LOCAL FEATURES TO A GLOBAL VIEW:

RECOGNITION OF OCCLUDED OBJECTS

BY SPECTRAL MATCHING USING PAIRWISE

FEATURE RELATIONSHIPS

WU JIA YUN
(M. ENG., CHONGQING UNIVERSITY)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2012

Acknowledgement

I would like to express my deep gratitude to my supervisor, Professor Lim Kah Bin. His integral view on research and his untiring support have made a deep impression on me. It is a great pleasure for me to pursue my PhD degree under his supervision. I would like to thank my friends for their hospitality when I first arrived in Singapore. For my colleagues: Zhao Meijun, Wang Qing and Wang Daolei, I am thankful for their discussions and advice on my research. Thanks also go to my lab-mates: Wu Yue, Wu Zimei and Bai Fengjun for their support and company during my stay in NUS. I am very grateful to the examiners of this thesis for their reviews and helpful feedbacks on this thesis.

The financial support of National University of Singapore is gratefully acknowledged.

Table of Contents

Summary

Object recognition has extensive applications in many areas, such as visual inspection, part assembly, artificial intelligence, etc. Although humans perform object recognition effortlessly and instantaneously, implementation of this task on machines is very difficult. The problem is even more complicated when the object of interest is partially occluded in the scene. Many researchers have dedicated themselves into this area and made great contributions in the past few decades, many amongst which are feature based algorithms. However, these existing algorithms have various shortcomings and limitations, such as their limited applications to gray images without background disturbance, and the lack of global inference about target objects.

In this research, our algorithms to solve the recognition of occluded object problem are formulated as a local to global strategy, namely making a recognition decision based on local information collected. Since global information is no longer reliable for the recognition of occluded objects, local features are extracted. Feature types and locations are not specified. Instead, we would like to gather as much information as possible. Since a global decision is made based on local information, this local to global nature of occlusion recognition has brought us to spectral matching, for its ability to determine global structural properties of graphs. For our occlusion recognition algorithms, encoding feature geometric relationship into graph is important to retain global structure of possible target object or its parts. However, spectral algorithms respond badly to corrupted data set, such as occluded objects, where ambiguous connections are generated. Therefore, our efforts are focused on how to reduce interactions of features from different objects before attempting to

solve occlusion recognition problem. Reducing feature interactions for spectral correspondence is the key for our algorithms to recognize occluded objects.

We propose to reduce feature interactions based on intermediate-level vision cues: grouping and segmentation. With our feature interaction formulations, inter and intra feature relationships are established to indicate their possibility to come from the same object. By combining feature interactions with spectral matching, our algorithms take into consideration, the feature geometric and appearance relationships, integrating low-level, intermediate-level vision cues into higher level vision tasks. On the other hand, the applications of our occlusion recognition algorithms are extended in dynamic scenes, where occlusion rates vary with time. Possible object regions are first extracted based on local saliency. Without assumptions on object appearance, our method is attention-guided. With the obtained regions, approaches have been proposed to integrate motion and stereo information into recognition, implying the cooperation between multiple vision applications. All of these efforts are made to reduce interactions between different objects, which serve as priors to guide our matching, recognition and pruning searching space.

List of Figures

List of Tables

List of Symbols

Graph for the k^{th} dataset:	$\mathbf{G}^k = (\mathbf{V}^k, \mathbf{E}^k)$
Adjacency matrix element:	w_{ij}
Proximity matrix	$\mathbf{M} : (m_{ij})$
Eigen-value:	λ
Eigen-vector:	$\mathbf{v}$
Modal matrix:	$\mathbf{Mo}$
Similarity edge attributes:	$\mathrm{S}(\cdot,\cdot)$
Binary matrix:	$\mathbf{C}_{n\times n'}$
Binary vector in quadratic program:	$\mathbf{C}_{nn'\times 1}$
Feature set:	$\mathbf{F}^k = \left[f_1^k, \quad f_2^k \quad \cdots \quad f_n^k\right]^T, f_i^k = (des_i^k, x_i^k)$
Image patch:	R_i
Color co-occurrence histograms	$\mathbf{CH}_i$
Patch texture descriptor	T_i
Quantized color set:	$\widetilde{C}$
Predefined distance set:	$\widetilde{D}$
Number of quantized color and distance:	$n_{\widetilde{c}}$ and $n_{\widetilde{d}}$
Matrices intersection:	$\mathrm{I}(:,:)$
Mean vector over region	$\overline{f}$
Texture similarity	$\rho(:,:)$
Appearance similarity:	$\widetilde{p}(:,:)$
Appearance weighting parameters	σ_1, σ_2
Matching constraint	$\Theta(:,:)$
Diagonal element in **M**	$\Gamma(:,:)$
Off-diagonal elements in **M**	$\widetilde{\Gamma}(:,:)$
Euclidean distance between features	$\mathrm{d}(:,:)$
Angle to the *x*-axis:	α

Local support region:	D
Confidence weighting in **M**	$\widetilde{w}$
Geometric weighting parameters	σ_3, σ_4
Intra-proximity	$\mathbf{I}_{\mathbf{intra}}$
Inter-similarity	$\mathbf{I}_{\mathbf{inter}}$
Feature association score:	θ
Feature association score vector	$\boldsymbol{\theta}$
Area of region	A
Area factor; Neighboring effect; Central effect	$\varepsilon_1, \varepsilon_2, \varepsilon_3$
Weighting parameters	$\sigma_{\varepsilon_2}, \sigma_{\varepsilon_3}$
Average pixel values	$\bar{s}_i$
Prominence of a salient region:	S_p
Camera intrinsic matrix	**K**
Homography matrix	**H**
Translation matrix	**T**
Rotation matrix	**R**

Chapter 1

Introduction

1.1 Background

An object recognition system is to locate and recognize objects of interest in a scene image taken in the real world, using object models which are known as priors. Object recognition has extensive applications in many areas, such as visual inspection, part assembly, artificial intelligence, etc. Although humans perform object recognition effortlessly and instantaneously, implementation of this task on machines is very difficult. It is a major and also challenging task in computer vision. Many researchers have dedicated themselves into this area of research and made great contributions in the past few decades.

The object recognition problem can be defined as a labeling problem based on models of known objects, shown in Figure 1.1. These model images for the individual objects are stored in the database and are identified with different labels. Given an image (scene image) containing one or more objects of interest and a set of labels corresponding to a set of models known to the recognition system (model images), this system should match and assign correct label (s) to the objects in the scene image. For an individual object, its model images are taken from different views and the number of views is pre-determined. In an arbitrary scene image, objects of interest

may appear differently from their models stored in database, in terms of scale, orientation and position.

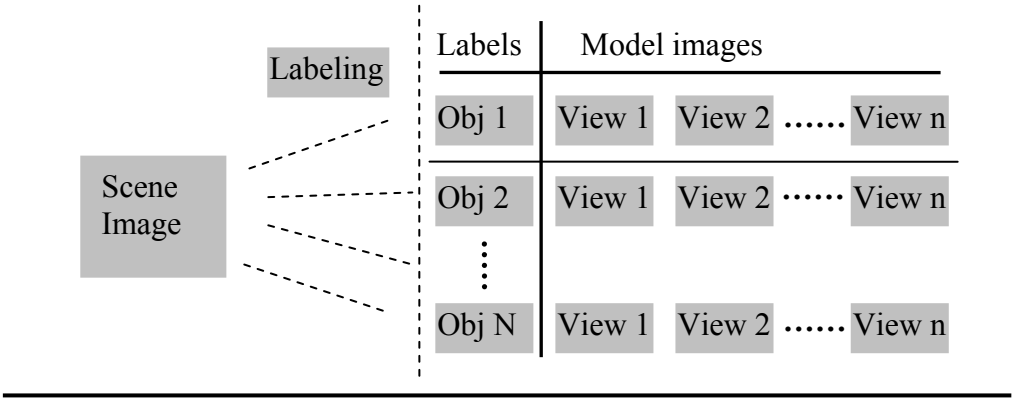


Figure 1.1 Recognition as a labeling problem

A vision recognition system is to emulate human recognition process: humans can only recognize objects (scene) that they have seen before (database). Specifically, for feature based object recognition, images are defined by features, quantifying their distinguishable characteristics. To recognize objects of interest in a scene, features extracted from the objects are matched with the feature sets of all model images in the database. Those correctly matched features are called correspondences. In this thesis, we would like to define that correspondence refers to a match in features found between scene and model images. The number of correspondences found indicates the confidence of recognizing the object in the scene.

Object recognition task can be challenging, because performances of recognition systems vary with vision applications, one of which is when the objects are partially occluded. While humans can easily distinguish different objects in occlusion scenarios, the state of the art recognition algorithms are far behind in this ability. General feature based recognition algorithms suffers from a major setback when applied in occluded object recognition, i.e., incomplete and/or deformed features.

However, what makes recognition of occluded objects task challenging is the interactions of features from different objects in the scene. This problem is ill-posed, because there is no formal definition of what constitutes an object category. Philosophically, any arbitrary definition could characterize a category (e.g. things with a complex shape). While people largely agree on common and useful object categories, it is still not clear which are the objects' properties that help us to separate them when they are partially overlapping in the scene. In some extreme cases, there are not enough visible features left on the object of interest to be recognized, i.e., not enough local information to form a global decision about this object, causing recognition failure.

In this thesis, we propose that feature interactions could be a key to solve occlusion recognitions. In the rest of this thesis, feature interactions from different objects are denoted as feature interactions for simplicity. If feature interactions can be reduced based on feature relationships that indicate feature groups are from the same regions or objects, recognition search space could be pruned. In fact, relationships between features have been employed, instead of features themselves in object recognition for its importance in object recognition [1] [2] [3]. This is motivated by research in cognitive science hypothesizing that human based category recognition depends on pairwise relationships between object parts [33]. With this knowledge, relationships between features could carry more important information than local and first-order features alone.

Feature relationships are pairwise in nature. Geometric constraints between feature pairs have been applied in correspondence problem [13] [43] [15]. For good matching performance, it is important to take into consideration not only the first-order local

appearance between individual features, but also the higher order geometric and appearance consistency between groups of features. Other research topics such as image segmentation and perceptual grouping are also exploiting pairwise similarities on pixel level or in feature space. In view of the above reasoning, the algorithms proposed in this thesis, for occluded object recognition, will base strongly on the relationships between features and the subsequent reduction in their interactions.

1.2 Problem descriptions

In natural scenes, it is commonly found that objects are occluded by or occluding other objects or surfaces, as shown in Figure1.2. These situations result in partially visible objects in the scene, which is referred to as occlusion in the rest of this thesis. For human beings, to recognize an object under severe occlusion could also be a difficult task.

Figure 1.2 Objects occluded by or occluding other objects or surfaces

Occlusion is a challenging problem, as it can severely compromise performances of many computer vision applications, such as path planning in autonomous robot navigation, target tracking and recognition in visual surveillance and segmentation from overlapping tissues in medical diagnosis, as shown in Figure 1.3.

Figure 1.3 Occlusions in computer vision applications

Occlusion is also a major challenge for the accurate computation of visual correspondence, such as stereo matching and image registration. In these cases, occluded pixels are visible in only one image, so there are no corresponding pixels in the other image. For 3-D reconstruction, it is particularly important to obtain good results at discontinuities, which are places where occlusions often occur. Ideally, a pixel in one image should correspond to at most one pixel in the other image, and a pixel that corresponds to no pixel in the other image should be labeled as occluded.

Specifically, occlusion affects feature based object recognition in the two main aspects, namely feature detection and description. Firstly, since the most sophisticated edge detector responses strongly to edges which do not correspond to any physical occlusion, occlusion boundaries cannot be detected from a single image [63] [215]. Therefore, it is difficult to reliably detect object boundary fragments for shape information. On the other hand, feature detectors equally respond to appearance of edges and occlusion boundaries, leaving a mixed set of features to represent the original object. As for description, occlusion violates the underlining assumption for aggregation process in feature description. Due to the lack of information, most processing techniques aggregate information spatially when describing individual feature point or region. This aggregation is the result of discrete patches centered at feature points in object recognition or region formulated by graphic models connecting neighborhood pixels such as Markov, Conditional or Discriminative

Random Field [160] [162]. Each of these techniques implicitly assumes that all the nearby or connected pixels are from the same object. This assumption is violated at occlusion boundaries, where information from the two different surfaces is collected within a single image patch. Then, patches of image data which may also cross object boundaries are used to create descriptor vectors for matching. An example is shown in Figure 1.4.

Figure 1.4 Information aggregations by discrete patches [83]

In particular, for scale-invariant methods, larger neighborhoods are taken when the scale of detection and description increases, resulting in many unusable large-scale features which contain information from (multiple) objects and background. Equivalently, as an object appears smaller and smaller within a scene, we must rely more heavily on its larger scale features (relative to observed size of the object) for matching. In both situations, occlusion could affect the feature description, leaving ambiguous feature set for the object of interest.

Since feature set extracted which originally represents one particular object is contaminated by occlusion, they are no longer reliable to represent this object. That is the reason why global features are avoided in occlusion recognition, since they are more easily corrupted, such as shape information. Researchers have developed

algorithms using local features to deal with this problem. In fact, sophisticated features detectors are sensitive themselves. Because of the ambiguity in feature detections and descriptions when occlusion involved, features from different objects tend to interact and therefore local feature matching alone is not enough to be used as evidence for object recognition.

1.3 Feature based recognition process

For a feature based object recognition system, the recognition process can be summarized in Figure 1.5.

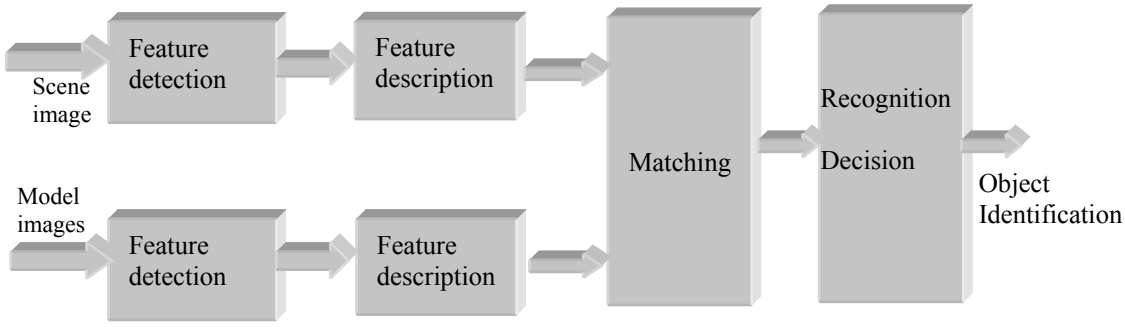


Figure 1.5 Three phases of feature based object recognition

This operation typically involves three distinct steps. Firstly, a feature detector identifies a set of image locations presenting rich visual information and whose spatial location is well defined. It is computationally inefficient to characterize the entire image. Therefore feature detector is to locate feature points of interest which most distinctly identify the target object. The spatial extent or scale of the feature may also be identified in this first step, as well as the local shape near the detected location. The second step is to determine the feature description. A vector is computed from the image to characterize local visual appearance near the location of the detected feature point. The image is characterized around each feature point in an invariant fashion. In the third step "Matching", a given feature is associated with one or more features in

model images stored in the database. Important aspects of matching are criteria to decide whether two features should be associated or considered to be related. From Figure 1.5, in order to reach a formal recognition decision, good structure of database and efficient searching algorithm are also required.

To identify 3-D objects of interest, a dictionary or a lookup table is built based on features extracted from model images for all known objects. Then, features extracted from a scene image are matched against model features. Subsequently, a geometric consistency model is then applied to all matching feature pairs to remove inconsistent outliers and to determine the object location in the scene and its 3-D pose with respect to model views.

There are various ways to match a given image feature to the established feature dictionary. A simple method is to find any database feature that has a description vector with a Euclidean distance to the given feature's description vector below some threshold value. This is a simple and fast method but lacks accuracy due to the nature of the local description vectors. Not all local feature description vectors are equally discriminatory, meaning that a single threshold value of the Euclidean distance is unsuitable for determining how a feature matches against all the remaining features. For instance, in the most popular SIFT matching algorithm, a nearest neighbor method is proposed for matching features. For a given image feature, it is necessary to find the nearest neighbor feature in the database, as well as the second nearest neighbor feature, using the description vectors to determine distance. The image feature is said to match the nearest neighbor feature if the distance between them is less than a certain ratio of the distance between it and its second nearest neighbor feature. Usually, the ratio was experimentally found to give the best trade off between

false positives and false negatives. The underlying justification for this approach is that the density of features near a given feature in a database is an indication of how discriminatory that feature is. A disadvantage of this approach is the difficulty of efficient nearest neighbor database search. In the case of a high dimensional search space, a modification of the *kd*-tree search was developed, called the Best Bit First algorithm [62], which can find the approximate nearest neighbor in a high dimensional space efficiently. It works by prioritizing the search of bins in a *kd*-tree towards those that are closest to the query location, and terminate the search after a given maximum number of bins have been examined. Using this method it is possible to find a nearest neighbor in a 128-dimensional space with a speedup of two orders of magnitude over traditional *kd*-tree search, with less than 5% incorrect matches.

Our algorithm aims to match features extracted in a scene image with occluded objects with model database. Our basis of occluded object recognition integrates the first-order and second-order feature relationships in terms of geometry, color and texture.

1.4 Our scheme

Researchers have developed some algorithms to deal with occlusion recognition problem. They either restricted themselves in 2-D environment or used color information alone [51] to recognize occluded object of interest. There will be a detailed overview of previous works in the next chapter.

To recognize a partially occluded object, a local to global strategy is required. Local information is the reliable source of information for object representation under occlusion, but the recognition should consider the global information of the object of

interest. In other word, evidence should be locally gathered to form a global decision on object identification. This local to global nature of occlusion recognition problem has brought us to spectral matching, by which global structure of the object is preserved through considering relationships between local features. It is natural to encode various feature relations in a graph, where nodes are associated unary features and edges second-order or higher order relationships between the features. With the feature relationship graph, graph based matching techniques, such as spectral matching, have been exploited to find correspondences for recognition. Correspondences from matching graphs could provide a global view of the target object or part, because the structure of model features should be maintained by its matched scene features with respect to both local appearance and relative spatial relationships. This property of spectral matching will be introduced in detail in Chapter 3.

Feature relationships are important in our formulation to solve occlusion recognition problem, because they represent local information as well as potential global description of target object. In the context of occlusion, feature relationships are closely related to feature interactions, since the original feature relationships for one object have been corrupted by interactions between features from different objects. It is crucial to handle feature interactions to reduce the recognition ambiguity for occluded objects. In our proposals, feature interactions could be reduced by measuring inter and intra image feature similarities. By measuring similarities of neighboring features, the possibility for two features to be from the same object or region is calculated. By measuring affinity between features and model images, the possibility for features to be associated with object of interest is established. These measurements serve as an indicator for feature interactions reduction. In fact, these

possibilities can be interpreted as constraints to correspondence formulations, by which various feature relationships can be taken into consideration. They also serve as priors on some level, thereby establishing priors about which pairs of features are likely to be part of the same object.

With the knowledge on the power of feature interactions, they are introduced into the recognition of occluded objects by developing rich constraints on matching process. We shall approach this objective based on spectral matching correspondence algorithm and feature interactions reduction based on appearance information. Intuitively, these two aspects could significantly benefit the recognition of occluded objects. If the feature interactions are reduced, recognition search space could be pruned, leading to higher confidence in the recognition of occluded object. In the meanwhile, query and models could have very different appearances from different viewpoints, which could be partly solved by a robust correspondence matching method for a better recognition rate.

In addition, we have considered situations where occlusion is present in a dynamic way, i.e., occlusion rates varying with time. In this context, region information is further introduced for its likely containing object of interest. By propagating region information throughout the image sequence, our algorithms to recognize occluded object could be extend to video.

1.5 Contributions

1) Introducing graph matching into recognition of occluded objects

We have studied the problem of recognition of occluded objects, commonly known as occlusion recognition. In occlusion scenario, global features are corrupted and

unreliable for the recognition task. To successfully recognize occluded objects, a global decision has to be made based on locally gathered information. This local to global nature of occlusion recognition problem has brought us to graph matching theory. Novel algorithms are proposed to handle occlusions based on graph matching, which has long been an open issue for graph matching algorithms. Popular spectral algorithms are evaluated according to their performances under different occlusion rates.

2) Interaction reduction for occlusion recognition

We propose that reducing feature interactions is a key to recognize occluded objects. Accordingly, this problem is addressed by soft grouping based on appearance similarity as well as associating features with objects of interest. They are two approaches handling feature interactions in a button-up or top-down fashion.

Local appearance information is interpreted as the statistical probability of two features to be from the same object or parts. This probability could serve as priors for further procedures. Compared with methods which learn priors from training set, our formulation suggests an efficient and one-shot-recognition solution. Besides the bottom-up-processing of linking features in scene image, top-down reasoning was proposed by associating individual features with the object of interest. In this context, the selection of potential assignments is with physical meaning, which is considered as one step further than *k*-nearest neighbors used in Spectral Matching (Marius 2005). Moreover, the occlusion rate has been integrated into recognition confidence.

3) Simultaneous extraction of possible object regions based on visual attention

In order to extend our algorithms in dynamic scenes, region information has been introduced. Region information was extracted based on local saliency and therefore the region extraction is attention guided for possible object regions. Extracting potential object regions based on local saliency for further recognition follows the trend in computer vision area: integrating low-level (saliency map) and intermediate-level (foreground) vision cues into higher level vision task (recognition).

4) Recognizing occluded objects in dynamic scenes

Based on these possible object regions, motion smoothness constraint for the same object is applied through image sequence from surveillance camera. Even when object region is not informative in a frame for recognition (object of interest is under severe occlusion), recognition can still be performed by matching with its associated regions in other frames. This approach is referred to as "Trace back" In the other dynamic scenes taken by movable camera platform; we implement the idea of 'Take a look around'. With this platform, binocular stereo vision system is able to access different views of an occlusion scene. Disparity map is refined by combining with region indication, which is considered to be equivalent to introducing appearance information into stereo algorithms. A viewpoint is updated by moving the vision system towards and along the direction in which the size of object region grows, implying reduced occlusion rates for successful recognition.

Our occlusion recognition algorithms perform without the assumption of knowing where and how the occlusion occurred, which is the assumed knowledge for many other occlusion recognition algorithms. In our formulation to recognize occluded object in dynamic scenes, motion and stereo information are introduced into recognition, which implies the integration of multiple vision applications.

1.6 Thesis Outline

A model based object recognition system has two distinct phases of operation. During the training phase, the system is exposed to a set of objects that it is expected to be recognized. This exposure allows the recognition system to build an internal representation, called model, for each object. During the next phase, known as recognition phase, the representation for the unknown object is obtained and a best match search amongst the stored models is conducted to determine the identity of the object. Therefore, there are three factors that characterize a model based object recognition system. These are (a) representation, (b) matching scheme, and (c) recognition decision. Our proposed algorithms utilize feature based representation and no special features are required. Component (b) and (c) will be covered and the rest of this thesis is organized as follows:

Chapter 2 gives a literature review on related research in object recognition and especially the recognition of 2-D partial occluded objects which employ traditional techniques. We are inspired to reduce feature interactions to achieve successful recognition based on intermediate vision tasks. Furthermore, the local to global nature of occlusion recognition problem drives us to spectral matching subjected to feature interactions reduction.

In Chapter 3, spectral matching algorithms are introduced. Correspondence from graph matching is formulated as quadratic integer programming. The two main elements in graph matching are considered: weighting function and discrete approximation methods. Multiple weighting functions have also been investigated. Popular spectral algorithms with quadratic formulation are evaluated based on their performances under different degrees of occlusions and outliers.

In Chapter 4, appearance information has been employed to reduce feature interactions. Our proposed methods are formulated by considering pairwise feature appearance relationships, in a bottom-up or top-down fashion. These proposed methods are evaluated on their ability to reduce feature interactions, respectively.

In Chapter 5, our algorithms in occlusion recognition have been proposed. Spectral matching algorithms could be introduced to occlusion recognition based on our formulations of reducing feature interactions. The proposed algorithms are evaluated on their performances under various occlusion rates. Comparisons are made between our methods and the state of the art recognition algorithms.

Chapter 6 has introduced region information for its usefulness in extending our recognition algorithms in dynamic scenes. A modified foreground subtraction method has been proposed to work with visual attention based saliency. The extraction of foreground regions is then attention guided, indicating possible object regions in individual frames.

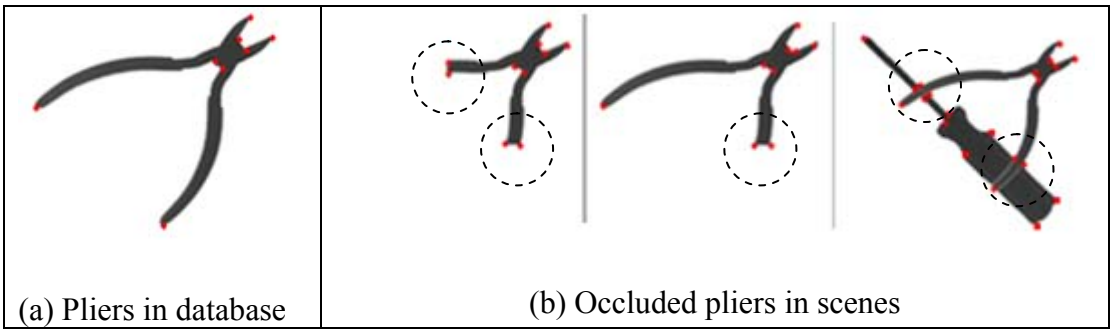
In Chapter 7, our algorithms to recognize occluded objects are extended to video or image sequences based on the region information. With respect to different applications, two approaches have been proposed to propagate the region information throughout the image sequences, thus extending the field of application of our algorithms in occluded object recognition.

Chapter 8 concludes and summarizes the contributions of the research work presented in this thesis. Some limitations of our proposed recognition approach are discussed. Potential future works are also presented.

Chapter 2

Literature review

For object recognition algorithms, if the object of interest is fully visible, a feature based recognition system can proceed as shown in Figure 1.5. However, if object of interest is partially visible (occluded by other objects), the recognition task is more challenging and requires special strategies. The difficulty lies on the fact that the original feature set to represent the object has been corrupted, where ambiguous features are generated and features from different objects tend to interact. This fact is illustrated in Figure 2.1, where red dots denote extracted features and circles indicate feature set corruption due to occlusion.



(a) Pliers in database (b) Occluded pliers in scenes

Figure 2.1 Original feature set corrupted by occlusions

Occlusions originated from overlapping objects commonly happen in practical vision applications. For instance, in an industrial assembly setting, component parts are moving on a conveyor belt for visual inspection as shown in Figure 2.2 (a). For cases such as parts overlapping each other, the vision system must be able to recognize each of the occluded objects correctly, other than to reject them as a single unidentifiable part. Similar situation occurs when a robot needs to pick up an

industrial part in a bin, which is filled with different parts overlapping each others as shown in Figures 2.2(b).

(a) Conveyor belt for visual inspection

(b) Picking an object from a bin

Figure 2. 2 Occlusion scenarios in industrial assembly setting

Existing algorithms to recognize occluded objects can be divided into two categories according to their choices of descriptive tokens used for matching. Some methods rely on computing local distinctive features that represent a genuine segment of this object or a distinguishable small area of a region. The rest depends on feature relationship schemes. In early occlusion recognition algorithms, occluded objects and model objects are described either by object boundary segments with a sequence of control or dominant points or by using local features such as corners and holes and their relationships [9]. In this chapter we will briefly review and discuss existing algorithms in occlusion recognition and related work with respect to our algorithm. Note, however, that further discussion of related work is distributed throughout the reminder of this thesis.

2.1 Occlusion recognition by local geometric features

Algorithms for object recognition can generally be classified into those which use image global features and those which use local features. Global features refer to

properties of an image as a whole, such as colour histogram, outline shape, and texture [67] [68], as well as characteristics of the entire region or boundary, for instance, area moments (Hu [175]; Teh et al. [176]; Khotanzad et al. [177]), curve moments (Chen [178]; Zhao et al. [179]) and Fourier descriptors (Persoon [172]; Richard et al. [173]; Etesami et al. [174]). The drawbacks of using global features for object recognition include sensitivity to clutter and occlusion, and difficulty in localizing an object in an image. Object recognition algorithms based on global features fail to work when partial occlusion takes place, where global features are severely contaminated.

On the other hand, local features describe image patches around points of interest. Local image features, such as edge and corner, are specially extracted for object recognition, because they are more robust to occlusion and clutter compared to global features [69] [70]. Early occlusion recognition algorithms focus on developing local features which are robust to occlusion. Typically, these works restricted themselves to 2-D object recognition, assuming that all the real world objects are viewed by a camera directly located on top of them. Local primitives in spatial domain were used such as corners, holes, and curve segments (e.g. Lines and arcs). Some researchers (Ansari et al. [180]; Han et al. [206]; Lamdan et al. [181]; Tsang et al. [58]; Chandran et al. [207]; Zhang et al. [182]) used dominant point based recognition methods to recognize partially occluded objects. Tan in [10] proposed a method for transforming the dominant points to a canonical form and performed the matching between occluded and model objects by matching the canonical forms under affine transformation. However, dominant points alone were insufficient to form a completely integrated representation of an object and the extraction of dominant point was sensitive to noise contamination.

Other researchers have utilized polygonal approximations to represent the object boundary as a string of line segments for occlusion recognition (Bhanu et al. [16]; Price [14]; Ayache et al. [167]; Bhanu et al. [9]; Grimson et al. [168]; Liu et al. [17]). To recognize objects which are not polyhedral, researchers described them by circular arcs (Turney et al. [169]; Knoll et al. [208]; Kalvin [170]; Ettinger [171]). Other methods introducing segments in spectral domain consisted of Fourier descriptors (Persoon [172]; Richard et al. [173]; Etesami et al. [175]) and Wavelet descriptors (Lim et al. [12]), which are less sensitive to noise compared to the features in spatial domain.

The underlining efforts in these 2-D occlusion recognition algorithms were to keep original object boundary segments away from ambiguous ones generated by occlusion, which might cause recognition failure when incorrectly matched with information stored in database. The feature interactions due to occlusion might be reduced to a certain degree in their settings. However, these local descriptions lack global interpretations of the object structure. Recognition algorithms based on these representations hardly differentiate different objects with similar shapes. Therefore, 2-D occlusion recognition methods had limited success for objects with arbitrary shapes. Although the combination of local features could generate more geometric representation of the object boundary, it was difficult to meet representation requirements in invariance and completeness for objects with arbitrary shape.

2.2 Occlusion recognition by feature geometric relationships

Other work employs less distinctive descriptors and rely more on the relationships between them to realize object recognition and localization. Feature relationships were considered in early schemes, including Grimson [78], Ikeuchi and Horn [210];

Faugeras and Hebert [183]; Gaston and Lozano-Perez [209]; Grimson and Lozano-Perez [101]; Stockman and Esteva [211]; and Brou [212]. It had been introduced by Grimson and Lozano-Perez [101] that geometric relationships between simple features could be utilized to recognize occluded objects.

Liu et al. in [17] developed a technique based on distance transformation in matching sequences of consecutive segments between a pair of contours. Lamdan et al. improved their former technique by using the methodology of geometric hashing in [65]. Tsang et al. in [58] introduced a new scheme capable of combing the isolated clusters for object classification using three point matching and distance transformation. Price in [14] suggested a method that builds a disparity matrix that contains the results of comparing the boundary segments of an object image with an occluded image. From this matrix the sequence of compatible segments is found. Based on the clustering structure introduced in [59], Bhanu and Ming [9] further added the length of boundary segments and then formed clusters for the objects which were under occlusion. In [61], model images and occluded object were represented as sequences of geometric graphs formed by connected lines. Subsequently, the occluded object was recognized by comparing properties of these graphs, where three connected edges from these two graphs were manually matched as priors. Background had not been considered in these algorithms and local features (contour, points and lines) were assumed to be lying on the foreground. These algorithms were mostly to produce individual correspondences, without considering the global configuration of object. In fact, the locally matched features do not necessarily lead to successful recognition, because of the contamination of extracted feature set due to occlusion. In the meanwhile, there is no investigation on the confidence of recognition, which

should be considered for the recognition of occluded object, since the number of available features decreases with the increase in occlusion rates.

Other popular methods which exploit feature relationships are Hough Transform [33] [41] and RANSAC (RANdom SAmple Consensus) [42]. They have been widely used to enforce spatial consistency between local features. Hough Transform often requires a careful selection of parameters (e.g. resolutions of sub-divisions) and it breaks down easily in the presence of clutter. RANSAC is more resistant to clutter. It generates random matching hypotheses for small number of anchor points and then evaluates the hypotheses with all the points. As it becomes increasingly slow with more outliers in the image, heuristics and General Hough Transform have been introduced in [42] to preprocess the local matches and prune the unpromising ones. However, these remedies do not offer great help when local feature matching becomes increasingly ambiguous. Therefore, these two methods are not suitable for occlusion recognition with unpredictable occlusion rates.

2.3 Recognition of occluded object by local feature relationships

We are motivated by the success of the existing methods reviewed in Section 2.2 and realize that local features and feature relationships are important in solving occlusion recognition problem. Both of them should be considered because of the local to global attribute of occlusion recognition: a global decision based on local information. Even though local features are reasonable sources for object representation under occlusion, global structure of this object is indispensable for a recognition decision. The global structure of the object of interest could be retained by feature relationships, which is why feature relationships are essential to our approaches. Matching local features between scene and model images as well as

maintaining the relationships between them, could keep the recognition from failure caused by the interactions of features from other objects in the scene.

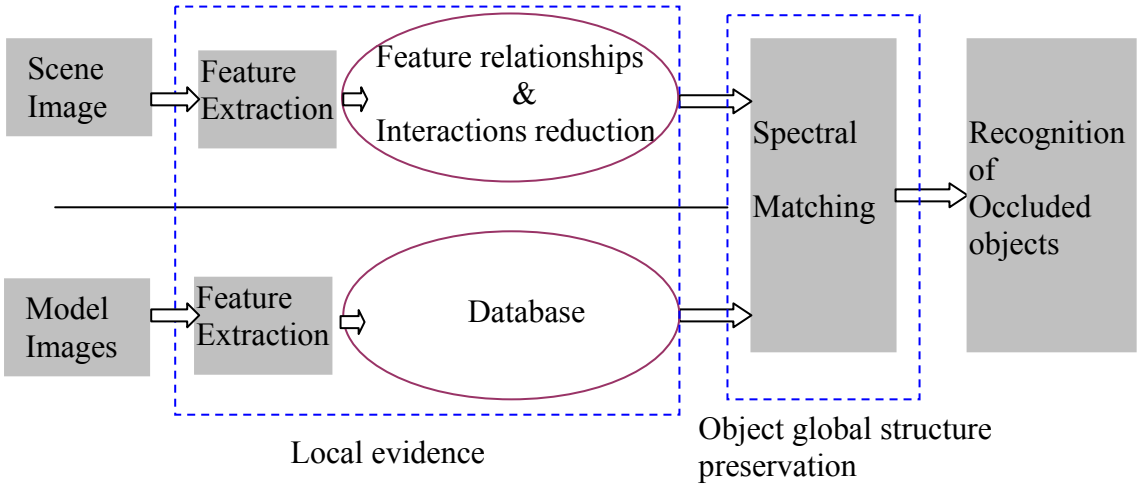


Figure 2.3 Pipeline of our algorithm

Our algorithms are to produce a global decision based on the local evidence. The flow chart of our algorithm is shown in Figure 2.3. In the following subsections, we shall review related work with respect to each component shown in the above flow chart.

