Delft University of Technology

Large Scale Image Retrieval for Location Estimation

Li, Xinchao

DOI
10.4233/uuid:0d09c0dc-fcb7-4598-90e0-d2a53e675cc3

Publication date
2016

Document Version
Final published version

Citation (APA)
Li, X. (2016). *Large Scale Image Retrieval for Location Estimation*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:0d09c0dc-fcb7-4598-90e0-d2a53e675cc3

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

LARGE SCALE IMAGE RETRIEVAL FOR LOCATION ESTIMATION

LARGE SCALE IMAGE RETRIEVAL FOR LOCATION ESTIMATION

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
woensdag 12 oktober 2016 om 12:30 uur

door

Xinchao LI

Master of Science in Information Science and Engineering,
Shandong University
geboren te Jinan, China.

This dissertation has been approved by the

promotors: Prof. dr. A. Hanjalic and Prof. dr. M.A. Larson

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. dr. A. Hanjalic	Delft University of Technology
Prof. dr. M.A. Larson	Radboud University
Independent members:	
Prof. Dr.-lng. habil. B. lonescu	University Politehnica of Bucharest
Prof. dr. R.C. Veltkamp	Utrecht University
Prof. dr. M. Worring	University of Amsterdam
Prof. dr. ir. G.J.P.M. Houben	Delft University of Technology
Dr. S. Papadopoulos	Centre for Research & Technology Hellas
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, reserve member

Copyright © 2016 by Xinchao Li

ISBN 978-94-92516-11-4

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To my parents

献给我的父亲母亲

CONTENTS

Summary **v**

Samenvatting **vii**

1 Introduction **1**
1.1 Why Do We Need Geo-annotated Multimedia? 2
1.2 On Automatic Geo-annotation of Social Media 2
1.3 Thesis Focus 3
1.4 Thesis Scope 6
1.5 Thesis Contribution and Layout 7
1.6 How to Read the Thesis 10
1.7 List of Publications Related to the Thesis 10
References 13

2 Geo-Visual Ranking **15**
2.1 Introduction 16
2.2 Rationale and Contribution 18
2.3 Related Work 20
2.3.1 Geo-constrained content-based location prediction 20
2.3.2 Content-based location prediction without geo-constraints 22
2.4 Geo-Visual Ranking (GVR) 23
2.4.1 Candidate image selection 23
2.4.2 Location extraction 25
2.4.3 Location ranking 26
2.4.4 Reducing the effect of high-volume uploads 28
2.5 Experimental Setup 28
2.5.1 Dataset. 28
2.5.2 Computing visual similarity 29
2.5.3 Experimental comparison 30
2.5.4 Evaluation procedure 31
2.6 Experimental Results 32
2.6.1 Stage 1: Comparative analysis of GVR, VisNN and MSC 32
2.6.2 Stage 2: Comparative analysis of GVR and state of the art 34
2.6.3 Stage 3: Applicability scope 36
2.7 Conclusion 37
References 41

3 Pairwise Geometric Matching 45
3.1 Introduction 46
3.2 Related Work and Contribution 47
3.2.1 Exploiting individual correspondences 48
3.2.2 Exploiting multiple correspondences 48
3.3 Correspondence Problem Formulation 49
3.4 Pairwise Geometric Matching 50
3.4.1 1vs1 matching 50
3.4.2 Hough voting 51
3.4.3 Integrating global and pairwise geometric relations 51
3.5 Experimental Setup 52
3.5.1 Object retrieval framework 52
3.5.2 Experimental protocol 53
3.5.3 Datasets 53
3.6 Experiments 53
3.6.1 Impact of the parameters 53
3.6.2 Pairwise image matching 53
3.6.3 Spatial verification for object retrieval 54
3.6.4 Run time efficiency 57
3.7 Discussion 58
References 63

4 Geo-distinctive Visual Element Matching 65
4.1 Introduction 66
4.2 Rationale and Contribution 68
4.3 Related Work 71
4.3.1 Geo-constrained content-based location estimation. 72
4.3.2 Geo-unconstrained content-based location estimation 72
4.3.3 Geo-distinctive visual element modeling 73
4.4 Geo-distinctive Visual Element Matching 74
4.4.1 Location as visual element cloud. 75
4.4.2 Geo-distinctiveness modeling 77
4.4.3 Visual matching per location. 79
4.5 Experimental Setup 81
4.5.1 Dataset. 82
4.5.2 Computing visual similarity 82
4.5.3 Experimental design 83
4.6 Experimental Results 83
4.6.1 Impact of the parameters 84
4.6.2 Geo-constrained location estimation 85
4.6.3 Geo-unconstrained location estimation 86
4.6.4 Comparison with the state-of-the-art 89
4.7 Conclusion 89
References 93

5 Discussion and Future Work 97
5.1 Discussion . . . 98
5.2 Directions for Future Research . . . 100
5.3 Conclusion . . . 102
References . . . 103

Acknowledgements i

Curriculum Vitae iii

SUMMARY

THE geo-graphical location at which an image or video was taken is a key piece of multimedia information. Such geo-information has become an indispensable component of systems enabling personalized and context-aware multimedia services. The research reported in this thesis investigates how to automatically derive geo-information from multimedia content. In particular, it focuses on the challenge of estimating the geo-coordinates of the location of an image solely on the basis of its visual content.

The goal of the research is to develop a scalable visual content-based location estimation system for images and to investigate the possibilities to improve its accuracy and reliability to a substantial extent. The system should be applicable in both the geo-constrained scenario, in which the multimedia item is taken at one of a previously defined set of locations, and the geo-unconstrained scenario, in which the multimedia item could have been taken anywhere in the world.

The thesis makes two different kinds of contributions. The first is high-level framework design. We develop a generic large-scale image retrieval-based framework for location estimation. The second is optimization of specific components of the system. We develop two approaches, geometric verification and geo-distinctive visual element matching, that address specific challenges faced by our retrieval-based framework. The resulting system makes location estimation more tractable in case of large image collections, and also more reliable. Our experimental results demonstrate that the system leads to an overall significant improvement of the location estimation performance and redefines the state-of-the-art in both geo-constrained and geo-unconstrained location estimation.

Based on the findings presented in this thesis, we make recommendations for future research directions, which we think are substantial and promising for large scale image retrieval and geo-location estimation.

SAMENVATTING

De geografische opnamelocatie van een afbeelding of video is belangrijke multimedia-informatie. Zulke geografische informatie is inmiddels een essentiële component in systemen die gepersonaliseerde en contextgevoelige multimediadiensten aanbieden. Het onderzoek in deze dissertatie houdt zich bezig met de vraag hoe geografische informatie automatisch uit multimedia-inhoud kan worden afgeleid. In het bijzonder ligt de focus op de uitdaging om, puur op de basis van visuele inhoud, geografische coördinaten van de opnamelocatie van een afbeelding in te schatten.

Het doel van het onderzoek is de ontwikkeling van een schaalbaar systeem voor locatie-inschatting op basis van visuele inhoud. Daarnaast richt het onderzoek zich erop om de mogelijkheden te verkennen om de nauwkeurigheid en betrouwbaarheid van dit systeem substantieel te verbeteren. Het systeem moet zowel toepasbaar zijn in de situatie waarin de opnamelocatie van het multimedia-item binnen een eerder vastgelegde verzameling van mogelijke locaties valt, als in de situatie waarin het multimedia-item op elke locatie in de wereld opgenomen kan zijn.

De dissertatie levert twee hoofdbijdragen. De eerste bijdrage is het ontwerp van een overkoepelend systeemraamwerk. We ontwikkelen hierbij een generiek raamwerk voor locatie-inschatting, gebaseerd op zoekmachinetechnieken voor grote beeldbanken. De tweede bijdrage is de optimalisatie van specifieke componenten in het systeem. We ontwikkelen twee aanpakken, geometrische verificatie en geo-discriminative matching van visuele elementen, waarin specifieke uitdagingen van onze raamwerk worden behandeld. Het resulterende systeem maakt de automatische inschatting van opnamelocatie haalbaarder en schaalbaarder voor grote beeldbanken, en daarnaast ook betrouwbaarder. De uitkomsten van onze experimenten demonstreren dat het systeem voor een significante prestatieverbetering in locatie-inschatting leidt, en dat de wetenschappelijke stand van zaken in lokaliseringsalgoritmen voor vastgelegde locatieverzamelingen en vrije locatiemogelijkheden door het systeem zijn hergedefinieerd.

Op basis van de inzichten in deze dissertatie doen we aanbevelingen voor toekomstige onderzoeksrichtingen, waarvan we menen dat die belangrijk en veelbelovend zijn voor het zoeken van afbeeldingen op grote schaal en de inschatting van geografische opnamelocaties.

1

INTRODUCTION

1.1. WHY DO WE NEED GEO-ANNOTATED MULTIMEDIA?

WE have witnessed rapid development and widespread usage of personal multimedia capturing devices such as cameras, phones and tablets over the past years. In combination with the immense popularity of social media, this development has enabled and stimulated generation and exchange of multimedia content on the Internet at a massive scale. User-generated multimedia, and in particular images and videos, has become an important aspect of our expression and interaction, complementing and enhancing the traditional communication channels.

The relevance and significance of user-generated multimedia in this respect has further grown with the increasing ease with which users can annotate content that they have captured. Annotations serve to accompany multimedia content with additional descriptive information, commonly referred to as *metadata*. For example, textual metadata in the form of *tags* or *captions* may be added to provide additional information about the captured content (e.g., what is displayed) or about the context in which the content was captured (e.g., an event at which an image was taken). However, metadata can also serve to provide "technical" information about the captured images or video, for instance, the time of capture, information about the creators, view count or sharing history. An important type of metadata belonging to this category is *geo-information*. This type of information is often expressed as geo-coordinates, i.e., the latitude and longitude of the location of the captured visual scene. In this thesis, we refer to geo-information expressed as geo-coordinates as *geo-location*. Information about the geo-location at which an image or a video was taken can assist in a wide range of applications involving user-generated content. For example, one can find popular objects and events in a particular area [1, 2], generate representative and diverse views of a geo-location [3], and recommend virtual tours by presenting information mined from user-generated travelogues and photos [4].

Our use of the term *geo-location* emphasizes that we are interested in the position of a location on a map. We exclude from the scope of inquiry other location-related information such as type of location, as determined by the function of a location (e.g., amusement park, outdoor market, or forest). We also exclude other socially or politically determined aspects of location (e.g., the boundaries of neighborhoods within a city, or the position of the border between two countries). We often use the word "location" to discuss our approaches, but in the context of this thesis "location" should be interpreted as "geo-location".

With the increasing demands of users for personalized and context-aware multimedia services, geo-information has became an indispensable component of systems enabling such services. The research reported in this thesis looks into the possibilities to facilitate automatic geo-annotation of user-generated multimedia, and specifically of the *social images* uploaded and shared on social media platforms.

1.2. ON AUTOMATIC GEO-ANNOTATION OF SOCIAL MEDIA

Many modern mobile devices make it possible via their GPS modules to automatically assign geo-coordinates to images/videos during capture. If this functionality is not used, an alternative is the use of location-aware interfaces that are designed for users to carry out manual geo-annotation of the content that they create, e.g., the geo-tagging possibilities offered by Flickr. Still, however, it is estimated that less than 10% of the images shared on social media are geo-tagged [5][6], which significantly reduces the foundations for developing the above-mentioned multimedia services. The reasons for this low percentage are potentially various. For example, just like with offline tagging in general, users may be insufficiently motivated to apply such interfaces to enrich their image collections. Offline tagging is typically found tedious and time-consuming [7, 8].

As an alternative to GPS or manual geo-annotation, increasing research attention has been devoted to techniques that automatically estimate geo-locations for images. Such approaches are commonly referred to as *location estimation (prediction)* techniques [6, 9–11]. There are multiple resources that can be exploited as clues for location estimation, ranging from content to metadata. As one example, we provide Fig. 1.1, which depicts a video shared on Flickr. People can infer the location of the video, illustrated in Fig. 1.2, from the visual content (if they remember the particular scene depicted in the video), from the acoustic content (if they recall the sound from the specific clock tower), from the location-specific tags (such as “Italy”, “Tuscany” and “Florence”), or from the other media items in the owner’s album which were taken around the same time as the video (if their locations are known).

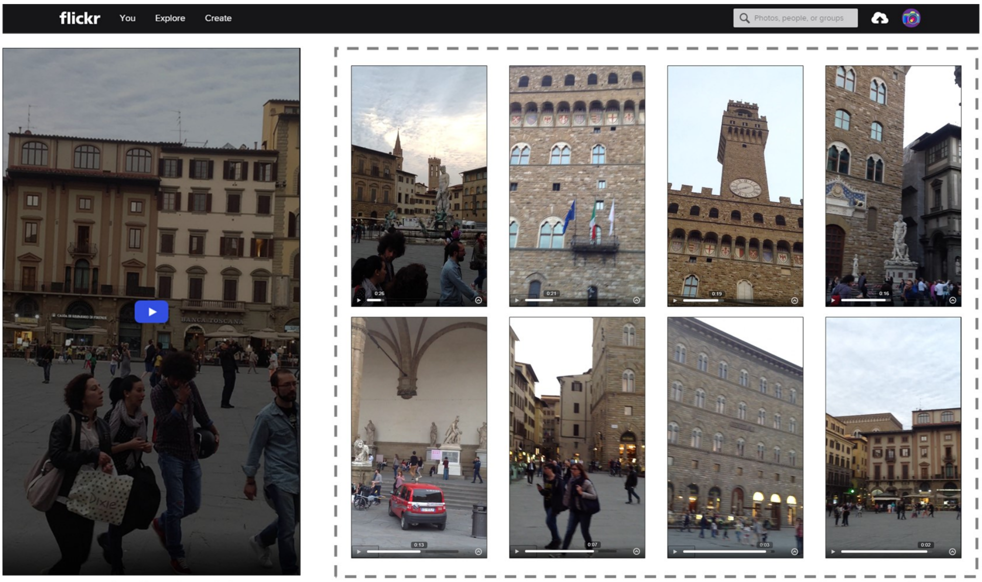
Among all the modalities within the multimedia, textual metadata that often accompany social media usually include place names and other location-specific terms, e.g., over 13% of image tags on Flickr could be classified as locations using WordNet as reported in [12]. As textual metadata is contributed by people, who can provide accurate and high level information about the image, textual metadata has served as the basis for a broad range of geo-location estimation algorithms (e.g., [6, 13]). However, the drawback of textual annotation-based location estimation is that annotations need to be manually created by users before prediction can be carried out.

Much of the research effort seeking reliable alternatives to textual metadata has therefore focused on exploiting the visual content of images directly to estimate the location of the depicted scene [9, 14, 15]. These approaches have the advantage of not depending on the availability or resolution of textual or audio metadata. Despite the numerous methods proposed in this direction, the challenge of estimating image location from the analysis of its visual content remains considerable. The research reported in this thesis aims at bringing the research community a substantial step further in pursuing this challenge.

1.3. THESIS FOCUS

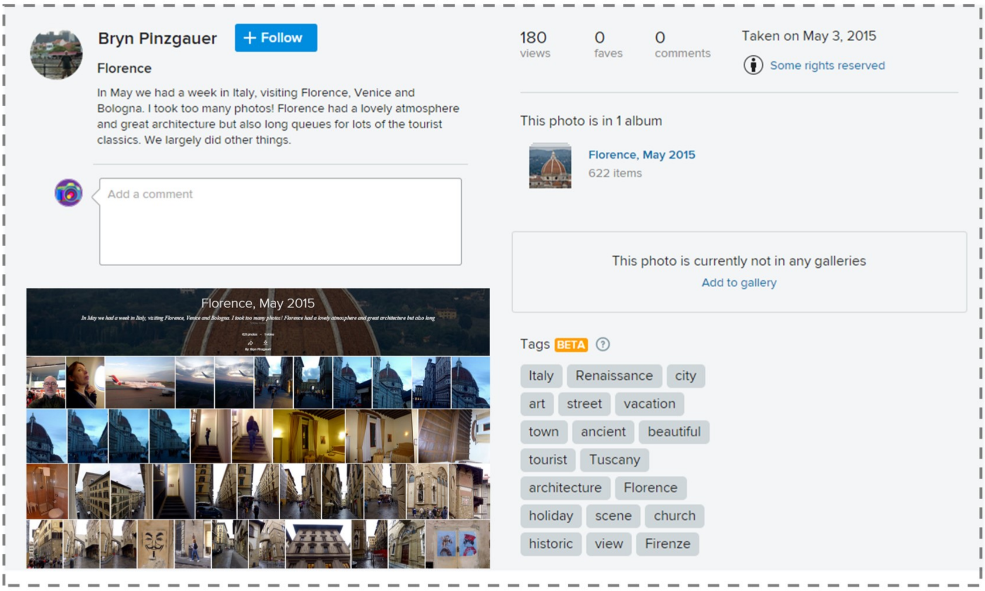
The challenge addressed in this thesis is illustrated in Fig. 1.3 and can be formulated as follows: *“given the visual content of an image, determine the geo-coordinates of the location of the depicted scene”.*

This challenge is substantial due to several reasons. First, one and the same visual scene can be captured under strongly varying conditions determined by the level or type of light, distortions, zoom or occlusion. Second, depending on the capture angle and direction, different scenes can be captured at one and the same location. For instance, standing on a particular spot on a beach, one can take a photo of the sea, but also of the beach or of the street running in parallel with the shore. This means that there is no unique link between the visual scene and a location. Third, the number of different unique scenes and locations worldwide is effectively infinite. Due to these reasons, most of the work in this direction has reported attempts which first make the challenge tractable before performing



One Video Shared on Flickr

Exemplar Frames of the Video



Bryn Pinzgauer

Florence

In May we had a week in Italy, visiting Florence, Venice and Bologna. I took too many photos! Florence had a lovely atmosphere and great architecture but also long queues for lots of the tourist classics. We largely did other things.

Florence, May 2015

180 views 0 faves 0 comments

Taken on May 3, 2015

Some rights reserved

This photo is in 1 album

Florence, May 2015
622 items

This photo is currently not in any galleries

Tags

Italy, Renaissance, city, art, street, vacation, town, ancient, beautiful, tourist, Tuscany, architecture, Florence, holiday, scene, church, historic, view, Firenze

Accompanying Metadata

Figure 1.1: Example of a video shared by a user on Flickr.

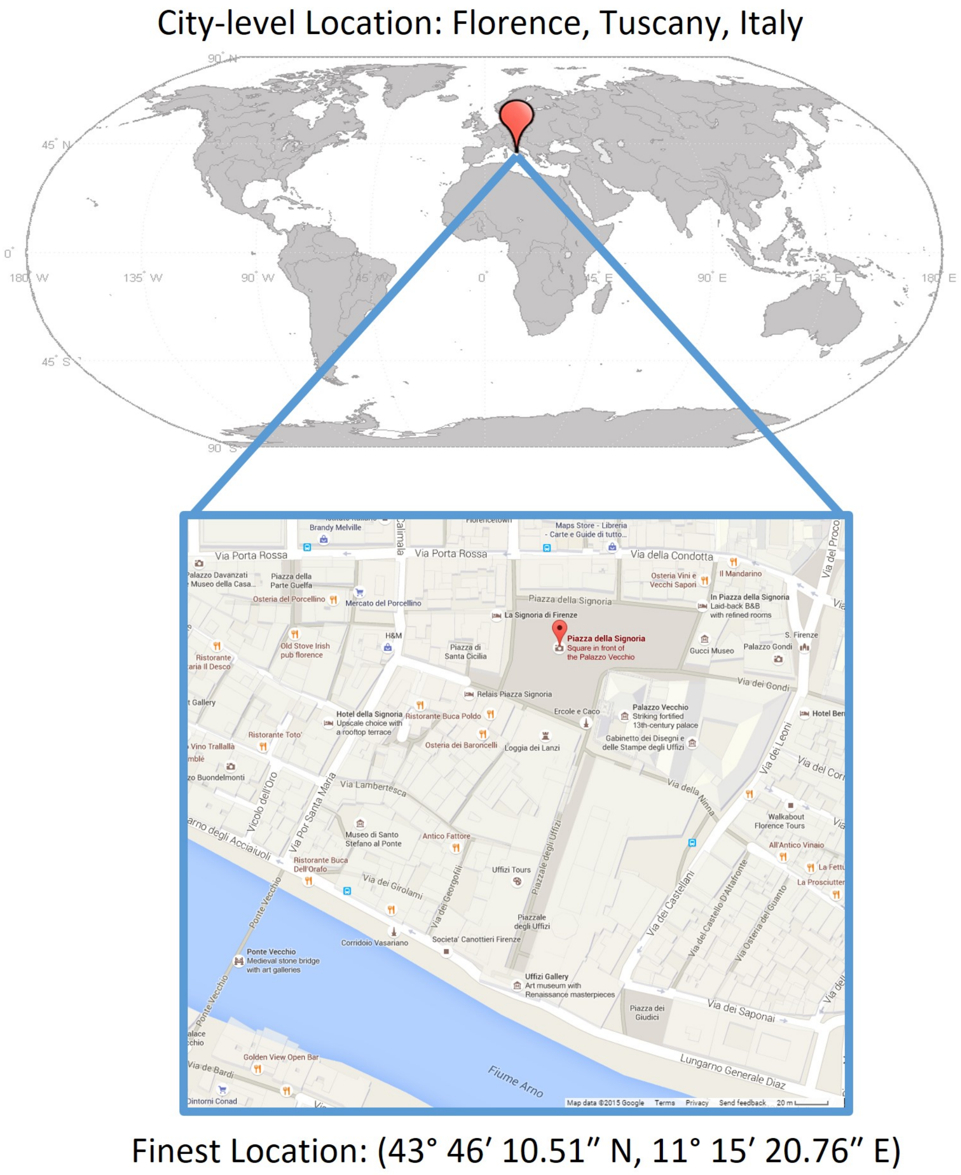


Figure 1.2: Illustration of the location of the video in Fig. 1.1.

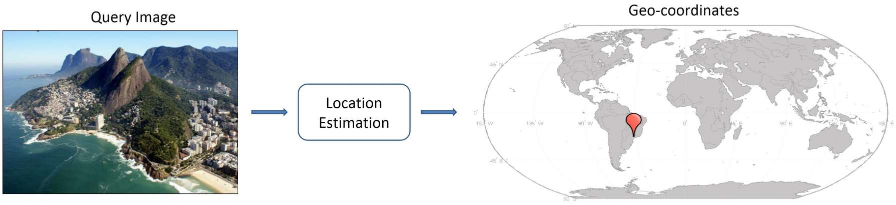


Figure 1.3: Illustration of the challenge of estimating the geo-location of an image solely using its visual content.

location estimation. These approaches typically attempt to narrow the domain of estimation and tackling the task in a *geo-constrained* way. They either estimate location within a geographically constrained area (e.g., in downtown San Francisco [15, 16]), within a set of predefined regions (e.g., ca. 1.5k places of interest around the world [17]), or by reducing the task to specific landmark recognition [14, 18].

Due to the difficulty of the challenge, there have only been a few attempts to tackle the geo-location estimation problem in a *geo-unconstrained* way, that is, where the target location can be any place around the world, for example [9]. Although the challenge is considerable, a recent survey [19] has indicated that there are still ample opportunities to address it that are waiting to be explored. In view of these considerations, and of the need for general solutions to location estimation that can operate on a global scale, we have focused our research on estimating geo-location of images without geo-constraints. Our expectation is that the solutions arising from our research should work well in both scenarios, that is for both geo-constrained and geo-unconstrained location estimation.

1.4. THESIS SCOPE

Two general approaches can be followed in order to infer the location information from the visual content of images:

- **Model-based (classification-based) approach**: Models are generated for a set of predefined locations by letting the system learn the visual characteristics of the location. Then, it is still possible to estimate the location based on the match between the visual characteristics of the target image and the learned characteristics of locations.
- **Memory-based (search-based) approach**: Here, the target image serves as query that is used to search a collection of geo-annotated images. Based on the predefined models of computing the relevance to the query, the geo-coordinates from the collection image(s) landing on the top of the result list are used to compute the geo-coordinates of the target image.

Model-based approaches collect location-related visual clues from different training images, and make a compact representation for individual locations. Their disadvantage is that it is highly problematic to formulate a location as a single class. If one divides the world into a limited number of regions, e.g., cities, then the visual diversity of the images collected from such regions, which are relatively large, will make it virtually impossible to learn a reliable (sufficiently discriminative) location model. In addition, the model tends to learn the frequently photographed visual scenes characteristic of one location, but not the rarely photographed ones. For this reason, model-based approaches struggle to handle queries that capture a rarely photographed visual scene. In contrast, memory-based approaches only require a single relevant photo for a given query; they do not care whether this photo captures a frequently photographed visual scene or a rarely photographed one. One single matched relevant photo is enough to estimate the location of the given query. Furthermore, modeling large regions with model-based approaches also implies an upper limit to the precision of the estimated locations. This limit would make such estimation meaningless in many application scenarios. On the other hand, if each single geo-tagged image is defined as a location on the earth, then model-based approaches become equivalent to memory-based approaches.

These considerations lead us to choose memory-based location estimation as the scope of our research. The objectives of our research within this scope are to explore the possibilities to develop a search-based location estimation framework and to exploit these possibilities to significantly improve the location estimation performance compared to the widely used reference methods. The emphasis of our research is on neutralizing the main bottleneck of the search-based approach, namely its heavy dependence on the presence of geo-tagged images in the collection that have been taken at or very close to the location of the target image and substantially resemble the visual scene existing in the target image. The specific contributions of the thesis are described in detail in the following section.

1.5. THESIS CONTRIBUTION AND LAYOUT

To understand how a memory-based approach works, it is informative to think about a process a human might use for identifying the location of an image Given a target image, we would try to link elements of the depicted scene with the scenes we saw before. Transferred to the system level, the visual content of a given image ("what we see now") can be submitted as a query to assess its match with the visual content of other images ("what we saw before") that are geo-annotated ("for what we know where it was taken"). If there are geo-annotated images in the collection whose visual content is sufficiently similar to that of the target image, the geo-coordinates of these images can be used to estimate the geo-coordinates of the target image.

We implement this rationale from different perspectives, starting from designing a general search-based location estimation framework and then improving specific framework components on the basis of the findings coming out from framework assessment at various stages. Our aim is to maximize the robustness and scalability of location estimation. Robustness means that the system needs to be able to properly estimate the location even if the query image and its collocated relevant images are overlapping only in a small fraction of scene elements and under varying capture conditions. Scalability means that the developed solution needs to enable efficient location estimation at a global scale relying on a large-scale collection of geo-annotated images.

In view of above, we organize the thesis into three parts as illustrated in Fig. 1.4 and represented by the three subsequent technical chapters. In **Chapter 2**, we unravel the problem of location inference from visual content, introduce the search-based approach and propose a novel way of implementing it, namely in the form of a *Geo-Visual Ranking (GVR) method* that takes into account the ambiguity in how visual content reflects a location. The rationale underlying the *GVR* method is that, compared to the images from a wrong loca-

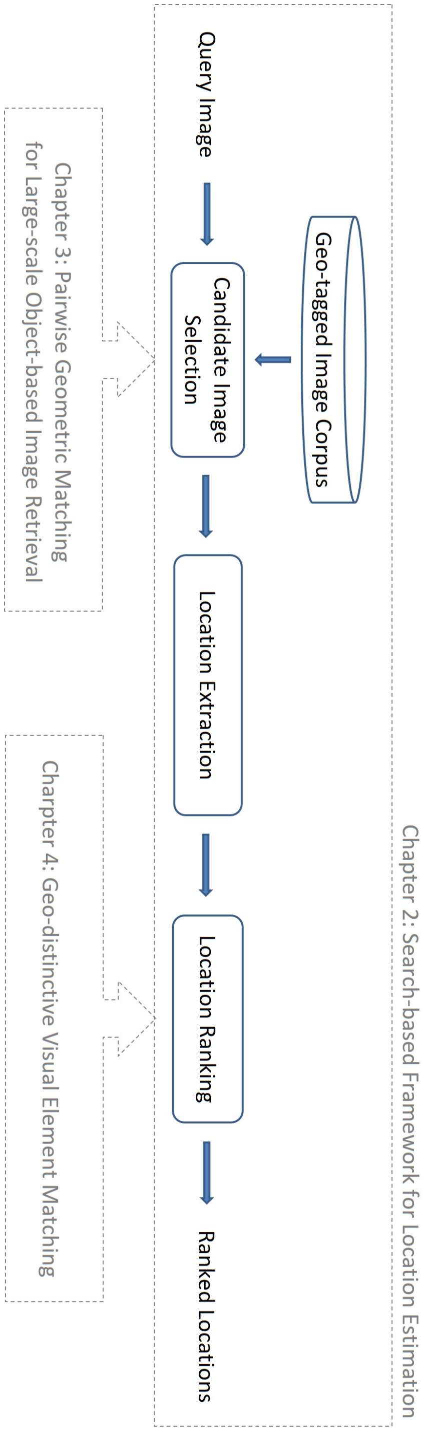

Query Image
Geo-tagged Image Corpus
Candidate Image Selection
Location Extraction
Location Ranking
Ranked Locations
Chapter 2: Search-based Framework for Location Estimation
Chapter 3: Pairwise Geometric Matching for Large-scale Object-based Image Retrieval
Charpter 4: Geo-distinctive Visual Element Matching

Figure 1.4: Illustration of the three parts of the thesis addressed in Chapters 2, 3 and 4, and their interrelations.

tion, more images from the true location will likely contain more elements of the visual content of the query image. We hypothesize that, for this reason, the evidence from the set of visually similar images from a wrong location is too weak to compete with the set captured at the true location, independently of the set size. Finally, we investigate the effect of different visual representations on location estimation within our framework. We find that although global features are known to be effective for retrieving semantically and structurally similar scenes, it is challenging to exploit them to improve the prediction effectiveness. We attribute this fact to the weakness of the relationship between scene types (e.g., “beach”, “city”, “landscape”), which global features are known to differentiate well, and specific locations. In contrast, local representations can establish stronger links between photos with same objects captured at one particular location, and can, in this way, generate more reliable prediction, exceeding the ability of global representations.

The findings of Chapter 2 lead us to focus on deriving location information from the objects captured in the images, or in other words using the *object-based image retrieval* approach. With object-based image retrieval we understand the problem of finding images that contain the same object(s) or scene elements as in the query image, however, possibly captured under different conditions in terms of rotation, viewpoint, zoom level, occlusion or blur. Many object-based image retrieval approaches and methods [20–22] have been proposed in recent literature, largely inspired by the pioneering work of Sivic and Zisserman [23] and built on the bag-of-features (BOF) principle for image representation. These retrieval systems generally consist of two main stages:

- Initial ranking stage, where the ranking of images from the collection is based on visual similarity computed on visual feature statistics measured in different images,
- Spatial verification stage, where the initial ranked list is re-ranked by applying geometric constraints to assess the reliability of visual correspondences between images.

