Intelligent Scene Analysis and Recognition

FINAL PROJECT REPORT

(28-Feb-2008 $\sim$ 30-Mar-2010)

Kai-Kuang Ma
School of Electrical and Electronic Engineering
Nanyang Technological University

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**
4. TITLE AND SUBTITLE **Intelligent Scene Analysis and Recognition**		5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **School of Electrical and Electronic Engineering,Nanyang Technological University,Singapore, ,**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**43**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Abstract

Knowing the name of the location and relative position towards the landmarks not only facilitates the end-user's navigation, but also provides the possibility to offer follow-up geographical services. *Visual Location Recognition and Registration* (VLRR) is addressed in this report, which refers to the problem of predicting the name and relative position of the location only using the captured query image. This problem is almost ill-posed, because on one hand, there is no formal definition of what constitutes a location and it is still not clear which are the location's properties that helps us to perform the recognition. On the other hand, in order to determine the relative position of the end-user, image registration of large viewpoint variation is required, which itself severely suffers from the well-known matching ambiguity.

To solve the first difficulty, *Bag-of-Features* (BoF) model based on *visual codebook* is used, where the codebook is obtained by performing an unsupervised clustering on local image features extracted from the training images. Consequently, each location and query image can be efficiently represented by the corresponding histograms of the appearance of *visual words* in the codebook. Finally a classifier is designed to make the final decision based on the similarity of those histograms via a supervised learning. However, this BoF model lacks of being aware that different visual words actually provide different discrimination power in the sequential location classification. Therefore, a simple and novel weighting scheme, called *Visual Words Aggregation Weighting* (VWAW) is proposed and we assume those visual words which are cluster centers of highly aggregated local image features while with less neighboring words to be more important than others. These two assumptions are reasonable in the sense that highly aggregated cluster center usually has smaller clustering error and the visual word with less neighborhood is more discriminant and robust. By applying this weighting scheme onto the *visual word histogram*, the experimental results indicate considerable improvement compared with the state-of-art weighting algorithms.

In order to address the second difficulty, we propose a novel and effective image registration approach via region pairing. Image pixels are firstly grouped into local homogenous regions using image over-segmentation and point-to-point feature matching is performed to obtain the putative matches. We then express the region pairing model via a *bipartite graph* whose vertices represent the regions having putative matches inside and the edges represent the similarity measurement between them. Finally graph matching is utilized to discover the one-to-one region correspondence and thus provides a helpful feedback for discarding mismatches in the point-to-point putative matches. Moreover the experimental results demonstrate our proposed significantly boosts the percentage of correct matches compared with initial SIFT matching.

Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency under the grant number. The author would like to sincerely thanks for their consistent support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

It is generally acknowledged that there is an accelerated growing rate for the utilization of widely deployed personal mobile device and wireless network. Therefore, the integration of computation intensive algorithm on these mobile devices to achieve information sharing and analysis remains quite promising. Among all, mobile positioning, which provides its user with his or her geographical location and follow-up services, receives enormous attention and is broadly applied in the real-time navigation system designed for mobile device. However the conventional GPS-based positioning system is vulnerable to either environmental disturbance (e.g. bad weather conditions) or multi-path effect caused by satellite signal reflecting off from the surrounding objects, such as buildings, water, trees, etc. Furthermore, such GPS-based approach also requires the mobile device being equipped with the expensive GPS module. An alternative way is to fully exploit the images captured by the inexpensive built-in camera on the mobile device and estimate its location through the reference images tagged by GPS coordinates in the pre-established database. Therefore, in this report, we describe the problem of *Visual Location Recognition and Registration* (VLRR) and present a solution approach for mobile positioning in urban city environment when GPS information is either unavailable or fairly incorrect due to its limitation mentioned above.

One possible application scenario of VLRR can be best illustrated in Figure 1.1. The end-user, that's say a tourist with a camera-mounted mobile phone in the Oxford University, would capture an image of her surrounding scenery and upload it to our VLRR system through wireless network in order to determine her location. After processing, our system would provide the end-user not only the name of location, but also point-to-point matches of captured query image and reference image of the same location in the pre-established database for the end-user to decide the relative position.

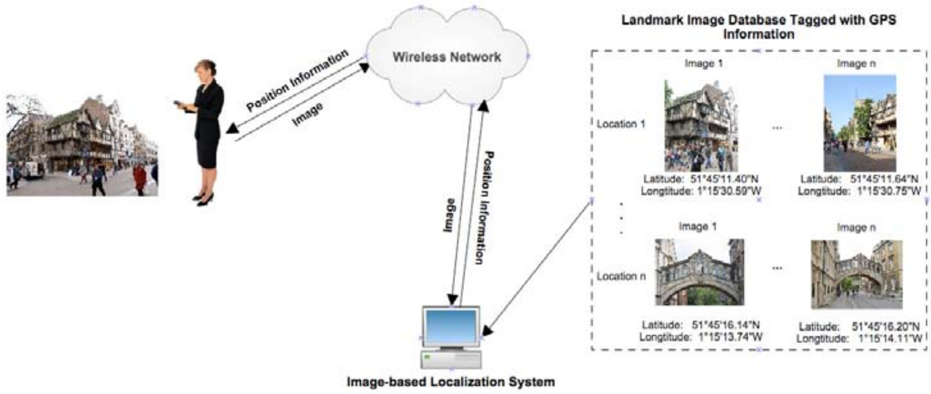


Figure 1.1: An application scenario of image-based localization for mobile positioning

1.2 Contributions and Outline

Our location recognition and registration presented in this report are based on local image features which is expected to be invariant towards geometric transformation (e.g., viewpoint variation) and photometric changes (e.g., different illuminance conditions). Let us first briefly discuss the major issues associated with each component.

When dealing with location recognition, we are usually most concerned with the issue of location representation, since there is no unified definition what constitutes a location and the problem becomes more challenge as the query image can be easily suffered from different illuminance conditions, global scale changes, rotations and affine deformations, to object occlusion with background clutter. In Chapter 2, *Bag-of-Features* model is introduced and used to represent each location as a distribution of local image features in the form of *visual codebook*. Recent research and experiments show that a proper weighting scheme of visual codebook can significantly improve the overall classification result of scene recognition [1] [2] [3]. Therefore, in this report, *Visual Words Aggregation Weighting* is proposed and those visual words which are cluster centers of highly aggregated local image features while with less neighboring words are highly emphasized.

Another theme that runs through this report is image registration or image matching of large viewpoint variation. The main difficulties is how to eliminate the ambiguity of local image features when we try to establish the their correspondences on the interesting points. In Chapter 3, we propose a novel while effective image registration approach through region pairing. Image pixels are firstly grouped into local homogenous regions using image over-segmentation and point-to-point feature matching is performed to obtain the putative matches. We then express the region pairing model via a bipartite graph whose vertices represent the regions having putative matches inside and the edges represent the similarity measurement between them. Finally graph matching is utilized to discover the one-to-one region correspondence and thus provides a helpful feedback for discarding mismatches in the point-to-point putative matches.

Our main contributions of this report can be summarized as follows:

- The proposed *Visual Words Aggregation Weighting* in section 2.3 for *Bag-of-Features* model is the first one to use the distribution of local image features belonging to the same cluster as well as the inter-word relationship to serve as the importance evaluation criteria of the visual word.

- In section 3.3, the proposed image registration via region pairing is a simple, yet effective approach, which is capable of achieving significant higher performance on the image pairs with large viewpoint variation. In particular, our method exceeds state-of-art SIFT matching [4] on the EPFL dataset [5].

The rest of this report is organized as follows.

Chapter 2 first reviews existing computer vision literature on location recognition and *Bag-of-Features* model, then it presents our *Visual Words Aggregation Weighting* scheme for appropriate visual word weights assignment. Chapter 3 describes our image registration via region pairing in details after a comprehensive review of the existing methods for feature mismatch rejection. Finally, Chapter 4 closes the report with a summary of our contributions and discussion of possible extensions and future research directions.

Chapter 2

Location Recognition based on Bag-of-Features Model

2.1 Introduction

In the 1990s, several researchers have considered "semantic" location recognition tasks such as distinguishing city views from landscapes [6] [7] and indoor from outdoor images [8]. Subsequent computational approaches to location description and recognition have often drawn their inspiration from the literature on human perception. For example, it is known that people can recognize locations by considering them in a "holistic" manner, without having to recognize individual objects [9]. Recently, it has been shown that human subjects can perform high-level location recognition tasks extremely rapidly [10] and in the near absense of attention [11]. Renninger and Malik [12] propose an orderless bag-of-textons model to replicate human performance on rapid location recognition tasks. Another perceptually inspired approach, due to Oliva and Torralba [13], computes a low-dimensional representation of a location based on several global properties such as "openness" and "ruggedness". A few recent local recognition approaches [14] [15] try to find effective intermediate representations in terms of basic natural texture categories, such as water, sky, sand, grass, foliage, etc. A major drawback of these methods is that these categories must be learned in a fully supervised fashion, which requires human participation either to hand-segment the training image and label their constituent categories or to provide numerous sample patches of each category. Fei-Fei and Perona [16] present an alternative unsupervised approach that represent locations as mixtures of small number of "themes", or characteristic textures. Meanwhile, several vision researchers have proposed orderless *Bag-of-Features* model [17] [18] [19], which is obtained by extracting scale or affine invariant local features from images, quantizing them into "visual words", and learning the distributions of these words for each locations. Despite its simpleness, this model shows robustness under various imaging conditions and particularly works well for handling the problem of partial occlusion and background clutter, which occur regularly in location recognition task [20] [21] [22].

