $O(s_i)$  are spatially connected then
    Merge regions  $t_a, \dots, t_b$  as a new region  $t_{a,\dots,b}$ 
    Replace regions  $t_a, \dots, t_b$  by  $t_{a,\dots,b}$  in region set  $T$ 
    Replace region pairs  $(s_i, t_a), \dots, (s_i, t_b)$  by a new region pair  $(s_i, t_{a,\dots,b})$ 
  end if
end for
for  $j = 1$  to the updated cardinality of region set  $T$  do
  For region  $t_j$ , find its associate region set  $O(t_j)$ .
  if regions  $s_c, \dots, s_d$  in region set  $O(t_j)$  are spatially connected then
    Merge regions  $s_c, \dots, s_d$  as a new region  $s_{c,\dots,d}$ 
    Replace regions  $s_c, \dots, s_d$  by  $s_{c,\dots,d}$  in region set  $S$ 
    Replace region pairs  $(s_c, t_j), \dots, (s_d, t_j)$  by a new region pair  $(s_{c,\dots,d}, t_j)$ 
  end if
end for

```

Fig. 3.6). This objective can be fulfilled by exploiting a search technique to conduct an optimal search. For that, a *region-to-region* similarity measurement metric is required to be developed and used for facilitating the search and generating one-to-one CRPs. This is described at length in the following section.

3.3 Proposed Region-to-region Similarity Measurement Metric

The proposed metric is used to individually measure the degree of similarity between the two regions for each candidate region matching pair. Three measuring aspects are considered for the design of *region-to-region* similarity measurement metric: 1) photometric consistency, 2) neighborhood consistency, and 3) point-to-point matching pair density, which will be detailed in the following sub-sections,

Figure 3.6. An example of an one-to-one coherent region pair by applying our proposed mismatch removal method.

respectively. Note that the last two aspects are proposed in this thesis. These three measurements will be linearly combined and exploited to measure the degree of similarity between the two regions for each candidate region matching pair.

3.3.1 Photometric Consistency

Ferrari *et al.* [10] proposed a photometric consistency criterion to measure the amount of similarity between region i in one image and region j in the other image on their texture and color separately, followed by combining these two measurements together as the final measurement. For each candidate region matching pair containing K point-to-point matching pairs, it is suggested in this thesis that the region's photometric consistency measurement should be equal to the average of the photometric consistency values individually measured over all these K pairs.

In [10], both the texture similarity measurement and the color similarity measurement are individually performed for each point-to-point matching pair over the two support regions. Note that since these two elliptical support regions usually have different size, a normalization is required to normalize them into a unit circle each and denoted as p and q , respectively. The texture similarity measurement, denoted as $NCC(p, q)$, can be computed by taking the absolute value of the *normalized cross-correlation* (NCC) measurement [50] between p and q . As to the

color similarity measurement, $dRGB(p, q)$ is calculated by taking the average of the pixel-wise intensity differences measured between p and q in the RGB color space. Finally, the photometric consistency η_1 of each point-to-point matching pair can be obtained by combining the texture and color similarity measurements together as [10]

$$\eta_1 = 1 + NCC(p, q) - \frac{dRGB(p, q)}{100}. \quad (3.2)$$

Note that the value of $NCC(p, q)$ lies in the range $[0, 1]$. In practice, the value of $dRGB(p, q)$ usually takes a value between 0 to 100. Consequently, the value of η_1 is approximately in the range of $[0, 1]$.

Finally, the photometric consistency, denoted as α_1 , for each candidate region matching pair (i, j) can be calculated by averaging the photometric consistency η_1 of all K point-to-point matching pairs that are contained in the candidate region matching pair (i, j) ; that is,

$$\alpha_1(i, j) = \frac{1}{K} \sum_{k=1}^K \eta_1. \quad (3.3)$$

3.3.2 Neighborhood Consistency

It has been observed that applying the Ferrari *et al.*'s photometric consistency criterion as described in Section 3.3.1 to the normalized support regions might still lead to incorrect region-to-region correspondences due to insufficiently large area of observation, simply based on the support regions. Fig. 3.7 illustrates a simple example to demonstrate this issue. As shown in Fig. 3.7(a) and Fig. 3.7(b), the smaller yellow-colored ellipses indicate two support regions centered at their respective feature points. Apparently, this is a mismatch after examining the image content of larger regions surrounding these two support regions. Indeed, by

(a) Image 1

(b) Image 2

(c) Support region in Image 1

(d) Support region in Image 2

(e) Extended support region in Image 1

(f) Extended support region in Image 2

Figure 3.7. A simple illustration showing why a checking of the neighborhood of each support region might be useful on the establishment of more reliable region-to-region correspondences. The smaller ellipses indicate the support regions centered at the respective two feature points, and the larger ellipses indicate the *extended* support regions which are obtained by extending the area coverage of the corresponding support regions by a factor of two. Note that all the support regions and the extended support regions are normalized into unit circles, followed by cutting a square image patch with the size of 100×100 for demonstration as shown in (c)-(f).

examining these two normalized support regions¹ as shown in Fig. 3.7(c) and Fig. 3.7(d), respectively, one can see that these two image regions bear a high degree of similarity. However, by extending the area coverage of each support region outward two times as the larger yellow-colored ellipses shown in Fig. 3.7(a) and Fig. 3.7(b), the *extended* support regions (also with normalization) present drastically different image content as shown in Fig. 3.7(e) and Fig. 3.7(f).

For the computation of the neighborhood consistency for each point-to-point matching pair, both the texture and the color similarity measurements are individually performed over the two normalized extended support regions, which are denoted as $\hat{p}$ and $\hat{q}$, respectively. For the measurement of the texture similarity, it has been discovered through our investigation that exploiting the Euclidian distance between the two SIFT feature descriptors $\hat{f}$ and $\hat{g}$ (respectively established on $\hat{p}$ and $\hat{q}$) could provide a much more robust measurement than that of utilizing the NCC measurement as described in Section 3.3.1. For the measurement of color similarity, the same computation of $dRGB$ as described in Section 3.3.1 will be applied to $\hat{p}$ and $\hat{q}$. Finally, the neighborhood consistency η_2 for each point-to-point matching pair can be obtained by combining the texture and the color similarity measurements together as

$$\eta_2 = \exp \left(-\|\hat{f} - \hat{g}\|_2 - \frac{dRGB(\hat{p}, \hat{q})}{100} \right), \quad (3.4)$$

where $\|\cdot\|_2$ represents the l_2 -norm used to calculate the Euclidian distance between the two SIFT feature descriptors $\hat{f}$ and $\hat{g}$, and the exponential function is incorporated here to guarantee that the value of η_2 will fall within the range of $[0, 1]$.

Finally, the measurement of the proposed *neighborhood consistency*, denoted as

¹The normalization operation is performed by first transforming each elliptical support region into a circular region, followed by resizing it into a unit circle. Only a square image patch with the size of 100×100 cut from the unit circle is shown here for demonstration.

α_2 , for each candidate region matching pair (i, j) is the average of the neighborhood consistency η_2 measured over all K point-to-point matching pairs that are contained in the candidate region matching pair (i, j) ; that is,

$$\alpha_2(i, j) = \frac{1}{K} \sum_{k=1}^K \eta_2. \quad (3.5)$$

3.3.3 Point-to-point Matching Pair Density

Intuitively, it makes sense to conjecture that if the density of point-to-point matching pairs contained in a candidate region matching pair (i, j) is high, the region correspondence established between region i and region j should be considered as reliable; that is, the candidate region matching pair (i, j) is more likely a correct match. The point-to-point matching pair density $\alpha_3(i, j)$ is therefore defined as

$$\alpha_3(i, j) = \frac{K}{\max(Area(i), Area(j))}, \quad (3.6)$$

where $Area(i)$ and $Area(j)$ are utilized to calculate the area of region i and region j , respectively. The ‘area’ here can be easily obtained by simply counting the total number of image pixels that are contained in each region.

Based on (3.3), (3.5), and (3.6), the proposed region-to-region similarity measurement metric $e_{i,j}$ for regions i and j of the candidate region matching pair (i, j) is given by ²

$$e_{i,j} = \lambda_1 \alpha_1(i, j) + \lambda_2 \alpha_2(i, j) + \lambda_3 \alpha_3(i, j), \quad (3.7)$$

where the three weighting parameters $\lambda_1 = 5$, $\lambda_2 = 2$ and $\lambda_3 = 4$ are empirically

²Note that $e_{i,j}$ is simply a scalar, denoting the degree of *region* similarity measured between region i from one image and region j from the other image. One should *not* view it as a matrix element.

determined in advance. Note that, if region pair (i, j) is not a candidate region matching pair, then $e_{i,j} = 0$ will be set.

3.4 Generation of Coherent Region Pairs via the Bipartite Graph

3.4.1 Model Representation

The established candidate region matching pairs as discussed in Section 3.2.2 can be modelled by the *bipartite graph* [41] for the generation of one-to-one *coherent region pairs* (CRPs); that is, all the regions from the established candidate region matching pairs are denoted as the *vertices* of the graph (e.g., regions s_1 , t_1 and $t_{2,3}$ in Fig. 3.4(c)). These vertices can be divided into two sets, X and Y , where set X consists of all the regions from Image 1 (i.e., region s_1 in Fig. 3.4(c)) and set Y consists of all the regions from Image 2 (i.e., regions t_1 and $t_{2,3}$ in Fig. 3.4(c)). For each candidate region matching pair (say, region pair (s_1, t_1)), a *link* is assigned to connect its associated two vertices (i.e., vertices s_1 and t_1).

Further note that each region from set X may have more than one corresponding region in set Y , and vice versa. However, our goal is to establish *one-to-one* region correspondences only (i.e., CRPs). To achieve this goal, the Hungarian method [41] is exploited to conduct an optimal search over the entire candidate region matching pairs. The search criterion is to find the one-to-one region-to-region correspondences that altogether yield the largest total degree of region similarity among all possible establishments of CRPs, since the larger the total degree of region similarity, the more reliable the identified set of CRPs being served as the reference information to judge whether each point-to-point matching pair is a correct match to retain or a mismatch to remove. In the following, the mathematical formulation for the

above-mentioned ideas will be given.

3.4.2 Bipartite Graph Matching

The *incidence vector* [41] (denoted as $\mathbf{z}$) is defined as a binary vector with the dimension equal to the cardinality of set X (i.e., the number of regions in set X) multiplied by the cardinality of set Y (i.e., the number of regions in set Y). Each element $z_{i,j} \in \mathbf{z}$ indicates whether region i (or the i th vertex) from set X and region j (or the j th vertex) from set Y has been established as a region-to-region matching pair; if so, the element $z_{i,j}$ is assigned with a value 1, or a value 0, otherwise. By exploiting the incidence vector $\mathbf{z}$, it becomes mathematically convenient to compute the total degree of region similarity (denoted as C) based on $z_{i,j}$ and $e_{i,j}$; that is, $C = \sum_j \sum_i e_{i,j} z_{i,j}$. Furthermore, it is our goal to have all the generated CRPs being a one-to-one correspondence each, this is equivalent to say that $\sum_j z_{i,j} = 1$, for all i , and $\sum_i z_{i,j} = 1$, for all j . These two conditions will be collectively imposed as a constraint during the optimal search using the Hungarian method [41].

In summary, the problem of establishing the one-to-one CRPs becomes an optimization problem on finding the optimal incidence vector $\mathbf{z}^*$ by maximizing the total degree of region similarity under the constraint of one-to-one region correspondence; that is,

$$\arg \max_{\mathbf{z}} C = \arg \max_{\mathbf{z}} \sum_j \sum_i e_{i,j} z_{i,j},$$

subject to

$$\sum_j z_{i,j} = 1, \forall i \text{ and } \sum_i z_{i,j} = 1, \forall j. \quad (3.8)$$

In the graph theory, such optimization problem in (3.8) is called the *bipartite graph*

matching, and the *optimal* incidence vector $\mathbf{z}^*$ could be determined by a combinatorial optimization algorithm, called the Hungarian method [41], in polynomial time.

The basic idea of the Hungarian method is to treat (3.8) as a classic assignment problem [41] in the optimization theory as follows: imagine each element in set X is a girl and each element in set Y is a boy, $e_{i,j}$ is then assumed to represent the degree of fondness between the i th girl in set X and the j th boy in set Y . The objective of the assignment problem is trying to find a good partner for each boy and girl so that they are all satisfied with such assignment. As pointed out in [41], the assignment problem has been well studied in literature and can be iteratively solved by the augmenting path method [41].

3.4.3 Proposed Process for Selecting the Best Matching Pairs

It is worthwhile mentioning that all the point-to-point matching pairs contained in the CRPs after mismatch removal are regarded as correct point-to-point matching pairs. However, depending on the application's requirements, the targeted or actually needed number of point-to-point matching pairs could be (much) smaller than the available ones. In this case, a hierarchical selection process is proposed here to select the top matches according to the desired number as follows.

The CRP with the highest degree of region similarity yielded will be chosen first. If the number of point-to-point matching pairs contained in this CRP is smaller than the targeted number, all the point-to-point matching pairs will be chosen without further applying any ranking process among them. This selection process will be continued for the next CRP with the second highest degree of region similarity, and so on, until the total number of point-to-point matching pairs contained in all the

CRPs identified so far is more than the targeted number. In this case, a ranking process among the point-to-point matching pairs contained in the current CRP will be performed according to the Euclidean distance measured between the two SIFT feature descriptors in each point-to-point matching pair. Note that, the smaller the Euclidean distance is, the higher the ranking of the point-to-point matching pair would be. According to this rule, only the desired number of point-to-point matching pairs with higher ranking in the current CRP will be chosen to meet the targeted number of point-to-point matching pairs.

3.5 Experimental Results

To evaluate the performance of our proposed bipartite graph model-based mismatch removal method, a public-domain dataset has been downloaded from [20] for conducting simulation experiments. The dataset consists of six well-known wide-baseline image pairs, denoted as *oase*, *auto*, *church*, *tree*, *mex* and *wash*, respectively. Each image pair has two images containing the same scene but captured under a large viewpoint variation. To demonstrate, these six image pairs are shown in Fig. 3.8.

For the generation of SIFT-based point-to-point matching pairs, the Hessian affine feature detector [31] is exploited to detect feature points, at which the SIFT feature descriptors are established [17]. The point-to-point matching pairs are therefore formed based on the nearest neighbor matching criterion [17] as stated in Chapter 1. The simulation results are documented in Table 3.1. Alongside, the number of correct matches is also provided for each test case and used as the ground truth; these correct matches are visually identified by us.

As shown in Table 3.1, one can observe that a large number of point-to-point mismatches are incurred. Even for the best case as shown in the image pair *mex*,

(a) oase

(b) auto

(c) church

(d) tree

(e) mex

(f) wash

Figure 3.8. A demonstration of six wide-baseline image pairs downloaded from [20]. Each image pair has two images containing the same scene but captured under a large viewpoint variation.

Table 3.1

THE EVALUATION OF THE SIFT-BASED POINT-TO-POINT MATCHING PAIRS ESTABLISHED ON THE SIX WIDE-BASELINE IMAGE PAIRS DOWNLOADED FROM [20]. THE CORRECT MATCHES ARE VISUALLY IDENTIFIED BY US AND USED AS THE GROUND TRUTH.

Image Pairs	Number of SIFT-based Matching Pairs	Number of Correct Matches	Precision
oase	568	101	17.78%
auto	732	65	8.88%
church	281	109	38.79%
tree	450	172	38.22%
mex	315	133	42.22%
wash	394	112	28.43%
Average	457	115	29.05%

only 42.22% matching pairs are correct, not to mention the worse case delivered by the image pair *auto*, where the precision falls drastically down to 8.88%. Such poor accuracy on establishing point-to-point matching pairs can be further verified by Fig. 3.9(a) and Fig. 3.9(c), where the point-to-point correspondence links by applying the nearest neighbor matching criterion are plotted across the two images under matching. In fact, those matching pairs inevitably cause tremendous errors for the follow-up image processing tasks, such as 3D reconstruction, motion estimation, and 2D/3D registration, to name a few. This justifies that an effective mismatch removal method is highly instrumental to those applications.

To conduct performance evaluation of mismatch removal methods, both the topological filter proposed by Ferrari *et al.* [10] and our proposed bipartite graph model-based mismatch removal method are independently applied to the identical sets of SIFT-based point-to-point matching pairs as established perviously. For a fair comparison between these two mismatch removal methods, only a fixed number of top (i.e., the most matched) point-to-point matching pairs are kept in each

(a) point-to-point matching pairs by applying the nearest neighbor matching criterion

(b) Top 120 matching pairs by applying the proposed method

(c) point-to-point matching pairs by applying the nearest neighbor matching criterion

(d) Top 120 matching pairs by applying the proposed method

Figure 3.9. An illustration of point-to-point matching pairs by applying the nearest neighbor matching criterion and our proposed mismatch removal method on image pair *mex* shown in (a)-(b) and image pair *church* shown in (c)-(d).