The spatial verification stage is the key to achieving a high precision for object-based image retrieval, especially when searching in large, heterogeneous image collections [24]. In order to improve the scalability and robustness of object-based image retrieval in our *GVR* framework, in **Chapter 3**, we present a novel *Pairwise Geometric Matching* method for the spatial verification stage. It uses global scale and rotation relations to enforce the local consistency of geometric relations derived from the locations of pairwise correspondences. The results presented in this chapter indicate the suitability of the proposed pairwise geometric matching method as a solution for large-scale object retrieval at an acceptable computational cost.

While having a robust and scalable object-based image retrieval system as module in our framework is a necessary condition for successful location estimation, it is not a sufficient one for achieving the desired level of performance. Since some objects may be common to different visual scenes, e.g., common static objects and mobile objects, an additional adaptation of the framework is required to make it focus on the scene-distinctive objects only. Therefore, in **Chapter 4**, we present a novel *Geo-distinctive Visual Element Matching* method to further improve the robustness of our location estimation framework. It explores and exploits geographical distinctiveness of visual elements found in the query image, and it further strengthens the support for finding the true location by devising an aggregated visual representation of a location that combines all visual elements from the query found in the images of that location. The proposed method makes the location estimation more tractable in case of a large image collection, but also more reliable, which leads to an overall significant improvement of the location estimation performance and redefines the state-of-the-art in both geo-constrained and geo-unconstrained location estimation. **Chapter 5** concludes the thesis with a summary of achieved results and an outlook towards the still open research challenges in the domain of automatic geo-annotation of social images.

1.6. HOW TO READ THE THESIS

The technical part of this thesis consists of original publications that have been adopted as chapters 2, 3 and 4. The publications' references are given at the beginning of each chapter. As a consequence of working with original publications, the notation and terminology may vary slightly across chapters. For the same reason, the introductory parts and related work sections in the chapters addressing the same general topic may be similar in terms of argumentation and the material they cover. We retain the original form of the publications so that it is clear that the authoritative reference is the reference provided at the beginning of each chapter.

1.7. LIST OF PUBLICATIONS RELATED TO THE THESIS

The following papers have been published by the author of this thesis while pursuing a Ph.D. degree in the Multimedia Computing Group at the Delft University of Technology. Those publications directly serving as chapters of the thesis are indicated accordingly.

JOURNAL

- **Xinchao Li**, Martha Larson and Alan Hanjalic, Global-Scale Location Prediction for Social Images using Geo-Visual Ranking, *IEEE Transactions on Multimedia*, 17(5): 674-686, 2015. (Full paper)**—[Chapter 2]**

- **Xinchao Li**, Martha Larson and Alan Hanjalic, Geo-distinctive Visual Element Matching for Location Estimation of Images, submitted to *IEEE Transactions on Multimedia.* (Full paper)**—[Chapter 4]**

CONFERENCE

- **Xinchao Li**, Martha Larson and Alan Hanjalic, Pairwise Geometric Matching for Large-scale Object Retrieval, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, Boston, US, 2015. (Full paper)**—[Chapter 3]**

- **Xinchao Li**, Martha Larson and Alan Hanjalic, Geo-visual ranking for location prediction of social images, *Proc. International Conference on Multimedia Retrieval (ICMR '13)*, Dallas, US, 2013. (Full paper)

WORKSHOP

- **Xinchao Li**, Peng Xu, Yue Shi, Martha Larson and Alan Hanjalic, Simple Tag-based Subclass Representations for Visually-varied Image Classes, *Proc. International Workshop on Content-based Multimedia Indexing (CBMI '16)*, Bucharest, Romania, 2016.
- **Xinchao Li**, Michael Riegler, Martha Larson and Alan Hanjalic, Exploration of feature combination in geo-visual ranking for visual content-based location prediction, *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, 2013.
- **Xinchao Li**, Claudia Hauff, Martha Larson and Alan Hanjalic, Preliminary Exploration of the Use of Geographical Information for Content-based Geo-tagging of Social Video, *Proc. MediaEval 2012 Workshop*, Pisa, Italy, 2012.
- Jaeyoung Choi and **Xinchao Li**, The 2014 ICSI/TU Delft Location Estimation System, *Proc. MediaEval 2014 Workshop*, Barcelona, Spain, 2014.

REFERENCES

[1] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proc. SIGIR '07*, 2007.

[2] Jaeyoung Choi, Eungchan Kim, Martha Larson, Gerald Friedland, and Alan Hanjalic. Evento 360: Social event discovery from web-scale multimedia collection. In *Proc. MM '15*, 2015.

[3] S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *IEEE Trans. Multimedia*, 15(4):921–932, 2013.

[4] Qiang Hao et al. Travelscope: standing on the shoulders of dedicated travelers. In *Proc. MM '09*, 2009.

[5] Jaeyoung Choi and Gerald Friedland. *Multimodal Location Estimation of Videos and Images.* Springer, 2014.

[6] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing Flickr photos on a map. In *Proc. SIGIR '09*, 2009.

[7] G. Schindler, P. Krishnamurthy, R. Lublinerman, Yanxi Liu, and F. Dellaert. Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In *Proc. CVPR '08*, 2008.

[8] Ning Zhou, W.K. Cheung, Guoping Qiu, and Xiangyang Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1281–1294, 2011.

[9] J. Hays and A.A. Efros. IM2GPS: estimating geographic information from a single image. In *Proc. CVPR '08*, 2008.

[10] G. Friedland, O. Vinyals, and T. Darrell. Multimodal location estimation. In *Proc. MM '10*, 2010.

[11] Martha Larson et al. Automatic tagging and geotagging in video collections and communities. In *Proc. ICMR '11*, 2011.

[12] Börkur Sigurbjörnsson and Roelof Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. WWW '08*, 2008.

[13] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proc. ICMR '11*, 2011.

[14] Yunpeng Li, D.J. Crandall, and D.P. Huttenlocher. Landmark classification in large-scale image collections. In *Proc. ICCV '09*, 2009.

[15] D.M Chen et al. City-scale landmark identification on mobile devices. In *Proc. CVPR '11*, 2011.

[16] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *Proc. CVPR '13*, 2013.

[17] Jing Li et al. GPS estimation for places of interest from social users' uploaded photos. *IEEE Trans. Multimedia*, 15(8):2058–2071, 2013.

[18] Weiqing Min, Changsheng Xu, Min Xu, Xian Xiao, and Bing-Kun Bao. Mobile landmark search with 3d models. *IEEE Trans. Multimedia*, 16(3):623–636, 2014.

[19] Martha Larson et al. The benchmark as a research catalyst: Charting the progress of geo-prediction for social multimedia. In *Multimodal Location Estimation of Videos and Images*. Springer, 2015.

[20] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV '07*, 2007.

[21] Herve Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Proc. CVPR '09*, 2009.

[22] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR '12*, 2012.

[23] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV '03*, 2003.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR '07*, 2007.

2

GEO-VISUAL RANKING

In this chapter, we introduce a generic framework that takes a visual-only, search-based approach to the prediction of the geo-location of social images. The target image is used as a query, and a geo-location is predicted based on the evidence collected from images retrieved from a background collection of images, already annotated with geo-location. The main novelty of the approach is that it leverages evidence from images that are not only geographically close to the target location, but also have sufficient visual similarity to the query image within the considered image collection. Our method is evaluated experimentally on a public dataset of 8.8 million geo-tagged images from Flickr, released by the MediaEval 2013 evaluation benchmark. Experiments show that the proposed method delivers a substantial performance improvement compared to the existing related approaches, particularly for queries with high numbers of neighbors.

This chapter is published as "Global-Scale Location Prediction for Social Images using Geo-Visual Ranking", by X. Li, M. Larson and A. Hanjalic in the *Multimedia, IEEE Transactions on*, vol. 17, no. 5, pages 674-686, 2015.

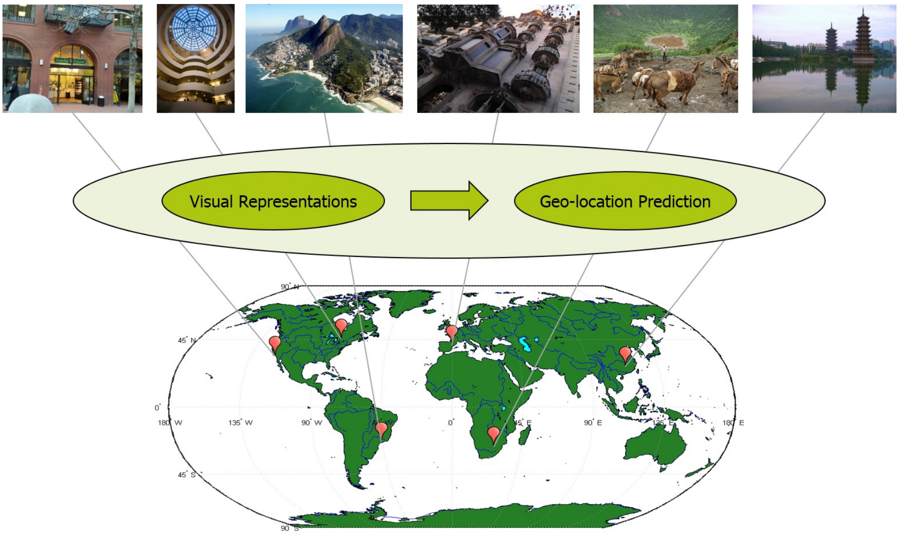


Figure 2.1: An illustration of the challenge of predicting the geo-location of an image automatically at global scale based on its visual content and independently of the availability and informativeness of textual metadata.

2.1. INTRODUCTION

THROUGH rapid development and widespread usage of capture devices such as cameras, phones and tablets, generation of *social images* in recent years has exploded. We define social images as photos that are either taken to share with other users on social media platforms like Flickr[1], or uploaded on such platforms for personal reasons, like memory preservation. In addition to the visual content of images and their textual metadata (e.g., tags), *geo-information*—information about the geographic locations at which they were taken and typically represented by geo-coordinates—is also important for supporting users in searching, browsing, organizing and sharing their photo collections. More specifically, geo-information can assist in finding popular objects and events in a particular area [1], in generating representative and diverse views of a location [2, 3], and in making virtual tours by presenting information mined from user-generated travelogues and photos [4].

While many modern mobile devices make it possible to assign geo-coordinates to images during capture, most social images are still shared and uploaded without this information [5]. As an alternative to manual geo-annotation supported by location-aware interfaces (e.g., the geo-tagging possibilities offered by Flickr), increasing research attention has been devoted to techniques that automatically estimate geo-locations of social images. Such approaches are commonly referred to as *geo-location prediction* techniques [5–9].

Textual metadata that often accompany social images may include place names and other location-specific terms and in this way help inform the geo-location prediction process [6, 8]. For this reason, text has served as the basis for a broad range of proposed geo-location predication algorithms (e.g., [5, 10]). However, the drawback of textual annotation is that it needs to be manually created first by the user. In addition, users adding tags to

---

[1] http://www.flickr.com/

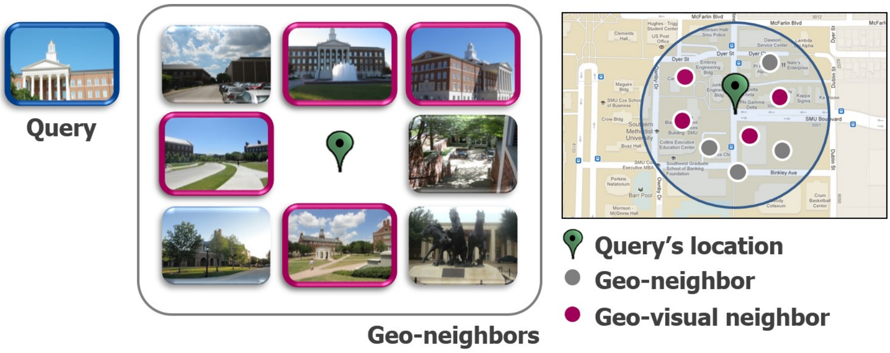


Figure 2.2: Geo-visual neighbors of a query image are images that have both a high visual similarity to the query image and also close geographic proximity to a candidate image.

images does not guarantee that useful location-related information is provided. In other words, although text can be useful for automated location prediction if available and sufficiently informative, uncertainty remains as to its availability and informativeness for an arbitrary image.

Alternative approaches have relied on the visual content of images only. They have the advantage of not depending on the availability of metadata. However, images taken at a location demonstrate a high degree of visual variability. For this reason, it is not surprising that the majority of such approaches narrow the domain of prediction and tackle the task within a geographically constrained area [11–13], or reduce this to specific landmark recognition [14–16].

The concerns about the availability and informativeness of textual metadata led us to set as our mission in this chapter the development of an approach that relies solely on visual representations of images, as depicted in Fig. 2.1. However, in contrast to the majority of related approaches mentioned above, our overall goal is to investigate the possibilities to improve geo-location prediction accuracy and reliability to a substantial extent and to achieve this at global scale, i.e., beyond the current, typically constrained, application scenarios. As indicated in recent surveys on this challenge [17, 18], there are still ample opportunities waiting to be explored. Specifically, the novelty of our work lies in addressing the shortcomings of the existing approaches to global-scale, visual-only geo-location prediction, and results in substantial performance improvement compared to these approaches.

We also note here that the problem of visual-only geo-location prediction at global scale is of larger interest to the multimedia community, since, as this chapter witnesses, it typifies a problem addressable with a search-based, i.e., ranking, solution. Search-based approaches have been heralded as holding promise for tackling large scale image annotation problems [19], and have also been successful for image classification (e.g., [20]). However, despite their simplicity and elegance, they do not provide a one-size-fits-all solution. Rather, much more work is necessary to understand when and why they work, which motivated the detailed research questions addressed in this chapter.

The remainder of the chapter is organized as follows. In the following section, we briefly describe the contribution of the chapter and define the main research questions to which we will provide answers. Then, in Section 2.3, we elaborate on the related work, including how the approach presented here matured with respect to our preliminary efforts. In Section 2.4, we provide a detailed explanation of our proposed method for geo-location prediction. While Section 2.5 details the experimental setup for assessing this approach, Section 2.6 presents and analyzes the results of experimental evaluation, both in terms of the obtained performance with respect to the state of the art and the impact of the availability of the information that we rely on when generating predictions. Section 2.7 concludes the chapter and provides an outlook towards future work.

2.2. RATIONALE AND CONTRIBUTION

In this chapter, we propose a novel method for predicting geo-coordinates of social images at global scale using visual content only. With the proposed method, we specifically address the shortcomings of previous approaches to this challenge, which can be grouped into two main categories.

The first category exploits simple pairwise content-based image similarities, and is characterized by *single nearest-neighbor (1-NN) approaches* [7, 21–25]. These approaches query a geo-tagged image collection and retrieve the image with the highest visual proximity to the query image. Once this image has been identified, its geo-coordinates are propagated to the query image. The main weakness of this approach is its high sensitivity to false positives, that is, to images that were not taken at the query image location, but which are visually similar to the query image nonetheless.

The second category includes *clustering approaches* [7], [26], [24]. Such approaches retrieve images that are visually similar to the query image and group them into clusters based on their geo-coordinates. The geo-coordinates of the centroid of the cluster that contains the most images are adopted as the geo-coordinates of the query image. Compared to the first category of approaches, the underlying idea here is to use more evidence than a single reference image to improve the reliability of inferring the geo-coordinates for the query image. However, in practice, this strategy may also work less well than the 1-NN approach. For instance, if the cluster containing images taken at the query image location (i.e., the true cluster) contains fewer images compared to other clusters, then it will not be chosen, and the inferred location will be incorrect.

To address the weaknesses of these two categories of approaches, we propose a *Geo-Visual Ranking (GVR)* approach. Instead of relying only on the 1-NN image, or on the biggest cluster of visual neighbors of the query image, we search for *geo-visual neighbors* of the query image. As illustrated on the example in Fig. 2.2, geo-visual neighbors are those images that are sufficiently visually similar to the query image and are also taken at the same location as the query image. The advantage of working with geo-visual neighbors is illustrated in Fig. 2.3. This figure shows a query image that is found to be visually similar to two geo-tagged social images taken at different locations (which we refer to as *candidate images*). The 1-NN approach faces difficulty in this situation as the probability of selecting the wrong reference image from the two candidates may be high. Under our approach, however, the selection of one of the two locations is informed by the sets of images (here referred to as *candidate geo-visual neighbors*) found at both locations (here referred to as *candidate locations*). Their contribution to the decision is based not on their number (i.e., the amount of supporting evidence as used by clustering approaches), but on their com-

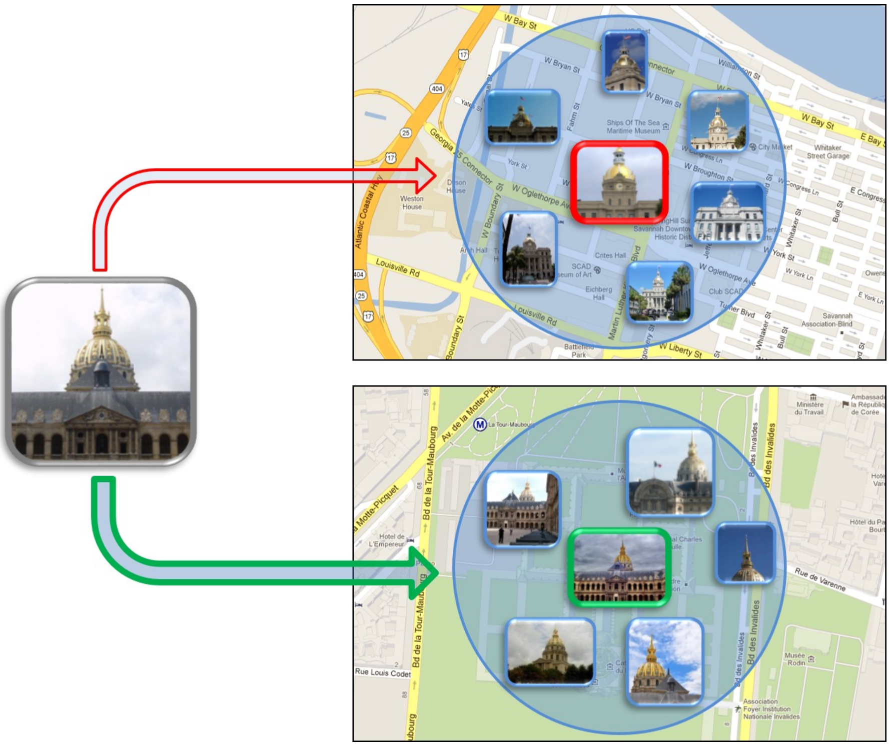


Figure 2.3: The principle underlying geo-visual ranking (*GVR*): The query image (left) matches two geo-tagged candidate images equally well. Each candidate image marks one candidate location and is accompanied by other images similar to the query and taken at that location. All images at a candidate location form the set of candidate geo-visual neighbors of the query image. The incorrect (upper) match is distinguished from the correct (lower) one by assessing the visual proximity of geo-visual neighbor sets to the query image.

bined visual proximity to the query image, aggregated over all images from a set. Use of the set's visual proximity makes it possible to point to the right candidate image (indicated by the thicker arrow in Fig. 2.3), despite the fact that this candidate image has a smaller set of geo-neighbors than the other candidate image. The rationale here is that, compared to the images from the wrong location, more images from the true location will likely contain more elements of the visual content of the query image. We hypothesize that this will make the set of candidate geo-visual neighbors at a wrong location too weak to compete with the set from the true location, independently of the set size.

The method proposed here represents an extension and a substantial improvement over our previous work [22, 23], which documented a first exploration of the idea of geo-visual ranking. In order to mature the initial idea, we improved the visual representation of the images and the image matching strategy. These improvements are critical because they led to a significantly better initial list of candidate images and to an improved set of candidate locations. These improvements are non-trivial because they had a wider impact on the system. Specifically, in order to translate the improvement of the quality of the initial list of candidate images into an increase in the accuracy of geo-location prediction, it was necessary to develop high-performance location extraction and location ranking steps (generation and assessment of geo-visual neighborhoods). In sum, this chapter goes above and beyond our previous work and provides answers to four research questions:

- **RQ1**: Is the *GVR* paradigm conceptually superior to *1-NN* and *clustering* paradigms? (Section 2.6.1)
- **RQ2**: How does *GVR* perform with respect to state-of-the-art methods? (Section 2.6.2)
- **RQ3**: What is the source of relative advantages of *GVR* compared to *1-NN* and *clustering* paradigms? (Section 2.6.3)
- **RQ4**: What is scope of the applicability of the proposed method? (Section 2.6.3)

2.3. RELATED WORK

The challenge of estimating the geo-location of an image using only its visual content has drawn increasing research attention over the past years. Work addressing this challenge has been pursued along two major directions: *geo-constrained* prediction, where the possible locations at which the target image could have been taken are limited to a defined geographic range or a set of predefined locations, and *geo-unconstrained* prediction, assuming that the target image could have been taken anywhere around the globe. We briefly elaborate on the reported achievements in both directions.

2.3.1. GEO-CONSTRAINED CONTENT-BASED LOCATION PREDICTION

Early work on geo-location prediction focused on street-level location prediction. Zhang and Kosecka [11] used a matching technique based on SIFT features [27] to select images with the views closest to the target image. The estimated location is then generated by performing position triangulation on the two best reference views selected by the camera motion estimation. Since the images used were densely sampled along the street, their system could achieve relatively precise prediction: estimation errors were less than 16m. Steinhoff et al. [28] applied fast nearest neighbor search within a collection of photos represented by local image features to achieve realtime location estimation on mobile devices. Experiments were conducted in an urban environment covering an area of a few city blocks. The reported accuracy is comparable to that of a GPS. Chen et al. [14] investigated the problem of city-scale landmark recognition for cell-phone images. They collected 150k panoramic images of San Francisco using surveying vehicles, which were further converted into 1.7 million perspective images. A vocabulary-tree-based retrieval scheme based on SIFT features [27] was built to approach this task. Gronat et al. [13] tried to attack this city-scale location recognition problem from the classification point of view. They modeled each geo-tagged image in the collection as a class, and learned a per-example linear SVM classifier for each of these classes with a calibration procedure that makes the classification scores comparable to each other. Due to the high computational cost in both off-line learning and online querying phases, the experiment was conducted on a limited dataset of $25k$ photos from Google Streetview taken in Pittsburgh, U.S., covering roughly an area of $1.2 \times 1.2 km^2$.

Another group of methods addresses the problem of landmark location prediction. Li et al. [16] proposed an approach to automatically mine popular landmarks from a large-scale

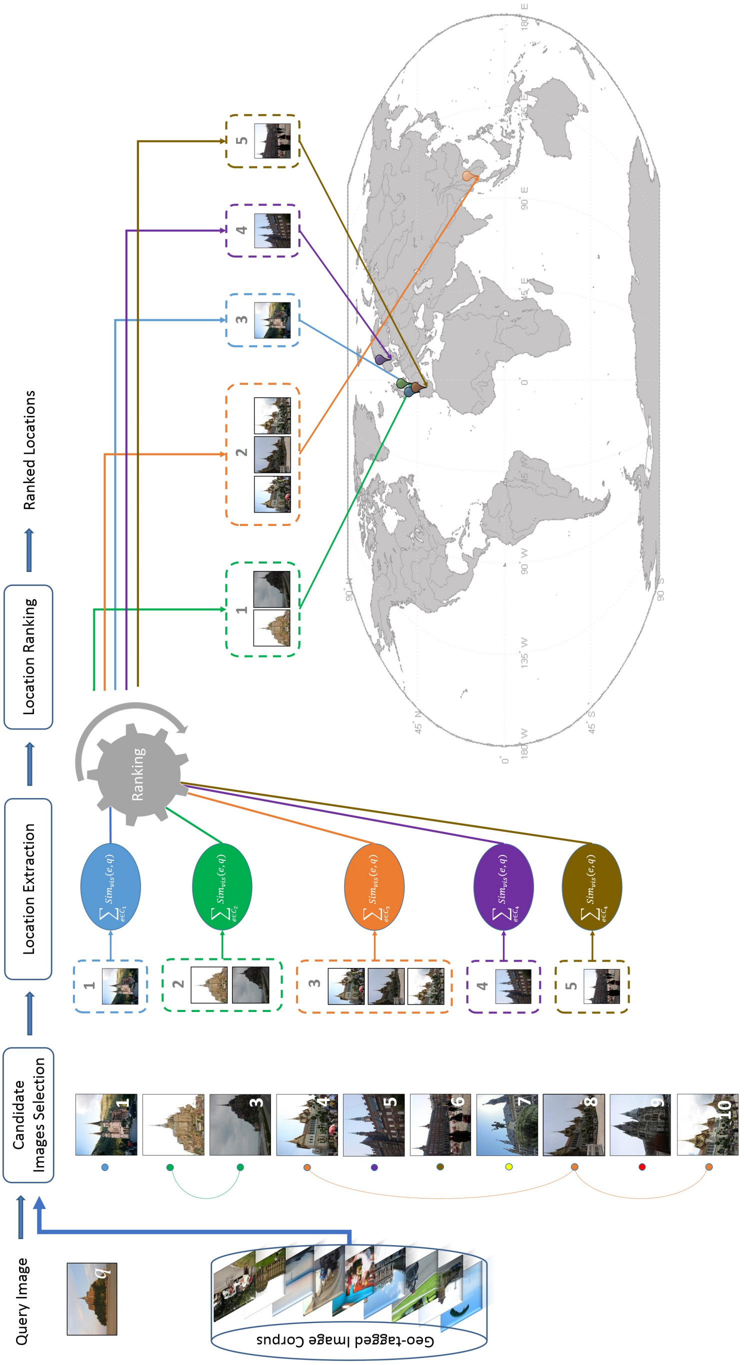

Query Image
q
Geo-tagged Image Corpus
Candidate Images Selection
Location Extraction
Location Ranking
Ranked Locations
1
2
3
4
5
6
7
8
9
10
$\sum_{e \in C_1} Sim_{vis}(e,q)$
$\sum_{e \in C_2} Sim_{vis}(e,q)$
$\sum_{e \in C_3} Sim_{vis}(e,q)$
$\sum_{e \in C_4} Sim_{vis}(e,q)$
$\sum_{e \in C_5} Sim_{vis}(e,q)$
Ranking
0°
45° N
90° N
180° W
135° W
90° W
45° W
0°
45° E
90° E
180° E
45° S
90° S

Figure 2.4: The proposed *Geo-Visual Ranking (GVR)* represented by three main steps: candidate image selection, location extraction and location ranking.

Flickr dataset and perform object recognition using a multi-class support vector machine for top 500 discovered landmarks. Although the landmarks are located all over the globe, this approach is geo-constrained because it limits the prediction to a finite set of locations. Similarly, Kalantidis et al. [15] also mined representative scenes from the geo-tagged photos from 22 European cities and then proposed an approach to estimate the location depicted in the target image by matching it with these representative scenes. Finally, Li et al. [21] proposed a hierarchical method to estimate a photo's geo-location. This approach matches the visual content of an image against a hierarchical structure mined in a set of images from about $1.5k$ predefined places of interest.

One of the main bottlenecks in finding the matching geo-tagged images for the target image is the absence of ground-level reference photos. In order to tackle this problem, Lin et al. [29] introduced a cross-view feature translation approach to learn the relations among three visual aspects: ground-level images, aerial images and land cover attribute images. The experiments performed on a $40km \times 40km$ region around Charlston, U.S., showed the potential of this approach to geo-localize a target image in the absence of geo-visual neighbors. The success of the approach is, however, limited to scenes that correlate well with aerial images and land cover attributes.

The approaches mentioned above served as a source of inspiration for the choice of visual features used in our own approach. The challenge we address is then how to deploy these features effectively for image similarity assessment in a general case, i.e., when the target location is not constrained to a set of predefined locations typically characterized by specific visual scenery elements.

2.3.2. CONTENT-BASED LOCATION PREDICTION WITHOUT GEO-CONSTRAINTS

Compared to the effort that has been devoted to geo-constrained location prediction, there has been relatively less work dedicated to predicting locations at the global scale. This can be explained by the challenge of the task. If we consider all social images that have been taken at arbitrary locations around the world as candidate images representing the target location, the virtually infinite and, consequently, unknown range of the visual content covered by these images makes it difficult to define an effective strategy to assess their correspondence to the query image.

A major contribution in this direction was the approach by Hays and Efros 2008 [7], which we refer to as *MSC* in this chapter. They deployed various global visual representations to model the visual scene similarity between images and employed the Mean Shift Clustering approach to estimate the location. Further contributions can be found among the submissions to the Placing Task of the MediaEval 2013 multimedia evaluation benchmark, which addressed the challenge of location prediction of social images [30]. From those approaches that made use of visual features, we mention here the approach by Li et al. [31], who combined ranked lists of candidate images created using various global visual representations, e.g., color and edge directivity descriptor (CEDD), scalable color (SCD) and border/interior pixel classification descriptor (BIC), to create an overall ranked list. The top ranked candidate image is used as the source of the geo-prediction, making this approach a variant of 1-NN. Kordopatis-Zilos et al. [24] deployed compact visual representations, SURF+VLAD vectors, to calculate visual similarities between images and applied an incremental spatial clustering scheme to find the most probable location. Davies et al. [26] proposed a multimodal version of *MSC* [7] that uses both local and global visual representations, including LSH-SIFT and PQ-CEDD, to obtain different sets of candidate locations. Geo-predictions are generated by selecting the mode with the highest probability. This work was later extended to [32] with incorporation of textual metadata. Our own contribution to the MediaEval 2013 (Li et al. [22]) deploys a combination of local and global visual representations within the geo-visual ranking system originally proposed in [23].

Because the idea behind the clustering approach *MCS* [7] is closest to the one underlying our proposed method, we chose this method as one of the main reference methods in our experimental comparative study. Additionally, we also include our previous work [22] and [23], which represent our own initial exploratory work and served as progenitor to the approach proposed here. Especially [22] is a valuable reference method since it was the best performing visual-only approach at the MediaEval 2013 Placing Task.