2.4 Feature detectors and descriptors in object recognition system

In the past years, object recognition systems have been continuously developed and improved. An ideal system should be able to detect a large number of meaningful features in a typical image, and match them reliably across different views of the same scene/object in the image. Critical issues in detection, description and matching are robustness with respect to viewpoint and lighting changes, the number of features detected in a typical image, the frequency of false alarms and mismatches, and the computational cost of each step. Different applications weigh these requirements differently.

Performances of a number of feature detectors and descriptors were evaluated and compared by Schmid in [69] and Mikolajczik and Schmid in [70]. Even though these evaluations relied on the use of images of flat scenes, or in some cases synthetic images, they have shed lights on the characteristics of different detection and description techniques. The relative performances have also been assessed based on 3-D objects in [71]. When considering 3-D images, instead of planar scenes, features detected are generated by either surface markings or the geometric shape of the object. The former are often associated with smooth surfaces, they are usually located far from object boundaries in order to avoid occlusion boundaries and have been shown to have a high stability across viewpoints [69][70]. Their shape change may be modeled by an affine transformation, resulting in the development of affine invariant detectors (Garding and Lindeberg in [184]; Lindeberg in [129]; Baumberg in [185]; Schaffalitzky and Zisserman, in [186]; Mikolajczyk and Schmid in [187]). The latter are associated with high surface curvature and are located near edges, corners and folds of the object. Due to occlusion and complexity of local shapes, these features have a much lower stability with respect to viewpoint changes. For features considered in this thesis, their locations in the image are not restricted to avoid the effects of occlusion, instead, feature are extracted to gather as much local information as possible.

In feature based object recognition system, geometry and appearance are the two major sources for feature extraction. Traditional feature based geometric approaches have been developed for recognition, such as alignment (Ayache and Faugeras [167]; Faugeras and Hebert [183]; Grimson and Lozano-Perez [188]; Huttenlocher and Ullman [189]; Lowe [190]) or geometric hashing (Thompson and Mundy [166]; Lamdan and Wolfson [164]; Lamdan and Wolfson [165]). Various subsets of

geometric image features are extracted before using pose consistency constraints to confirm or discard competing match hypotheses. However, they largely ignore the rich source of information contained in the image brightness and/or color pattern. On the other hand, appearance-based methods, originally proposed in the context of face recognition (Turk and Pentland [191]; Pentland et al. [192]; Belhumeur et al. [193]) and 3-D object recognition (Murase and Nayar [194]; Selinger and Nelson [195]), integrate geometric reasoning into a classical pattern recognition framework (Duda et al. [196]) which exploits the discriminatory power of relatively low-dimensional, empirical models of global object appearance in classification tasks. However, they typically deemphasize the combinatorial aspects of the search involved in any matching task, which limits their ability to handle occlusion and clutter [64]. Viewpoint and/or illumination invariants (or invariants for short) provide a natural indexing mechanism for object recognition tasks. Unfortunately, although planar objects and certain simple shapes, such as bilateral symmetries (Nalwa [197]) or various types of generalized cylinders (Ponce et al. [198]; Liu et al. [199]), admit invariants, general 3-D shapes do not (Burns et al. [200]). This is the main reason why invariants became less popular after an intense flurry of activity in the early 1990s (Mundy and Zisserman [201]; Mundy et al. [202]).

In our proposed algorithms to solve occlusion recognition problem, both appearance and geometric information are exploited. Because of the nature of occlusion problem, the local appearance provides an effective filter for selecting promising matching candidates, thus reducing feature interactions from different objects in the scene. On the other hand, the spatial relationships help the matching algorithms to discard geometrically inconsistent candidate matches, leading to a global decision for recognition. In the following sections, we shall first discuss correspondence by

spectral graph matching based on spatial relationships between features, and then followed by possible solutions for local feature interaction reduction.

2.5 Correspondence from Graph matching

When the internal structures of two data sets cannot be ignored, they are often considered not simply as data sets, but as two separate graphs. To find correspondences of features extracted in scene and model images, spatial relationships between them can be well formulated as a geometric graph, which forms the basis of graph matching. Correspondence by graph matching has been extensively studied in computer vision since the early 70's. Current work in computer vision using graph matching usually use local appearance information for node attributes of the graph and geometrical relationships (described by distances and angles) for higher order attributes [13] [9] [19] [53]. This is the reason why graph matching fits our algorithms to encode both local appearance and spatial relationships between features for occluded object recognition.

During the past years, improvements in graph matching have been made mainly in three aspects: graph model, similarity function and optimization. In the next section, we shall review briefly previous researches which are relevant to our study, i.e. similarity measurements and spectral approximation.

2.5.1 Different similarity measures of graphs

Previous graph models with relatively weak unary and pairwise attributes did not include as much relevant information as possible, such as discriminative local features or powerful geometric relationships, other than the simple weights or pairwise distances on edges. Different similarity measurements have been defined for graph

establishments. Some of the early work such as [18] compared structural representations by counting consistent sub-graphs. This measure was refined in [56]. Graphs were also compared using the string edit distance [110]. The edit distance between two graphs is defined as a weighted sum of the costs of applying different edit operators, such as insert, delete and re-label nodes and edges, in order to transform one graph into the other. The edit distance was extensively used and refined in graph matching [29] [28] [111]. Early work based on more principled statistical measures include the entropy distance for structural graph matching [112] and information theoretic approaches such as [24]. Most recent work in computer vision involving graph matching [13] [43] [15] use objective functions that fit the integral quadratic programming formulation or Semi-definite Programming (SDP) techniques [20] [21]. The similarity measures, such as the graph edit distance or counting the number of consistent sub-graphs, were sometimes too convoluted and lacked a clear, intuitive insight [83]. More importantly, they lack a configuration for multiple cues combination, namely coding only one type of pairwise constraint for every operation.

2.5.2 Spectral approximation for correspondence

Graph matching is in general NP-hard (Non-deterministic Polynomial-time hard). Many different approximate algorithms have been proposed, such as genetic algorithms [22] [114] [119] [123], probabilistic formulations [23] [120], expectation-maximization algorithms [46] [66] [149], decision trees [150], neural networks [151] [152] [153] and spectral techniques [154] [155] [156] [157] [158].

As an approximate solution of graph matching formulated as Integer quadratic programming, spectral matching brings several improvements over previous methods in [1-10], which made it quite popular for solving computer vision tasks [83]. It has

wide applicability because there is no constraint imposed on the type of unary and pairwise scores in similarity measurements. Spectral matching methods offer an attractive route to correspondence matching since they provide a representation that can be used to characterize graph structure at the global level [73].

Spectral graph matching requires entries to be positive and to increase with the quality of the agreement, which can be easily accommodated by all graph matching applications. Spectral approximations [13] [43] reformulate graph matching problem, as it drops the integer constraints in the optimization stage and then obtains the discrete solution by binarizing a continuous result. Therefore, it lowers the complexity of the problem from NP-hard to a deterministic low-order polynomial. These properties contribute to the popularity of spectral matching as the preferred choice for many computer vision applications. Leordeanu et al. [13] propose a spectral analysis method for promoting feature matching accuracy. Its solution is based on the expected geometric alignments between correct correspondences against the accidental alignments of incorrect ones. This idea inspires the appropriate design of meaningful first and second-order scores. Cour et al. [43] add affine constraints to the spectral matching formulation in [13] and propose a normalization procedure to improve the matching accuracy. However, these feature matching algorithms are often very accurate for features with the highest similarities, but the matching accuracy falls rapidly when the undesired match number increases, especially for contaminated data sets. Leordeanu et al. in [159] propose to approach the discrete solution iteratively in the optimization stage. Recent research [19] employs sparse reliable correspondence priors based on spectral matching to manually guide the feature matching process. These spectral algorithms are introduced in detail in Chapter 3.

Our matching algorithms take the configuration of spectral approximation of graph matching, which is formulated as integer quadratic programming. In this formulation, not only individual feature similarity but also pairwise agreements on candidate correspondences are taken into consideration. Therefore, local object appearance as well as the object geometric structure is captured. Finally, a global decision on recognition is reached based on locally gathered information.

2.6 Feature interaction reduction based on intermediate-level vision

As for occlusion recognition, spectral matching technique itself is not suitable, where the original presentation of an object is contaminated by occlusion. Occlusion still remains an open issue for graph matching. Therefore strategies which reduce adverse effects imposed by occlusion for spectral matching should be taken into consideration. Given the nature of occlusion recognition problem (features from different objects tend to interact), we propose to reduce feature interactions by integrating intermediate-level vision cues.

Many researchers have found that intermediate-level vision applications are helpful for higher level vision tasks, such as recognition. They can provide powerful priors and direct the higher level recognition on the right path. Efforts have been made to employ intermediate-level cues to benefit higher level vision topics by researchers as reported in [40] [113] [115] [116] [117] [118]. In our research, intermediate-level vision provides useful information to integrate multiple cues into our strategies of reducing feature interactions to solve occlusion recognition.

Our algorithms to reduce feature interactions are closely related to feature similarities measurements. These measurements indicate how likely features are part

of the same object can be established, resulting in the reduction of feature interactions. Our proposed algorithms in Chapter 4 are inspired by the intermediate-level cues such as perceptual grouping and image segmentation, where pairwise feature or pixel similarities are commonly involved.

2.6.1 Feature interaction reduction by perceptual grouping

Perceptual grouping is to group features which are likely to belong to the same object based on cues that do not include the knowledge about the specific object or object category. As recognized early in [44], the grouping process gives the recognition algorithm a rough idea on where objects are and how features should be grouped together. Without perceptual constraints imposed by low-level grouping, an object detection process can produce many false positives in a cluttered scene. With the help of reducing feature interactions by grouping, matching algorithm might ignore most of the irrelevant background clutter and objects without interest.

Perceptual grouping was recognized in computer vision earlier on as having the potential to improve both matching and recognition. Marr, [47] argued, based on neurophysiologic studies of human vision system, that vision system should be able to recover the 3-D shape of objects without knowledge about the objects' class or about the scene. It seems clear that for humans, knowledge about the object class and the scene are not necessary for shape perception [147]. However, if one is able to perceive 3-D shape of a scene with respect to its own reference frame, it would be relatively easy to figure out which features in the scene should be grouped together [147]. We consider grouping as one of the ways of constraining the matching/recognition search space by considering only features that are likely to

come from the same object, which could reduce the feature interactions when occlusion happens.

Grouping is based on cues such as color, texture, shape and perceptual principles that apply to most objects in general, regardless of their category or specific identity. Generally, there are two main types of perceptual grouping:

1. Geometry based: Gestalt-like cues (between line segments or contours), such as proximity, good continuation and parallelism/perpendicularity.

2. Appearance based: objects tend to have unique and relatively homogenous appearance distributions.

Most of the existing grouping methods tend to make a hard decision (yes or no) about which group the features belong to, such as [44] [54] [55] [92], but the truth is that sometimes it is impossible to divide features into their correct groups without the knowledge of the specific grouping level. For example, is the wheel of a bicycle a separate object or is it part of the whole bicycle? Both situations can be true, depending on what we are looking for. While perceptual grouping alone should correctly separate most objects, it sometimes does not have enough information to make the decisions. On the other hand, grouping is supposed to serve the higher level process, therefore it is important to keep most of the grouping information around and transmit it to the higher recognition processes. Therefore, it could be more practical for grouping information to be used as a possibility, indicating how likely the features belong to the same object (i.e., soft grouping), rather than a hard decision.

2.6.2 Feature interaction deduction based on image segmentation

While it is often considered a separate field within computer vision, image segmentation could be considered as a special case of perceptual grouping, i.e., grouping pixels or image patches according to their similarity of appearance. Segmentation algorithms have been developed to help recognize target objects by finding possible object regions [40] [115]. Intuitive segmentation methods, such as edge detection, histogram based method, region growing and region splitting, were sensitive to noise such as occlusion and quantization errors as well as the predetermination of the termination criteria. Current segmentation techniques take on a bottom-up approach (Normalized Cuts by Shi & Malik [8]; Jigsaw approach by Borenstein and Ullman [213]; Implicit Shape Model by Leibe and Leonardis [214]), where image properties such as feature similarity (brightness, texture, motion etc), boundary smoothness and continuity are used to detect perceptually coherent units.

Segmentation is considered to be beneficial to object recognition for constructing object hypotheses in practice from the image itself without assuming impractical amount of hypotheses. However, segmentation algorithms usually have bottom-up configuration with unary and pairwise terms that use information only from low and local levels. Without utilizing any prior about the scene, image segmentation responds to poor data conditions, such as shadows and occlusions. The main challenge is how to combine bottom-up segmentation and top-down object model to efficiently generate a (small) set of feasible object hypothesis. Since segmentation can provide relatively global information (regions) for object recognition, techniques have been developed to combine segmentation with recognition in [116] [117] [118], integrating possible object regions for recognition.

Graph based segmentation has been receiving increasing interests over the past years. The image is formulated as Markov Random Field or a weighted graph with edges measuring similarity between adjacent pixels. This graph is then partitioned into disjoint sets (segments) with homogeneity. Saliency based segmentation, i.e. salient region detection, [99] [97] [121] have proposed to segment regions that likely contain target objects. In [38], segmentation has been employed in feature space to segment feature clusters, where efforts have been made to cluster features that are likely from the same object, other than finding actual object boundaries.

2.7 Conclusions

-a glimpse to our proposed algorithms

Existing occlusion recognition algorithms and current image representations in object recognition system have been reviewed. For our solution to occlusion recognition problem, relevant work in graph matching and intermediate-level vision cues are investigated. The knowledge gained from the reviews described in this chapter forms the basis of the algorithms developed in this thesis.

In our work, we propose to exploit image information involving both geometry and appearance to facilitate occlusion recognition problem. The object global structure is preserved by local feature relationships under the configuration of spectral graph matching. Thus, our local to global concept for occlusion recognition is established. Local information and pairwise feature relationships are addressed throughout this thesis. Depending on the nature of occlusion problem, our algorithms propose to reduce feature interactions by integrating intermediate-level vision cues: soft grouping and possible object region segmentation. In our proposed algorithms, we will show

how our feature interaction reduction formulations work with spectral graph matching for occlusion recognition problem. In addition, we propose to integrate other vision applications, such as motion and stereo vision, into recognition, by which the cooperation between multiple vision tasks is established.

Chapter 3

Spectral correspondence by pairwise feature geometry

There are many tasks in computer vision which require efficient techniques for finding consistent correspondences between two sets of features, such as object recognition, shape matching, wide baseline stereo vision, and 2-D and 3-D image registration. The problem of correspondence is referred to the finding of a mapping between one set of data and another set of data. When the internal structure of these data sets cannot be ignored, they are often considered not simply as data sets, but as two separate graphs. As a result, the correspondence problem is commonly perceived as graph matching. In this setting, graph vertices represent features extracted from each instance (e.g. a scene image and a model image). Therefore, the problem of graph matching is to find a mapping between the two sets of nodes that preserves the relationships between them as much as possible.

In our work on recognition of occluded objects, we employ graph based method. The geometric relationships between features are represented by a graph. In doing so, the global structure of possible target objects or parts are retained. Accordingly, correspondence search between scene and model images will result in a global decision based on locally gathered information.

3.1 Correspondence from spectral approximation of graph matching

Geometric relationship between feature points has been enforced into correspondence problem [15] [43] [53], where feature geometric relationship can be

formulated as a spatial graph. Conventional graph matching methods rely on local structural decompositions; and correspondence analysis is achieved through the iterative propagation of local consistency constraints with the hope of achieving global consistency. Therefore, the main difficulty of the graph based correspondence problem is the combinatorial complexity. In Figure 3.1, an example of the categorical product graph is shown, and all possible matching is presented in the categorical product graph, as introduced in [19]. The accuracy is traded against the efficiency gains achieved by using increasingly localized structures.

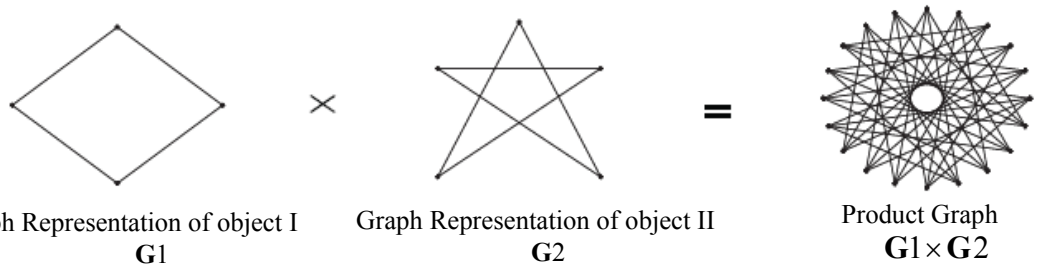


Figure 3.1 Combinatorial complexity of graph matching

As an approximate solution for graph matching, spectral graph theory is applied to a family of techniques that aims to characterize the global structural properties of graphs using the Eigen-values and Eigen-vectors of either the adjacency matrix or the closely related Laplacian matrix [1]. The main idea behind spectral graph theory is to use the distribution of Eigen-values to provide a compact summary of graph structure [73].

3.1.1 Notations and graph construction

Given two sets of features from two images to be matched which are denoted as $\mathbf{F}^1_{N_1\times 1} = \begin{bmatrix} f^1_1 & f^1_2 & \cdots & f^1_{N_1} \end{bmatrix}^T$ and $\mathbf{F}^2_{N_2\times 1} = \begin{bmatrix} f^2_1 & f^2_2 & \cdots & f^2_{N_2} \end{bmatrix}^T$ where $f^k_i = (des^k_i, x^k_i)$,

des_i^k and x_i^k are the i^{th} feature description vector and its location in the k^{th} image, with $k = \{1,2\}$, respectively.

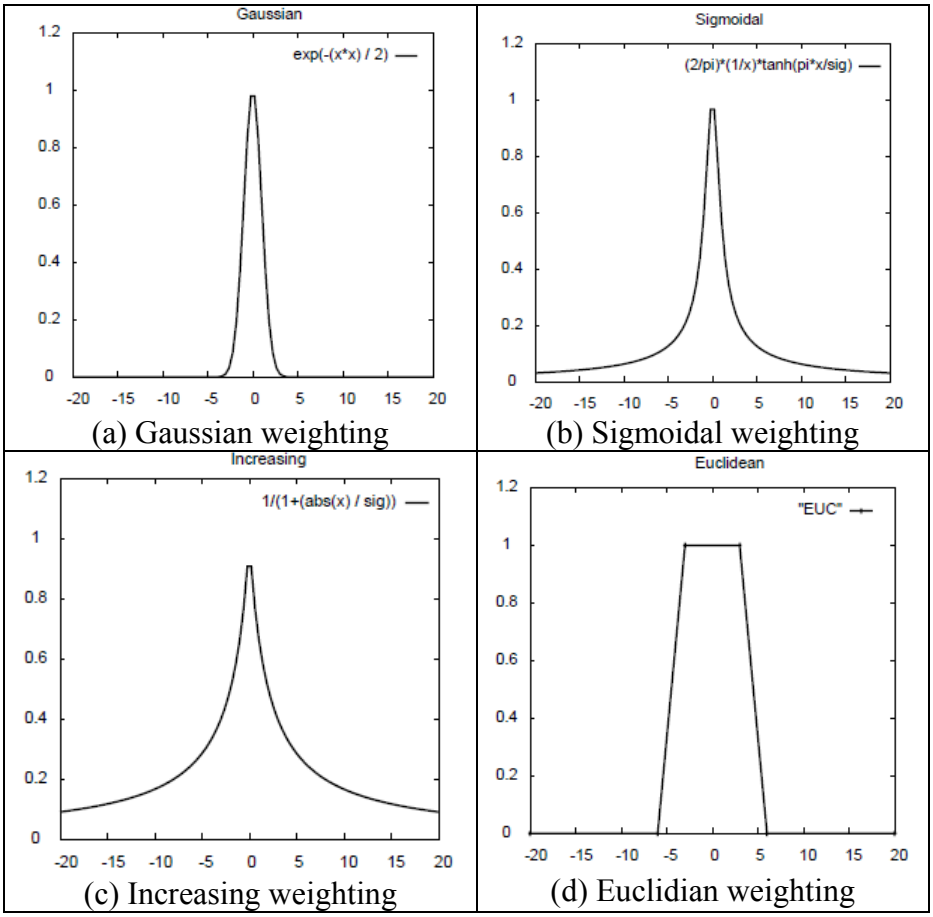
Traditionally in graph matching algorithm, an undirected spatial graph is defined as $\mathbf{G}^k = (\mathbf{V}^k, \mathbf{E}^k)$ with vertex set $\mathbf{V}^k$ and edge set $\mathbf{E}^k$ for the k^{th} image. The edges in $\mathbf{E}^k$ reflect the geometric neighboring relations among features in the k^{th} image, and is defined in terms of *n* nearest neighbor or an ε-ball distance criteria in the feature position space. In addition, the adjacency matrix is directly based on the edge information and each entry w^k is defined for the graph $\mathbf{G}^k$:

$$\text{Graph adjacency matrix: } w_{ij}^k = \begin{cases} 1 & \textit{If } x_i^k \textit{ and } x_j^k \textit{ are connected} \\ 0 & \textit{otherwise} \end{cases} \tag{3.1}$$

The spectral approach to correspondence commences by enumerating a proximity matrix $\mathbf{M}^k : (m_{ij}^k)$. This is an emphasized version of the graph adjacency matrix. Rather than setting the elements to unity or zero depending on whether or not there is a connecting edge between a pair of nodes, elements in the proximity matrix are weights that reflect the strength of a pairwise adjacency relationship (how the weights are determined will be explained in Section 3.1.2). The underlying fact is that in graph matching, global consistency is invariably assessed using only the edges of the graphs. However, the increased fidelity of representation achieved using a global spectral representation must be weighed against their relative fragility to the addition of noise and clutters.

3.1.2 Different weighting functions

Various alternative ways have been suggested to construct the proximity matrix. The role of the weighting function is to model the probability of adjacency relations between vertices in the graph. There are many choices of possible weighting functions described in the literature. However, they can be classified according to a broad-based taxonomy from their derivative [73].



(a) Gaussian weighting (b) Sigmoidal weighting

(c) Increasing weighting (d) Euclidian weighting

Figure 3.2 Weighting functions for constructing proximity matrix

If the derivative is monotonically increasing, then the weighting function is increasing. If the derivative is asymptotically constant, then the weighting function is sigmoidal. Finally, if the derivative asymptotically approaches zero then the weighting function is re-descending. In this section, we investigate several weighting

functions which fall into these different classes. The graphs of weighting functions are shown in Figure.3.2.

If i and j are two points and x_i and x_j are their positions, respectively, the corresponding element in the proximity matrix for these two vertices in a graph representation is given by different weighting functions as follows. σ controls the width of weighting kernel. In Figure 3.2, the *y*-axis is the variable m_{ij} and the *x*-axis is the variable *x*.

A. Gaussian weighting (Figure 3.2(a))

$$m_{ij} = \exp\left[-\frac{1}{2\sigma^2}\left\|x_i - x_j\right\|^2\right] \tag{3.2}$$

B. Sigmoidal weighting (Figure 3.2(b))

An example of hyperbolic tangent function is given for a sigmoidal weighting function:

$$m_{ij} = \frac{2}{\pi\left\|x_i - x_j\right\|}\log\cosh\left[\frac{\pi}{\sigma}\left\|x_i - x_j\right\|\right] \tag{3.3}$$

C. Increasing weighting (Figure 3.2.(c))

$$m_{ij} = \left[1 + \frac{1}{\sigma}\left\|x_i - x_j\right\|\right]^{-1} \tag{3.4}$$

D. Euclidean weighting function (Figure 3.2(d))

In order to investigate the effect of using a weighting function that decreases linearly with distance, we follow the formulation in [73]. A piecewise specification is used to define a trapezoid function and a Euclidean weighting function is formed based on it.

$$m_{ij} = K\left(\left\|x_i - x_j\right\|\right), \text{ where } K(\eta) = \begin{cases} 1 & if\ k < s_1 \\ 1 - \dfrac{1}{s_1 - s_2}[\eta - s_1] & if\ s_1 < k < s_2 \\ 0 & otherwise \end{cases} \tag{3.5}$$

3.1.3 Correspondences by Eigen decomposition

For formulated proximity matrices of two data sets, correspondences between these data sets can be located by Eigen decomposition method, as introduced below. Correspondences are located by comparing the ordered Eigen-vectors of the proximity matrices for different images. The Eigen-vectors of the proximity matrices can be viewed as the base vectors of an orthogonal transformation on the original data identities.

The modal structure for a data set is found by solving the Eigen-value equation

$$\left|\mathbf{M} - \lambda\mathbf{I}\right| = 0 \tag{3.6}$$

Where $\lambda_i, i = 1, \cdots, n$ is the i^{th} Eigen-value of matrix **M**, and its associated Eigen-vector $\mathbf{v}_i$ is computed using the equation

$$\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{3.7}$$

Eigen-vectors are ordered according to the values of their associated Eigen-values. The ordered column vectors are used to construct a modal matrix

$$\mathbf{Mo} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_n)^T$$

The column index of this modal matrix refers to the order of the Eigen-values while the row index is the index of the original dataset. This modal ordering process is repeated for both data sets to give two modal matrices:

$$\mathbf{Mo}_1 = (\mathbf{v}_1^1, \mathbf{v}_2^1, \cdots, \mathbf{v}_{n_1}^1)^T \text{ and } \mathbf{Mo}_2 = (\mathbf{v}_1^2, \mathbf{v}_2^2, \cdots, \mathbf{v}_{n_2}^2)^T$$

where $n_1 = |\mathbf{Mo}_1|$ and $n_1 = |\mathbf{Mo}_2|$. Since the two data sets are potentially with different sizes, we truncate the modes of the larger data set. This corresponds to removing the last $|n_1 - n_2|$ rows and columns of the larger matrix. The resulting matrix has $n = \min(n_1, n_2)$ rows and columns.

The modal matrices can be viewed as linear transformations of the original identities of the data sets for each image to the modal base representation. Row entries of the modal matrix correspond to the original data. The column entries measure how the original data identities are distributed among the different Eigen-modes. Therefore, two data sets are considered as correspondence if their values of the corresponding rows from the two modal matrices are close enough. Hence, correspondences can be located by searching for rows of the transformation matrices (modal matrices) which have the maximal similarity.

3.2 Integer quadratic programming for encoding pairwise relationships

3.2.1 Integer quadratic programming of graph matching

Mathematically, correspondence between two graphs in graph matching can be addressed as follow. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ and $\mathbf{G}' = (\mathbf{V}', \mathbf{E}')$ be two attributed graphs. To match these two graphs is to find a mapping between $\mathbf{V}$ and $\mathbf{V}'$ that best preserves the attributes between edges $w_{ij} \in \mathbf{E}$ and $w'_{i'j'} \in \mathbf{E}'$. Equivalently, we seek a set of correspondences, or matches $\mathbf{C} = \{i, i'\}$, so as to maximize the graph matching score, defined as:

$$\text{Max} \sum_{ii' \in \mathbf{C}, jj' \in \mathbf{C}} \text{S}(w_{ij}, w'_{i'j'}) \tag{3.8}$$

The function $S(\cdot,\cdot)$ measures the similarity between edge attributes. As a special term, $S(w_{ii}, w'_{i'i'})$ is simply the score associated with the match $\{i, i'\}$. Let $n = |\mathbf{V}|$ and $m = |\mathbf{E}|$ and likewise for n' and m'.

To employ pairwise information, Eq. (3.8) can be reformulated as an integer quadratic programming. Let us represent $\mathbf{C}$ as a binary matrix $c_{ii'} \in \{0,1\}$ with size $n \times n'$. $c_{ii'} = 1$ *iff* i matches i'. The matching constraint can be one-to-one or one-to-many. For one-to-one matching, it is $\sum_i c_{ii'} = 1$ and $\sum_{i'} c_{ii'} = 1$. Reshape elements of $\mathbf{C}$ to form a vector with the size of $(n \times n') \times 1$, which indicates a list of candidate correspondences. In general, the matching constraint (one-to-many) can be represented as an affine inequality constraint $\mathbf{AC} \le \mathbf{b}$. Then Eq. (3.8) can be formulated as quadratic programming:

$$\text{Max} (\mathbf{C}^T \mathbf{M} \mathbf{C}) \quad s.t. \quad \mathbf{AC} \le \mathbf{b}, \quad \mathbf{C} \in \{0,1\}^{nn'} \tag{3.9}$$

$\mathbf{M}$ is a $(n \times n') \times (n \times n')$ proximity matrix and each entry indicates relationships between two pairs of the candidate correspondences, i.e. $m_{ii',jj'} = S(w_{ij}, w'_{i'j'})$. Therefore, pairwise geometric relationships is integrated to form proximity matrix. In the quadratic programming, it takes into consideration both unary and second-order terms. This reflects the similarities in local appearance (candidate correspondences), as well as in the pairwise geometric relationships (pairwise agreements) between the matched features [159]. Actually, pairwise constraints in feature matching are gaining a widespread usage in computer vision, especially in object matching and recognition [15] [43].

3.2.2 Proximity Matrix M from pairwise geometry

To encode the pairwise feature relationships between two sets of features, we introduce the similarity graph, denoted as a triplet $\mathbf{G}^{12} = (\mathbf{V}^1, \mathbf{V}^2, \mathbf{E}^{12})$, The similarity graph $\mathbf{G}^{12}$ is a bipartite graph, and the proximity matrix $\mathbf{M}$ of $\mathbf{G}^{12}$ are computed from geometric agreements between feature pairs. Within this similarity graph, nodes are the potential correspondences $(f_i^1, f_{i'}^2)$ and weights on edges are the agreements between pairs of potential correspondences, $\{(f_i^1, f_{i'}^2), (f_j^1, f_{j'}^2)\}$, presenting pairwise geometric agreements between potential correspondences, shown in Figure 3.3. Note that for brevity, i and j are the i^{th} and j^{th} features in Image 1, represented by f_i^1 and f_j^1; i' and j' represent the features $f_{i'}^2$ and $f_{j'}^2$ in Image 2, respectively. This graph can be described mathematically by its proximity matrix $\mathbf{M}$, where entries represent agreements between two pairs of correspondences.

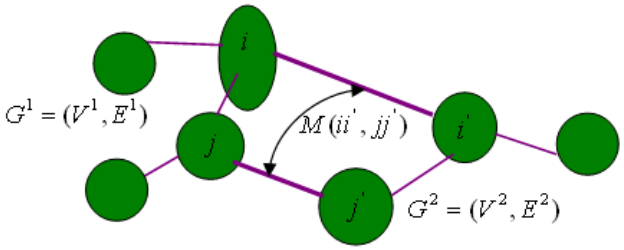


Figure 3.3 Similarity graph based on pairwise feature interactions

Specifically, each entry of $\mathbf{M}$ involves two pairs of candidate correspondences, i.e., off diagonal elements of $\mathbf{M}$ contains the relationships between the feature pairs $\{(f_i^1, f_{i'}^2), (f_j^1, f_{j'}^2)\}$ and diagonal terms represent the individual similarity between features f_i^1 and $f_{i'}^2$.

Features mentioned here could consist of points, lines, shape descriptors or points

of interest, depending on specific applications. For problems where the features are non- discriminative (e.g. points), it is the features pairwise geometric information that helps in finding the correct correspondence. When discriminative features are extracted (e.g. points of interest) then both the geometry and the properties of each individual feature can be used. In the meanwhile, our formulation can accommodate different kinds of correspondence mapping constraints, such as one-to-one, allowing a data feature to match at most one model feature (commonly used), or one-to-many, allowing a feature from one data set to match several features from the other data set (used in shape matching [15]).

3.2.3 Spectral approximation for integer quadratic programming

Correspondence problem between graphs using pairwise relationships is now formulated as an integer quadratic programming. To obtain the correspondence is to find the indicator vector $\mathbf{C}$ that maximizes the quadratic score function, in Eq. 3.10. It can be addressed in the most recent and general form (with one-to-one matching constraint):

$$\mathbf{C}^* = \arg\max(\mathbf{C}^T\mathbf{M}\mathbf{C}) \qquad s.t.\ \mathbf{A}\mathbf{C} = 1,\ \mathbf{C} \in \{0,1\}^n \tag{3.10}$$

where $\mathbf{A}\mathbf{C} = 1$ is the one-to-one matching constraints and $\mathbf{A}$ is a constraint matrix; $\mathbf{C} \in \{0,1\}^n$ is the integer constraints and n is the number of candidate correspondences.

Eq. (3.10) requires that $\mathbf{C}$ is an indicator vector such that $c_{ii'} = 1$ if feature f_i^1 is matched to feature $f_{i'}^2$ and zero otherwise. Different matching constraints, such as many-to-one, can be accommodated in the same formulation, by appropriately setting the constraints matrix $\mathbf{A}$. Usually one-to-one constraints are imposed on $\mathbf{C}$ such that

one feature from one image can be matched to at most one other feature from the other image.

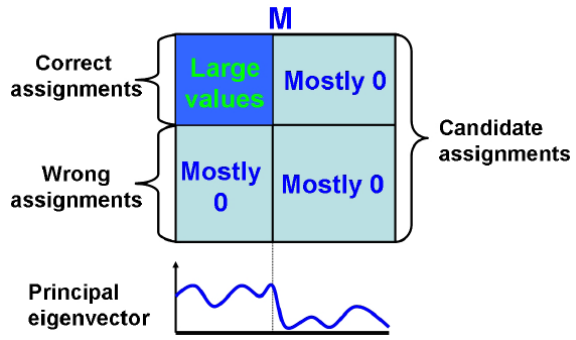
The difficulty of this problem depends on the structure of the matrix **M**, but in most cases it is NP-hard. There is no efficient algorithm that can guarantee optimality bounds. A lot of efforts have been spent in finding good approximate solutions by relaxing the integer one-to-one constraints in order to be able to find optimal solutions to the new problem efficiently [83]. Spectral approximation of this problem, referred to as spectral matching, takes advantage of the particular properties of geometric matching. It is expected that the optimum solution of the relaxed problem to be close to that of the original problem with integer constraints (Eq. (3.10)). We shall show several leading spectral approximation algorithms for this problem.