Table 3.2

A PERFORMANCE COMPARISON OF TWO MISMATCH REMOVAL METHODS: THE TOPOLOGICAL FILTER [10] VERSUS OUR PROPOSED BIPARTITE GRAPH MODEL-BASED METHOD. THE PRECISION IS COMPUTED BY DIVIDING THE NUMBER OF CORRECT MATCHES BY THE TOTAL TARGETED NUMBER, 120. THESE SIX WIDE-BASELINE IMAGE PAIRS ARE DOWNLOADED FROM [20].

Image Pairs	Ferrari <i>et al.</i> [10]		Proposed Method	
	Number of Correct Matches	Precision	Number of Correct Matches	Precision
oase	71	59.33%	76	63.3%
auto	47	39.17%	49	40.83%
church	93	77.5%	99	82.5%
tree	96	80.0%	97	80.8%
mex	94	78.3%	103	85.8%
wash	72	60.0%	77	64.1%
Average	79	65.73%	84	69.56%

method, since by this way we can only examine the precision (i.e., the percentage of correct matches). In our experiment, the top 120 is chosen because this number is around the average number of the correct matches as shown in Table 3.1. For the experimental results obtained from the Ferrari *et al.*'s method [10], the top 120 point-to-point matching pairs are simply selected based on the ranking of the Euclidean distance measured between the two SIFT feature descriptors for each point-to-point matching pair. In our case, the top 120 matching pairs are identified based on the hierarchical selection process as described in Section 3.4.3.

For the performance evaluation, the generated top 120 matching pairs are visually verified against the ground truth of correct matches as described previously. Consequently, the precision can be simply calculated by dividing the number of correct matches by the total targeted number, 120. These simulation results are summarized in Table 3.2, from which one can observe that the average precision of

point-to-point matching pairs has been significantly increased up to 65.73% by the topological filtering method [10] and 69.56% by our proposed mismatch removal method (that is, improved by 2.26 and 2.39 times, respectively). Furthermore, our proposed mismatch removal method consistently outperforms the Ferrari *et al.*'s topological filtering method in all test image pairs by an additional 3.83% improvement on average precision. As demonstrated in Fig. 3.9(b) and Fig. 3.9(d), our proposed mismatch removal method is able to effectively remove the incorrect matches in point-to-point matching pairs that are established based on the nearest neighbor matching criterion.

3.6 Summary

In this chapter, a novel bipartite graph model-based mismatch removal method is proposed to significantly reduce the erroneous SIFT-based point-to-point correspondences often encountered in the wide-baseline image matching. The turnkey novelty lies in the use of the proposed one-to-one coherent region matching pairs (CRPs) as the reference information to judge whether each existing point-to-point matching pair, constructed based on the SIFT feature descriptors, is a correct match to retain or a mismatch for removal. To achieve this goal, the SIFT-based point-to-point matching pairs are first utilized to guide the merging of the segmented image regions that are spatially connected to form the candidate region matching pairs. A bipartite graph model is then developed to represent the established candidate region matching pairs. In general, this bipartite graph consists of one-to-many and many-to-one region correspondences. The Hungarian method, together with the developed region-to-region similarity measurement metric, are applied to the bipartite graph model for searching the optimal one-to-one CRPs by maximizing the total degree of region similarity over the entire bipartite graph model under

the constraint that each of the established region-to-region correspondence must be kept in one-to-one during the search. The generated CRPs are used to determine whether each point-to-point matching pair is a correct match or simply a mismatch.

The performance of our proposed method is evaluated and further compared with a state-of-the-art algorithm using six well-known wide-baseline image pairs. Simulation results have clearly shown that our proposed method is able to significantly reduce the mismatches incurred in the SIFT-based point-to-point matching pairs and consistently outperform the method under comparison.

Chapter 4

Common Visual Pattern Discovery via Directed Graph Model

From the standpoint of high-level recognition, the generic objective of *common visual pattern discovery* is to identify all the visual objects that are commonly presented in two images under comparison. This process could be exploited to facilitate human beings or machines on recognizing the visual objects co-existing in both images. However, such seemingly simple task could oftentimes require laborious efforts even for human beings to perform. For instance, how easily can you identify a pair of game disks (as visual objects) co-existing in both images as shown in Fig. 4.1? Furthermore, if there are such pairs, how many of them? In fact, the above-mentioned task has potential applications in many image processing and computer vision systems, such as near-duplicate image detection [51] [52] [53], object discovery [13] [54] [55] [56], content-based retrieval [57] and image segmentation [58] [9] [59] [60].

Figure 4.1. A demonstration on how difficult to determine whether there is any visual object commonly presented in both images and, if any, how many of them?

4.1 Overview of Common Visual Pattern Discovery

The difficulty of common visual pattern discovery can be alleviated by applying a robust feature point matching method (e.g., [13] [61]), which is aimed to establish as many correct point-to-point correspondences between two images under comparison as possible. However, common visual pattern discovery (or identification) is, in fact, beyond the above-mentioned feature point matching problem with two distinct accounts as follows.

First, in common visual pattern discovery, the generation of a ‘large’ set of point-to-point correspondence links across the two images is *not* necessarily required; in fact, it is less desirable, even the algorithm is able to do so. Instead, a ‘sufficient’ number of point-to-point correspondences with a fairly high degree of confidence on these established links are mostly desired. Refer to Fig. 4.1 for example, one can imagine that even a single point-to-point correspondence that correctly links the same game disk presented in these two images is enough to help human being on quickly identifying that pair of game disks. That is, so far as the point-to-point correspondence establishment is concerned, it is the ‘quality’ (i.e., correctness) issue,

not the ‘quantity’ issue. Second, unlike the feature point matching, whose final results are point-to-point correspondences only, the final results of common visual pattern discovery are the point-to-point correspondents grouped into different sets, and hopefully each set exactly corresponds to one visual object. Obviously, without such grouping, it is impossible to distinguish the different visual objects that are commonly presented in two images, which is undesirable from the application point of view.

Further note that certain adverse factors are often encountered and need to be addressed properly, such as scale change, viewpoint variation, indiscriminate feature representation, and so on. All these factors will lead to miss detection or yield incorrect common visual pattern discovery results. In this chapter, a *directed* graph (or *digraph*) model-based approach is exploited for finding common visual patterns between two images under comparison. The turnkey solution lies in the fact that there are *two* link weights associated to each link in the directed graph model, and each link weight indicates what is the degree of association from the standpoint of one vertex ‘seeing’ the other vertex. With such established graph model, the discovery of a common visual pattern boils down to finding a strongly-associated subgraph from the digraph at an assumed scale-change factor. In order to discover all common visual patterns, a range of scale-change values will be experimented in our work. Hopefully, each identified strongly-associated sub-graph corresponds to *one* common visual pattern.

For the first challenge regarding the computation of the two link weights for each link, we consider the influence of the link weights inserted from the neighboring vertices for each vertex of the link. For that, a novel *generalized median* measurement method (coined as the *n-ranking* process in our work) and a *link weight enhancement* technique are developed in our work. The entire setup facilitates the computation of the required two link weights for each link in the digraph with multiple advantages

that will be discussed in detail later on. For the second challenge regarding how to efficiently identify the strongly-associated subgraph, the non-cooperative game theory [62] is exploited to handle the non-symmetric weighted adjacency matrix encountered in our approach. Extensive simulation results show that the proposed digraph-model-based method outperforms three comparable state-of-the-art algorithms [12] [14] [63], especially under the presence of large viewpoint variations and with challenging image contents.

The remaining of this chapter is organized as follows. Section 4.2 briefly introduces the research background of common visual pattern discovery. Section 4.3 discusses the basic framework of the conventional graph model-based approach followed by a succinct review of the comparable state-of-the-art methods. Section 4.4 describes the inherited limitation of the conventional undirected graph model on reflecting the degree of association among the vertices and introduces our proposed directed graph model (or digraph) in detail. Section 4.5 explains how to identify the strongly-associated subgraphs from the digraph by using the non-cooperative game theory. The above-mentioned link weight computation and subgraph extraction are performed at each assumed scale change factor and such experiments are conducted over a range of scale change factor values. Section 4.6 provides extensive simulation results with benchmark comparisons. Conclusions are drawn in Section 4.7.

4.2 Background

There are two general approaches to identify visual patterns that are commonly presented in a set of images: 1) methods based on Bag-of-Words (BoW) model [64]; 2) robust feature point matching methods. Methods based on BoW first quantize the continuous feature descriptors established on each image under comparison into discrete visual words that are predefined in a codebook (or the so-called *dictionary*)

through vector quantization. Each region from the image can then be effectively represented by the statistical distribution of the visual words (or the so-called *visual word histogram*). Finally, the matching of any two image regions could be performed by comparing the similarity of their visual word histograms. One of the main disadvantage of this approach is that its performance is heavily influenced by the quantization loss in the vector quantization. Therefore, how to improve the performance of vector quantization so that the quantized feature descriptor could be more discriminative becomes the main challenge of this approach. To tackle this problem, two main research directions have been proposed. The former one aims to alleviate the difficulty by using the soft assignment of visual words [65] [66] [67] [68]. In this paradigm, the quantization value of each feature descriptor is a weighted combination of several neighboring visual words rather than a single visual word in the conventional methods. By doing so, the quantization loss is reduced which leads to increased performance of common visual pattern discovery. For the latter research direction, researchers try to address the challenge by modeling visual word co-occurrences [69] [70] [71] [72] [73] through data mining techniques. The basic idea of this kind of methods lies in the observation that the co-occurrence of specific visual words could provide much better discrimination of different visual patterns.

Another main disadvantage of the methods based on BoW model is that it usually requires a large number of images to pre-determine the dictionary. Therefore, this approach may not be suitable for the common visual pattern discovery among a small number of images (e.g., common visual pattern discovery between two images as discussed in this chapter). In this case, the second approach based on feature point matching would be more widely applied. Among all the methods belonging to this category, recent research work suggest that the graph model-based

approach could provide satisfying results with reasonable computation complexity [12] [74] [75] [14]. In the following, the basic framework of the conventional graph model-based approach and its associated terminologies will be established and described for laying the foundation of the proposed directed graph model-based approach.

4.3 Undirected Graph Model: Conventional Approach

In this section, the basic framework of the graph-model-based approach and its associated terminologies will be established for the development of this chapter. Some state-of-the-art *undirected* graph-model-based methods (conventional approach) will be succinctly discussed as the background information for understanding our proposed *directed* graph model approach. An equivalent graph-model-based approach, called *graph matching*, will be highlighted in the last sub-section to augment the appreciation and insights of the various graph-model-based approaches.

4.3.1 Framework Overview

First, a set of feature points will be detected in each image independently, followed by generating 128-dimension *feature descriptors* (or feature vectors), one for each detected feature point. For that, the well-known *scale invariant feature transform* (SIFT) [17] is exploited in our work to generate feature descriptors. Further assume that there are M and N feature points generated in these two images, respectively. To further establish the so-called *feature-point correspondence* encountered in any application that requires feature-point matching, there will be $M \times N$ possible combinations on such point-to-point correspondence establishment from one image to

the other, exhaustively. Now, the whole issue boils down to identify as many *correct* point-to-point correspondences as possible based on these total combinations; this will be discussed in detail later on.

In the context of graph model, each above-mentioned feature-point correspondence is treated as a *vertex*. The relationship between any two vertices (i.e., two feature-point correspondence pairs) can be quantified by assigning a *link weight* on the link that connects these two vertices in order to reflect their *degree of association*. If this quantitative measurement satisfies a pre-determined thresholding criterion, these two vertices are considered *strongly-associated*. Consequently, it is confident to conclude that both vertices (or point-to-point correspondence links) are correct ones, and this means that two links are correctly established between the same visual object in the two images under comparison. Thus, the problem of common visual pattern discovery becomes to identify and extract such strongly-associated vertices from the established undirected graph model; the entire set of such identified vertices is treated as one strongly-associated *subgraph*. One way to solve this subgraph-identification problem is to mathematically formulate it as an integer quadratic programming problem in the optimization theory [76] [77], and this will be exploited and discussed in Section 4.5.

4.3.2 Undirected Graph Model Approach

To identify correct point-to-point correspondence links, the *pairwise spatial consistency* principle is commonly exploited [12] [78] [63] [14]. The basic idea of this principle is as follows. When a visual object that commonly appears in both images is with a similarity-transformation relationship between the two images, the Euclidian distance measured between any two feature points located on the same

visual object in one image should be proportional to that of two corresponding feature points of the same object in the other image only by a real-valued constant. Furthermore, this constant is equal to the *ratio of scaling factors* (or the so-called *scale-change factor* in [14]) of the visual object across the two images under comparison. Note that this information is not known in advance.

Consequently, the degree of pairwise spatial consistency (or inconsistency, for that matter) of any two vertices (i.e., two feature-point correspondence pairs) by incorporating an assumed scale-change factor will be served as the *link weight* of these two vertices to reflect how likely (or unlikely) these two vertices are both correct feature-point correspondences.

To judge whether the given two vertices are two correct feature-point correspondences, simply based on the pairwise spatial consistency principle, is a *binary* decision. Thus, two kinds of decision errors could be incurred: false alarm and miss detection. For the false alarm, it means that even the two vertices satisfy the pairwise spatial consistency principle, but they might still turn out to be two incorrect feature-point correspondences. For the miss detection, it means that even the two vertices don't satisfy the pairwise spatial consistency principle, but they are actually two correct feature-point correspondences. In the real practice of common visual pattern discovery, these two kinds of errors are commonly encountered due to the two main adverse factors: *geometric deformation* and/or indiscriminate representation of feature descriptors (or the so-called 'outliers' in [12]), respectively. For the former, the geometric deformation is usually incurred when two images were captured from two different viewpoints (or viewing angles). Therefore, simply using the information of scale-change factor alone will tend to yield more incorrect feature-point correspondences. The latter is due to the limitation of SIFT algorithm and/or complicated image contents, such as the ones with cluttered background and/or with repetitive patterns (causing strong 'self-similarity' in this case), to

name a few, indiscriminate feature descriptors will be inevitably generated by the SIFT algorithm.

It is worthwhile mentioning here that the methodologies developed for addressing common visual pattern discovery problem using undirected graph-model approach can be broadly classified into two categories. One is on the computation or enhancement of link weights so that the graph model becomes more robust to combat various adverse factors as mentioned just now. The other is focused on the development of more accurate and/or faster algorithms to approximately solve the integer quadratic programming problem mentioned in Section 4.3.1.

Based on the foundation of pairwise spatial consistency principle, Cour *et al.* [63] propose a normalization process to enhance the link weights of the undirected graph model. By iteratively reducing the link weight on the link that connects any two *uninformative vertices* being identified [63], these vertices will eventually not be treated as strongly-associated vertices any more. Rather than using the pairwise spatial consistency principle, Duchenne *et al.* [75] extend the consideration of *two* feature points (i.e., ‘pairwise’) to *three* feature points at a time, for establishing the so-called *point triplet spatial consistency*. As a result, more robust link weights on the proposed undirected *hypergraph* can be computed as demonstrated in [75].

Regarding the second methodology mentioned above, a comprehensive survey of the related methods can be found in [79]. Note that the integer quadratic programming problem is an NP-hard problem, therefore an approximation (or the so-called *relaxation*) method is needed to make this problem computational tractable. After relaxation, different optimization methods can be applied to solve the ‘relaxed’ problem in order to get an approximate solution to the original integer quadratic programming problem. Such techniques that can be found in the existing literature include the graduated assignment algorithm [80], the spectral method [12], a max-product belief propagation algorithm (denoted as COMPOSE) [81], semi-definite

programming-based method [82] [83], and a method based on the evolutionary stable strategies [84].

Completely different from the existing works mentioned above that are all based on the undirected graph model, a *directed* graph (or *digraph*) model-based method is proposed in this chapter by taking the advantage of the digraph that has *two* link weights on each link, instead of *one* link weight on each link as inherited in the conventional graph model. Consequently, the resultant two link weights are able to reflect the different degrees of association from the standpoint of how one vertex ‘sees’ the other one, so far as the degree of association is concerned. Note that the two vertices of our proposed directed graph model will be considered as ‘strongly-associated’ only when both link weights ‘view’ each other in high degree of association; i.e., both vertices reach a consensus that they are strongly associated to each other. Due to this more stringent checking on the confirmation of the degree of association between each pair of two vertices under consideration, our proposed digraph-based approach becomes more robust to combat various adverse circumstances, such as large viewpoint variations and indiscriminate feature descriptors; all these are commonly encountered in common visual pattern discovery application.