2.4. GEO-VISUAL RANKING (GVR)

The problem of predicting geo-location g of a target (query) image q can be seen as the problem of determining the location g among the set of considered candidate locations G, which is associated with the strongest evidence of being the correct geo-location of q. Since we rely on visual information only, we assume that the location at which the image is taken is also reflected in the visual content of the image. We also assume that query q could have been taken anywhere in the world, and that set G does not a priori privilege specific locations over others. In that case, the estimated location $\tilde{g}$ can be found as

$$\tilde{g} = \arg\max_{g \in G} Score(g, q) \tag{2.1}$$

where the function $Score(g, q)$ is defined to quantify the affinity between q and g. We note that relative likelihood of the locations $g \in G$ can also be estimated using a dedicated model that takes into account the domain knowledge in a given use case. This consideration is, however, beyond the scope of this chapter.

Our proposed approach for estimating $\tilde{g}$ is illustrated in Fig. 2.4 and consists of three main steps. In the first, *candidate image selection* step, for a given query image, we first retrieve from the collection of geo-tagged images a ranked list C of candidate images that are most visually similar to the query. Then, in the *location extraction* step, based on geo-distribution of these candidate images, candidate locations are extracted that form the set G. Each location from G is represented by images from the list C that form the corresponding location cluster. As introduced in Section 2.2, we refer to these image sets as sets of candidate geo-visual neighbors of the query image. For a location g, we denote this set as C_g. Finally, $Score(g, q)$ is modeled by the visual proximity between the sets C_g and the query q and is used to rank the candidate locations for the purpose of selecting the most likely (top-ranked) one ($\tilde{g}$) to be adopted for the query image. This last step is referred to as *location ranking*. In the following subsections, we elaborate in detail on each of the steps.

2.4.1. CANDIDATE IMAGE SELECTION

Given a set of geo-tagged images crawled from the web, the goal of this step is to select those images that, based on their visual content, are most likely to have been taken at the same location as the query image. Since this set of candidate images serves as input for

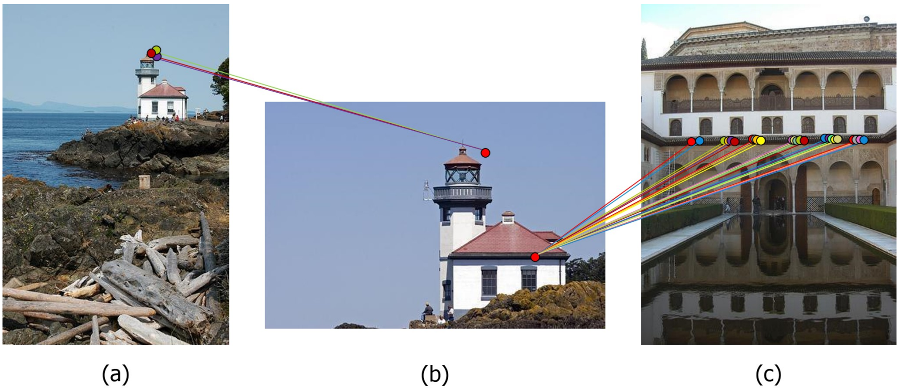


Figure 2.5: Illustration of two cases of invariant region matching between two images. Because we allow multiple matches per region, many matches can be identified, as shown by the links between images (b) and (c). There the lower right region in image (b) has found matches with 17 regions in image (c). On the other hand, for the upper left region in image (b) only 3 matching regions in image (a) were found.

all further steps, the quality of this set is critical for the success of our approach. While we considered different visual features and matching strategies that have been proposed in recent literature ([7, 22, 23, 33–36]), our exploratory experiments led us to develop a more effective methodology that better meets the specific requirements of the geo-location matching task addressed in this chapter.

Conceptually, we search for invariant regions in the images and consider matches between invariant regions of two images as evidence that the images' visual content reflects the same location in the physical world, possibly captured under different conditions, e.g., capturing angle, scale or illumination. In order to identify the invariant regions and assess their matches, we use the standard bag-of-visual-words paradigm, which scales up well to a large-scale datasets [33, 35, 37].

We formulate the visual similarity $Sim_{vis}(e,q)$ between images e and q as

$$Sim_{vis}(e,q) = \sum_{m \in \mathrm{M}} W_m \tag{2.2}$$

where M is the set of matches found between two images. W_m is the weight of each match m. It is computed using the method presented in [34], and is based on the distance of the underlying visual words in the feature space, as well as the geometric consistency of the invariant regions represented by the visual words. In addition, in order to take into account the quantization noise of visual words and to capture the geometric relation between matches, we add Hamming embedding [38], multiple assignment [35, 38] and Hough pyramid matching [34] to our bag-of-visual-words scheme.

Compared to the traditional ways of computing the similarity in Eq. 2.2, we refine the set of region matches to be taken into account by focusing on those that support the expectation of finding the same characteristic scene elements in both images if they are taken at the same physical location. As illustrated in Fig. 2.5, in case of the same location (images (a) and (b)), such elements would be more-or-less uniquely linked to each other. On the other hand, numerous matches found between a region in one image and many regions

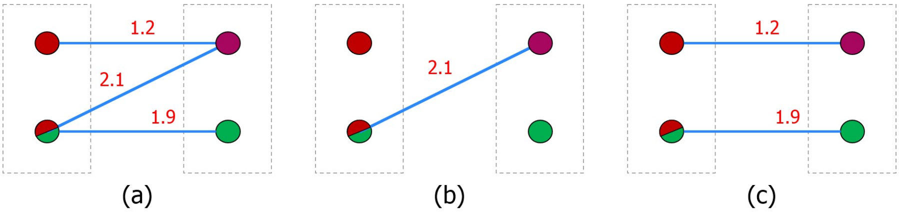


Figure 2.6: Illustration of two different strategies for filtering out multiple alternative region matches between two images. Case (a) shows the original matches generated by the traditional bag-of-visual-words method with multiple assignment on the query side (left image). The lower point in the query image represents a region that matches two different visual words marked with red and green in the right image. Case (b) illustrates the strategy by Jegou et al. [35] that focuses on the strongest matches between the regions in two images. Case (c) is the proposed '1vs1' strategy that balances filtering out of the matches with preserving as many informative matches as possible.

in another image (Fig. 2.5, images (b) and (c)) are typically a consequence of the ambiguities in computing the image similarity and reduce the probability that two images show the same scene. In this way, multiple matches per region may negatively bias the similarity computation process in Eq. 2.2, leading to wrong geo-location prediction. In order to prevent this negative bias, we propose to add the one-to-one mapping constraint on the matched regions between two images to guarantee that one region in image A can only have one matched region in image B, and vice versa. In general, this can be formulated as an assignment problem, where one can minimize the overall distance between two region sets by using the Hungarian algorithm with the computing time in $O(m^3)$ for set with m features [39]. As finding optimal matches is time consuming, one can aim at an approximate solution.

A representative method to implement the one-to-one mapping constraint in image similarity computation was proposed by Jegou et al. [35], who addressed the effect of burstiness of visual words on image matching. They proposed to reduce this effect by choosing the strongest match per region first and then discard all the other matches associated with matched regions. However, as can be seen from Fig. 2.6 (case (b)), this strategy may result in insufficient number of matches for reliable image similarity computation. Compared to this, and as also illustrated by the case (c) in Fig. 2.6, in our refinement approach, that we refer to as '1vs1', we focus on preserving those matches that have the potential to inform the assessment of the relation between two images in terms of their geographic proximity. We first allow the matches between the regions that originally have few matching links assigned (i.e., potential unique matches), and then discard other matches that contain regions belonging to the allowed match. We continue this process until no more regions need to be processed. As indicated by the experimental results summarized in Table 2.1 (Section 2.5), this '1vs1' strategy has the potential to outperform the one of Jegou et al. [35] in a realistic use case.

2.4.2. LOCATION EXTRACTION

Given a ranked list of candidate images, the next step is to derive a set G of candidate locations. Since multiple images from the list C could have been taken at the same location, we propose a method that can gradually build the set G from the geo-coordinates found by moving down the list. If new geo-coordinates are found within the distance d of an already selected candidate location, the geo-coordinates of this location are updated by calculating the centroid of the geo-coordinates of all images at that location, otherwise a new candidate location is created. We set the distance d such to meet the maximum allowed prediction deviation of the system and thus equal to the evaluation radius introduced in Section 2.5.4. We also note that the trivial realization of this approach, namely considering the location of each individual image as the candidate location, leads to the 1-NN approach, which we described earlier and that we will also use later on as one of the baselines in our experimental comparative analysis.

The process of building the set G is steered by two parameters:

- N - the number of top-ranked images in the list C that we consider a reliable set of candidate images, and
- G_{max} - the maximum number of candidate locations that we consider reliable to enter the selection process in Eq. 2.1.

The rationale behind specifying N is to prevent the system from considering images from the list C that visually deviate too much from the query image and, for this reason, may introduce noise into the set of candidate images with which we work. Setting of the parameter G_{max} further helps reach this goal since it prevents that the number of candidate locations becomes unreasonably high. If G_{max} candidate locations are found before the entire top-N part of the list C has been exploited, then no further candidate locations are created. In that case, the size of the set G becomes equal to G_{max} and we only allow already found locations to be further populated by going further down the top-N list. Alternatively, if less than G_{max} candidate locations are found in top-N images, then we work with this smaller number of locations only. Setting of the parameters N and G_{max} will be discussed in Section 2.6.

2.4.3. LOCATION RANKING

The step explained above could already be deployed to generate a ranked list of candidate locations, for instance by linking the rank of each candidate location to the rank of its image positioned highest in the list C. However, this would make the *GVR* approach conceptually equal to the 1-NN category of approaches and would prevent it from making use of all the available information derived from the geo-visual context of the candidate locations and, consequently, from making more reliable predictions. We therefore allow all of the images in the set C_g to contribute to the cumulative visual proximity of C_g to the query q and compute the $Score(g,q)$ determining the rank of the location g as:

$$Score(g,q) = \sum_{e \in C_g} Sim_{vis}(e,q) \tag{2.3}$$

We note here that another possibility for ranking score computation would involve normalizing the sum in Eq.2.3 by the size of the C_g set. A priori, it seems that such an approach might potentially help in the situations in which an incorrect location is more heavily populated by images (e.g., many images of popular landmarks) than the true location. In such a case, if the images on the wrong location are not significantly dissimilar from the target

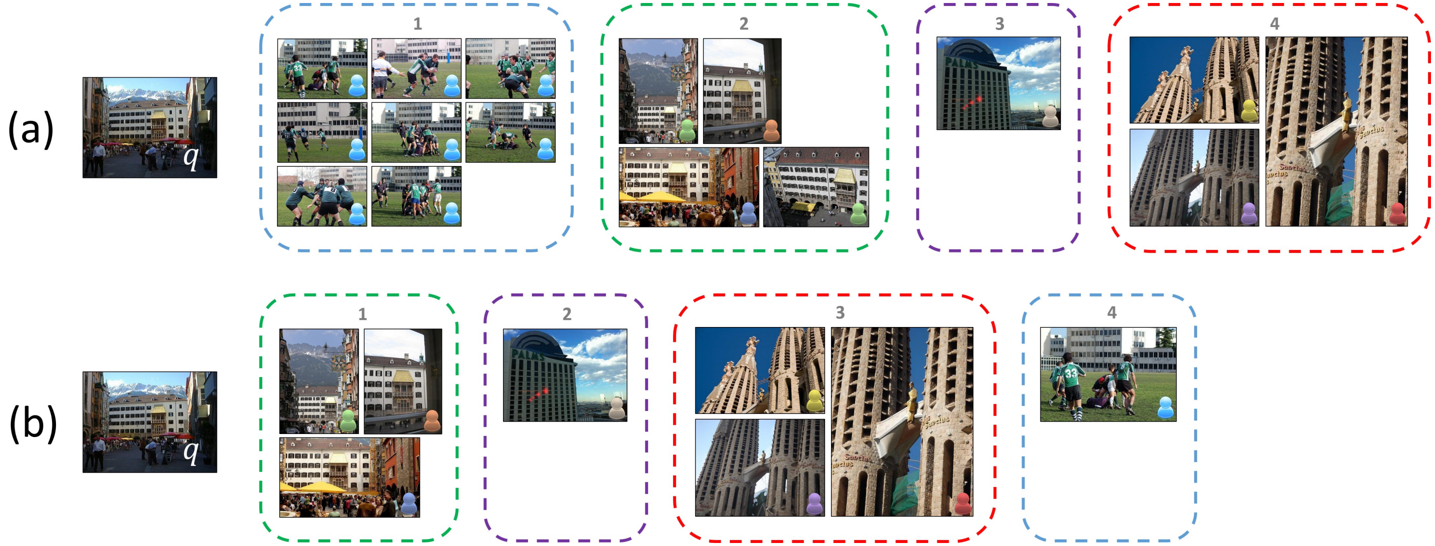


Figure 2.7: Illustration of the effect of high-volume uploads on the *GVR* method performance. Photos contributed by the same user in one geo-visual neighborhood are marked with the same color of the person icon. Lists (a) and (b) are, respectively, the ranked lists of the candidate locations without and with the constraint regarding high-volume uploads. Applying this constraint enables the true location to be ranked first in the list (b).

image, Eq.2.3 would allow the incorrect location to outrank the correct location. However, such a normalization also has the potential to make negative impact. In particular, in the case that the correct location has more geo-visual neighbors than the incorrect one, but if the visual match of the images at the correct location is insufficiently strong, then the incorrect location could possibly outrank it. Since it is difficult, if not impossible, to make a reliable prior estimate about which of the two cases would dominate in the actual geo-prediction problem, we tested this experimentally by comparing the performance of *GVR* using both the Eq.2.3 and its normalized version. The performance is computed on our test data collection as a part of our overall experimental study that is explained in detail in sections 2.5 and 2.6. The results showed that the score computation using Eq.2.3 leads to 25% improvement in *GVR* performance, compared to its normalized version. This indicates that Eq.2.3 better reflects the phenomena present in a typical large-scale Internet image collection related to geo-location prediction. On the basis of this result, we confidently adopt Eq.2.3 for score computation as a part of our *GVR* framework, and do not consider its normalized version in the remainder of this chapter.

2.4.4. REDUCING THE EFFECT OF HIGH-VOLUME UPLOADS

The reliability of the candidate location list can be negatively influenced by the tendency of social media users to upload many images taken at the same location, for instance those related to a specific event attended and intensively photographed by a user. High-volume uploads of individual users damage prediction because they lead to a disproportionately high number of images in the set C_g of a false location, which may overwhelm the otherwise lower visual similarity between images in that set and the query compared to the true location. For this reason, when applying Eq.2.3, a false location can be found to match the query better than the true location. An illustration of this case is given in Fig. 2.7 using the true example generated with the method presented in this chapter.

In order to reduce the negative effect of high-volume uploads, we introduce a constraint that requires that sets C_g contain at most one image from any given user. In addition to handling the high-volume upload problem, we point out that this constraint also promotes the independence in the evidence (i.e., geo-visual neighbors have been captured by different users, and also during different photo-taking events) that contributes to assessing candidate locations, leading to a more robust prediction. The positive effect of this constraint is also illustrated on the example in Fig. 2.7.

2.5. EXPERIMENTAL SETUP

In this section, we describe the setup of our experimental framework for assessing the performance of the proposed *GVR* method. In the following subsections, we will elaborate on the details regarding all aspects of this framework, including the dataset we used, features we selected to measure visual similarity of images, reference methods from literature that we deploy for comparative analysis and the assessment criteria.

2.5.1. DATASET

To assess the performance of the proposed *GVR* method, we carry out experiments on an image collection that is based on the dataset released by the MediaEval 2013 Placing Task [30]. To create this dataset, geo-tagged images were randomly selected from Flickr, but in a way that maintained the global geographic coverage and retained the original user structure within the online image sharing network. Since the dataset release included only the metadata and not the images themselves, we re-crawled Flickr to collect the images using the links in the metadata. Because some images were removed from Flickr after the dataset was collected, the final collection we worked with contained 8,799,260 images.

We adopt the same training/test set split as used in the MediaEval 2013 Placing Task. There, 261,892 images served as test queries, leaving 8,537,368 as the training set used to generate the predictions. The split was created such that the set of users who contributed the test images were excluded from the set of users who contributed the training images. This constraint makes it impossible for the algorithm to leverage the fact that a single user often uploads (near-) duplicate images. It makes the task more challenging, but also ensures that it is more realistic. We further divided the test set of 261,892 images into two partitions, 10% serving as a development partition, which we use to tune parameters, and the rest of 90% serving as a test partition.

2.5.2. COMPUTING VISUAL SIMILARITY

In our experiments, we compare the cases in which image similarity is computed using local and global features, either separately or in combination. We elaborate on the deployed local and global image representations and related image matching strategies in more detail in the following subsections.

IMAGE SIMILARITY BASED ON LOCAL FEATURES

In choosing local feature-based image representation, we followed a standard line of reasoning. Since SURF [40, 41] has been reported to be faster and more compact than SIFT [42], we use SURF to find and describe invariant regions in the image. To further speed up the retrieval and improve the accuracy, we also adopt the state-of-the-art technique proposed in [38], which represents subregions of the feature space by signatures, and compares descriptors not only based on their visual words, but also based on the distance between their subregions within the feature space. To address the quantization noise introduced by visual word assignment, we adopt the strategy used in [35, 38], which assigns a given descriptor to several nearest visual words. As this multiple assignment strategy significantly increases the number of visual words per image, e.g., on average 4.2 visual words per descriptor, we only apply this at the query side.

We deploy the BoofCV software to extract SURF descriptors with default parameters and use exact k-means to cluster these descriptors and generate visual words. As described in [33, 34], the bag-of-visual-words-based system can have a different performance depending on whether the visual words vocabulary is trained on an image set with or without test queries, i.e., whether the vocabulary is *specific* or *generic*. To mimic the situation in a real retrieval system, we use a separate set of $50k$ randomly selected images from Flickr to train the *generic* $20k$ vocabulary set and use it in all experiments.

To select the strategy for image matching to work with, we did a preliminary experimental study using the feature-extraction procedure and system setup mentioned above and involving the proposed '1vs1' and two alternative image matching strategies from [35] and [43]. Since the image matching problem stated in Section 2.4.1 is closely related to object retrieval, namely the problem of finding the images containing the same objects or scene elements as in the query image, we tested these implementations against two standard datasets used for this purpose, namely the *Oxford* dataset [33] and the *Holiday* dataset [38], with or without 1M randomly selected images from Flickr as distractors. The mean average precision (mAP) for different datasets is reported in Table 2.1.

Compared with [35] and [43], '1vs1' achieved comparable performance on generic vocabularies and even outperformed these alternative methods slightly. Note that the high performance achieved by [43] on the *Oxford* dataset is mainly due to the specific features optimized for unrotated images. However, this gain is at the cost of worse performance for rotated images, e.g., on the *Holiday* dataset. As we do not restrict the geo-location estimation task to unrotated images only, we particularly focused on the performance on the *Holiday* dataset, which contains images that have undergone various transformations related to rotations, viewpoint change and blur. The *Holiday* dataset also includes a large variety of scene types (e.g., not only buildings, as in the *Oxford* dataset, but also landscape, animal, flowers and indoor scenes), better indicative for the strategy to be selected. The results in Table 2.1 made us adopt the proposed '1vs1' strategy to implement the candidate image selection step using local features, which will be used in all further experiments reported in this chapter.

As a side remark, we note that more recent alternative strategies for image matching using visual words have been proposed in [34] and [33]. We also compared '1vs1' with these strategies and the performance was rather close (slightly better than [34] and slightly worse than [33]). However, we did not report these results in Table 2.1 because these two alternatives were assessed on *Oxford* dataset only, which we found too limiting for our comparative analysis.

Table 2.1: mAP comparison on *Oxford* and *Holidays* for generic vocabularies

	1vs1	[35]	[43]
$Holidays$	**0.879**	0.848	0.780
$Holidays+1M$	**0.820**	0.791	–
$Oxford$	0.657	0.685	**0.822**
$Oxford+1M$	**0.571**	0.542	–

IMAGE SIMILARITY BASED ON GLOBAL FEATURES

To compute visual similarity based on global visual features, we chose GIST representation [36], which has been shown to be effective in retrieving semantically and structurally similar scenes [7, 44]. For each image, we resize it to 375 × 500, which is the most common image size in the collection, and then create the GIST descriptor in 5 × 5 spatial resolution with 4 scales and 6 orientations at each scale. After this, each image is represented by a 600 dimensional vector and we use L2 distance to compare these vectors. These settings are the same as in *MSC* [7], which we will introduce as one of our reference methods in the following section.

2.5.3. EXPERIMENTAL COMPARISON

We assess the proposed *GVR* method through a comparative experimental analysis, which we perform in three stages. In the first stage, we compare the performance of our proposed

GVR approach with two other categories of methods introduced in Section 2.2. For each category, we select one representative method which is not restricted to specific regions or landmarks. These methods are:

- *VisNN*: Our implementation of the 1-NN approach, which uses the geo-location of the image visually most similar to the query image as the predicted location. As already indicated in Section 2.4.2, we deploy here the trivial realization of our method for location extraction, where the location of each individual image in the list C of candidate images is considered as one candidate location. Choosing for our own implementation here allows us to experiment with both local and global image representations when selecting candidate images. In this way, we are able to compare this method with GVR consistently and fairly.

- *MSC*: Method used in [7], which performs mean-shift clustering on the geo-locations of the top N images most similar to the query and then ranks these clusters by their size. The centroids of the ranked clusters are used as the predicted locations. Different from the original work in [7], where only global image representations were used, we evaluate this clustering method using both local and global image representations. Again, this allows us to make a fair comparison of this method with *VisNN* and *GVR* since the visual image representations used are consistent. In addition, as the kernel bandwidth of mean-shift clustering used in *MSC* is tuned for coarse-grained location estimation, i.e., city level, with an estimation range of $25km$, we use much smaller bandwidths to maximize the performance with respect to the required prediction deviation.

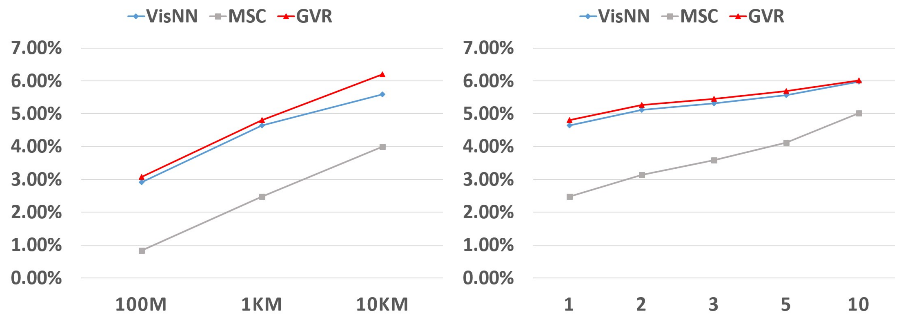
As the main result of the first stage, we will show that *GVR* outperforms the other two alternatives when local, global and combined local-global feature representations are used. The results of this stage are reported in Section 2.6.1 and serve to provide an answer to **RQ1**, i.e., investigation of the validity of the proposed *GVR* paradigm.

In the second stage, and as already indicated in Section 4.3, we compare *GVR* with our previous work [22] and [23]. The goal of this experiment is twofold. First, it will help us gain more insight into the effect of combining local and global features. Second, since the method [22] was the best performing visual-only approach at MediaEval 2013 Placing Task, this comparison will also provide an answer to **RQ2**, i.e., what the improvement of *GVR* method is with respect to the state of the art.

In the third stage, we investigate the relative advantages and applicability of *GVR*, by analyzing the dependence of the performance of *GVR*, *VisNN* and *MSC* on the number of geo-visual neighbors that are available at a location and exploited to generate predictions. The results of this stage are reported in Section 2.6.3 and address **RQ3** and **RQ4**, i.e., reveal the reasons for the relative performance among the three methods, and also provide indication of the applicability scope of the proposed *GVR* method.

2.5.4. EVALUATION PROCEDURE

To evaluate the performance of the proposed system, we adopt the procedure standardly used in the literature. We start by defining an evaluation radius r_{eval}. This radius controls the evaluation precision and the tolerance to data noise in the ground truth, which is generated by a GPS device or through manual labeling. An image is considered to be correctly



(a) (b)

Figure 2.8: General performance on local features. (a) HR@1 with respect to different evaluation radiuses, (b) HR@k performance for varying k and for the evaluation radius of $1km$.

predicted if its predicted geo-coordinates fall within r_{eval} around the ground truth location. Formally expressed, the correctness of an image with respect to an evaluation radius r_{eval} is calculated by the evaluation function $f_{r_{eval}}$,

$$f_{r_{eval}}(g,\tilde{g}) = \begin{cases} right\,,\ geoDist(g,\tilde{g}) \le r_{eval} \\ wrong\,,\ otherwise \end{cases} \tag{2.4}$$

where $geoDist(g,\tilde{g})$ designates the geographical distance between g and $\tilde{g}$.

We use the Hit Rate at top K ($HR@K$) as the criterion to assess the quality of prediction. Given a query, the system returns a ranked list of possible locations. Then, $HR@K$ measures the proportion of queries that are correctly located in the top K locations. Specifically, $HR@1$ represents the ability of the system to output a single accurate prediction.

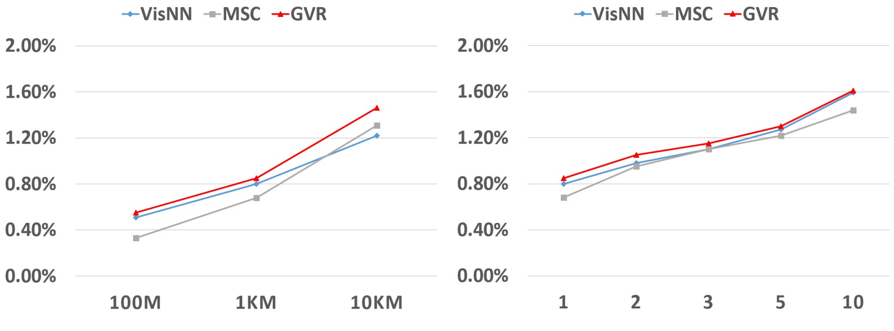
2.6. EXPERIMENTAL RESULTS

We implemented our *GVR* framework by constructing a Map-Reduce-based structure on a Hadoop-based distributed server containing 90 nodes with 8 cores each. The overall run time to build the initial visual rank (the candidate image selection step) for all $261k$ queries on a dataset of 8.8 million photos is about 91 hours, which corresponds to about 1.26 seconds per query image. The overall run time for location extraction and location ranking for all $261k$ queries is 5 minutes, which is 2 minutes faster than for the *MSC* method.

In this section, we report the results of our experimental study, which compares our *GVR* method with the reference methods. We deploy different settings using local and global features for image representation, both separately and in combination, and in the three stages defined above.

2.6.1. STAGE 1: COMPARATIVE ANALYSIS OF GVR, VISNN AND MSC

We use our development partition to tune the parameters of *GVR*. We set two parameters: the number N of top-ranked images in the list C and the maximum allowed number of candidate locations G_{max}, both defined in Section 2.4.2. The tuning is performed per experiment (i.e., separately for local, global and combined local and global features) using grid



(a) (b)

Figure 2.9: General performance on global features. (a) HR@1 with respect to different evaluation radiuses, (b) HR@k performance for varying k and for the evaluation radius of $1km$.

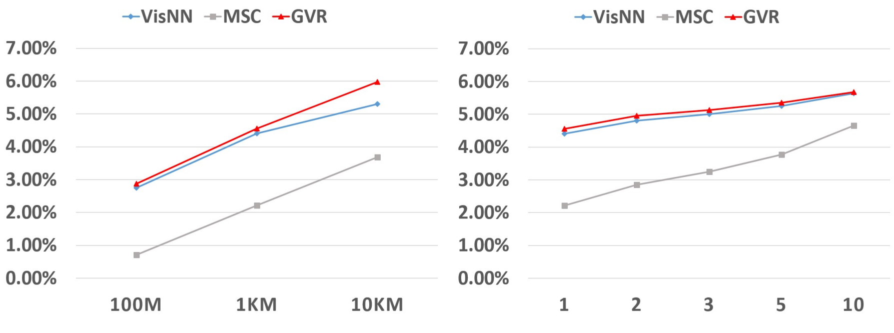
search, through which the best parameter combination is found. For both G_{max} and N we considered the values 10, 20, 50 and 100 and only those combinations where $N > G_{max}$. In the second and third stage we used the optimized best-performing version of *GVR*.

In order to have a fair comparison with *GVR*, we also searched for the optimal value of N for *MSC*. We noticed, however, that the performance of *MSC* continuously decreases for the increasing value of N, never outperforming either of the other two methods. We therefore adopted the same value of N for *MSC* as the one that was found optimal for *GVR*. While the value of N does not affect the general conclusion regarding the relative performance of *MSC*, using the same value of N for both *MSC* and *GVR* proved to be beneficial for the analysis we perform in the next section as it helps identify the reasons for the relative performance among the three methods.

Fig. 2.8 shows the performance of different methods using local features with different values of r_{eval} (Fig. 2.8.a) and different hit rates (Fig. 2.8.b). This figure reveals that *GVR* consistently outperformed both *VisNN* and *MSC* across the board. The average gain in performance was 5% over *VisNN* and 80% over *MSC*. The parameter combination found optimal for this experiment was $G_{max} = 20$ and $N = 100$.

The performance of different methods based on global features is illustrated in Fig. 2.9. In general, all three methods performed significantly worse compared to the previous experiment where local features were used. This indicates that global features alone may not be discriminative enough for location prediction. Although global features such as GIST are known to be effective in retrieving semantically and structurally similar scenes [7, 44], scenes like beaches and forests can appear at many places around the world and have similar general visual characteristics for different locations. In this case, even though the scenes can be matched well by their category, it is difficult for global features to pinpoint one precise location. The parameter combination found optimal for this experiment was $G_{max} = 10$ and $N = 20$.

In view of the results obtained using local and global features separately, we further explored whether they could be combined for more reliable location prediction. For this purpose, we expand the scheme of our *GVR* approach by an additional ranking step. We first follow the procedure from Section 2.4.1 to create the ranked list of candidate images.



(a) (b)

Figure 2.10: General performance on the combination of local and global features. (a) HR@1 with respect to different evaluation radiuses, (b) HR@k performance for varying k and for the evaluation radius of $1km$.

This list is generated based on image comparison using local features. We then take the t top ranked images from this list and rank them again using global features. In the end, we select the top N ranked photos as the final selected candidate set. Although different reranking methods can be deployed here, we chose this simple concatenation of two ranking steps in order to prevent that images too low in the list C are moved to the top of the reranked list. Our procedure can, however, still be seen as reranking, since the initial ranking influences the set of images that propagate to the next step. The prediction results for three methods are reported in Fig. 2.10. Compared to the results in Fig. 2.8, combining local and global features slightly underperforms with respect to using the local features only. The parameter combination found optimal for this experiment was $t = 300$, $N = 100$ and $G_{max} = 20$.