Spectral Matching (SM) in [13] drops the constraints entirely and only incorporates it during the discretization step. The resulting objective function is:

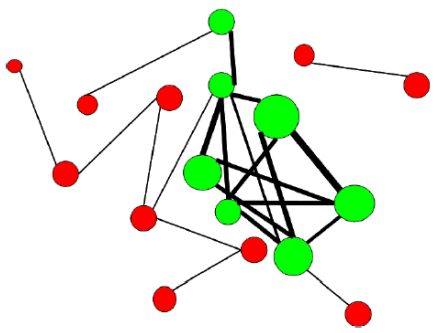
$$\mathbf{C}^* = \arg\max(\mathbf{C}^T\mathbf{M}\mathbf{C}) \qquad s.\,t.\ \ \|\mathbf{C}\| = 1 \qquad (3.11)$$

Since only the relative values between the elements of **C** matter, the norm of **C** is fixed 1. From the Rayleigh's ratio theorem, **C** that will maximize the dot-product $\mathbf{C}^T\mathbf{M}\mathbf{C}$ is the principal Eigen-vector of **M**. Since **M** has non-negative elements, by Perron-Frobenius theorem, the elements of **C** will be in the interval [0, 1]. Then the continuous solution of **C** is binarized by maximizing the dot-product with the leading Eigen-vector of **M**. The assumption is that **M** is a slightly perturbed version of an ideal matrix, with the rank =1, for which maximizing this dot product will give the global optimum of Eq. (3.11). Ideal matrix with a rank = 1 is shown in Figure 3.4, where correct correspondences are strongly connected to form the main cluster in **M**

and wrong correspondences are weakly connected to this main cluster.



(a) main cluster formed by correct correspondences



(b) weak connection between wrong correspondences and the main cluster

Figure 3.4 Ideal matrix with rank = 1

Spectral Graph Matching with Affine Constraints (SMAC) was developed later in [43] , which determines the optimal solution of a modified score function, with a tighter relaxation that imposes the affine constraints $\mathbf{AC} = \mathbf{b}$ (general form of mapping constraint) during optimization and dropping the binary constraints on C.

$$\mathbf{C}^* = \arg\max \frac{\mathbf{C}^T \mathbf{M} \mathbf{C}}{\mathbf{C}^T \mathbf{C}} \qquad s.t. \quad \mathbf{AC} = \mathbf{b} \qquad (3.12)$$

Their proposed solution is given by the leading Eigen pair of $\mathbf{P}_A \mathbf{M} \mathbf{P}_A \mathbf{C} = \lambda \mathbf{C}$, where $\mathbf{C}$ is scaled so that $\mathbf{AC} = \mathbf{b}$; $\mathbf{P}_A = \mathbf{I}_n - \mathbf{A}_{eq}^T (\mathbf{A}_{eq} \mathbf{A}_{eq}^T)^{-1} \mathbf{A}_{eq}$, $\mathbf{A}_{eq} = [\mathbf{I}_k, 0](\mathbf{A} - (1/\mathbf{b}_k)\mathbf{b}\mathbf{A}_k)$; $\mathbf{A}_k$ and $\mathbf{b}_k$ denote the last row of $\mathbf{A}$, $\mathbf{b}$, respectively; and k is the number of constraints. An important aspect is that $\mathbf{P}_A$ is in general, a full matrix even when $\mathbf{M}$ is sparse. Therefore, special care has to be exercised in the actual implementation by using Sherman-Morrison formula (for one-to-one matching) for operations of the type $\mathbf{Y} = \mathbf{P}_A \mathbf{C}$. The discrete solution is obtained to binarize the continuous $\mathbf{C}$.

3.2.4 Efficient Integer Projected Fixed Point algorithm (IPFP)

Spectral matching algorithms are sensitive to contaminations which could cause spurious connections or edges in the graphical model, and hence seldom leave a matrix with a rank of 1. Therefore, the approximated solution may deviate a lot from the actually global optimum. This is also the reason why the performances of spectral matching algorithms decline dramatically with increasing noise. In occluded object recognition task, noisy data will be expected. Therefore, a robust to noise and efficient matching algorithm is of great interest to us.

It was investigated in [83] that none of the previous spectral methods are concerned with the original integer constraints during optimization. Little computational time is spent in order to binarize the solution, based on the assumption that the continuous optimum is close to the discrete global optimum of the original combinatorial problem [83]. Their final post processing steps are usually a simple procedure where the continuous solution is binarized. In fact, the solution returned by current state-of-the-art algorithms, including spectral matching, can be significantly improved if special care is taken into consideration during the post-processing, i.e., binarization step.

Based on these investigations, an iterative algorithm can probably improve any continuous or discrete solution rapidly offered by some other graph matching methods. The iterative algorithm works by optimizing the original problem with its integer constraints. Each iteration consists of two stages. The first one optimizes in the discrete domain, a linear approximation of the quadratic function around the current solution. It gives a direction along which the second stage maximizes the original quadratic score in the continuous domain. The stage two can be viewed as a projection on the discrete domain and this algorithm is called Integer Projected Fixed

Point (IPFP) algorithm [83]. It aims to optimize the following continuous problem, in which the integer constraint from Eq. (3.10) is removed:

$$\mathbf{C}^* = \arg\max(\mathbf{C}^T\mathbf{M}\mathbf{C}) \qquad s.t.\ \ \mathbf{A}\mathbf{C} = 1,\ \ (c_i \in \mathbf{C}) \geq 0 \tag{3.13}$$

Table 3.1 Integer Projected Fixed Point program

Step 1. initialize $\mathbf{C}^* = \mathbf{C}_0$, $\mathbf{S}^* = \mathbf{C}_0^T\mathbf{M}\mathbf{C}_0$, $k = 0$, where $c_i \geq 0$ and $c_i \neq 0$
Step 2. let $\mathbf{B}_{K+1} = \mathrm{L}_d(\mathbf{M}\mathbf{C}_k)$, $\widetilde{A} = \mathbf{C}_k^T\mathbf{M}(\mathbf{B}_{k+1} - \mathbf{C}_k)$, $\widetilde{B} = (\mathbf{B}_{k+1} - \mathbf{C}_k)^T\mathbf{M}(\mathbf{B}_{k+1} - \mathbf{C}_k)$
Step 3. if $\widetilde{B} \geq 0$, set $\mathbf{C}_{k+1} = \mathbf{B}_{k+1}$, else let $r = \min\{-\widetilde{A}/\widetilde{B}, 1\}$ and set
$\mathbf{C}_{k+1} = \mathbf{C}_k + r(\mathbf{B}_{k+1} - \mathbf{C}_k)$
Step 4. if $\mathbf{B}_{k+1}^T\mathbf{M}\mathbf{B}_{k+1} \geq \mathbf{S}^*$, then set $\mathbf{S}^* = \mathbf{B}_{k+1}^T\mathbf{M}\mathbf{B}_{k+1}$, and $\mathbf{C}^* = \mathbf{B}_{k+1}$
Step 5.if $\mathbf{C}_{k+1} = \mathbf{C}_k$, stop and return the solution $\mathbf{C}^*$
Step 6. set $k = k + 1$ and go back to step 2.

Table 3.1 shows the steps of IPFP algorithm. In step 1, the quadratic score $\mathbf{C}_k^T\mathbf{M}\mathbf{C}_k$ is first approximated by the first-order Taylor expansion around the current solution $\mathbf{C}_k$: $\mathbf{C}^T\mathbf{M}\mathbf{C} \approx \mathbf{C}_k^T\mathbf{M}\mathbf{C}_k + 2\mathbf{C}_k^T\mathbf{M}(\mathbf{C} - \mathbf{C}_k^T)$. In step 2, two stages are introduced. Stage one: the continuous approximation is maximized within the discrete domain by the projection L_d under the one-to-one discrete constraints. Since all possible discrete solutions have the same norm, L_d boils down to finding the discrete vector $\mathbf{B}_{k+1} = \arg\max \mathbf{B}^T\mathbf{M}\mathbf{C}_k$. Stage two: the same discrete B_{k+1} maximizes the dot-product in the continuous domain, $\mathbf{A}\mathbf{C} = 1$, $\mathbf{C} > 1$.Step 3 and Step 4 are the updating rules. Finally, Step 5 is the termination criteria.

IPFP algorithm is in fact a relaxation problem of the original one by removing the

integer constraints. It is equivalent to the original problem, if the proximity matrix **M** is convex. The algorithm is basically a sequence of linear assignment (or independent labeling) problems, in which the next solution is found by based on the previous one. In practice, the algorithm converges in about 5 to 10 steps, which makes it very efficient. Step 3 ensures that the quadratic score increases with each iteration, and Step 4 guarantees that the binary solution returned is never worse than the initial solution.

The role of $\mathbf{B}_{k+1}$ is to provide a direction of largest possible increase (or ascent) in the first-order approximation, within both the continuous domain and the discrete domain simultaneously. Along this direction the original quadratic score can be further maximized in the continuous domain (as long as $\mathbf{B}_{k+1} \neq \mathbf{C}_k$). At step 3 we determine the optimal point along this discrete direction, also inside the continuous domain. The hope is that the algorithm will tend to converge towards discrete solutions that are (or are close to) the maxima of the relaxed problem.

Theoretical properties of IPFP have been analyzed. It is shown that IPFP has strong convergence and climbing guarantees, which is stated in [83] as follows:

1. The quadratic score $\mathbf{C}_k^T \mathbf{M} \mathbf{C}_k$ increases at every step k and the sequence of $\mathbf{C}_k$ converges.
2. The algorithm converges to a maximum of the relaxed problem.
3. If **M** is positive semi-definite with positive elements, then the algorithm converges in a finite number of iterations to a discrete solution, which is a maximum of the relaxed problem.

4. if **M** has non-negative elements and rank=1, then the algorithm will converge and return the global optimum of the original problem after the first iteration.

3.3 Performance evaluations of matching algorithms

In this section, the performances of the different graph matching algorithms are investigated both on synthetic images and real images. There are two aspects of these experimental studies. Firstly, the effects of varying weighting functions are investigated, from which the elements of the proximity matrix are calculated. Secondly, the robustness of these algorithms under occlusion and outliers is evaluated. In the following description, inliers are commonly referred to those points in a scene data set, which have corresponding points in the model data set. Conversely, those points without corresponding points in the model data set are known as outliers.

3.3.1 Choice of weighting functions

Our experiments are conducted with random point sets. In the first set of experiments, given a model data set Q of 2-D points, n_Q^i inliers are randomly selected in a given region of a plane. The n_Q^i points from Q are corrupted independently with white Gaussian noise $N(0,\sigma)$, and are then rotated and translated along with the remaining group of points to form the scene data set P. n_Q^o and n_P^o outliers are added to Q and P respectively, by randomly selecting points in the same region as the inliers in Q and P. These outliers have the same random uniform distribution over the x-y plane. The total number of points in Q and P are $n_Q = n_Q^i + n_Q^o$ and $n_P = n_P^i + n_P^o$, respectively The two data sets are shown within the purple and green rectangles in Figure 3.5.

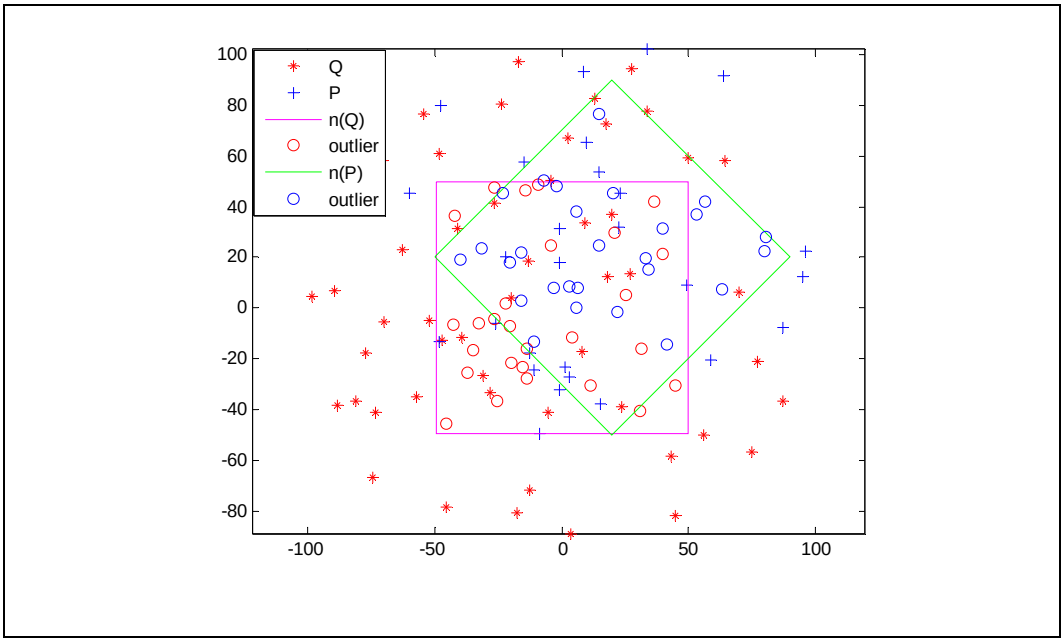


Figure 3.5 Data generation for testing different weighting functions

This is a difficult type of problem for two reasons. Firstly, the points are non-discriminative and they can be translated and rotated arbitrarily, so any of the points from P can potentially match any of the n_Q model points from Q. This maximizes the search space (the solution vector will have $n_Q \times n_P$ elements) and leaves the task of finding a solution entirely to the relative geometric information between pairs of points. Secondly, picking points randomly in the plane creates homogeneous data sets with an increased level of symmetry, which increases the chance of accidental agreements between wrong correspondences or between correct correspondences and incorrect ones. In addition, choosing outliers from the same distribution as the inliers, within the same region, increases the chance that geometric relationships are formed among outliers or between the outliers and the inliers. These geometric relationships could be similar as those built among the inliers only.

By matching the synthetic model and scene data sets, we investigate first the effect of varying weighting functions in the establishment of proximity matrix **M**. The aim is to determine which of the weighting functions returns correspondences

which are most robust to random measurement error or point-position jitter.

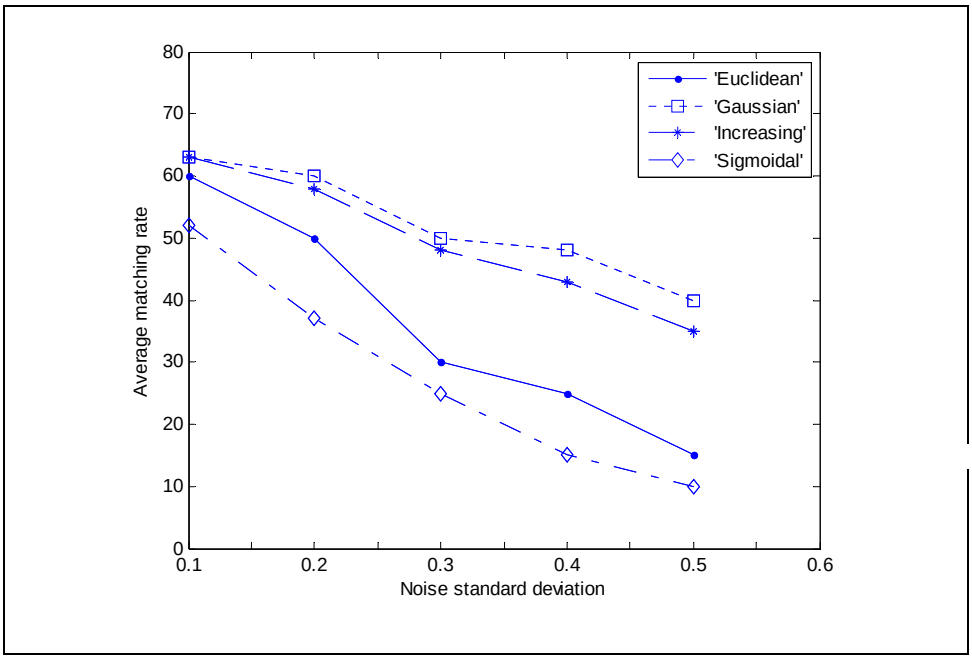


Figure 3.6 Average matching rate with different weighting functions

Figure 3.6 shows result of matching the two data sets. The *x*-axis is the standard deviation of the noise σ. Average matching rate given on the *y*-axis is the average performances of the three spectral algorithms (SM, SMAC, IPFP), on data set of 50, 100, 150 points and ratio of outliers/inliers = 50% in each set of *P* and *Q*. Figure 3.6 clearly shows that the best performance is obtained with the highest matching rates through the use of the Gaussian weighting function.

3.3.2 Robustness to occlusion and noise

With Gaussian weighting function, the performances of matching algorithms under occlusion are further investigated. Scene data is generated by corrupting the model data with noise and occlusion. In this regard, we introduce two parameters ρ and τ, both in percentage. ρ controls the percentage of inliers and τ represents the percentage of outliers. For a model set with *n* points, we control the rate of occlusion

with varying percentage number of ρ and τ, as follows:

1. Discarding a subset of size $(1-\rho)*n$ from the model set;
2. Perturbing the remaining data set with white Gaussian noise $N(0,\sigma)$;
3. Applying a rigid transformation (**R**, **T**) to the data set;
4. Adding $(\tau-\rho)*n$ spurious, uniformly distributed points;

Thus, after the above steps, the scene data set will have a total of $\tau*n$ points, among which only $\rho*n$ are inliers. The inliers are represented by the red '+'. These steps are illustrated in Figure 3.7.

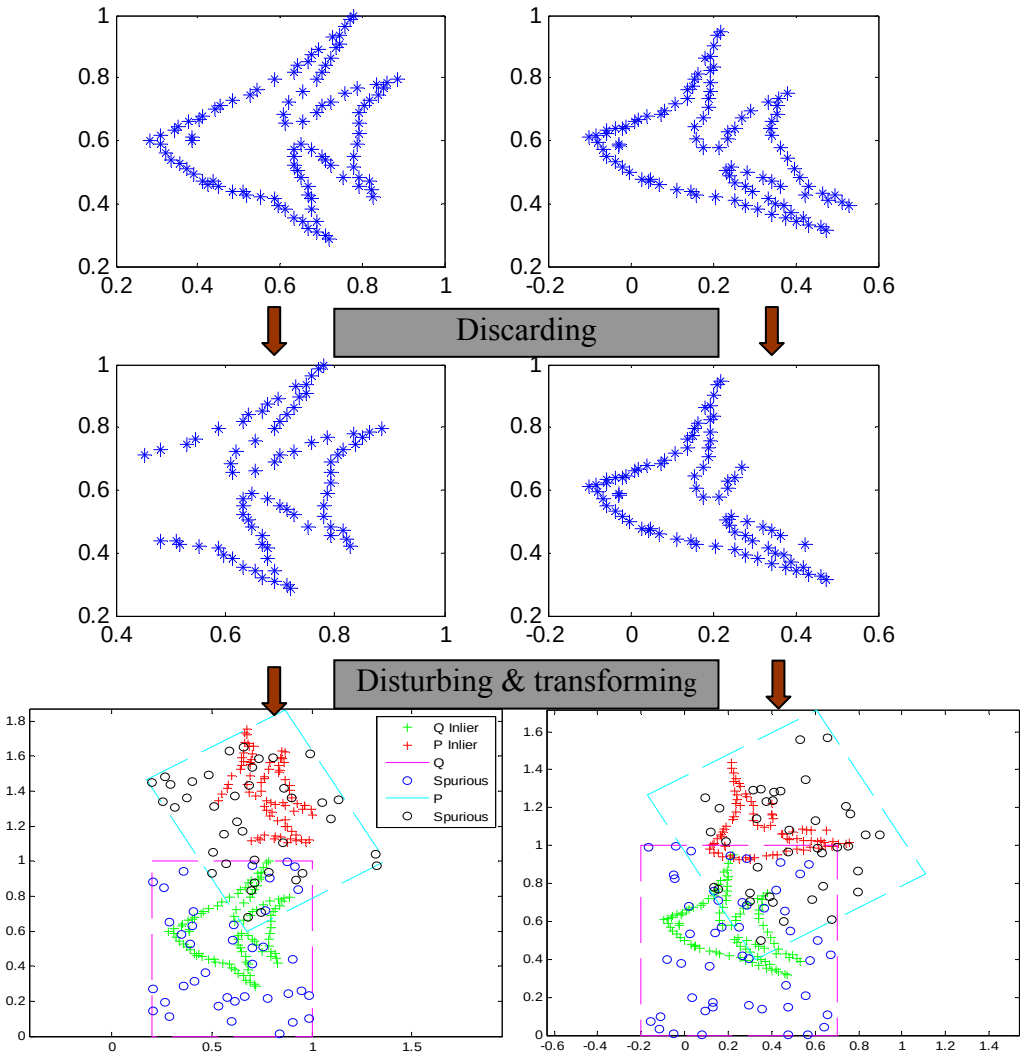


Figure 3.7 Generating corrupted scene data set from model sets

There are two data pairs generated in Figure 3.7. In our experiments, five models (fish model and Chinese characters from [122]) are selected. By varying occlusion rates $\rho \in [0.2,0.5]$ and outlier strength $\tau \in [1,1.5]$, six corrupted scene data sets are generated for each model. In total, there are thirty data pairs for carrying out the matching experiments.

With these data pairs, performances of four state of the art matching algorithms are compared: Spectral Matching (SM) [13], Spectral Matching with Affine Constraints (SMAC) [43], Graduated Assignment (GA) [123], Semi-definite Programming (SDP) [124] and Integer Projected Fixed Point algorithm (IPFP) [159]. First we form the pairwise proximity matrix. We left the matching score entirely to the pairwise geometric information, (namely diagonal elements are zero since there is no information on the individual correspondences), since points are non-discriminative. The off diagonal elements of the matrix represent agreement between candidate correspondences $(f_i^1, f_{i'}^2)$ and $(f_j^1, f_{j'}^2)$, we use the pairwise distances between points as [13]:

$$m(ii', jj') = \begin{cases} 4.5 - \dfrac{(d_{ij} - d_{i'j'})^2}{2\tilde{\sigma}^2} & if \left| d_{ij} - d_{i'j'} \right| < 3\tilde{\sigma} \\ 0 & otherwise \end{cases} \tag{3.14}$$

where d_{ij} is the Euclidean distance between points f_i and f_j; The parameter $\tilde{\sigma}$ controls the sensitivity of the agreement on deformations. We set $\tilde{\sigma} = 2$, as suggested in [13] for comparison.

Figure 3.8 shows the performances of the five algorithms. For each data pair, the average matching rate for each algorithm is collected over different noise levels

($\sigma \in [0.2,0.5]$). The strength of corruption is defined as $(\tau - \rho)/\rho$, which is the joint effects of occlusion and outliers.

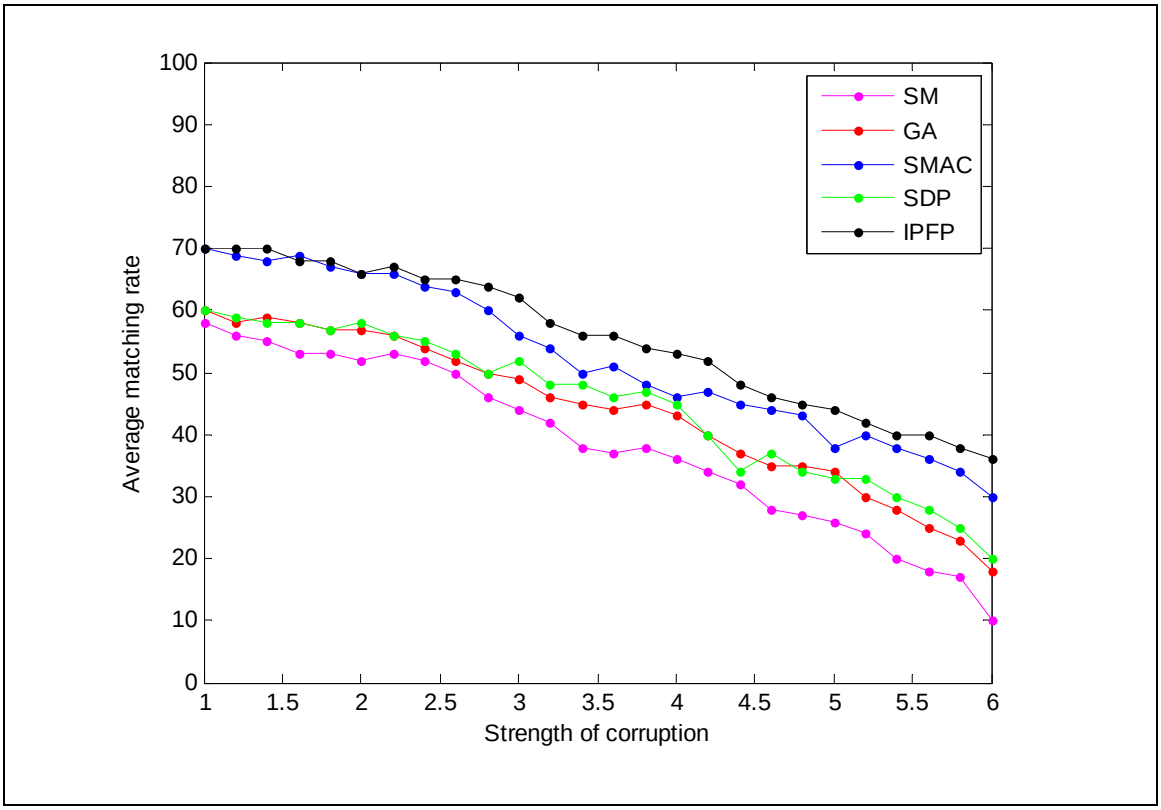


Figure 3.8 Comparison of matching performances

In all cases (GA, SDP, SM, SMAC and IPFP), the proximity matrix is normalized. Performances in matching using SMAC and IPFP algorithms are comparably good in our experiment settings. SDP is the most computationally expensive, because it involves the "squaring" of proximity matrix; and the most efficient algorithm is IPFP with less than 10 iterations to reach convergence. It is also noticed that when the corruption strength is low, the presented methods have strong resistance to noise. However, we observe that when the occlusion rate is large, methods exhibit weak resistance to image corruptions.

3.3.3 Spectral matching with IPFP as a post-processing step

We further investigate the performances of the matching algorithms using real images, where IPFP will be used as a post-processing step. Experiments are carried out on representative data set. We used cars and bicycles image pairs, extracted from the Pascal 2007 database [134], and face image pairs from Caltech-4 database[1].

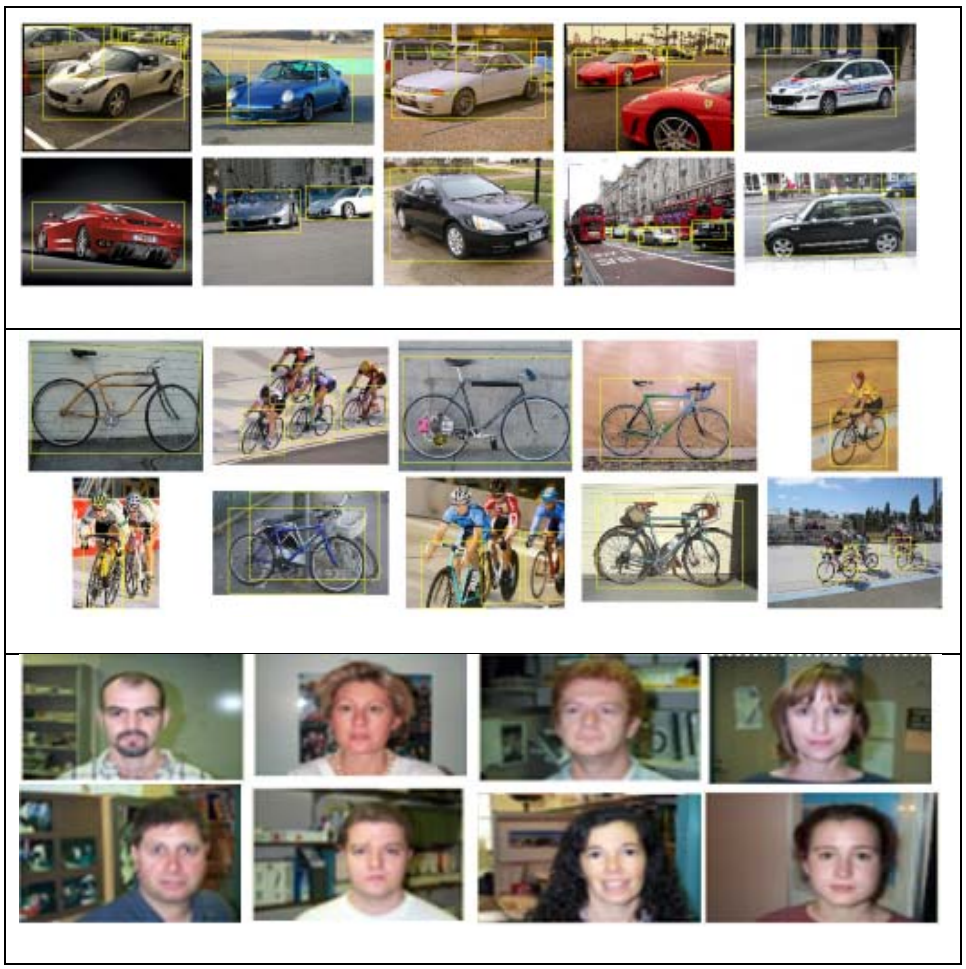


Figure 3.9 Sample images from Pascal 2007 and Caltech-4 database

We test on cars and bicycle dataset, following the same experimental setup with outliers as in [86]: all outlier features are allowed in the reference image, but no outliers are allowed to be present in the other image. These experiments are challenging due to large intra-category variations in shape in the Pascal 2007 database

[1] http://www.robots.ox.ac.uk/ vgg/data3.html

and also the presence of the large number of outliers (on average 5 times more outliers than inliers). The ground truth correspondences were manually selected. Features are extracted by sampling the contours. The pairwise agreement is calculated based on angle and distance relationship preservation, which will be defined by Eq. (4.9) in Chapter 4. Steps to calculate matching rate for an algorithm are listed in Table 3.2.

Table 3.2 Steps to calculate matching rate for a matching algorithm

Input: the i^{th} pair of image ($i = 1, \cdots, N$)

1. Manually select inliers, obtaining number of ground Truth correspondences: N_{gt};
2. Apply graph matching algorithm, obtaining number of correspondences: N_{ma};
3. Matching rate MR_i for i^{th} pair of image: $MR_i = \frac{\#(N_{ma} \in N_{gt})}{N_{ma}} \times 100\%$;
4. $i = i + 1$, repeat step 1 to 3.

Output : Matching rate for the matching algorithm: $MR = \frac{\sum_{i=1}^{N} MR_i}{N}$

Performances of the matching algorithms are averaged over *N=30* pairs of images, shown in Table 3.3. The difficulty of the matching problems is reflected by the relatively low matching rate of all algorithms, especially on the cars and bicycles image sets. It is also shown that graph matching algorithms are not desirable for corrupted data sets.

Table 3.3 Comparison of matching rates (%) on cars and bicycles datasets

Algorithms / Datasets	GA	SDP	SM	SMAC	IPFP
Cars	31.6	22.1	27.1	40.7	51
bicycles	33.2	27.4	29	34.5	35.2

Table 3.3 shows that IPFP consistently outperforms other algorithms in the

matching rate. On the other hand, we carry out the same experiments with IPFP as a post-processing step. Table 3.4 demonstrates that considerable improvements have been achieved with this post-processing step.

Table 3.4 Improvements of matching rates (%) with IPFP a post-processing step

Algorithms / Datasets	GA	SDP	SM	SMAC	IPFP
Cars: alone	31.6	22.1	27.1	**40.7**	51
Cars: + IPFP	43.2	47.4	39	**53**	51
Cars: Improvement	+11.6	+25.3	+11.9	**+10.3**	0
Bicycles: alone	33.2	27.4	29	**39.5**	43.2
Bicycles: + IPFP	42.1	38	40	**55**	43.2
Bicycles: Improvement	+8.9	+10.6	+11	**+15.5**	0
Faces: alone	45	40	47.1	**52**	60.2
Faces: + IPFP	59.4	57	59	**65.4**	60.2
Faces: Improvement	+14.4	+17	+11.9	**+13.4**	0

In Table 3.4, matching rates are given for the experiments with outliers on cars and bicycles from Pascal 07 database, and faces from Caltech-4 database. It is noted that IPFP by itself outperforms all the others without post-processing. When the solution of other algorithms becomes the input to IPFP, their performances are greatly improved. Among them, SMAC with IPFP as a post-processing step has the best performance.

3.4 Conclusions

According to the characteristics of occlusion recognition, we have chosen graph based correspondence because of its merits of finding global structure from local information. The general spectral matching theory is introduced. We employ the spectral correspondence based on the integer quadratic formulation to integrate

pairwise feature relationships. The two main components for spectral matching is the establishment of proximity matrix with different weighting functions and the approximation of discrete correspondences. Through our experiments setting, we have found that the Gaussian weighting function is the most suitable one to represent pairwise feature relationships and the iterative optimization formulation of IPFP outperforms other spectral matching algorithms. In particular, matching rates are dramatically improved for spectral algorithms with IPFP as a post-processing step. Therefore, we shall choose SMAC with a post-processing IPFP as the matching algorithm for our purpose of occlusion recognition.

Chapter 4

Reduction of feature interactions by pairwise appearance

Spectral matching algorithms have been introduced using pairwise feature geometry in the previous chapter. However, spectral matching itself is not suitable for occlusion recognition problem. In fact, occlusion has long been an open issue for graph matching algorithms, as ambiguous connections are generated by occlusion during the graph formulation. On the other hand, even though all the matching algorithms mentioned in Chapter 3 claim that they can accommodate different kinds of pairwise constraints, they do not have the possibility of integrating multiple pairwise constraints. Despite of the elegance of spectral matching, geometric information is not enough to represent object locally and form a comprehensive recognition decision. In our formulation, local appearance information is taken into consideration to work with spectral matching for recognizing occluded objects.

We have employed the local appearance to reduce feature interactions, which is addressed as a key issue in our solution to occlusion recognition problem. In this chapter, two methods are proposed to reduce feature interactions. They are Appearance Prior (A.P.) [125], measuring how likely features are from the same objects, and Feature Association (F.A.) [93], evaluating how well a feature is associated with the object of interest. Therefore, our approaches to reduce feature interactions have a bottom-up or top-down fashion.