Several approaches based on *Bag-of-Features* model [1] [23] [24] are based on the observation that each visual word actually carries different discriminative information for location recognition and an appropriate visual word weighting scheme is needed. They take advantage of this insight by directly incorporating the word weighting methods in text-based retrieval and promising experimental results further validate its feasibility. Our proposed *Visual Words Aggregation Weighting* presented in this chapter actually try to quantitatively measure the aggregation of both local image features belonging to the same visual word and the distribution of visual words in the codebook to perform weighting. In Section 2.4, the experimental results proves that this consideration becomes a source of additional discriminative power for location recognition task.

In the following chapter, we will first give a brief overview of the *Bag-of-Features* model used for location recognition. Then more details of the proposed *Visual Words Aggregation Weighting* will be provided.

2.2 Bag-of-Features Model for Location Recognition

As is shown in Figure 2.1, the *Bag-of-Features* model for location recognition actually consists of two stages: off-line training and on-line query. In off-line training, local training image features extracted by some feature detector and descriptor are grouped into specific numbers of clusters to form visual codebook. Then a visual words weighting scheme is carried out to measure the different discriminative power for location recognition embedded in each visual word. Successively, each local feature are assigned to the nearest visual word and therefore a weighted visual word histogram of each location is obtained. In on-line query, same strategy is used to derive the weighted visual word histogram of the query image based on the visual codebook and the weighting coefficients generated in off-line training. Finally, a bin-to-bin histogram comparisons between each location and the query image is performed by computing the cosine similarity [25] [26]. The most similar one would be chosen as the final recognition result and the corresponding location name is output.

2.2.1 Feature Detection

Feature detection which extracts a set of local interested points and salient image regions serves as the first step and also forms the basis of the *Bag-of-Features* model. The detected local features are expected to be invariant to geometric changes(e.g., various viewpoints) and photometric changes(e.g., different illumination conditions) so as to be robust in spite of various scale, illuminations and viewpoint conditions. Numerous research on different feature detectors which emphasize different aspects of invariance are reported in recent years [4]. And the most popular feature detectors which give sufficient performance results were shown in [27]. In this report, *Hessian-Affine* detector [28] is adopted due to its robustness towards moderate viewpoint variation and uniform illumination

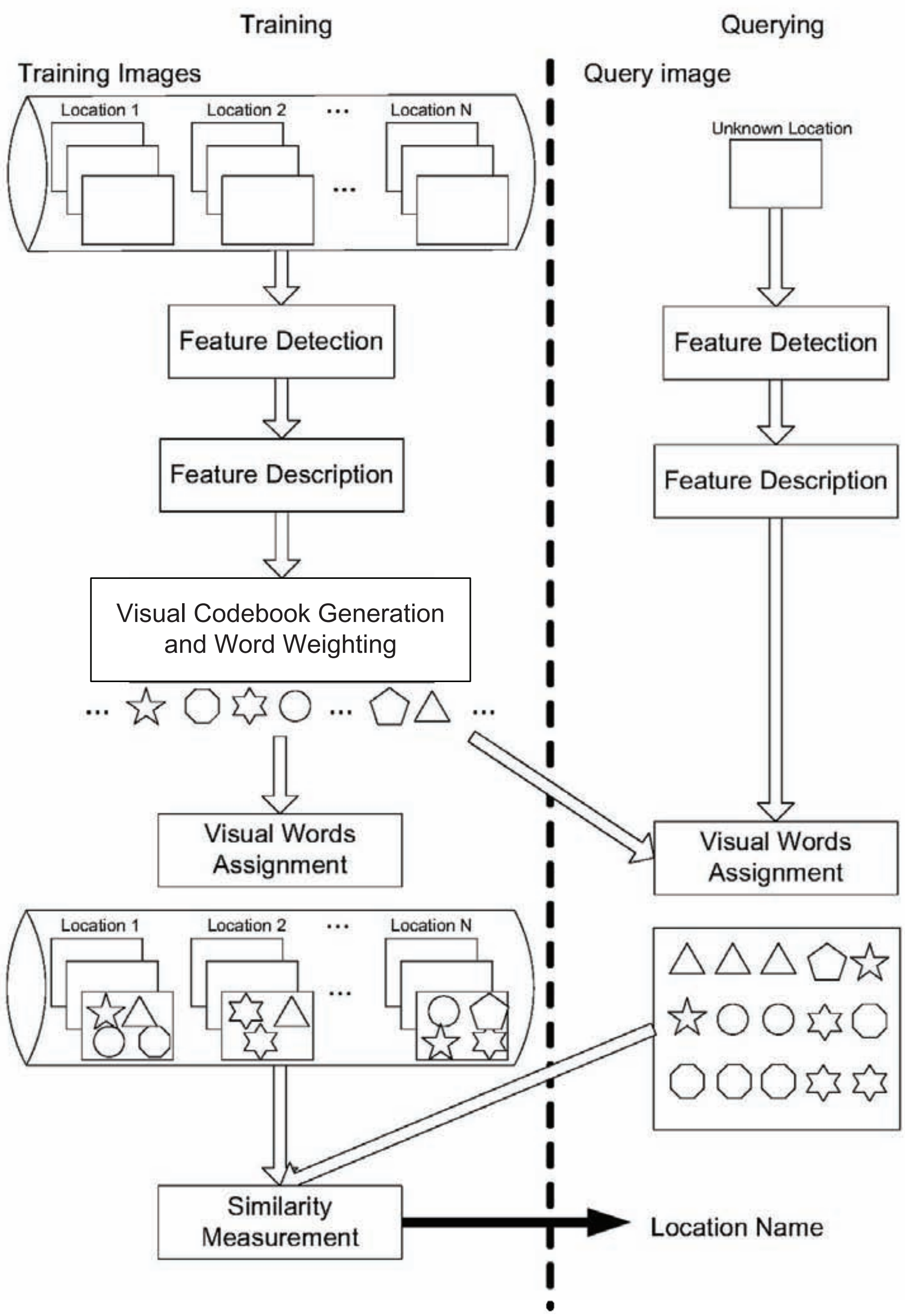


Figure 2.1: The Bag-of-Features model for location recognition

change, which occur quite often in the natural images captured by the mobile device.

The *Hessian-Affine* detector relies a multiple scale iterative algorithm to spatially localize and select scale and affine invariant points. At each individual scale, the *Hessian-Affine* detector chooses interest points based on the Hessian matrix defined in Equation 2.1:

$$H(\mathbf{x};\sigma^2) = \begin{bmatrix} L_{xx}(\mathbf{x};\sigma^2) & L_{xy}(\mathbf{x};\sigma^2) \\ L_{xy}(\mathbf{x};\sigma^2) & L_{yy}(\mathbf{x};\sigma^2) \end{bmatrix} \tag{2.1}$$

where $\mathbf{x} = (x, y)$, $L_{ab}(\mathbf{x})$ is second partial derivative in the a direction and $L_{ab}(\mathbf{x})$ is the mixed partial second derivative in the a and b directions. It is important to note that the derivatives are computed on the image $I(\mathbf{x})$ smoothed by a Gaussian kernel $g(\sigma^2))$ with the variance of σ^2: $L(\mathbf{x}) = g(\sigma^2) * I(\mathbf{x})$. The selection criterion for Hessian interest points is to find local extrema of both the determinant and the trace of the Hessian matrix shown in Equation 2.2:

$$\begin{aligned} DET(H(\mathbf{x})) &= \sigma^2(L_{xx}(\mathbf{x};\sigma^2)L_{yy}(\mathbf{x};\sigma^2) - L_{xy}^2(\mathbf{x};\sigma^2)) \\ TR(H(\mathbf{x})) &= \sigma(L_{xx}(\mathbf{x};\sigma^2) + L_{yy}(\mathbf{x};\sigma^2)) \end{aligned} \tag{2.2}$$

Furthermore *Hessian-Affine* detector uses an iterative shape adaptive algorithm to estimate the local affine transform for each interest point in order to obtain invariance against affine transformed images. As is shown in Figure 2.2, in spite of the affine deformation caused by two different viewpoints, the regions detected by the *Hessian-Affine* detector clearly correspond.