While the performance of all three methods varies depending on the deployed visual features, *GVR* was found to be superior to other two in all three experiments. This allows us to answer positively to **RQ1**.

2.6.2. STAGE 2: COMPARATIVE ANALYSIS OF GVR AND STATE OF THE ART

To investigate the effect of combining the local and global features in more detail, we conducted a comparison between *GVR* variants using the local only and local and global features, referred to as *Local*+ and *Global&Local*+, respectively, and our initial work on geo-visual ranking reported in [23] and [22], referred to as *Local* and *Global&Local*, respectively. The notation + is used to indicate that the method variant deploys the ‘1vs1’ matching constraint introduced and justified in Section 2.4.1. Methods *Local* and *Global&Local* can therefore be seen as the counterparts of *Local*+ and *Global&Local*+ as there no sophisticated image matching is deployed.

If we compare the methods using local features only to those relying on combination of local and global features (Fig. 2.11), in contrast to the positive gain of about +18% achieved by *Global&Local* over *Local*, the gain achieved by *Global&Local*+ over *Local*+ is negative, −6%. This indicates that combining local and global features can only provide added value if global features compensate for the suboptimal performance of local features. By applying ‘1vs1’ matching, some top-ranked, but incorrect candidate images, that potentially could

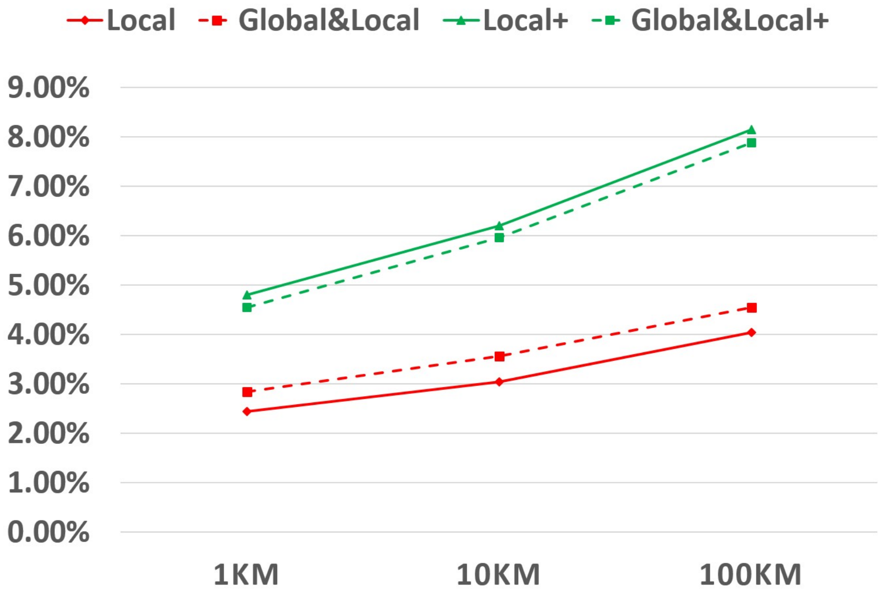
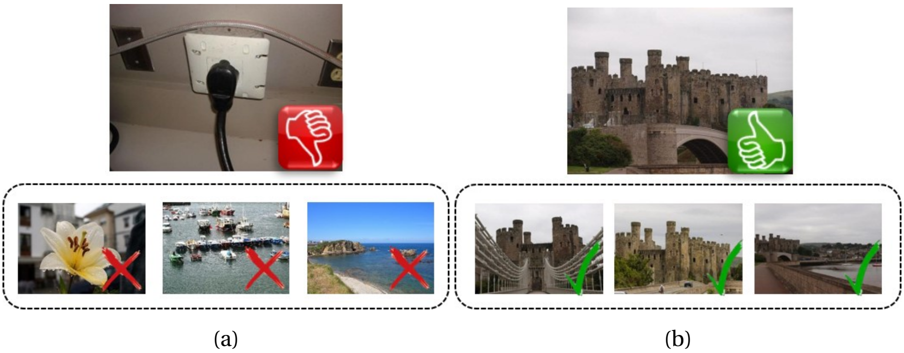


Figure 2.11: Comparison of the performance of four methods in terms of HR@1 with respect to different evaluation radiuses.

have been filtered out using global features, have already been removed. Since global features significantly underperform compared to local features in a general case, deploying global features to this optimized first step is then likely to make the end result worse. For example, a good candidate photo may only contain one part of the scene captured by the query and therefore have a different global image representation. In this case, although this photo may be ranked at the top based on local features, it could be filtered out in the second step due to wrong interpretation based on global features.

The comparison in Fig. 2.11 also provides another insight, namely on the improvement the proposed *GVR* approach in its most successful variant (*Local+*) achieves with respect to state of the art. We consider here the category of methods that rely solely on the social images and their visual representation. In this category, the state of the art is represented by our method, *Global&Local*, from [22] from the MediaEval 2013 Placing Task. For the evaluation radius of $1km$, *Local+* results in the performance gain of about 69% compared to *Global&Local*. This provides an answer to **RQ2**.

The fact that *Local+* is the best performing method highlights the contributions made in this chapter and consisting of image matching using '1vs1' strategy from Section 2.4.1 and the improvements in Location Extraction and Location Ranking introduced in Section 2.4, but also of reducing the effect of high-volume uploads as explained in Section 2.4.4 and illustrated in Fig.7. Regarding the latter, we compared the performance of *Local+* with and without this step. This experiment reveals that controlling for high-volume uploads (i.e., the version of *Local+* in Fig. 2.11) introduced a increase of 4.2% in HR@1 at $r_{eval} = 1km$.



(a) (b)

Figure 2.12: Illustration of two query photos and geo-tagged photos taken at the same location as the query. (a) there is no visual link between query photo and photos taken in the same location. (b) within the query's location, there are photos which have captured similar visual content as the query.

2.6.3. STAGE 3: APPLICABILITY SCOPE

The generally low prediction performance reported in Section 2.6.1 is due to the fact that a large majority of queries do not have any visually similar images taken in their geo-neighborhood. The incorrect location prediction for these images pulls the overall performance of all three compared methods down significantly. As illustrated in Fig. 2.12, if the query has no geo-visual neighbors, the probability increases that it will find visual matches from a wrong candidate location rather than from the right one, which leads to an incorrect location prediction. In contrast, a sufficient number of geo-visual neighbors increases the probability of finding the right visual match between the true location and the query, which again helps the method to make the right prediction. In this section, we zoom in onto the subset of the queries from our test set for which there is at least one geo-visual neighbor available. For these cases, we perform the comparative analysis again between the three methods from the previous section, with the objective of assessing, in a more reliable fashion, the ability of each of them to use the information that is available in the training set in order to make the right prediction.

In order to perform this experiment, we first need to find this query subset. Manual inspection of our data for this purpose would be too tedious and time consuming: we have $261k$ queries, where 35% of them have more than 1000 geo-neighbors. Instead, we used the proposed method for image similarity computation (Section 2.4.1), which deploys local image representation and the '1vs1' matching strategy, to help us create a rough judgment about geo-visual neighborhoods of our queries. For each query, we inspected the top-N ranked list of visually similar geo-tagged images and counted those found within the same radius from the query location as defined in Section 2.4.2. After experimenting with different values of N, we selected $N = 1000$ as the most suitable value. As illustrated in Fig. 2.13, we found that only about 20% of the queries had at least one geo-visual neighbor.

Fig. 2.14 breaks down the geo-location prediction performance across query groups characterized by different numbers of geo-visual neighbors. Apart from the fact that the performance of all three methods increases dramatically compared to that reported in previous experiments, an interesting pattern becomes visible that reveals the main value of the proposed *GVR* method. Note that this is the value already alluded to in the description

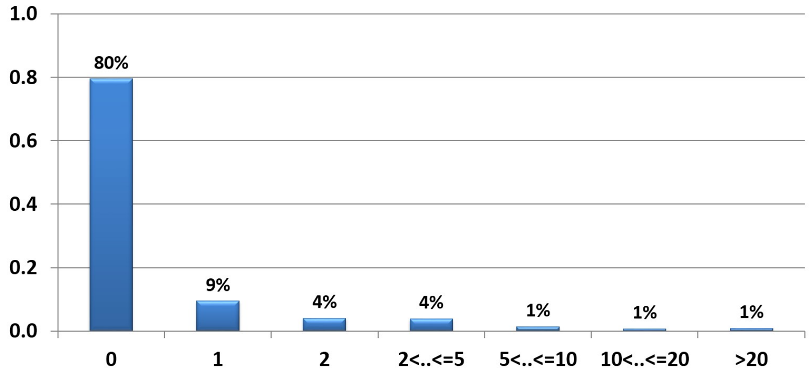


Figure 2.13: Distribution of queries over different numbers of geo-visual neighbors calculated over the social image collection.

of the rationale in Section 2.2.

If we compare the performance of *VisNN* and *MSC* across the range, we see that *VisNN* outperforms *MSC* for low numbers of geo-visual neighbors, while the opposite holds for higher numbers of geo-visual neighbors. This can be explained by the fact that *MSC* selects the candidate location with the largest cluster of candidate images as the predicted one. Then, the probability that the true location is selected decreases with the decreasing number of geo-visual neighbors of the query. In parallel, the probability increases that the *MSC* performance becomes lower than that of *VisNN*, which only needs to match a single geo-visual neighbor for the prediction. On the other hand, if there are more geo-visual neighbors, the available visual evidence makes *MSC* more reliable than the single-image evidence *VisNN* is based on. Furthermore, the performance distribution in Fig. 2.14 shows that *GVR* mimics the best-performing method in the given geo-visual neighborhood context: for few geo-visual neighbors, *GVR* performs like *VisNN*, while it becomes equivalent to *MSC* for more available geo-visual neighbors. This analysis explains the reasons underlying the relative performance among the three methods, providing an answer to **RQ3**.

In addition, for queries with middle level number of geo-visual neighbors, *GVR* outperforms both *VisNN* and *MSC*. This makes *GVR* best capable of making the most use of the available information in the given query context. As an illustration, Fig. 2.15 gives one true example using the result generated with these three methods for one query with middle level number of geo-visual neighbors. As the answer to **RQ4**, concerning the applicability of our proposed approach, it can be concluded that *GVR* can effectively be applied across the range of the size of geo-visual neighborhoods. *GVR* contrasts in this respect with both *VisNN* and *MSC*, which are best applicable solely for few or many available geo-visual neighbors, respectively.

2.7. CONCLUSION

We have presented a geo-visual ranking approach addressing the challenging task of predicting geo-locations of social images using only the visual content of images. The main contribution of the approach is that it improves over two major classes of previous ap-

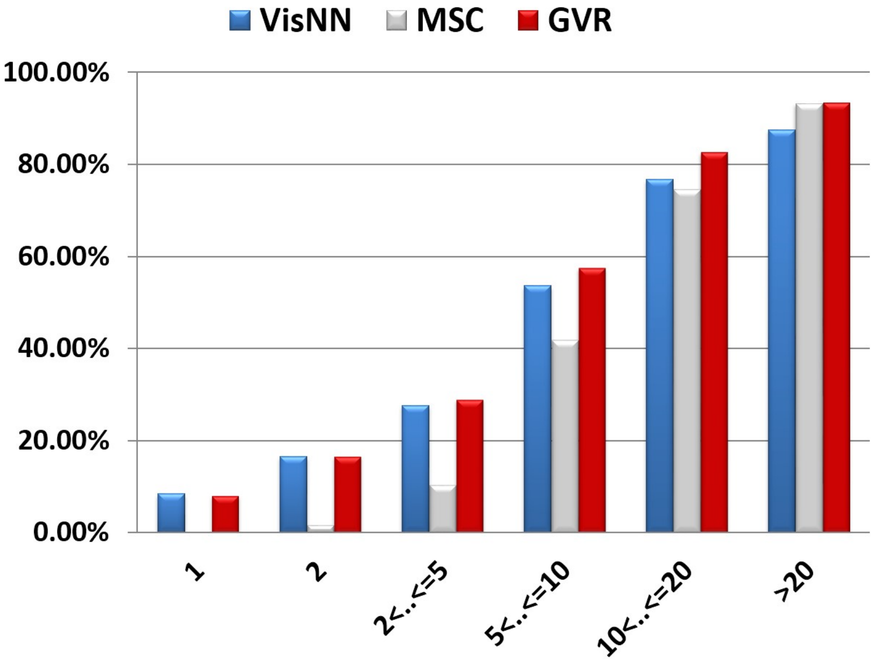


Figure 2.14: HR@1 with $r_{eval} = 1km$ for queries with different numbers of geo-visual neighbors.

proaches, addressing the disadvantages of both 1-NN and clustering. We carry out evaluation using the publicly available dataset from the MediaEval 2013 Placing Task containing 8.8M images. This data set does not set specific focus on frequently photographed areas or on a limited set of locations or landmarks. It therefore allows us to evaluate our approach as a geo-unconstrained prediction approach, i.e., given a photo, it predicts a location anywhere in the world. Compared with other methods that have been proposed to tackle the same problem, and evaluated on the same dataset, the proposed *GVR* method achieves sound performance for geo-location prediction and significantly outperforms state of the art in its approach category. The performance is especially high for images with many geo-visual neighbors in the collection. Crucially, however, the *GVR* approach also retains reliable performance for queries with a low number of geo-visual neighbors, which are highly problematic for the clustering-based *MSC* approach.

In terms of the roles of different visual representations of images for location prediction, we find that although global features such as GIST are known to be efficient for retrieving semantically and structurally similar scenes, it is challenging to exploit them to improve the prediction performance. We attribute this fact to the weakness of the relationship between scene types (which GIST is known to differentiate well) and specific locations. Because it is difficult for global features to pinpoint one precise location, it is also difficult to exploit them for geo-prediction. In contrast, local representations can establish stronger links between photos taken at one particular location, and can, in this way, generate relatively reliable prediction, exceeding the ability of global representations. The effectiveness of local

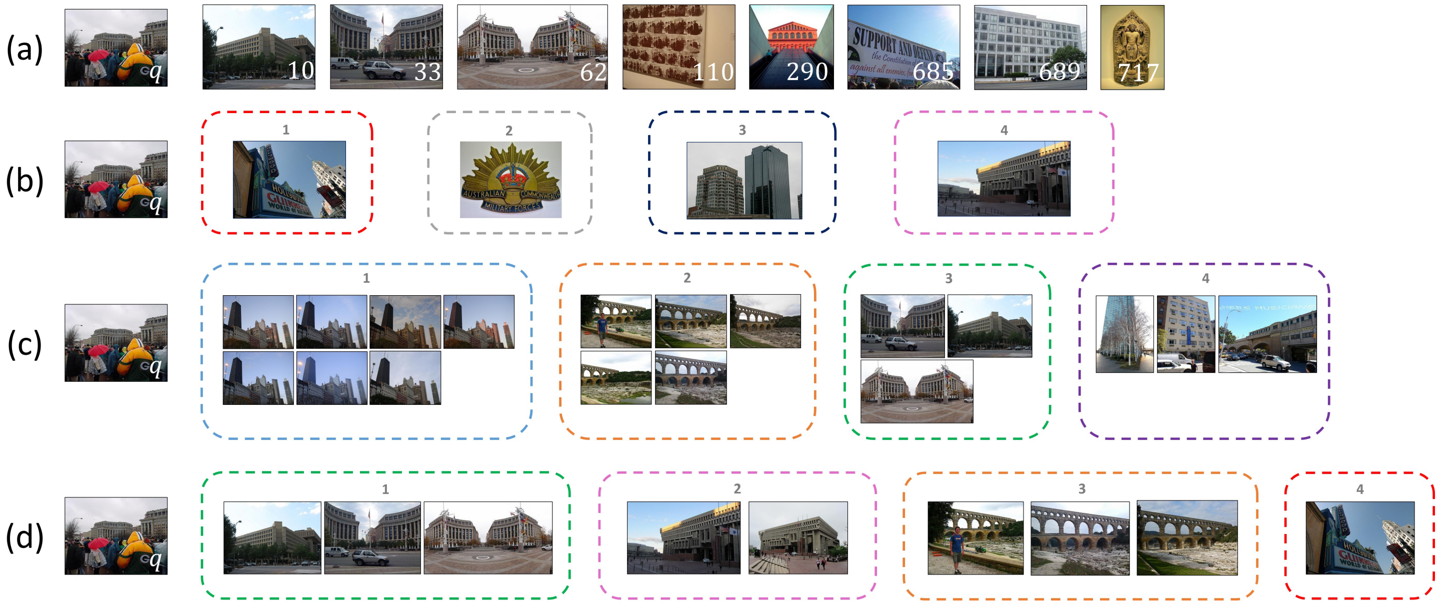


Figure 2.15: An illustration of the relative performance among the methods *VisNN*, *MSC* and *GVR*: (a) 8 geo-visual neighbors found by the system for a given query, (b) ranked candidate locations using *VisNN*, (c) ranked candidate locations using *MSC*, (d) ranked candidate locations using *GVR*.

features is also further improved if they are combined with sophisticated image matching strategies, like the '1vs1' strategy proposed in this chapter.

We note in closing that this chapter has shed light on the difficulty of the problem of geo-location prediction for social images. Fig. 2.12.a illustrated the problem of query photos for which there is no visual connection to the photos in the social image collection taken at the correct location. Fig. 2.13 reveals that there are only ca. 20% of the photos in our collection for which we can find at least one geo-neighbor. The converse problem is also worth consideration, namely, the occurrence of visually similar images taken at different locations. For example, a view onto an unbroken expanse of desert can be shot at multiple locations on the surface of the earth. Taking the next step forward in geo-location prediction for social images involves determining to which extent these issues characterize large collections of images taken by users. Our future work will move in this direction. Specifically, it will include investigation of geo-visual diversity and ambiguity in how visual content reflects the locations at which images were taken. We will explore these issues in greater depth to arrive at insight that would help us to further improve geo-location prediction of social images.

REFERENCES

[1] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proc. SIGIR '07*, 2007.

[2] Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proc. WWW '08*, 2008.

[3] Jiajun Liu et al. Presenting diverse location views with real-time near-duplicate photo elimination. In *Proc. ICDE '13*, 2013.

[4] Qiang Hao et al. Travelscope: standing on the shoulders of dedicated travelers. In *Proc. MM '09*, 2009.

[5] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing Flickr photos on a map. In *Proc. SIGIR '09*, 2009.

[6] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proc. WWW '09*, 2009.

[7] J. Hays and A.A. Efros. IM2GPS: estimating geographic information from a single image. In *Proc. CVPR '08*, 2008.

[8] Jaeyoung Choi et al. Human vs machine: establishing a human baseline for multimodal location estimation. In *Proc. MM '13*, 2013.

[9] Martha Larson et al. Automatic tagging and geotagging in video collections and communities. In *Proc. ICMR '11*, 2011.

[10] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proc. ICMR '11*, 2011.

[11] Wei Zhang and J. Kosecka. Image based localization in urban environments. In *Proc. 3DPVT '06*, 2006.

[12] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on Google Maps street view. In *Proc. ECCV '10*, 2010.

[13] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proc. CVPR '13*, 2013.

[14] D.M Chen et al. City-scale landmark identification on mobile devices. In *Proc. CVPR '11*, 2011.

[15] Kalantidis Yannis et al. VIRaL: Visual image retrieval and localization. *Multimedia Tools and Applications*, 51:555–592, 2011.

[16] Yunpeng Li, D.J. Crandall, and D.P. Huttenlocher. Landmark classification in large-scale image collections. In *Proc. ICCV '09*, 2009.

[17] Jiebo Luo et al. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools Appl.*, 51(1):187–211, 2011.

[18] Martha Larson et al. The benchmark as a research catalyst: Charting the progress of geo-prediction for social multimedia. In *Multimodal Location Estimation of Videos and Images*. Springer, 2015.

[19] X-J Wang, Lei Zhang, and Wei-Ying Ma. Duplicate-search-based image annotation using web-scale data. *Proceedings of the IEEE*, 100(9):2705–2721, 2012.

[20] Michael Riegler et al. How 'how' reflects what's what: Content-based exploitation of how users frame social images. In *Proc. MM '14*, 2014.

[21] Jing Li et al. GPS estimation for places of interest from social users' uploaded photos. *IEEE Trans. Multimedia*, 15(8):2058–2071, 2013.

[22] Xinchao Li, Michael Riegler, Martha Larson, and Alan Hanjalic. Exploration of feature combination in geo-visual ranking for visual content-based location prediction. In *Proc. MediaEval '13*, 2013.

[23] Xinchao Li, Martha Larson, and Alan Hanjalic. Geo-visual ranking for location prediction of social images. In *Proc. ICMR '13*, 2013.

[24] Giorgos Kordopatis-Zilos et al. CERTH at MediaEval Placing Task 2013. In *Proc. MediaEval '13*, 2013.

[25] Lin Tzy Li et al. A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications*, pages 1–37, 2013.

[26] Jamie Davies et al. Identifying the geographic location of an image with a multimodal probability density function. In *Proc. MediaEval '13*, 2013.

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

[28] Ulrich Steinhoff et al. How computer vision can help in outdoor positioning. In *Proc. AmI '07*, 2007.

[29] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proc. CVPR '13*, 2013.

[30] Claudia Hauff, Bart Thomee, and Michele Trevisiol. Working Notes for the Placing Task at MediaEval 2013. 2013.

[31] Lin Tzy Li et al. Multimodal image geocoding: the 2013 RECOD's approach. In *Proc. MediaEval '13*, 2013.

[32] Jamie Davies et al. Placing photos with a multimodal probability density function. In *Proc. ICMR '14*, 2014.

[33] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR '12*, 2012.

[34] Yannis Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.

[35] Herve Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Proc. CVPR '09*, 2009.

[36] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[37] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV '03*, 2003.

[38] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[39] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[40] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[41] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3):346–359, 2008.

[42] L. Juan and O. Gwun. A comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing*, 3(4):143–152, 2009.

[43] Michal Perd'och, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR '09*, 2009.

[44] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Trans. Graphics*, 26(3):4, 2007.

3

PAIRWISE GEOMETRIC MATCHING

The key to effective search-based geo-location prediction, introduced in the previous chapter, is the verification step in the retrieval process. In this chapter, we turn specifically to the issue of spatial verification. Specifically, we consider the pairwise geometric relations between correspondences and propose a strategy to incorporate these relations at significantly reduced computational cost, which makes it suitable for large-scale object retrieval. In addition, we combine the information on geometric relations from both the individual correspondences and pairs of correspondences to further improve the verification accuracy. Experimental results on three reference datasets show that the proposed approach results in a substantial performance improvement compared to the existing methods, without making concessions regarding computational efficiency.

This chapter is published as "Pairwise Geometric Matching for Large-scale Object Retrieval", by X. Li, M. Larson and A. Hanjalic in the *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, Boston, US, 2015.

3.1. INTRODUCTION

IN this chapter, we address the challenge of improving the efficiency and reliability of image matching in an object-based image retrieval scenario. Under object-based image retrieval, further referred simply to as "object retrieval", we understand the problem of finding images that contain the same object(s) or scene elements as in the query image, however, possibly captured under different conditions in terms of rotation, viewpoint, zoom level, occlusion or blur. Many object retrieval approaches and methods [1–5] have been proposed in recent literature, largely inspired by the pioneering work of Sivic and Zisserman [6] and built on the bag-of-features (BOF) principle for image representation. An analysis of the state-of-the-art reveals that these approaches and methods are typically centered around the idea of detecting and verifying correspondences between salient points in a given pair of images. The initial set of correspondences are detected based on matches between visual feature statistics measured in different images around found salient points. The correspondence verification step then serves to filter out unreliable correspondences. This verification is typically a spatial (geometric) one and involves geometric constraints to secure consistency of transformation of different image points. Spatial verification is the key to achieve high precision for object retrieval, especially when searching in large, heterogeneous image collections [6, 7].

A common way of verifying the initial correspondences is to apply a *geometric matching*. Geometric matching can be done either explicitly, by iteratively building an optimized transformation model and fitting it to the initial correspondences (e.g., RANSAC-based model fitting approaches [7, 8]), or implicitly, e.g., by verifying the consistency of the image points involved in the correspondences in the Hough transform space [9, 10]. Compared to these approaches, pairwise relative geometric relations between the correspondences have not been frequently exploited for spatial verification. This may be due to the fact that the typical number N of initially detected correspondences is usually large, resulting in high computational complexity of pairwise comparisons, which can be modeled as $\mathcal{O}(N^2)$. This complexity makes exploitation of pairwise relations less attractive when operating on large image collections. Exploiting these pairwise geometric relations could, however, further improve the performance of image matching as it brings valuable additional information about local object or scene constraints of the correspondences into the matching process. As illustrated in Figure 3.1, the geometric relations in terms of rotation and scaling between vectors formed by a pair of correspondences are closely related to the global geometric relations between images that are encoded in the transformation of the image regions surrounding the salient points. Our goal in this chapter is therefore twofold. First, we aim at generating the conditions under which pairwise geometric relations can be applied for spatial verification at a reasonable computational cost. Second, we aim at maximizing the benefit of involving these relations for improving the object retrieval performance.

We pursue the goal specified above by a novel *pairwise geometric matching* method that consists of three main steps. We first propose a one-versus-one ('1vs1') matching strategy for the initial correspondence set to handle the redundancy of one-to-many correspondences, which is a typical result of detecting correspondences between two images [3] [10]. By removing this redundancy, a new, significantly reduced correspondence set is generated. Then, similarly to [9, 11], we reduce this set even further, by deploying Hough voting in the scaling and rotation transformation space. After these two steps, a large fraction of

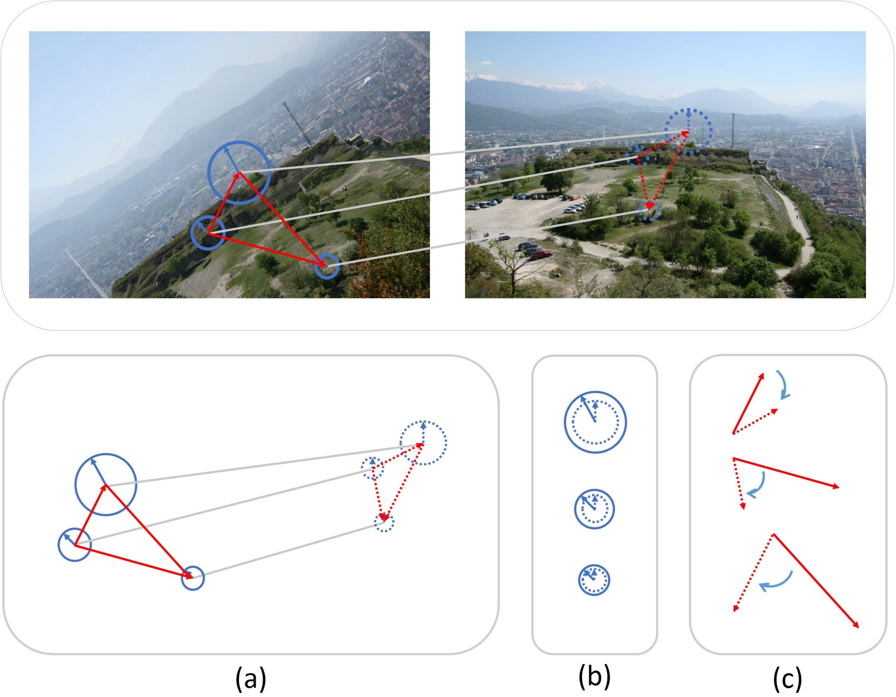


Figure 3.1: (a) Three correspondences found for two images, (b) global rotation and scale relations between images encoded in the transformation of the matched salient points from individual correspondences, (c) rotation and scale relations between vectors formed by pairwise salient points involved in the correspondences. Transformations in cases (b) and (c) are closely related to each other and can be used to emphasize each other for spatial verification.

original correspondences are filtered out, which enables us to exploit pairwise geometric relations for spatial verification at a significantly reduced computational cost. Finally, a simple pairwise weighting method is devised to incorporate both the global geometric relations derived from individual correspondences and the local pairwise relations of pairs of correspondences. As we will show by experimental results in Section 3.6, our proposed method makes the spatial verification of correspondences more tractable in case of a large image collection, but also more reliable, which leads to an overall significant improvement of the object retrieval performance compared to state-of-the-art methods.

3.2. RELATED WORK AND CONTRIBUTION

The existing work addressing the problem of verifying the geometric consistency within a set of correspondences can be grouped in two main categories. The first category comprises the methods exploiting *individual* point correspondences for spatial verification, while the methods from the second category exploit multiple correspondences for this purpose. We briefly analyze the representative methods from these categories and position our contribution with respect to them.

3.2.1. EXPLOITING INDIVIDUAL CORRESPONDENCES

Model-based methods. For two images capturing the same object, a limited number of correspondences can be deployed to estimate the geometric model transforming the points of one image into those of the other image [12]. Once the model is obtained, each correspondence can be assessed in how it fits this model. The key challenge here is how to do model estimation in the presence of noisy correspondences. One of the classical methods to pursue this challenge is RANSAC [13]. Over the years, several attempts have been made to improve its efficiency. For example, Chum et al. [8] managed to significantly speed up the model estimation by adding a generalized model optimization step when the new maximum of inliers is reached. This results in less iterations needed for model estimation to converge. Philbin et al. [7] exploited local appearance of matched image points to generate model hypotheses using a single correspondence, which significantly reduces the amount of possible model hypotheses. Different from RANSAC-based methods, Lowe [11] applied Hough transform to the geometric transformation space to find groups of consistently transformed correspondences prior to estimating the transformation model. In contrast to these model-based methods, which typically need complex iterative model optimization, we are targeting a more lightweight, model-free method.

Model-free methods. As an alternative to the methods discussed above, one can also implicitly verify the correspondences with respect to their consistency in the Hough transformation space. Avrithis and Tolias [10] exploited the relative geometric relations, i.e., scaling, orientation and location, between the local appearance of the matched points. Each correspondence generates one vote in the 4-dimensional transformation space and is then weighted by pyramid matching to capture its consistency with other correspondences. Jégou et al. [9] used the scaling and orientation relations between matched points to find the correspondences that agree with the dominant transformation found in the transformation space. Similarly, Zhang et al. [14] exploited the translation between matched points using Hough voting in a 2-dimensional translation space. Shen et al. [15] also exploited the translation using Hough voting. However, instead of using only the original query object, they applied several transformations with different rotations and scales to the query object, and searched for the best possible translation of these transformed query objects against a collection image. In this way, rotation and scaling invariance can be added to the system. Our proposed method belongs to this category of model-free approaches. However, in contrast to most of the existing work which focuses on individual correspondences, we are considering the pairwise relations between correspondences as well.