4.3.3 Graph Matching Method: An Equivalent Approach

It is worthwhile mentioning that there is another graph-model-based approach, called *graph matching* [85], which has been exploited for tackling common visual pattern discovery problem in [80] [12] [86] [78]. Instead of trying to identify the strongly-associated subgraph in a *single* undirected graph model, *two* undirected graphs (one from each image) need to be generated from the outset for conducting the graph matching process.

First, an undirected graph model is built upon all the feature points detected from one image, and the same process is applied to the other image as well. Consequently, if there are M and N feature points being detected from the two images, respectively, two undirected graphs with M and N vertices will be generated. Unlike the single undirected graph model established in Section 4.3.1, each vertex of the graph established in graph matching framework as described in this sub-section is a feature point detected from one image, rather than a feature-point correspondence across two images. The Euclidian distance between the two vertices (i.e., two feature points on the same image) is used as the link weight of the link that connects these two vertices. Note that the previously described pairwise spatial consistency principle can still be exploited to identify correct matches between the vertices of the two graphs, thus called *graph matching*; that is, a link (with two vertices) from one graph shall be matched to another link from the other graph provided that these two link weights satisfy the pairwise spatial consistency principle. Now, the problem of common visual pattern discovery becomes a graph matching problem which is aimed to identify all the correct matches between the vertices of the two graphs. From the viewpoint of graph theory, such problem could also be mathematically formulated as the integer quadratic programming problem as that in the undirected graph-model-based approach [80] [86]. As pointed out in [79], both the undirected graph-model-based approach and the graph matching-based approach are equivalent to each other despite their differences on graph-model formulation.

4.4 Directed Graph Model: Proposed Approach

4.4.1 Model Setup and Motivation

Given two sets of feature points, generated by SIFT [17] and respectively denoted as sets A and B , assume that there are M and N feature points in these two point sets. The product space $U = A \times B$ contains all $M \times N$ feature-point correspondence pairs; that is, each member in the set U is a two-tuple element (a, b) , where $a \in A$ and $b \in B$. The degree of similarity of two feature descriptors (128 dimensions each) established at feature points a and b can be quantitatively evaluated by measuring their Euclidian distance (i.e., the l_2 -norm); the shorter the Euclidian distance, the larger the degree of similarity. In this case, it indicates that the established feature-point correspondence will be considered as a correct one with a high degree of confidence. Therefore, such similarity measurement can be used to identify those weak or incorrect feature-point correspondences from the set U . For that, a pre-determined distance threshold $\varepsilon = 0.3$ is used in our work to prune the set U down to a (much) smaller set V . Consequently, all the ‘surviving’ vertices $v_i \in V$ after thresholding are considered as reliable feature-point correspondence pairs for facilitating the remaining processing, as follows.

To simplify the notation of mathematics without creating any ambiguity, we shall let the subscript index of the feature-point correspondence pair v_i (viewed as a vertex) equal to the subscript index of the feature point from the set A throughout this chapter. Accordingly, for every two vertices $v_i = (a_i, b_k)$ and $v_j = (a_j, b_l)$, the Euclidian distance between the spatial coordinates of the feature points a_i and a_j (measured in the set A) as well as the Euclidian distance between the spatial coordinates of the feature points b_k and b_l (measured in the set B) can be separately computed and denoted as $d_{i,j}$ and $d_{k,l}$, respectively. According to the previously-defined pairwise spatial consistency principle, if the vertices v_i and v_j are correct

point-to-point correspondences and the feature points a_i and a_j are from the same visual object in one image, then the feature points b_k and b_l must come from the same visual object appeared in the other image. Further note that each visual object has its own generic *scale* value from the viewpoint of its resided image. The two scale values of the same visual object appeared in their respective images under comparison will be different, provided that these two images are acquired from different viewpoints or viewing angles. Such *scale-change factor*, denoted as α , can be approximately obtained from the ratio of the above-mentioned distances; that is,

$$\alpha = \frac{d_{k,l}}{d_{i,j}}. \quad (4.1)$$

The constant α is a real-valued number, and it is consistent with the zoom-in (if $\alpha > 1$) or zoom-out (if $\alpha < 1$) factors of the same visual object judging from one image to the other image. This factor can be further utilized to compute the initial link weights that reflect the degree of pairwise spatial consistency among the vertices. That is, given two vertices v_i and v_j , the initial link weight $w_{i,j}$ of the link v_i-v_j is computed by [14]

$$w_{i,j} = \|d_{k,l} - \alpha d_{i,j}\|_2 = |d_{k,l} - \alpha d_{i,j}|. \quad (4.2)$$

Note that the smaller the initial link weight $w_{i,j}$ is, the more consistent the pairwise spatial relationship of vertices v_i and v_j would be; that is, vertices v_i and v_j are more likely to be the correct point-to-point correspondences that link the same visual object across two images.

Note that the values of α are not available, since the correct point-to-point correspondences are unknown and need to be established in the first place. Therefore, a reasonable range of α values ranging from 0.25 to 4 are exhaustively experimented in this chapter. Over this range, the step size of 0.2 is used for the sub-range of [1,

Figure 4.2. Examples of two graph models: (a) conventional *undirected* graph model, in which only one link weight is associated with each link (without direction). Here, all the link weight values are presumably computed based on Eq. (4.2); (b) *directed* graph model with *two* link weights on each link (with the pointing direction as indicated by the arrow). Here, the two link weight values are computed by applying Eqs. (4.3) and (4.4) developed in this chapter, respectively.

4] (for the zoom-in cases). The reciprocal of each scale-change factor value determined for the sub-range of $[1, 4]$ is respectively used for the sub-range of $[0.25, 1]$ (for the zoom-out cases) to ensure that various visual objects can be individually detected under different scale-change factors. Refer to (4.2), a smaller value of $w_{i,j}$ indicates that the scale-change factor α is a more correctly assumed value to use. However, this has created a fundamental problem as we have observed as follows.

It has been observed that the computation of $w_{i,j}$ by (4.2) only involves two vertices of the link without considering the link weights of their neighboring vertices at each side. Here, the *neighboring vertices* of vertex v_i for the link v_i-v_j is simply defined as those vertices that have direct link connections to vertex v_i besides vertex v_j . For illustration, a simple undirected graph model is shown in Fig. 4.2(a), which has four vertices: v_1 , v_2 , v_3 and v_4 , to reveal the fundamental issue encountered in the use of the undirected graph model. This highly motivates us to explore the directed graph (or digraph) model for tackling common visual pattern discovery problem. In Fig. 4.2(b), the link that connects each pair of two vertices is represented by a solid line with a link weight value as indicated on the respective

link. All the link weights here have presumably been computed by using (4.2). Notice that the computation of each link weight only involves the two vertices of the link. For example, the link weight $w_{1,2} = 3$ is only determined based on vertices v_1 and v_2 via (4.2), and it has nothing to do with their neighboring vertices v_3 and v_4 . However, we would like to point out here that such single link weight could be insufficient to accurately reflect the degree of pairwise spatial consistency between the two vertices as explained below.

Notice that vertices v_3 and v_4 are two neighboring vertices of vertex v_1 , and the involved links v_1-v_3 and v_1-v_4 have the link weights $w_{1,3} = 5$ and $w_{1,4} = 7$, respectively. Relatively speaking, the link weight $w_{1,2} = 3$, should be regarded as ‘small’ when compared with the link weights $w_{1,3} = 5$ and $w_{1,4} = 7$. Therefore, from the viewpoint of vertex v_1 , the link v_1-v_2 should be considered as pairwise spatial *consistent*. (Recall that the smaller the link weight $w_{i,j}$, the larger the degree of pairwise spatial consistency.) However, from the viewpoint of vertex v_2 , this conclusion is reserved, since compared with the link weights $w_{2,3} = 1$ and $w_{2,4} = 0.5$, the link weight $w_{1,2} = 3$ is considered ‘large,’ relatively speaking. Therefore, from the viewpoint of vertex v_2 , the link v_1-v_2 should be considered as pairwise spatial *inconsistent*. At this juncture, it should be convinced that using a *single* link weight has its inherited fundamental limitation on reflecting the degree of association between the two vertices under consideration that might arrive at a conflicting conclusion from the standpoints of the involved vertices regarding their pairwise spatial consistency. To address this issue, the *directed* graph (or *digraph*) model is investigated in this chapter, since it has *two* link weights associated with each link as indicated in Fig. 4.2(b); that is, one weight is assigned to the link with the direction pointing from vertex v_i to vertex v_j (denoted by $e_{i,j}$) and the other weight is assigned to the same link but with a reversed direction, pointing from vertex v_j to vertex v_i (denoted by $e_{j,i}$). That is, $e_{i,j}$ reflects the degree of association

from the ‘eyes’ of vertex v_i : how does it ‘see’ vertex v_j . The interpretation of $e_{j,i}$ is in the reverse way, likewise. This will completely eliminate the possibility of incurring the conflicting conclusion on pairwise spatial consistency as discussed above.

4.4.2 The Gaussian Link-Weight Mapping Function

Now, the entire issue boils down to how to compute the two link weights $e_{i,j}$ and $e_{j,i}$ for each link v_i-v_j in the digraph model. To achieve this goal, a novel *link-weight mapping function* is proposed here to map the initially computed $w_{i,j}$ to $e_{i,j}$ and to $e_{j,i}$, separately. The design of this mapping function considers two objectives: (1) it must be able to reflect the “relativeness” among all the neighboring link weights of each vertex under consideration; (2) it would be highly desirable to have an effect on enhancing the link-weight values. All these design considerations will facilitate the identification and extraction of the strongly-associated subgraph more accurately, which is to be discussed in detail in Section 4.5.

For accomplishing the first objective, a new ranking operation is proposed in this chapter, called the *n-ranking* process, which is operated on the weights of all the links that are connected to vertex v_i . That is, after the initial link weights between vertex v_i and each of its immediate neighboring vertices are individually computed by exploiting (4.2), these values are sorted in the increasing order, followed by taking the n th smallest value as the n th-ranked value, denoted as γ_i . In our work, the value of n has been empirically determined by taking 10% of the total number of the neighboring vertices of vertex v_i , followed by rounding up the result to the nearest integer¹. Similar to the *median* value encountered in statistics, the computed value γ_i is also quite robust to combat the presence of those link weight values

¹If 50% is chosen, the *n-ranking* process is equivalent to compute the *median* value. Therefore, the proposed *n-ranking* process can be viewed as a *generalized median* computation. This novel concept and process can be treated as a generic statistical metric and measurement that can be also exploited in other applications.

that are either too small or too big compared with other neighboring link weights. Therefore, the proposed n -ranking process offers a simple and yet effective way to dynamically reflect how large the neighboring link weights of vertex v_i , relatively speaking, through our proposed robust statistical measurement. Since different vertices usually have different neighboring vertices, thus, in general, the n th-ranked value γ_i of vertex v_i is not equal to the n th-ranked value γ_j of vertex v_j . For example, in Fig. 4.2(a), vertices v_1 and v_2 have the n th-ranked values $\gamma_1 = 5$ and $\gamma_2 = 1$, respectively, when $n = 2$ is assumed.

The purpose of achieving the second objective—enhancing the link weights—is to further facilitate the follow-up process on identifying the strongly-associated subgraph based on the digraph established at each assumed scale-change factor. In essence, the identification of such subgraph can be viewed as a two-class clustering operation where one cluster (i.e., a subgraph, in our case) contains those vertices that satisfy pairwise spatial consistent principle and collectively regarded as *one* common visual pattern. Hopefully, this pattern (or subgraph) corresponds to *one* visual object, from the application standpoint. All the remaining vertices that fail to satisfy the principle are classified into the other cluster and viewed as pairwise spatial inconsistent vertices. From the viewpoint of clustering, it is desirable to have a smaller intra-cluster distance and a larger inter-cluster distance. The fundamental concept of the link weight enhancement introduced in this chapter bears the same objective and effectiveness, which is described as follows.

The use of Gaussian function as the link-weight mapping function is proposed in this chapter, which is expressed as

$$e_{i,j} = \frac{1}{\gamma_i} \exp \left(-\frac{w_{i,j}^2}{2\gamma_i^2} \right). \quad (4.3)$$

For the generation of the other link weight $e_{j,i}$, it can be computed likewise for vertex v_j by simply replacing γ_i with γ_j in (4.2) to reflect the consideration of

“relativeness” of all immediate neighboring link weights at vertex v_j . That is,

$$e_{j,i} = \frac{1}{\gamma_j} \exp \left(-\frac{w_{i,j}^2}{2\gamma_j^2} \right). \quad (4.4)$$

Note that the *initial* link weights, $w_{i,j}$, that were obtained from (4.2) is a difference of two Euclidian distances (with incorporation of scale-change factor) in order to reflect *the degree of spatial consistency* of the two vertices v_i and v_j ; that is, the *smaller* the difference, the more likely the two vertices are correct point-to-point correspondences. On the other hand, the *relative* link weights, $e_{i,j}$ and $e_{j,i}$, carry the physical meaning of the *degree of association* between vertices v_i and v_j ; thus, the *larger* the weight values of both $e_{i,j}$ and $e_{j,i}$, the more likely the two vertices are strongly-associated to each other. That is, the physical meaning of the initial link weight $w_{i,j}$ and its resultant relative link weights $e_{i,j}$ and $e_{j,i}$ are in a reversed sense and trend; this matches the mathematical forms presented in (4.3) and (4.4).

In fact, to map $w_{i,j}$ to $e_{i,j}$ and to $e_{j,i}$, respectively, one can use a simple, say, linear decreasing function. The benefits of using the Gaussian function rather than a linear decreasing function for the mapping are as follows. First, the nonlinearity of the function provides an enhancement effect on the computation of the relative link weights, $e_{i,j}$ and $e_{j,i}$. This is because the *inflection point* of the Gaussian function is the point where the Gaussian function changes from concave to convex [87]. Furthermore, we set this inflection point exactly at $w_{i,j} = \gamma_i$ in (4.3) and $w_{i,j} = \gamma_j$ in (4.4) with reasons that should be clear from the next discussions.

In the following, let us use (4.3) for a detailed discussion, and the same reasoning is applicable to (4.4), likewise. In (4.3), the Gaussian function is a concave function for $w_{i,j} < \gamma_i$, and thus the relative link weight $e_{i,j}$ calculated by (4.3) should be larger than those of being calculated by using a linear decreasing function, which is assumed that it passes through the reflection point as well. On the other hand,

the Gaussian function becomes convex for $w_{i,j} > \gamma_i$, and the relative link weight $e_{i,j}$ calculated by (4.3) becomes smaller than those of being calculated by using the same linear decreasing function. Back to our example in Fig. 4.2(a), since $w_{1,2} < \gamma_1$, a larger link weight value (compared with that of being generated by using a linear decreasing function) will be assigned to $e_{1,2}$ to further enhance the degree of association between vertices v_1 and v_2 . That is, from the viewpoint of vertex v_1 , the link v_1-v_2 should be considered as *more* pairwise spatial consistent. On the contrary, since $w_{1,2} > \gamma_2$, a smaller value will be yielded and assigned to $e_{2,1}$. Therefore, from the viewpoint of vertex v_2 , the link v_1-v_2 should be considered as *less* pairwise spatial consistent.

In addition, the use of the Gaussian function can help reduce the computational complexity by taking the advantage of the Gaussian function's *three-sigma rule*, which means that all the relative link weights $e_{i,j}$ and $e_{j,i}$ can be set to zeros right away without further computation, once $w_{i,j} > 3\gamma_i$. For example, in Fig. 4.2(b), the link weights $e_{3,4}$ and $e_{4,3}$ are assigned with zeros according to this rule, since $w_{3,4} > 3\gamma_3$ and $w_{4,3} > 3\gamma_4$.

4.4.3 Matrix Representation of Directed Graph Model

According to the graph theory, a graph model with K vertices can be represented by a $K \times K$ *weighted adjacency matrix* E . The element from the i th row and the j th column of matrix E is the link weight $e_{i,j}$ of the directed graph model. Note that, all the diagonal elements $e_{i,i}$ (for $i = 1, \dots, K$) of matrix E are set to zeros, since there is no link pointing from any vertex v_i back to itself. According to this definition, the weighted adjacency matrix E of the directed graph model in Fig. 4.2(b) becomes

$$E = \begin{bmatrix} 0 & 0.1671 & 0.1213 & 0.0751 \\ 0.01111 & 0 & 0.1960 & 0.9925 \\ 0.1213 & 0.6065 & 0 & 0 \\ 0.0866 & 0.1425 & 0 & 0 \end{bmatrix}. \quad (4.5)$$

Unlike the undirected graph model whose the weighted adjacency matrix E is always *symmetric* (i.e., $w_{i,j} = w_{j,i}$), the weighted adjacency matrix E of directed graph model is, in general, *non-symmetric*; that is, $e_{i,j}$ is not necessarily equal to $e_{j,i}$. This would lead to the difficulty in solving the optimization problem on identifying the strongly-associated subgraph in our directed graph model that is to be discussed in detail in the following section.