In this chapter, we shall describe these two methods and show their effectiveness in

feature interaction reduction. Our frameworks to recognize occluded objects will be proposed in the next chapter by combining spectral matching with these strategies to reduce feature interactions.

4.1 Feature interaction reduction based on pairwise relationships

The performances of object recognition methods have been improved over the past years. However, they are far behind human ability to handle recognition under occlusions, which is subjected to interactions between features from different objects.

The state of the art algorithms in perceptual grouping and image segmentation [75] [76] could have reduced feature interactions, since ideally grouped features or segmented regions imply different objects, by which feature interactions from different objects could be eliminated. However, these algorithms are complicated procedures themselves and we are not looking for the exact object boundaries. Local relationships between features or pixels are important observations from grouping and segmentation, respectively. Recently, pairwise feature relationships are considered by various vision applications. For instance, occlusion boundaries are detected in [77] by comparing the motion information carried by patches from both sides of the extracted edges. This is based on the fact that image patches from the same object carry more similar motion status. In [38], features are grouped according to their similarity in appearance in order to constrain the matching process. Multiple cues are integrated for segmentation in [85], by considering the pixels relationships.

In fact, over or under grouping (segmentation) can easily occur, as the number of clusters or groups cannot be anticipated. There is a great chance for features not being completely classified. Therefore, the clustered features are “noisy” for the group-wise

matching as reported in [38]. Since the occlusion rate of the object of interest is also unpredictable, more "noisy" groups are expected when grouping methods are directly applied in clutter scenes. Therefore, a hard decision about grouping may not be helpful for the reduction of feature interactions.

In our algorithm, instead of being interested in forming exact feature groups based on perceptual grouping alone, we would choose to focus on the quality of pairwise relationships and determine how to integrate them into our recognition algorithm. Based on investigations in perceptual grouping, we are inspired to measure how likely features are from the same object by pairwise appearance similarity, which serves as priors in our occlusion recognition algorithms in order to reduce feature interactions. This idea is addressed as Appearance Prior, which is explained in detail in Section 4.3.

4.2 Feature association for feature interaction reduction

In fact, a smaller set of features which are more likely to be from object of interest could be suitable for our purpose to reduce feature interactions, thereby rejecting most background clutters. To implement this idea, a top-down procedure has been proposed by ranking features according to how well they are associated with the object of interest. The highly ranked features could be more likely to belong to object (s) of interest which is (are) partially visible in the scene. Our approach to associate feature points is inspired by the descriptor importance ranking method [145], where affinity between images was evaluated based on the number of shared important descriptors. However, their work did not produce point-to-point correspondence. Our formulation of feature associations carries physical meaning towards recognition, because the matching process will take place among feature points with high association scores.

In our formulation, patch information is used to associate detected feature points with object of interest. The reason is intuitive that features extracted from the target object in the scene should carry local patterns which can be generated by its model image. This feature association procedure to reduce feature interactions has a more specified purpose compared with the state of the art methods in perceptual grouping and image segmentation [25] [26]. Features with top association scores are selected in a top-down fashion, involving the object of interest. In [146], top-down segmentation relies on region associations with the model object but the image alignment is needed in advance.

The feature association is defined and implemented in Section 4.4, which is addressed as Feature Association for its ability to reduce feature interactions.

4.3 Feature interactions reduction by Appearance Priors

To reduce interactions between features from different objects, pairwise similarity information could provide priors, indicating how likely two feature points are from the same object. This probability information is referred to as Appearance Priors (A.P.).

Assume that features from the same object are more likely to carry similar local appearance, and then the pairwise appearance priors can be calculated by measuring the local appearance similarity carried by a pair of features.

For detected features in scene image $\mathrm{F} = [f_1, f_2, \cdots, f_n]^T$, where $f_i = \{des_i, (x_i, y_i)\}$, f_i and (x_i, y_i) are the i^{th} feature vector and its location, respectively. The corresponding description of each feature point has been

represented by its neighborhood $R_i, i \in [1,\cdots,n]$, also known as an image patch, in terms of color co-occurrence histograms CH_i and texture descriptor T_i. Accordingly, their appearance similarity is evaluated in terms of color and texture. I_{ij} and ρ_{ij} are the color similarity and texture similarity between features f_i and f_j, respectively, as shown in Figure 4.1. They will be explained in detail in the following subsections.

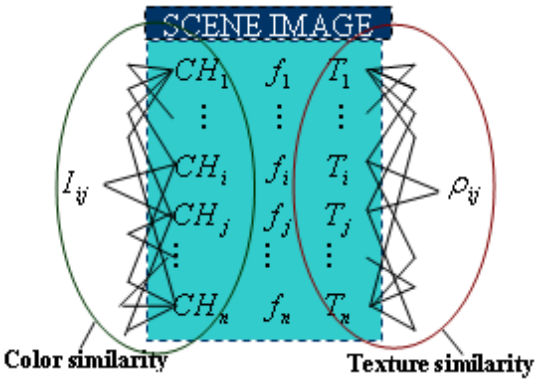


Figure 4.1 Appearance similarity in terms of color and texture

4.3.1 Color description by color Co-occurrence Histograms (CH)

The color co-occurrence histogram keeps track of the number of pairs of certain colored pixels that occur at certain separation distances in image space. As a holistic feature, the **CH** adds geometric information to the normal color histogram which abstracts away all geometry [51]. We represent $\mathbf{CH}_i$ for an image patch R_i (centered at (x_i, y_i)), where each entry counts the number of occurrences of color pairs $\tilde{c}_p$ and $\tilde{c}_q$, measured from a distance $\tilde{d}_k$, shown in Figure 4.2.

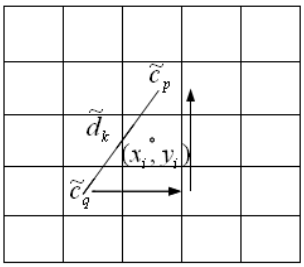


Figure 4.2 $\mathbf{CH}_i$ calculation in image patch R_i

We quantize the image into a set of $n_{\tilde{c}}$ representative colors $\tilde{\mathbf{C}} = \{\tilde{c}_1, \tilde{c}_2, \cdots, \tilde{c}_{n_{\tilde{c}}}\}$, after which every pixel is assigned a color from $\tilde{\mathbf{C}}$. The distances are predefined as a set of $n_{\tilde{d}}$ distance $\tilde{\mathbf{D}} = \{\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_{n_{\tilde{d}}}\}$. This quantization in color and the distance definition result in $\mathbf{CH}(\tilde{c}_p, \tilde{c}_q, \tilde{d}_k)$, where $\tilde{c}_p$ and $\tilde{c}_q$ are two colors from the color set $\tilde{C}$ and $\tilde{d}_k$ is the distance value in set $\tilde{D}$. For a given distance, the model image can directly be quantized by *k-means* algorithm with $n_{\tilde{c}}$ being the number of centers and Euclidean distance in normalized color space. Let $p_{\tilde{\xi}}, \xi \in [1, \cdots, n_i]$ be a random pixel in R_i, color co-occurrence matrix of R_i is defined as $\mathbf{CH}_i = (\tilde{c}_{pq})_{n_{\tilde{c}} \times n_{\tilde{c}}}$. Entry $\tilde{c}_{pq}$ specifies the probability of finding a pixel of color $\tilde{c}_p$ at a fixed distance $\tilde{d}_k$ from a given pixel of color $\tilde{c}_q$, as Markov Stationary Feature introduced in [126]:

$$\tilde{c}_{pq} = \#(p_1 = \tilde{c}_p, p_2 = \tilde{c}_q \,|\, p_1 - p_2 = \tilde{d}_k) / 2 \tag{4.1}$$

where $\tilde{d}_k$ indicates the distance between two colors $\tilde{c}_p$ and $\tilde{c}_q$; $\#(p_1 = \tilde{c}_p)$ denotes the number of pixels falling into histogram bin $\tilde{c}_p$. Subsequently, color Similarity between any pair of features is defined as the intersection of their corresponding **CH**s [51], using Eq. (4.2):

$$I_{ij} = \sum_{p=1}^{n_{\tilde{c}}} \sum_{q=1}^{n_{\tilde{c}}} \sum_{k=1}^{n_{\tilde{d}}} \min[(\mathbf{CH}_i(\tilde{c}_p, \tilde{c}_q, \tilde{d}_k), \mathbf{CH}_j(\tilde{c}_p, \tilde{c}_q, \tilde{d}_k))] \qquad i, j \in [1, \cdots, n] \qquad (4.2)$$

4.3.2 Texture similarity

The texture description of region R_i is denoted as T_i, responding to the covariance matrix of image derivatives in [127]. The ξ^{th} pixel in R_i, $\xi = [1,2,\cdots,n_i]$ can be mapped to a five dimensional feature vector f_ξ, using the first and second-order derivatives of the intensity *Img*, in both x and y direction.

$$f_\xi = [\frac{\partial \text{Img}}{\partial x}, \frac{\partial \text{Img}}{\partial y}, \frac{\partial^2 \text{Img}}{\partial x^2}, \frac{\partial^2 \text{Img}}{\partial y^2}, \frac{\partial^2 \text{Img}}{\partial xy}]$$

Texture descriptor T_i of R_i will have the form:

$$T_i = \frac{1}{n_i - 1} \sum_{\xi=1}^{n_i} (f_\xi - \overline{f})^T (f_\xi - \overline{f}) \qquad (4.3)$$

where $\overline{f}$ is the mean of vector f_ξ over region R_i.

This formulation has a comparative performance to the state of the art filter bank approaches in [161]. The texture similarity between two regions is defined by [127]:

$$\rho_{ij}(T_i, T_j) = \sqrt{\sum_{\varsigma=1}^{n_\lambda} [\ln \lambda_\varsigma(T_i, T_j)]^2} \quad , \; i, j \in [1, \cdots, n] \qquad (4.4)$$

Where $\lambda_\varsigma(T_i, T_j), \varsigma \in [1, \cdots, n_\lambda]; i, j \in [1, \cdots, n]$ are the generalized Eigen-values of T_i and T_j.

By combining color and texture information, we are able to separate groups of features with predefined number of clusters, as shown in Figure 4.3.

$$\tilde{p}(i,j) = \exp-(\frac{\rho_{ij}^2}{\sigma_1^2} + \frac{(I_{ij})^2}{\sigma_2^2}) \qquad i, j \in [1,\cdots,n] \qquad (4.5)$$

$\tilde{p}(i,j)$ indicates the appearance similarity between the i^{th} and j^{th} feature; I_{ij} is defined in Eq.(4.2); σ_1 and σ_2 are the weightings of each cue.

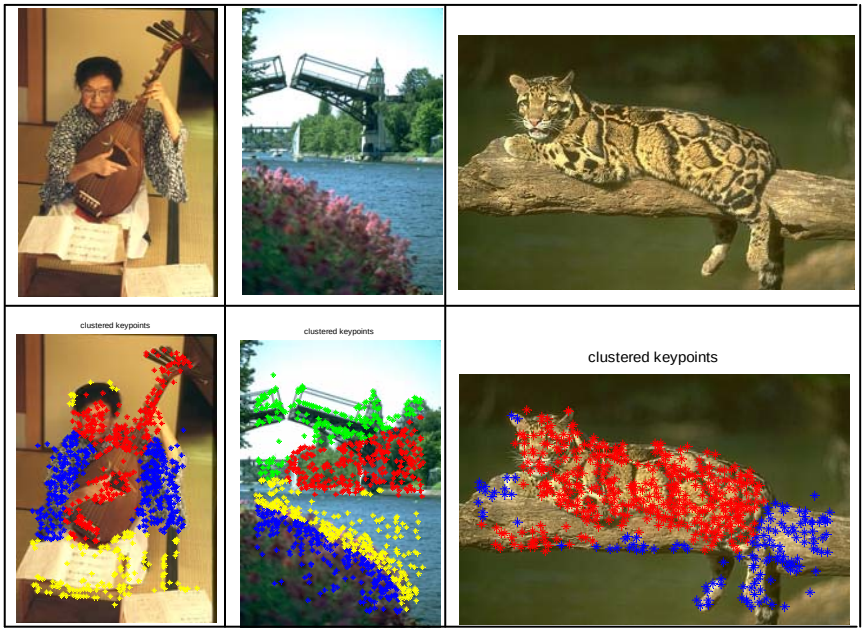
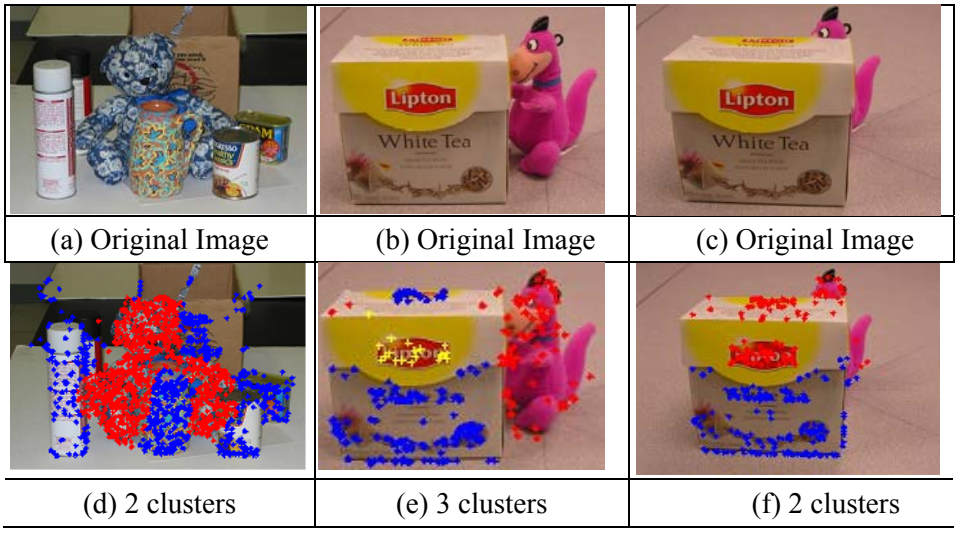
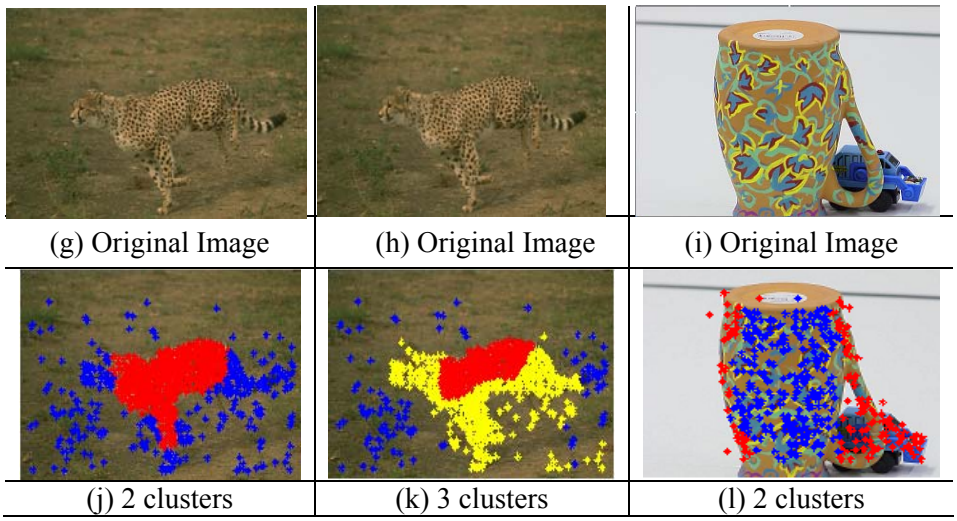


Figure 4.3 Clustering of feature points by color and texture

However, the predefined number of clusters could affect the clustering performance, as shown in Figure 4.4.



(a) Original Image	(b) Original Image	(c) Original Image
(d) 2 clusters	(e) 3 clusters	(f) 2 clusters



(g) Original Image | (h) Original Image | (i) Original Image

(j) 2 clusters | (k) 3 clusters | (l) 2 clusters

Figure 4.4 Appearance based feature clustering

For example, Figure 4.4 (g) and (h) are the same image. Their clustering results are inconsistent, shown in (j) and (k), from which the effects of predefined number of clusters are evident. Since no evidence could suggest that the entire object falls within one cluster, interactions between features from different objects still exist. Under occlusion, the remaining feature points of the object of interest can be easily clustered into a bigger group, as shown in (f) and (l). Therefore, appearance based feature clustering alone is not enough to guide the recognition of occluded objects. That is why the algorithm in [38] may not be suitable to handle occlusion problem. On the other hand, appearance difference can indeed serve as priors, indicating how likely features are from the same object.

4.4 Feature interactions reduction by Feature Association

4.4.1 Definition of Feature Association (F.A.)

For detected features in scene image $\mathrm{F} = [f_1, f_2, \cdots, f_n]^T$, $f_i = \{des_i, (x_i, y_i)\}$, des_i and (x_i, y_i) are the i^{th} feature vector and its coordinates, respectively. Intuitively, there are features which are more likely to come from the object of interest than the rest in an occluded scene. For features located on the object of interest, its

surrounding image properties (e.g. colors, intensity, and edge) are shared by its model image. As observed in Figure 4.5, the dinosaur is partially occluded, and image patches (defined in Section 4.3) are extracted to present local patterns. Some patches have higher co-occurrence with the model image based on the color information alone.

We define the Feature Association (F.A.) as:

> *One feature is associated with the object of interest if its local pattern is shared by the model image and its neighbours are also associated with the model.*

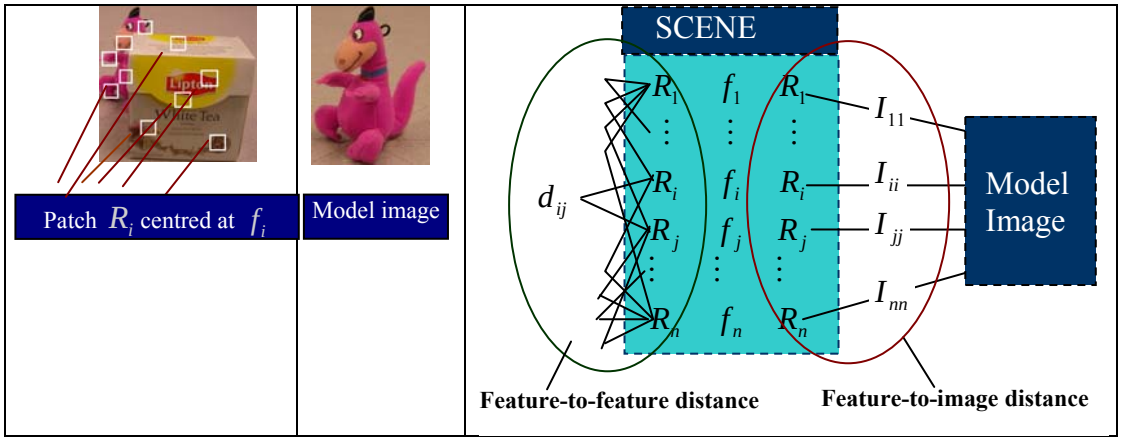


Figure 4.5 Feature-to-feature distance and feature-to-image distance

The association score for a feature depends not only on its affinity with model image, but also the proximity with neighboring features. Therefore, we have defined a feature-to-image distance (image patch to model image) and feature-to-feature distance (within scene image), shown in Figure 4.5. Note that, our definition of feature association score endows "feature importance" in [145] with physical meaning. It integrates top-down reasoning from the model image, while capturing internal geometric layouts of local features within scene image. Moreover, feature importance ranking under the setting of [145] is bi-directional and with no global view of local feature provided.

The association score for one feature in the scene can be expressed as:

$$\theta_i = \sum_{j=1}^{n} \theta_j \, \mathrm{d}(f_i, f_j)\, \mathrm{I}(R_i, \mathrm{Im}_{\mathrm{model}}) \tag{4.6}$$

Where n is the number of features in scene image; θ_i $(i = 1, \cdots, n)$ is the association score of the i^{th} feature f_i ; $\mathrm{d}(f_i, f_j)$ is the intra-proximity (Euclidean distance) between the i^{th} feature f_i and the j^{th} feature f_j, while $\mathrm{I}(R_i, \mathrm{Im}_{\mathrm{model}})$ is the inter-affinity between image patch R_i to the model image. Therefore, the association score of one feature is the weighted summation of that of the rest of the features.

Internal geometric layouts in the scene can be interpreted as the following matrix:

$$\mathbf{I}_{\mathrm{int}\,ra} = \begin{bmatrix} 1 & d_{12} & d_{13} & \cdots \\ d_{21} & 1 & d_{23} & \vdots \\ d_{31} & d_{32} & 1 & d_{ij} \\ \vdots & \cdots & d_{ji} & 1 \end{bmatrix}, \quad i, j = (1, \cdots n) \tag{4.7}$$

where, $d_{ij} = \mathrm{d}(f_i, f_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$

Intra- affinity between features and the model image takes the following form:

$$\mathbf{I}_{\mathrm{int}\,er} = \begin{bmatrix} I_{11} & 0 & \cdots & 0 \\ 0 & I_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & I_{ii} \end{bmatrix}, \quad i = (1, \cdots n) \tag{4.8}$$

where $I_{ii} = \mathrm{I}(R_i, \mathrm{Im}_{\mathrm{model}})$ and its calculation is specified in Eq.(4.9).

4.4.2 Implementation of Feature Association

To reduce the illumination effects, image is represented in a more robust way. Normalized *R* and *G* channels as well as gradients and Laplacians on these two channels are all employed. Each pixel can be defined by a 6-dimensional vector $[R_{norm}, G_{norm}, Grads\_R, Grads\_G, Lap\_R, Lap\_G]$ .Therefore, the image quantization proceeds in space with a dimension of six.

Affinity between an image patch and the model image is measured by the similarity of their color patterns, i.e. feature-to-image distance. Similar to the formulation of color relationships in Section 4.3.1, intersection of color co-occurrence matrices is used to measure this feature-to-image distance. Image patch and the model image are represented as co-occurrence matrices $\mathbf{CH}_{R_i}$ and $\mathbf{CH}_{\text{model}}$ [51], where each entry counts the number of occurrences of two quantization indexes $\widetilde{c}_p$ and $\widetilde{c}_q$ within the distance $\widetilde{d}_k$. The patch to image distance is defined by the intersections of their co-occurrence matrices:

$$\mathrm{I}(R_i, \mathrm{Im}_{\text{model}}) = \sum_{p=1}^{n_{\widetilde{c}}} \sum_{q=1}^{n_{\widetilde{c}}} \sum_{k=1}^{n_{\widetilde{d}}} \min[(\mathbf{CH}_{R_i}(\widetilde{c}_p, \widetilde{c}_q, \widetilde{d}_k), \mathbf{CH}_{\text{model}}(\widetilde{c}_p, \widetilde{c}_q, \widetilde{d}_k))], i \in [1, \cdots, n] \quad (4.9)$$

The meaning of the parameters is defined in section 4.3.1.

It is noticed that the colors in the scene image and the model image should not be quantized separately as even the same color will be assigned different indices in both the scene and model image, shown in Figure 4.6 (a) and (b).

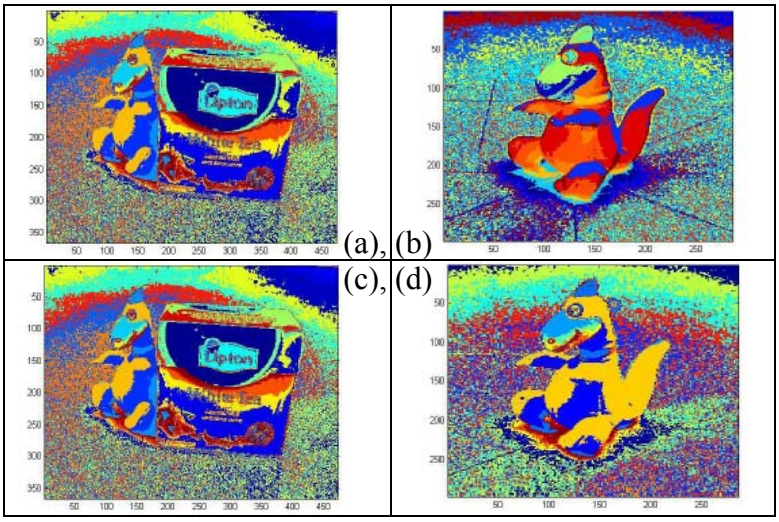


Figure 4.6 Color quantization by joint *k-means* clustering:
(a) and (b) are the results from separate *k-means*;
(c) and (d) are from joint *k-means*

In order to use color histogram to describe image patches, it is necessary that the same object is represented by similar colors in different images. A joint *k-means* clustering is proposed to quantize scene and model images. Usually, the scene image should be richer in colors and therefore the *k-means* algorithm is firstly applied to quantize the colors in the scene image. Secondly, the cluster centroids of the scene image are then used for clustering colors in the model image. The colors in the model image are assigned into clusters according to the minimum Euclidean distance criteria, based on the clustering scheme of the scene image, shown in Figure 4.6 (c) and (d). In other word, the model image is quantized towards the same cluster-centers obtained from the scene image.

In the meanwhile, we formulate the feature-to-feature distance $d_{ij} = \mathrm{d}(f_i, f_j)$ in the scene image by their geometric relationships, namely Euclidean distance. There are alternative ways to measure this proximity by the combination of color, texture and geometry information carried by image patches centered at feature points as in [38]. However, it is computationally intensive. The geometry alone is working well, as shown in Figure 4.7, where features associated with model image are marked in

blue or red.

The total association scores can be written as a vector: $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots \theta_n]^T$ and accordingly, Eq. (4.6) can be rewritten as

$$\boldsymbol{\lambda\theta} = \mathbf{I}_{\text{int}er}\mathbf{I}_{\text{int}ra}\boldsymbol{\theta} \tag{4.10}$$

where $\boldsymbol{\lambda}$ is a set of scalars that introduces the scale effect.

The association score vector $\boldsymbol{\theta}$ (with $\theta_i \geq 0$) has to satisfy Eq. (4.10). As matrix ($\mathbf{I}_{\text{int}er}\mathbf{I}_{\text{int}ra}$) is non-negative, the Perron-Frobenius theorem guarantees the non-negativity of the principle Eigen-vector. Therefore, $\boldsymbol{\theta}$ is obtained as the principle Eigen-vector of the ($\mathbf{I}_{\text{int}er}\mathbf{I}_{\text{int}ra}$) and then $\lambda_i \in \boldsymbol{\lambda}$ is the corresponding Eigen-value. The larger the elements of principle Eigen-vector is, the more associated the corresponding features are with the object of interest, carrying higher association scores.

The features are ranked according to their association scores. Those with higher scores are more likely to be lying on the object of interest. A number N_T is chosen to decide how many features with top association scores should be kept, while the remaining ones are truncated. Here the associated feature points with top association scores are shown in blue or red in Figure 4.7. The remaining features are trimmed off because of their low association scores.

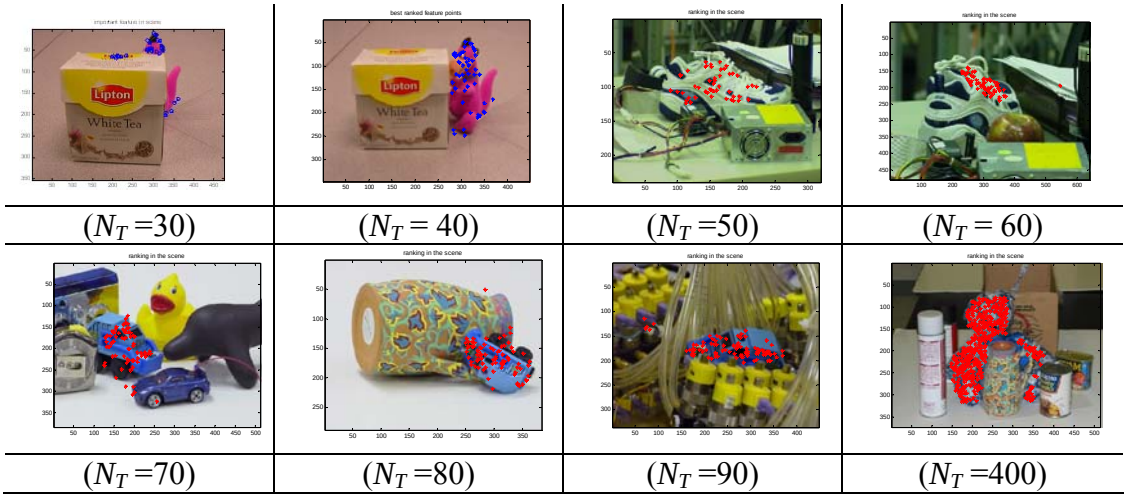


Figure 4.7 Features associated with objects of interest

In Figure 4.7, the remaining top feature points are all lying on the object of interest. In practice, it is difficult to decide on the value of N_T. Usually, the higher the value of N_T, the more likely that features from other objects are included. In our experiments, N_T. is 50% of the total number of features detected in the model. This problem-specific parameter can be obtained from ground truth correspondences.

4.5 Conclusions

Two approaches have been proposed in this chapter to reduce feature interactions which is addressed as a key issue in occlusion recognition problem. In our approaches, the appearance similarity could serve as priors on how likely two features are from the same object. In the meanwhile, features are associated with the object of interest based on the defined feature association scores. Effectiveness of Appearance Prior and Feature Association has been experimentally shown to reduce feature interactions. In the next chapter, how they could benefit the recognition of occluded objects is evaluated by working with spectral matching.

Chapter 5

Recognition of occluded objects in a scene

In our proposed solution to occlusion recognition problem, there are two important aspects, i.e., a local to global decision and a reduction of feature interactions. In Chapter 3, the spectral matching algorithm is chosen to arrive at a global recognition decision based on local evidence. Furthermore, appearance information has been employed to reduce feature interactions in Chapter 4. In this chapter, our algorithms to recognize occluded objects in a scene are proposed. They are designed by combining strategies for the reduction of feature interactions with the spectral matching algorithm. Thus, pairwise feature appearance and geometric relationships are encoded in our algorithms.

Our work is related to the algorithms presented in [220] [221], which also solve the correspondence problem by taking into consideration the pairwise feature appearance and geometric relationships. Their works report that the outliers /occlusions are handed successfully. Torresani et al. in [220] defined a complicated objective function with an occlusion cost term and then outliers are handled as part of the optimization process using dual decomposition. This method either has complicated objective function or complicated constraints, that it results in high computational cost and limits its applications on sparse feature sets. Instead, in our work, we handle occlusion by removing low-confident correspondences, which enables our algorithms to work for real world image matching problem. Our spectral matching method looks for a group of correspondences which maximizes the matching score when matching a

scene image to the model image. Therefore, a global maximum is guaranteed by the spectral matching method. On the other hand, Liu and Yan in [221] optimize the same quadratic function for correspondences when matching one image with multiple images. Their work is to find local maxima which indicate multiple groups of correspondences. Thus, their work to detect multiple visual patterns has different objective from our aim in occluded object recognition.

In this chapter, two algorithms which work independently are proposed for the recognition of occluded objects in a scene. Based on the work presented in Chapter 3 and 4, Algorithm 1 integrates Appearance Prior (A.P.) into the spectral matching algorithm; and Algorithm 2 combines Feature Association (F.A.) with the spectral matching algorithm. In the following sections, pairwise geometry is first formulated and local geometric consistency is then enforced by spectral matching algorithm. Performances of our proposed occlusion recognition algorithms are then evaluated and compared with the state of the art recognition algorithms.

5.1 Proximity matrix by pairwise geometric agreement

In real world applications, features that constitute certain kind of structures are often extracted together and thus they are often matched set by set [19]. The matching assumes that spatial consistency should be preserved by corrects correspondences in the two images. Based on graph theory, local geometric relationships between pairs of correspondences are then encoded into the proximity matrix, which includes both correct correspondences and incorrect ones. Provided that the feature interactions are eliminated, the matching problem would reduce to the clustering of correct correspondences in proximity matrix. In fact, appearance similarities may not be sufficient to separate different objects, which have been observed in Section 4.3.2,

thus leaving ambiguous connections in the proximity matrix. Therefore, we propose to combine appearance similarity with spectral matching algorithm, where the spatial consistency between correspondences is encoded by the proximity matrix.

5.1.1 Proximity matrix for spatial consistency

The proximity matrix is built by using relationships both between individual features and feature pairs as introduced in [13]. The two sets of features are denoted as $\mathbf{F}^1_{n_1\times 1} = [f^1_1, f^1_2, \cdots, f^1_{n_1}]^T$ and $\mathbf{F}^2_{n_2\times 1} = [f^2_1, f^2_2, \cdots, f^2_{n_2}]^T$. Based on mapping constraints (one-to-many), preliminary matching can be carried out between these two feature sets and the result can be written as:

$$\mathbf{C}_{n_1\times n_2} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n_2} \\ c_{21} & c_{22} & \cdots & c_{2n_2} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n_1 1} & c_{n_1 2} & \cdots & c_{n_1 n_2} \end{pmatrix} \tag{5.1}$$

where, $c_{rs} = \Theta_{rs}(f^1_r, f^2_s), r \in [1, \cdots, n_1]; s \in [1, \cdots n_2]$ and Θ_{rs} is defined as:

$$\Theta_{rs}(f^1_r, f^2_s) = \begin{cases} 1 & \textit{If matching constraints are satisfied} \\ 0 & \textit{Otherwise} \end{cases}$$

The total number of non-zero elements in **C** is n' ($n' \le n_1 \times n_2$), and a new matrix **L** of size ($n' \times 2$) can be formed by elements with non-zero value in **C**. Matrix **L** gives the list of candidate correspondences, where each row consists of paired features $(f^1_i, f^2_{i'})$, as candidate correspondences.

$$\mathbf{L}_{n'\times 2} = \begin{pmatrix} \vdots & \vdots \\ (f_i^1 & f_{i'}^2) \\ (f_j^1 & f_{j'}^2) \\ \vdots & \vdots \end{pmatrix}$$

With the list of candidate correspondences, proximity matrix **M** is established based on pairwise spatial consistency between correspondences, as follows.

$$\mathbf{M}_{n'\times n'} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1n'} \\ m_{21} & m_{22} & \cdots & m_{2n'} \\ \cdots & \cdots & \cdots & \cdots \\ m_{n'1} & m_{n'2} & \cdots & m_{n'n'} \end{pmatrix} \tag{5.2}$$

where element $m_{ab}, a, b \in [1, n']$ is defined as:

$$m_{ab} = \begin{cases} \Gamma(f_i^1, f_{i'}^2), a = b \\ \widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2)), a \neq b \end{cases} \quad a, b \in [1, \cdots, n']; i, j \in [1, \cdots, n_1]; i', j' \in [1, \cdots, n_2] \tag{5.3}$$

where $\Gamma(f_i^1, f_{i'}^2)$ are the diagonal elements in **M**, each of which measures the similarity between matched features in each row of matrix **L**; and $\Gamma(f_i^1, f_{i'}^2) = \left\| des_i - des_{i'} \right\|$, is the Euclidean distance between feature descriptors;

$\widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2))$ are the off-diagonal elements in **M**, each of which is the spatial consistency between two feature pairs, involving two rows in **L**. These elements measure how well the pairwise geometric relationship between two model features f_i^1 and f_j^1 is preserved by their potential correspondences $f_{i'}^2$ and $f_{j'}^2$ in scene image;

The preservation of pairwise geometric relationship is further defined in the next subsection.