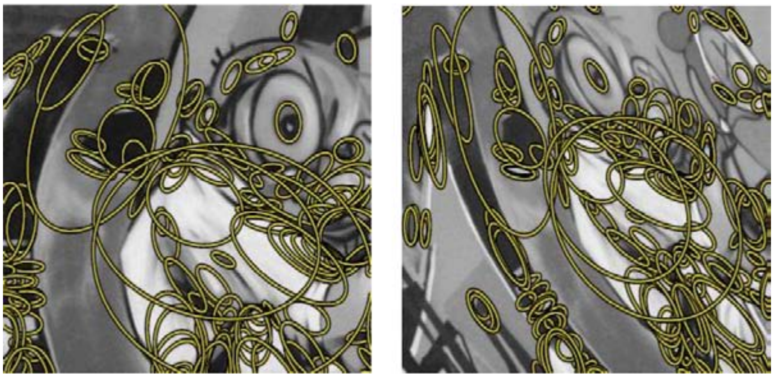


Figure 2.2: Local features detected by the *Hessian-Affine* detector from two different view

2.2.2 Feature Description

Once the local features in image such as interest points are detected, feature description which servers as the second step will be applied to depict the local neighborhood region of the salient points or patches as high-dimensional descriptor vectors. Such descriptors should be distinctive while keep invariant and robust to various image transformation such as affine distortions, scale changes, illumination changes or occlusions. Common feature descriptors include *differential invariants* [29], *steerable filters* [30], *moment invariants* [31], *complex filter* [32], *shape context* [33], *spin images* [34] and the recently developed *Scale Invariant Feature Transform*(SIFT) [4]. According to a fairly comprehensive review on these feature descriptors [35], SIFT descriptor outperform all the others as it is shown to be more robust to various image transformation like rotation, scale changes, affine transformation and illumination changes. The robustness and the distinctive character of the SIFT descriptor make it a good choice to be used in this report. And detailed mathematical investigation of this algorithm is provided as follows:

Developed by David Lowe [4], SIFT descriptor which belongs to a kind of distribution-based methods that represent certain region properties by multidimensional histograms, is a 3D histogram of gradient location and orientation, where location is quantized into a 4×4 location grid and the gradient angle is quantized into eight orientations. Thus, the resulting descriptor is of dimension $4 \times 4 \times 8$=128. For details, there are two steps for computing SIFT descriptor: Orientation assignment for interested points and 128 dimensional descriptor vectors generations. In the first step, the gradient magnitudes m and local orientations ϕ at the interest point $I(x, y)$ should be firstly calculated as:

$$\begin{aligned} m &= \sqrt{(I(x+1,y)-I(x-1,y))^2+(I(x,y+1)-I(x,y-1))^2} \\ \phi &= \tan^{-1}((I(x,y+1)+I(x,y-1))/(I(x+1,y)-I(x-1,y))) \end{aligned} \tag{2.3}$$

Both the magnitude and orientation calculations for the gradient are weighted by the Gaussian-smoothed image $L(x, y, \sigma)$ at the point's estimated scale σ so that all computations are performed in a scale-invariant manner. Then an orientation histogram with 36 bins is formed by adding each sample in a neighboring Gaussian-weighted circular window, with each bin covering 10 degrees. The peaks in this histogram correspond to dominant orientations. Once the histogram is filled, the orientations corresponding to the highest peak and local peaks that are within 80% of the highest peaks are assigned to the interest point. In the case of multiple orientations being assigned, an additional interest point is created having the same location and scale as the original interest point for each additional orientation. This first step is to assign orientations to points at particular scales, which ensured invariance to image location, scale and rotation. After that, in the second step, the descriptor vectors is computed as a set of orientation histograms on (4×4) pixel neighborhoods. These Histograms contain 8 bins each, and each descriptor contains a 4×4 array of 16 histograms around the interest point. This leads to a SIFT feature vector with $4 \times 4 \times 8 = 128$ elements. This vector is finally normalized to enhance invariance to changes in illumination. The computation process of SIFT descriptor is illustrated in Figure 2.3.

2.2.3 Visual Codebook Generation and Word Weighting

The third step, clustering the feature descriptors in their feature space into a specific number of clusters then generating appropriate visual word weights is considered as an important step

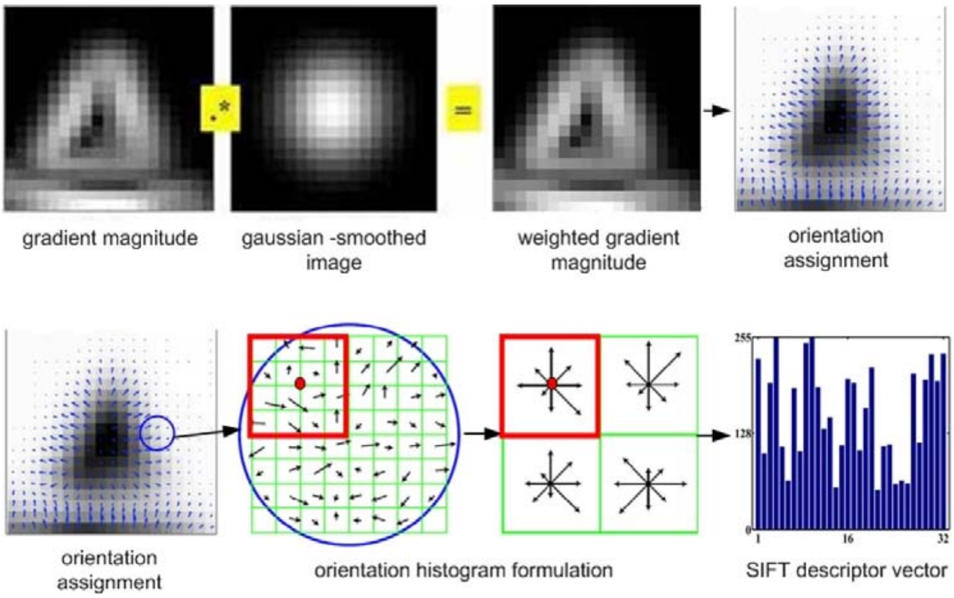


Figure 2.3: Illustration of the SIFT descriptor. After orientation assignment for interested points, a histogram of gradient orientation is formed in each grid cell. 2×2 grid is depicted for a better facility of illustration. 8 orientation bins are used in each grid cell therefore giving a descriptor of dimension 32 =(2×2×8).

since it would have a significant impact on the final performance of the whole location recognition. Different from the way that the codebook is obtained in case of text, which is the natural results from the words occurring in the document collections, the construction of visual codebook involves two critical issues: how to choose an efficient clustering algorithms especially for a large quantity of feature descriptors and the problem of setting up an optimal codebook size.

For the first issue, it is well known that generating clusters for large quantity of feature descriptors presents challenges to many conventional clustering algorithms [23]. The size of data essentially rules out the computation demanding methods such as *mean-shift* [36] and *spectral clustering* [37]. Although K-Means is the most popular clustering algorithm due to its simplicity for implementation [16] [19], the traditional "flat" K-Means clustering still has difficulty to scale to large feature descriptors which results in high computational complexity. In order to solve this problem, Hierarchial K-Means proposed by Nister and Stewenius [20] is utilized in this report. Unlike the "flat" K-Means defining K to be the final number of clusters, in Hierarchial K-Means, it is defined as the number of group partitioned in each iteration. First, an initial K-Means clustering runs on the feature descriptors with the output of K cluster centers. Then the feature descriptors are partitioned into K groups, where each group consists of those being closest to a particular cluster center. The same process is then applied to each group of the feature descriptors up to L times, recursively defining clusters by splitting each group into K new groups and resulting in K^L clusters. Figure 2.4 shows the illustration of the Hierarchical K-Means.

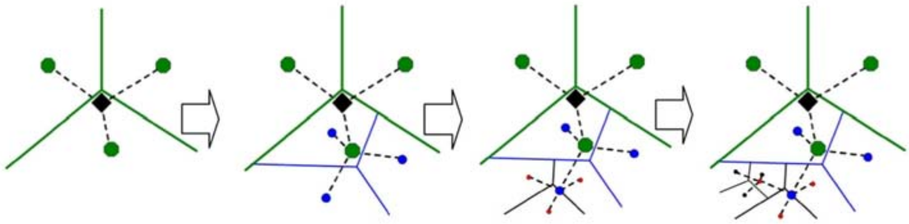


Figure 2.4: Illustration of Hierarchical K-Means. The hierarchical quantization is defined at each level by K clusters (in this case K=3)

For the second issue, since the visual codebook is built using cluster centers obtained from the clustering algorithm, the optimal number of clusters cannot be determined automatically. So setting a suitable size of codebook is of critical importance, involving the trade-off between discriminativity and generalizability. With a small codebook, the visual words is not very discriminative because dissimilar feature descriptors can be mapped to the same visual words. One the other hand, as the codebook size increases, the visual words becomes more discriminative, but meanwhile less generalizable and forgiving to noise, since similar feature descriptors may be mapped to different visual words. However, although some recent studies tend to show that larger dictionary size generally improve the quality of *Bag-of-Features* model, using a large codebook size also increase the cost associated with clustering feature descriptors, computing visual word weights, assigning visual words and running supervised classifiers. In this report, we use the codebook size 1000-5000 for the Paris database.

The problem of word weighting comes from the interesting observation that each visual word actually carries different discriminative information for location recognition therefore cannot be treated equally. Well-designed visual word weighting scheme can significantly improve the overall performance of location recognition and also alleviate the influence of visual codebook size on the recognition performance. In Section 2.3, detailed explanation of the proposed *Visual Word Aggregation Weighting* shall be given.

2.2.4 Visual Words Assignment

The final step is to assign visual word selected from the visual codebook to each feature descriptor and represent each location and query image by the weighted visual word histogram. For such visual word assignment, a *nearest neighbor* vector quantization [16] [23] [19] is used in this report. Based on this scheme, two feature descriptors are considered identical if they are assigned to the same visual word(cluster center). One the contrary, two feature descriptors assigned to different(even very close) visual words are considered totally different. After the vector quantization, a weighted visual word histogram of each location and query image can be derived by multiplying the frequency counts of the visual words with their corresponding weights obtained in the pervious step.

2.3 Proposed Weighting Scheme for Visual Codebook

The analogy between visual words in *Bag-of-Features* model and text words in *Bag-of-Words* model [38] provides opportunities for migrating text-based retrieval techniques to solve problems in image data. Consequently, as is shown in table 2.1 , some word weighting schemes are also employed to improve the recognition performance in image domain. Up to our knowledge, *term frequency-inverse document frequency* (TF-IDF) weighting is the most popular scheme applied to image recognition based on *Bag-of-Features* model [23], [24]. In TF-IDF, the weight w_{ij} of the visual word v_i in class d_j consists of two parts: *term frequency* (TF) and *inverse document frequency* (IDF). On one hand, TF is the number of occurrences of visual word v_i in class d_j, assuming visual words with high frequency of occurrences in the same class should be regarded as important. on the other hand, IDF measures the ratio of classes having visual word v_i to the total number of classes, indicating the more classes visual word v_i belong to, the less discriminative it would be in the following classification, therefore, v_i should be regarded less important than the other visual words. TF-IDF weighting scheme is actually quite similar towards the basic principle of OTSU threshold and we could further summarize its basic idea into the following two assumptions:

- Visual word with higher frequency of within-class occurrences should be assigned a larger weight.
- Visual word with less frequency of between-class occurrences should also be assigned a larger weight.