4.5 Strongly-Associated Subgraph Identification

4.5.1 Problem Formulation for Identifying Strongly-Associated Subgraph

In general, a visual object commonly presented in the two images under comparison are acquired from two different viewpoints and oftentimes undergo a scale change. However, such scale change information is normally not available. Thus, a range of real-valued scale-change factors α , ranging from 0.25 to 4 as explained in Section 4.4, are exhaustively tried out in (4.2). For each assumed α under experiment, a new set of initial link weights $w_{i,j}$ for the entire graph model will be computed, followed by conducting the link-weight mapping from $w_{i,j}$ to the relative link weights $e_{i,j}$ via (4.3) and to $e_{j,i}$ via (4.4), respectively, for the establishment of our proposed digraph. For each generated digraph, the subgraph with the *largest* average link weight will be identified from the digraph. Furthermore, if this largest amount is

greater than a pre-defined threshold $\eta = 0.11$ as empirically determined in this chapter, the corresponding subgraph would be considered as a *strongly-associated* subgraph. All the vertices belonging to this subgraph will be collectively treated as a *common visual pattern*, and hopefully, they are all from the same visual object. On the other hand, if this largest amount is less than or equal to η , it is considered that there is no strongly-associated subgraph that can be identified and extracted at the given scale-change factor. In the following, we shall discuss how to identify the subgraph that yields the largest average link weight from each digraph. Suppose that there are L vertices belonging to a subgraph that have been identified from a digraph that contains K vertices in total. The *indicator vector* $\mathbf{x}$, which is a binary vector with the dimension of K , can be used to represent this subgraph. Each element $x_i \in \mathbf{x}$ indicates whether the i th vertex v_i of the digraph is considered as a vertex of the subgraph; if so, x_i is assigned with a value 1 or a value 0, otherwise. That is, if there are L elements in $\mathbf{x}$ that has been assigned with a value 1, the sum of all the vector components x_i of $\mathbf{x}$ should be equal to L . The purpose of exploiting the indicator vector $\mathbf{x}$ lies on the mathematical convenience for computing the average link weight $C = \frac{1}{L^2} \mathbf{x}^T E \mathbf{x}$ for a subgraph with L vertices; hence, there are L^2 link weights. Therefore, the problem of identifying the subgraph with the largest average link weight could be tackled by solving the following optimization problem with respect to $\mathbf{x}$:

$$\arg \max_{\mathbf{x}} C, \text{ subject to } \sum_{i=1}^K x_i = L \text{ and } x_i \in \{0, 1\}. \quad (4.6)$$

To solve the optimization problem in (4.6), the spectral method [12], which exploits the property of Rayleigh quotient, can be applied to the weighted adjacency matrix E , provided that it is symmetric. The optimal solution of $\mathbf{x}$ (denoted as $\mathbf{x}^*$)

becomes the eigenvector of the largest eigenvalue of matrix E . In [14], Liu *et al.* apply the replicator equation used in the evolution game theory to iteratively find the optimal solution $\mathbf{x}^*$. Furthermore, the issue of the convergence has been addressed in [88] for the symmetric and nonnegative matrix E only. For the conventional undirected graph model, this condition is easily satisfied since there is only one link weight on each link that guarantees the matrix E being *symmetric*. However, as pointed out in Section 4.4.3, the weighted adjacency matrix E resulted from our directed graph model is *non-symmetric*, and this makes both methods introduced in [12] and [14] totally not applicable. In the following, a numerical method for solving (4.6) that involves non-symmetric adjacency matrix E is proposed, which is based on *evolutionary stable strategy* (ESS) [84] of the non-cooperative game theory.

4.5.2 Proposed Numerical Method: An Evolutionary Approach

Note that in (4.6), each element x_i in $\mathbf{x}$ is an integer which only takes the value of either 0 or 1. From the standpoint of the optimization theory, Eq. (4.6) could only be solved by exhaustively trying all possible combinations of binary values (0 and 1) for $\mathbf{x}$ [89]. This kind of exhaustive search will result in extremely high computational complexity and becomes an NP-hard problem. To ‘bypass’ this computational issue, a *relaxation* process is proposed to solve (4.6) in [84] by introducing a dummy variable $\mathbf{y} = \frac{\mathbf{x}}{L}$ to re-express (4.6) as the optimization problem of solving $\mathbf{y}$:

$$\arg \max_{\mathbf{y}} C = \mathbf{y}^T E \mathbf{y}, \text{ subject to } \sum_{i=1}^K y_i = 1 \text{ and } y_i \in [0, 1]. \quad (4.7)$$

It should be pointed out here that both (4.6) and (4.7) are aimed to find the largest average link weight C . However, each element $y_i \in \mathbf{y}$ in (4.7) is not an integer but a real-valued number between 0 and 1. Due to this difference, it is further proved

in [84] that the optimization problem as formulated in (4.7) is no longer an NP-hard problem, and its optimal solution $\mathbf{y}^*$ can be efficiently obtained by applying an iterative method based on the ESS. However, as the number of vertices (i.e., L) in the strongly-associated subgraph is not known in advance, therefore, the optimal solution $\mathbf{x}^*$ to (4.6) can not be directly obtained from the optimal solution $\mathbf{y}^*$ of (4.7) via $\mathbf{x}^* = L\mathbf{y}^*$. To tackle this problem, a binarization operation applied to $\mathbf{y}^*$ is proposed in [84] with the purpose of finding a good approximation of $\mathbf{x}^*$. In fact, this binarization operation is a thresholding process: If the element $y_i^* \in \mathbf{y}^*$ is greater than or equal to a pre-defined threshold τ (where $\tau = 0.5$ as empirically determined in [84]), the corresponding element $x_i^* \in \mathbf{x}^*$ is set to 1, or to 0, otherwise. Now, the question on solving (4.6) is boiled down to how to solve the optimization problem in (4.7). According to [84], the optimization problem in (4.7) could be solved by exploiting the ESS. For that, some basic concepts and definitions of the ESS shall be introduced first in the following, followed by the detailed explanations on how to use them to solve (4.7).

Given a weighted adjacency matrix E , a differential model is used to depict the change of the state $\mathbf{z}$ of a biology system. The state transition from time t to time $t + 1$ during the evolution process could be established as follows [88]:

$$z_i(t + 1) = z_i(t) \frac{(E\mathbf{z}(t))_i}{\mathbf{z}(t)^T E\mathbf{z}(t)}, \text{ for } i = 1, \dots, K; \quad (4.8)$$

where $(E\mathbf{z}(t))_i$ is the i th component of the resultant vector by multiplying vector $\mathbf{z}(t)$ by the weighted adjacency matrix E . The K -dimensional vector $\mathbf{z}(t)$ is used to represent the state of the biological system at time t , and each of its component $z_i(t)$ represents a salient attribute of the biological system at that time (such as size, temperature, humidity of the biological system). Similar to the $y_i \in \mathbf{y}$ in (4.7), each component $z_i(t) \in \mathbf{z}(t)$ is a real number with the range of value $[0, 1]$,

and the sum of all the vector components of $\mathbf{z}(t)$ is equal to 1. Further note that, starting with some initial state $\mathbf{z}(0)$, once the state of the biological system reaches to an equilibrium state (i.e., Eq. (4.8) converges or remains unchanged) through the evolution process for a period of time t_s , the K -dimensional vector $\mathbf{z}(t_s)$ at the equilibrium state is called an *asymptotically stable stationary point*, which corresponds to the *maximizer* (i.e., the $\mathbf{y}^*$) of (4.7). In this case, the resultant average link weight C is the maximum and thus the solution to (4.7) is determined. This connection has been mathematically proven in [84] and stated as follows: if the weighted adjacency matrix E used to establish the differential model in (4.8) (and in (4.7) as well) is *non-negative*, the local maximizer of (4.7) is also the asymptotically stable stationary point of (4.8).

It is expected that the asymptotically stable stationary point as discussed above could be a *local* point (or the local maximizer of Eq. (4.7)), rather than a global one, since the initial state $\mathbf{z}(0)$ imposed for the computation of (4.8) was randomly generated. To increase the chance of obtaining global stationary point, a sufficient number of trials using different randomly-generated initial state $\mathbf{z}(0)$ can be considered. As suggested in [84], this number can be set to K (i.e., the vector dimension of $\mathbf{z}(t)$) to have a reasonable tradeoff between the performance and computational complexity. Finally, among all the identified stationary points computed from (4.8), the one that leads to the largest average link weight C is considered as the global maximizer, or the optimal solution $\mathbf{y}^*$ of (4.7).

4.5.3 Topological Splitting

It should be pointed out here that when multiple non-overlapping visual objects presented in two images undergo the *same* scale change with the same factor α as defined in Eq. (4.1), our proposed method developed up to this point would not be

able to further differentiate them as two different visual objects. For illustration, an example is given in Fig. 4.3(a), in which two compact disks (as visual objects) undergo the same scale change with the factor $\alpha = 0.625$, besides some translation and rotation. Since the goal of our proposed method at this stage is to find the most strongly-associated subgraph at each assumed scale-change factor (ranging from 0.25 to 4) only, the resultant feature-point correspondences would treat both compact disks as *one* visual object instead of two visual objects as indicated in Fig. 4.3(b). To overcome this limitation, an additional process is needed to further discriminate these visual objects. For that, a novel and yet computationally simple algorithm is proposed here, called the *topological splitting*, to determine whether there is a need to further partition each strongly-associated subgraph into multiple smaller subgraphs so that each subgraph corresponds to one visual object. The proposed topological splitting algorithm works as follows.

Since each SIFT feature descriptor is established based on one circular region or the so-called *supported region* [17], the motivation of our proposed topological splitting is to fully utilize such existing information obtained previously. For those feature points detected from the same visual objects, their associated supported regions tend to aggregate together and highly overlapped. Otherwise, these regions will be far apart, provided that the visual objects are also far separated. By utilizing these observations, our proposed topological splitting algorithm can be designed as follows: If the supported regions from any two vertices of the strongly-associated subgraph overlap with each other, these two vertices will be classified into the same smaller subgraph. Otherwise, these two vertices will be separated and classified into two different smaller subgraphs; for example, see Fig. 4.3(c). The limitation of this method lies in the assumption that different visual objects undergoing the same scale change should be spatially located far apart. If this assumption gets violated, our proposed method would be unable to discriminate the different visual

(a)

(b)

(c)

Figure 4.3. A demonstration of a limitation on discriminating two visual objects (i.e., compact disks) when they are undergoing the same scale change. (a) Two visual objects having the same scale-change factor $\alpha = 0.625$; (b) The obtained common visual pattern discovery result by using our proposed digraph approach and being viewed as *one* visual object. Note that the circles are the supported regions generated by the SIFT algorithm [17]; (c) *Two* visual objects are identified by further applying our developed topological splitting.

objects.

4.6 Experimental Results and Discussions

Extensive simulation experiments are conducted to comprehensively evaluate our proposed *digraph*-based approach on the discovery of common visual patterns of two images under comparison. Furthermore, adverse factors or ‘interferences’ such as geometric deformation and indiscriminate representation (or outlier) will be considered in our simulations for the study of performance evaluation. As pointed out in [12], for a pair of typical natural images, it is very difficult to quantitatively measure the amount of geometric deformation and the number of outliers contained within them. For that, Leordeanu *et al.* [12] proposed a simulation approach by generating two artificial point sets that simulate detected feature points as if they were generated by the SIFT algorithm. Various amounts of geometric deformation and different number of outliers are considered in this simulation. The ground truth of point-to-point correspondence pairs is obtainable and can be used to conduct objective performance evaluation. The experimental results of our proposed method are compared with those obtained by applying three comparable state-of-the-art *graph* approaches proposed by Leordeanu *et al.* [12], Liu *et al.* [14], and Cour *et al.* [63], respectively. In the following, we shall describe how to generate the artificial point sets in detail first.

4.6.1 Objective Evaluation: Experimental Setup and the Ground Truth

To demonstrate, a square-shaped visual object is presented in Image 1 of Fig. 4.4. By rendering this visual object with a certain amount of rotation and translation,

Figure 4.4. An illustration of two point sets A and B that are generated to simulate the SIFT feature points detected on the square-shaped visual object commonly presented in Image 1 and Image 2, respectively. Note that, the two point sets contain two categories of points: matched and mismatched. The former ones are represented by the red dots, while the latter ones are represented by the blue stars.

the same visual object is presented in Image 2 of the same figure. Two point sets, denoted as A and B , are generated to simulate two sets of SIFT-generated feature points as if they were independently detected in Image 1 and in Image 2, respectively. Further assume that there are M and N points in these two sets (i.e., $M = 9$ and $N = 11$ as shown in Fig. 4.4). Note that these points can be classified into two non-overlapping categories so far as the point-to-point correspondences are concerned: *matched* and *mismatched*. Since any pair of matched SIFT feature points must be an *one-to-one* correspondence, the total number of such points in Image 1 and in Image 2 must be identical (denoted as P). For the mismatched SIFT feature points due to indiscriminate representation (i.e., the so-called *outliers* in [12]), they could be one-to-many or many-to-one correspondences. Therefore, the total numbers of such points in the two images are oftentimes *not* identical; let these two numbers be denoted as Q_1 in Image 1 and Q_2 in Image 2, respectively. In sum, $M = P + Q_1$ and $N = P + Q_2$. Now, the question is boiled down to: How to generate these two categories of simulated feature points in order to reflect various degrees of interference as mentioned earlier? This shall be discussed in the following.

To generate the simulated ‘matched’ (i.e., the first-category) points, P points are

randomly chosen from the square-shaped visual object in Image 1 as if they were the actual feature points detected by the SIFT algorithm. Since the object's rotation angle and translation displacement are already determined for simulation and known in prior, the spatial coordinates of the corresponding positions of these P points located in Image 2 can be mathematically computed as the ground truth of the first-category points in Image 2. In this case, the pairwise spatial consistency principle will be guaranteed on these P pairs of point-to-point correspondences. However, in practice, the so-called 'matched' SIFT feature points in Image 2 could be easily deviated from the ground truth due to the geometric deformation mentioned in Section 4.3.2. To simulate this situation, a Gaussian white noise $N(0, \sigma)$ is added onto the ground truth to perturb the computed spatial coordinates using different values of σ to simulate the 'actual' spatial coordinates of the first-category points in Image 2. On the other hand, the second-category points can be simply generated by randomly selecting Q_1 and Q_2 points on the square-shaped visual object in Image 1 and in Image 2, respectively. For the example illustrated in Fig. 4.4, we have $P = 6$, $Q_1 = 3$, and $Q_2 = 5$; hence, $M = 9$, and $N = 11$.

By generating the point sets A and B using the above-mentioned simulation process, this will allow us to quantitatively measure the feature-point correspondence performance under the conditions of the above-mentioned two adverse factors—geometric deformation and indiscriminate representation (or outlier). By adjusting the value of the variance σ of the Gaussian white noise added to the ground truth, various amounts of the geometric deformation could be generated; the larger the value of σ is, the stronger the geometric deformation would be. As to the number of feature points simulated for the second category, they are the number of outliers in the points sets A (i.e., Q_1 in Image 1) and B (i.e., Q_2 in Image 2), respectively. The larger the values of Q_1 and Q_2 are, the more outliers would present in the two point sets.

4.6.2 Objective Evaluation: Experimental Results

Four experiments are conducted to comprehensively evaluate the performance of our proposed method based on the generated point sets A and B as described perviously. The number of correct point-to-point correspondences is fixed and set to 15 (i.e., $P = 15$). For simulating adverse effects, this can be achieved by setting different values to σ (for creating geometric deformation), Q_1 and Q_2 (for creating outliers). Our proposed method, Leordeanu *et al.*'s method [12], Liu *et al.*'s method [14], and Cour *et al.*'s method [63] are independently applied to each case of point sets and run for 30 independent trials each. For each trail, the number of correctly identified point-to-point correspondences by applying each method is recorded. By simply taking the average of these numbers recorded from the 30 tails, the quantitative performance evaluation result of each method applied to the current point sets could be obtained. As shown in Fig. 4.5, the blue line with dots, the green line with triangles, and the magenta line with diamonds denote the performance curves resulted by [12], [14], and [63], respectively, while the performance curve of our proposed method is represented by the red line with squares.