5.1.2 Pairwise geometry preservation

For general features, such as corners, pairwise geometric relationship is defined by the segment which links two features in a defined neighbourhood as shown in Figure 5.1:

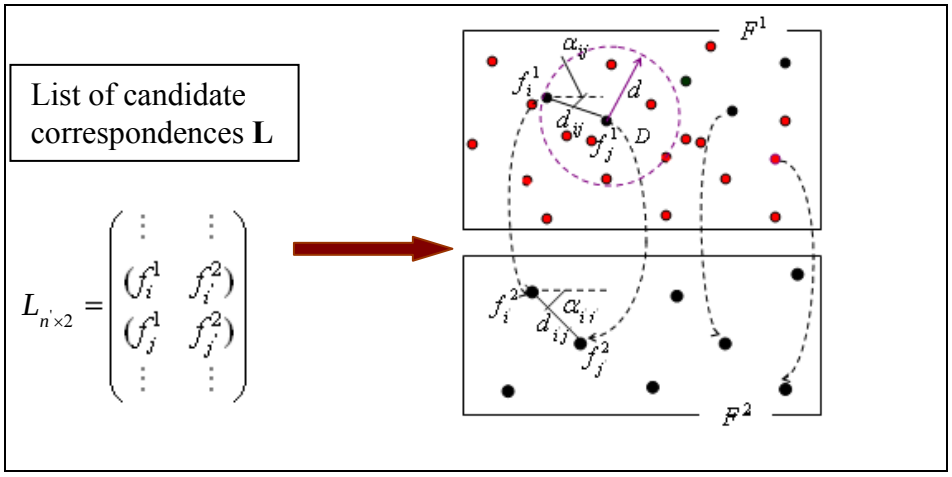


Figure 5.1 Pairwise geometric relationship

The geometric relationship between f_i^1 and f_j^1 is defined by the segment which links these two points with its length d_{ij} and the angle made with x-axis α_{ij}. This geometry is then preserved by the segment linking their correspondences $f_{i'}^2$ and $f_{j'}^2$, using its length $d_{i'j'}$ and the angle made with x-axis $\alpha_{i'j'}$. Therefore, this preservation is defined by $\tilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2))$, using angle and distance differences.

$$\tilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2)) = \begin{cases} \exp-\left(\dfrac{(d_{ij}-d_{i'j'})^2}{\sigma_3^2} + \dfrac{(\alpha_{ij}-\alpha_{i'j'})^2}{\sigma_4^2}\right) & d_{ij} \le d \\ 0 & otherwise \end{cases} \tag{5.4}$$

Where σ_3 and σ_4 weigh the change in relative length and in orientation, respectively; note that d is the radius of a local support region D.

$\widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2))$ are considered as the local geometric consistency ($d_{ij} \le d$). Geometric relationships between two close features tend to be more correctly persevered than those between distant features. When $d_{ij} > d$, the geometric relationships between two faraway features are set to zero. Note that d is determined during the experiments which will be presented later.

Since a small diagonal element $\Gamma(f_i^1, f_{i'}^2)$ means a higher confidence in the correspondence between the features, other correspondences in agreement with them should also have a higher confident vote. Therefore, additional weights are given to off-diagonal elements in **M** which involve small diagonal elements. In particular, the off-diagonal elements are emphasized by the reciprocal of the square root of the product of two diagonal elements involved. Our proximity matrix will then take the following form:

$$\mathbf{M}_{n' \times n'} = \begin{pmatrix} m_{11} & \widetilde{w}_{12} m_{12} & \cdots & \widetilde{w}_{1n'} m_{1n'} \\ \widetilde{w}_{12} m_{21} & m_{22} & \cdots & \widetilde{w}_{2n'} m_{2n'} \\ \cdots & \cdots & \cdots & \cdots \\ \widetilde{w}_{1n'} m_{n'1} & \widetilde{w}_{2n'} m_{n'2} & \cdots & m_{n'n'} \end{pmatrix} \tag{5.5}$$

where, $\widetilde{w}_{ij} = 1/\sqrt{m_{ii} \cdot m_{jj}}$ and **M** is still symmetric.

Our algorithms to recognize occluded objects are proposed by combining the spectral matching algorithm with some strategies to reduce feature interactions. The two proposed algorithms are introduced in the next sections.

5.2 Algorithm 1: Combining geometry with Appearance Prior

Our Algorithm 1[125] to recognize occluded object in a scene is to combine feature interactions reduction based on Appearance Prior (A.P.), proposed in Section 4.3 with

the spectral matching algorithm. This spectral algorithm alone is not suitable for occlusion recognition; however, pairwise appearance similarity could serve as its priors about how likely a pair of features are from the same object (probably the object of interest). The effectiveness of A.P. to reduce feature interactions has been shown in Section 4.3.2 and its short-fall to guide matching process is also analyzed.

In this algorithm, the color and texture similarities between features serve as priors to qualify the possibility that the features are from the same object. A.P. is introduced into the spectral matching algorithm by multiplying the elements in the proximity matrix $\widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2))$, defined in Eq. (5.4), with the appearance similarity $\widetilde{p}(f_i^1, f_j^1)$, defined in Eq. (4.5):

$$\widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2)) = \widetilde{\Gamma}((f_i^1, f_{i'}^2),(f_j^1, f_{j'}^2)) * \widetilde{p}(f_i^1, f_j^1) \qquad (5.6)$$

The physical meaning behind Eq. (5.6) is that the local geometry between a pair of features is preserved by the elements in the proximity matrix, which is explained in section 5.1.2, and this preservation will be given larger weights if these two features are from the same objects. The preservation of local geometry is further illustrated in Figure 5.2, where the local geometric relationship between a pair of feature points in the left car is preserved by their correspondences in the red truck. If the two features in the left image carry similar appearance information, this geometric preservation will be given a larger weight by Eq. (5.6).

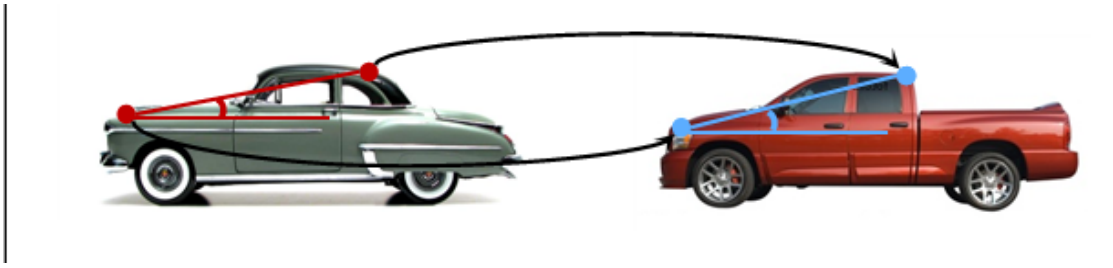


Figure 5.2 Pairwise geometry preservation

5.3 Algorithm 2: Combining geometry with Feature Association

In our Algorithm 2 [93], the spectral matching algorithm is combined with feature interactions reduction based on Feature Association, proposed in Section 4.3. F.A. has shown its ability to locate features from object of interest. However, its performance drops in situations where similar local information is shared by the background and target object. Therefore, pairwise geometric consistency is combined with Feature Association to further constrain the matching process. This algorithm to recognize occluded object is formulated as a two-stage strategy.

Given $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots \theta_{n_1}]^T$ obtained from Eq. (4.5), features with lower association score should be trimmed off because their chances to belong to the object of interest are low, indicating a lower chance for them to form correct correspondences. The original feature set is trimmed as $\mathbf{F}^1 = [f_1^1, f_2^1, \cdots, f_l^1]^T$ with association scores $\boldsymbol{\theta}' = [\theta_1, \theta_2, \cdots \theta_l]^T, l < n_1$. The new feature set $\mathbf{F}^1 = [f_1^1, f_2^1, \cdots, f_l^1]^T$ and feature set of model image $\mathbf{F}^2 = [f_1^2, f_2^2, \cdots, f_{n_2}^2]^T$ are paired up, resulting in $z \le l \cdot n_2$ pairs of candidate correspondences. Subsequently, the spatial consistency matrix $\mathbf{M}$ can be formulated as follows:

$$\mathbf{M}_{z\times z} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1z} \\ m_{21} & m_{22} & \cdots & m_{2z} \\ \cdots & \cdots & \cdots & \cdots \\ m_{z1} & m_{z2} & \cdots & m_{zz} \end{pmatrix} \tag{5.7}$$

where components in **M** take the same definition as in Eq. (5.3).

The performances of the two proposed algorithms are shown in the next section and comparisons will be made with the state of the art recognition algorithms.

5.4 Experiments

We test our algorithms in the recognition of occluded objects under different occlusion rates. In our database, there are four target objects (one toy dinosaur and three toy cars). Training images of an object of interest are taken at a fixed interval of 15° for a total of 24 views, with a uniform background. Therefore the size of our database is 4 (objects) *24(views) =96 images. Taking toy dinosaur as an example, Figure 5.3 shows that the dinosaur is under different occlusion rates in the scene images. It is shown that the dinosaur is occluded by a box by minor (20%), moderate (40% and 50%) and severe occlusion rates (70%). In our experiments, scene images contain one or more target objects (out of 4) with various occlusion rates (20% to 70%).

(a) 20%	(b) 40%	(c)50%	(d)70%

Figure 5.3 Measurements of different occlusion rates

5.4.1 Parameters setting

In our experimental settings, the number of clusters for image quantization and the size of image patch are parameters to be determined. To describe local pattern, the number of clusters (i.e., colors) and the patch size interact with each other. Fewer colors or smaller patches result in the loss of the local information; while too many colors or bigger patch sizes introduce noise. Figure 5.4 shows the matching performance (in terms of matching rate) vs. the number of clusters and the patch size (in number of pixels for the length). The experiments carried out to arrive at the results are conducted in the following steps:

(i) Keep one of the parameter fixed (number of clusters, say).

(ii) Perform the experiments by varying the values of the other parameter (image patch size in this case)

(iii) Determine the matching rate by taking the average of results obtained in (ii).

(iv) Use another value for the number of clusters and repeat steps (i) to (iii).

The algorithms perform well for patch sizes of 40 to 60 pixels and 30 to 70 clusters, as shown in Figure 5.4. In our subsequent experiments, we used the patch size of 40 pixels and 50 clusters for image quantization, due to the tradeoff between speed and accuracy. Better results may be obtained by optimizing these parameters.

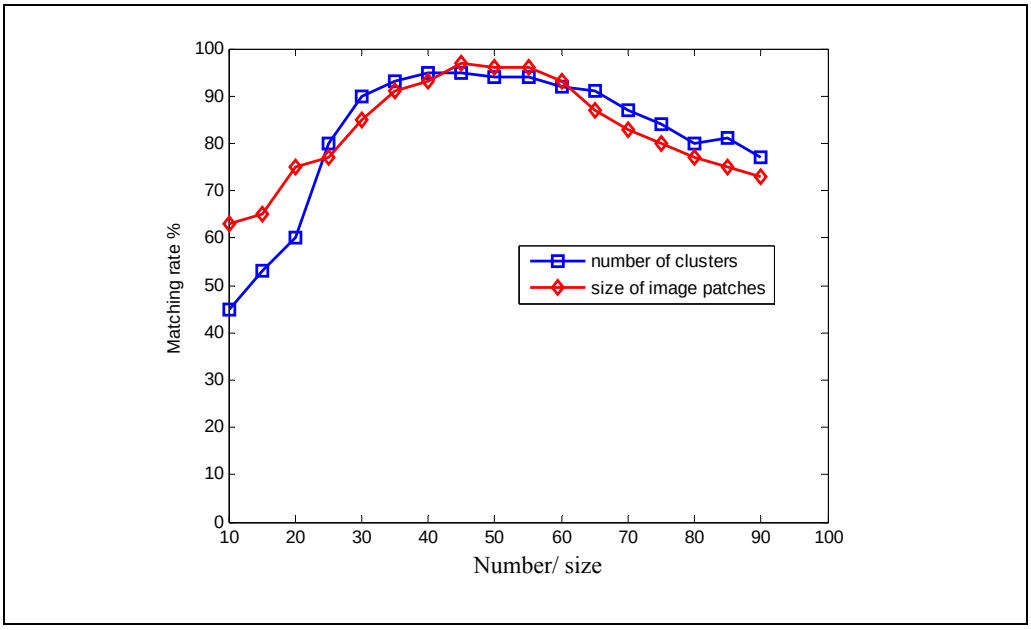


Figure 5.4 Matching rate vs. patch size and number of clusters

The size of local support region for each feature point (d in Eq. (5.4)) is set to cover its 10 nearest neighbors. The ranges of weights for length and angle difference are narrowed by automatic scale selection [22]. In our experiments, neither the length nor the angle difference is favored and they are equally set as $\sigma_3 = \sigma_4 = 5$, which are determined based on ground truth correspondences. Particular tasks or problem domains may benefit from individually weighting. To evaluate the performance of the similarity measure, we follow the work reported in [48]. We set $\sigma_1 = \sigma_2 = 0.3$ as the values that would maximize the *F*-measure over a set of 50 randomly selected training images from the Berkeley Segmentation Database (BSD) [57]. The *F*-measure is given by $F(p,r) = pr/(0.5*p + 0.5*r)$, where p is the precision defined as the probability that two features declared to be in the same group are indeed in the same group, and r is the recall defined as the probability that a same group pair is detected [38].

5.4.2 Recognition performances of Algorithm 1

Performances of Algorithm 1 are evaluated under different occlusion rates. For comparison, performances of four other matching algorithms with and without Appearance Prior (A.P.) under different occlusion rates are shown in Table 5.1. The four chosen graph matching algorithms are Spectral Matching with Affine Constraints (SMAC), Spectral Matching (SM), Graduated Assignment (GA) and Semi-definite Programming (SDP). Since they are all maximizing the same quadratic function, they are working with Integer Projected Fixed Point algorithm (IPFP) as a post-processing step for comparison. SIFT features are extracted and matched. Ground truth is manually provided. Procedures to calculate the matching rate are given in Table 3.2.

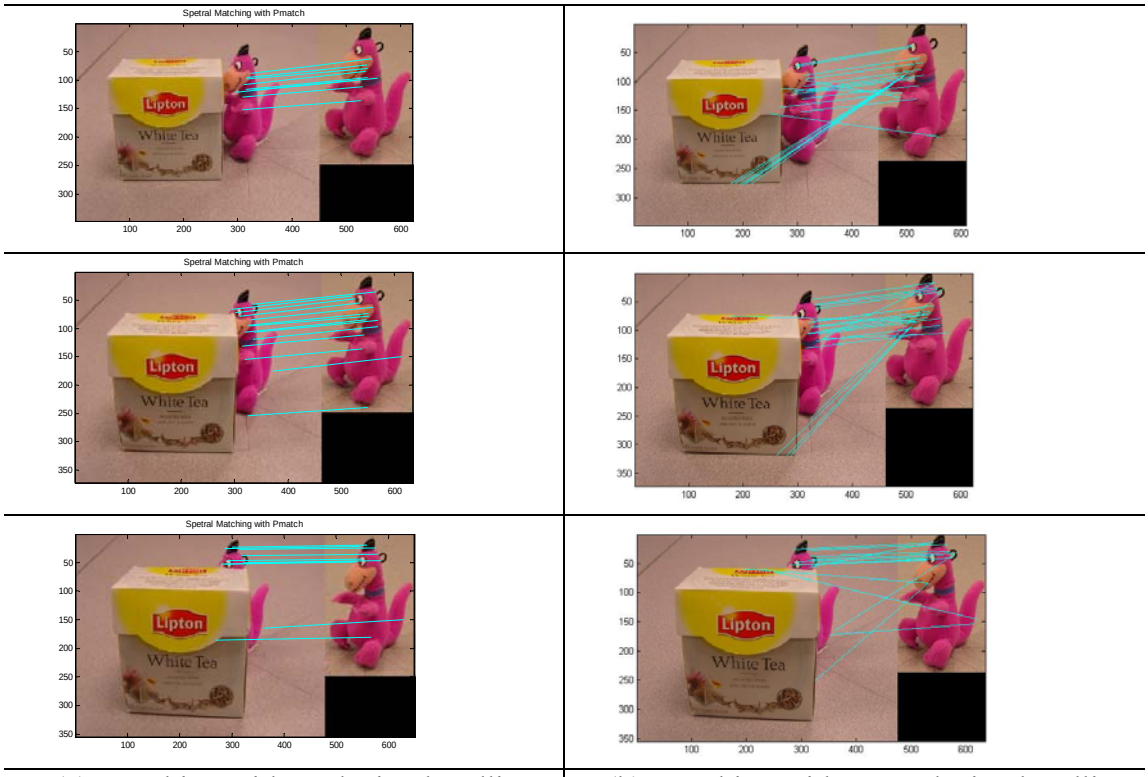
Table 5.1 Recognition rates (%) comparison w/o A.P.

Algorithms / Occlusion rate(%)	GA+IPFP		SDP+IPFP		SM+IPFP		SMAC +IPFP	
	Without A.P.	With A.P.	Without A.P.	With A.P.	Without A.P.	With A.P.	Without A.P.	With A.P.
30	*70.4*	*80.3*	*70.3*	*82.3*	*73.7*	*84.5*	*85.2*	***97.1***
40	*60.3*	*64.2*	*66.2*	*76.7*	*70*	*75.2*	*73.4*	***92.7***
50	*38.5*	*43.1*	*44.1*	*54.8*	*54.3*	*58.1*	*63.1*	***90***
70	*26.1*	*30.2*	*36.8*	*40.3*	*43.6*	*47*	*55.6*	***84.7***

It is observed in Table 5.1 that graph matching algorithms alone are sensitive to occlusion which causes ambiguous connections in the proximity matrix. Therefore, their performances decline with the increase in occlusion rates. With feature interaction reduction by Appearance Prior, their performances have been improved. Our proposed Algorithm 1 has a clearly better tolerance of occlusion, as indicated in the rightmost column in Table 5.1.

Figure 5.5 gives a closer look on the difference that can be made by occlusion

handling. Correspondences between images are linked by green lines. In the occlusion scenario, features tend to be incorrectly matched (by SIFT matching algorithm) because of the similar local appearance in Figure 5.5 (b). However, with our algorithms, a global reasoning based on local information is introduced to enforce correct correspondences as shown in Figure 5.5 (a).



(a) Matching with occlusion handling (b) Matching without occlusion handling

Figure 5.5 Matching w/o occlusion handling

5.4.3 Recognition performances of Algorithm 2

Algorithm 2 has been formulated as a two-stage strategy in Section 5.3. In order to evaluate its performances on our database, the confidence of recognition needs to be first redefined by taking into consideration the occlusion rate.

Existing algorithms perform matching discriminatively and ambiguous matches are often discarded because they are assumed to arise from background clutter. This is

exemplified by the Euclidean distance based nearest neighbor approach in [42]. In this context, object is recognized if the rate of correctly matched features is above a predefined threshold, i.e., confidence of recognition. Unfortunately, when the object of interest is under severe occlusion (rate of occlusion is over 50%), there is only a very small portion of features detected from the object of interest. Given that all these features are correctly matched, the rate of matched features for the object of interest is too low for it to be considered recognized, leading to a low confidence of recognition. Therefore, whether the object of interest is recognized is related to its occlusion rate.

To recognize the occluded object in a scene, the definition of confidence of recognition should be redefined according to our two-stage strategy. In our proposed algorithm, only features that are highly associated with the object of interest are considered for matching, which can be interpreted as taking the occlusion rate into consideration.

$$Refined\ recognition\ confidence = \frac{Number\ of\ correct\ matches}{Number\ of\ associated\ features} \tag{5.7}$$

Table 5.2 shows the performances of matching algorithms with or without Feature Association (F.A.) under various occlusion rates.

Table 5.2 Recognition rates (%) comparison w/o F.A.

Algorithms / Occluded Area (%)	GA+IPFP		SDP+IPFP		SM+IPFP		SMAC+IPFP	
	Without F.A.	With F.A.	Without F.A.	With F.A.	Without F.A.	With F.A.	Without F.A.	With F.A
30	*70.4*	*82.2*	*70.3*	*83.4*	*73.7*	*85.5*	*85.2*	***98.3***
40	*60.3*	*67.8*	*66.2*	*77*	*70*	*76.3*	*73.4*	***94.5***
50	*38.5*	*45.1*	*44.1*	*55.1*	*54.3*	*58.4*	*63.1*	***92***
70	*26.1*	*30*	*36.8*	*37.6*	*43.6*	*45.2*	*54*	***86.9***

The same four graph matching algorithms, as introduced in Section 5.4.2, are chosen for comparison. In our experiments, the object of interest is recognized if the redefined recognition confidence (Eq. (5.7)) is greater than 0.7. In Table 5.2, it is shown that performances of matching algorithms are improved by Feature Association, where occlusion rate is taken into consideration in the recognition confidence. Our Algorithm 2 has the best performance under occlusion, as shown in the rightmost column in Table 5.2.

It is not practical to predict the number of features to be trimmed off, i.e., to determine the rate of occlusion before recognition. In our experiment, 50% of the number of features extracted from the model image is considered as our reference, i.e., the value of N_T = 50% defined in Section 4.4.2. There could be an optimal choice for this parameter either through machine learning or optimizing an object function to recognize a specific object class.

We have listed the performances of our two proposed algorithms (Algorithms 1 and 2) in Table 5.3. It is shown that both A.P. and F.A. are able to improve the performances of the spectral graph matching algorithms. However, it is not fair to directly compare the performances of Algorithms 1 and 2, since they are working under different principles, i.e., integrating A.P. or F.A, respectively. Algorithm 2 takes advantage of supervised recognition, introducing top-down reasoning when model views are provided. It redefines the recognition confidence with the consideration of the occlusion rate, which has received little attentions in the field of occlusion recognition research. On the other hand, Algorithm 1 could work in unsupervised fashion where no model views are available.

Table 5.3 Recognition by our two proposed algorithms

Algorithms / Occluded area (%)	GA+IPFP		SDP+IPFP		SM+IPFP		Proposed Algorithm 1	Proposed Algorithm 2
	With A. P.	With F.A.	With A. P.	With F.A.	With A.P.	With F.A.		
30	*80.3*	*82.2*	*82.3*	*83.4*	*84.5*	*85.5*	***97.1***	***98.3***
40	*64.2*	*67.8*	*76.7*	*77*	*75.2*	*76.3*	***92.7***	***94.5***
50	*43.1*	*45.1*	*54.8*	*55.1*	*58.1*	*58.4*	***90***	***92***
70	*30.2*	*30*	*40.3*	*37.6*	*47*	*45.2*	***84.7***	***86.9***

5.4.4 Recognition performance comparison

The performances of our algorithms are further compared with the state of the art recognition algorithms. We adapt our matching algorithms to the 3-D object recognition configuration introduced by Lowe [74], in order to recognize the 3-D pose of the target, which is briefly reviewed as follows.

The 3-D presentation of an object is given by a set of model views which describe the appearance of the object from a range of significantly different locations around the view sphere. Individual model view is given by clustering training images that fit the same similarity transformation built from their matched features. Therefore, when matching a scene image to objects models, this matching is propagated through the model views to perform object identification and 3-D pose recognition. The same matching method is used for matching training images for recognition, i.e., matching scene image with training images.

In the study of the 3-D object recognition using our algorithms, we use images from the Ponce object recognition database [64]. We find it suitable for our study of occlusion recognition, because it contains cluttered test shots with multiple objects as shown in Figure 5.6.

Figure 5.6 Sample images from Ponce object recognition database

We make no assumptions on where and what are the occluded objects to be recognized in a scene images. In fact, we attempt to recognize multiple objects in scenes, where objects are overlapping each other, leaving us both occluded and occluding objects. The recognition is completed when all objects in a scene are recognized. The performances of the five state of the art recognition algorithms have been reported on Ponce object recognition database (Ferrari et al. [204], Lowe [42], Mahamud & Hebert [203], Moreels et al. [205], and Rothganger et al. [64]). All the algorithms achieved recognition rates of 90% and above with the false detection rates below 10%. Compared with these algorithms, our algorithms together with Lowe's and Fred's algorithm have the configuration of combining the information associated with multiple views in the recognition process, while the rest simply use multiple views to build object models. Mahamud & Hebert [203], Moreels et al. [205] and Rothganger et al. [64] considered all training pictures independently, which essentially reduces object recognition to image matching.

In our proposed algorithms, both appearance and geometric information are taken into consideration. The Appearance Prior is an important cue for the association of features according to their appearance similarity, indicating possible object regions. On the other hand, Feature Association efficiently provides an appearance-based filter for possible model searching and the geometric verification reasons for the most positive view. Our algorithms obtains about 95% recognition rate when applied to scenes with occluded 3-D objects, compared to result reported by the state of the art recognition techniques [64].

Some of the matching results are shown in Figure 5.7, where occluded objects are correctly matched to model views. Existing matching algorithms tend to fail in challenging cases such as Figure 5.7(b), (c), (d), (e) and (f), where the background has structured patterns. For instance, the colorful vase near the bear in (d), introduces ambiguous matches.

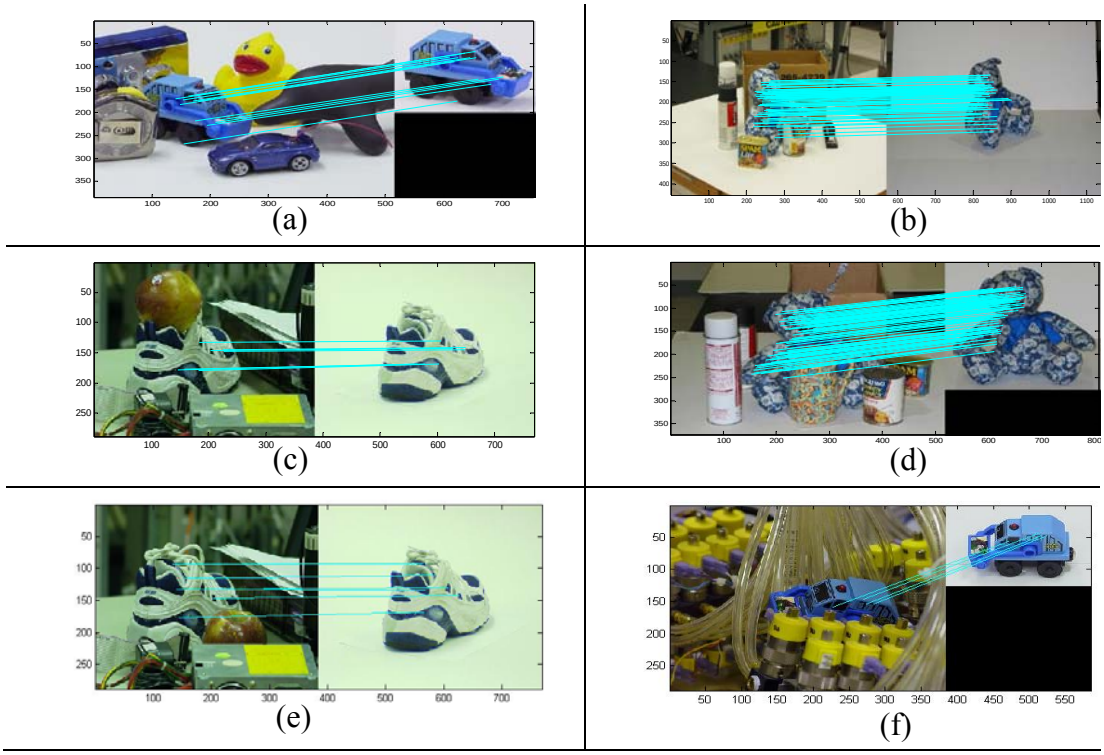


Figure 5.7 Matching of occluded objects

5.4.5 Effect of Feature Association in occlusion recognition

Since Algorithm 2 is a two-stage strategy, its performance is contributed by the two stages together. Therefore, further investigations on whether F.A. alone boosts the occlusion recognition need to be conducted.

It is noted that our F.A. does not set requirements on types of features. In this section, effect of F.A. in recognition is further evaluated, by comparing with sophisticated feature descriptors. Since the "greedy" version of RANSAC (with N = 20) gives the best performance of Local Affine Invariance (L.A.I) [64], we also use this RANSAC as the baseline of comparison to demonstrate the effectiveness of Feature Association (F.A.). The comparison is shown in Table 5.4. In our recognition experiments, eight object models were matched against a set of 51 scene images. Each scene image contains instances of up to five objects of interest, even though most of them only contain one or two. The recognition rate for each object of interest is defined as the number of correct recognitions out of the total number of recognitions processed.

Table 5.4 Comparison of recognition rates using the greedy RANSAC

Methods	Apple	Bear	Rubble	Salt	Shoe	Spidey	Truck	Vase	Mean	Time
RANSAC (SIFT)	8/11	11/11	9/9	10/10	7/9	4/4	12/12	12/12	94%	3.7s
RANSAC (L.A.I.) 8×8 resolution	8/11	11/11	9/9	10/10	7/9	4/4	12/12	12/12	94%	1.9s
RANSAC (L.A.I.) 16×16 resolution	9/11	11/11	9/9	10/10	5/9	4/4	11/12	12/12	91%	1.6s
RASANC (F.A.)	8/11	11/11	9/9	10/10	7/9	4/4	12/12	12/12	94%	**1.5s**

As shown in Table 5.4, the recognition rates are compared between RANSAC (SIFT), RANSAC (L.A.I) and RANSAC (F.A.), i.e., RANSAC combined with color

SIFT descriptor, Local Affine Invariant [64], and Feature Association respectively. It is shown that RANSAC (F.A.) achieves comparable performances with RANSAC (L.A.I) and RANSAC (SIFT). However, its computation efficiency out performs the other algorithms by 0.4s to 2.2s.

For an 512×512 image with moderate occlusion, the computational cost for recognition of all objects (two or three) in each image is about 1.4 seconds on 2.4 GHz Pentium computer, with about 0.8 second required for feature association, and about 0.6 second to perform indexing and least-squares verification of its 3-D pose. This computational cost is slightly lower than Lowe's 1.5 seconds, which is one of the most efficient 3-D object recognition algorithms. Moreover, our method is more stable when occlusion rate is higher, where SIFT feature matching is less reliable.

5.5 Conclusions

In this chapter, two novel algorithms are proposed to solve occlusion recognition problem, where strategies to reduce feature interactions are combined with the spectral matching algorithms which enforce geometric consistency.

In Algorithm 1, Appearance Prior has been integrated into the spectral matching algorithm. This could suggest a possible way for facilitating unsupervised object recognition, where no labeled training images are provided. This method can be further introduced to recognize a specific object class. In Algorithm 2, Features Association can be interpreted as a top-down procedure, associating features with the object of interest which is partially visible in the scene. Its facilitation for occlusion recognition has been manifested by baseline comparison. This association stage can also be viewed as one step further than *k*-nearest neighbors to form the initial graph

by common graph matching algorithms [13] [19] [72]. Our strategy redefines the recognition confidence with the consideration of occlusion rate, which has received little attentions in the field of occlusion recognition.

In our experiments, we have shown that graph matching algorithms themselves are not desirable for occluded object recognition. However, their performances are dramatically improved by introducing A.P. or F.A. into these graph matching algorithms. The experiments carried out show that our proposed algorithms can achieve performances comparable to the state of the art recognition algorithms.

Chapter 6

Local saliency to foreground object regions

In Chapter 5, two algorithms based on the principle of reducing feature interactions from different objects to recognize occluded objects in a scene have been presented,. In this and the next chapters, we would extend the applicability of our algorithms in dynamic scenes by introducing region information into consideration.

Our proposed methods are valid when occluded objects are still recognizable under occlusion rates of 70% or more. However, when occlusion situations arise in a dynamic way, the object of interest could appear in the scenes with varying occlusion rates ranging from 0 to 100% in a random manner. An example is shown in Figure 6.1, depicting a man walking behind a tree. It is noted that the occlusion rate is too high for our algorithms to process in some frames of the image sequence. In addition, in a dynamic scene, an object moving behind several objects would be occluded from time to time with varying occlusion rate.

Figure 6.1 Frames with varying occlusion rates (man walking behind a tree)

In order to widen the applications of our proposed algorithms, we would like to extend them to recognize occluded objects in dynamic scenes, where the observations of the changing scenes are made possible by using a surveillance camera or movable vision platform to capture the image sequence. From the captured sequence, it is observed that the object of interest is unrecognizable in some frames, but recognizable in others. This observation has motivated us. If the information of a target object is propagated throughout the image sequence, information needed in one frame for recognition could be collected from other frames which have lower occlusion rates. In order to implement this idea, possible object regions should be first located in individual frames, and this region information must be propagated. In this chapter, the algorithm to extract possible object regions is proposed and the said information propagation will be addressed in Chapter 7.

Region information has been known to improve recognition performance, as reported in [102] [116] [117] [118]. Region information is important because it is independent of object-category and it can provide powerful priori knowledge to direct the higher level recognition to be on the right path. Region information can also narrow the search space for matching.

In this chapter, a foreground subtraction algorithm is modified to combine with visual attention based saliency, in order to extract possible object regions. The object information is then propagated by two proposed approaches to realize our purpose of extending our occlusion recognition algorithms in dynamic scenes, which is elaborated Chapter 7.