3.2.2. EXPLOITING MULTIPLE CORRESPONDENCES

In contrast to rich previous work focusing on individual correspondences, the information encoded in groups of correspondences has remained less exploited for spatial verification. Some related methods implicitly encode the spatial-order information of the correspondences. Wu et al. [16] bundled the local features according to their location and captured the relative order consistency of the correspondences along the X- and Y-coordinates in each image. As this simple way of capturing order consistency cannot support complex geometric transformations, it is primarily suitable for problems of near-duplicate detection. Compared to this, Cao et al. [17] encoded the spatial-order relation between local features by ordering them in a set of linear and circular directions, so rotation can be handled as well. Instead of relying on the ordering of the correspondences, we deploy a more subtle information for spatial verification, namely the rotation and scaling relations between the vectors formed by salient points involved in correspondences. This is likely to make spatial verification more reliable.

We are not the first ones exploiting pairwise geometric relations between correspondences. Carneiro and Jepson [18] employed a pairwise semi-local spatial similarity to capture the pairwise relations of correspondences and grouped them using connected component analysis based on the pairwise similarity matrix. This work was further combined with a probabilistic verification method in [19] to increase the proportion of correct matches in the correspondence set. Likewise, by building a pairwise similarity matrix of correspondences, Leordeanu and Hebert [20] employed a spectral method to greedily recover inliers and find the strongly connected cluster within the correspondence set. These works are related to our approach as they all exploit the pairwise relation between correspondences. However, these methods were designed to exploit the pairwise relations directly from the initial correspondences. As discussed earlier in this chapter, the complexity of spatial verification in this case becomes too high to be applicable in the case of a large image collection. Compared to these methods, our contribution is twofold. First, we significantly reduce the number of correspondences and in this way make the proposed spatial verification more tractable. Second, our pairwise geometric matching method combines both the global geometric relations derived from individual correspondences and the local pairwise relations of pairs of correspondences for improved object retrieval performance.

3.3. CORRESPONDENCE PROBLEM FORMULATION

We start out from a standard representation of an image using local features. This representation typically involves detection of salient points in the image and representation of these points by suitable feature vectors describing local image regions around these points. For instance, in the SIFT [11] scheme, which is widely deployed for this purpose, salient points are detected by a Difference of Gaussians (DOG) function applied in the scale space. The points are then represented by local feature vectors $\mathbf{f} = [\mathbf{x}, \theta, \sigma, \mathbf{q}]$, where $\mathbf{x}$, θ and σ stand for the spatial location, dominant orientation and scale of the represented region around the point, respectively, and $\mathbf{q}$ is the feature description of the region. Given the images F and $\tilde{F}$, and their salient points with indexes i and m and represented by feature vectors $\mathbf{f}_i$ and $\tilde{\mathbf{f}}_m$, respectively, we define the initial set $\mathbf{C}$ of correspondences c_{im} between them as

$$\mathbf{C} = \{(\mathbf{f}_i, \tilde{\mathbf{f}}_m, W_{ini}(c_{im}) | \Phi(\mathbf{f}_i, \tilde{\mathbf{f}}_m) = 1\} \tag{3.1}$$

Here, $\Phi(.) \in \{0, 1\}$ is the binary matching function serving to judge whether two image points capture the same object point in the physical world. For instance, in the BOF scheme, this function is typically computed as $\Phi = \delta(u(\mathbf{q}_i) - u(\tilde{\mathbf{q}}_m))$, where $u(\mathbf{q}_i)$ is the quantized cluster center of the description vector $\mathbf{q}_i$ of local feature $\mathbf{f}_i$ and where $\delta(.)$ is the Kronecker delta. Furthermore, $W_{ini}(c_{im})$ is the weight initially assigned to a correspondence c_{im} and representing the proximity between two points in the local feature space. For instance, the weight can be computed in terms of the statistical distinctiveness of the quantized visual feature center within the image collection, e.g., using the inverse document frequency (*idf*) scheme applied in the BOF context [6]. As an alternative, this weight can also be computed using Hamming distance employed in the Hamming Embedding scheme [3, 9].

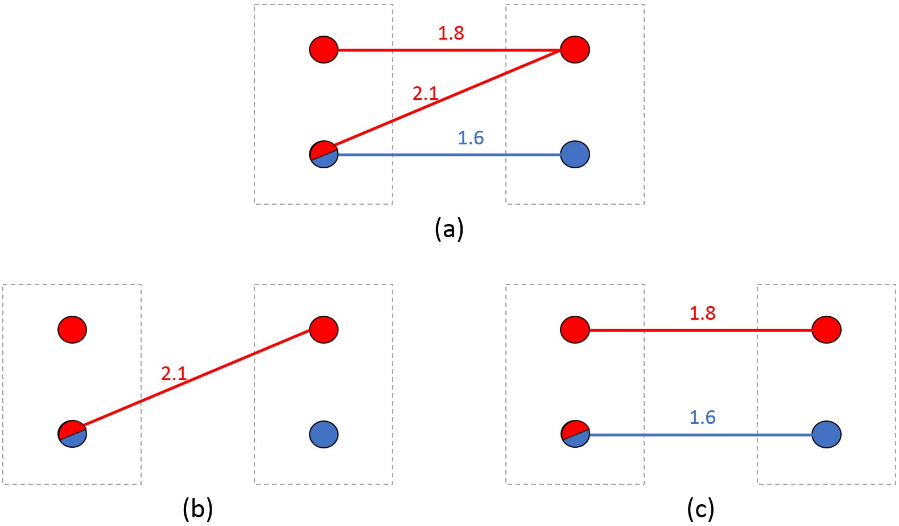


Figure 3.2: Illustration of two different strategies for filtering out multiple alternative correspondences. Case (a) shows the original correspondences. The lower point in the query image (left image) represents a point that matches two different points marked with red and blue in the right image. Case (b) illustrates the strategy by Jegou et al. [3] that focuses on the strongest correspondences. Case (c) is the proposed '1vs1' strategy that balances filtering out of the correspondences with preserving as many informative correspondences as possible.

3.4. PAIRWISE GEOMETRIC MATCHING

In this section we describe the three steps of our proposed pairwise geometric matching method: (a) applying the '1vs1' matching constraint, (b) Hough voting and (c) integrating global and pairwise geometric relations.

3.4.1. 1VS1 MATCHING

The initial correspondence set **C** usually contains a large portion of outliers, or incorrect correspondences, and can include multiple mappings for one single point, i.e., the burstiness phenomenon observed in [3]. However, object matching implies that one object point in one image can only have one corresponding point in another image. Therefore, the final verified correspondence set should only contain unique correspondences between points.

To achieve this, one can formulate an assignment problem, where one can minimize the overall distance between two point sets by using the Hungarian algorithm with the computing time in $O(N^3)$ for set with N features [21]. As finding optimal matches is time consuming, one can aim at an approximate solution. For instance, Jégou et al. [3] proposed to choose the strongest match per point first and then discard all the other matches associated with matched points. However, as can be seen from Figure 3.2 (case (b)), this strategy may result in insufficient number of matches for geometric check. In order to generate a more robust solution, we devise the '1vs1' matching strategy and apply it to the initial correspondence set **C**.

As illustrated by the case (c) in Figure 3.2, in our approach we focus on preserving as many correspondences as possible to maximally inform the assessment of the relation between two images. We first start from the point that originally has fewest matching correspondences assigned (i.e., potential unique matches), select the one with the highest weight, and then discard other matches that contain points belonging to this selected correspondence. We continue this process until no more points need to be processed. In this way, we generate a correspondence set $\mathbf{C}_{1vs1}$ that serves as input for further steps.

3.4.2. HOUGH VOTING

We now depart from the set $\mathbf{C}_{1vs1}$ and follow the strategy from [9, 11] to apply a Hough voting scheme in search for dominant ranges of the target transformation parameters, specifically for the rotation and scaling, in the transformation space. Then, we further reduce the number of correspondences by filtering out those that are not consistently transformed within these ranges.

Each correspondence, c_{im}, stands for a transformation from point i in image F to point m in image $\tilde{F}$. The rotation and scaling relations for this correspondence are denoted, respectively, by

$$\theta = \theta_m - \theta_i, \;\; \sigma = \sigma_m / \sigma_i \tag{3.2}$$

Each correspondence gives a vote in the 2-dimensional rotation-scaling transformation. The dominant ranges of these two transformation parameters, denoted as B_ϑ and B_ς, emerge as the corresponding ranges of the largest bin in the 2-dimensional voting histogram. The correspondences with votes falling in this largest bin are considered to most reliably reveal the transformation between two images. They form the set $\mathbf{C}_{R\&S}$, which serves as input into the last step of the proposed method.

3.4.3. INTEGRATING GLOBAL AND PAIRWISE GEOMETRIC RELATIONS

We start out from the correspondences included in the set $\mathbf{C}_{R\&S}$ and assess the match between images F and $\tilde{F}$ based on pairwise geometric relations between the correspondences. These pairwise geometric relations are derived from the rotation and scaling relations between the corresponding vectors connecting the correspondences in the two images. Given the correspondences, c_g and c_h, which connect point i in image F to point m in image $\tilde{F}$, and point j in image F to point n in image $\tilde{F}$, respectively, we can generate vector $\mathbf{v}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ in image F and vector $\tilde{\mathbf{v}}_{mn} = \mathbf{x}_m - \mathbf{x}_n$ in image $\tilde{F}$. The pairwise geometric relations between the two vectors in terms of rotation and scaling can then be defined as

$$\begin{aligned} \theta_{gh} &= \arccos\left(\frac{\mathbf{v}_{ij} \cdot \tilde{\mathbf{v}}_{mn}}{||\mathbf{v}_{ij}|| \cdot ||\tilde{\mathbf{v}}_{mn}||}\right) \cdot \operatorname{sgn}(\mathbf{v}_{ij} \times \tilde{\mathbf{v}}_{mn}) \\ \sigma_{gh} &= \frac{||\tilde{\mathbf{v}}_{mn}||}{||\mathbf{v}_{ij}||} \end{aligned} \tag{3.3}$$

where θ_{gh} and σ_{gh} are the counterclockwise rotating angle and the scaling factor from $\mathbf{v}_{ij}$ to $\tilde{\mathbf{v}}_{mn}$, respectively.

Each correspondence c_g is then weighted by its pairwise rotation and scaling consistence with other correspondences:

$$W_{PG}(c_g) = \sum_{c_h \in \mathbf{C}_{R\&S}, h \neq g} f(\theta_{gh}, \sigma_{gh}) \tag{3.4}$$

where

$$f(\theta_{gh}, \sigma_{gh}) = \begin{cases} 1, & \text{if } \theta_{gh} \in B_{\vartheta},\ \sigma_{gh} \in B_{\varsigma} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

We note that the weights computed using Eq.3.4 combine together the information on geometric relations obtained from individual correspondences, as imposed by the rotation and scale range limits B_{ϑ} and B_{ς} in Eq.3.5, and from the pairs of correspondences, as indicated by vector relations in Eq.3.3. The final matching score between two images is obtained as the sum of the weights $W_{PG}(c_g)$ of all correspondences from the set $\mathbf{C}_{R\&S}$:

$$S(F, \tilde{F}) = \sum_{c_g \in \mathbf{C}_{R\&S}} W_{PG}(c_g) \tag{3.6}$$

3.5. EXPERIMENTAL SETUP

3.5.1. OBJECT RETRIEVAL FRAMEWORK

We evaluate our proposed pairwise geometric matching method in an object retrieval context. For this purpose, we implemented an object retrieval system based on the classical bag-of-feature-based scheme [6] and considering recent advances in realizing this scheme [2, 3, 9]. To make the system scalable to large image collections, we implemented it using a Map-Reduce-based structure on a Hadoop-based distributed server[1].

Local descriptors and visual words: we use Hessian-affine detector [22] to detect salient points and compute SURF descriptors [23] for these points. As described in [4, 10], the bag-of-feature-based system performs differently depending on whether the visual words vocabulary is trained on an image set with or without test data, i.e., whether the vocabulary is *specific* or *generic*. To mimic the situation in a real retrieval system, we use a separate set of $50k$ randomly selected images from Flickr to learn the *generic* vocabulary set with exact k-means and use it in all experiments.

Weighting the initial correspondences and calculating initial ranking score: As indicated in Section 3.3, the initial set of correspondences can be weighted using different methods. We deploy two common weighting schemes:
(1) ***BOF***: We use the square of the inverse document frequency (*idf*) of the visual word associated with a correspondence as the matching weight. The initial ranking score for the retrieved images is obtained as the sum of the weights of all correspondences, divided by the L2 norm of the bag-of-feature vector.
(2) ***HE***: We employ the Hamming Embedding (HE)-based method proposed in [3] to weight the matched features based on the Hamming distance between their signatures. When calculating the initial ranking score, the burst weighting scheme developed in [3] is employed to handle the burstiness phenomenon in the initial ranking phase.

Multiple assignment (MA): To take into account the quantization noise introduced by a bag-of-feature image representation, we adopted the method from [3] to assign a descriptor to multiple visual words and applied it on the query side only to reduce the computational cost.

---

[1] This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

3.5.2. EXPERIMENTAL PROTOCOL

We assess the proposed method through a comparative experimental analysis and by following similar protocol and criteria as in [10]. We use the *precision-recall curve* to evaluate the pairwise image matching performance and use *mean average precision* (mAP) to evaluate the improvement in object retrieval using the proposed spatial verification method. In the experiments, we use three variants of our implemented object retrieval system based on the two weighting schemes introduced in Section 3.5.1: (1) ***BOF***, with a *generic* vocabulary of 100K, as also deployed in [10], (2) ***HE***, with a *generic* vocabulary of 20K and with 64-bit Hamming signature and (3) ***HE+MA***, which is equivalent to *HE* combined with multiple assignment. This is the same setting as in [3]. We further denote our proposed pairwise geometric matching method as (***PGM***) and its three steps described in sections 3.4.1, 3.4.2 and 3.4.3 as ***1vs1***, ***HV*** and ***PG***, respectively. We refer to the three system realizations incorporating *PGM* as *BOF+PGM*, *HE+PGM* and *HE+MA+PGM*.

We compare these system realizations with state-of-the-art methods both integrally and by adding individual steps one by one in order to assess the contribution of each step to the overall object retrieval performance. We use three state-of-the-art methods as baselines that we refer to as ***HPM*** [10], ***SM*** [20] and ***FSM*** [7]. With respect to *HPM*, we do the comparison directly by integrating the binary code of [10] into our system. As this binary code does not support Hamming embedding, we only integrate it into the *BOF* setting, which is referred to as *BOF+HPM*. Regarding *SM* and *FSM*, as there were no original implementations available for them, the comparison is only indirect, using the experimental results reported in [10] that were obtained on the same datasets as in this chapter.

3.5.3. DATASETS

We conduct the experiments on three publicly available datasets commonly used in the related work, namely *Oxford* [7], *Holidays* [24] and *Barcelona* [25]. To mimic the large-scale image retrieval scenario, we follow the same strategy used in [3, 10] to add distractors to dataset images. We crawled 10 million geo-tagged photos from Flickr for this purpose. These photos are distributed all around the world, except for Oxford and Barcelona regions.

3.6. EXPERIMENTS

3.6.1. IMPACT OF THE PARAMETERS

We start our series of experiments by evaluating the impact of two main parameters, namely the bin sizes of rotation and scale used in Hough voting, on the system performance. These parameters control the trade-off between filtering out the mismatches and remaining tolerant to nonrigid object deformations. We evaluate these parameters in the object retrieval scenario using the *HE+MA* system implementation. Based on the results in Table 3.1, we choose the bin size of 30 degrees for rotation and 0.2 for logarithmic scale as they are best performing across the two datasets, and we adopt these parameter values for all subsequent experiments.

3.6.2. PAIRWISE IMAGE MATCHING

To assess the *PGM* method, we follow the same experimental procedure as in [10], which enumerates all pairs of images in the *Barcelona* dataset and classifies each image pair to be

Table 3.1: mAP comparison of *PGM* on *Oxford* and *Holidays* datasets with different bin sizes for rotation and scale.

	Oxford			Holidays		
	0.1	0.2	0.3	0.1	0.2	0.3
15	0.725	0.734	0.730	0.882	**0.893**	0.888
30	0.735	**0.737**	0.731	0.883	0.892	0.890
45	0.728	0.732	0.724	0.886	0.888	0.882

relevant or irrelevant based on whether its matching score is higher than a threshold. There are in total 927 images in the *Barcelona* dataset, which form $927 \times 927 = 859329$ image pairs, and among which 74,075 image pairs are relevant according to the ground truth. Figure 3.3 shows the precision-recall curves computed for various realizations of our system. Regarding the state-of-the-art, we compare our method directly with *HPM* and indirectly with *SM* based on the results reported in [10] and using similar basic system configuration. For recall of 0.9, *BOF+PGM* achieves the precision of 0.68, which is better than 0.42 achieved by *BOF+HPM* or 0.2 achieved by *SM*. We note that according to Figure 3.3, our method can achieve even better performance (precision of 0.83 at recall 0.9) if the best performing system variant is deployed.

3.6.3. SPATIAL VERIFICATION FOR OBJECT RETRIEVAL

We now evaluate the proposed method in the object retrieval context. For each query image, top-1000 ranked images are selected to perform spatial verification. Since the rank order of these images is adjusted based on verification, we refer to this set of top-1000 images as the *reranking range*. We first evaluate *PGM* against the original datasets without distractors. According to Table 3.2, *PGM* clearly outperforms the baselines. Figure 3.4 shows examples of ranked images obtained using *PGM* and *HPM*.

Table 3.2: mAP comparison of different spatial verification schemes. All results are generated under the same conditions: reranking on top $1K$ ranked photos from BOF using SURF feature and *Single Assignment* on $100K$ vocabulary.

	FSM[1]	HPM[1]	HPM	PGM
Oxford	0.503	0.522	0.525	**0.609**
Holidays	-	-	0.734	**0.825**
Barcelona	0.827	0.832	0.888	**0.900**

[1] The results are from [10].

Figure 3.5 illustrates the system performance with different sizes of image database. The binary code of *HPM* needs to keep all the index information in the memory, which in the case of a database of 10 million images, leads to memory consumption that is too large. For this reason, *HPM* is not included at this scale. The curves in the figure indicate the improvement of the performance after adding each of the steps of our method to the basic *BOF* system configuration. Step-for-step improvement is not clearly evident in the

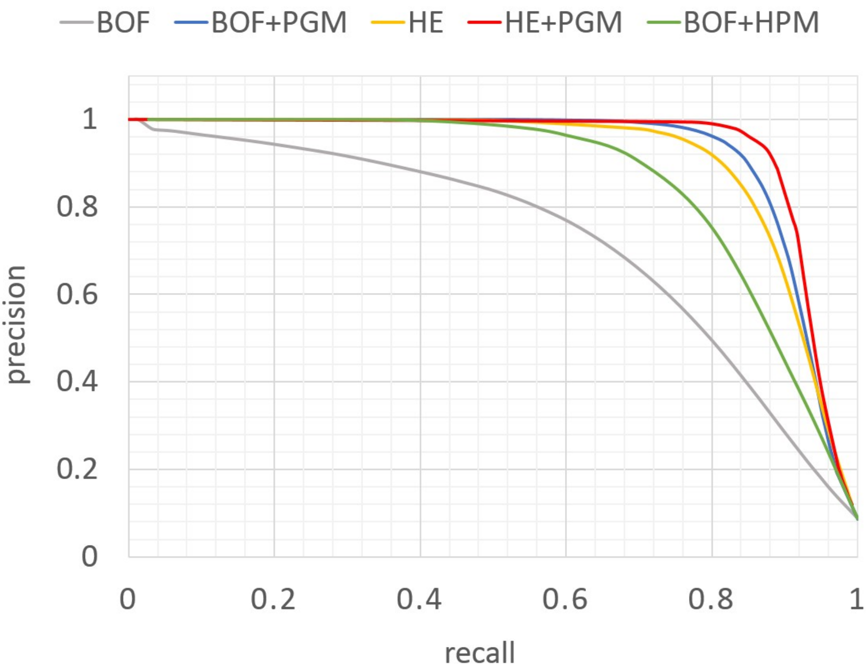


Figure 3.3: Precision-recall curves over all pairs of images in the *Barcelona* dataset.

case of the *HE* system configuration. This is because in this configuration the 'burstiness' phenomenon is handled in the initial retrieval phase using burst weighting [3]. Therefore, the *1vs1* and *HV* steps cannot bring much additional improvement. *PG*, on the other hand, becomes the key step to improve over *HE*.

Regarding the comparison with the best performing baseline, *HPM*, we observe that *BOF+PGM* (cf. *+PG* in Figure 3.5) consistently outperforms *HPM* at each scale. Furthermore, as a flat and much simplified version of *HPM*, *BOF+1vs1+HV* (cf. *+HV* in Figure 3.5) can still achieve comparable performance. This is mainly because, in contrast to detecting conflicts at the visual word level in *HPM*, the proposed *1vs1* matching strategy operates at the point level, which makes it more accurate.

In addition, we observe that the improvement of *BOF+PGM* over *HPM* shrinks with the increasing scale of image collection. Due to the increasing number of distractor images in this case, the number of true-matching photos included in the (in this case fixed) reranking range is likely to decrease. However, within this range, it becomes increasingly easy to separate true matches from the false ones using spatial verification, with the consequence that all verification methods start performing similarly. As illustrated in Figure 3.6, the improvement achieved by PGM becomes significant again when we increase the reranking range with increasing image collection scale.

In the next experiment, we compare our best performing system variant, *HE+MA+PGM* with other state-of-the-art image retrieval systems in a similar setting: constructing the system on *generic* vocabulary, employing *multiple assignment*, using any form of spatial ver-

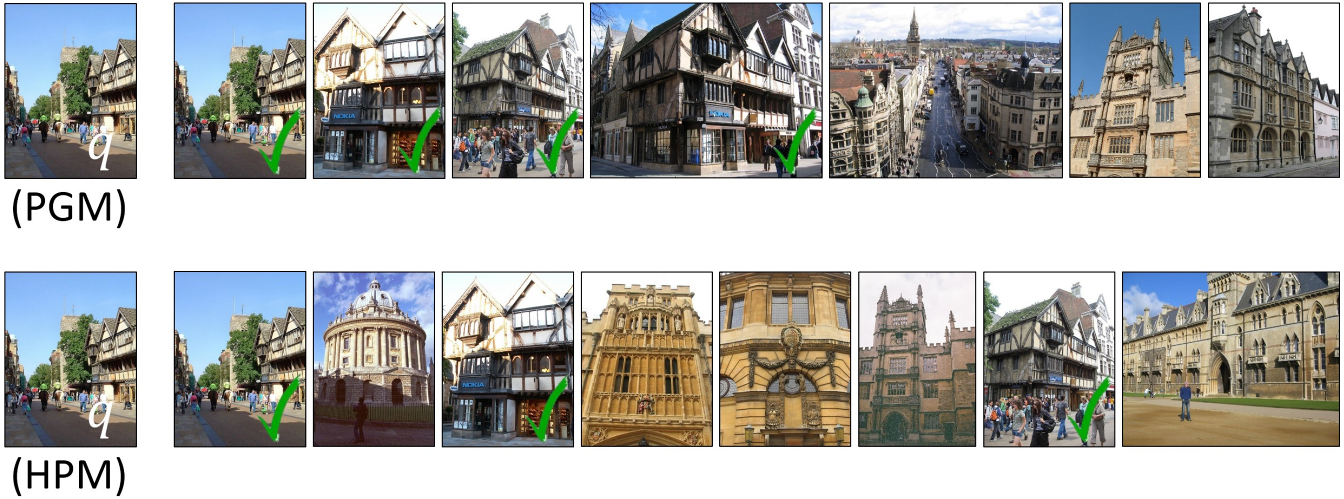


Figure 3.4: Exemplar ranking result for *PGM* and *HPM*.

ification, and without query expansion. As summarized in Table 3.3, our system achieves state-of-the-art performance for image retrieval. The high performance achieved by [26, 27] on the *Oxford* dataset is mainly due to use of superior features, which can efficiently represent unrotated photos. This gain is, however, at the cost of worse performance for rotated photos, e.g., on the *Holiday* dataset. We note that we did not add query expansion [1, 28] and incremental spatial verification [28] into our system, as they usually require re-calculating the correspondences for the new expanded query. We believe, however, that the proposed pairwise geometric matching method is compatible with these schemes.

Table 3.3: mAP comparison of different image retrieval system on generic vocabulary with spatial verification on top 200 (SP200) or top 1000 (SP1000) ranked photos.

	SP	Oxford	Holidays
Jégou et al. [3]	200	0.685	0.848
Philbin et al. [2]	200	0.598	-
HE+MA+PGM	200	**0.691**	**0.892**
Perd'och et al. [26]	1000	0.725	0.769
Mikulík et al. [27]	1000	**0.742**	0.749
HE+MA+PGM	1000	0.737	**0.892**

Table 3.4: Computing time and mAP comparison of *PGM* and *HPM* with spatial verification against all database images.

	Oxford		Holidays		Barcelona	
	Time[1]	mAP	Time[1]	mAP	Time[1]	mAP
PGM	**2.2**	**0.635**	**1.2**	**0.825**	1.1	**0.900**
HPM	2.8	0.527	1.7	0.734	**0.85**	0.888

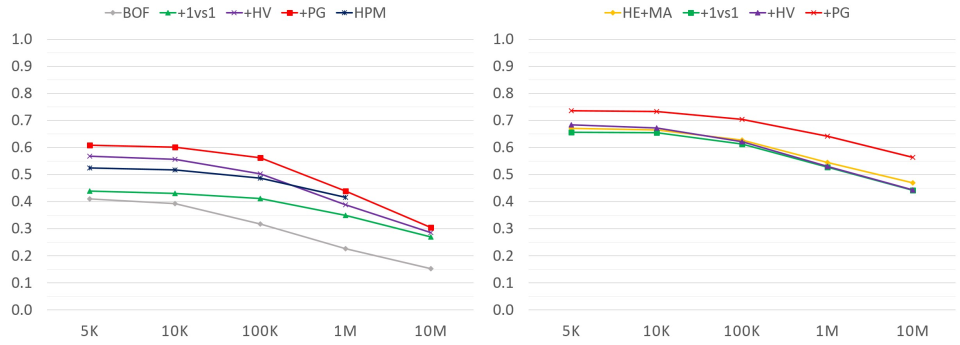
[1] average matching time per pair of images in ms.

3.6.4. RUN TIME EFFICIENCY

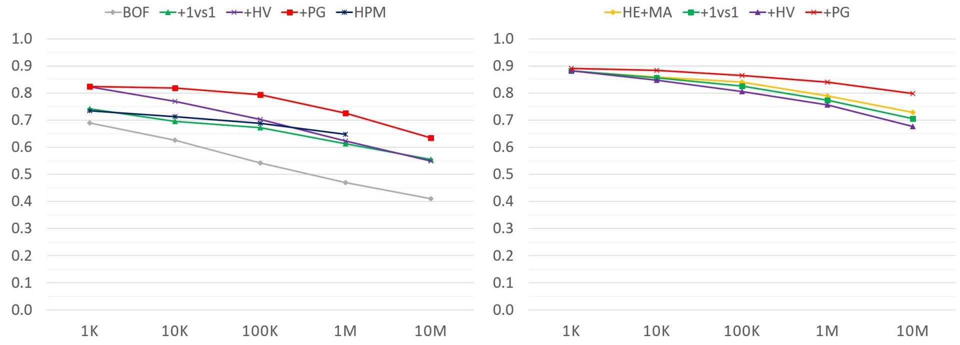
In the last experiment, we evaluate the run time efficiency of our system. To do this, we conduct spatial verification against all database images. We first analyze the effect of the two filtering steps, *1vs1* amd *HV*, on reducing the size of the correspondence set. As illustrated in Figure 3.7, for about 60% of the image pairs, only 20% of matches remained to be checked after these two filtering steps, which dramatically reduces the influence of the pairwise operation on the overall run time. To evaluate the overall run time efficiency, we implement a toy version of our system in Java in a single-thread fashion to be comparable with the available binary code from *HPM*, and test it on a 2.3GHz 8-core processor. As summarized in Table 3.4, *PGM* achieves comparable run time efficiency, while significantly improving the performance. We also evaluate the query time of the entire retrieval system with spatial verification on top-1000 ranked images in the *BOF* setting. *PGM* achieves 2.7s, 1.6s and 0.7s for Oxford, Holidays and Barcelona datasets, respectively. In contrast, HPM consumes 2.9s, 2.7s and 0.7s.

3.7. DISCUSSION

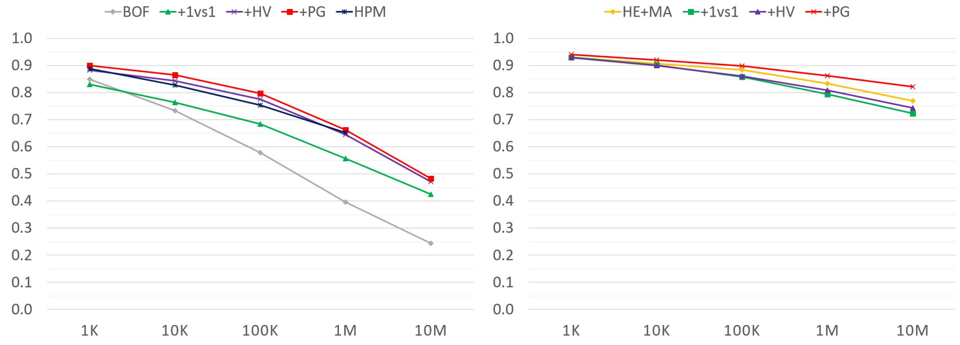
The results presented in the previous section indicate the suitability of the proposed pairwise geometric matching method as a solution for large-scale object retrieval at an acceptable computational cost. The superiority of *PGM* compared to the state-of-the-art solutions becomes evident in a context in which a high number of outliers in the initial correspondences generated by *BOF* and errors in detected features' scale, rotation and position hinder the fit of a specific model (e.g., RANSAC). *PGM* encodes not only scale and rotation information derived from the local points, but also their locations. This is achieved by using global scale and rotation relations to enforce the local consistency of geometric relations derived from the locations of pairwise correspondences. By mapping locations of points to pairwise rotation and scale, the approach is more tolerant to the detection noise. At the same time, using a number of filtering steps, *PGM* significantly reduces the number of correspondences that must be considered, which makes it possible for PGM to maintain high image matching reliability at a substantially reduced computational cost.