Table 2.1: Weighting schemes migrated from text-based retrieval(v_i: the ith visual word; d_j: the jth class; w_{ij}: weight for visual words t_i in class d_j; tf_{ij}: the number of occurrences of visual word v_i in class d_j; N: total number of classes; n_i: the number of classes having visual word v_i; TF: term frequency; IDF: inverse document frequency.)

Weighting Scheme	w_{ij}
Binary [39]	1 if v_i is in d_j, 0 otherwise
TF [40]	tf_{ij}
TF-IDF [24] [23]	$tf_{ij} \cdot log\frac{N}{n_i}$

However, unlike the text words, visual words generated by *Bag-of-Features* model lack of sematic meanings and are also prone to be inaccurate, this fundamental difference significantly prevents the performance grain obtained by foregoing mentioned weighting schemes directly migrated from text retrieval. For example, in Figure 2.5, four query images of three different locations are represented by *Bag-of-Features* model using four-word visual codebook for illustration purpose. By applying those weighting methods directly from text retrieval, such as TF-IDF, the similarity of four histograms are equally small. However, it is obviously that image 3 and image 4 are more alike since it is the same car captured under different viewpoint and environment. In other words, the weighting scheme for the visual words should reflect such differences (i.e., compared with histogram 1 and 2, histogram 3 and 4 should get a much higher similarity score).

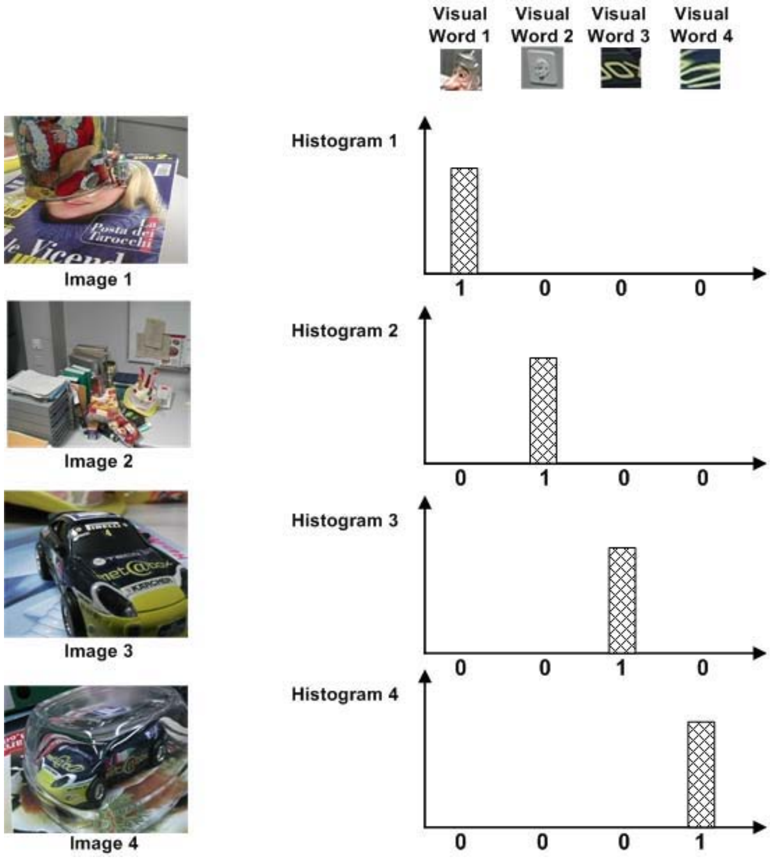


Figure 2.5: Failure of the weighting method directly from texture retrieval when the visual words is over-discriminative

In order to address the problem illustrated in Figure 2.5, we try to combine the TF-IDF weighting with the aggregation measurement of both local image features around a specific visual word as well as its neighboring visual words in the codebook. Such combination not only reflects the visual words' within-class and between-class importance, but also capture the inter-words relationship, which is essential to compensate the side-effect caused by inappropriate codebook generation. Inspired by this, a simple while novel *Visual Word Aggregation Weighting* scheme is proposed in this report.

Local Image Feature Aggregation Measurement

Given a visual codebook $V = \{v_1, v_2, \cdots, v_n\}$ with the size of n, the local image features f of training images can be partitioned into $O = \{o_1, o_2, \cdots, o_n\}$, where for each o_i, it contains k_i local features which should be assigned to the visual word v_i. Then we define the average quantization error for visual word v_i as follows:

$$e(v_i) = \frac{1}{k_i} \sum_{f_j \in o_i} \|f_j - v_i\|_2 \tag{2.4}$$

From Equation 2.4 we can see the average quantization error $e(v_i)$ actually serves as a perfect aggregation measurement of local image features indicating how compactly they are distributed around the corresponding visual word. Smaller this value is, more compact the distribution would be. Consequently this visual words should be regarded as a better cluster center and be more discriminative and robust in the following classification. Therefore we define local image feature aggregation weighting as follows:

$$w_{local}(v_i) = log(\frac{1}{e(v_i)} + 1) \tag{2.5}$$

Inter-words Aggregation Measurement

For a given visual word v_i, we define the inter-words aggregation measurement in the form of average distance between the visual word v_i and its τ-nearest neighborhood $N(v_i, \tau)$:

$$c(v_i) = \frac{1}{\tau} \sum_{v_j \in N(v_i,\tau)} \|v_i - v_j\|_2 \tag{2.6}$$

From the Equation 2.6, it is clear that the larger $c(v_i)$ is, the less neighboring visual words appear around v_i. As a matter of fact, this situation is more preferred by the classifier as illustrated in Figure 2.6. Among the four visual words, visual word 3 and 4, which are closely located in the feature space, are most unreliable ones, since they actually depict the same part of a car label just under different viewing condition. In other word, these two visual words are over-discriminative and should be merged into one. Such over-discriminative will lead to a high probability of incorrect visual word assignment during vector quantization process of *Bag-of-Features* model and result in a worse recognition performance. Therefore, a lower importance weight should be given towards them. We then define the inter-words aggregation weighting as follows:

$$w_{inter}(v_i) = log(c(v_i) + 1) \tag{2.7}$$

Combine Inter-word Relationship into TF-IDF

Finally we aim to achieve a combined weighting scheme which not only capture the inter-word relationship, but also emphasize the visual words' within-class and between-class importance. Let d_j represents the jth class and the final weighting value w_{ij} is given by:

$$w_{ij} = w_{local}(v_i) \cdot w_{inter}(v_i) \cdot (tf_{ij} \cdot log\frac{N}{n_i}) \tag{2.8}$$

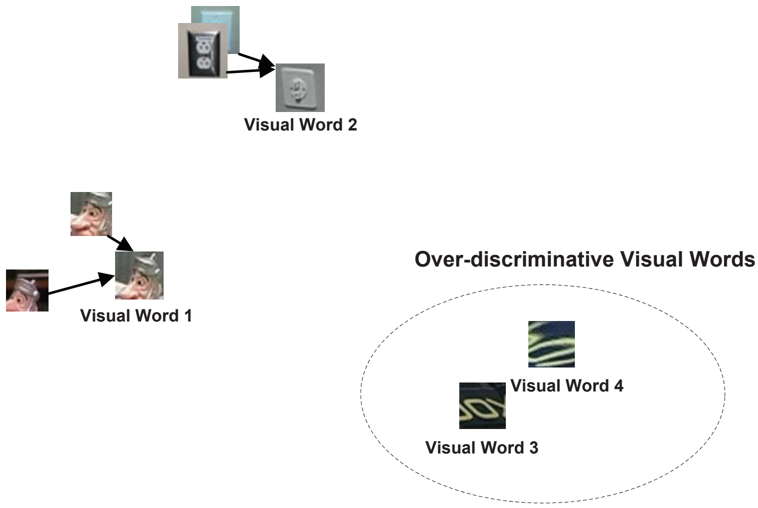


Figure 2.6: The illustration of over-discriminative visual words

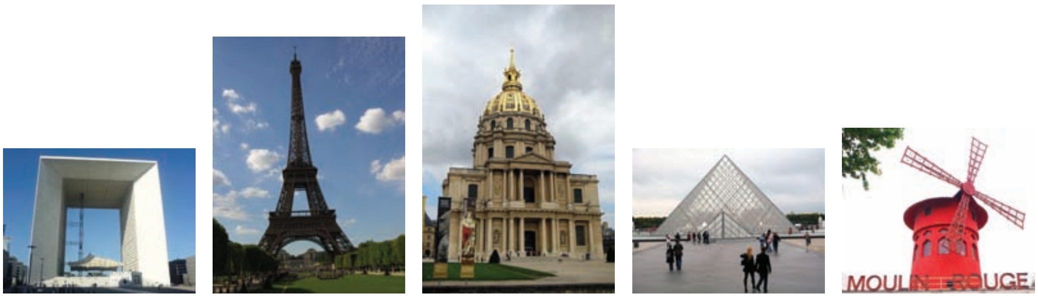
where tf_{ij} is the number of occurrence of visual word v_i in class d_j, N is the total number of classes and n_i is the number of class having visual word v_i. In this weighting definition, $w_{local}(v_i)$ up-weights the visual words which is a good cluster center form by aggregated local features, while $w_{inter}(v_i)$ degrades the visual words which might be over-discriminative in the visual codebook. Meanwhile tf_{ij} emphasize visual words occurring frequently in a particular class and $log\frac{N}{n_i}$ down-weights visual words appearing frequently between different classes.