6.1 Visual attention based saliency

If no assumptions are made on object appearance, visual attention would be a useful cue to locate possible object regions which are likely to attract human attentions. Visual attention is to identify fixation points that a human viewer would focus on at the first glance. The extensive psychophysical literature on visual search has demonstrated that basic visual features can capture and guide attention [141]. When we look at a scene without any *à priori* knowledge or intended task, our attention will be attracted to some locations mostly because of their saliency, defined by contrasts in color, intensity, orientation etc. [89][90]. In the early stage of visual processing, these low-level features can "pop-out" automatically. This "pop-out" (pre-attentive capture) suggests that features can drive the allocation of attention. Models such as Treisman's feature integration theory [133] or Wolfe's guided search model [139] produce human-like search behavior using only low-level feature information. Computational implementation of this class of model is known as "saliency map," a spatial ranking of conspicuous visual features that could be candidates for covert or overt attention (Itti & Koch [136]; Itti, Koch, & Niebur [100]; Koch & Ullman [90]). In fact, fixated locations by human can sometimes be better distinguished by object-level information (regions) then by image saliency (pixels). Algorithms have been developed to find salient regions based on saliency map. For instance, the Itti's saliency map has been extensively explored by some algorithms to extract salient regions. Han et al. [94] use Markov Random Field to integrate the seed values from Itti's saliency map along with low-level features of color, texture, and edges to grow the salient object regions. Ko and Nam [95] utilize a Support Vector Machine trained on image segment features to select the salient regions of interest using Itti's map, which are then clustered to extract the salient objects. Ma and Zhang [217] use fuzzy growing on their saliency

maps to confine salient regions within a rectangular region. Rathu et al. in [218] reformulate the saliency map as an energy function based on conditional random field model and the global minimum can be computed via graph cuts [219]. While these algorithms are struggling to find well-defined region boundary, the link between salient regions to possible object regions are not clear. In this thesis, the extraction of regions is based on local saliency and therefore is attention guided.

6.2 Foreground subtraction based on color histogram

M. Leordeanu in [147] has proposed to find the foreground by comparing color histograms from inside and outside a bounding box which is centered on the foreground. If the predefined bounding box covers an object well, window based approaches could distinguish the foreground from the background, which is the similar idea used in [27]. In a more general sense, if a correct bounding box is available (covering object well), color likelihood is a possible tool to separate the foreground and the background by comparing the histogram computed for the interior of the region of interest (the given window area), and that computed from the remaining region of the image. Algorithms, such as GrabCut [142] and Lazy Snapping [143], use color histograms to separate the foreground from the background given minimal user input. In [147], an algorithm has been proposed to subtract the foreground by a defined bounding box without clues about the foreground size and the approximate shape under three assumptions.

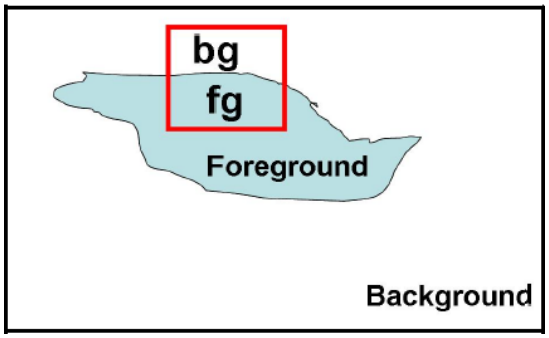


Figure 6.2 Bounding box (red), centered on the foreground [147]

As shown in figure 6.2, with a wrong bounding box in red, the foreground can be found if the following assumptions hold:

1. *The area of the foreground is smaller than that of the background.*
2. *The majority of pixels inside the bounding box belong to the foreground.*
3. *The color distributions of the background and foreground are independent of position.*

The first assumption is true for most objects in images. The second assumption could be met in practice if the center of the bounding box is considered to be on the foreground. The third assumption implies a uniform distribution of color in the foreground and the background, which is not true for general cases. At a given pixel, if a bounding box is chosen that meets all the above assumptions, it is possible to find which other pixels in the image are likely to be on the same object. This method gives global results, as it is able to group pixels that are far away in the image.

This is an ambitious yet hard to implement algorithm. The subtracted foreground corresponds to a specified bounding box. This subtraction imposes an assumption on the foreground appearance, which is similar to the appearance inside the box. On the other hand, it is an impractical decision to extract only one region (foreground) as object of interest. It is more practical to retain multiple regions.

In this thesis, local saliency is detected by Itti's model. This foreground subtraction algorithm is modified by introducing local saliency and we shall show these assumptions relaxed in our formulation. Considering these salient pixels as seed points, multiple regions are extracted simultaneously with respect to these pixels, unlike methods in [89] [97] [99] which assume a single dominant object in an image. Compared with the previous works, our contribution is the global measurement for multiple regions based on local saliency.

6.3 Multiple regions extraction based on visual attention

Experimental evidence suggests that attention can be tied to objects, object parts, or groups of objects [137] [138]. The usefulness of Itti's saliency mode has been demonstrated in various contexts. Even though its ability to serve as a front-end for object recognition is limited by the fact that salient pixels from this method do not cover the entire salient regions [37], the salient pixels fall well within salient regions [216].

6.3.1 Itti's model for saliency detection

The Itti's saliency model is shown pictorially in Figure 6.3. This computation model is briefly reviewed in this section.

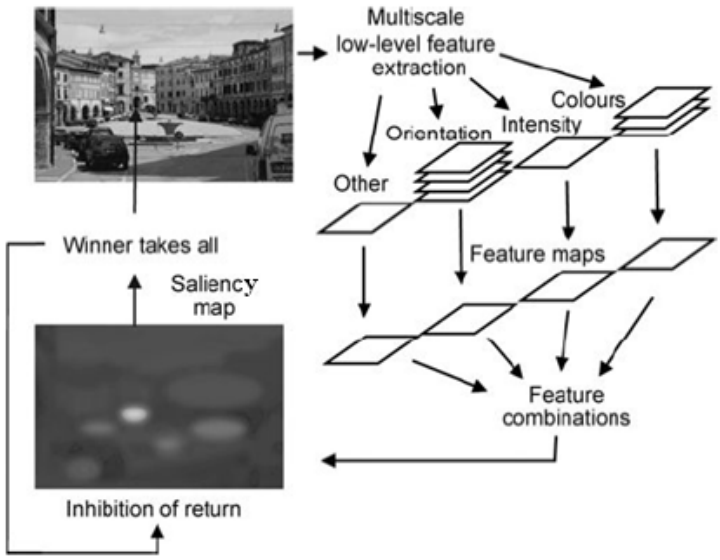


Figure 6.3 Itti's saliency model

In the implementation of the saliency model, low-level features are extracted in parallel across the extent of the viewed scene. Local competition across image space and feature scales results in feature maps for luminance contrast, color contrast, and orientation contrast [132]. These individual feature maps are combined [140] by weighted sum to create an overall distribution of local feature contrast, known as the "saliency map."

Attention is then allocated to the locations in the scene according to the saliency in the computed map using a winner-takes-all principle. The winning location of this process is attended to. In order to avoid attention becoming "stuck" at the most salient location, a local, transient inhibition is applied to each attended location. In each iteration of the model, winner-takes-all principle selects the most salient location, which is followed by the inhibition at the attended location. Thus, a relocation of attention is effectively represented.

6.3.2 Discontinuity preserving smoothing by Mean Shift

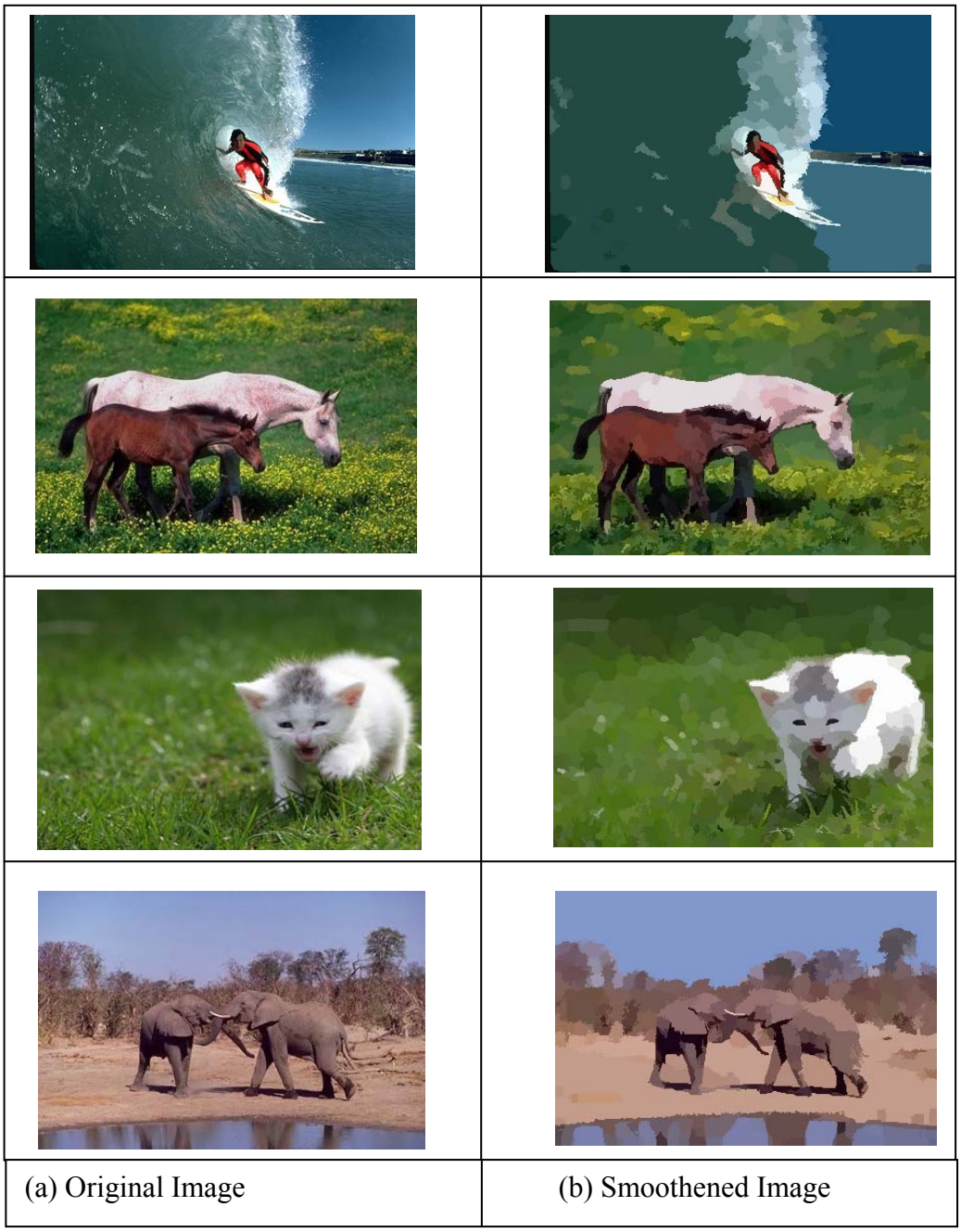
Based on the detected salient locations, methods have been developed to determine

the salient regions which could possibly contain objects of interest. Mostly, the feature map which contributes the most to the salient location is attended to, and the salient region is then segmented in this map.

In fact, the segmented region around the winner location shows the homogeneity of feature distribution. Usually these homogeneous regions are within a small area. Even though the salient locations are successfully detected, the detected salient regions most likely contain a portion of desired object if any. Smoothing is required to make the region extraction less sensitive to feature discontinuity.

A discontinuity preserving smoothing strategy is provided by the Mean Shift method. This method, proposed in [103], is a statistically iterative algorithm and has been used to analyze feature space successfully [104]. In this algorithm, the controlling parameter is the bandwidth and therefore the image quantization is not affected by the pre-defined cluster numbers.

To display the validity of this algorithm, original images and the smoothed images are shown in Figure 6.4. Images are from the Berkeley Segmentation Database (BSD) [57].



(a) Original Image (b) Smoothened Image

Figure 6.4 Discontinuity preserving smoothing

6.3.3 Combining visual saliency with foreground subtraction

In our formulation, the salient location is considered as a seed point and its corresponding foreground region is extracted by the foreground subtraction algorithm. With the knowledge of seed points and smoothed image, the original assumptions in [147] are relaxed by the following one:

A bounding box is centered at the detected salient point, covering majority of its associated homogeneous region.

The foreground associated with this salient point can be extracted by comparing color histograms of the regions inside and outside this bounding box. The proof is given below.

For a given color c:

Let $P(c/fg)$ and $P(c/bg)$ be the probabilities for c on the foreground and background, respectively.

Let $P(c/box)$ and $P(c/\sim box)$ be the computed probabilities for c inside and outside the bounding box, respectively.

We would like to prove the expression:

$$P(c/fg) > P(c/bg) \Longleftrightarrow P(c/box) > P(c/\sim box) \tag{6.1}$$

The physical meaning behind this expression basically shows that whenever a color c is more often found on the foreground than in the background, it is also true that c will be more likely to be found inside the bounding box than the outside of it.

The proof is straight forward:

$$P(c/box) = P(c/fg, box)P(fg/box) + P(c/bg, box)P(bg/box) \tag{6.2}$$

$$P(c/\sim box) = P(c/fg, \sim box)P(fg/\sim box) + P(c/bg, \sim box)P(bg/\sim box) \tag{6.3}$$

It is true that the color distribution is independent of location inside the homogeneous feature region, which is associated with the salient point. In the

meanwhile, this feature region contributes to the local saliency, providing the most promising features that best differentiate foreground from the background. Therefore, we have $P(c/\ fg, box) = P(c/\ fg)$ and $P(c/bg, box) = P(c/bg)$. Then Eq. (6.2) and Eq. (6.3) can be rewritten as:

$$P(c/box) = P(c/\ fg)P(\ fg/box) + P(c/bg)P(bg/box) \tag{6.4}$$

$$P(c/\sim box) = P(c/\ fg)P(\ fg/\sim box) + P(c/bg)P(bg/\sim box) \tag{6.5}$$

Considering $P(bg/box) = 1 - P(\ fg/box)$ and $P(bg/\sim box) = 1 - P(\ fg/\sim box)$, Eq. (6.4) and (6.5) will take the following forms:

$$P(c/box) = P(\ fg/box)(P(c/\ fg) - P(c/bg)) + P(c/bg) \tag{6.6}$$

$$P(c/\sim box) = P(\ fg/\sim box)(P(c/\ fg) - P(c/bg)) + P(c/bg) \tag{6.7}$$

Finally, $$\frac{P(c/box) - P(c/\sim box)}{(P(c/\ fg) - P(c/bg))} = P(\ fg/box) - P(\ fg/\sim box) \tag{6.8}$$

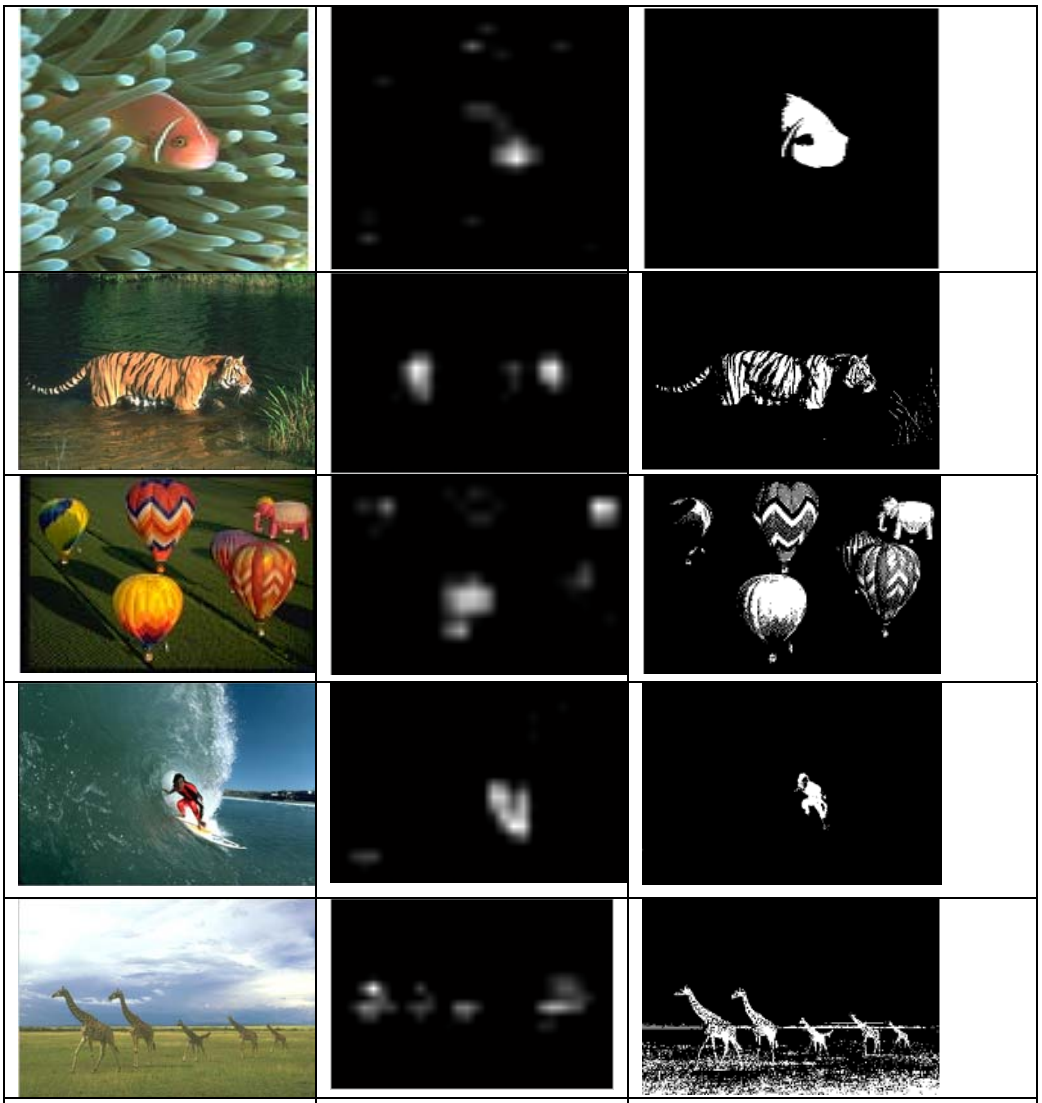
Under our assumption: the majority of this box belongs to the foreground, therefore:

$$P(\ fg/box) > 0.5 > P(\ fg/\sim box)$$

Whenever $P(c/\ fg) > P(c/bg)$ is true, $P(c/box) > P(c/\sim box)$ holds even for the box with the wrong size and vice versa. Expression (6.1) is hence proved!

Subsequently, each pixel of color c in the image is given the posterior value $P(c/box)/(P(c/box) + P(c/\sim box))$ at its location. This posterior probability is greater than 0.5 whenever the pixel of color c is more likely to locate in the foreground associated with the seed. Finally, a foreground region can be defined by the largest connected component touching the interior of the bounding box.

By locating the bounding boxes at various salient locations, multiple regions are extracted. Our strategy extracts the relevant regions by combining local saliency with statistic global description of colors in the foreground. Therefore, our strategy follows a local to global configuration. Some results are shown in Figure 6.5.



(a) Original image (b) Itti's saliency map (c) Foreground regions

Figure 6.5 Foreground regions extracted based on local saliency

6.4 Prominence evaluation of foreground regions

Among the regions that are simultaneously extracted with respect to the salient points, there could be some regions which are too trivial to contain any object of

interest, such as a small region from a "bright spot". Therefore, these regions are evaluated with respect to their prominence, as it is more likely for a prominent region to contain object of interest. The prominence evaluation is based on the following factors:

1. Area factor. Larger regions have greater prominence. This is represented as area factor which is simply the ratio of the area of a region to the area of the whole image and is computed as:

$$\varepsilon_1(R_i) = \frac{A_i}{A} \tag{6.9}$$

where A_i is the area of the i^{th} region R_i and A is the total image area.

2. Neighboring effect: In human vision system, the attractiveness of a unit is affected by the nearer neighbors much more than by farther ones. Thus, this property could be represented as a distance factor, which is an exponential function of the spatial distance between the i^{th} region R_i and j^{th} region R_j. It is written as:

$$\varepsilon_2(R_i, R_j) = 1 - \exp(-d(R_i, R_j) / 2\sigma_{\varepsilon_2}{}^2) \tag{6.10}$$

where $d(R_i, R_j)$ is the relative spatial distance between regions R_i and R_j, normalized to [0, 1].

3. Central effect: There is a strong tendency for observers to fixate the center of images irrespective of the scene's content as suggested in [96] [135]. A position factor is used to evaluate the central effect, which is represented as an exponential function:

$$\varepsilon_3(R_i) = 1 - \exp(-d(R_i, center) / 2\sigma_{\varepsilon_3}{}^2) \tag{6.11}$$

where $d(R_i, center)$ is the distance from the region R_i to the image center $center$, and it is normalized to [0, 1].

Considering all the factors above, the overall prominence score of a foreground region is defined as:

$$S_p(R_i) = \omega_{in} \cdot \bar{s}_i \cdot \varepsilon_1(R_i) \cdot \varepsilon_3(R_i) + \omega_{out} \sum_{j=1}^{N-1} (\bar{s}_j \cdot \varepsilon_1(R_i) \cdot \varepsilon_2(R_i, R_j)) \quad (6.12)$$

The first term represents the location and the size of the candidate region R_i, i.e., intra factor, while the second term takes care of the relationship between R_i and its neighbors i.e., inter factor. ω_{in} and ω_{out} are the weights of intra and inter factor respectively; $\bar{s}_i$ and $\bar{s}_j$ are the average pixel values in the foreground region R_i and R_j. Regions are then ranked according to its prominence score, which indicates the possibility for this region to contain object of interest. Two examples are shown in Figure 6.6, where numbers are assigned to the extracted foreground regions, which represent the ranking of their prominence scores.

Figure 6.6 Prominence ranking of foreground regions

6.5 Conclusions

Our intention to extract foreground is to find possible object regions based on visual

attentions. Salient locations are first found by the saliency model of bottom-up attention. Subsequently, the foreground subtraction method is modified to extract regions associated with individual salient points. These regions are further ranked by a prominence score for their possibility to host objects of interest. Our proposed approach suggests a way by which the recognition can be benefited from unsupervised training and learning multiple objects.

In the next chapter, we shall propose to use the region information obtained by the algorithm presented in this chapter to extend our algorithms of occlusion recognition in dynamic scenes.

Chapter 7

Recognition of occluded objects in dynamic scenes

With the concept of region information introduced in Chapter 6, we can now extend our proposed occlusion recognition algorithms to deal with situations involving dynamic scenes. In this chapter, approaches to propagate the region information will be proposed, so that information needed for recognition in one frame can be collected from the other frames where occlusion rates are lower.

In this chapter, two approaches are proposed to propagate region information acquired from the early frames to the later ones in an image sequence captured from a dynamic scene. There are two ways to acquire an image sequence, which could be taken by using a surveillance camera or a movable vision platform. In the case of a surveillance system, image sequence or video is taken by a still camera and a "Trace back" approach is proposed to enable the recognition process, regardless of the occlusion rates in the individual frames. When using a movable platform, disparity map is provided at each view position and a "Take a look around" approach is proposed to implement a view-updating process, which works by looking for frames where recognition algorithms can be applied successfully.

7.1 "Trace back" approach to integrate motion information

Video stream or image sequence is an available source in many computer vision systems. Image sequence provides information which varies with time. Changes of

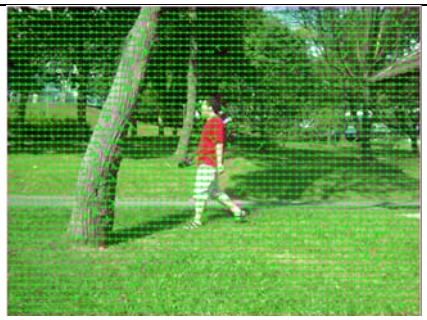
occlusion rates could be minor in consecutive frames, containing no perceptive changes in object appearance. With the consistency in consecutive frames, possible objects regions extracted could be associated throughout the image sequence, where invisible parts in one frame could be visible in other frames. Therefore, the information of the target object in one frame could be retrieved from other frames in order to improve the results of recognition. This approach is referred to as "Trace back".

7.1.1 Association of regions by motion smoothness constraint

Motion is a powerful cue, because the relative motion discrepancies at the depth discontinuities may occur at the occlusion boundaries [109]. The occlusion boundaries are considered as hints for object segmentation, which possibly benefits object recognition. Motion segmentation algorithm [128] can be used to find independently moving objects, under varying assumptions on the object movements. On the other hand, it is noted that most of the visible boundaries can be the occlusion boundaries, or they can also be edges between patches of different colors and/or texture in an object Therefore, multiple appearance cues are combined with local motion cues in [63] to detect occlusion. In [63], appearance and motion cues are combined in order to locate object regions through frames in video. In Chapter 6, multiple regions are extracted in individual frames and evaluated based on their prominence. However, it is impractical to decide which regions contain object of interest by their prominence scores alone.

In our proposed "Trace back" approach, a motion smoothness constraint is introduced to associate regions throughout the frames in the sequence. It is reasonable to assume that object of interest has no sudden change of movement in consecutive

frames, which is referred to as motion smoothness. In the case of a surveillance camera system, the object is visible in several consecutive frames due largely to the camera's wide field of view. Local consistency of appearance and movement of this object are shown in consecutive frames. Therefore, the region that contains object of interest in one frame would have a corresponding region along its direction of motion in the next frame, i.e., its motion correspondence as shown in Figure 7.1. Figure 7.1(a), is the motion estimate of frame 1 based on optical flow [107] and (b) shows its motion correspondence (in blue) in frame 2.



(a) Motion estimation in frame 1

(b) Motion correspondence in frame 2

Figure 7.1 Motion correspondence

For the region extracted (by algorithm proposed in Chapter 6) in frame 1, if its motion correspondence overlaps the region extracted in frame 2 (by algorithm proposed in Chapter 6) by 80% of its area, these two regions are considered associated, indicating that they are from the same object. Therefore, the region information has been propagated in these two frames.

Through a video stream, motion smoothness constraint is applied to associate regions in each frame, leaving groups of regions which could indicate the same object in the video. Regions which accidentally appear in several frames are ruled out. The scheme for region association in our work is shown in Figure 7.2.

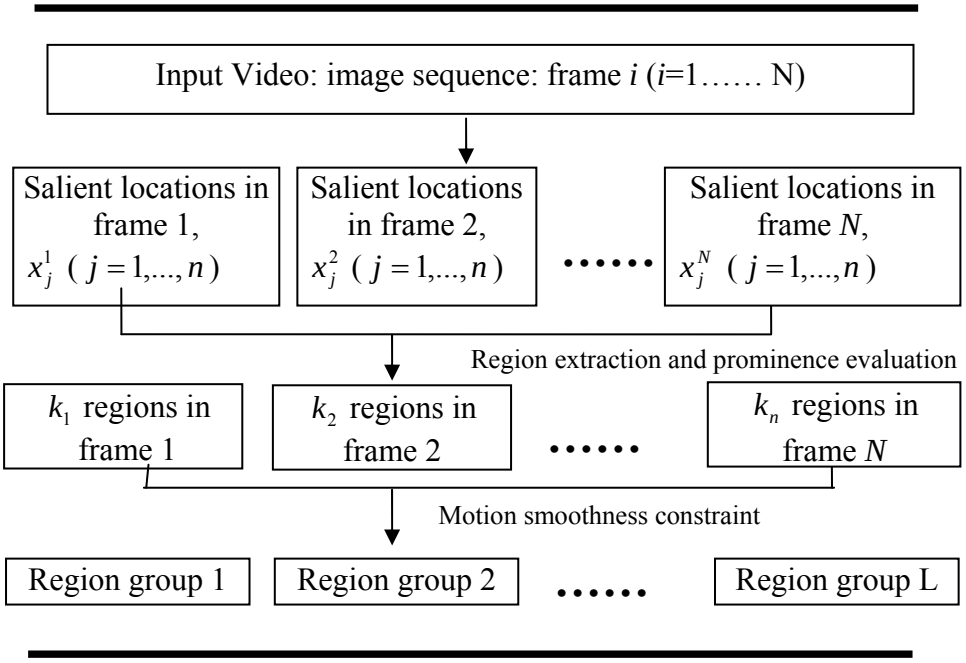


Figure 7.2 Scheme to associate regions based on motion smoothness

For individual frames, we fix the number of salient locations as $n = 5$. However, the foreground regions extracted may take a number of $k_1, k_2, \cdots, k_n \le 5$, as some salient location may not have a local support region, such as a single bright spot in the scene. These extracted regions from the consecutive frames are associated based on the motion smoothness constraint. In our association scheme, multiple objects can be accommodated, which results in having multiple regions groups. For single object of interest, there could only be one region group.

In the next section, these associated regions are further grouped according to their appearance information; and finally the description for each group is formed to enable object-level representation.

7.1.2 Recognition based on grouped regions

The object of interest can be partially or totally occluded in some frames of an image sequence. Since different parts of an object may not be visible simultaneously

in a single frame, regions associated by motion may display varying appearances for the same object, for example, the front and back of a car, or the front and side of a face. To complete the description of this object, object-level grouping of these regions is conducted based on appearance. This will enable object-level matching and recognition throughout a video clip, where the object may be occluded by different rates on different frames.

To explain the object-level region grouping, we define the object-level grouping as the determination of the set of regions which (a) last for a significant number of frames (motion association), and (b) move (semi-rigidly) together throughout the frames (appearance consistency). The novel idea implemented here is to use both motion and appearance consistency in order to group objects throughout the image sequence. Regions which are associated by motion smoothness will be further compared based on their appearance. Therefore, the objects of interest can be described by grouped regions through frames, among which the objects of interest might be under severe occlusions. In our experiment, regions with top prominence scores in each frame are first associated through a video clip. If the regions have corresponding regions in the 4 consecutive frames, they are considered as the potential object of interest. These associated regions are further grouped based on appearance similarity as object regions, in order to complete the object-level description. It is shown in Figure 7.3 that the relevant region in each frame is marked within a red rectangle (by using the algorithm in Chapter 6) and then the regions from the same object are tracked and identified throughout the image sequence, based on motion smoothness constraint and object-level grouping based on appearance. Therefore, the grouped tracked regions, i.e., all red rectangles in Figure 7.3, represent the same object of interest throughout the sequence.

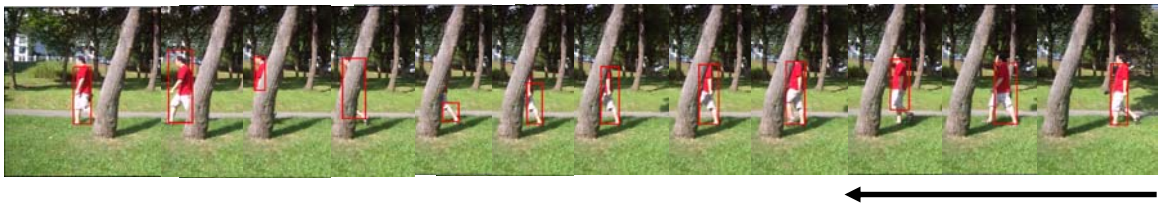


Figure 7.3 Grouped object regions through image sequence

In order to recognition the object of interest in the image sequence or video, the model images are matched with the grouped regions. If there is a confident match (by our recognition algorithms) between the model images and a region group, frames which contain these regions are traced back, as the object of interest has been recognized in these frames. Therefore, for frames which contain little information about the object of interest, i.e., a small part of the object due to high occlusion rates, our proposed recognition algorithms can still proceed by using object information from other frames, based on the grouped regions.

7.2 "Take a look around" approach to integrate stereo information

Region information introduced in Chapter 6 has been employed to occluded object recognition from image sequence captured by a surveillance camera. In this thesis, another situation is also considered, where a movable camera platform is used, such as an end-effectors-mounted camera on a robot. In this case, multiple views of a scene are accessible.

In an object recognition system, model images of objects of interest are usually taken from different viewpoints, and stored in the database. The number of views is determined by the tradeoff between the computational cost and the recognition confidence. Recognition uncertainty may be quite severe in a densely cluttered scene with a limited number of model images in the model database. The ideal situation is

that multiple views of the same scene can be collected in-situ or when the views are captured. During this process, the system will search for a view position where the object of interest is recognizable with the given model database of the object in question. This is to mimic human vision system: if we are not sure about the identity of an object in a cluttered environment, we will take a look round it and stop when the recognition succeeds. This implies a view-updating process, i.e., a better view is found by looking around to reduce occlusion rate. We would like to implement this view-updating process in computer vision system, which is referred to as "Take a look around" approach.

7.2.1 Regions from disparity map

Given that the movable platform with a stereo vision system (two cameras) is available, partially occluded objects may be successfully recognized by moving the stereo cameras to different viewpoints, in order to find the view where the objects are the least occluded for recognition purpose. With regions obtained as candidate object regions using the method presented in Chapter 6, the occlusion recognition algorithms are combined with the functionality of the movable camera platform system, shown in Figure 7.4. The moving trajectory of the platform is set such that the cameras always face the object of interest.

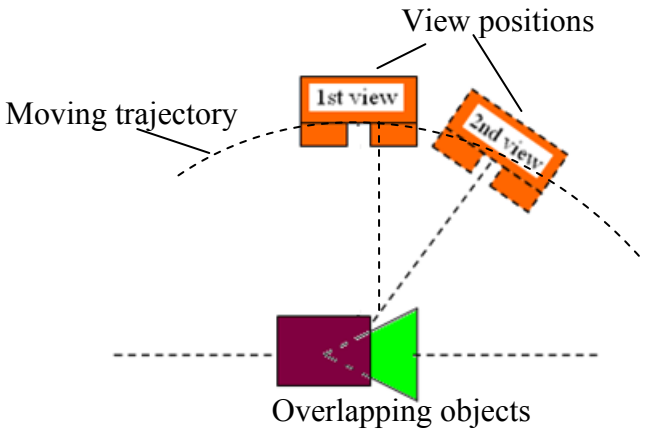


Figure 7.4 Views acquisitions by movable camera platform

It is difficult to obtain a 360° view around a scene. Therefore, an angular interval is required to obtain distinct viewpoints while being physically achievable in a typical indoor environment. We choose to process images at every 20°, facing the region of interest. This interval may be adjusted for different applications.

With the use of the movable platform, the stereo vision system can view the scene at different viewing positions. From one viewing position, the stereovision system can produce a disparity map. Theoretically, regions from the disparity map respond to different objects for their physical distances to the camera. In fact, disparity map from stereo matching itself is complicated and thus extraction of object regions in the disparity map of an arbitrary scene relies on the reliability and performance of stereo algorithms. We, on the other hand, did not set requirements on the performances of stereo algorithms. We shall obtain the object region from the refined disparity map, which will be introduced in Section 7.2.2. We shall also show how to update viewpoints based on the object region from the refined disparity map, after the introduction of stereo system calibration.