These four experiments are:

- *Experiment 1): Geometric deformation only*

In this case, the outlier issue is not encountered; thus, $Q_1 = Q_2 = 0$. That is, no second-category points will be added to the point sets A and B , respectively. The robustness of the four methods to the geometric deformation is evaluated by performing the experiments where the value of σ increases from 0 to 10 with the step size of 1 (i.e., 11 cases) to simulate various amounts of geometric deformation. As shown in Fig. 4.5(a), when a small and moderate amount of geometric deformation is encountered (i.e., $0 \leq \sigma \leq 6$), all the four methods could perform equally well with almost all the 15 correct

(a) Geometric deformation only (i.e., $Q_1 = Q_2 = 0$)

(b) Outliers only (i.e., $\sigma = 0$)

Figure 4.5. (Continue)

(c) A fixed number of outliers (i.e., $Q_1 = Q_2 = 15$) (d) A fixed degree of geometric deformation (i.e., $\sigma = 5$)

Figure 4.5. Performance comparison among Leordeanu *et al.* [12], Liu *et al.* [14], Cour *et al.* [63] and our proposed method. In all our experiments conducted, the number of correct correspondences is 15 (i.e., $P = 15$) as the ground truth. Multiple pairs of point sets A and B will be generated through simulation to incorporate various combinations of the geometric deformation and outlier that are often encountered in the natural images.

point-to-point correspondences being identified. However, when the amount of geometric deformation becomes large, say, $7 \leq \sigma \leq 10$, one can see that the proposed method consistently outperforms the other three methods.

- *Experiment 2): Outliers only*

In this case, the geometric deformation issue is ruled out; thus, $\sigma = 0$. That is, no Gaussian white noise will be added to the mathematically computed spatial coordinates (i.e., the ground truth). The robustness of the four methods to the outliers is evaluated by performing the experiments where the number of outliers Q_1 and Q_2 increases from 0 to 60 with the step size of 10. As shown in Fig. 4.5(b), our proposed method could identify almost all the 15 correct point-to-point correspondences without being affected by the different numbers of outliers ranging from 0 to 60. However, the same performance could be obtained by [12], [14], and [63] only when the number of outliers is small (e.g., for [12], the number of outliers needs to be less than 10, while for [14] and [63], the number of outliers needs to be less than 20.). As the number of outliers continues to grow, their performances drop drastically. When the number of outliers becomes large (i.e., $40 \leq Q_1 = Q_2 \leq 60$), the average numbers of correct point-to-point correspondences identified by [12], [14], and [63] have dropped to approximately 34%, 37%, 55% of that being identified by our proposed method, respectively.

- *Experiment 3): A fixed number of outliers with various amounts of geometric deformation*

In this experiment, the number of outliers is set to $Q_1 = Q_2 = 15$, and the value of σ increases from 0 to 10 with the step size of 1. As shown in Fig. 4.5(c), when a small and moderate amount of geometric deformation is encountered (i.e., $0 \leq \sigma \leq 6$), our proposed method is able to outperform

[12] and [63], while yielding similar performance as that of [14]. When the amount of geometric deformation becomes larger (i.e., $7 \leq \sigma \leq 10$), the superiority of our proposed method is more obvious—the average numbers of correct point-to-point correspondences identified by [12], [14], and [63] are only approximately 47%, 63%, 61% of that being identified by our proposed method, respectively.

- *Experiment 4): A fixed amount of geometric deformation with different numbers of outliers*

In this experiment, the amount of geometric deformation is fixed by setting $\sigma = 5$, and the values of Q_1 and Q_2 still increase from 0 to 60 with the step size of 10. As shown in Fig. 4.5(d), our method could produce similar performance to that of [14], when the number of outliers is less than 10. In all the other test cases, our proposed method substantially outperforms the methods proposed by [12], [14], and [63], on identifying correct point-to-point correspondences. For example, when $Q_1 = Q_2 = 40$, the numbers of correct point-to-point correspondences identified by [12], [14], and [63] are only approximately 40%, 47%, 62% of that being identified by our proposed method, respectively.

4.6.3 Performance Evaluation with Natural Images

In this section, the performance evaluation of our proposed method is conducted and further compared with Liu *et al.*'s method based on six thoughtfully chosen image pairs as shown in Fig. 4.6. For both methods, SIFT was used to generate the local feature descriptor for each detected feature point. After independently applying each method, the entire set of point-to-point correspondences as identified at each assumed scale change factor across the test image pair is viewed as *one* common visual pattern. For presentation, successfully identified common visual

(a) Liu *et al.*'s method applied to toy 1

(b) Our proposed method applied to toy 1

Figure 4.6. (Continue)

(c) Liu *et al.*'s method applied to book

(d) Our proposed method applied to book

Figure 4.6. (Continue)

(e) Liu *et al.*'s method applied to game disk

(f) Our proposed method applied to game disk

Figure 4.6. (Continue)

(g) Liu *et al.*'s method applied to magazine

(h) Our proposed method applied to magazine

Figure 4.6. (Continue)

(i) Liu *et al.*'s method applied to toy 2

(j) Our proposed method applied to toy 2

Figure 4.6. (Continue)

(k) Liu *et al.*'s method applied to toy 3

(l) Our proposed method applied to toy 3

Figure 4.6. Common visual pattern discovery results based on six thoughtfully-chosen image pairs. In general, (a)-(f) tend to yield more *outliers* due to a severely cluttered background, while (g)-(l) contain more *geometric deformations* due to large view point variations.

patterns are discriminated by using different colors. Hopefully, each pattern labeled with the same color indeed corresponds to a true visual object that is semantically meaningful to the human being. Further insights and discussions regarding Fig. 4.6 are provided as follows.

Note that for the three test image pairs in Fig. 4.6(a)-(f), each visual object commonly shared by the two images under test only undergoes a translation, rotation and scale change without any viewpoint variation. For example, the beverage bottle in Fig. 4.6(a) and Fig. 4.6(b) undergoes such similarity transform only. Consequently, one of the two adverse circumstances—geometric deformation caused by the viewpoint variation is not encountered in this case. However, since all the image contents present significant amount of cluttered background, it is expected that the resulted initial feature point matching pairs in set V (define in Section 4.4.1) should contain a large number of incorrect matches. Consequently, many outliers are encountered in all three image pairs. As shown in Fig. 4.6(b), Fig. 4.6(d) and Fig. 4.6(f), our method could successfully discover multiple common visual patterns of the visual objects in all three test image pairs, while Liu *et al*'s method produces one incorrect common visual pattern as indicated by the black circles in Fig. 4.6(e).

For the remaining three test image pairs in Fig. 4.6(g)-(l), all the visual objects commonly shared by the two images under test undergo significant viewpoint variations. For this kind of test image pairs, both adverse circumstances (i.e., geometric deformation and outlier) could be easily encountered that lead to severe spatial inconsistency. As illustrated in Fig. 4.6(h), Fig. 4.6(j) and Fig. 4.6(l), our method could still successfully discover multiple common visual patterns in each of the three test image pairs. However, for Liu *et al*'s method, the common visual pattern indicated by the black circles in Fig. 4.6(i) and Fig. 4.6(k), respectively, is obviously incorrect.

4.6.4 Near-duplicate Image Retrieval

In this section, the proposed digraph model-based method is further applied to the task of near-duplicate image retrieval, which is beneficial to many real-world multimedia applications, such as video search, copyright infringement detection, to name a few. The performance of our proposed method is evaluated on the well-known Columbia dataset [51], which contains 150 near-duplicate image pairs and 300 non-duplicate images with the same size of 352×264 . The performance of our proposed method is further compared with that of Liu *et al.*'s method [14] and two other state-of-the-art near-duplicate image retrieval techniques proposed by Wu *et al.* [90] and Zhu *et al.* [91], respectively.

Note that the evaluation method proposed in [92] is adopted in this section to conduct the simulation experiments. To be more specific, each image in 150 near-duplicate image pairs (or 300 images in total) will be independently treated as a query image. The retrieval task is to find the relevant images in remaining 599 images and rank them according to the their similarity towards the query image. For our proposed digraph-model based method, this similarity is measured by the total number of point-to-point correspondences established on the common visual patterns of the query image and the retrieved image. The larger this number is, the two images are more likely to be near-duplicate image pair.

As pointed out in [91], two stages of image retrieval are utilized to achieve better performance: 1) coarse retrieval stage based on the global feature (297-dimensional feature descriptor); 2) refined retrieval stage based on the local feature, such as SURF feature descriptors used in [91]. Since our proposed method is based on SIFT feature descriptor, which is a local feature in essence. For fair comparison, the same coarse retrieval stage as stated in [91] is also used before applying our proposed method.

Figure 4.7. Performance comparisons among Liu *et al.*'s method [12], Wu *et al.*'s method [90], and Zhu *et al.*'s method [91] on Columbia dataset for near-duplicate image retrieval.

Finally, the *cumulative accuracy* versus *rank* curve is utilized to objectively evaluate the performances of different near-duplicate image retrieval algorithms. The cumulative accuracy measures the percentage of successful retrieval (or finding the correct near-duplicate image) among all the 300 queries. Rank is the number of images retrieved by each algorithm. As demonstrated in Fig. 4.7, our proposed method consistently outperforms all the other three methods under the evaluation. This observation further verifies that our proposed method could reliably identify the common visual patterns co-existed in the natural image pairs.

4.7 Summary

In this chapter, a novel directed graph (or digraph) model-based method is proposed to discover the visual patterns commonly presented in the two images under comparison. Each SIFT-descriptor-based point-to-point correspondence is established as a vertex, and the key success of our approach lies at the consideration of the neighboring vertices for each vertex of the link to quantitatively reflect the degree of association among the vertices more accurately. For that, instead of assigning a single initial link weight to each link of the graph, two *relative* link weights are generated for each direction of the link by exploiting our developed *n-ranking* process (i.e., a generalized median measurement) and a novel link weight enhancement technique. Consequently, the resultant two link weights are more robust to combat various interferences commonly encountered in common visual pattern discovery application, such as scale change, viewpoint variation, and indiscriminate feature representation. A common visual pattern is then equivalent to the strongly-associated subgraph identified from each directed graph newly established at an assumed scale-change factor by applying the non-cooperative game theory that is able to handle the non-symmetric weighted adjacency matrix encountered in our approach. In case there are multiple visual objects undergo the same scale-change factor across two images, our proposed topological splitting algorithm might be able to further discriminate them, provided that these objects are not too close to each other.

To evaluate and demonstrate the efficacy and robustness of our proposed method on discovering common visual patterns, the performance of our proposed method is evaluated and further compared with the state-of-the-art algorithms [12] [14] [63] [90] [91]. For standing out the above-mentioned adverse effects, artificially generated point sets that simulate the SIFT feature points are used to conduct the

objective evaluation first, followed by conducting the performance evaluation on six thoughtfully chosen natural image pairs and Columbia dataset, respectively. Extensive simulation results obtained in both performance evaluation stages clearly demonstrate that our proposed method is more capable to handle the adverse circumstances, and leads to accurate discovery of the common visual patterns.

Chapter 5

Feature Histogram Equalization: A Feature Contrast Enhancement Technique

Different from the pervious two chapters which aim to refine the initial feature point matching pairs according to the application's requirements, this chapter will be focusing on how to increase the precision of the initially established SIFT-based point-to-point matching pairs without any further refinement process. To achieve this goal, a novel feature contrast enhancement technique, called *feature histogram equalization* (FHE), is proposed. The fundamental idea of the FHE is to independently re-distribute and modify the vector-component values of feature descriptors along each vector dimension for better discrimination; consequently, the feature contrast is enhanced and the precision of the initial feature point matching pairs could be improved.

The remaining of this chapter is organized as follows. In Section 5.1, an overview of the existing techniques to improve the precision of the initial feature point matching pairs is presented. In Section 5.2, the concept of the proposed feature contrast

is introduced, followed by a detailed explanation regarding why the feature contrast might need to be enhanced for better discrimination of the feature descriptors. Section 5.3 describes how to form a set of component-wise feature histograms based on all the feature descriptors obtained from the the two images under matching. Section 5.4 explains how to conduct feature histogram equalization for increasing feature contrast at length. In Section 5.5, extensive simulations are conducted to evaluate the performance of the proposed FHE on the SIFT feature descriptors to demonstrate its efficacy in improving the accuracy of the initial feature point matching pairs.

5.1 Background and Literature Review

In order to improve the precision of the initial feature point matching pairs, several local feature enhancement techniques have been recently introduced. These techniques can be broadly categorized into two approaches—one is on the enhancement of feature point detection and local feature descriptor computation (for the first stage in Fig. 1.2), and the other is on the improvement of local feature descriptor matching (for the second stage in Fig. 1.2).

The published results of the first approach are fairly extensive and can be traced a long way back in the open literature. As discussed in Chapter 1, the establishment of local feature consists of two steps: 1) identifying the regions of interest using a certain feature detector, 2) generating a feature descriptor based on each detected regions. For the first step, one of the earliest and most well-known feature detector is Harris corner detector [29]. Mikolajczyk *et al.* [34] provided a comprehensive evaluation of the most popularly used feature detectors and came to the conclusion that different feature detectors show their complementary properties and the so-called all-purpose detector didn't exist. More recently, *features from accelerated segment*

test (FAST) criterion [93] for feature detection has drawn increasing attention in many state-of-the-art methods that require the real-time implementations. Mair *et al.* [94] further extended Rosten *et al.*'s research work [93] and proposed *adaptive and generic corner detection based on the accelerated segment test* (AGAST) for a better performance in the corner detection.

For the second step regarding the generation of the feature descriptors, *scale invariant feature transform* (SIFT) [17] was ranked as the best feature descriptor in a survey [35] for its highly discriminative power and robustness to viewpoint variations and illumination changes. There are two research directions to improve the SIFT feature descriptor. One is to generate new robust feature descriptors that are invariant for more severe geometric transformations and illumination changes. For example, the *affine-SIFT* feature descriptor proposed by Morel and Yu [95] improved the SIFT feature descriptor on handling the affine transformation incurred between the same object presented in the two images under matching. This was achieved by adding in two additional parameters to reflect the amount of viewpoint changes. The values of these two parameters were determined by conducting multiple experiments of initial feature point matching using different parameter values over a range of possible values, from which the parameter values that generate the most accurate initial feature point matching pairs were selected as the final ones.

Another research direction is to develop new feature descriptors that can be computed more efficiently for faster initial feature point matching with little compromise on the performance of SIFT feature descriptor. PCA-SIFT [96] exploited the well-known dimensionality reduction technique, *principle component analysis* (PCA), to reduce the dimension of the SIFT feature descriptor from 128 to 36. However, the reduced time in the establishment of the initial feature point matching pairs was almost offset by the increasing time spent on PCA in forming the PCA-SIFT feature descriptors. Alternatively, *speeded up robust feature* (SURF) [21] utilized

integral images to speed up the computation of convolutions to establish SIFT-like local gradient histograms. The resulted 64-dimension SURF feature descriptors were proven to yield comparably accurate initial feature point matching pairs as the SIFT feature descriptors did by the simulation results in [21].

For the second approach that aims to improve the feature matching criterion, Cao *et al.* [97] proposed to exploit the Chi-square distance instead of using the Euclidean distance to measure the similarity of two feature descriptors. Simulation results reported in [97] have shown that more accurate initial feature point matching pairs were established. Dame *et al.* [98] further proposed to utilize another metric, called *mutual information* (MI), as the feature matching criterion. Consequently, the proposed method was far less sensitive to the illumination variations and the non-linear geometric transformation incurred between the two images under matching.

Completely different from the existing methods, a novel approach for the enhancement of feature-descriptor representation is proposed, which is in essence a *post-processing* method. That is, given a set of feature descriptors that have been generated by any feature descriptor generation algorithm (say, the SIFT technique as adopted in this thesis), are we still able to further modify them for improving their feature contrast (or, equivalently, increasing the vector distances among themselves for improving their discriminative ability)? For that, *feature histogram* and *feature histogram equalization* (FHE) are proposed in this thesis for the first time, as follows.

The fundamental idea of the proposed FHE is in parallel with that of the *image histogram equalization* (IHE). While the IHE is exploited to re-distribute image intensity values over the image's full dynamic range to better visualize the image contents of low-contrast images (refer to Figure 5.1 for a demonstration), the FHE modifies the vector-component values of feature descriptors to increase the feature

Figure 5.1. A gray-scale image (534×800) and its histogram are shown in (a) and (b), respectively. After applying the *image histogram equalization* (IHE) process to (a), the equalized image and its histogram are shown in (c) and (d), respectively. Note that the IHE stretches the original image intensity values over its full dynamic range, from 0 to 255, for an 8-bit gray-scale image.

contrast. This is especially beneficial for those images with low feature contrast. That is, the IHE is beneficial for improving visualization of image objects, while the FHE is for better discriminating the detected feature points.