(a) BOF&Oxford (b) HE&Oxford

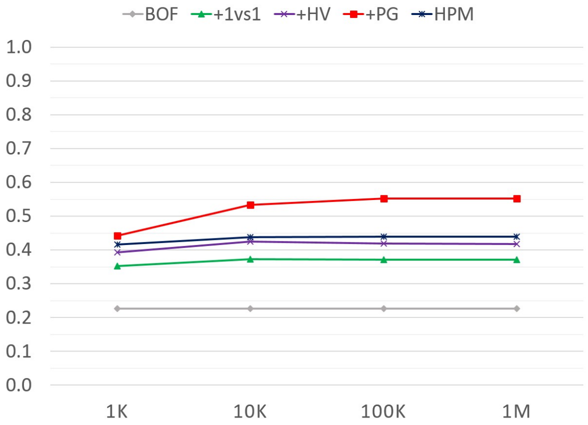


(c) BOF&Holidays (d) HE&Holidays

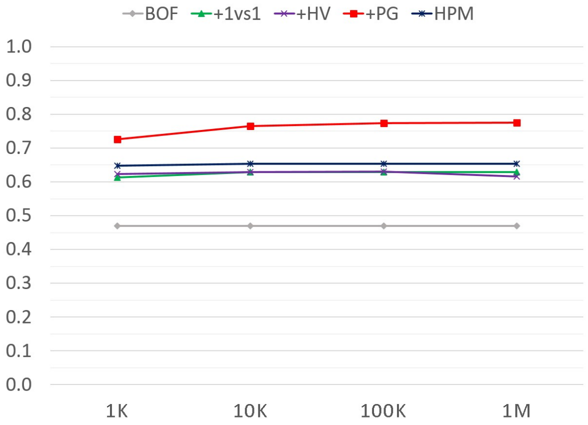


(e) BOF&Barcelona (f) HE&Barcelona

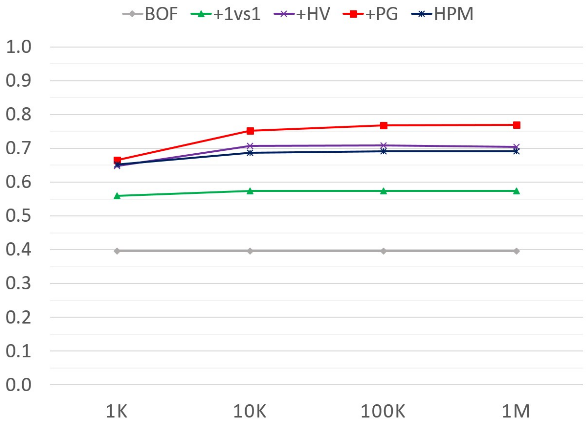
Figure 3.5: mAP of *BOF*-based and *HE*-based systems against different sizes of image database with fixed reranking range.



(a) BOF&Oxford

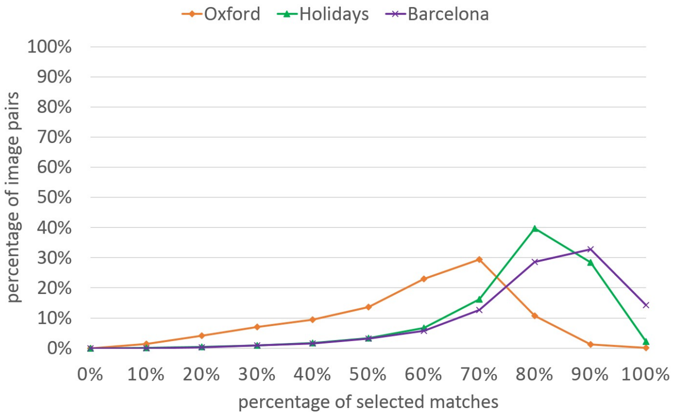


(b) BOF&Holidays

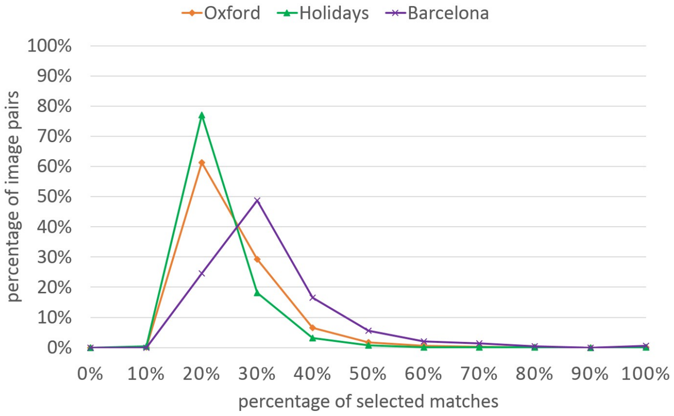


(c) BOF&Barcelona

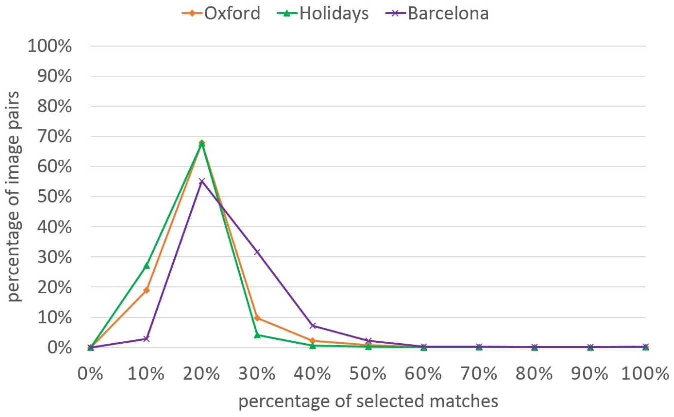
Figure 3.6: mAP of *BOF*-based system against 1M image database with different reranking ranges.



(a) 1vs1



(b) HV



(c) 1vs1 and HV

Figure 3.7: Distribution of the percentage of selected matches after *1vs1* and *HV* steps, taken individually and together.

REFERENCES

[1] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV '07*, 2007.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR '08*, 2008.

[3] Herve Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Proc. CVPR '09*, 2009.

[4] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR '12*, 2012.

[5] Linjun Yang, Bo Geng, Yang Cai, Alan Hanjalic, and Xian-Sheng Hua. Object retrieval using visual query context. *IEEE Trans. Multimedia*, 13(6):1295–1307, 2011.

[6] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV '03*, 2003.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR '07*, 2007.

[8] Ondrej Chum, Jirı Matas, and Stepan Obdrzalek. Enhancing ransac by generalized model optimization. In *Proc. ACCV '04*, 2004.

[9] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[10] Yannis Avrithis and Giorgos Tolias. Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International Journal of Computer Vision*, 107(1):1–19, 2014.

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

[12] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.

[14] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *Proc. CVPR '11*, 2011.

[15] Xiaohui Shen, Zhe Lin, Jonathan Brandt, Shai Avidan, and Ying Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *Proc. CVPR '12*, 2012.

[16] Zhong Wu, Qifa Ke, M. Isard, and Jian Sun. Bundling features for large scale partial-duplicate web image search. In *Proc. CVPR '09*, 2009.

[17] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Spatial-bag-of-features. In *Proc. CVPR '10*, 2010.

[18] Gustavo Carneiro and Allan D Jepson. Flexible spatial models for grouping local image features. In *Proc. CVPR '04*, 2004.

[19] G. Carneiro and AD. Jepson. Flexible spatial configuration of local image features. *IEEE Trans. PAMI*, 29(12):2089–2104, 2007.

[20] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proc. ICCV '05*, volume 2, pages 1482–1489 Vol. 2, 2005.

[21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[22] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3):346–359, 2008.

[24] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV '08*, 2008.

[25] G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *Proc. ICCV '11*, 2011.

[26] Michal Perd'och, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR '09*, 2009.

[27] Andrej Mikulík, Michal Perdoch, Ondřej Chum, and Jiří Matas. Learning a fine vocabulary. In *Proc. ECCV '10*. 2010.

[28] Ondrej Chum, Andrej Mikulik, Michal Perdoch, and Jiri Matas. Total recall ii: Query expansion revisited. In *Proc. CVPR '11*, 2011.

4

GEO-DISTINCTIVE VISUAL ELEMENT MATCHING

In this chapter, we further improve our visual-only, search-based framework for geo-location prediction of social images. Specifically, we focus on the geo-distinctiveness of visual elements within an image collection, and introduce an approach called distinctive visual element matching (DVEM). This approach uses representations that are specific to the query image whose location is being predicted. These representations are based on visual element clouds, which robustly capture the connection between the query and visual evidence from candidate locations. We then maximize the influence of visual elements that are geo-distinctive because they do not occur in images taken at many other locations. We carry out experiments using two large-scale, publicly-available datasets: the San Francisco Landmark dataset with 1.06 million street-view images and the MediaEval 2015 Placing Task dataset with 5.6 million geotagged images from Flickr.

This chapter is submitted as "Geo-distinctive Visual Element Matching for Location Estimation of Images", by X. Li, M. Larson and A. Hanjalic to the *Multimedia, IEEE Transactions on*, 2016.

4.1. INTRODUCTION

INFORMATION about the location at which an image was taken is valuable image metadata. Enriching images with geo-coordinates benefits users by supporting them in searching, browsing, organizing and sharing their images and image collections. Specifically, geo-information can assist in generating visual summaries of a location [1, 2], in recommending travel tours and venues [3, 4], in photo stream alignment [5], and in event mining from media collections [6, 7].

While many modern mobile devices can automatically assign geo-coordinates to images during capture, a great number of images lack this information [8]. Techniques that automatically estimate the location of an image [8–12] have been receiving increasing research attention in recent years. Specifically, predicting geographic location solely from visual content holds the advantage of not depending on the availability of the textual annotation. The challenge of visual content-based geo-location estimation derives from the relationship between visual variability and location. Images taken at a single location may display high visual variability, whereas images taken in distinct locations may be unexpectedly similar.

The core idea underlying our approach to this challenge is depicted in Fig. 4.1, which illustrates the pattern of visual matching that we will exploit in this chapter. Inspecting each column of images in turn, we can see similarities and differences among the areas of the images marked with colored boxes. These areas contain visual elements that match between query image (top row) and the location images (lower rows). We use the term *visual element* to denote a group of pixels (i.e., an image neighborhood) that is found around salient points and that also can automatically be identified as being present in multiple images, i.e., by means of visual matching.

Moving from left to right in the figure, we notice that the areas matched in the first two locations (left and middle columns) share similarity. Here, the visual elements contained in these areas correspond to FedEx trucks, street lights, and fire escapes. The locations in these two columns are *different* from the query location. These visual matches introduce visual confusion between the query image and images taken at other locations. In contrast, location in the third column is the *same* as the query location. The matching areas contain visual elements correspond to specific, distinguishing features of the real-world location, not found in other locations, in this case, elements of the architecture. We call such visual elements *geo-distinctive*.

This chapter introduces a visual matching approach to image geo-location estimation that exploits geo-distinctive visual elements, referred to as *distinctive visual element matching* (DVEM). This approach represents a contribution to the line of research dedicated to developing search-based approaches to visual-content-based geo-location estimation for images. Under search-based geo-location estimation, the target image (whose geo-coordinates are unknown) is used to query a *background collection*, a large collection of images whose geo-coordinates are known. Top-ranking results from the background collection are processed to produce a prediction of a location, which is then propagated to the target image. As is customary in search-based approaches, we refer to the target image as the *query image*. The DVEM approach represents a significant extension to our generic *geo-visual ranking* framework [13] for image location estimation.

As will be explained in detail in Section 4.2 and 4.3, DVEM represents a considerable

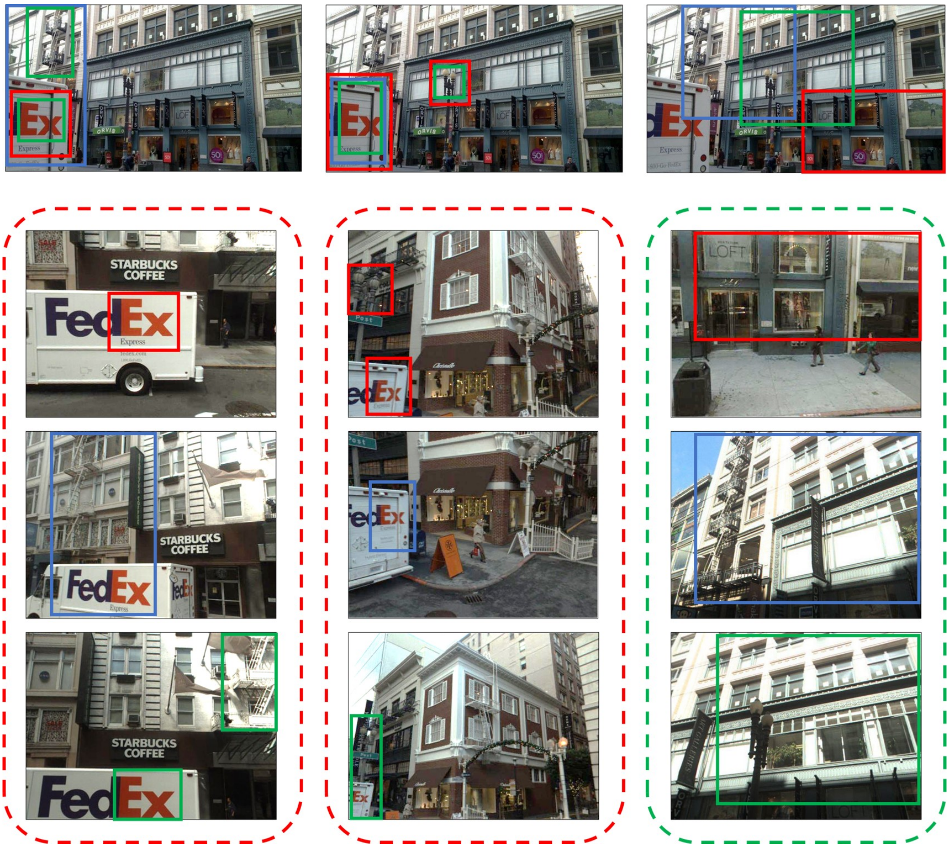


Figure 4.1: Colored boxes indicate match between areas of a query image (top row) and location images taken at three different locations (columns). Note how these areas differ when the location is different from the query location (left and middle columns) and when it is the same (right column).

advancement of the state of the art in search-based approaches to visual-content-based image geo-location estimation. In a nutshell, the innovation of DVEM is its use of a visual representation that is ‘contextual’ in that it is specific to the query image. This representation is computed in the final stage of search-based geo-location, during which top-ranked results are processed. The key is that the representation is not fixed in advance, but rather is calculated at prediction time, allowing it to change as necessary for different queries. This factor sets DVEM apart from other attempts in the literature to exploit geo-distinctiveness, which pre-calculate representations based on the background collection, rather than zeroing in on visual information most important for an individual query. The experimental results we present in this chapter demonstrate that DVEM can achieve a substantial improvement for both major types of image geo-location prediction covered in the literature: geo-constrained and geo-unconstrained.

The remainder of the chapter is organized as follows. In Section 4.2, we present the rationale underlying our proposed approach, DVEM, and describe its novel contribution in more detail. Then, in Section 4.3, we provide an overview of the related work in the domain of image location estimation and position our contribution with respect to it. Section 4.4 describes the DVEM approach in detail. Our experimental setup is explained in Section 4.5 and Section 4.6 reports our experimental results. Section 4.7 concludes the chapter and provides an outlook towards future work.

4.2. RATIONALE AND CONTRIBUTION

The fundamental assumption of content-based geo-location estimation is that two images that depict the same objects and scene elements are likely to have been taken at the same location. On the basis of this assumption, search-based geo-location estimation exploits image content by applying object-based image retrieval techniques. The rationale for our approach is grounded in a detailed analysis of the particular challenges that arise when these techniques are applied to predict image location. We examine these challenges in greater depth by returning to consider Fig. 4.1. In Section 4.1, we have already discussed the existence of confounding visual elements in images from the wrong location (left and middle columns), and also of characteristic visual elements in images from the true location (right column). We now look again at these cases in turn.

Geo-distinctivness. Images taken at a wrong location (Fig. 4.1 left and middle) capture a underlying reality that is different from the reality captured by the query. The figure shows two typical sources of confounding visual elements. First, elements corresponding to real-world objects that are able to move from one location to the other, such as a FedEx truck. Second, elements corresponding to objects that are identical or highly similar and occur at multiple locations, such as the fire escapes and the street lamps. A third case (not depicted) occurs when objects or scene elements at different locations appear having similar visual elements in images due to the way in which they were captured (i.e., perspective, lighting conditions, or filters).

Our approach is based on the insight that confounding visual elements will occur in many locations that are *not* the true location of the image. DVEM is designed to limit the contribution of visual elements that occur in many locations, and instead base its prediction on visual elements that are discriminative for a specific location.

Location representation. Images taken at the true location (Fig. 4.1 right column) imply a related set of challenges. Conceptually, to relate a query image and its true location, we would like to count how many visual elements in the query correspond to real-world aspects of the location. Practically, however, such an approach is too naïve, since we cannot count on our image collection to cover each location comprehensively. Further, we face the difficulty that the true-location images in our background collection may have only a weak link with the query image. Specifically for the example in Fig. 4.1, the variation in perspective is significant between the query and images from the true location (right column), which will heavily weaken their visual correspondences. We again must deal with the same set of factors that give rise to confounding visual elements, mentioned above: camera angle, zoom-level, illumination, resolution, and filters. These also include the presence of mobile objects such as pedestrians, vehicles, and temporary signs or decorations. We have no control over the presence of these distractors, but we can seek to reduce their impact, which will in turn limit their contribution to the match between query and wrong locations.

DVEM builds on the practical insight that we should focus on aggregating evidence strength, rather than merely counting visual elements common between a query and a location. In particular, we aim to integrate two tendencies, which are illustrated by the right column of Fig. 4.1. Here, it can be seen that the match between query image and true location involves (a) a wider variety of different visual elements than matches with wrong locations and (b) visual elements that are distributed over a larger area within the image. These tendencies can be considered as reflections of the common sense expectation that the number of ways in which a query can overlap with true-location images is much larger than the number of ways in which a query can overlap with wrong-location images.

Connection with search-based geo-location estimation. Next we turn to describe how DVEM extends our general *geo-visual ranking* (GVR) framework [13]. As previously mentioned, DVEM contributes to the processing step in a search-based geo-location estimation pipeline. Fig. 4.2 depicts the GVR framework in the top row, and the DVEM extension in the bottom row. The dashed line indicates the steps that compose DVEM and the arrow show that it replaces the Location Ranking step of GVR.

Here, we provide a brief review of the functioning of GVR. In the Candidate Image Selection step, we use the query image to query a background collection (corpus) of geo-tagged images, i.e., images annotated with geo-coordinates. In the Location Extraction step, we group the retrieved images according to their locations, creating image sets corresponding to candidate locations. This information serves as input into DVEM.

The three steps of DVEM are designed to address the challenges covered at the beginning of the section, and incorporate both geo-distinctiveness and location representation:

- *Location as Visual Element Cloud* builds a ‘contextual’ query-specific representation of each candidate-location image set that reflects the strength of the visual evidence relating that image set to the query.
- *Geo-Distinctiveness Modeling* captures the ability of visual elements to discriminate the image sets of the candidate locations that are competing for a given query.
- *Visual Matching per Location* calculates the ranking score for each candidate location with the target to incorporate both the distinctiveness of visual elements and the matching strength between visual elements and the location.

These steps are explained in more detail in Section 4.4, which also includes further motivating examples.

Novel contributions. As stated in the introduction, the novel contribution of DVEM is its use of query-specific, ‘contextual’, visual representations for geo-location estimation. No collection-wide representation of location is needed. Instead, flexible representations are built at prediction time that aggregate evidence for ranking a location optimally against its specific competitors for each query. The implications of this contribution are best understood via a comparison with classical information retrieval. DVEM can clearly claim the traditional vector space model with TF-IDF weighting scheme used in information retrieval as a progenitor. TF-IDF consists of a Term Frequency (TF) component, which represents the contents of items (documents), and an Inverse Document Frequency (IDF) component, which discriminates items from others in the collection [14]. DVEM uses the same

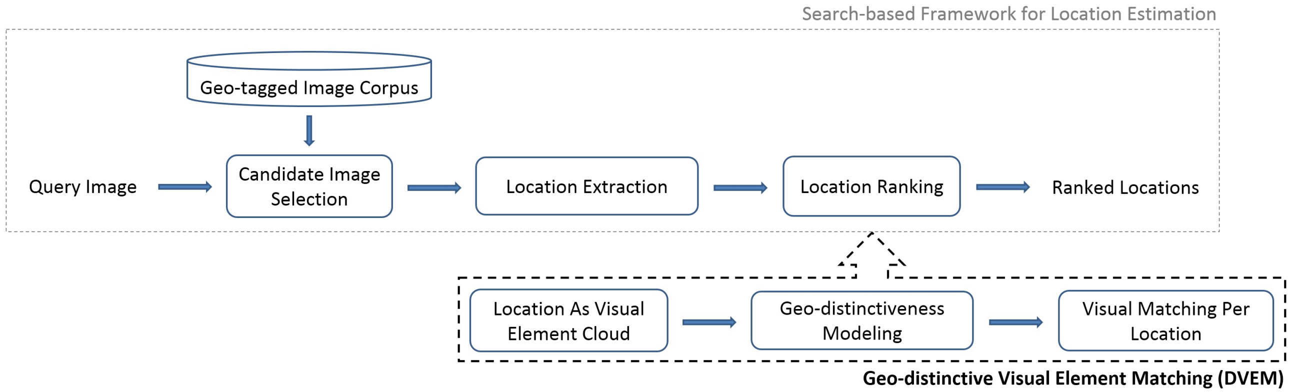

Search-based Framework for Location Estimation
Geo-tagged Image Corpus
Query Image
Candidate Image Selection
Location Extraction
Location Ranking
Ranked Locations
Location As Visual Element Cloud
Geo-distinctiveness Modeling
Visual Matching Per Location
Geo-distinctive Visual Element Matching (DVEM)

Figure 4.2: Our propsed *Geo-distinctive Visual Element Matching* (DVEM) approach, depicted with its integration as the location ranking step of the generic search-based location estimation framework [13].

basic principle of combining a representative component, the visual element cloud, and a discriminative component, geo-distinctiveness modeling. However, its application of these principles is unique, and differentiate DVEM from the ways in which TF-IDF has been deployed for bag-of-feature-based image retrieval in the past.

- DVEM moves matching from the level of the item (i.e., individual image), to the level of the candidate image set. The visual element cloud generated from the candidate image set makes it possible for individual visual elements to contribute directly to the decision, compensating for the potentially weak visual link of any given location image with the query.

- DVEM dispenses with the need to define individual locations at the collection level offline at indexing time. Instead DVEM defines ‘contextual’ visual representations of locations over the candidate image sets, which represent the images most relevant for the decision on the location of a particular query at prediction time.

The use of ‘contextual’ visual representations of locations that are created specifically for individual queries have two important advantages. First, these representations involve only images that have been visually verified in the candidate image selection step. Since images that are not relevant to the location estimation decision are not present in the candidate image set, the location representations can focus on the ‘contextual’ task of ranking the competing locations to make the best possible decision for a given query, improving robustness. Second, the number of competing locations for any given query is relatively low, meaning that the geo-distinctiveness calculation is computationally quite light. This solves the problem of making geo-distinctiveness computationally tractable. It allows DVEM to scale effortlessly as the number of possible candidate locations grows to be effectively infinite in the case of geo-location estimation at global scale.

As we will show by experimental results in Section 4.6, these advantages delivers an overall significant improvement of the location estimation performance compared to state-of-the-art methods.

4.3. RELATED WORK

Visual-only geo-location estimation approaches can be divided into two categories. The first is *geo-constrained* approaches. Such approaches estimate geo-location within a geographically constrained area [15, 16] or a finite set of locations [17–20]. The second is *geo-unconstrained* approaches, which estimate geo-location at a global scale [10, 13]. The challenge of geo-unconstrained geo-location estimation is daunting: a recent survey [21] indicated that there are still ample opportunities waiting to be explored in this respect. In this work, our overall goal is to substantially improve the accuracy of image location estimation using only their visual content, and to achieve this improvement in both the geo-constrained and geo-unconstrained scenarios. As demonstrated by our experimental results, DVEM’s representation and matching of images using geo-distinctive visual elements achieves a substantial performance improvement compared to existing approaches to both geo-constrained and geo-unconstrained location estimation.

4.3.1. GEO-CONSTRAINED CONTENT-BASED LOCATION ESTIMATION

City-scale location estimation. Chen et al. [15] investigated the city-scale location recognition problem for cell-phone images. They employed a street view surveying vehicle to collect panoramic images of downtown San Francisco, which were further converted into 1.7 million perspective images. Given a query image taken randomly from a pedestrian's perspective within the city, a vocabulary-tree-based retrieval scheme based on SIFT features [22] was employed to predict the image's location by propagating the location information from the top-returned image.

Gopalan [23], using the same data set, modeled the transformation between the image appearance space and the location grouping space and incorporated it with a hierarchical sparse coding approach to learn the features that are useful in discriminating images across locations. We choose this dataset for our experiments on the geo-constrained setting, and use this approach as one of our baselines. The other papers that evaluate using this data set are the aggregated selective matching kernel purposed by Tolias et al. (2015) [24], the work exploiting descriptor distinctiveness by Arandjelović and Zisserman (2014) [25], the work exploiting repeated pattens by Torii et al. (2013) [16], the graph based query expansion method of Zhang et al. (2012) [26] and the initial work of Chen et al. (2011) [15]. The experiments in Section 4.6.4 makes a comparison with all of these approaches.

The DVEM is suited for cases in which there is no finite set of locations to apply a classification approach. However, we point out here, that classification approaches have been proposed for geo-constrained content-based location estimation. Gronat et al. [27] modeled each geo-tagged image in the collection as a class, and learned a per-example linear SVM classifier for each of these classes with a calibration procedure that makes the classification scores comparable to each other. Due to high computational cost in both off-line learning and online querying phases, the experiment was conducted on a limited dataset of $25k$ photos from Google Streetview taken in Pittsburgh, U.S., covering roughly an area of $1.2 \times 1.2 km^2$.

Beyond city scale. Authors that go beyond city scale, may still address only a constrained number of locations. Kalantidis et al. [18] investigate location prediction for popular locations in 22 European cities using *scene maps* built by visually clustering and aligning images depicting the same view of a scene. Li et al. [17] constructed a hierarchical structure mined from a set of images depicting about 1,500 predefined places of interest, and proposed a hierarchical method to estimate image's location by matching its visual content against this hierarchical structure. Our approach resembles [18] in that we also use sets of images to represent locations. Note however that in DVEM location representations are created specifically for individual queries at prediction time, making it possible to scale beyond the fixed set of locations.

4.3.2. GEO-UNCONSTRAINED CONTENT-BASED LOCATION ESTIMATION

Estimating location from image content on a global scale faces serious challenges. First, there is effectively an infinite number of locations in the world. Second, geo-unconstrained location prediction is generally carried out on large collections of user-contributed social images. As a consequence, less photographed locations are underrepresented. These challenges imply that geo-unconstrained location estimation cannot be addressed by training a separate model for each location. Finally, the visual variability of images taken a given location is often high, and is also quite erratic. For instance, images taken at a location of a monument that is a tourist attraction will probably focus on some aspects of the monument, limiting the scope of the captured visual scene. However, images taken at an arbitrary beach may be taken from any view point to capture a wide variety of the visual scene. This variability can heavily hinder inference of location-specific information from the visual content of images, and exacerbates the difficulty of linking images showing different aspects of a location.

The problem of geo-unconstrained content-based image location estimation was first tackled by Hays and Efros [10]. They proposed to use visual scene similarity between images to support location estimation with the assumption that images with higher visual scene similarity were more likely to have been taken at the same location. In recent years, research on geo-unconstrained location prediction has been driving forward by the MediaEval Placing Task [21]. The Placing Task result most relevant to DVEM is our submission to the 2013 Placing Task [28]. This submission deployed a combination of local and global visual representations within the GVR system [29], and out-performed other visual-content-based approaches that year. Here, we will focus on 2015, the most recent edition of the Placing Task [30], which received three submissions using visual-content-based approaches. Kelm et al. [31] exploited densely sampled local features (pairwise averaged DCT coefficients) for location estimation. Since this submission is not yet a functional, mature result, it is not considered further here. Li et al. [32] employed a rank aggregation approach to combine various global visual representations in a search-based scheme, and used the top ranked image as the source for location estimation. Instead of using hand-crafted features, Kordopatis-Zilos et al. [33] made use of the recent developments in learning visual representations. They fed a convolutional neural network with images from 1,000 points of interest around the globe and employed it to generate the CNN features. Location is then estimated for the query image by finding the most probable location among the most visually similar photos calculated based on their proximity in the feature space.

Our DVEM is related to these approaches in the sense that it is data driven and search based. However, these approaches depend on finding significant image-level matches between the query image and individual images in the background collection. They do not attempt to compensate for the possibility that the match between the query image and individual images taken at the true location might be minimal, due to the way in which the image was taken, or exploit geo-distinctiveness.

4.3.3. GEO-DISTINCTIVE VISUAL ELEMENT MODELING

As discussed in Section 4.2, in a classical information retrieval system, document (item) distinctiveness is traditionally computed off-line during the indexing phase at the level of the entire collection [14]. This approach is also used in the classical bag-of-feature-based image retrieval system. For example, in [34], the distinctiveness of each visual word is generated from its distribution in the image database. Note that our system uses the approach of [34] in the Candidate Image Selection step (first block in Fig. 4.2), as a standard best practice. Our novel use of geo-distinctiveness goes above and beyond this step, as described in the following.

The key example of the use of distinctiveness for content-based geo-location estimation is the work of Arandjelović and Zisserman [25], who modeled the distinctiveness of each local descriptor from its estimated surrounding local density in the descriptor space. This approach differs from ours in two ways: first, we use geo-distinctiveness, calculated on the basis of individual locations, rather than general distinctiveness and, second, we use geo-metrically verified salient points, rather than relying on the visual appearance of the descriptors of the salient points. As we will show by experimental results in Section 4.6, which uses Arandjelović and Zisserman [25] as one of the baselines, this added step of geo-distinctive visual elements matching significantly improves location estimation

Where geo-distinctivenss has been used in the literature, it has been pre-computed with respect to image sets representing a pre-defined inventory of locations. Typical for such approaches is the work from Doersch et al. [35]. They built a collection of image patches from street view photos of 12 cities around the world, and mined the image patches that are location-typical—both frequent and discriminative for each city—based on the appearance similarity distribution of the image patches. Similarly, Fang et al. [36] incorporated the learned geo-representative visual attributes into the location recognition model in order to improve the classification performance. These learned geo-representative visual attributes were shown to be useful for city-based location recognition, i.e., to assign a given image to one of the cities. However, this approach faces a significant challenge. As the number of pre-defined locations grows larger, there are less geo-representative visual attributes exist per location. For this reason, the risk increases that a query image contains few of the location-typical elements that have been found for the location at which it was taken. In our work, instead of extracting location-typical features from the image collection and using them to assess the query, we turn the approach around. Our approach is to focus on the visual elements that we extract from the query, and to model their geo-distinctiveness around candidate locations for this particular query at prediction time.