7.2.2 Stereo system calibration

Camera calibration is to find parameters that relate the world coordinates to the

image coordinates. A camera is modeled by the usual pinhole camera principle. The relationship between a 3-D point $[X, Y, Z, 1]^T$ and its image projection or coordinates $[u, v, 1]^T$ is given by:

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}\begin{bmatrix} \mathbf{R} & \mathbf{T} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \qquad \text{where } \mathbf{K} = \begin{bmatrix} \alpha_x & \gamma & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7.1}$$

K contains intrinsic parameters. $\alpha_x = f \cdot m_x$ and $\alpha_y = f \cdot m_y$ which represent focal length in terms of pixels, and m_x and m_y are the scale factors relating pixels to real-world distance; the subscripts "*x*" and "*y*" indicate the quantity in the two respective directions; γ represents the skew coefficient between the *x* and *y* axis; (u_0, v_0) is the center of the image plane, i.e., Principle Point. [**R T**] are the extrinsic parameters which denote the coordinate system transformations from 3-D world coordinates to 3-D camera coordinates. **K [R T]** forms the projection matrix which associates points in a camera's image space with locations in the 3-D world space. Camera calibration methods, such as Direct Linear Transformation (DLT), Roger Tsai's algorithm [49] and Zhang's algorithm [52], are designed to determine all parameters in **K**, **R** and **T**.

In Zhang's calibration method [52], where plane object captured is set to $Z = 0$. Eq. (7.1) is then simplified to Eq. (7.2), with $Z = 0$. Therefore, the projection matrix (**K[R T]**) is reduced to a 3×3 homography matrix **H**. The homography matrix (H) can then be processed into intrinsic parameter (**K**), rotation (**R**), and translation (**T**) matrices, which is deduced as follows.

$$z_c\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}\begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3 & \mathbf{t} \end{bmatrix}\begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \mathbf{K}\begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t} \end{bmatrix}\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{7.2}$$

Let $\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3 \end{bmatrix} = \mathbf{K}\begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t} \end{bmatrix}$, then we have $\mathbf{r}_1 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_1$ and $\mathbf{r}_2 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_2$.

here, λ_c is an arbitrary scalar;

$\mathbf{r}_1$ and $\mathbf{r}_2$ are rotational vectors about two neighboring axes of a Cartesian coordinate system, and hence they are orthonormal:

$$\mathbf{r}_1^T\mathbf{r}_2 = 0 \text{ and } \|\mathbf{r}_1\| = \|\mathbf{r}_2\| = 1 \tag{7.3}$$

Substituting $\mathbf{r}_1 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_1$ and $\mathbf{r}_2 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_2$ in Eq. (7.3), two constraints imposed on **K** are obtained as shown in Eq. (7.4), given one homography:

$$\begin{aligned} &\mathbf{h}_1^T\mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{h}_2 = 0 \\ &\mathbf{h}_1^T\mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{h}_1 = \mathbf{h}_2^T\mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{h}_2 \end{aligned} \tag{7.4}$$

Because a homography has 8 degrees of freedom and there are 6 extrinsic parameters (3 for rotation and 3 for translation), we can only obtain 2 constraints on the intrinsic parameters based on Eq. (7.4). For the five unknown parameters in **K**, at least three images of the model plane are required to solve for all the intrinsic parameters by Eq. (7.4). Once the intrinsic parameters are obtained, the extrinsic parameters are readily computed by Eq. (7.5):

$$\mathbf{r}_1 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_1; \mathbf{r}_2 = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_2; \mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2; \mathbf{t} = \frac{1}{\lambda_c}\mathbf{K}^{-1}\mathbf{h}_3 \tag{7.5}$$

Once the extrinsic parameters for each camera are known, the stereovision system can be calibrated, characterizing the relative position and orientation of the two cameras (a translation T transformation and a rotation R transformation), shown in Figure 7.5. For each camera, the projection of the world point $\mathbf{P}_w$ is presented as $\mathbf{p}_c$ in camera coordinates by their extrinsic parameters as shown in Eq. (7.6):

$$\mathbf{p}_{c1} = \mathbf{R}_1\mathbf{P}_w + \mathbf{T}_1 \quad \text{and} \quad \mathbf{p}_{c2} = \mathbf{R}_2\mathbf{P}_w + \mathbf{T}_2 \tag{7.6}$$

Thus, $\mathbf{p}_{c1} = \mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{p}_{c2} + \mathbf{T}_1 - \mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{T}_2$, and the transformation between two cameras is obtained by the following expression:

$$\mathbf{R} = \mathbf{R}_1\mathbf{R}_2^{-1} \text{ and } \mathbf{T} = \mathbf{T}_1 - \mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{T}_2 \tag{7.7}$$

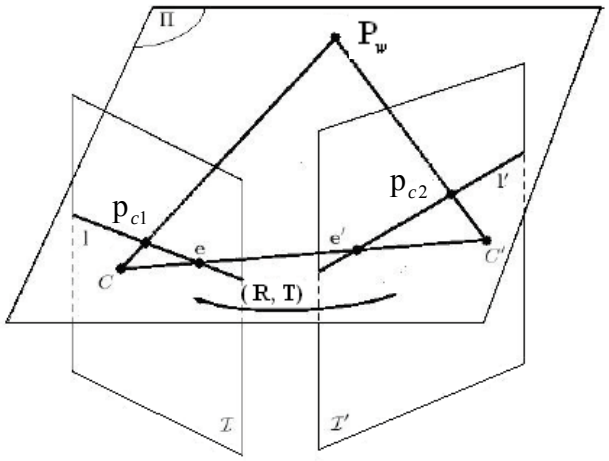
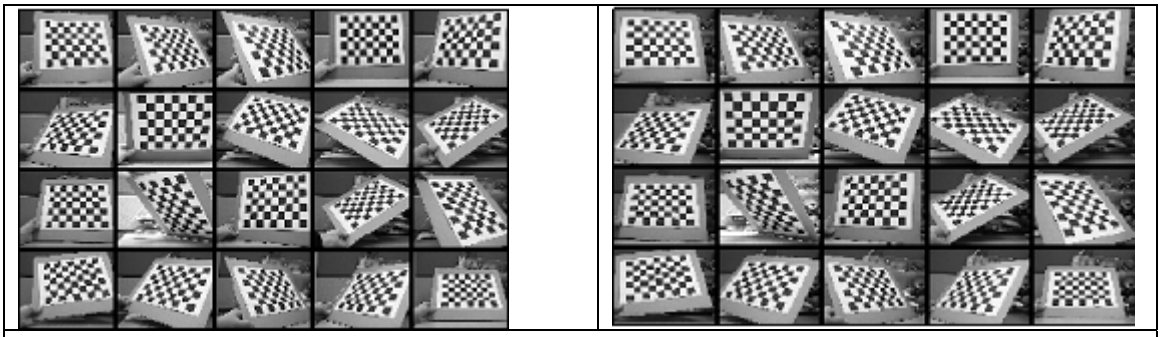
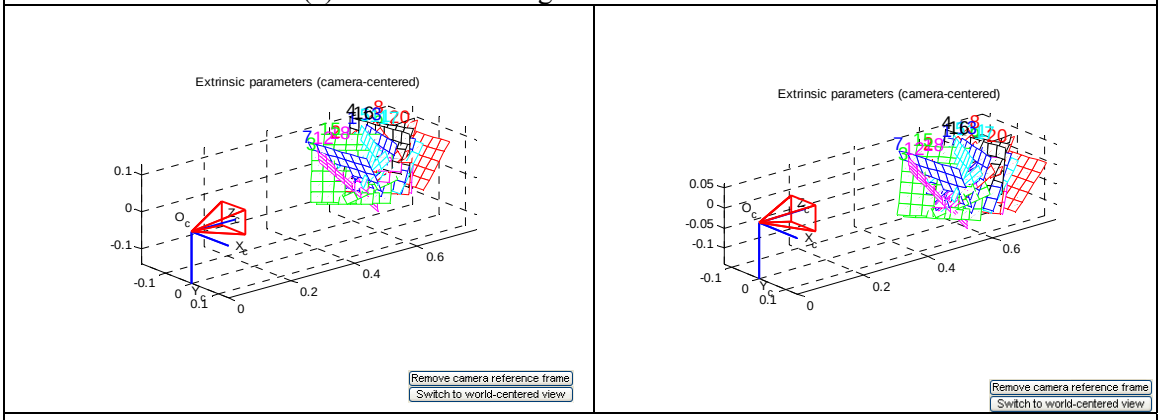


Figure 7.5 Epipolar geometry

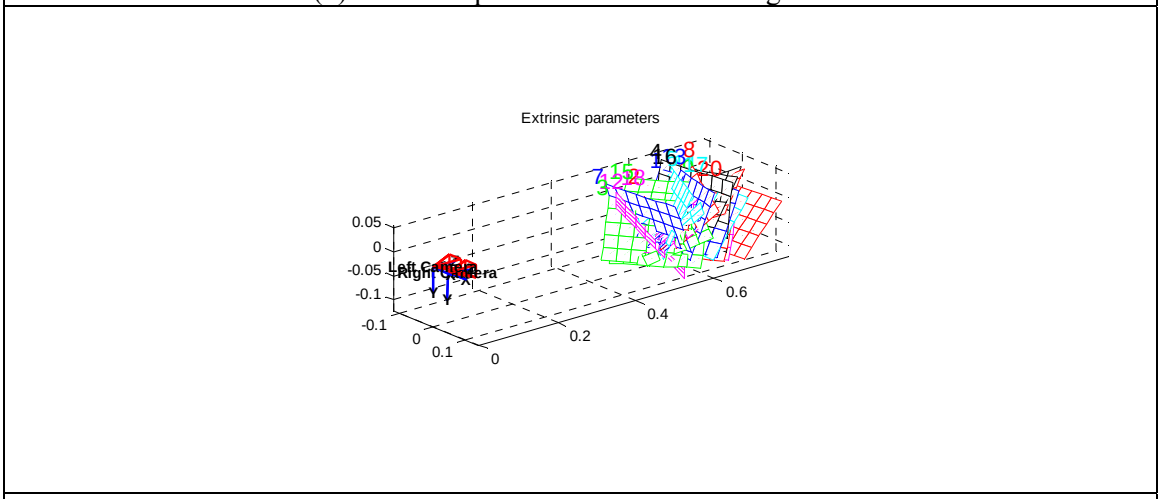
We adapt the steps stipulated in [52] and calibrate our system with 20 pairs of chessboard images with varying positions, shown in Figure 7.6 (a). The calibration results are pictorially shown in Figure 7.6 (b) and (c). The results of the intrinsic and extrinsic parameters are extracted from these figures and presented in Table 7.1.



(c) Chessboard images for two cameras calibration



(d) Extrinsic parameters for left and right camera



(c) Relative positions of the two cameras

Figure 7.6 Stereo system calibration

The calibration result is listed in Table 7.1. The intrinsic camera parameters (Focal Length, Principle Point, distortion) for each camera are shown in Table 7.1. In addition, the extrinsic parameters **R** and **T** are obtained by Eq. (7.7), which characterize the relative position and orientation of the two cameras.

Table 7.1 Calibration parameters for the stereo system

Left camera			Right camera		
Focal Length	Horizontal	$\alpha = 494.165$ (pixel)	Focal Length	Horizontal	$\alpha = 497.50942$ (pixel)
	Vertical	$\beta = 492.576$ (pixel)		Vertical	$\beta = 495.84945$ (pixel)
Skew	$\gamma = 0$		Skew	$\gamma = 0$	
Principle point	u_0 =149.47882 v_0 =72.59524		Principle point	u_0 =155.81019 (pixel) v_0 =133.27308 (pixel)	
Distortion (radial)	$k_1 = -0.11504$ $k_2 = 0.23891$		Distortion (radial)	$k_1 = -0.14627$ $k_2 = 0.23646$	
Extrinsic parameters	Rotation		**R**=[1 -0.0124 -0.0082 0.0119 1 -0.0608 0.0089 0.0607 1]		
	Translation		**T** = [-0.04045 -0.00148 0.00272]		

Based on the calibration parameters, image pairs taken by the two cameras can be rectified by the following way. Given a pair of stereo images, rectification determines a transformation between the image planes such that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes. Subsequently, correspondence searching is carried out along these lines (in green). An example of a rectified image pair in one view is shown in Figure 7.7.

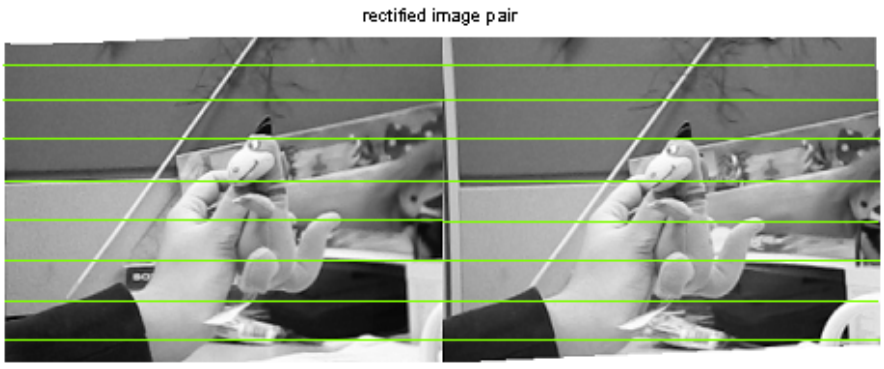


Figure 7.7 Rectified image pair in one view

7.2.2 Object region from the refined disparity map

The main advantage of rectification is that the stereo correspondences search is reduced to a 1-D search problem along the horizontal raster lines of the rectified images. The disparity map can then be obtained from a pair of rectified images, using self-adapting dissimilarity measure [148]. The disparity map measurement is shown in Figure 7.8 and it computes the difference in image location of the object seen by the left and right camera.

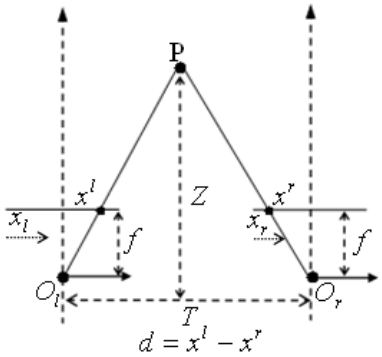
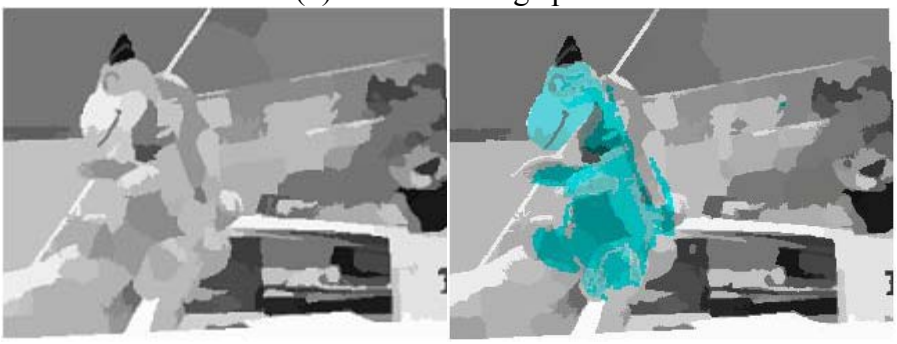


Figure 7.8 Disparity measurements

In general cases, regions can be found from the disparity map, which carry similar disparity values. However, stereo algorithms generate noisy regions due to their own imperfections as shown in the left image of Figure 7.9 (c).

(a) Original image pair

(b) Rectified image pair



(c) Disparity map and object region from refined disparity map

Figure 7.9 Object region from refined disparity map

With the region extracted by the algorithm proposed in Chapter 6, the object region can be more correctly separated from the disparity map, shown in the right image of Figure 7.9 (c). The region information is incorporated into the disparity map by assigning each pixel there with a region indication, which results in a refined disparity map. The refined disparity map is then segmented and the largest region is defined as the object region. Assuming the object of interest is the largest region is reasonable because human is looking for views where object of interest is less occluded and provides dominate information.

Object region defined in this manner comes from the combined consideration of the disparity discontinuity and local appearance. Our object detection based on the disparity map with region indications prompts for a view-updating procedure, by which a better view is determined from multiple views to reduce the occlusion rate for recognition.

7.2.3 View updating by growing an object region

For occluded objects, a better view can be expected when multiple views are available. A better view is defined as one where the occlusion rate is reduced to an extent for successful recognition. The idea to look for a better view is to mimic human response to the recognition of occluded objects. When the object of interest is unrecognizable because of occlusion, we would take a look around it till we can recognize it. With movable vision platform, this idea can be implemented by a "Take a look around" approach. Among the views of an occluded scene, our "Take a look round" approach looks, view by view, for the one where the occluded object can be successfully recognized. This can be regarded as a view-updating process.

In particular, a view is updated if the size of the detected object region grows. The object region is detected from the refined disparity map in one view position. As the views change, the occlusion rates for the object of interest vary accordingly, implying the change of the size of the detected object region. For two consecutive views, the flowchart for view updating is shown in Figure 7.10. If the detected object region is growing for one object, the second view is considered a better view. If there is no apparent change of the object region, the next two views (at 20 degree interval) will be considered before changing the direction of movement of the system. The view-updating process terminates when our occlusion recognition algorithms show a confident recognition at a updated view. The confident recognition is determined if our refined recognition confidence (defined in Eq. (5.7)) is larger than 0.7 in Algorithm 2 or there are at lease 3 matched features in Algorithm 1 to generate object 3D pose as stated in [74]. If there are no confident recognition at all the viewing positions, this view-updating process also terminates. Such a situation could occur

when there is no change in occlusion rates for the object of interest along the moving trajectory of the platform.

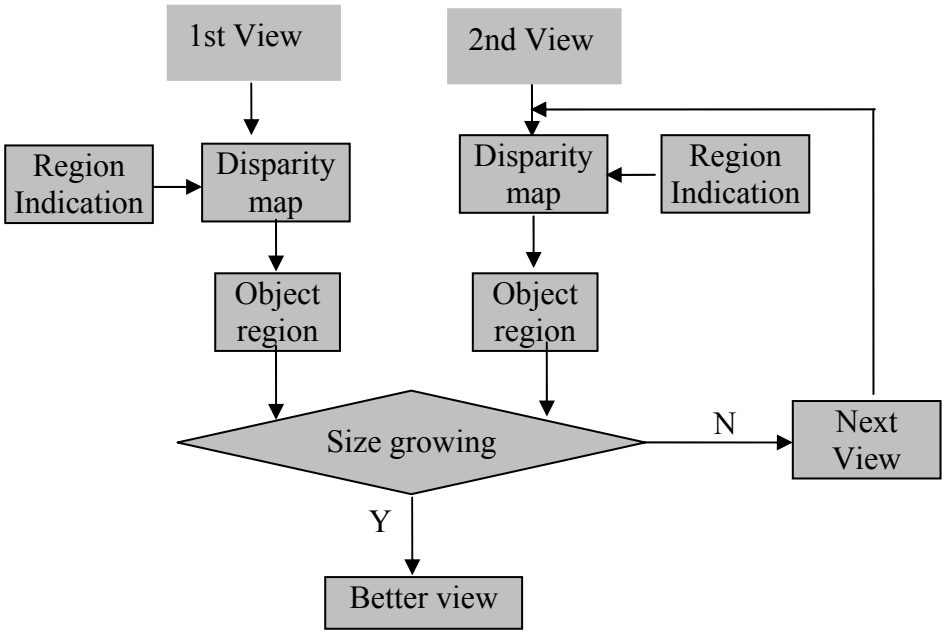


Figure 7.10 View updating between two views

A view-updating example is shown in Figure 7.11. The stereovision system rotates around the occluded target (pink dinosaur) to generate two views. It is noticed in Figure 7.11(e) and (g) that the disparity map is quite poor for the unstructured scenes to separate regions. Fortunately, our strategy of determining object regions does not depend on the quality of disparity map alone. Combining the visual attention based region with the disparity discontinuity, the object regions are defined from the refined depth map, marked in green in Figure 7.11 (f) and (h). The growing of the size of the detected object region in view 2 suggests that view 2 is a better view than view 1, where occlusion rate is reduced. Thus, we have completed one step of "Take a look round".

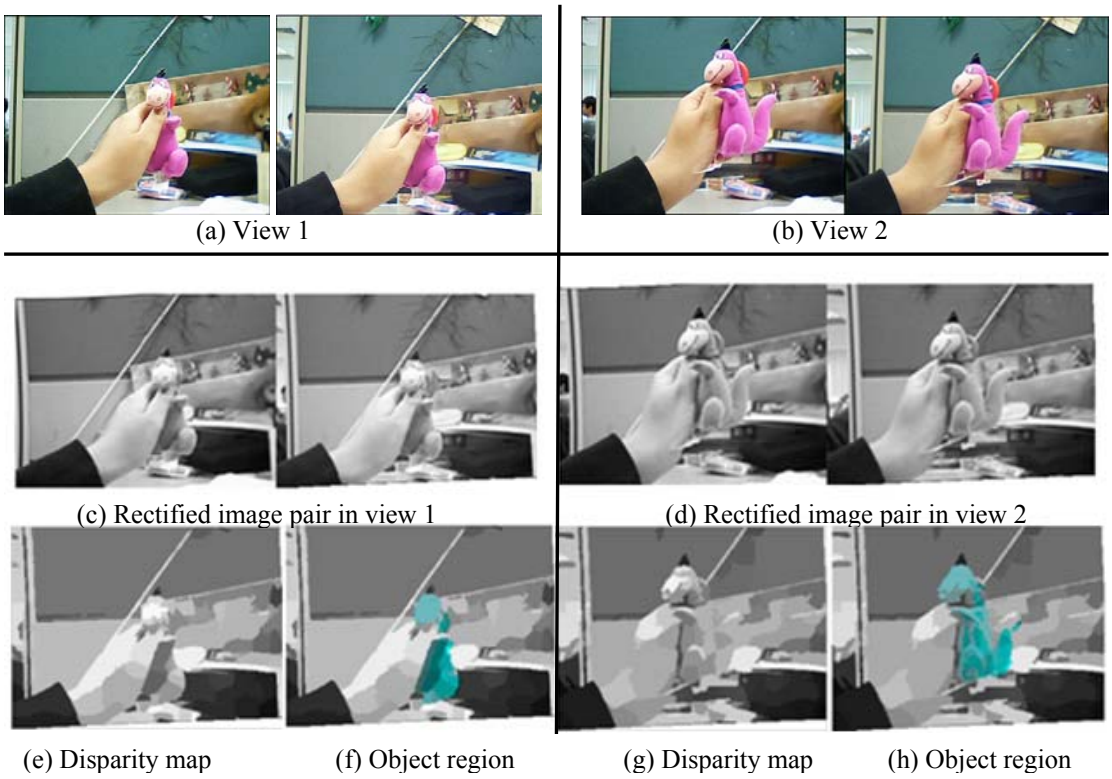


Figure 7.11 A better view by "Taking a look around"

7.3 Conclusions

In this chapter, we have extended our recognition algorithms in dynamic scenes by two approaches based on the region information obtained in Chapter 6. In our experiments, visual information is carried by video stream from surveillance camera and multiple views from movable vision system. We have utilized motion and stereo information in these two situations to help the recognition task.

In the surveillance system, motion smoothness constraint is applied to the consecutive frames to associate the candidate object regions. With our "Trace back" approach, occluded objects can be recognized even when it is completely "blacked out" in one frame, because its associated regions in other frames can always be used for recognition. On the other hand, with the help of the movable vision platform, multiple views of the scene are accessible. We mimic the human nature response that

we take a look around when the object of interest is unrecognizable due to occlusion. For multiple views obtained by this platform, our “Take a look around” approach employs a view-updating process, which is used to search for a better view until the occlusion is reduced enough to obtain a successful recognition. Our strategy to obtain object regions combines the depth discontinuity with the region indications. It also suggests a possible way to introduce appearance cue into stereo vision, in order to refine the disparity map. This “Take a look around” method updates a view when the size of detected object region grows. This updating procedure terminates when the object is successfully recognized, with the stipulated criteria presented in Chapter 5.

Chapter 8

Conclusions and discussions

This thesis addresses the problem of recognition of occluded objects. The algorithms are proposed with respect to the characteristics of occlusion recognition problem: features from different objects in a scene tend to interact, making the recognition task by matching local features alone difficult. Our principle in the recognition of occluded objects, then, is to reduce local interactions between features from different objects and make a global recognition decision based on locally gathered information. Amongst the objects in a given scene, we take into consideration their pairwise feature relationships, local to global matching strategy and the visual attention guided extraction of regions containing possible objects, in order to reduce the local interactions. Our main goal is to develop efficient algorithms for matching and recognition of partially occluded objects. In our recognition algorithms, how well the features are matched is measured at the local (first-order) level, and how well their geometric structure and appearance consistency are preserved is assessed at a higher order level. The proposed algorithms introduce the conventional graph matching into occlusion recognition, integrating low-level and intermediate-level vision cues into recognition. Furthermore, we propose to recognize occluded objects in dynamic environments, thereby extending the applications of our occlusion recognition approaches.

8.1 Summary of the thesis

In Chapter 1, the objective of this study was stated as the recognition of occluded objects in a cluttered scene. In such a situation, occlusion cannot be detected in a single image and the assumption for aggregation process in feature description is violated. Therefore, the presence of occlusion could result in the failure of feature based object recognition algorithms**.** Based on the studies carried out in this thesis, the solution to the occlusion recognition problem can be found by using algorithms which reduce the interactions of features from different objects. Furthermore, the local to global nature of occlusion recognition problem leads us to spectral matching.

In Chapter 2, we have reviewed the existing algorithms for occluded object recognition. Their solutions are, either to design specific features or to make use of the geometric relationships between features. To generalize our study, feature detectors and descriptors developed in object recognition systems were also reviewed. Based on these literatures, our algorithms would consider features for both their geometric and appearance information. Moreover, the spatial structure of target object could be preserved by graph matching. Subsequently, relevant work in graph matching was reviewed. As for feature interaction reduction methods, we were interested in integrating the intermediate-level vision cues into our schemes. Therefore, the perceptual grouping and image segmentation methods were reviewed.

In Chapter 3, the correspondence problem was formulated as a quadratic programming of graph matching based on the pairwise feature geometry. The integer quadratic programming as well as its spectral relaxation algorithms was introduced. These spectral approximation algorithms have been investigated with respect to their performances under different degrees or rates of occlusions and outliers. Through

experiments, proper weighting function has been determined to be Gaussian. Spectral Matching with Affine Constraints (SMAC), with Integer Projected Fixed Point (IPFP) as a post-processing step, has shown to be the most effective and stable algorithm under various occlusion rates. This spectral matching algorithm (i.e., SMAC+IPFP) has then been employed for our recognition purpose.

In Chapter 4, the other concern of our solution to the recognition of occluded objects has been addressed, i.e., feature interactions. Due to features interactions, ambiguous connections would be present in the proximity matrix during the graph formulation. We proposed to reduce feature interactions by Appearance Prior as well as Feature Association. These two approaches indicate how likely the pairwise features come from the same object or the object of interest, respectively. Inter and intra feature relationships are established and integrated. Both of these two approaches are effective in reducing feature interactions, which has been experimentally verified.

Two algorithms have been proposed in Chapter 5 to adapt the spectral matching algorithm for our work. They are the result of introducing the strategies of feature interactions reduction into spectral matching. Therefore, the geometric and appearance information of the object of interest are integrated. Algorithm 1 combined spectral matching with Appearance Prior (A.P.). It is shown that the performances of matching algorithms have been improved by A.P. and our algorithm has the best performances in occlusion recognition, compared to the other matching algorithms. In Algorithm 2, a two-stage algorithm has been proposed by integrating Feature Association (F.A.) into spectral matching. It gives the best performance among all the compared matching algorithms. Specifically, the effect of F.A. alone in boosting

occlusion recognition has been evaluated.

In Chapter 6, we have been motivated to consider the intermediate vision cue, i.e., the region, in order to extend our recognition algorithms to be used in dynamic scenes. The foreground subtraction algorithm has been modified by introducing visual attention based saliency. Salient locations are returned by saliency model and then multiple foreground regions are obtained simultaneously with respect to the detected salient locations. These regions are then ranked based on their prominence. The regions with high prominence scores are considered as possible object regions.

In Chapter 7, we extended the application of our occluded object recognition algorithms into dynamic environments in which the occlusion rates vary with time. Motion and stereo information have been introduced into the recognition task based on region information (i.e. possible object regions) which is defined and discussed in Chapter 6. With motion smoothness constraint, region information has been propagated through the image sequence by our proposed "Trace back" approach. Therefore, objects of interest can be recognized throughout this video even if the object is totally obscured in some of the frames. On the other hand, we proposed to use a 'Take a look around' approach to implement a view-updating process. The object region is obtained from the refined disparity map, where the stereo depth is combined with regions information. A view is updated if the object region grows in this view, implying the reduction of the occlusion rate.

8.2 Contribution &limitations

The contributions and lessons learnt from the work are presented as follows:

1) Introducing graph matching into recognition of occluded objects

We have studied the problem of recognition of occluded objects, commonly known as occlusion recognition. In occlusion scenario, global features are corrupted and unreliable for the recognition task. To successfully recognize the occluded objects, a global decision has to be made based on locally gathered information. This local to global nature of occlusion recognition problem has brought us to the graph matching theory. Chapter 5 addressed the novel algorithms to handle occlusions by graph matching, which has long been an open issue for the graph matching algorithms. Popular spectral algorithms are evaluated according to their performances under different occlusion rates.

2) Interaction reduction for occlusion recognition

We proposed that reducing feature interactions is a key to recognize occluded objects. Accordingly, we have addressed this problem by soft grouping based on the appearance similarity as well as the associating features with objects of interest. There are two approaches proposed to handling feature interactions in a bottom-up or top-down fashion, respectively.

In Chapter 4, local appearance information was interpreted as the statistical probability of two features to be from the same object or parts. This probability could serve as priors for further procedures, i.e., Appearance Prior. Compared to the methods which learn priors from training set, our Appearance Prior formulation suggests an efficient and one-shot-recognition solution. Besides the bottom up processing of linking features in a scene image, top-down reasoning was proposed by associating the individual features with the object of interest. In this context, the selection of potential correspondences is given physical meaning, which is considered as one step further than *k*-nearest neighbors used in Spectral Matching (Marius 2005).

Moreover, the occlusion rate has been taken into consideration by redefining the recognition confidence.

However, these two kinds of priors only work when there are still certain amounts of feature points lying on the object of interest. In extreme situations, very few features could be detected on the object of interest, where the occlusion rate is over 70%, the object surfaces are texture-less or the pictorial properties are distributed uniformly. This is a common dilemma which may be faced by the feature based object algorithms. In these cases, specially designed features are required.

3) Simultaneous extraction of possible object regions based on visual attention

To further explore the problem of reducing interactions, region information was extracted based on local saliency. Region extraction is attention guided for possible object regions with no assumptions on the appearance of the object of interest. The foreground subtraction strategy has been modified by introducing visual saliency, so that multiple regions are extracted simultaneously. These regions are further ranked based on their prominence scores. Regions with high scores are considered as candidate object regions for recognition. Extracting potential object regions based on local saliency for further recognition follows the trend in computer vision area: integrating low-level (saliency map) and intermediate-level (foreground) vision cues into higher level vision task (recognition).

4) Recognizing occluded objects in dynamic scenes

Occlusion recognition is also explored in dynamic environments, where occlusion rates change in a dynamic way. Based on these possible object regions (regions with high prominence scores), motion smoothness constraint for the same object is applied

through the image sequence taken by surveillance camera. Even when object region is not informative in a frame for recognition (object of interest is under severe occlusion), recognition can still be performed by matching with its associated regions in other frames. This approach is referred as “Trace back”. In dynamic scenes taken by movable camera platform, we implement the idea of ‘Take a look around’. With this platform, binocular stereo vision system is able to access different views of an occlusion scene. Disparity map is refined by integrating it into region indication, which is considered to be equivalent to introducing appearance information into stereo algorithms. A viewpoint is updated by moving the vision system towards and along the direction in which the size of object region grows, implying reduced occlusion rates for successful recognition.

Our occlusion recognition algorithms perform without the assumption of knowing where and how the occlusion occurred, which is the assumed knowledge for many other occlusion recognition algorithms. In our formulation to recognize occluded object in dynamic scenes, motion and stereo information are introduced into recognition, which implies the integration of multiple vision applications.

8.3 Future work

The work presented in this thesis can be extended into several important avenues which are presented below. In addition, they might be combined, eventually to solve many other similar problems.

1. Handling random occlusion with a pruned search space

Our work focuses on occlusions caused by objects overlapping and its applications can be found in situations like autonomous robot navigation, industrial inspection and

security surveillance. There are other types of occlusions, known as random occlusions and corruptions, which may be caused by large changes of illumination and deformation. For instance, sparse coding algorithms introduced in [130] [131] are used to handle occlusion in face recognition, where local change of illumination in faces causes occlusion. These algorithms are mathematically elegant and efficient to solve recognition problem by L1 optimization. However, the implementation of these methods requires the object of interest to be the dominating part or the only object in the image, such as face images. Nevertheless, the sparse coding algorithms could be a promising tool to recognize occluded objects from cluttered background with a pruned search space through our feature association strategy or region indication, which could suggest another research direction.

2. Occlusion detection and evaluation

We have realized that the recognition rate is closely related to the occlusion rate. For single image, the refined recognition rate has taken the occlusion rate into consideration with our formulation in Chapter 5. With image sequence, there is more evidence for occlusion detection. It is important to evaluate the extent of occlusions before embarking in designing and implementing any vision applications. When occlusion rate is over certain value, the applications, such as object tracking and recognition, should be terminated to save computational cost.

List of Publications:

Conference papers:

[1] Wu Jiayun and Lim Kah Bin, Chen Xiao, Recognition of Occluded Objects by Feature Interactions, IEEE Conference on Automation and Mechatronics, Singapore, 2010

[2] Lim Kah Bin and Wu Jiayun, Recognizing occluded objects by Spectral Matching, IEEE Conference on Computer Design and applications, Xi'an, China,2011

[3] Wu Jiayun and Lim Kah Bin, A color grouping method for detection of foreground regions, Twelfth International Conference on Control, Automation, Robotics and Vision, 2012 (accepted).

Journal papers:

[1] Wu Jiayun and Lim Kah Bin, A spectral technique to recognize occluded objects, In IET Image Processing, 6(2): 160–167, 2012.