2.4 Experimental Results

To verify the performance of our proposed *Visual Words Aggregation Weighting* scheme, experiments on Paris database for location recognition are conducted and detailed implementations and results are listed in this section.

2.4.1 Database

The original Pairs database consists of 6412 images collected from *Flickr* by searching 12 particular locations of Pairs, where the average size of each image is approximately 768 $\times$ 1024 pixels [1]. Among them, we select 1080 images belonging to 9 different locations, as is shown in Figure 2.7. And for each location we assign 120 images, from which 20 images are randomly chosen as the training images and the rest images are all considered as queries.



(a) (b) (c) (d) (e)

(f) (g) (h) (i)

Figure 2.7: Pairs database from [1] (a) La Defense, (b) Eiffel Tower, (c) Hotel des Invalides, (d) Louvre, (e) Moulin Rouge, (f) Arc de Triomphe, (g) Notre Dame, (h) Pantheon, (i) Sacre Coeur

2.4.2 Results and Discussions

We compare the proposed *Visual Words Aggregation Weighting* scheme with the state-of-art TF-IDF weighting method [24] [23]on Pairs database. Figure 2.8 demonstrates the variation of average recognition rate of 9 locations towards different visual codebook sizes ranging from 1000 to 5000 at the interval of 500 and the experimental data are all summarized in Table 2.2. It is obvious that our proposed *Visual Words Aggregation Weighting* which takes the compactness of visual words' distribution into consideration yields superior performance to the popular TF-IDF weighting method. Meanwhile, another interesting observation from Figure 2.8 is that compared with TF-IDF, the average recognition rate of proposed weighting scheme is less fluctuate towards the change of visual codebook size. It is generally acknowledged that there is still no consensus how to set the appropriate size of visual codebook [24] [40] [41] of *Bag-of-Features* model. However by adopting our proposed weighting scheme, this issue can be alleviated or even sidestepped to some extent. Figure 2.9 shows a recognition rate comparisons between our proposed weighting scheme and TF-IDF weighting on each of the 9 locations in the Paris database when the visual codebook size is set to be 4000. It is apparently that our proposed weighting scheme significant improves the recognition rate for 8 out of total 9 locations and slightly drops for only one location. In summary, our proposed *Visual Words Aggregation Weighting* scheme provides a superior performance compared with the current state-of-arts TF-IDF method.

Table 2.2: The average recognition rate comparisons of two weighting schemes: Proposed Visual Words Aggregation Weighting and TF-IDF on Pairs database

Visual Codebook Size	Proposed	TF-IDF
1000	67.2%	66.9%
1500	69.2%	67.3%
2000	69.7%	68.7%
2500	70.1%	69.2%
3000	70.5%	69.8%
3500	71.5%	69.9%
4000	72.0%	70.4%
4500	71.7%	69.1%
5000	71.2%	68.3%
Mean Average Recognition Rate	**70.34%**	**68.84%**

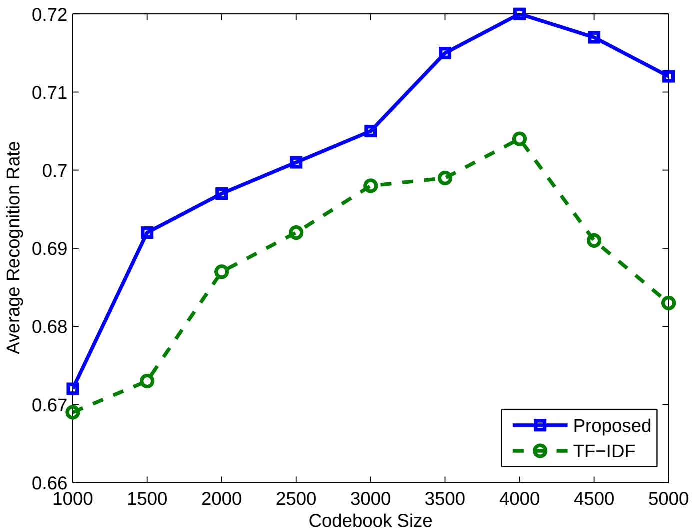


Figure 2.8: The average recognition rate curve comparisons of two weighting schemes: proposed Visual Words Aggregation Weighting and TF-IDF on Pairs database

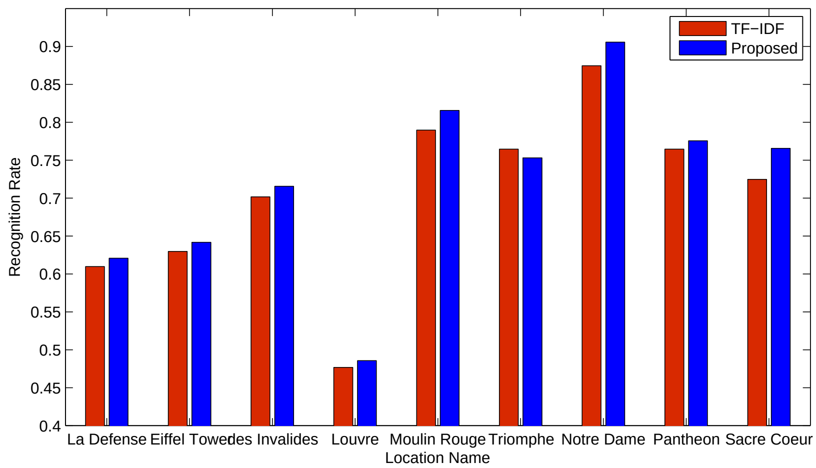


Figure 2.9: The average recognition rate comparisons of two weighting schemes: proposed Visual Words Aggregation Weighting and TF-IDF on each location of Pairs database, the size of codebook is set to be 4000

Chapter 3

Image Registration via Region Pairing

3.1 Introduction

Image registration of large viewpoint variation, or so called wide-baseline image matching, is one of fundamental challenges in computer vision and is experiencing a growing interest for the task of object recognition [4] [42], structure and motion estimation [43] [32], and so forth. The predominant framework is to establish the correspondence of interest points by means of extracting distinctive local image features, which are expected to be invariant towards various geometric and photometric changes. However, these feature-based approaches usually suffer from the ambiguity of local image features. Therefore, the search for similar image feature between the reference image and target image typically leads to a high percentage of mismatches in putative correspondence. As a result, how to discard those mismatches becomes one of the central problems to be solved.

One typical way to address this problem is to assume that all image features in the reference image undergo a rigid transformation [44] [45]. However, this method is unapplicable when non-rigid deformation exists. A more flexible scheme is presented by Ferrari *etal* [42], where the author propose an algorithm consisting of two operations: expansion and contraction of correspondence points and mismatches are slowly removed. The main problem with this method is high computational complexity of the whole algorithm. Recently, information about the geometric and spatial layout of image features have also been considered for proposing robust image matching algorithms [46] [47] [48] [49]. These methods formulate image matching as a graph matching problem by defining a cost function which evaluate both the appearance similarity and the pairwise geometric compatibility between image features. Although all these methods demonstrate a promising results under the assumption of low percentage of mismatches in putative correspondence, in the real case of wide-baseline matching, this percentage is often above 50% due to large viewpoint variation, significant clutter and occlusion occurring between the image pair.

In this work, we try to eliminate this matching ambiguity by adding a spatial support from image segmentation, which aims to find the corresponding region pairs that are both coherent within the images and similar across images. We refer these region pairs as *Common Saliency Region Pair*(CSRP) and defines them as follows:

1. Each region in the pair should be coherent in both photometry and geometry.

2. Putative matches falling within CSRP are regarded as correct point-to-point matches.

3. One-to-one constraint of CSRP: a common saliency region cannot be matched to two common saliency regions which are on the same image.

4. Compared with the mismatched CSRP, the corrected matched CSRP always has a higher similarity measurement.

We intend to determine CSRP from the support of putative matches within these regions by formulating this problem into bipartite graph matching. The main advantage of our formulation is to combine the inter image similarity with the intra image similarity in a systematic way to impose the constrain of feature spatial configuration. Thus, global common saliency region correspondence and local point-to-point feature correspondence would have a mutual verification to maximal reduce the mismatches in putative correspondence.

3.2 Problem Formulation

We formulate wide-baseline image matching via region pairing as follows. The local image feature for reference image I_r and target image I_t is described as f_r and $\widetilde{f}_t$ respectively. Then putative match set P_{rt} representing a initial correspondence between the reference image feature set O_r and target image feature set O_t could be denoted by

$$P_{rt} = \{(f_r, \widetilde{f}_t) | \widetilde{f}_t \in O_t, f_r \in \gamma(\widetilde{f}_t, O_r), s_f(f_r, \widetilde{f}_t) > \tau_s\} \tag{3.1}$$

where $\gamma(\cdot)$ is the nearest neighbor matching between the target image feature $\widetilde{f}_t \in O_t$ and the reference image feature set O_r. Function $s_f(\cdot)$ represents the ratio test [50] of two features and τ_s is the predefined ratio threshold.