To demonstrate the efficacy and effectiveness of the proposed feature contrast enhancement, we begin with the generation of SIFT descriptors [17], followed by applying the developed FHE process to *equalize* them. That is, all the involved SIFT descriptors are subjected to be equalized by the developed FHE algorithm for yielding a ‘better’ representation in the sense that the *equalized* SIFT descriptors are

more discriminative from one to another (i.e., yielding stronger feature contrast). This equalization process is conducted along each vector dimension (called *feature slice* in our work), independently.

As to the targeted application system, we focus on the *feature-point correspondence*¹, which attempts to establish as many correct point-to-point correspondence pairs as possible between two images with similar image contents under matching (e.g., [17] [96] [31] [35]). The main reason that we consider this system application is due to the fact that feature-point correspondence is commonly used by many image processing and computer vision related tasks, such as wide-baseline image matching [99] [20], object recognition [10] [11], common visual pattern discovery [12] [13] [14], and three-dimensional scene reconstruction [6] [7], to name a few. Extensive simulation results will show that the proposed FHE is a simple and effective tool for the enhancement of feature contrast.

5.2 What is Feature Contrast?

Feature contrast is about the image content from the ‘high-level’ standpoint, but through the processing of feature descriptors at the ‘low-level’ representation. To appreciate this newly-introduced concept, the best way is to begin with two opposite scenarios—one demonstrates *low* (or poor) feature contrast as shown in Figure 5.2, and the other is shown in Figure 5.3 for appreciating what is *high* (or strong) feature contrast.

In Figure 5.2(a), the image pair “Wall” contains many repeated patterns (i.e., bricks) under matching for the establishment of point-to-point correspondence pairs. Clearly, such kind of image is inherited with strong *self-similarity* in its image

¹This has been commonly referred to as ‘feature correspondence’ in many existing works; in our view, this is vague and a misnomer, in fact. Although each detected feature point has an associated feature descriptor, however it is the *points* are matched, not *features*.

Figure 5.2. A demonstration of *low* feature contrast through the image pair “Wall.” A large number of repeated patterns (i.e., bricks) are presented in the image pair (a) and (b); this shows that the image content of “Wall” is inherited with strong *self-similarity*. To further appreciate this aspect, a set of zoom-in pictures of several local regions centered on various feature points, including the point “A” in (a), are shown in (c) and (d), respectively. Note that all the regions are re-sized to 85×85 for a better visualization.

content, and obviously this yields a fairly poor *feature contrast* and less distinctive feature descriptor representation. To conduct feature-point correspondence on such kind of image pair, one can see that many regions of interest (as highlighted by the yellow circles in Figure 5.2(b)) look fairly similar to that of the feature point “A” as labeled in Figure 5.2(a), for example. Such strong self-similarity can be further verified through their zoom-in images as presented in Figures 5.2(c) and Figure 5.2(d), respectively. Therefore, it is expected that the feature descriptors generated in these feature-point locations will be quite similar to each other and hard to be differentiated among themselves. Consequently, it becomes quite difficult to determine which feature point in Figure 5.2(d) should be linked to the feature point “A” in Figure 5.2(c) as a correct correspondence pair.

Compared with the “Wall,” the image pair “Graf” as shown in Figures 5.3(a) and

Figure 5.3. A demonstration of *high* feature contrast through the image pair “Graf.” One can observe that the “Graf” is full of distinctive shapes and regions and therefore presents much less degree of self-similarity, compared with that shown in Figure 5.2. Similarly, a set of zoom-in pictures of several local regions centered on various feature points, including the point “B” in (a), are shown in (c) and (d), respectively. Note that all the regions are re-sized to 85×85 for a better visualization.

5.3(b), on the other hand, does not possess such strong self-similarity in their image content, since the “Graf” image contains many distinctive shapes and regions. In this case, it is quite obvious that “Graf” possesses strong *self-dissimilarity*. Indeed, most regions of interest detected in “Graf” are quite different from one to another, as indicated in Figure 5.3(d). This makes the follow-up point-to-point correspondence matching process across the image pair become much easier on deciding which feature point in Figure 5.3(d) should be made a correspondence link to the feature point “B” in 5.3(c), for example. All these shed the light that the feature contrast of the image “Graf” is much higher than that of the image “Wall,” and why the coined feature contrast is of great importance.

It is important to note that while the IHE is exploited to enhance low contrast images for improving the *visibility* of image contents, our proposed FHE, on the other hand, has a totally different objective: the FHE is used to stretch the vector

distances among the involved feature descriptors further apart from each other for improving *discrimination* among the involved feature descriptors. Consequently, the feature contrast is effectively enhanced, and it is beneficial to the follow-up processing tasks that involves pattern discrimination in essence. More experimental simulation results and discussions will be provided in Section 5.5 on the feature-point correspondence application using several commonly-used test image pairs, including “Wall” and “Graf” demonstrated here.

5.3 Feature Histogram

It should be borne in mind from the outset that the proposed FHE is *not* the well-known IHE under a different disguise, nor a direct exploitation of it for our demonstrated feature-point correspondence application either; in fact, they are totally different in multiple accounts, starting from their main objective as previously highlighted—the IHE is purely for image visualization purpose, and the proposed FHE is developed as a generic *feature enhancement* tool for potential applications in the future. Other fundamental differences between the IHE and the FHE will be made more clear along the way. In this section, we shall expound on the very first question—What is the *feature histogram*? This will be defined, and how to construct the histograms will be developed in the framework of feature-point correspondence application that aims to establish point-to-point correspondences between two similar images, as follows.

Given a pair of images under feature-point correspondence matching process, feature descriptors will be independently generated for each image using the well-known SIFT method [17]. Further suppose there are M feature descriptors being generated from both images altogether. Obviously, M is a variable constant, depending on the image content—a more complicate image content tends to yield a

larger value of M ; on the other hand, images with ‘simpler’ image content will have a smaller value of M . Unlike the image-intensity value, which is a *scalar*, each feature descriptor is an N -dimensional *vector* (e.g., there are 128 vector components in a SIFT feature descriptor; thus, $N = 128$). Therefore, it is impossible to exploit the concept and methodology of the image histogram directly to the feature descriptor. In our case, N feature histograms will be independently constructed—one for each vector dimension across all the M feature descriptors involved. This shows the first fundamental difference between the image histogram and the feature histogram—while only a *single* histogram is formed to represent the contrast of image intensity, there will be N feature histograms to collectively represent the contrast of the total M ’s N -dimensional feature descriptors. In the following, how to generate one feature histogram shall be explained in detail.

For ease of presentation, let the i th feature histogram (for $1 \leq i \leq N$), denoted as h_i , represent the statistical distribution of the i th vector dimension of the feature descriptors. All the i th component from each of M feature descriptors forms a set C_i , which is denoted as the i th *feature slice*. That is, each element $C_i(j)$ in the set C_i is the i th vector component of the j th feature descriptor (where $1 \leq j \leq M$); it is a scalar and real-valued. Similar to the image histogram, the feature histogram h_i is the statistical distribution of C_i . However, unlike the image intensity which is an *integer* value, each element in the feature slice C_i is a *real* number. Therefore, for each vector dimension or for the feature slice C_i , two challenges are further imposed on the formation of the feature histogram: 1) how to identify the full dynamic range, and 2) how to determine the number of histogram bins. These are further detailed in the following.

In the conventional image histogram, the full dynamic range is determined by the total number of distinctive image pixel values that could be possibly encountered in an image random field; that is, this is solely determined by the bit length chosen

for image representation, and it is totally independent from the image content [50]. Therefore, for all the images with the same bit-length representation, their full dynamic range is *constant*. For example, the full dynamic range for an 8-bit gray-scale image is $\{0, 1, \dots, 255\}$. However, as mentioned previously, each element in the feature slice is a *real* number; therefore, the collection of distinctive values in the feature slice should be $[0, +\infty)$, which is obviously not a ‘proper’ full dynamic range to be used for the feature slice. To solve this problem, it is proposed in this thesis that the full dynamic range for the feature slice C_i is between the minimum value and the maximum value of the elements in the feature slice C_i . According to this definition, the full dynamic range for different feature slices will be different. This can be considered as the second fundamental difference between the image histogram and the feature histogram.

Regarding the number of histogram bins as the third fundamental difference between the image histogram and the feature histogram, the histogram bins in the image histogram is simply set to the number of possible pixel values; thus, it is set to 256 bins for an 8-bit gray-scale image, for example. However, for the feature histograms, this becomes a quantization problem that one can fix either the bin number or the bin width, given a determined dynamic range. In this thesis, we choose the former, mainly due to the dynamic range of feature slices as defined above are not constant, besides fixing the bin number for each slice has advantage on ease of implementation and practical use. For that, the bin number has been empirically determined as $L = 50$ and constantly exploited in each feature slice. Now, the 128 feature histograms can be explicitly constructed as follows. For $i = 1, 2, \dots, N = 128$ and $k = 1, 2, \dots, L = 50$,

$$B_i(k) = \begin{cases} [\min(C_i) + (k - 1)\Delta_i, \min(C_i) + k\Delta_i], & \text{if } 1 \leq k < L; \\ [\min(C_i) + (L - 1)\Delta_i, \max(C_i)], & \text{if } k = L; \end{cases} \quad (5.1)$$

Algorithm 2 Computing N 's Feature Histograms

INPUT: The M 's feature descriptors with N -dimension each extracted from the two images, and the number of feature histogram bins L .

OUTPUT: The N 's feature histograms $h_1, \dots, h_N$.

{Note that $N = 128$ (for SIFT), $L = 50$, and M is a variable constant depending on the image pair experimented.}

for $i = 1$ to N **do**

 Calculate the bin width Δ_i for the feature histogram h_i by (5.1).

for $k = 1$ to L **do**

 Establish the k th bin $B_i(k)$ by (5.1).

 Calculate $r_i(k)$, which is the number of the elements in C_i that belong to $B_i(k)$.

 Calculate the k th element of the feature histogram h_i by (5.2).

end for

end for

$$\text{where } \Delta_i = \frac{\max(C_i) - \min(C_i)}{L}.$$

For the i th feature slice, $B_i(k)$ and Δ_i are introduced here to represent the k th bin and the width of each bin, respectively. Functions $\max(\cdot)$ and $\min(\cdot)$ are used to identify the maximum value and the minimum value from slice C_i , respectively.

Finally, the feature histogram h_i of the i th feature slice C_i with the total number of L bins could be established as a discrete function as

$$h_i(k) = \frac{r_i(k)}{M}, \quad (5.2)$$

where $r_i(k)$ is the cardinality number of the k th bin $B_i(k)$. Note that the total number of the elements in C_i is M . Therefore, $r_i(k)$ needs to be divided by M so that the sum of all $h_i(k)$ (for $1 \leq k \leq L$) of the feature histogram is equal to 1; i.e., $\sum_{k=1}^L h_i(k) = 1$. Therefore, h_i is the *probability density function* (pdf) of the statistical distribution of the total M 's element values from feature slice C_i . The computation of the feature histogram is summarized in Algorithm 2.

Comments: Unlike the formation of conventional image histogram with one image as the input and one histogram being produced at the output, *two* images are involved as the input in our chosen feature-point correspondence application that

leads to 128 histograms being generated at the output as above-mentioned. In this case, each histogram is computed along one feature vector dimension, or equivalently speaking, along one feature slice. However, one should note that the proposed feature histogram and its equalization are generic concepts and methodology, and these might need to be modified for various applications. For example, in our chosen feature-point correspondence application, the feature descriptors obtained from two input images must be *jointly* processed on the formation of 128 feature slices and their follow-up feature histograms, respectively. In fact, our investigation on processing the two images *separately* has yielded a much inferior performance on point-to-point feature-point correspondence results. One intuitive explanation to these outcomes is that: Since equalization can be viewed as a certain ‘normalization’ process, a joint treatment of the two images under matching is equivalent to putting them on the same ‘scale’ for normalization. On the contrary, if these two images are processed separately in the formation of feature histograms, this has an effect that these two images are normalized on two different scales. Consequently, this makes point-to-point feature-point correspondence matching becomes meaningless and leads to a much inferior performance as expected. In the following, we shall discuss how to conduct feature histogram equalization in detail.

5.4 Feature Histogram Equalization (FHE)

As mentioned previously, the sole objective of the IHE process is to stretch the dynamic range of image intensity values to its *full* range (e.g., from 0 to 255 for an 8-bit gray-scale image) for the purpose of *visualizing* low-contrast image contents. Since the FHE has a totally different objective and the dynamic range is real numbered and varies from slice to slice, the design methodology of the FHE process is not so straightforward. In fact, how to conduct an effective (hopefully,

optimal) equalization process for feature descriptors is an open question, and, in fact, a challenging task. For that, we have observed that there are two contradictory objectives need to be addressed on the design of FHE process, as follows.

Similar to the image histogram (refer to Figure 5.1), general speaking the more evenly the feature histogram is distributed, the higher the contrast of the feature descriptors would be resulted. After equalization, the resulted feature histogram in each feature slice should be more ‘flat’ and thus closer to a uniform distribution. But, this mismatches to the technical nature and requirements of the feature histograms. Simply imagine that when all the *equalized* feature histograms become too much the same as a uniform distribution, the shape or contour of the original feature histogram (such as peaks and valleys as their characteristics or signatures will be significantly changed and even completely lost. Such *over-equalized* feature descriptors are undesired, as it will make discrimination among them becoming more difficult. In fact, this will inevitably deteriorate the performance of feature-point correspondence application. To solve this dilemma, the i th *equalized* feature histogram (denoted by $\hat{h}_i$) in one hand should be stretched as much it can (thus, more close to a uniform distribution, denoted by u), but on the other hand, $\hat{h}_i$ should maintain sufficient similarity to the original feature histogram h_i ’s contours. To simultaneously satisfy these two contradictory objectives, a Lagrange multiplier optimization formulation, which is inspired by an image histogram modification proposed in [100], is established as follows:

$$\hat{h}_i = \arg \min_h \|h - h_i\|_2^2 + \lambda \|h - u\|_2^2 \quad (5.3)$$

where $\|\cdot\|_2$ denotes the l_2 norm and the Lagrange multiplier λ regulates the above-mentioned two objectives. Note that if $\lambda = 0$, $\hat{h}_i = h_i$ is obviously the solution, since the l_2 -norm distance is zero in this case. This is equivalent to say that ‘no

equalization' is applied to the feature descriptors (i.e., $\hat{h}_i = h_i$); thus, the feature contrast remains unchanged. On the other hand, when λ is large, the last term of (5.3) dominates the l_2 -norm distance measurement; hence, making histogram h more close to the uniform distribution u will help to minimize the total l_2 -norm distance in (5.3). In this case, the i th equalized feature histogram tends to be *over-equalized*, as discussed earlier. Therefore, there exists an optimum value of λ over the range of $[0, +\infty)$ that could provide the best trade-off between the two objectives mentioned previously.

In order to find a closed-form analytical solution of (5.3), it could be alternatively written in a quadratic form as

$$\hat{h}_i = \arg \min_h [(h - h_i)^T (h - h_i) + \lambda (h - u)^T (h - u)], \quad (5.4)$$

which can be easily solved by setting the first derivative of the quadratic term to 0, and the solution of (5.4) can be arrived at

$$\hat{h}_i = \frac{1}{1 + \lambda} h_i + \frac{\lambda}{1 + \lambda} u. \quad (5.5)$$

From (5.5), one can observe that the i th equalized feature histogram is a weighted combination of the original feature histogram h_i and a uniform distribution u .

Since the optimal λ is hard to derive, we resort to experimental approach instead. For that, different λ values ranging from 0 to 10 with the step size of 1 are individually experimented in this thesis. Base on these experimental results, the best performance is incurred when $\lambda = 2$. By substituting the determined $\lambda = 2$ into (5.5),

$$\hat{h}_i = \frac{1}{3} h_i + \frac{2}{3} u. \quad (5.6)$$

In order to perform the equalization process, the *cumulative distribution function* (cdf) of the feature histogram needs to be introduced here. As pointed out in Section 5.3, the feature histogram is the probability density function of the statistical

distribution of the total M 's element values from each feature slice C_i . Therefore, the corresponding cdf H_i of the i th original feature histogram h_i and the cdf $\hat{H}_i$ of the i th equalized feature histogram $\hat{h}_i$ could be defined, respectively, as follows. For $k = 1, 2, \dots, L$,

$$H_i(k) = \sum_{j=1}^k h_i(j), \quad (5.7)$$

$$\hat{H}_i(k) = \sum_{j=1}^k \hat{h}_i(j). \quad (5.8)$$

Based on (5.6), $\hat{h}_i$ can be obtained from h_i and u ; thus, $\hat{H}_i$ can be computed from (5.8). Hence, the equalization process boils down to finding a transform function (denoted by g) between $H_i(k)$ and $\hat{H}_i(k)$ by histogram specification [50]. For $m, k = 1, 2, \dots, L$, the transform function g is well-known and can be easily arrived at

$$m = g(k) = \hat{H}_i^{-1} [H_i(k)], \quad (5.9)$$

where $\hat{H}_i^{-1}(\cdot)$ is the inverse function of $\hat{H}_i(\cdot)$. That is, the equalization operation first computes the bin index m of the new (i.e., equalized) histogram via (5.9), followed by computing the value of $\hat{C}_i(j)$ according to

$$\hat{C}_i(j) = C_i(j) + (m - k)\Delta_i, \text{ for } C_i(j) \in B_i(k). \quad (5.10)$$

The developed FHE algorithm is summarized in Algorithm 3.