[2] Lim Kah Bin and Wu Jiayun, Recognition of occluded objects by reducing feature interactions, Image and Vision Computation, 2012 (submitted)

Reference:

[1] F. R. K. Chung, Spectral Graph Theory. In *CBMS Series, Vol. 92, American Mathematical Society, Providence, RI, 1997.*

[2] L. Hagen and A. Kahng, New Spectral Methods for Radio Cut Partitioning and Clustering. In *IEEE Transactions on Computer-Aided Design, 11(9): 1074-1085 1992.*

[3] M. Meila and J. Shi, Learning Segmentation by Random Walks. In *Advances in Neural Information Processing, 2000.*

[4] M. Meila and J. Shi, a Random Walks View of Spectral Segmentation. In *proceedings of AI and STATISTICS (AISTATS) 2001.*

[5] A. Y. Ng and M. Jordan et al., On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing, 2002.*

[6] U. Von. Luxburg, A tutorial on Spectral Clustering. In *Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics, 2006.*

[7] M. Newman and D. Watts et al., Random Graph Models of Social Networks. In *Proceeding of National Acad Sciences, 2002.*

[8] J. Shi and J. Malik, Normalized Cuts and Image Segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):888-905, 2000.*

[9] B. Bhanu and J. C. Ming, Recognition of Occluded Objects: a Cluster-Structure Algorithm. In *Pattern Recognition, 20(2):199-211, 1987.*

[10] D. Verma and M. Meila, Comparison of Spectral Clustering Algorithms. In *Advances in Neural Information Processing, 2003.*

[11] T. H. Du and K. B. Lim et al., 2-D Occluded Object Recognition Using Wavelets. In *proceeding of the 4th International Conference on Computer and Information Technology (CIT'04), Wuhan, China, 2004.*

[12] K. B. Lim and T. H. Du, a Wavelet Approach for Partial Occluded Object Recognition. In *proceeding of the 1st International Symposium on Digital Manufacture (ISDM'2006), Wuhan, China, 2006.*

[13] M. Leordeanu and M. Hebert, a Spectral Technique for Correspondence Problems using Pairwise Constraints. In *International Conference on Computer Vision, 2005.*

[14] K. Price, Matching Closed Contours. In *IEEE Workshop on Computer Vision, 1984.*

[15] A. Berg and T. Berg et al., Shape Matching and Object Recognition Using Low Distortion Correspondences. In *International Conference on Computer Vision, 2005.*

[16] B. Bhanu and O. D. Faugeras, Shape Matching of Two Dimensional Objects. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(2): 137-155, 1984.*

[17] H. C. Liu and M. D. Srinath, Partial Shape Classification Using Contour Matching in Distance Transformation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(11):1072-1079, 1990.*

[18] L. Shapiro and R. Haralick, a Metric for Comparing Relational Descriptions. In IEEE Transactions on Pattern Analysis and Machine Intelligence, *7(1):90-94, 1985.*

[19] H. Wang and S. C. Yan et al., Correspondence Propagation with Weak Priors. In *IEEE* Transactions on *Image Process. 18(1):140-50, 2009.*

[20] H. Bai and E. Hancock, Graph Matching Using Spectral Embedding and Semidefinite Programming. In *British Machine Vision Conference, 2004.*

[21] M. Kumar and P. Torr et al., Solving Markov Random Fields Using Second Order Cone Programming Relaxations. In *International Conference on Computer Vision and Pattern Recognition, 2006.*

[22] C. Boeres, Heursticas Para Reconhecimento de Cenas por Correspondencia de Grafos. In *PhD thesis, Universidade Federal do Rio de Janeiro, Brazil, 2002.*

[23] E. R. Hancock and J. Kittler, Edge-labeling Using Dictionary-based Relaxation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(2):165–181, 1990.*

[24] K. Boyer and A. Kak, Structural Stereopsis for 3D Vision. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(2):144-166, 1998.*

[25] M. Galun and E. Sharon et al., Texture Segmentation by Multiscale Aggregation of Filter Responses and Shape Elements. In *International Conference on Computer Vision, 2003*.

[26] M. Carcassoni and E. R. Hancock, Point Pattern Matching with Robust Spectral Correspondence. In *International Conference on Computer Vision and Pattern Recognition, 2000.*

[27] R. Collins and Y. Liu et al., Online Selection of Discriminative Tracking Features. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10):1631–1643, 2005.*

[28] H. Bunke, Error Correcting Graph Matching: On the Inuence of the Underlying Cost Function. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):917–922, 1999.

[29] H. Bunke and K. Shearer, a Graph Distance Metric Based on the Maximal Common Subgraph. In *Pattern Recognition Letters, 19(3): 255–259, 1998.*

[30] R. C. Bolles and R. A. Cain, Recognizing and Locating Partially Visible Objects: The Local-Feature-Focus Method. In *International Journal of Robotics Research, 1(3):57-82, 1982.*

[31] E. Grimson, Recognition and Localization of Overlapping Parts from Sparse Data. In *Technical Report AI Memo 763, AI Laboratory, MIT, 1987.*

[32] D. Walther and U. Rutishauser et al., Selective Visual Attention Enables Learning and Recognition of Multiple Objects in Cluttered Scenes. In *Computer Vision and Image Understanding, 100 (1-2): 41-63, 2005.*

[33] J. M. González-Linares and N. Guil et al., an Efficient 2D Deformable Objects Detections and Location Algorithm. In *Pattern Recognition, 36(11):2543-2556, 2003.*

[34] Y. Deng and B. S. Manjunath et al., an Efficient Color Representation for Image Retrieval. In *IEEE Transactions on Image Processing, 10(1):140–147, 2001.*

[35] D. Walther and C. Koch, Modeling Attention to Salient Proto-objects. In *Neural Networks, 19(9):1395–1407, 2006.*

[36] W. J. Christmas and J. Kittler et al., Structural Matching in Computer Vision Using Probabilistic Relaxation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(8):749-764, 1995.*

[37] U. Rutishauser and D. Walther et al., Is Bottom-Up Attention Useful for Object Recognition? In *International Conference on Computer Vision and Pattern Recognition, 2004.*

[38] F. J. Estrada and P. Fua et al., Appearance-based Keypoint Clustering. In *International Conference on Computer Vision and Pattern Recognition, 2009*

[39] S. Sarkar and K. Boyer, Quantitative Measures of Change Based on Feature Organization: Eigenvalues and Eigenvectors. In *Computer Vision and Image Understanding, 71(1):110-136, 1998.*

[40] B. C. Russell and A. A. Efros et al., Using Multiple Segmentations to Discover Objects and Their Extent in Image Collection. In *International Conference on Computer Vision and Pattern Recognition, 2006.*

[41] C. Schimid and R. Mohr, Local Gray Value Invariants for Image Retrieval. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5):530-535, 1997.*

[42] D. G. Lowe, Distinctive Image Features from Scale-invariant Keypoints. In *International Journal of Computer Vision, 60(2):91–110, 2004.*

[43] T. Cour and P. Srinivasan et al., Balanced Graph Matching. In *Advances in Neural Information Processing,* 2006.

[44] D. Lowe, Perceptual Organization and Visual Recognition. *Kluwer Academic, Boston, 1985.*

[45] D. Crandall and D. Huttenlocher, Weakly Supervised Learning of Part-based Spatial Models for Visual Object Recognition. In *European Conference on Computer Vision, 2006.*

[46] A. D. J. Cross and E. R. Hancock, Graph Matching with A Dual-step EM Algorithm. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1236-1253, 1998.*

[47] D. Marr, Vision. *Freeman Publishers, 1982.*

[48] C. Fowlkes and D. Martin et al., Learning Affinity Functions for Image Segmentation: Combining Patch-based and Gradient-based Approaches. In *International Conference on Computer Vision and Pattern Recognition, 2003.*

[49] R. J. Tsai, a Versatile Camera Calibration Technique for High Accuracy 3D Machine Vision Metrology Using Off the Shelf Cameras and Lenses. In *IEEE Journal of Robotics and Automation, 3(4):323-346, 1987.*

[50] V. Kolmogorov and R. Zabih, Computing Visual Correspondence with Occlusions via Graph Cuts. In *International Conference on Computer Vision, 2001.*

[51] P. Chang and J. Krumm, Object Recognition with Color Co-occurrence Histograms. In *International Conference on Computer Vision and Pattern Recognition, 1999.*

[52] Z. Zhang, a flexible new technique for camera calibration. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000.*

[53] O. Duchenne and F. Bach et al., A tensor-based Algorithm for High-Order Graph Matching, In *International Conference on Computer Vision and Pattern Recognition, 2006.*

[54] S. Sarkar and P. Soundararajan, Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(5):504–525, 2000.*

[55] D. Jacobs, Robust and Efficient Detection of Salient Convex Groups. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(1):23-37, 1996.*

[56] M. A. Eshera and K. Fu, An Image Understanding System Using Attributed Symbolic Representation and Inexact Graph-matching. In *Journal of the Association for Computing Machinery, 8(5):604–618, 1986.*

[57] D. Martin and C. Fowlkes et al., A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *International Conference on Computer Vision and Pattern Recognition, 2001.*

[58] P. Tsang and P. Yuen et al., Classification of Partially Occluded Objects Using Three Point matching and Distance Transformation. In *Pattern Recognition, 27(1): 27- 40, 1994*.

[59] S. Chang and F. Hsuan et al., Fast Algorithm for Point Pattern Matching: Invariant to Translations, Rotations and Scale Changes. In *Pattern Recognition, 30(2): 311–320, 1997.*

[60] T. Leung and J. Malik, Detecting, Localizing and Grouping Repeated Scene Elements from An Image. In *European Conference on Computer Vision, 1996.*

[61] Y. E. Sonbaty and M. A. Ismail, Matching Occluded Objects Invariant to Rotations, Translations, Reflections, and Scale Changes. In *Lecture Notes in Computer Science, 2003,*

[62] J. Beis and D. Lowe, Shape Indexing Using Approximate Nearest-neighbour Search in High-dimensional Spaces. In *International Conference on Computer Vision and Pattern Recognition, 1997*

[63] A. N. Stain and M. Hebert, Local Detection of Occlusion Boundaries in Video. In *British Machine Vision Conference, 2006.*

[64] F. Rothganger and S. Lazebnik et al., 3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints. In *International Journal of Computer Vision, 66(3):231–259, 2006.*

[65] Y. Lamdan and H. J. Wolfson, Geometric Hashing: a General and Efficient Model-based Recognition Scheme. In *International Conference on Computer Vision, 1998.*

[66] A. W. Finch and R. C. Wilson et al., Symbolic Graph Matching with the EM algorithm. In *Pattern Recognition, 31(11):1777–1790, 1998.*

[67] T. Gevers and A. Smeulders, Color-based Object Recognition. In *Pattern Recognition, 3(2): 453-464, 1999.*

[68] H. Murase and S. K. Nayar, Visual Learning and Recognition of 3-D Objects from Appearance. In *Internationa*l *Journal of Computer Vision 14(1): 5–24, 1995.*

[69] C. Schmid and R. Mohr, Evaluation of Interest Point Detectors. In *International Journal of Computer Vision, 37(1):151–172, 2000.*

[70] K. Mikolajczyk and C. Schmid, a Performance Evaluation of Local Descriptors. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10):1615–1630, 2005.*

[71] P. Moreels and P. Perona, Evaluation of Features Detectors and Descriptors Based on 3D Objects. In *International Conference on Computer Vision, 2005*

[72] R. Zass and A. Shashua, Probabilistic Graph and Hypergraph Matching. In *International Conference on Computer Vision and Pattern Recognition, 2008.*

[73] M. Carcassoni and E. R. Hancock, Spectral Correspondence for Point Pattern Matching, In *Pattern Recognition, 36(1):193-204, 2003.*

[74] D. J. Lowe, Local Feature View Clustering for 3D Object Recognition. In *International Conference on Computer Vision and Pattern Recognition, 2001.*

[75] M. Galun and E. Sharon et al., Texture Segmentation by Multiscale Aggregation of Filter Responses and Shape Elements. In *International Conference on Computer Vision, 2003.*

[76] G. Guy and G. Medioni, Inferring Global Perceptual Contours from Local Features. In *International Journal of Computer Vision, 13(9):920–935, 1996.*

[77] A. Stein and M. Hebert, Local Detection of Occlusion Boundaries in Video. In *Image and Vision Computing, 27(5):514–522, 2009.*

[78] W. E. L. Grimson, Object Recognition by Computer: The Role of Geometric Constraints. *the MIT Press, Cambridge, 1990.*

[79] T. Tuytelaars and L. Van Gool, Wide Baseline Stereo Matching Based on Locally Affine Invariant Regions. In *British Machine Vision Conference, 2000.*

[80] T. Tuytelaars and L. Van Gool, Matching Widely Separated Views Based on Affine Invariant Regions. In *International Journal on Computer Vision, 59(1):61-85, 2004.*

[81] J. Besag, on the Statistical Analysis of Dirt Pictures. In *Journal of the Royal Statistical Society. Series B (Methodological), 48(3): 259-302, 1986.*

[82] L. Herault and R. Horaud et al., Symbolic Image Matching by Simulated Annealing. In *British Machine Vision Conference, 1990.*

[83] M. Leordeanu, Spectral Matching, Learning, and Inference using Pairwise Interactions. In *PhD theis, CMU, 2009.*

[84] R. Zass and A. Shashua, Probabilistic Graph and Hypergraph Matching. In *International Conference on Computer Vision and Pattern Recognition, 2008.*

[85] J. Malik and S. Belongie et al., Texture, Contours and Regions: Cue Integration in Image Segmentation. In *International Conference on Computer Vision, 1999.*

[86] S. F. B. Duc and J. Bigun. Face Authentication with Gabor Information on Deformable Graphs. In *IEEE Transactions on Image Processing, 1999.*

[87] I. Sethi and N. Ramesh, Local Association Based Recognition of Two Dimensional Objects. In *Machine Vision and Applications 5(2):265-276, 1992.*

[88] R. Desimone and J. Duncan, Neural Mechanisms of Selective Visual-attention. In *Annual Review of Neuroscience, 1(8):193– 222, 1995.*

[89] L. Itti and C. Koch, Computational Modelling of Visual Attention. In *Nature Reviews Neuroscience, 2(3):194–203, 2001.*

[90] C. Koch and S. Ullman, Shifts in Selective Visual-attention-towards the Underlying Neural Circuitry. In *Human Neurobiology 4 (4): 219–227, 1985.*

[91] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *International Conference on Computer Vision and Pattern Recognition, 2001.*

[92] A. Etemadi and J. P. Schmidt et al., Low-level Grouping of Straight Line Segments. In *British Machine Vision Conference, 1991.*

[93] J. Y. Wu and K. B. Lim, a Spectral Technique to Recognize Occluded Objects. In *IET Image Processing, 6(2): 160 – 167, 2012.*

[94] J. Han and K. Ngan, et al., Unsupervised Extraction of Visual Attention Objects in Color Images. In *IEEE Transactions on Circuits and Systems for Video Technology, 16(1):141–145, 2006.*

[95] B. C. Ko and J. Y. Nam. Object-of-Interest Image Segmentation Based on Human Attention and Semantic Region Clustering. In *Journal of Optical Society of America A, 23(10):2462– 2470, 2006.*

[96] B. T. Vincent and R. J. Baddeley et al., Do We Look at Lights? Using Mixture Modeling to Distinguish Between Low- and High-level Factors in Natural Image Viewing. In *Visual Cognition, 17(6), 856–879. 2009.*

[97] X. Hou and L. Zhang, Saliency Detection: A Spectral Residual Approach. In *International Conference on Computer Vision and Pattern Recognition*, 2007

[98] K. E. A Van de Sande and T. Gevers et al., Evaluation of Color Descriptors for Object and Scene Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9): 1582-1596, 2010.*

[99] T. Liu and J. Sun et al., Learning to Detect A Salient Object. In *International Conference on Computer Vision and Pattern Recognition, 2007.*

[100] L. Itti and C. Koch et al., A Model of Saliency Based Visual Attention for Rapid Scene Analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254-1259, 1998.*

[101] W. Grimson and T. Lozano-Perez, Recognition and Localization of Overlapping Parts from Sparse Data in Two and Three Dimensions. In *IEEE Conference on Robotics and Automation, 1985.*

[102] A. Toshev and J. Shi et al., Image Matching via Saliency Region Correspondences. In *International Conference on Computer Vision and Pattern Recognition, 2007.*

[103] K. Fukunaga and L. D. Hostetler, the Estimation of the Gradient of a Density Function. with Applications in Pattern Recognition. In *IEEE Transactions on Information Theory, 21(1):32–40, 1975.*

[104] D. Comaniciu and P. Meer, Mean shift: A Robust Approach toward Feature Space Analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(51):603-619, 2002.*

[105] H. Zabrodsky and S. Peleg, Attentive Transmission. In *Journal of Visual Communication and Image Representation, 1(2):189–198, 1990.*

[106] J. Luo and A. Singhal et al., A Computational Approach to Determination of Main Subject Regions in Photographic Images. In *Journal of Image Vision Computing, 22(3):227-24, 2004.*

[107] D. J. Fleet and Y. Weiss, Optical Flow Estimation. In *Mathematical models for Computer Vision: The Handbook. Springer, 2005.*

[108] D. Feldman and D. Weinshall, Motion Segmentation Using an Occlusion Detector. In *European Conference on Computer Vision, 2006*

[109] A. N. Stain, Occlusion Boundaries: Low-Level Detection to High-Level Reasoning. In *PhD Thesis, CMU, 2008.*

[110] A. Sanfeliu and K. S. Fu, a Distance Measure between Attributed Relational Graphs for Pattern Recognition. In *IEEE Trans. Systems, Man and Cybernetics, 1(3):353- 362, 1983.*

[111] S. Tirthapura and D. Sharvit et al., Indexing Based on Edit-distance Matching of Shape Graphs. In *international Symposium on Voice, Video, and Data Communications, Boston, 1998.*

[112] A. Wong and M. You, Entropy and Distance of Random Graphs with Application to Structural Pattern Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 7(5):566- 609, 1985.*

[113] C. Fowlkes and J. Malik, How Much Does Globalization Help Segmentation. In *Technique Report No. UCB/CSD-4-1340, Berkeley, 2004.*

[114] K. Khoo and P. Suganthan, Evaluation of Genetic Operators and Solution Representations for Shape Recognition by Genetic Algorithms. In *Pattern Recognition Letters, 23(13):1589–1597, 2002.*

[115] C. Pantofaru, Studies in Using Image Segmentation to Improve Object Recognition. In *PhD Thesis, CMU, 2008.*

[116] V. Ferrari and T. Tuytelaars et al., Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views. In *International Journal of Computer Vision 67(2):159– 188, 2006.*

[117] C. Pantofaru and C. Schmid et al., Object Recognition by Integrating Multiple Image Segmentations. In *European Conference on Computer Vision, 2008.*

[118] S. X. Yu and R. Gross et al., Concurrent Object Recognition and Segmentation by Graph Partitioning. In *Advances in Neural Information Processing, 2002.*

[119] R. C. Wilson and E. R. Hancock, Structural Matching by Discrete Relaxation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (6):634-648, 1997.*

[120] R. C. Wilson and E. R. Hancock, Bayesian Compatibility Model for Graph Matching. In *Pattern Recognition Letters, 7(3):263–276, 1996.*

[121] C. Guo and Q. Ma, et al., Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform. In *International Conference on Computer Vision and Pattern Recognition, 2008.*

[122] H. L. Chui and A. Rangaranja, a New Point Matching Algorithm for Non-rigid Registration. In *International Conference on Computer Vision and Pattern Recognition, 2000.*

[123] S. Gold and A. Rangarajan, a Graduated Assignment Algorithm for Graph Matching. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(4):377-388, 1996.*

[124] C. Schellewald and C. Schnorr, Probabilistic Subgraph Matching Based on Convex Relaxation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2005.*

[125] J. Y. Wu and K. B. Lim, Recognition of Occluded Objects by Feature Interactions. In *IEEE International Conference on Automation and Mechatronics, Singapore, 2010.*

[126] J. G. Li and W. X. Wu et al., One Step beyond Histograms: Image Representation using Markov Stationary Features. In *International Conference on Computer Vision and Pattern Recognition, 2008.*

[127] O.Tuzel and F.Porikli et al., Region Covariance: a Fast Descriptor for Detection and Classification. In *European Conference on Computer Vision, 2006.*

[128] S. Ogale and C. Fermller et al., Motion Segmentation Using Occlusions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence,27(6):988-992, 2005*

[129] T. Lindeberg, Feature Detection with Automatic Scale Selection. In *International Journal of Computer Vision, 30(2):79–116, 1998.*

[130] J. Wright and A. Y. Yang et al., Robust Face Recognition via Sparse Representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210-227, 2008.*

[131] M. Yang and L. Zhang, Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In *European Conference on Computer Vision, 2010.*

[132] W. Watler and M. Hayhoe et al., Eye Guidance in Natural Vision: Reinterpreting Salience. In *Journal of Vision, 11(5): 1–23, 2011.*

[133] A. Treisman and G. Gelade, a Feature-Integration Theory of Attention. In *Journal of Cognitive Psychology, 12(1):97–136, 1980.*

[134] M. Everingham and L. Van Gool et al., the PASCAL Visual Object Classes (VOC) Challenge. In *International Journal of Computer Vision, 88(2), 303-338, 2010.*

[135] B. W. Tatler, the Central Fixation Bias in Scene Viewing: Selecting an Optimal Viewing Position Independently of Motor Biases and Image Feature Distributions. In *Journal of Vision, 7(14): 1–17, 2007.*

[136] L. Itti and C. Koch, a Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention. In *Vision Research, 40 (10-12):1489–1506, 2000.*

[137] J. Duncan, Selective Attention and the Organization of Visual Information. In *Journal of Experimental Psychology: General, 113 (4):501–517, 1984.*

[138] P. Roelfsema and V. Lamme et al., Object-based Attention in the Primary Visual Cortex of the Macaque Monkey. In *Journal of Nature, 395 (6700):376–381, 1998.*

[139] J. M. Wolfe, Guided Search 4.0: Current progress with a Model of Visual Search. In *W. Gray (Ed.), Integrated models of cognitive systems (pp. 99–119). New York: Oxford. 2007.*

[140] L. Itti and C. Koch. Feature Combination Strategies for Saliency-based Visual Attention Systems. In *Journal of Electronic Imaging, 10(1):161–169, 2001.*

[141] J. M. Wolfe, What Can 1 Million Trials Tell Us about Visual Search? In *Psychological Science, 9(1): 33–39, 1998.*

[142] C. Rother and V. Kolmogorov et al., Grabcut: Interactive Foreground Extraction using Iterated Graph Cuts. In *ACM Transactions on Graphics, 23(3):309–314, 2004.*

[143] Y. Li and J. Sun et al., Lazy Snapping. In *ACM Transactions on Graphics, 23(3):303-308, 2004.*

[144] T. Ojala and M. Pietikäinen et al., Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7): 971–987, 2002.*

[145] S. Kosinov and E. Bruno et al., Spatial Consistent Partial Matching for Intra and Inter Image Prototype Selection. In *Signal Processing: Image Communication, 23(7):516-524, 2008.*

[146] M. S. Cho and K. M. Lee, Partially Occluded Object-Specific Segmentation in View-Based Recognition. In *International Conference on Computer Vision and Pattern Recognition, 2007.*

[147] M. Leordeanu, Pairwise Grouping Using Color. In *CMU technical report, 2008.*

[148] A. Klaus and M. Sormann et al., Segment-Based Stereo Matching using Belief Propagation and a Self-Adapting Dissimilarity Measure. In *International Conference on Pattern Recognition, 2006.*

[149] B. Luo and E. R. Hancock, Structural Graph Matching using the EM Algorithm and Singular Value Decomposition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (10):1120 – 1136, 2001.*

[150] B. T. Messmer and H. Bunke, a Decision Tree Approach to Graph and Subgraph Isomorphism Detection. In *Pattern Recognition, 32(1):1979–1998, 1999.*

[151] L. Y. Lyul and P. R. Hong, A Surface-based Approach to 3D Object Recognition using a Mean Field Annealing Neural Network. In *Pattern Recognition, 35 (2):299-316, 2002.*

[152] D. Riviere and J. Mangin et al., Auto-matic Recognition of Cortical Sulci of the Human Brain using a Congregation of Neural Networks. In *Medical Image Analysis, 6(2):77–92, 2002.*

[153] P. Suganthan and H. Yan, Recognition of Handprinted Chinese Characters by Constrained Graph Matching. In *Image and Vision Computing, 16(3):191-201, 1998.*

[154] S. Umeyama, an Eigen Decomposition Approach to Weighted Graph Matching Problems. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(5):695-703, 1988.*

[155] G. Scott and H. Higgins, An Algorithm for Associating the Features of 2 Images. In *Proceedings of the Royal Society of London Series B-Biological, 244(1309):21-26, 1991.*

[156] G. Scott and H. Higgins, Feature Grouping by Re-localization of Eigenvectors of the Proximity Matrix. In *British Machine Vision Conference, 1990.*

[157] L. Shapiro and J. Brady, Feature-based Correspondence -an Eigenvector Approach. In *Image and Vision Computing, 10(2):283–288, 1992.*

[158] A. Shokoufandeh and S. Dickinson et al., Indexing Using a Spectral Encoding of Topological Structure. In *International Conference on Computer Vision and Pattern Recognition, 1999.*

[159] M. Leordeanu and M. Hebert et al., An Integer Projected Fixed Point Method for Graph. In *Advances in Neural Information Processing, 2009.*

[160] S. Kumar and M. Hebert, Discriminative Random Fields. In *International Journal of Computer Vision, 68(2):179-202, 2006.*

[161] R. Manduchi and P. Perona et al., Efficient Deformable Filter Banks. In *IEEE Transactions on Signal Processing, 46(4):1168-1173, 1998.*

[162] J. Lafferty and A. McCallum et al., Conditional Random Fields: Probabilistic Models for Segmentation and Labeling Sequence Data. In *International Conference on Machine Learning, 2001.*

[163] C. Liu, Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. In *PhD Thesis, Massachusetts Institute of Technology, 2009.*

[164] Y. Lamdan and H. J. Wolfson, Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. In *International Conference on Computer Vision, 1988.*

[165] Y. Lamdan and H. J.Wolfson, On the Error Analysis of Geometric Hashing. In *Conference on Computer Vision and Pattern Recognition, 1991.*

[166] D. Thompson and J. Mundy, Three-dimensional Model Matching from an Unconstrained Viewpoint. In *International Conference on Robotics and Automation, 1987.*

[167] N. Ayache and O. D. Faugeras, Hyper: A New Approach for the Recognition and Positioning of Two-dimensional Objects. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(1): 44–54, 1986.*

[168] W. E. L. Grimson, Correspondence: On the Recognition of Curved Objects. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(6): 632-643, 1989.*

[169] J. L. Turney and T. N. Mudge et al., Recognizing Partially Occluded Parts. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 7(4):410-421, 1985.*

[170] A. Kalvin and E. Schonberg et al., Two-dimensional, Model Based, Boundary Matching Using Footprints. In *International Journal of Robotics Research, 5(4): 38-55, 1986.*

[171] G. J. Ettinger, Large Hierarchical Object Recognition Using Libraries of Parameterized Model Sub-parts. In *International Conference on Computer Vision and Pattern Recognition, 1988*.

[172] E. Persoon and K. S. Fu, Shape Discrimination Using Fourier Descriptors. In *IEEE Transactions on* System Man Cybern, *7(3):170-179, 1977.*

[173] C. W. Jr. Richard and H. Hamami, Identification of Three Dimensional Objects using Fourier Descriptors of the Boundary Curve. In *IEEE Transactions on System Man Cybern, SMC-4(4), 371-378. 1974.*

[174] F. Etesami and J. Uicker, Automatic Dimensional Inspection of Machine part Cross-Section using Fourier Analysis. In *Computer Vision, Graphics, and Image Processing, 29(2): 216-247, 1985.*

[175] M. K. Hu, Visual Pattern Recognition by Moment Invariants. In *IRE Transaction on Information Theory, 8(2): 179-187, 1962.*

[176] C. H. Teh and R. T. Chin, On Image Analysis by the Method of Moments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(4): 496 - 513, 1988.*

[177] A. Khotanzad and Y. H. Hong, Invariant Image Recognition by Zernike Moment. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(1):489-497, 1990.*

[178] C. Chen, Improved Moment Invariants for Shape Discrimination. In *Pattern Recognition, 26(5): 683-686, 1993.*

[179] D. M. Zhao and J. Chen, Affine Curve moment Invariants for Shape Recognition. In *Pattern Recognition, 30(6): 895-901, 1997.*

[180] N. Ansari and E. J. Delp, Partial Shape Recognition: A Landmark-based Approach. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(5):470-483 1990.*

[181] Y. Lamdan and J. T. Schwartz et al., Affine invariant model-based object recognition. In *IEEE Transactions on Robotics and Automation, 6(5):578-589, 1990.*

[182] J. Zhang and X. Zhang et al., Object Representation and Recognition in Shape Spaces. In *Pattern Recognition, 36(5): 1143-1154, 2003.*

[183] O. D. Faugeras, and M. Hebert, the Representation, Recognition and Locating of 3-D Objects. In *International Journal of Robotics Research, 5(3):27–52, 1986.*

[184] J. Garding and T. Lindeberg, Direct Computation of Shape Cues Using Scale-adapted Spatial Derivative Operators. In *International Journal of Computer Vision, 17(2):163–191, 1996.*

[185] A. Baumberg, Reliable Feature Matching Across Widely Separated Views. In *International Conference on Computer Vision and Pattern Recognition, 2000.*

[186] F. Schaffalitzky and A. Zisserman, Multi-view Matching for Unordered Image Sets, or "How do I organize my holiday snaps?". In *European Conference on Computer Vision, 2002.*

[187] K. Mikolajczyk and C. Schmid, An affine invariant interest point detector. In *European Conference on Computer Vision, 2002.*

[188] W. E. L. Grimson and T. Lozano-Perez, Localizing Overlapping Parts by Searching the Interpretation Tree. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9(4):469–482, 1987.*

[189] D. P. Huttenlocher and S. Ullman, Object recognition using alignment. In *International Conference on Computer Vision, 1987.*

[190] D. G. Lowe, the Viewpoint Consistency Constraint. In *International Journal of Computer Vision, 1(1):57–72, 1987.*

[191] M. Turk and A. Pentland, Eigenfaces for Recognition. In *Journal of Cognitive Neuroscience, 3(1): 71–86, 1991.*

[192] A. Pentland and B. Moghaddam et al., View-Based and Modular Eigenspaces for Face Recognition. In *International Conference on Computer Vision and Pattern Recognition, 1994.*

[193] P. N. Belhumeur and J. P. Hespanha et al., Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711–720, 1997.*

[194] H. Murase and S. K. Nayar, Visual Learning and Recognition of 3-D Objects from Appearance. In *International Journal of Computer Vision, 14(1): 5–24, 1995.*

[195] A. Selinger and R. Nelson, a Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition. In *Computer Vision and Image Understanding, 76(1): 83–92, 1999.*

[196] R. O. Duda and P. E. Hart et al., Pattern Classification. *Wiley-Interscience, Second edition, 2001.*

[197] V. S. Nalwa, Line-Drawing Interpretation: A Mathematical Framework. In *International Journal of Computer Vision, 2(1): 103–124, 1988.*

[198] J. Ponce and D. Chelberg et al., Invariant Properties of Straight Homogeneous Generalized Cylinders and their Contours. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(9): 951–966, 1989.*

[199] J. Liu and J. Mundy et al., Efficient Recognition of Rotationally Symmetric surfaces and Straight Homogeneous Generalized Cylinders. In *International Conference on Computer Vision and Pattern Recognition, 1993.*

[200] J. B. Burns and R. S. Weiss et al., View Variation of Point-Set and Line-Segment Features. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(1): 51–68, 1993.*

[201] J. L. Mundy and A. Zisserman, Geometric Invariance in Computer Vision. *MIT Press, 1992.*

[202] J. L. Mundy and A. Zisserman et al., Applications of Invariance in Computer Vision. In *Second Joint European-U. S. Workshop, Portugal, 1993.*

[203] S. Mahamud and M. Hebert, the Optimal Distance Measure for Object Detection. In *International Conference on Computer Vision and Pattern Recognition, 2003.*

[204] V. Ferrari and T. Tuytelaars et al., Simultaneous Object Recognition and Segmentation by Image Exploration. In *European Conference on Computer Vision, 2004.*

[205] P. Moreels and M. Maire et al., Recognition by Probabilistic Hypothesis Construction. In *European Conference on Computer Vision, 2004.*

[206] M. H. Han and D. Jang, The Use of Maximum Curvature Points for the Recognition of Partially Occluded Objects. In *pattern recognition, 23(1):21-33, 1990.*

[207] S. Chandran and S. K. Kim et al., Parallel Computational Geometry of Circular and Line Segments. In *Image Vision and Computing, 1(4): 71-83, 1996.*

[208] T. Knoll and R. Jain, Recognizing Partially Visible Objects Using Feature Indexed Hypotheses. In *IEEE Journal of Robotics and Automation, 2(1):3-13, 1986.*

[209] P. C. Gaston and T. Lozano-Perez, Tactile Recognition and Localization Using Object Models: The Case of Polyhedra on a Plane, In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(3):257-266, 1984.*

[210] K. Ikeuchi and B. Horn et al., Picking Up an Object from a Pile of Objects. In *Proceedings of the 1st International Symposium on Robotics Research, 1983.*

[211] G. Stockman and J. C. Esteva, 3D Object Pose form Clustering with Multiple Views. In *Pattern Recognition Letters, 3(4):279-286, 1985.*

[212] P. Brou, Using the Gaussian Image to Find the Orientation of Objects. In *International Journal of Robotics Research, 3(4):89–125, 1984.*

[213] E. Borenstein and S. Ullman, Class-specific, Top-down Segmentation. In *European Conference on Computer Vision, 2002.*

[214] B. Leibe and A. Leonardis et al., Combined Object Categorization and Segmentation With an Implicit Shape Model. In *European Conference on Computer Vision workshop on statistical learning in computer vision, 2004.*

[215] A. N. Stein and M. Hebert, Occlusion Boundaries from Motion: Low-Level Detection and Mid-Level Reasoning. In *International Journal on Computer Vision*, *82(2): 325-357, 2009.*

[216] R. Achantay, S. Hemamiz et al., Frequency-tuned Salient Region Detection. In *International Conference on Computer Vision and Pattern Recognition, 2009.*

[217] Y. F. Ma and H. J. Zhang. Contrast-based Image Attention Analysis by Using Fuzzy Growing. In *ACM International Conference on Multimedia, 2003.*

[218] V. Kolmogorov and R. Zabih, What Energy Functions can be Minimized via Graphcuts? In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2 (6): 147–159, 2004*

[219] E. Rathu and J. Kannala et al., Segmenting Salient Objects from Images and Videos. In *European Conference on Computer Vision, 2010.*

[220] L. Torresani and V. Kolmogorov, Feature Correspondence via Graph Matching: Models and Global optimization. In *European Conference on Computer Vision, 2008.*

[221] H. Liu and S. Yan, Common Visual Pattern Discovery via Spatial Coherent Correspondences. In *International Conference on Computer Vision and Pattern Recognition, 2010.*