Furthermore, over-segmentation [51] is performed on both reference image and target image to obtain two homogenous region sets R_r and R_t, our goal is to determine CSRP from these two region sets by inspecting the mutually coherent putative feature correspondence and finally reject the mismatches in P_{rt}.

3.3 Proposed Graph Model for Region Pairing

We use bipartite graph $G(V, E)$ to model our region pairing problem to determine CSRP. A bipartite graph is a graph whose vertices can be divided into two disjoint sets S and T such that every edge connects a vertex in S to one in T. Therefore, in the following section, we would first introduce how we define and group the vertices of the graph. Then detailed explanation shall be given on how we establish the edges and assign their associated weights.

3.3.1 Vertices of Bipartite Graph

In order to well define the vertices of bipartite graph, we shall first introduce the following three definitions:

Definition 1: If there is a point-to-point putative feature match $(f_r, \widetilde{f_t})$ falling within the region r_i and $\widetilde{r_j}$ in the reference image I_r and target image I_t individually, the region pair $(r_i, \widetilde{r_j})$ is called a *putative region match*. And region $\widetilde{r_j}$ is called the *putative correspondence region* of region r_i.

As is shown in Figure 3.1, a putative region match is actually a spatial expansion of putative feature match from single pixel towards a homogenous region.

Figure 3.1: An example of putative region match. The yellow cross indicates the point-to-point feature putative match. The red and blue regions are the resulting putative region match.

Definition 2: For a given region in either reference image I_r or target image I_t, if there exist putative regions matches, the connected putative correspondence regions are called *unreliable segments*. Otherwise they are called *reliable segments*.

The intrinsic idea behind the reliable and unreliable segments is to try to overcome the disadvantage of image over-segmentation that successive images of the same scene do not always yield to the same image segments due to illuminance, viewpoint or scale variations. Typically, a region in the reference image may always be segmented into two or more smaller segments in the target image and vice versa. This phenomenon can be favourably detected and compensated by unreliable segments, indicating mutual connected putative correspondence regions should be representing a same region and thus need to be merged together to form a new reliable segments.

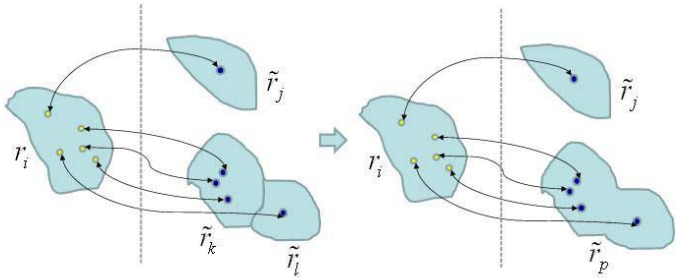


Figure 3.2: An illustration of merge operation on unreliable segments.

In Figure 3.2, a simple illustration of above-mentioned merge operation on unreliable segments is given. For the given region r_i in the reference image, region $\widetilde{r_j}$, $\widetilde{r_k}$ and $\widetilde{r_l}$ are the putative correspondence regions in the target image. Furthermore, regions $\widetilde{r_k}$ and $\widetilde{r_l}$ are connected, therefore we refer them as unreliable segments and a merge operation is performed to obtain a new reliable segment $\widetilde{r_p}$ beside the original one $\widetilde{r_j}$. Meanwhile, after this merge operation, all the regions engaged in putative region matches are reliable segments. As a matter of fact, for the given putative region matches of the reference image and target image, we could always find unique *reliable putative region matches*, whose definition is given below, by a series of merge operation indicated in Algorithm 1.

Definition 3: If all the regions in the putative region matches of the reference image I_r and target image I_t are reliable segments, these putative region matches are called *reliable putative region matches*.

With the help of the above three definitions, we could further define the vertices of our bipartite graph as follows:

Vertices of Bipartite Graph: The vertices of bipartite graph V are composed by two disjoint sets S and T, which are the regions of the reliable putative region matches lying in the reference image I_r and target image I_t respectively.

3.3.2 Edge and Weights of Bipartite Graph

Since the vertices of bipartite graph all come from the reliable putative region matches, it is very natural to assign an edge for each match $(r_i, \widetilde{r_j})$. The main problem is how we evaluate the consistency between these two regions and assign an appropriate weight towards the edge.

Photometric Consistency: This value encodes the photometric consistency between two regions. Let us assume for each reliable putative region match $(r_i, \widetilde{r_j})$, there are L putative point-to-point feature matches falling inside. In this work, we use affine covariant region detector [52] [53] as feature detector and scale invariant feature transform (SIFT) [4] as feature descriptor. Thus, we could represent the kth putative feature match as $C_k(f_i, \widetilde{f_j}) = (x_i, \widetilde{x_j}, H_{ij})$, where x_i, $\widetilde{x_j}$ and H_{ij} denotes the coordinate of feature f_i in the reference image, coordinate of feature $\widetilde{f_j}$ in the target image and homography to transform the covariant region p_i around feature f_i to covariant region

Algorithm 1 Algorithm to Obtain Reliable Putative Region Matches by Merge Operation

INPUT: The putative region matches $M = (S, T)$, where the region set in reference image I_r is $S = \{r_1, r_2, \cdots, r_n\}$ and region set in target image I_t is $T = \{\widetilde{r_1}, \widetilde{r_2}, \cdots, \widetilde{r_n}\}$

OUTPUT: The reliable putative region matches R

for $i = 0$ to n **do**

For the region r_i in the reference image I_r, find its putative correspondence regions $PR(r_i)$ in the target image I_t.

if region $\widetilde{r_k}, \cdots, \widetilde{r_l}$ in $PR(r_i)$ is connected **then**

Merge region $\widetilde{r_k}, \cdots, \widetilde{r_l}$ together as a new region $\widetilde{r_m}$ and replace the corresponding putative region match in M and T.

end if

end for

for $i = 0$ to length of set T **do**

For the region $\widetilde{r_i}$ in the target image I_t, find its putative correspondence regions $PR(\widetilde{r_i})$ in the reference image I_r.

if region $r_k, \cdots, r_l$ in $PR(\widetilde{r_i})$ is connected **then**

Merge region $r_k, \cdots, r_l$ together as a new region r_m and replace the corresponding putative region match in M and S.

end if

end for

Let $R = M$

$\widetilde{p_j}$ around feature $\widetilde{f_j}$, respectively.

Then for each putative feature match $C_k(f_i, \widetilde{f_j})$, let us define the similarity of its two covariant regions by

$$s_{photo}(C_k(f_i, \widetilde{f_j})) = 1 + NCC(p_i, \widetilde{p_j}) - \frac{dRGB(p_i, \widetilde{p_j})}{100} \tag{3.2}$$

where NCC is the normalized cross-correlation between the gray level image region, while dRGB is the average pixel-wise Euclidean distance in RGB color-space after independent normalization of the 3 colorbands for photometric invariance [42]. Before computation, region p_i and $\widetilde{p_j}$ should be normalized to unit circles and rotated to the same dominant orientation using homography mapping H_{ij}. Furthermore, we could define the overall photometric similarity of putative region match $(r_i, \widetilde{r_j})$ as

$$sim_{photo}(r_i, \widetilde{r_j}) = \frac{1}{L}\sum_{k=1}^{L} s_{photo}(C_k(f_i, \widetilde{f_j}))^2 \tag{3.3}$$

Region Context Consistency: This value verifies the region context similarity of the given putative region match. And Figure 3.3 demonstrates a simple example why photometric consistency alone cannot produce reliable region match. As is shown in Figure 3.3(a) and Figure 3.3(b), the smaller yellow ellipses indicate the covariance regions $p_i, \widetilde{p_j}$ of putative feature match $(f_i, \widetilde{f_j})$. It is apparent that this is a false match, resulting in a false putative region match easily. However, if we normalize these two regions using homography H_{ij} and present the results in Figure 3.3(c) and Figure 3.3(d). We would surprisingly discover that these two image patches are quite similar, in other words, the photometric similarity measurement of Equation 3.2 would be very high, thus leads to a contrary conclusion.

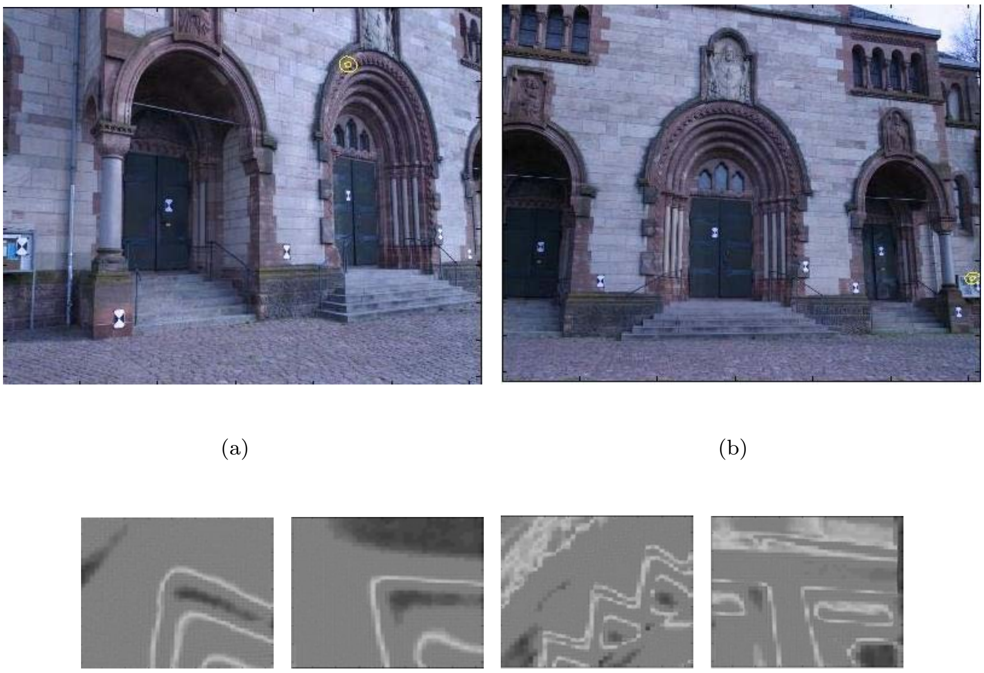
In order to solve this problem, we enlarge the covariance regions ω times, and define the new regions $q_i, \widetilde{q_j}$ as *neighborhood regions*, indicated by the big yellow ellipses in Figure 3.3(a) and Figure 3.3(b). We then perform the same normalization and obtain two drastically different image patches shown in Figure 3.3(e) and Figure 3.3(f), leading to the correct conclusion that $(f_i, \widetilde{f_j})$ is a false match. From here we could see the consistency of neighborhood regions is a necessarily complementary measurement for reliable region match.