Algorithm 3 Feature Histogram Equalization

INPUT: The N 's feature histograms $h_1, \dots, h_N$ individually formed from feature slices $C_1, \dots, C_N$ with M 's elements each, which is the total number of feature descriptors extracted from two images.

OUTPUT: The N 's equalized feature slices $\hat{C}_1, \dots, \hat{C}_N$.

for $i = 1$ to N **do**

 Calculate the i th equalized feature histogram $\hat{h}_i$ by (5.5).

 Calculate the cumulative distribution function H_i and $\hat{H}_i$ of h_i and $\hat{h}_i$ by (5.7), respectively.

for $k = 1$ to L **do**

 Determine the bin index m mapped from the bin index k by (5.9).

for $j = 1$ to M **do**

 Calculate $\hat{C}_i(j)$ by (5.10).

end for

end for

end for

5.5 Experimental Results

5.5.1 Applying the Proposed FHE to Feature-Point Correspondence Application

To demonstrate its performance and potential, the proposed FHE is applied to feature-point correspondence application in this thesis, while this generic ‘tool’ can be exploited in any feature-descriptor-based applications. The objective of feature-point correspondence is to establish as many point-to-point correspondence pairs between two images under matching as possible. To be more specific, each correspondence pair is a link, connecting a specific location of a feature point (or an interest point) identified from one image to another location on the other image, to denote that these two image locations are, in fact, the same point or location from the standpoint of ‘high-level’ image contents. It is expected that, for those image pairs with strong self-similarity in their image contents (i.e., in poor feature contrast, such as the ‘brick’ patterns presented in Figure 5.2), the proposed FHE could effectively increase their feature contrast and thus yield much appreciable

performance gain in terms of establishing more correct point-to-point correspondence pairs. On the contrary, if the contents of two images are already in high feature contrast (as previously demonstrated in Figure 5.3), the performance gain contributed by the FHE should be small and much less appreciable, since the established feature descriptors are already fairly distinct from each other. In this case, the discrimination among these feature descriptors becomes quite easy, and many correct point-to-point correspondence pairs have been established in the first place already.

One should bear in mind that, in any case, the system performance resulted by exploiting the proposed FHE should *not* become inferior to that without applying the FHE, to say the least. This trend is very much the same as that of applying the IHE process to stretch the dynamic range of a given gray-scale image. That is, distinct contrast improvement can only be appreciated when the equalization process is applied to low-contrast images; the lower the image contrast, the stronger of such appreciation. For those images that are already in high contrast, applying the IHE won't make appreciable or any improvement on image contrast. More details are presented in this section as follows.

Given a pair of images under matching, the *affine-invariant Harris detector* [31] is first used in this thesis to find the *feature points* (or the so-called 'interest points') on each image independently; the positions of these points, in fact, locate at the center of those identified elliptical-shaped image regions (or the so-called *support region*), respectively. The SIFT feature descriptors [17] are then independently generated at each feature point based on all the image pixels covered by its associated elliptical-shaped support region, respectively. In our focused feature-point correspondence application, each feature point detected from one image is subjected to identify its corresponding feature point on the other image to establish the correspondence pair. This process is normally done by exhaustively searching all possible

pair-wise combination of feature points across the two images under matching, based on the vector-distance measurement of their associated feature descriptors. At the end of the search, the one that yield the shortest Euclidian distance is considered as the correct matched pair.

5.5.2 Simulation Setup and the Ground Truth Generation

Whether an established correspondence pair is a correct one or not can only be answered provided that the ground truth is made available. For that, a public domain dataset [35] that could be utilized to generate the ground truth is used to conduct objective performance evaluation. This dataset consists of 48 images in total, covering eight different scenes (i.e., there are six images from each scene category). The first image (denoted as I_1) in each scene category is served as the ‘reference’ image, while the remaining five images (i.e., $I_2, \dots, I_6$) are generated by successively varying the image contents by imposing a certain type of geometric transformation or some illumination changes only. More specifically, viewpoint changes are made for test image pairs “Graf,” “Wall,” and “Bark;” zoom and rotation for “Boat;” illumination change for “Leuven;” blurring effect for “Bikes” and “Tree;” and distortion incurred by JPEG compression for “Ubc.” Based on these data, the ground truth can be established through computational photography, as follows.

In the computational photography, the spatial location of an image pixel is represented by a three-dimensional (3-D) *homogeneous coordinate system* [101] that can be obtained from the two-dimensional (2-D) Cartesian coordinate system by simply inserting an additional dimension and filled with the value of 1. For example, for an image pixel at the location of (3, 4) in the Cartesian coordinate system, it becomes (3, 4, 1) in the homogenous coordinate system. The benefit of using the

homogenous coordinate system lies in the mathematical convenience on computing the ground truth of the correct correspondence of a feature point, as follows. Given a feature point with the coordinate $\mathbf{x}$ from one image, the spatial location $\mathbf{y}$ of its correct corresponding image pixel on the other image (i.e., the ground truth) can be easily computed by conducting projective transformation; i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}$, where matrix $\mathbf{A}$ is a 3×3 *projective transformation matrix* (or the so-called *homography*), and the 3×1 vectors, $\mathbf{x}$ and $\mathbf{y}$, are the coordinates of two feature points located in the homogenous coordinate system. Note that the corresponding 3×3 projective transformation matrices that perform I_1 -to- I_6 transformation in each scene category can be obtained from [35] for generating the ground truth. For the test image pairs without any geometric change involved (e.g., “Leuven”, “Bikes”, “Tree” and “Ubc”), the projective transformation matrix is simply the identity matrix each.

The computed ground truth as above-mentioned can be used to conduct objective performance evaluation for feature-point correspondence application. However, as pointed out in [31], due to the limitation of the feature detector’s accuracy, given an feature point from one image, the spatial location of the correct corresponding point identified by the feature detector in the other image is usually in the neighborhood of the computed ground truth, rather than exactly falling on it. Consequently, simply based on the ground truth of the given feature point could not reliably decide whether the corresponding feature point in the other image is a correct correspondence or not. To solve this problem, the ground truth of support region is a reliable supporting information that can be used to judge whether the established correspondence pair of two feature points is correct or not [31]. More details are provided as follows.

Note that each detected feature point has its associated support region. Therefore, given a feature point from one image, we can individually compute the corresponding location of each image pixel from its associated support region using

the projective transformation in the homogenous coordinate system as discussed previously. All such computed ground-truth locations together can form a support region in the other image, denoted as *support-region ground truth*. It is suggested in [31] that if the support region of the corresponding feature point in the other image has an overlap with the computed support-region ground truth by 50% at least, this signifies that these two feature points should be considered as a correct point-to-point matching pair. In this thesis, this criterion is used for the performance evaluation of the feature-point correspondence application.

5.5.3 Simulation Results and Discussions

Figure 5.4 demonstrates a set of *precision-versus-recall* performance curves generated by the software downloaded from [35]. Note that these performance curves are commonly used to evaluate the performance of feature-point correspondence method based on the SIFT feature descriptors. Here, we shall compare the resulted performance with and without using our proposed FHE to appreciate how much additional performance gain can be obtained. All experiments are conducted based on the reference image I_1 and its most different image I_6 in each scene category in order to investigate the most challenging correspondence scenarios. Simulation results have shown that with incorporation of our proposed FHE, the feature-point correspondence performance can be greatly improved especially when the test image pair contains a large amount of self-similar regions (i.e., the image content is inherited with strong self-similarity or in poor feature contrast). For example, for the image pairs “Boat” and “Bikes,” there are approximately 20% and 15% improvement in precision when the recall is fixed at 0.2, as shown in Figure 5.4(d) and Figure 5.4(f), respectively. However, if the test image pair contains only a few or even no self-similar regions, our proposed FHE would deliver small or even

(a) Graf

(b) Wall

(c) Bark

Figure 5.4. (Continue)

(d) Boat

(e) Leuven

(f) Bikes

Figure 5.4. (Continue)

(g) Trees

(h) Ubc

Figure 5.4. The *recall* (horizontal axis) versus *precision* (vertical axis) curves resulted by exploiting the initial feature point matching on the generated SIFT feature descriptors *with* and *without* incorporating the proposed FHE for various test cases: (a)-(c) viewpoint change, (d) zoom and rotation, (e) illumination change, (f)-(g) blurring, and (h) JPEG compression distortion. The test image pairs I_1 and I_6 are constantly used for testing the largest content variations in each scene category. Note that the range of precision in (a) is fairly small—in the range of $[0, 0.02]$ (i.e., showing nearly the same performance), while the remaining plots are constantly in the range of $[0, 1]$, showing appreciable or even significant precision improvement.

no performance improvement as expected; for example, the image pair “Graf,” as shown in Figure 5.4(a). Note that this is analogous to the use of IHE to an image that is already in high contrast or has been equalized already. In such case, there will be either small or no improvement on the visibility of image objects at all. For the image pair contains moderate amount of self-similar regions, the performance gain by using our proposed FHE method would yield a performance improvement between the above-mentioned two extremal cases. This could be best illustrated by the image pair “Ubc” in Figure 5.4(h), which has approximately 7% increase in precision when the recall is fixed at 0.2. These simulation results altogether convincingly demonstrate that the proposed FHE is able to provide a graceful performance improvement depending on different degrees of feature contrasts.

5.6 Summary

In this chapter, a novel and generic post-processing tool, called *feature histogram equalization* (FHE), is proposed for improving the representation of the generated feature descriptors. The entire development of the FHE is motivated by and slated on a fundamental concept introduced in this thesis for the first time, called *feature contrast*. It has been observed that feature contrast is inherited in all feature descriptors as a generic attribute on feature representation. Just like image contrast, which can be enhanced by exploiting the well-known *image histogram equalization* (IHE), so does the feature contrast by applying the developed equalization process to the feature histograms. Although the basic idea or ‘spirit’ of the FHE is in parallel with the IHE, their objectives are completely different—while the IHE modifies the image-intensity values to increase the visibility of low-contrast image contents, the FHE alters the vector-component values along each vector dimension (called

feature slice) to increase the feature contrast. This is equivalent to stretch the vector distances among the *equalized* feature descriptors further apart from each other. As a result, their associated feature points could be better discriminated to facilitate the targeted application. Furthermore, an optimal equalization algorithm is also derived in this chapter, with an empirically-determined parameter value suggested for practical usage.

To demonstrate its usefulness and effectiveness, the FHE is applied to feature-point correspondence application—aiming to identify as many correct point-to-point correspondence pairs as possible between two different images under matching. Extensive experimental results have clearly shown that our proposed method is able to significantly improve the performance of feature-point correspondence application especially when the image contents contain substantial amount of self-similar regions (i.e., presenting poor feature contrast). Just like the IHE that has been widely exploited as an effective front-end pre-processing tool in numerous image processing and computer vision tasks, it is expected that the proposed FHE will yield the same contribution and impact as a generic and powerful post-processing tool for many feature-descriptor-based applications along the way. Furthermore, the realization of FHE is quite simple with extremely low computational complexity, and it is inherited with parallel processing framework, since all the feature slices can be processed simultaneously and totally independent from each other. All these make the proposed FHE fairly attractive to many image processing and computer vision tasks that require maintaining real-time performance.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

Feature point matching, which is one of the fundamental and challenging problems in image processing and computer vision, is thoroughly investigated in this thesis. The goal of feature point matching is to automatically establish point-to-point correspondences between the two images acquired from the same scene but through two (very) different viewpoints. It is well-recognized that the SIFT-based feature point matching pairs initially established according to the nearest neighbor matching criterion tend to yield a large number of mismatches due to the limitations of SIFT technique and the simple matching criterion utilized. This is especially easy to incur when the two images under matching undergo a large viewpoint variation and/or contain a severely cluttered background. To tackle this problem, three thrust contributions are made in this work and presented in Chapters 3, 4, and 5, respectively. Further note that each contribution here addresses a different fundamental issue; the first two are from the viewpoint of the application's requirement, while the last one is from the fundamental aspect of feature representation.

For the first contribution as discussed in Chapter 3, the mismatch removal is

highly beneficial to the applications that require a sufficiently large number of point-to-point matching pairs in order to establish a certain geometric relationship or transformation between the given two images or two cameras. After identifying and removing those incorrect feature point matching pairs, the performance of those applications could be effectively improved. For that, a new bipartite graph model-based mismatch removal method is proposed in this work, and the main idea of this method is to use the proposed one-to-one coherent region pairs (CRPs) as the reference information to verify whether each initially established SIFT-based point-to-point matching pair is a correct match to retain or a mismatch to remove. To achieve this goal, the SIFT-based initial feature point matching pairs are first utilized to guide the merging of the segmented image regions that are spatially connected in order to form the candidate region matching pairs. A bipartite graph model is then developed to represent the candidate region matching pairs, followed by applying the Hungarian method to the model for identifying the optimal one-to-one CRPs. The optimality is achieved by maximizing the total degree of region similarity over all the candidate region matching pairs under the constraint that each possibly constructed region-to-region correspondence must be one-to-one. Note that the region similarity between the two regions for each candidate region matching pair is measured by the proposed region-to-region similarity measurement metric, which is a weighted combination of the following three aspects: photometric consistency, neighborhood consistency and point-to-point matching pair density. The performance of our proposed method is evaluated and further compared with a state-of-the-art algorithm [10] using six well-known wide-baseline image pairs. Simulation results have clearly shown that our proposed method is able to significantly reduce the mismatches incurred in the initially established SIFT-based point-to-point matching pairs and also outperforms the selected state-of-the-art method with an additional 3.83% improvement on average precision.

For the second contribution as detailed in Chapter 4, a novel *directed graph* (or *diagraph*) model-based correct match identification method is proposed to identify the visual patterns commonly shared in the two images under comparison. In our approach, each SIFT-based initial feature point matching pair is treated as a vertex of the diagraph model. The key novelty of our proposed method is to consider the neighboring vertices of each vertex of the link for quantitatively reflecting the degree of association between the two vertices of the link more accurately. To achieve this goal, unlike the conventional undirected graph model-based methods that assign a single link weight to each link of the graph, two link weights are generated for each direction of the link by exploiting our developed n -ranking process and a novel link weight enhancement technique. Consequently, the entire diagraph with two link weights is more robust on combating various interferences commonly encountered in the application of common visual pattern discovery, such as large viewpoint variation and indiscriminative feature descriptors. Based on the link weights generated at each assumed scale change factor, the strongly-associated subgraph is then identified from the directed graph by applying the non-cooperative game theory for handling non-symmetric adjacency matrix issue. All the vertices (i.e., point-to-point correspondences) belonging to the subgraph are collectively treated as one common visual pattern; hopefully, this set of vertices corresponds to one visual object. If this is not the case, our proposed topological splitting algorithm might be able to further discriminate them. The performance of our proposed method is evaluated and further compared with three state-of-the-art algorithms [12] [14] [63]. The quantitative evaluation is conducted first by utilizing the *simulated* SIFT feature points for standing out the technical challenges, followed by conducting the performance evaluation using six thoughtfully-chosen natural image pairs. Extensive simulation results in both evaluations clearly demonstrate the superiority of our proposed method on handling the adverse circumstances (e.g., viewpoint variations

and a severely cluttered background) for achieving more accurate discovery of the common visual patterns.

For the third contribution as elaborated in Chapter 5, *feature contrast* is first introduced as a measurement of the degree of self-similarity contained in an image pair. For the image pair with strong self-similarity, the established SIFT feature descriptors would be quite similar to each other and difficult for discrimination. Consequently, the feature contrast is low, and the precision of the initial feature point matching pairs will tend to drop drastically. In order to effectively increase the feature contrast, the SIFT feature descriptors obtained from an image pair will be jointly used to form a set of component-wise feature histograms. The proposed FHE is then applied to each feature histogram independently to re-distribute and modify the vector-component values of the SIFT feature descriptors along each vector dimension (called *feature slice*). Consequently, the feature contrast among the equalized feature descriptors is enhanced and becomes more distinct for better discrimination. Extensive simulation results have clearly shown that our proposed method is able to significantly improve the precision of the initial feature point matching pairs without any further refinement process, especially when the image pair contains substantial amount of self-similarity regions. This can be fully appreciated from the observation that approximately up to 20% improvement on the precision is accomplished when the recall is fixed at 0.2 in our experiments. However, as expected, if the image pair contains only a few or even no self-similarity regions, our proposed method should at least deliver the same performance as that without exploiting the FHE.