4.4. GEO-DISTINCTIVE VISUAL ELEMENT MATCHING

In this section, we present DVEM in depth, providing a detailed description of the components depicted in Fig. 4.2. We start with the GVR framework [13] (Fig. 4.2, top row), the generic search-based location estimation pipeline upon when DVEM builds. The framework was described in Section 4.2. Here, we provide the necessary additional detail.

The first step of GVR is Candidate Image Selection, and serves to retrieve, from the collection of geo-tagged images, a ranked list of candidate images that are most visually similar to the query q. In contrast to the original version of GVR, our new pair-wise geometrical matching approach is used for this step [37]. The result is a ranked list of images that have been visually verified, ensuring that we can be confident that their visual content is relevant for the decision on the location of the query image. We limit the ranked list to the top 1000 images, since this cutoff was demonstrated to be effective in [13]. In the second step, Location Extraction, candidate locations are created by applying a geo-clustering process to the ranked list (see [13] for details), resulting in the set G of candidate locations. The set of images I_g associated with each location g in G is referred to as the *candidate location image set*. In the third step, Location Ranking, visual proximities for each g are calculated on the basis of sets I_g and the query q, resulting in $Score(g,q)$. Finally, $Score(g,q)$ is used to rank the locations g in G. The top-ranked location provides the geo-location estimate, and is propagated to the query image. As previously mentioned, DVEM replaces the Location

Ranking step of GVR. Specifically, it contributes an advanced and highly effective method for calculating $Score(g,q)$. The remainder of this section discusses each of the steps of DVEM (bottom row Fig. 4.2) in turn.

4.4.1. LOCATION AS VISUAL ELEMENT CLOUD

The visual element cloud is a representation of I_g that aggregates the evidence on the strength of the visual link between I_g and the query q. The cloud, illustrated in Fig. 4.3, serves as a representation of the location g in terms of visual elements that occur in the query. For the first step of creating the cloud, we adopt the standard approach (as used, e.g., with SIFT) of detecting salient points in the images using a salient point detector and representing these points with feature vectors (i.e., descriptors) describing the local image neighborhoods around the points. The size of the neighborhood is determined by the salient point detector.

Next, we calculate correspondences between the salient points in the query and in the individual images on the basis of the local image neighborhoods of the points. Then, we apply geometric matching, which secures the consistency of transformation between different salient points. In this work, we use PGM [37], as applied in the Candidate Image Selection step, but another geometric verification approach could also be chosen. The result of geometric matching is a set of one-to-one correspondences c between salient points in the query and in the individual images I_g (cf. Fig. 4.3a), and a set of matching scores $IniScore(c)$ associated with the correspondences c. The *visual elements* are the salient points in the query image that have verified correspondences in I_g. Note that our use of one-to-one correspondences ensures that a visual element may have only a single correspondence in a given image. As will be seen in detail below, the matching score $IniScore(c)$ allows us to incorporate our confidence concerning the reliability of the visual evidence contributed by individual visual elements into the overall $Score(g,q)$, which is used to rank the location.

Finally, we aggregate the visual elements and their scores per image in I_g in order to generate the visual element cloud (cf. Fig. 4.3b). Formally expressed, the visual element cloud $\mathbf{S}_g$ for location g is calculated as:

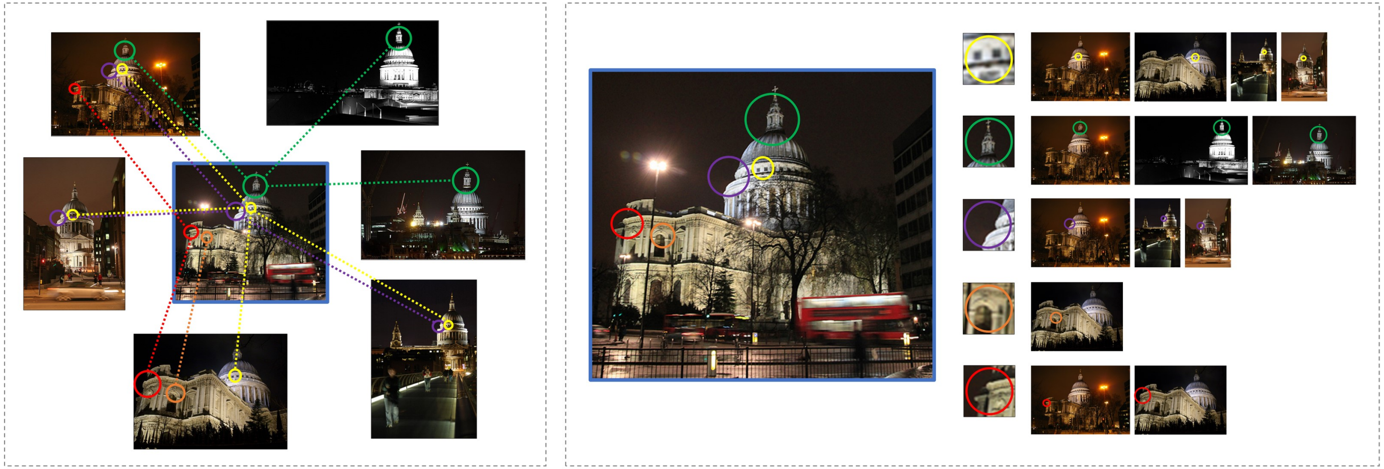
$$\mathbf{S}_g = \{\mathbf{W}_e | e \in \mathbf{E}_g, \mathbf{W}_e = \{w(e)_j | j = 0, 1 \ldots m(e)\}\} \tag{4.1}$$

Here, $\mathbf{E}_g$ is the set of visual elements that occur in the query and link it with the images I_g representing location g. $\mathbf{W}_e$ is the set of weights $w(e)_j$ of correspondences between the visual element e appearing in the query and the j^{th} image in I_g in which it also appears. The total number of images which have correspondences involving element e in the set I_g is denoted by $m(e)$.

The weights $w(e)_j$ are obtained by using a Gaussian function to smooth the initial matching score, $IniScore(c)$, of the correspondence c in which the j^{th} appearance of the visual element e is involved, and is denoted as

$$w(e)_j = 1 - \exp(-\frac{IniScore(c)^2}{\delta^2}). \tag{4.2}$$

Here, δ controls the smoothing speed as shown in Fig. 4.4.



(a)

(b)

Figure 4.3: Illustration of the visual element cloud. Figure (a) shows the correspondences c between the query image (center) and images taken in one location. The relationship between the visual element cloud constructed for this location and the query is illustrated in Figure (b). The cloud is represented by the visual elements from the query and the images of the location these elements appear in

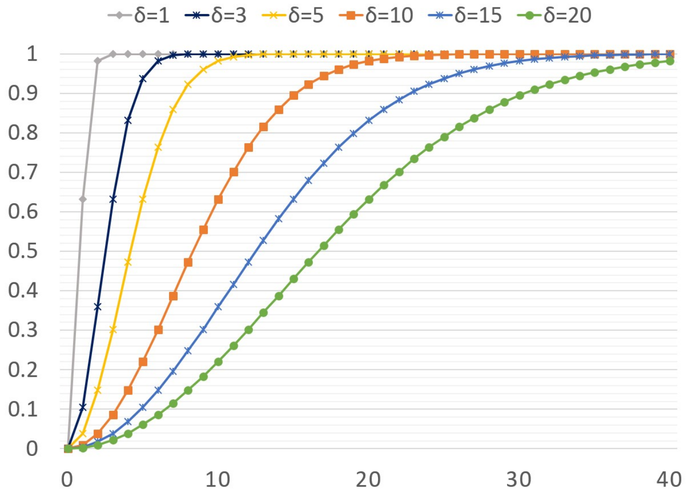


Figure 4.4: Matching score smoothing function $w(e)_j$ vs. $IniScore(c)$ for various δ.

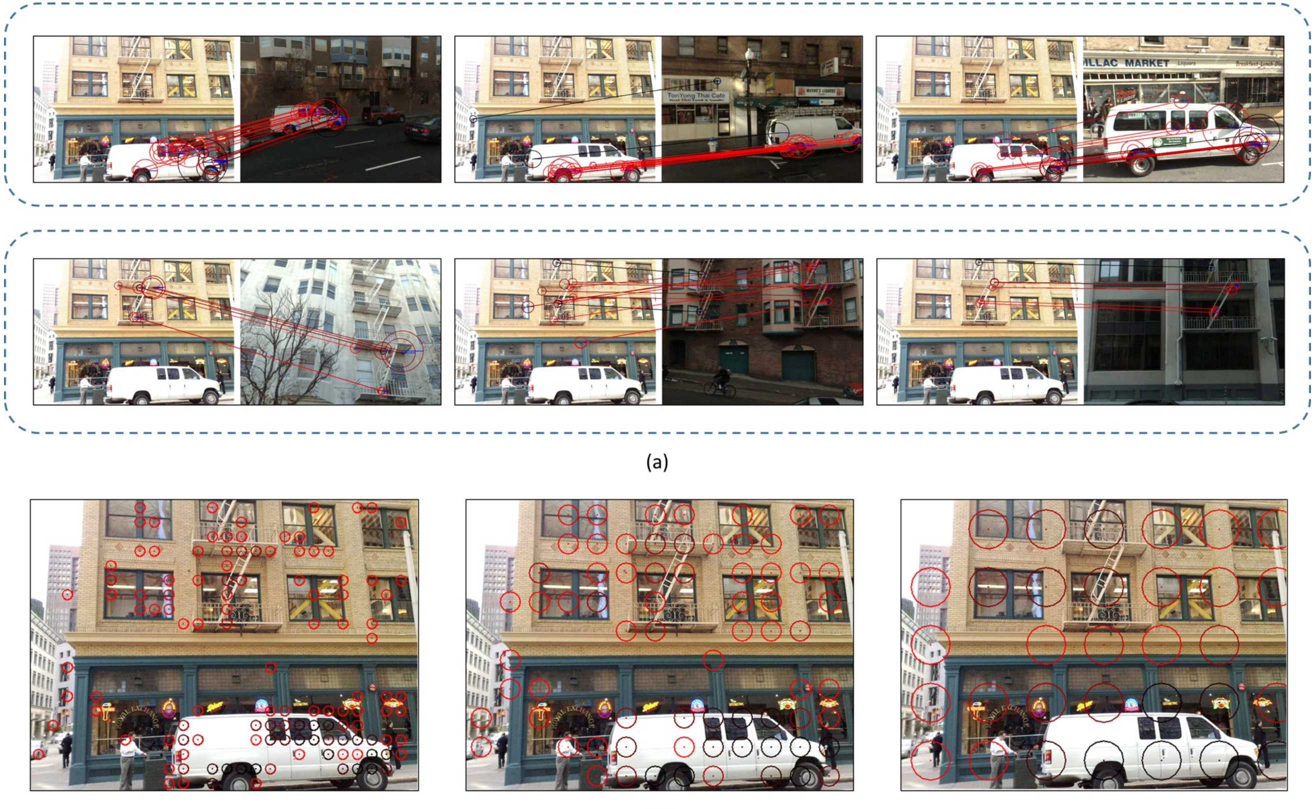
The δ parameter is set according to the general, data-set independent, behavior of the geometric verification method that is employed. Note that when $\delta = 1$ the values of $w(e)_j$ are effectively either 0 or 1, meaning that visual elements either contribute or do not contribute, rather than being weighted.

4.4.2. GEO-DISTINCTIVENESS MODELING

We start our explanation of geo-distinctiveness modeling with an illustration of the basic mechanism. Fig. 4.5(a) (top two rows) contain pairs of images. They show the correspondences between the query image (lefthand member of each pair) with images taken at locations other than the query location (righthand member of each pair). As in the case of the visual element cloud, these correspondences pick out the visual elements that we use for further modeling.

Fig. 4.5(b) (bottom row) shows how the geo-distinctiveness weights are calculated. The image is divided into regions, and a geo-distinctiveness weight is calculated per region. The three versions of the query image represent three different settings of region size, indicated by the increasing diameters of the circles. In the figure, the center of the circle indicates the center of the region, and the color indicates the weight. The color scale runs from red to black, with red indicating the most geo-distinctive regions. Examination of Fig. 4.5(b) shows the ability of geo-distinctiveness weights to focus in on specific, distinguishing features of the real world location. Visual elements corresponding to common objects occurring at multiple locations (e.g., the white delivery van and fire escape) automatically receive less weight (i.e., as shown by black).

Expressed formally, geo-distinctiveness is calculated with the following process. We divide the query image, of size $w \times h$, into non-overlapping small regions with size $\tilde{a} \times \tilde{a}$, $\tilde{a} = min(w/a, h/a)$. For completeness note that we allow right and bottom regions to be smaller than $\tilde{a} \times \tilde{a}$, in the case that w or h is not an integer multiple of a.



(a)

(b)

Figure 4.5: Illustration of geo-distinctiveness modeling. Figure (a) shows how visual elements corresponding to two common objects in the query image (white delivery van and fire escape) give rise to strong matches with images from different locations. The geo-distinctiveness of these visual elements in the query image under different region resolution is shown in Figure (b), with the color changing from black to red to represent the increase of geo-distinctiveness.

We then transfer the scale represented by each visual element from the level of the neighborhood of a salient point to the level of an image region. We carry out this transfer by mapping visual elements to the regions in which they are located. Note that the consequence of this mapping is that all visual elements contained in the same image region are treated as the same visual element. The effect of the mapping is to smooth the geo-distinctiveness of the visual elements in the query image. Changing a will change the size of the region, and thereby also the smoothing. The effect can be observed in Fig. 4.5(b), e.g., the fire escape at the top middle of the photo is less discriminative (the circle turns black) as the area becomes larger.

For each visual element e in each image in the image set I_g for location g in G, we calculate a geo-distinctiveness weight W_{Geo}. Recall that e in each image in I_g stands in a one-to-one correspondence c with a visual element in the query image. W_{Geo} is then defined as

$$W_{Geo}(e) = \begin{cases} \log(N/n(r(e))), & \text{if } n(r(e)) < \vartheta \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where N is the total number of location candidates (i.e., $|G|$), $r(e)$ is the image region of the query containing the visual element corresponding to e, and $n(r(e))$ is the total number of locations in G with an image from their image set I_g that is involved in a correspondence with any visual element occurring in the query region $r(e)$. Finally, ϑ is a threshold completely eliminating the influence of elements that have correspondences with many locations in G. The effect of parameters a and ϑ is discussed in the experimental section.

4.4.3. VISUAL MATCHING PER LOCATION

We start our discussion of visual matching by considering the patterns of visual elements associated with a true match between an query image and a location. First, we investigate whether or not we can indeed expect more visual elements in true-location visual element clouds compared to wrong-location visual element clouds. We carry out the analysis on two datasets, the San Francisco Landmark dataset and the MediaEval '15 Placing Task dataset, the geo-location estimation image sets used in our experiments, which will be described in detail in Section 4.6. Results are shown in Fig. 4.6. Here, we see that the ratio between the number of unique visual elements in a wrong-location cloud and a true-location cloud is mainly distributed between 0 and 1. The observation holds whether the top-10 ranked wrong locations are considered (solid line), or whether only the wrong location with the most visual elements is considered (dashed line). This analysis points to the ability of the number of visual elements to distinguish true from wrong locations, and motivates us to include aggregation of visual elements as part of our visual matching model.

Next, we return to our earlier statement (Section 4.2) that we expect the match between queries and a true location to display (a) a wider variety of visual elements, and (b) visual elements that are distributed over a greater area of the image, than in the case of a match with a false location. These expectations are borne out in our calculations of visual correspondences, as illustrated in Fig. 4.7. The images from the true location (lefthand side) capture a broad and diverse view of the scene and thus match different regions of the query image, e.g., the column and the bridge, as opposed to the images taken at a wrong location (righthand side) that only have correspondences with few specific visual elements, e.g., the top of the column. This pattern leads us to not simply aggregate visual elements, but se-

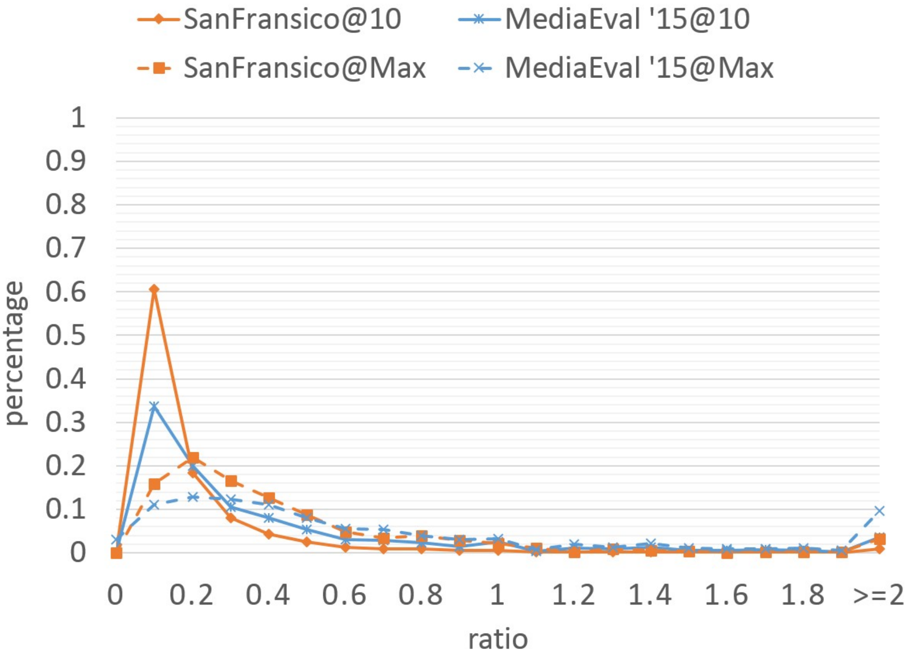


Figure 4.6: Distribution of the ratio of number of unique visual elements between wrong location and true location. The scheme with @10 means the results are calculated based on the top-10 wrong locations in the initial ranked list for each query. The scheme with @Max means the results are calculated based on the wrong location that has the maximum number of visual elements among all wrong locations in the initial ranked list.

lect them in a particular way. Specifically, for a given area of the image query, only a single visual element is allowed to contribute per location. This approach rewards locations in which visual elements are diverse and distributed over the query image.

Expressed formally, visual matching uses the following procedure. We divide the query, of size $w \times h$, into regions $\tilde{b} \times \tilde{b}$, $\tilde{b} = min(w/b, h/b)$. This splitting resembles what we used for geo-distinctiveness modeling, but serves a separate purpose in the current step. Then, in order to calculate the match between q with a candidate location image set I_g, we iterate through each region of the query image. For each region, we select the single visual element e that has the strongest matching score with images from a given location. Recalling that $\mathbf{W}_e$ are the weights of the visual correspondences with the query for image set I_g representing location g, the strongest matching score is expressed as $\tilde{w}_e = \max(\mathbf{W}_e)$. The result is a set of k visual elements. Note that although the same query image regions are used for all locations, k may vary per location, and is less than the total number of query regions in the cases where some query regions fail to have links in terms of visual elements with a location.

The final visual proximity between location g and query image q combines a visual representation of the location g and of the query q. The representation of the query uses the visually distinctive weights $W_{Geo}(e)$ from Eq. 4.3: $\mathbf{r}_q = (W_{Geo}(0), W_{Geo}(1), ..., W_{Geo}(k))$.

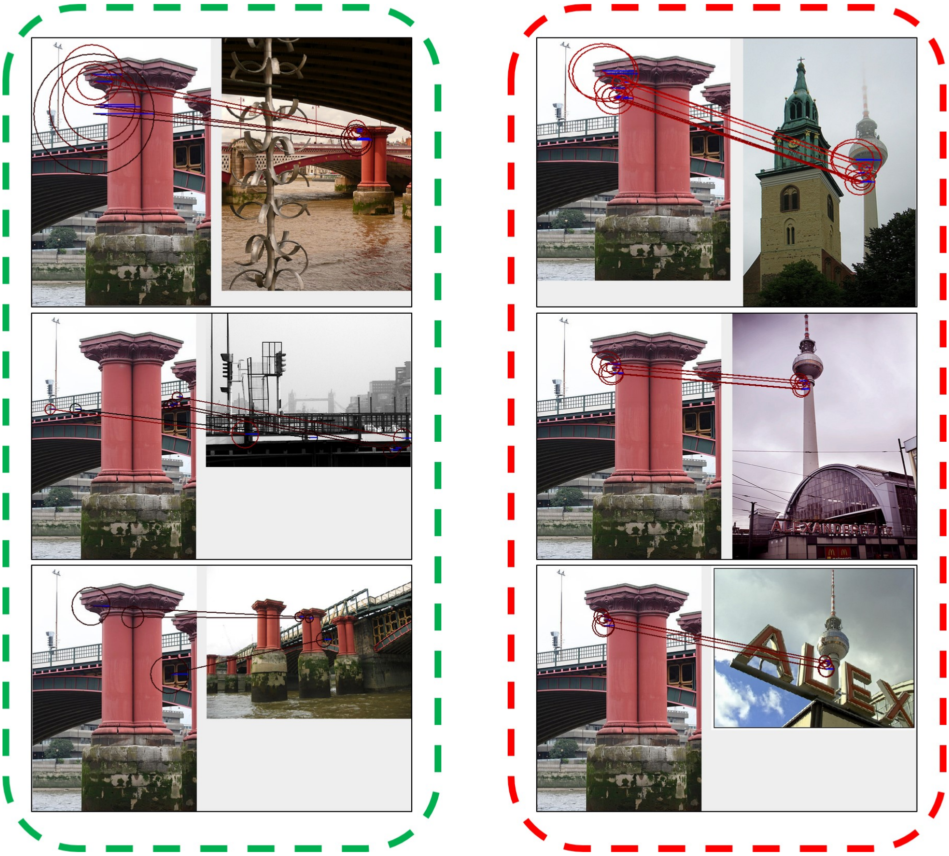


Figure 4.7: Illustration of the initial correspondence set between the query image and the photos in two different locations with the color intensity from black to red representing the increase of the strength of the initial matching score. The left photo set is from the same location as the query image.

The representation of the location combines these weights with visual matching weights $\tilde{w}_e$: $\mathbf{r}_g = (\tilde{w}_0 W_{Geo}(0), \tilde{w}_1 W_{Geo}(1), ..., \tilde{w}_k W_{Geo}(k))$. The combination is calculated as,

$$Score(g,q) = \mathbf{r}_q \cdot \mathbf{r}_g = \sum_{e \in \mathbf{E}_g} \tilde{w}_e W_{Geo}(e)^2 \tag{4.4}$$

The final location estimation for the query is calculated by ranking the locations by this score, and propagating the top-ranked location to the query.

4.5. EXPERIMENTAL SETUP

In this section, we describe the setup of our experimental framework for assessing the performance of DVEM. This provides the background for our experimental results of parameter selection (Section 4.6.1), geo-contrained location estimation (Section 4.6.2), geo-unconstrained location estimation (Section 4.6.3), and our comparison with the state of the art (Section 4.6.4).

4.5.1. DATASET

We carry out experiments on two image datasets that are commonly used in location estimation, one for the geo-constrained, and one for the geo-unconstrained image geo-location prediction scenario.

San Francisco Landmark dataset [15]: This dataset is designed for city-scale location estimation, i.e., geo-constrained location estimation. The database images (background collection) are taken by a vehicle-mounted camera moving around downtown San Francisco, and query images are taken randomly from a pedestrian's perspective at street level by various people using a variety of mobile photo-capturing devices. We use 1.06M perspective central images (PCI) derived from panoramas as the database photos, and the original 803 test images as queries. For our detailed experiments in Sections 4.6.1 and 4.6.2 we use 10% of the test images for development, and report results on the other 90% of the test images. The ground truth for this dataset consists of building IDs. The geo-location of an image is considered correctly predicted if the building ID is correctly predicted.

MediaEval '15 Placing Task dataset [30]: This dataset is designed for global scale location estimation, i.e., geo-unconstrained location estimation. It is a subset of the YFCC100M collection [38], a set of Creative Commons images from Flickr, an online image sharing platform. The background collection and the query images were randomly selected in a way that maintained the global geographic distribution within the online image sharing community. The MediaEval 2015 Placing Task dataset is divided into 4.6M training and 1M test images. Here again for our detailed experiments in Sections 4.6.1 and 4.6.3 we use 2% of the test set for development, and report results on the other 98% of the test set. The ground truth for this dataset consists of geo-coordinates, either recorded by the GPS of the capture device or assigned by hand by the uploading users. An image is considered to be correctly predicted if its predicted geo-coordinates fall within a given radius r_{eval} of the ground truth location. r_{eval} controls the evaluation precision and the tolerance of the evaluation to noise in the ground truth.

4.5.2. COMPUTING VISUAL SIMILARITY

Conceptually, we consider the visual matches between different areas of two images as evidence that their visual content reflects the same location in the physical world, possibly differing as to how they are captured, e.g., capturing angle, scale or illumination. In order to identify these areas and assess the strength of the link between their occurrences in images, we deploy our recently-developed image retrieval system [37]. This system is based on pairwise geometric matching technology and is built upon the standard bag-of-visual-words paradigm. The paradigm is known to scale up well to a large-scale datasets [34, 39, 40]. To further speed up retrieval and improve accuracy, we use pairwise geometric matching in the following pipeline of state-of-the-art solutions:

- Features & Vocabularies: Since up-right Hessian-Affine detector and Root-SIFT [40] have proven to yield superior performance, we use this feature setting to find and describe invariant regions in the image. We use exact k-means to build the specific visual world vocabularies with the size of 65,536 based on the features from the training images.

- Multiple Assignment: To address the quantization noise introduced by visual word assignment, we adopt the strategy used in [41], which assigns a given descriptor to several of the nearest visual words. As this multiple assignment strategy significantly increases the number of visual words per image, we only apply this at the query side.

- Initial ranking: We adopt the Hamming Embedding technique combined with burstiness weighting proposed in [39] in the initial ranking phase.
- Geometric verification: To find the reliable correspondences for DVEM, the pairwise geometric matching technology [37] is employed for fast geometric verification, which is reported to be the state-of-the-art in image retrieval in terms of speed and accuracy. In the experiment conducted on the development set, we found that due to a high inter-similarity of the street view images taken in downtown San Francisco, removing the correspondences with low matching score generated by pairwise geometric matching can generally help to improve the estimation. Here the threshold is set to 4.

The ranked list resulting from this computation of visual similarity is used in the Candidate Image Selection step (cf. Fig. 4.2) and for two baselines, as discussed next.

4.5.3. EXPERIMENTAL DESIGN

We carry out two different sets of evaluations that compare the performance of DVEM to the leading content-based approaches to image geo-location estimation. The first set (Sections 4.6.2 and 4.6.3) assesses the ability of DVEM to outperform other search-based geo-location estimation approaches, represented by VisNN and GVR:

- *VisNN*: Our implementation of the 1-NN approach [10], which uses the location of the image visually most similar to the query image as the predicted location. It is a simple approach, but in practice has proven difficult to beat.
- *GVR*: Method used in [13], which expands the candidate images by their locations and uses the overall visual similarity of images located in one location as the ranking score for that location. This method is chosen for comparison since it has been demonstrated to outperform other state-of-the-art approaches for geo-unconstrained location estimation [28, 29].

The second set of evaluations (Section 4.6.4) compares our methods with other state-of-art methods, which do not necessarily use a search-based framework.

Our evaluation metric is Hit Rate at top K ($HR@K$). Recall that given a query, the system returns a ranked list of possible locations. $HR@K$ measures the proportion of queries that are correctly located in the top K listed locations. Specifically, $HR@1$ represents the ability of the system to output a single accurate estimate.

4.6. EXPERIMENTAL RESULTS

We implemented our DVEM framework on top of the object-based image retrieval system [37] by constructing a Map-Reduce-based structure on a Hadoop-based cluster[1] containing 1,500 cores. The initial visual ranking (the candidate image selection step) takes

---

[1] This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

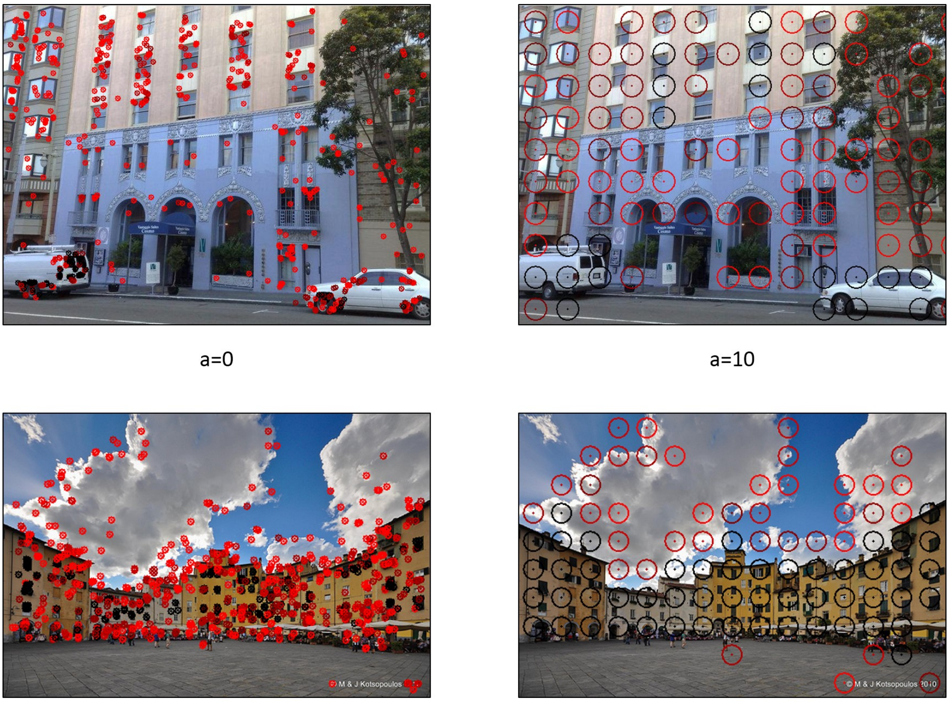
about 105 mins for San Francisco dataset (803 queries on a collection of 1.06M photos) and about 88 hours for the MediaEval '15 dataset (1M queries on a collection of 4.6M photos). The DVEM stage is executed after the initial visual ranking, and takes 2.4ms per query.