We, therefore, construct a new descriptor pairs $(U_i, \widetilde{U_j})$ on the normalized image patches $q_{ni}, \widetilde{q_{nj}}$. $(U_i, \widetilde{U_j})$ comprises 16 SIFT-like gradient orientation histograms with 8 bins [4] extracted on the given image patches. And we could further define the following similarity measurement on neighborhood regions for each putative feature match $C_k(f_i, \widetilde{f_j})$:

$$s_{neighbor}(C_k(f_i, \widetilde{f_j})) = \frac{1}{2}(\|U_i - \widetilde{U_j}\|_2 + \frac{dRGB(q_{ni}, \widetilde{q_{nj}})}{100}) \tag{3.4}$$

Finally we could define the overall region context consistency of putative region match $(r_i, \widetilde{r_j})$ as

$$sim_{neighbor}(r_i, \widetilde{r_j}) = \frac{1}{L}\sum_{k=1}^{L} s_{neighbor}(C_k(f_i, \widetilde{f_j}))^2 \tag{3.5}$$



(a) (b)

(c) (d) (e) (f)

Figure 3.3: An illustration for region context consistency. (a) reference image, (b) target image, (c) normalized covariance region in reference image, (d) normalized covariance region in target image, (e) normalized neighborhood region in reference image, (f) normalized neighborhood region in target image

Maximality of Putative Feature Matches: This value represents the relative number of putative feature matches lying in the corresponding putative region match. It is obvious that the higher this value is, the more likely putative region match is to be correct. Therefore, this value is defined by the following equation

$$sim_{num}(r_i, \widetilde{r_j}) = \frac{L}{max(Area(r_i), Area(\widetilde{r_j}))} \tag{3.6}$$

where $Area(\cdot)$ is to calculate the area of the input region.

From Equation 3.3, 3.5, and 3.6, the edge weight for each putative region match is given by

$$E(r_i, \widetilde{r_j}) = \lambda_{photo} sim_{photo}(r_i, \widetilde{r_j}) + \lambda_{neighbor} sim_{neighbor}(r_i, \widetilde{r_j}) + \lambda_{num} sim_{num}(r_i, \widetilde{r_j}) \tag{3.7}$$

3.4 Bipartite Graph Matching

From the definition of bipartite graph $G(V, E)$ in the above section, it is not very difficult to find this graph is actually featured by *many-to-many correspondence*. In order to determine CSRP, which is obligated by *one-to-one constraint*, bipartite graph matching is used to find the best one-to-one match M between the disjoint vertex sets S and T belonging to graph $G(V, E)$.

Suppose the numbers of vertices in S and T are m and n respectively, $E_{i,j}$ represents the edge weight between the ith vertex in set S and jth vertex in set T. If these two vertices do not have any edge between them, E is set to be 0. We further define a *incidence vector* x by

$$x_{i,j} = \begin{cases} 1, & (i,j) \in M \\ 0, & otherwise \end{cases} \tag{3.8}$$

Hence the above bipartite graph matching problem can be converted into a integer program listed below:

$$max \sum_{j=1}^{n} \sum_{i=1}^{m} E_{i,j} x_{i,j}$$

subject to

$$\begin{aligned} &\sum_{j=1}^{n} x_{i,j} = 1, i \in S \\ &\sum_{i=1}^{m} x_{i,j} = 1, j \in T \end{aligned} \tag{3.9}$$

We then use Hungarian method to find the optimal incidence vector x in Equation 3.9. And CSPR is determined by Equation 3.10

$$CSPR = \{(i,j) \mid (i,j) \in M, E_{i,j} \geq \tau_r\} \tag{3.10}$$

where τ_r is a predefined threshold for the edge weights. And the putative feature matches within CSRP are all regarded as our final matches.

3.5 Experimental Results

3.5.1 Implementation Parameters

For local image feature extraction, we use Hessian-Affine detector [53] with SIFT as a feature descriptor. After putative feature matching, potential mismatches are rejected through the ratio test with the threshold $\tau_s = 0.8$. In our experiments, the parameters in edge weight of bipartite graph are fixed as follows: $\lambda_{photo} = 13$, $\lambda_{neighborhood} = 2$, $\lambda_{num} = 4$. And threshold for the edge weights τ_r is set to the value for output final top 100 best matches.

3.5.2 Results and Discussions

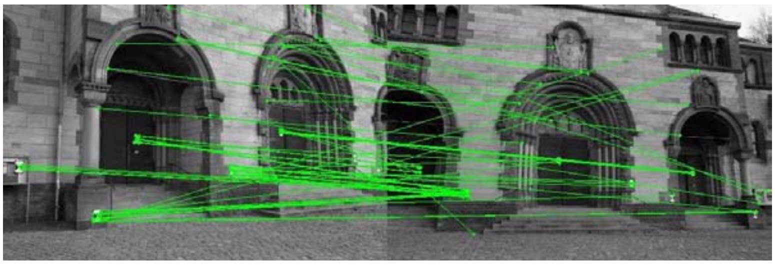
To evaluate the performance of our proposed algorithm, we test it on 3 sequences: *fountain, castle and Herz-Jesu* from the dataset created by EPFL [5]. Each sequence contains several images of the same building captured under different viewpoint at the interval of approximate 20 degrees. We further select 4 image pairs from each sequence with the viewpoint variation around 80 degrees as our test data and some of them are shown in Figure 3.4. Those images pose several challenges on the wide-baseline matching, like significant viewpoint difference, uniform color and texture as well as many repetitive structure. Consequently, the correct match for a given point may usually fail when SIFT matching is applied alone, as is shown in Figure 3.5(a). However, our proposed method on one hand efficiently bundles the features lying in the similar color and texture region together to perform matching, while on the other hand remarkably eliminates the matches containing feature points in different image segments. In this way, the matching ambiguous is greatly reduces as in Figure 3.5(b).

To compare quantitatively the difference between the SIFT matching and our proposed method, we count how many correct matches are there in the final top 100 best matches from these two methods respectively and Table 3.1 shows the percentage of correct matches of 4 sequences. It is obvious that the number of correct matches in all sequence is around 2 times higher than that in the SIFT matching.

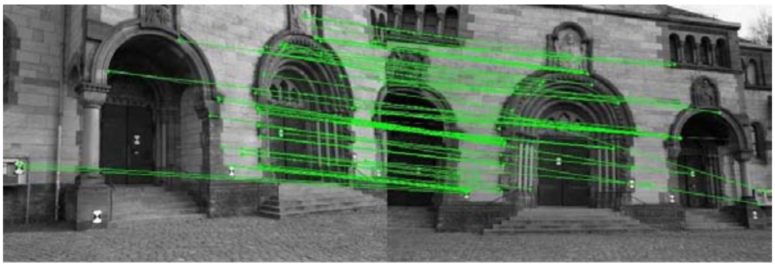
Figure 3.4: Image 1 and 5 of test sequence *fountain*, *castle* and *Herz-Jesu*

Table 3.1: Percentage of Correct Matches on EPFL Dataset of Top 100 Matches

Sequences	Image Number	SIFT	Proposed Method
fountain	1, 5	72%	86%
	2, 6	67%	95%
	3, 7	54%	94%
	4, 8	57%	100%
castle	1, 5	37%	78%
	2, 6	36%	71%
	3, 7	37%	74%
	4, 8	31%	68%
Herz-Jesu	1, 5	35%	68%
	2, 6	42%	89%
	3, 7	33%	72%
	4, 8	22%	54%



(a)



(b)

Figure 3.5: Top 100 matches of image 3 and 6 of sequence *Herz-Jesu* by (a) SIFT matching, (b) our proposed method

Chapter 4

Conclusions and Future Works

4.1 Conclusions

In this report, two difficulties of *Visual Location Recognition and Registration* are presented. The first one is how to best represent each location and architect an accurate and robust location recognition system. While the latter one is how to eliminate the matching ambiguity often occurred in the image registration of large viewpoint variation. Therefore, in this section we would briefly recapitulate our main contributions in solving the above two problems and highlight the most important experimental findings.