6.2 Future Works

6.2.1 Fast Feature Point Matching

The main focus of this thesis is trying to improve the accuracy of the final point-to-point correspondences established by a feature point matching process for satisfying different application's requirements. However, as more and more practical feature point matching systems are, or will be, running on small portable devices, such as smart mobile phone, tablet PC, to name a few. The computational complexity and memory consumption become two important criteria for the performance evaluation of the feature point matching process. Recently, some research works have already been carried out to design more computationally efficient and memory saving feature detector [93] [94], feature descriptor [102] [103] [104] [105], and feature point matching criterion [106]. Although the above-mentioned methods indeed accelerate the speed of feature point matching and reduce the required memory, the precision of the final point-to-point correspondences is also suffered to a great extent. Consequently, an 'optimal' fast feature point matching algorithm, which can deliver a balanced tradeoff among precision, computational complexity and memory requirement still needs to be further researched.

6.2.2 Feature Enhancement

In this thesis, a novel feature contrast enhancement technique, called *feature histogram equalization* (FHE), is proposed and proven to be an effective approach to increase the precision of the initial feature point matching pairs. This might offer a novel perspective on improving the performance of local feature descriptors for many other image processing and computer vision applications, such as object classification [107], nearly duplicated video retrieval [108], to name a few. As in the

existing literature, most of the research to improve the performance of local feature descriptors for these tasks is still focusing on proposing new SIFT-like feature descriptors [35] [109], which are more discriminative and more robust to viewpoint variations and illumination changes. It would be worthwhile to see how much performance improvement can be delivered by exploiting the feature enhancement technique for the above-mentioned feature-descriptor-based applications.

Publications

1. Chen Wang and Kai-Kuang Ma, “Common Visual Pattern Discovery via Directed Graph Model,” submitted to *IEEE Transactions on Image Processing*. (Under Review)
2. Chen Wang and Kai-Kuang Ma, “Feature Histogram Equalization: A Feature Contrast Enhancement Technique,” submitted to *Journal of Visual Communication and Image Representation*. (Under Review)
3. Chen Wang and Kai-Kuang Ma, “Bipartite Graph Model-based Mismatch Removal for Wide-baseline Image Matching,” submitted to *Journal of Visual Communication and Image Representation*. (Under Review)
4. Chen Wang and Kai-Kuang Ma, “Common Visual Pattern Discovery via Directed Graph Model,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 2957–2960, 11–14 Sep. 2011.
5. Chen Wang, Kai-Kuang Ma, Teck-Khim Ng, and Dong Guo “Mismatch Removal for Wide-baseline Matching via Coherent Region-to-Region Correspondence,” in *Proceedings of the Fourth Pacific-Rim Symposium on Image and Video Technology*, pp. 101–106, 14–17 Nov. 2010.
6. US Patent (Provisional Cover): “Feature Histogram Equalization: A Feature

Contrast Enhancement Technique,” Chen Wang and Kai-Kuang Ma, Application No. 61561557, filed by NTU, 18 Nov. 2011

Bibliography

- [1] D. Skea, I. Barrodale, R. Kuwahara, and R. Poeckert, “A control point matching algorithm,” *Pattern Recognition*, vol. 26, no. 2, pp. 269–276, 1993.
- [2] J. Starink and E. Backer, “Finding point correspondences using simulated annealing,” *Pattern Recognition*, vol. 28, no. 2, pp. 231–240, 1995.
- [3] P. J. Besl and H. D. McKay, “A method for registration of 3-D shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [4] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong, “A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,” *Artificial Intelligence*, vol. 78, pp. 87–119, 1995.
- [5] O. Enqvist, K. Josephson, and F. Kahl, “Optimal correspondences from pairwise constraints,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2009.
- [6] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz, “Multi-view stereo for community photo collections,” in *Proceedings of the IEEE International Conferences on Computer Vision*, 2007, pp. 1–8.

- [7] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski, “Building Rome in a day,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 72–79.
- [8] E. Borenstein and S. Ullman, “Learning to segment,” in *Proceedings of the European Conference on Computer Vision*, 2004.
- [9] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, “Using multiple segmentations to discover objects and their extent in image collections,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [10] V. Ferrari, T. Tuytelaars, and L. Gool, “Simultaneous object recognition and segmentation from single or multiple model views,” *International Journal of Computer Vision*, vol. 67, pp. 159–188, 2006.
- [11] M. Cho, Y. M. Shin, and K. M. Lee, “Co-recognition of image pairs by data-driven monte carlo image exploration,” in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 144–157.
- [12] M. Leordeanu and M. Hebert, “A spectral technique for correspondence problems using pairwise constraints,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2005.
- [13] H.-K. Tan and C.-W. Ngo, “Common pattern discovery using earth mover’s distance and local flow maximization,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2005, pp. 1222–1229.
- [14] H. Liu and S. Yan, “Common visual pattern discovery via spatially coherent correspondences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1609–1616.

-
- [15] C. Geng and X. Jiang, “Face recognition using SIFT features,” in *Proceedings of the IEEE International Conference on Image Processing*, 2009, pp. 3313–3316.
- [16] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney, “Surf-face: Face recognition under viewpoint consistency constraints,” in *Proceedings of the British Machine Vision Conference*, vol. 2, 2009, p. 7.
- [17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.
- [18] G. Carneiro and A. Jepson, “Multi-scale phase-based local features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 736–743.
- [19] T. Kadir and M. Brady, “Saliency, scale and image description,” *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.
- [20] T. Tuytelaars and L. Van Gool, “Matching widely separated views based on affine invariant regions,” *International Journal of Computer Vision*, vol. 59, pp. 61–85, 2004.
- [21] T. T. H. Bay, A. Ess and L. V. Gool, “SURF: Speeded up robust features,” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [22] P. Torr and D. Murray, “The development and comparison of robust methods for estimating the fundamental matrix,” *International Journal of Computer Vision*, vol. 24, no. 3, pp. 271–300, 1997.

- [23] O. Chum and J. Matas, “Matching with PROSAC-progressive sample consensus,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. Ieee, 2005, pp. 220–226.
- [24] M. Agrawal, K. Konolige, and M. R. Blas, “CenSureE: Center surround extremas for realtime feature detection and matching,” in *Proceedings of the European Conference on Computer Vision*, 2008.
- [25] C. Schmid and R. Mohr, “Local grayvalue invariants for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.
- [26] G. Carneiro and A. D. Jepson, “Flexible spatial configuration of local image features,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2089–2104, 2007.
- [27] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.
- [28] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.
- [29] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proceedings of the Fourth Alvey Vision Conference*, 1988, pp. 147–151.
- [30] T. Lindeberg, “Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention,” *International Journal of Computer Vision*, vol. 11, no. 3, pp. 283–318, 1993.

- [31] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.
- [32] J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda, “Uniqueness of the gaussian kernel for scale-space filtering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 1, pp. 26–33, 1986.
- [33] E. J. Pauwels, L. J. Van Gool, P. Fiddelaers, and T. Moons, “An extended class of scale-invariant and recursive scale space filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 691–701, 1995.
- [34] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schafalitzky, T. Kadir, and L. Gool, “A comparison of affine region detectors,” *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.
- [35] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [36] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, “CHoG: Compressed histogram of gradients A low bit-rate feature descriptor,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 2504–2511, 2009.
- [37] L. Van Gool, T. Moons, and D. Ungureanu, “Affine/photometric invariants for planar intensity patterns,” in *Proceedings of the European Conference on Computer Vision*. Springer, 1996, pp. 642–651.

- [38] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.
- [39] W. T. Freeman and E. H. Adelson, “The design and use of steerable filters,” *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 13, no. 9, pp. 891–906, 1991.
- [40] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets, or How do I organize my holiday snaps?,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2002, pp. 414–431.
- [41] R. Diestel, *Graph Theory*. Springer, 2005.
- [42] M. Fischler and R. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [43] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 508–515.
- [44] D. Nistér, “Preemptive RANSAC for live structure and motion estimation,” *Machine Vision and Applications*, vol. 16, no. 5, pp. 321–329, 2005.
- [45] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, “Spatial coding for large scale partial-duplicate web image search,” in *Proceedings of the international conference on Multimedia*. New York, NY, USA: ACM, 2010, pp. 511–520.
- [46] W. Zhou, H. Li, Y. Lu, and Q. Tian, “Large scale image search with geometric coding,” in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1349–1352.

- [47] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu, “Bipartite graph reinforcement model for web image annotation,” in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 585–594.
- [48] Y. Gao, Q. Dai, M. Wang, and N. Zhang, “3d model retrieval using weighted bipartite graph matching,” *Signal Processing: Image Communication*, vol. 26, no. 1, pp. 39–47, 2011.
- [49] P. Felzenszwalb and D. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.
- [50] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.
- [51] D. Zhang and S. Chang, “Detecting image near-duplicate by stochastic attributed relational graph matching with learning,” in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 877–884.
- [52] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, “Near-duplicate keyframe identification with interest point matching and pattern learning,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 1037–1048, 2007.
- [53] W.-L. Zhao and C.-W. Ngo, “Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection,” *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 412–423, 2009.
- [54] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine, “Unsupervised object discovery: A comparison,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 284–302, 2010.

- [55] H. Kang, M. Hebert, and T. Kanade, “Discovering object instances from scenes of daily living,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 762–769.
- [56] A. A. E. Hongwen Kang, Martial Hebert and T. Kanade, “Connecting missing link: Object discovery from sparse observations using 5 million product images,” in *Proceedings of the European Conference on Computer Vision*, 2012.
- [57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [58] S. Todorovic and N. Ahuja, “Extracting subimages of an unknown category from a set of images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 927–934.
- [59] M. Cho, Y. Shin, and K. Lee, “Unsupervised detection and segmentation of identical objects,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1617–1624.
- [60] S. Vicente, C. Rother, and V. Kolmogorov, “Object cosegmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 2217–2224.
- [61] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1. IEEE, 2005, pp. 462–469.
- [62] T. Basar and G. J. Olsder, *Dynamic noncooperative game theory*. Society for Industrial and Applied Mathematics, 1999.

- [63] T. Cour, P. Srinivasan, and J. Shi, “Balanced graph matching,” in *Advances in Neural Information Processing Systems 19*, B. Scholkopf, J. Platt, and T. Hofmann, Eds. Cambridge, MA: MIT Press, 2007.
- [64] J. Sivic and A. Zisserman, “Video data mining using configurations of view-point invariant regions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2004, pp. I–488.
- [65] Y. Jiang, C. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM, 2007, pp. 494–501.
- [66] T. Tuytelaars and C. Schmid, “Vector quantizing feature space with a regular lattice,” in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
- [67] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, “Visual word ambiguity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.
- [68] P. Koniusz and K. Mikolajczyk, “Soft assignment of visual words as linear coordinate coding and optimisation of its reconstruction error,” in *Proceedings of the IEEE International Conference on Image Processing*, 2011, pp. 2413–2416.
- [69] J. Yuan, Y. Wu, and M. Yang, “Discovery of collocation patterns: from visual words to visual phrases,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

- [70] T. Mita, T. Kaneko, B. Stenger, and O. Hori, “Discriminative feature co-occurrence selection for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1257–1269, 2008.
- [71] S. Ito and S. Kubota, “Object classification using heterogeneous co-occurrence features,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 209–222.
- [72] M. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1745–1752.
- [73] J. Yuan, M. Yang, and Y. Wu, “Mining discriminative co-occurrence patterns for visual recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 2777–2784.
- [74] A. C. Berg, T. L. Berg, and J. Malik, “Shape matching and object recognition using low distortion correspondence,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2005.
- [75] O. Duchenne, F. Bach, I. Kweon, and J. Ponce, “A tensor-based algorithm for high-order graph matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [76] R. Horaud and T. Skordas, “Stereo correspondence through feature grouping and maximal cliques,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 11, pp. 1168–1180, 1989.
- [77] M. Pelillo, K. Siddiqi, and S. Zucker, “Matching hierarchical structures using association graphs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1105–1120, 1999.

- [78] L. Torresani, V. Kolmogorov, and C. Rother, “Feature correspondence via graph matching: Models and global optimization,” in *Computer Vision—ECCV 2008*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008, vol. 5303, pp. 596–609.
- [79] D. Conte, P. Foggia, C. Sansone, and M. Vento, “Thirty years of graph matching in pattern recognition,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 03, pp. 265–298, 2004.
- [80] S. Gold and A. Rangarajan, “A graduated assignment algorithm for graph matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 377–388, 1996.
- [81] G. Elidan, J. Duchi, D. Tarlow, and D. Koller, “Using combinatorial optimization within max-product belief propagation,” in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, vol. 19. The MIT Press, 2007, pp. 369–376.
- [82] P. H. Torr, “Solving markov random fields using semi definite programming,” in *Proceedings of Artificial Intelligence and Statistics*, 2003.
- [83] C. Schellewald and C. Schnörr, “Probabilistic subgraph matching based on convex relaxation,” in *Proceedings of Energy minimization methods in computer vision and pattern recognition*. Springer, 2005, pp. 171–186.
- [84] A. Torsello, S. R. Buló, and M. Pelillo, “Grouping with asymmetric affinities: A game-theoretic perspective,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

- [85] M. A. Fischler and R. Elschlager, “The representation and matching of pictorial structures,” *IEEE Transactions on Computers*, vol. C-22, no. 1, pp. 67–92, 1973.
- [86] T. Caetano, T. Caelli, D. Schuurmans, and D. Barone, “Graphical models and point pattern matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1646–1663, 2006.
- [87] M. Berger and B. Gostiaux, *Differential geometry: manifolds, curves, and surfaces*. Springer, 1988.
- [88] J. Weibull, *Evolutionary game theory*. MIT Press, 1997.
- [89] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer-Verlag, 2006.
- [90] X. Wu, W.-L. Zhao, and C.-W. Ngo, “Near-duplicate keyframe retrieval with visual keywords and semantic context,” in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 162–169.
- [91] J. Zhu, S. C. Hoi, M. R. Lyu, and S. Yan, “Near-duplicate keyframe retrieval by nonrigid image matching,” in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 41–50.
- [92] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, “Near-duplicate keyframe identification with interest point matching and pattern learning,” *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 1037–1048, 2007.
- [93] E. Rosten and T. Drummond, “Machine learning for high speed corner detection,” in *Proceedings of the European Conference on Computer Vision*, 2006.

- [94] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, “Adaptive and generic corner detection based on the accelerated segment test,” in *Proceedings of the European Conference on Computer Vision*, 2010.
- [95] J. M. Morel and G. Yu, “ASIFT: A New Framework for Fully Affine Invariant Image Comparison,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.
- [96] Y. Ke and R. Sukthankar, “PCA-SIFT: a more distinctive representation for local image descriptors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 506–513.
- [97] L. Cao, J. Luo, F. Liang, and T. Huang, “Heterogeneous feature machines for visual recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 1095–1102.
- [98] A. Dame and E. Marchand, “Optimal detection and tracking of feature points using mutual information,” in *Proceedings of the IEEE International Conference on Image Processing*, 2009.
- [99] V. Ferrari, T. Tuytelaars, and L. V. Gool, “Wide-baseline multiple-view correspondences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 718–725.
- [100] T. Arici, S. Dikbas, and Y. Altunbasak, “A histogram modification framework and its application for image contrast enhancement,” *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 1921–1935, 2009.
- [101] R. Hartley and A. Zisserman, *Multiple View Geometry in computer vision*. Cambridge University Press, 2003.

- [102] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “BRIEF: Binary robust independent elementary features,” in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 778–792.
- [103] S. Leutenegger, M. Chli, and R. Siegwart, “BRISK: Binary robust invariant scalable keypoints,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2548–2555.
- [104] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.
- [105] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast Retina Keypoint,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [106] S. Taylor, E. Rosten, and T. Drummond, “Robust feature matching in $2.3\mu\text{s}$,” in *Proceedings of the IEEE CVPR Workshop on Feature Detectors and Descriptors: The State Of The Art and Beyond*, June 2009.
- [107] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.
- [108] H. Shen, J. Liu, Z. Huang, C. Ngo, and W. Wang, “Near-duplicate video retrieval: Current research and future trends,” *IEEE Transactions on Multimedia*, no. 99, p. 1, 2011.
- [109] E. Tola, V. Lepetit, and P. Fua, “Daisy: An efficient dense descriptor applied to wide-baseline stereo,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2010.