Location Recognition based on Bag-of-Features Model: Chapter 2 has presented an orderless *Bag-of-Features* model for recognizing different locations through the images captured under large scale changes, rotations and even perspective distortions. The use of visual word histogram has enabled us to best characterize each location under a wide range of geometric variability and therefore a fast and reliable classification can be performed. Being aware that different visual words actually carry different discrimination information in the sequential location classification, we propose a simple while novel visual word weighting scheme: *Visual Words Aggregation Weighting*, which considers those visual words being the cluster centers of highly aggregated local image features while with less neighboring words to be important than others. By applying this weighting scheme, a weighted visual histogram would be obtained and used in the following location classification and the experiment results in Section 2.4.2 show a significantly improvement in the final location recognition rate when compared with the state-of-art TF-IDF weighting algorithm.

Image Registration via Region Pairing: Chapter 3 has presented a novel image registration algorithm via region pairing, which intends to eliminate the matching ambiguity by adding a spatial support from exploring the similarity within the image. First, image over-segmentation is used to group image pixels into local homogenous regions and point-to-point feature matches are established. Then we propose a bipartite graph model to represent the relationship between those coherent segments and finally graph matching is utilized to obtain one-to-one region pairing results, which significantly facilitates rejecting the mismatches in the pervious point-to-point feature correspondence. In Section 3.5.2, a number of image pairs of different sequences featuring severe viewpoint variation, uniform color and texture as well as self-similarities are tested and the obtained experiment results indicate our algorithm is adequate for the use in the real world situations.

4.2 Future Works

In the future, we will work towards improving the flexibility, discriminative power, and expressiveness of *Bag-of-Features* model to further increase the accuracy and robustness of location recognition.

Feature Representation: The salient region detectors and appearance-based descriptor used in this report are all best at capturing stable and distinctive texture patterns. Consequently, our proposed method would always fail for those location, which do not characterized in their texture and appearance, such as in Figure 4.1. In order to overcome this difficulty, a shape detector and descriptor should be developed as a complimentary of texture [54].

(a) (b)

Figure 4.1: Locations lacking texture information challenge the capabilities of our current representation methods in the Bag-of-Features model: (a) Plaster sculpture exhibition in the National Gallery, London; (b) Louvre glass pyramid

Sparse Coding for Codebook Generation: How to discriminately design the visual codebook still remains as the central problem for the *Bag-of-Features* model. Recent research [55] based on *sparse codes* of SIFT features indicates that by relaxing the restrictive cardinality constraint of the K-means during the codebook generation and using *max* spatial pooling instead of computing histogram, the new image representation captures more salient properties of visual patterns and turns out to work much better even with linear classifier. Therefore, further research of this in empirical study and theoretical understanding is an interesting direction.

Bibliography

[1] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] J. Sivic and A. Zisserman, “Efficient visual search for objects in videos,” vol. 96, no. 4, 2008, pp. 548 – 566.

[3] V. C. P. Tirilly and P. Gros, “A review of weighting schemes for bag of visual words image retrieval,” IRISA, Tech. Rep., 2009.

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[5] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, “On benchmarking camera calibration and multi-view stereo for high resolution imagery,” in *Proceedings of Computer Vision and Pattern Recognition*, 2008, pp. 1 –8.

[6] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image features,” in *Proceedings of International Conference on Computer Vision*, 2005, pp. 1458–1465.

[7] A. Vailaya, A. Jain, and H. J. Zhang, “On image classification: city images vs. landscapes,” *Pattern Recognition*, vol. 31, no. 12, pp. 1921 – 1935, 1998.

[8] M. Szummer and R. W. Picard, “Indoor-outdoor image classification,” in *IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 42–51.

[9] I. Biederman, *Aspects and extension of a theory of human image understanding*, Z. Pylyshyn, Ed. Ablex Publishing Corporation, 1998.

[10] D. F. S. Thorpe and C. Marlot, “Speed of processing in the human visual system,” *Nature*, vol. 381, pp. 520–522, 1996.

[11] C. K. L. Fei-Fei, R. VanRullen and P. Perona, "Natural scene categorization in the near absense of attention," in *Proceedings of the National Academy of Sciences USA*, vol. 99, no. 14, 9601, 2002, p. 9596.

[12] L. Renninger and J. Malik, "When is scene identification just texture recognition," *Vision Research*, vol. 44, pp. 2301–2311, 2003.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[14] L. Lu, K. Toyama, and G. D. Hager, "A two level approach for scene recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 688–695.

[15] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Pattern Recognition Symposium, DAGM*, 2004.

[16] L. Fei-fei, "A bayesian hierarchical model for learning natural scene categories," in *In CVPR*, 2005, pp. 524–531.

[17] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision*, 2004, pp. 1–22.

[18] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan, "Categorizing nine visual classes using local appearance descriptors," in *In ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[19] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *In IEEE International Conference on Computer Vision*, 2005, pp. 370–377.

[20] D. Nistr and H. Stewnius, "Scalable recognition with a vocabulary tree," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[21] S. Lazebnik and R. Maxim, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, 2009.

[22] S. Lazebnik and M. Raginsky, "Learning nearest-neighbor quantizers from labeled data by information loss minimization," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, March 2007 2007, pp. 1–8.

[23] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of Computer Vision and Pattern Recognition*, 2007.

[24] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," vol. vol.2, 2003, pp. 1470 – 1477.

[25] J. M. S. Belongie and J. Puzicha, "Image retrieval using both color and texture features," *The Journal of China Universities of Posts and Telecommunications*, vol. 14, no. 1, pp. 94–99, 2007.

[26] A. K. M. S. S. A. Vadivel, "Performance comparison of distance metrics in content-based image retrieval applications," in *Proceedings of the International Conference on Information Technology*, Bhubaneswar, India, December 2003, pp. 1–6.

[27] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, p. 2005, 2005.

[28] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of European Conference Computer Vision*. Springer Verlag, 2002, pp. 128–142.

[29] J. J. Koenderink and A. J. van Doom, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, no. 6, pp. 367–375, 1987.

[30] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 891–906, 1991.

[31] L. J. V. Gool, T. Moons, and D. Ungureanu, "Affine/ photometric invariants for planar intensity patterns," in *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume I*, 1996, pp. 642–651.

[32] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets," in *Proceedings of the 7th European Conference on Computer Vision*, 2002, pp. 414–431.

[33] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509–522, 2002.

[34] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1265–1278, 2005.

[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[36] D. Comaniciu, P. Meer, and S. Member, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[37] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.

[38] D. Lewis, "Navie bayes at forty: the independence assumption in information retrieval," in *Proceedings of the European Conference on Machine Learning*, 1998, pp. 4–15.

[39] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *In Proceedings of the European Conference on Computer Vision*, 2006, pp. 490–503.

[40] C. Schmid, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[41] S. L. J.-G. Zhang, M. Marszalek and C. Schmid, "Local features and kernels for classification of texture and object categories: An in-depth study," INRIA, France, Tech. Rep., 2005.

[42] V. Ferrari, T. Tuytelaars, and L. Gool, "Simultaneous object recognition and segmentation from single or multiple model views," *International Journal of Computer Vision*, vol. 67, pp. 159–188, 2006.

[43] P. Pritchett and A. Zisserman, "Matching and reconstruction from widely separated views," in *Proceedings of the European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, 1998, pp. 78–92.

[44] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions via graph cuts," in *Proceedings of International Conference on Computer Vision*, 2001, pp. 508–515.

[45] O. Chum and J. Matas, "Matching with prosac: Progressive sample consensus," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 220–226.

[46] O. Duchenne, F. Bach, I. Kweon, and J. Ponce, "A tensor-based algorithm for high-order graph matching," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1980–1987.

[47] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proceedings of International Conference on Computer Vision*, 2005.

[48] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondence," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 26–33.

[49] B. Fischer, S. Member, and J. M. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 513–518, 2003.

[50] D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of International Conference on Computer Vision*, 1999, pp. 1150–1157.

[51] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[52] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004, british Machine Vision Computing 2002.

[53] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[54] M. Stark, M. Goesele, and B. Schiele, “A shape-based object class model for knowledge transfer,” in *In Twelfth IEEE International Conference on Computer Vision*, 2009.

[55] J. Yang, K. Yu, Y. Gong, and T. S. Huang, “Linear spatial pyramid matching using sparse coding for image classification.” in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 1794–1801.

Project Manpower

- Chen Wang: PhD Student (August 2008 $\sim$ Present), School of Electrical and Electronic Engineering, Nanyang Technological University Singapore. Tentative thesis title *Visual Location Recognition and Registration*.
- Jing Zhu: Graduated M.Eng Student (August 2007 $\sim$ June 2009), School of Electrical and Electronic Engineering, Nanyang Technological University Singapore. Thesis title *Content-based Feature Weighting for Scene Recognition*.
- Baojiang Zhong: Research Scientist, Temasek Lab Nanyang Technological University Singapore.

Publication

- Wang Chen and Kai-Kuang Ma, "Wide-baseline Image Matching via Region Pairing", *British Machine Vision Conference 2010*, Aberystwyth, United Kingdom, August 2010. (To Be Submitted)
- Baojiang Zhong and Kai-Kuang Ma, "A Succinct Shape Matching Algorithm Based on Rectangularized Curvature Scale-space Map", *European Computer Vision Conference 2010*, Hersonissos, Greece, September 2010. (Under Reviewing)
- Baojiang Zhong and Kai-Kuang Ma, "On Convergence of Smoothed Planar Curves", *IEEE Transaction on Image Processing.* (Accept)
- Baojiang Zhong, Kai-Kuang Ma and Wenhe Liao, "Scale-Space Behavior of Planar-Curve Corners", *IEEE. Transaction on Pattern Recognition and Machine Intelligence*, vol. 31, no. 8, pp. 1517-1524, August 2009.