In this section, we report the experimental results and compare our DVEM method with reference methods in both areas of geo-constrained and geo-unconstrained location estimation. We use part of the test data (10% for San Francisco dataset and 2% for MediaEval '15 dataset) as development partition to set the parameters of DVEM, and use the rest of the test data to evaluate the system. Recall that the parameters are the image region size a defined in Section 4.4.2, the frequent threshold ϑ defined in Eq. (4.3) and the image region size b defined in Section 4.4.3. The parameter δ defined in Eq. (4.2) is set empirically to 5 based on the general observation that the initial correspondence score generated by pairwise geometric matching [37] usually reflects a strong match when it is above 10. As previously mentioned, the number of top-ranked images from the image retrieval system, which are used to generate the candidate locations set G, is set to 1000. Note that we use the same G for GVR.

4.6.1. IMPACT OF THE PARAMETERS

We start our series of experiments by evaluating the impact of a, b, ϑ on the system performance using our development partition. We explore the parameter space with grid search, as shown in Table 4.1. For both a and b, we considered the values 0, 30, 20 and 10 (Table 4.1, top). Note that $a = 0$ means that the system assigns a different geo-distinctiveness weight to each individual visual element, and $a = 30, 20, 10$ are regions increasing in size. Similarly, $b = 0$ means that system deploys all visual elements appearing in the images of a given location for query-location matching, and $b = 30, 20, 10$ are regions increasing in size. After 10 performance dropped dramatically, and these values were not included in the table. We choose $a = 10, b = 20$ as an operating point for the San Francisco dataset and $a = 0, b = 30$ for the MediaEval '15 dataset. For ϑ, we considered the values 4, 5, 6 and 7, but found little impact (Table 4.1, bottom). We choose $\vartheta = 5$ for the San Francisco dataset and $\vartheta = 6$ for the MediaEval dataset.

We notice that the performance is mainly influenced by the parameter a, which is used to smooth the geo-distinctiveness of the visual elements in the query. The optimal values for parameter a are different on the two datasets. An investigation of the difference revealed that it can be attributed to the difference in the respective capture conditions. The examples in Fig. 4.8 illustrate the contrast. The queries in the San Francisco dataset (Fig. 4.8, top) are typically zoomed-in images, taken on the street with a limited distance between the camera and the captured object (e.g., car or building). High levels of zoom results in the salient points that correspond to object details, e.g., a single tire on a car can have multiple salient points assigned to it. Such a high resolution of salient points may confuse object matching and is for this reason not productive for location estimation. For this reason, it appears logical that a value of a that leads to a higher level of grouping of salient points for the purposes of geo-distinctiveness assessment leads to the best performance. In contrast, the queries in the MediaEval '15 dataset that have the best potential to be geo-located (Fig. 4.8, bottom) are mostly zoomed-out images capturing a scene from a distance. The level of detail is much less than in the previous case, and the salient points tend to already pick out object-level image areas relevant for location estimation. Aggregat-



a=0

a=10

a=0

a=10

Figure 4.8: Illustration of the geo-distinctiveness of visual elements under different region resolutions for query images from San Francisco dataset (top row) and MediaEval '15 dataset (bottom row). The color changing from black to red indicates an increase in geo-distinctiveness

ing the salient points together through image splitting like in the previous case would have a negative effect, as it would reduce the resolution of salient points too drastically, leading to a loss of geo-relevant information. For this reason, it is logical that the parameter value $a = 0$ is the optimal one, reflecting that no image splitting should be carried out.

4.6.2. GEO-CONSTRAINED LOCATION ESTIMATION

The performance of different methods on the San Francisco Landmark dataset is illustrated in Fig. 4.9. DVEM consistently outperforms both VisNN and GVR across the board, with the performance gain of 3% and 4% for HR@1 with respect to the revised ground truth released in April 2014 (Fig. 4.9.b).

GVR performs even worse than VisNN with respect to the revised ground truth. This is due to the fact that in the street-view dataset the database images are captured by the survey vehicle, which can make multiple near-duplicate images per location. When a location contains same visual elements of the query image, e.g., the white van in Fig. 4.5b, the summed visual similarity of images taken in this location will heavily influence the estimation. In contrast, DVEM can handle this situation since it differentiates visual elements

Table 4.1: HR@1(%) comparison of DVEM on San Francisco (revised ground truth) and MediaEval ’15 datasets ($r_{eval} = 1km$) with different a, b, and ϑ.

	San Francisco				MediaEval ’15			
	$\vartheta = 5$				$\vartheta = 6$			
a \ b	0	30	20	10	0	30	20	10
0	81.3	80	82.5	81.3	8.1	**8.2**	8.1	8
30	80	80	81.3	80	8	7.9	7.9	7.8
20	81.3	81.3	80	80	7.8	7.8	7.8	7.6
10	80	82.5	**83.8**	**83.8**	7.2	7.3	7.3	7.2

	$a = 10, b = 20$				$a = 0, b = 30$			
ϑ	4	5	6	7	4	5	6	7
	83.8	**83.8**	**83.8**	82.5	8.1	8.1	**8.2**	8.1

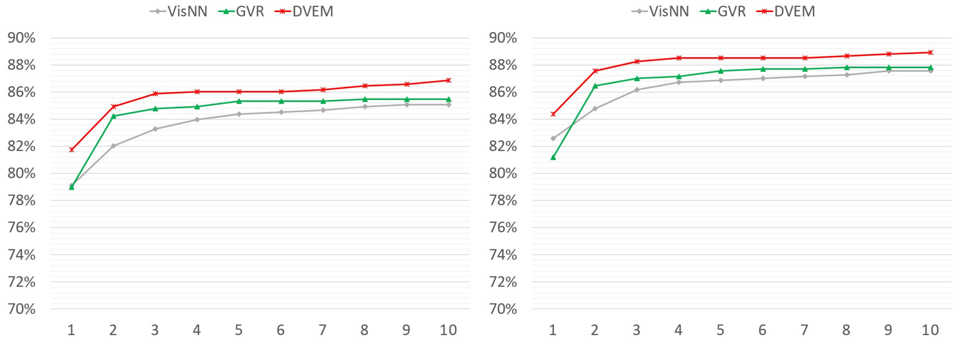
based on their geo-distinctiveness and eliminates the influence of redundancy by matching not at the image level, but rather at the level of the visual element cloud.

We note that, as 52 out of 803 (6.5%) query images do not correspond in location to any images in the database collection. Consequently, the maximal performance that can be reached is 93.5%. In addition, the ground truth is automatically labeled based on building ID, which is generated by aligning images to a 3D model of the city consisting of 14k buildings based on the location of the camera [15]. This introduces noise into the ground truth. We conducted a manual failure analysis on the 74 queries for which DVEM makes wrong estimation with respect to HR@1. We found that for 9 queries, the ground-truth database images are irrelevant, and for 32 queries, the database images located in the top-1 predicted location are relevant, but their building ID is not included in the ground truth. This makes the maximum performance that could be achieved by DVEM an HR@1 of 88.3%.

4.6.3. GEO-UNCONSTRAINED LOCATION ESTIMATION

Fig. 4.10 shows the performance of different methods with different evaluation radiuses (Fig. 4.10a.) and different hit rates (Fig. 4.10b.) on the MediaEval ’15 Placing Task dataset. This figure demonstrates that DVEM consistently outperforms both VisNN and GVR. The gain in performance is 12% over VisNN and 5% over GVR for HR@1.

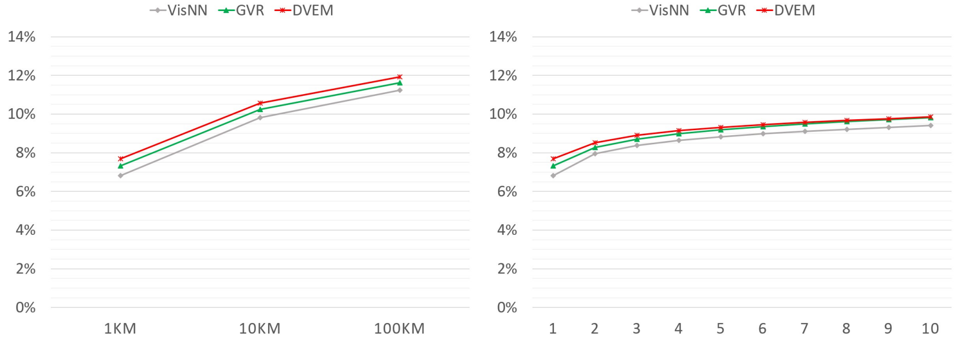
Next we turn to investigate in more detail why VisNN is outperformed by GVR, which is in turn outperformed by our new DVEM approach. In general, GVR outperforms VisNN because it can leverage the existence of multiple images from the true location that are visually similar to the query. GVR fails, however, when wrong locations also are associated with multiple images that are visually similar to the query. DVEM, however, is able to maintain robust performance in such cases. Fig. 4.11 contains an example that illustrates the difference. The query q is shown on the left. VisNN is illustrated by row (a), which contains the top-10 images returned by VisNN. There is no correct image for the query location among



(a) original ground truth

(b) revised ground truth

Figure 4.9: HR@k performance for varying k on the SanFrancisco street view dataset. (a) performance with respect to the original ground truth, (b) performance with respect to the revised ground truth released on April 2014.



(a)

(b)

Figure 4.10: Performance on the MediaEval '15 Placing Task dataset. (a) HR@1 with respect to different evaluation radiuses, (b) HR@k performance for varying k and for the evaluation radius of $1km$.

them. This reflects that the collection lacks a single good image-level visual match for the query. GVR is illustrated by row (b), which contains five sets of images from the five top-ranked candidate locations. We see the top-1 candidate location image set contains many images similar to the query, although it is not the true location. Instead, the true location, whose candidate location image set also contains many images, is ranked second. DVEM is illustrated by row (c), which again contains five candidate location image sets. This time, the correct location is ranked first. We can see that the DVEM decision avoided relying too heavily on the distinctive floor pattern, which is common at many tourist locations, and cause GVR to make a wrong prediction. Instead DVEM is able to leverage similarity matches involving diverse and distributed image areas (such as the ceiling and the alcoves in the walls), favoring this evidence over the floor, which is less geo-distinctive.

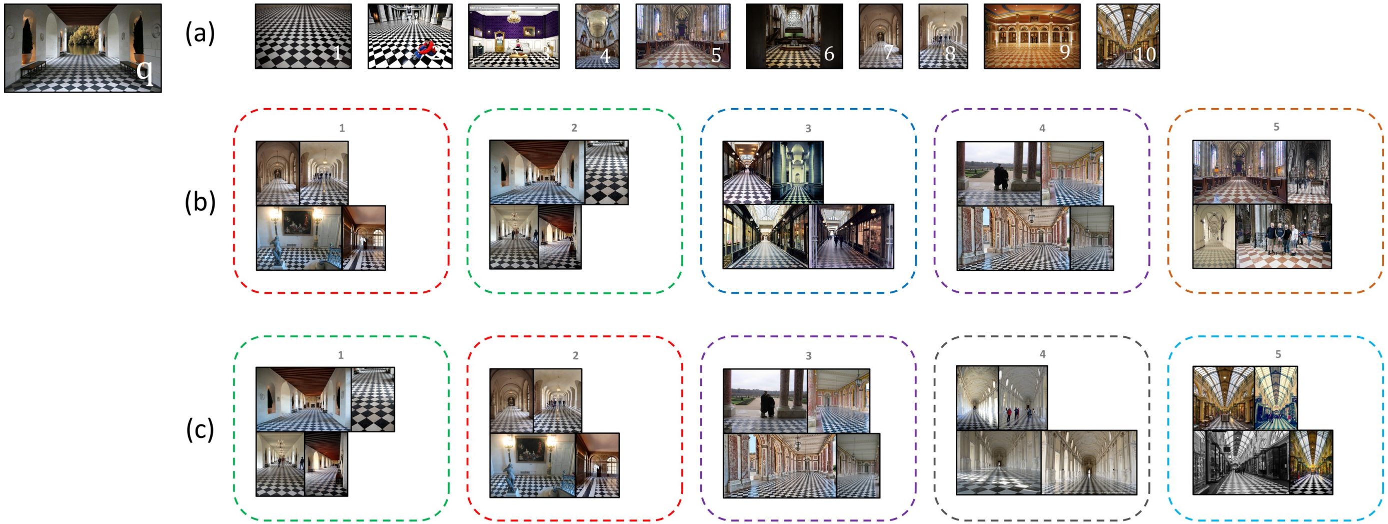


Figure 4.11: Illustration of the relative performance among the methods VisNN, GVR and DVEM on the MediaEval '15 Placing Task dataset: (a) the initial visual rank of top-10 most similar photos for a given query, the location of the top ranked photo is the result of VisNN, (b) ranked candidate locations using GVR, (c) ranked candidate locations using DVEM. There are maximum 4 photos shown for each location.

4.6.4. COMPARISON WITH THE STATE-OF-THE-ART

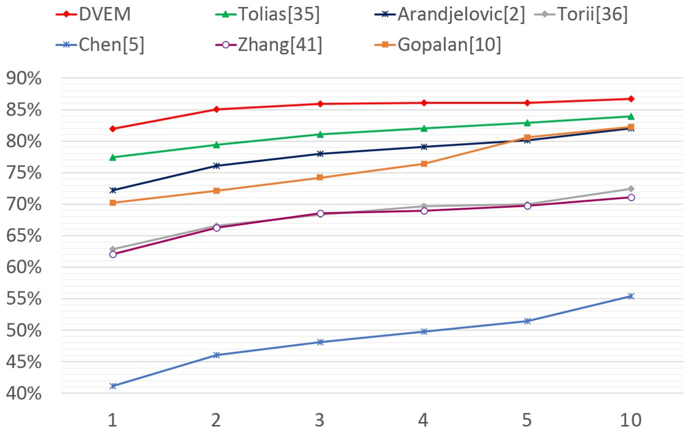
In this experiment, we compare DVEM with other state-of-the-art location estimation systems regarding both the geo-constrained and geo-unconstrained case. We compare our results with the top results that have been reported by other authors on the two experimental datasets that we use.

As the reference methods for geo-constrained location estimation, we use the work of Gopalan (2015) [23], Tolias et al. (2015) [24], Arandjelović and Zisserman (2014) [25], Torii et al. (2013) [16], Zhang et al. (2012) [26] and the initial work of Chen et al. (2011) [15]. Results are reported on the entire test set as defined with the San Francisco dataset release. This set is identical to the sets on which these authors report their results. The results in Fig. 4.12 demonstrate that our proposed DVEM approach outperforms the state-of-the-art on the *San Francisco* dataset. For completeness, we include additional discussion of our experimental design. The papers cited in Fig. 4.12 use a variety of tuning methods, which are sometimes not fully specified. We assume that these tuning methods are comparable to our choice, namely to use, 10% of the test data (Section 4.6.1). Referring back to Table 4.1, we can see that our demonstration of the superiority of DVEM is independent of this assumption. In the table, we see that the difference in performance for DVEM for the best and the worst parameter settings is less than 4% absolute. If the performance of a very poorly tuned version of DVEM falls by this amount, it still remains competitive with well-tuned versions of the other approaches in Fig. 4.12. This assures us that the superiority of our approach does not lie in our choice of tuning.

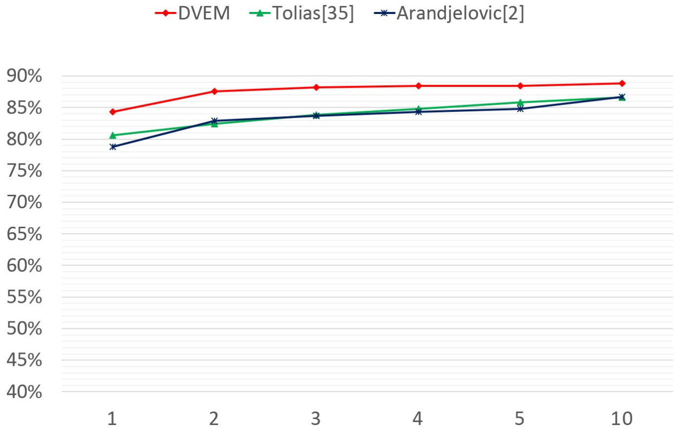
For geo-unconstrained location estimation, we compare our method to Li et al. [32], and the neural network-based representation-learning approach by Kordopatis-Zilos [33]. Results are reported on the entire test set as defined by the data release made by the MediaEval 2015 Placing Task. The results in Fig. 4.13 show that our DVEM system redefines the state-of-the art on the MediaEval '15 dataset. Again, for completeness, we include additional discussion of our experimental design. The submissions to the MediaEval 2015 Placing Task are not allowed to tune on the test data. They do, however, have access to a leaderboard which includes 25% of the test data. In 2015, teams made a limited number of submissions to the leader board (<= 3). Our experimental design was different in that we tuned on 2% of the test data. However, again referring back to Table 4.1 we can see the magnitude of the advantage that this choice gave us. The worst parameter settings yielded performance that was lower than that of the best parameter settings by 1% absolute. If the performance of a very poorly tuned version of DVEM falls by this amount, it would still outperform its competitors in Fig. 4.13. We point out that the independence of the superiority of DVEM from the way in which the parameters are set can be considered a reflection of an observation already made above: the choice of the critical parameter a is dependent on how data was captured in general (i.e., zoom-in vs zoom-out) and not on the specific composition of the dataset.

4.7. CONCLUSION

We have presented a visual-content-based approach for prediction of the geo-locations of images, based on common sense observations about challenges presented by visual patterns in image collections These observations led us to propose a highly transparent approach that represents locations using visual element clouds representing the match be-



(a) original ground truth



(b) revised ground truth

Figure 4.12: HR@k performance for varying *k* on the SanFransico street view dataset. (a) performance with respect to the original ground truth, (b) performance with respect to the revised ground truth released on April 2014.

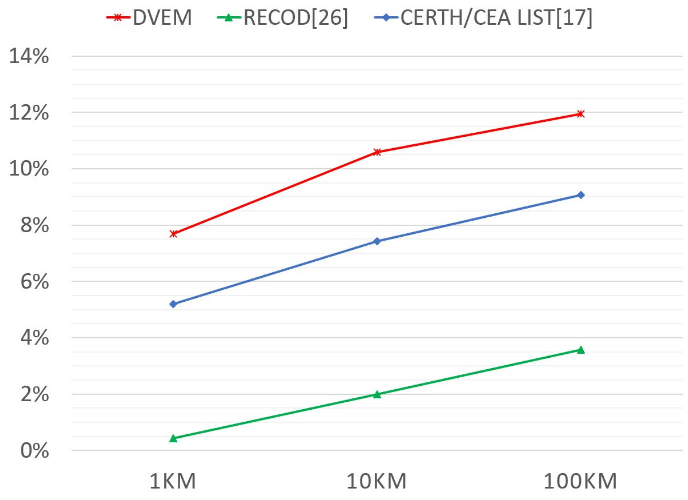


Figure 4.13: HR@1 performance with respect to different evaluation radiuses on the MediaEval '15 Placing Task dataset.

tween a query and a location, and leveraging geo-distinctiveness. Our evaluation, conducted on two publicly available datasets, demonstrates that the proposed approach achieves performance superior to that of state-of-the-art approaches in both geo-constrained and geo-unconstrained location estimation.

We close with two additional observations about the value of the proposed DVEM approach moving forward. A key challenge is that the distribution of image data used for geo-unconstrained location prediction is highly sparse over many regions. This sparsity has led to the dominance of search-based approaches such as DVEM over classification approaches, already mentioned above. An additional consequence, we expect, is that the search-based framework will remain dominant, and that new, deep-learning approaches will contribute features, as in [33], which can enhance, but will not replace, DVEM. Note that because DVEM calculates representations over a 'contextual' image set, rather than the whole collection, it is not forced to pre-define locations of a particular scale. The result is that DVEM is able to apply geo-distinctiveness to predict the location of images on a continuous scale, limited only by the visual evidence present in the data set.

REFERENCES

[1] S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *IEEE Trans. Multimedia*, 15(4):921–932, 2013.

[2] Jiajun Liu et al. Presenting diverse location views with real-time near-duplicate photo elimination. In *Proc. ICDE '13*, 2013.

[3] Yan-Ying Chen, An-Jung Cheng, and W.H. Hsu. Travel recommendation by mining people attributes and travel group types from community-contributed photos. *IEEE Trans. Multimedia*, 15(6):1283–1295, 2013.

[4] Xiangyu Wang, Yi-Liang Zhao, Liqiang Nie, Yue Gao, Weizhi Nie, Zheng-Jun Zha, and Tat-Seng Chua. Semantic-based location recommendation with multimodal venue semantics. *IEEE Trans. Multimedia*, 17(3):409–419, 2015.

[5] Jianchao Yang, Jiebo Luo, Jie Yu, and T.S. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Trans. Multimedia*, 14(6):1642–1651, 2012.

[6] Junsong Yuan, Jiebo Luo, and Ying Wu. Mining compositional features from GPS and visual cues for event recognition in photo collections. *IEEE Trans. Multimedia*, 12(7):705–716, 2010.

[7] Jaeyoung Choi, Eungchan Kim, Martha Larson, Gerald Friedland, and Alan Hanjalic. Evento 360: Social event discovery from web-scale multimedia collection. In *Proc. MM '15*, 2015.

[8] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing Flickr photos on a map. In *Proc. SIGIR '09*, 2009.

[9] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proc. WWW '09*, 2009.

[10] J. Hays and A.A. Efros. IM2GPS: estimating geographic information from a single image. In *Proc. CVPR '08*, 2008.

[11] Tao Guan, Yunfeng He, Juan Gao, Jianzhong Yang, and Junqing Yu. On-device mobile visual location recognition by integrating vision and inertial sensors. *IEEE Trans. Multimedia*, 15(7):1688–1699, 2013.

[12] Martha Larson et al. Automatic tagging and geotagging in video collections and communities. In *Proc. ICMR '11*, 2011.

[13] Xinchao Li, Martha Larson, and Alan Hanjalic. Global-scale location prediction for social images using geo-visual ranking. *IEEE Trans. Multimedia*, 17(5):674–686, 2015.

[14] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[15] D.M Chen et al. City-scale landmark identification on mobile devices. In *Proc. CVPR '11*, 2011.

[16] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *Proc. CVPR '13*, 2013.

[17] Jing Li et al. GPS estimation for places of interest from social users' uploaded photos. *IEEE Trans. Multimedia*, 15(8):2058–2071, 2013.

[18] Kalantidis Yannis et al. VIRaL: Visual image retrieval and localization. *Multimedia Tools and Applications*, 51:555–592, 2011.

[19] Yunpeng Li, D.J. Crandall, and D.P. Huttenlocher. Landmark classification in large-scale image collections. In *Proc. ICCV '09*, 2009.

[20] Weiqing Min, Changsheng Xu, Min Xu, Xian Xiao, and Bing-Kun Bao. Mobile landmark search with 3d models. *IEEE Trans. Multimedia*, 16(3):623–636, 2014.

[21] Martha Larson et al. The benchmark as a research catalyst: Charting the progress of geo-prediction for social multimedia. In *Multimodal Location Estimation of Videos and Images*. Springer, 2015.

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

[23] Raghuraman Gopalan. Hierarchical sparse coding with geometric prior for visual geo-location. In *Proc. CVPR '15*, 2015.

[24] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, pages 1–15, 2015.

[25] Relja Arandjelović and Andrew Zisserman. Dislocation: Scalable descriptor distinctiveness for location recognition. In *Proc. ACCV '14*, 2014.

[26] Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N Metaxas. Query specific fusion for image retrieval. In *Proc. ECCV '12*, 2012.

[27] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proc. CVPR '13*, 2013.

[28] Xinchao Li, Michael Riegler, Martha Larson, and Alan Hanjalic. Exploration of feature combination in geo-visual ranking for visual content-based location prediction. In *Proc. MediaEval '13*, 2013.

[29] Xinchao Li, Martha Larson, and Alan Hanjalic. Geo-visual ranking for location prediction of social images. In *Proc. ICMR '13*, 2013.

[30] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. The placing task at mediaeval 2015. In *MediaEval 2015 Workshop*, 2015.

[31] Pascal Kelm et al. Imcube @ MediaEval 2015 Placing Task: A Hierarchical Approach for Geo-referencing Large-Scale Datasets. In *Proc. MediaEval '15*, 2015.

[32] Lin Tzy Li et al. RECOD @ Placing Task of MediaEval 2015. In *Proc. MediaEval '15*, 2015.

[33] Giorgos Kordopatis-Zilos et al. CERTH/CEA LIST at MediaEval Placing Task 2015. In *Proc. MediaEval '15*, 2015.

[34] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. ICCV '03*, 2003.

[35] Carl Doersch et al. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[36] Quan Fang, Jitao Sang, and Changsheng Xu. Giant: Geo-informative attributes for location recognition and exploration. In *Proc. MM' 13*, 2013.

[37] Xinchao Li, Martha Larson, and Alan Hanjalic. Pairwise geometric matching for large-scale object retrieval. In *Proc. CVPR '15*, 2015.

[38] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[39] Herve Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Proc. CVPR '09*, 2009.

[40] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR '12*, 2012.

[41] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

5

DISCUSSION AND FUTURE WORK

5.1. DISCUSSION

THE objectives of the research reported in this thesis were to develop a scalable visual content-based location estimation system for images, to investigate the possibilities to improve its accuracy and reliability to a substantial extent, and to achieve this in both the geo-constrained and geo-unconstrained scenario. We pursued these objectives from different perspectives ranging from high level framework design to optimization of specific components of the system. We organized these perspectives as the three main chapters in this thesis. While Chapter 2 introduced a generic framework for visual content-based location estimation, Chapter 3 and Chapter 4 focused on the development of two key components within the framework. Specifically, Chapter 3 covered geometric verification for finding reliable candidates for relevant geo-annotated images in the collection, and Chapter 4 covered geo-distinctive visual element discovery and matching for selecting the candidate location that most likely corresponds to the visual scene depicted in the query image. In this section, we reflect on the algorithmic solutions that we developed and the results that we obtained.

Our research can be considered to have been motivated by two underlying questions:

- How to approach the design of a general framework for automatically estimating the location of a visual scene depicted in an image?
- How to make such a framework capable of estimating locations at a global scale, i.e., when the target location is not constrained to a set of predefined locations typically characterized by specific visual scene elements?

Following the rationale given in Section 1.4, we focused on a search-based framework for location estimation, in which the target image serves as a query to be matched with the geo-annotated images in the available large-scale collection. Then, based on the visual matching between the query and collection images, information is derived about the most likely location present in the image collection that may resemble the visual scene of the query image. For this method to work adequately, both scalability and robustness need to be considered with great care. Scalability implies that the location estimation of any given query image can be carried out quickly, independently, of the complexity of the depicted scene and the size of the reference collection of geo-annotated images. Robustness is related to the ability of the system to handle the absence of a one-to-one relation between the location and the visual properties of an image. Two images taken at the same location can be visually completely different, while visually similar images can be found that were taken at different locations. In order to neutralize the robustness concern, we developed, in Chapter 2, a geo-visual ranking method that incorporates the fact that, compared to the images from the wrong location, more images from the true location will likely contain more elements of the visual content of the query image. Note that the geo-visual ranking method can also be seen as a simplified graph that explores the first order neighborhood relation in terms of geo-location and visual content. This method is preferred over exploring a full graph, which is usually time consuming on large scale image dataset. The evaluation carried out on a publicly available dataset containing 8.8M images demonstrates that the devised geo-visual ranking approach achieves sound performance for geo-location estimation and significantly outperforms the state-of-the-art in its approach category.

The results presented in Chapter 2 indicate that the proposed geo-visual ranking framework is conceptually suitable as a solution for geo-unconstrained location estimation. However, in order to elevate its robustness to an acceptable level, deeper investigation was needed regarding ways of representing images for visual matching and of assessing the quality of the match. Already in Chapter 2, it was shown that local features for image representation are more suitable than global features, which led us to consider the search-based geo-location estimation problem to be related to an object-based image retrieval problem. However, what remained open was the question of how to improve the efficiency and reliability of the object-based image retrieval system in the context of a large-scale image collection. This question was addressed in Chapter 3, in which we focused on spatial verification as being the key component for achieving high precision in an object-based image retrieval system. Spatial verification is used to re-rank the initial ranked list of images visually matching the query. The re-ranking is based on geometric constraints that are deployed to assess the validity of matches between the corresponding objects in two images. We found that the high number of outliers in the initial correspondences generated by bag-of-features and errors in the scale, rotation and position of the detected features hinder the fit of a specific transformation model (e.g., RANSAC-based mode fitting). As an alternative, we devised a model-free method which implicitly verifies the correspondences with respect to their consistency in the geometric transformation space. The devised method uses global scale and rotation relations to enforce the local consistency of geometric relations derived from the locations of pairwise correspondences. In this way, it encodes not only the scale and rotation information derived from the local points, but also their locations. By mapping locations of points to pairwise rotation and scale, the approach is more tolerant to detection noise. At the same time, using a number of filtering steps, the devised method significantly reduces the number of correspondences that must be considered, which makes it possible to maintain high image matching reliability at a substantially reduced computational cost. The experimental results on three publicly available datasets indicate that the proposed method makes the spatial verification more tractable in case of a large image collection, but also more reliable, which leads to an overall significant improvement of the object retrieval performance compared to state-of-the-art methods.

With the improved object-based image retrieval system devised in Chapter 3, we moved forward towards our overall goal in this thesis, that is, to develop a scalable visual content-based location estimation system for images. Because it is common for the query image to capture a scene comprising multiple objects, and each of them can be found in some locations, we faced a new challenge that was investigated in Chapter 4, namely how to differentiate between the objects contained in the query image that are useful or not useful to the geo-location estimation problem and how to combine the different location clues from the relevant objects, possibly collected over several images from the collection, to make the final location estimation. In order to address this challenge, we investigated the geographical (geo-)distinctiveness of the visual elements contained in the query image, and developed a geo-distinctive visual element matching method as a component of the location ranking step of our general search-based location estimation framework from Chapter 2. We approached this by following two main principles. First, we searched for ways in which geo-distinctiveness of visual elements can be modeled reliably and computed efficiently to be taken into account in the assessment of the match between two images. Second, we searched for a comprehensive visual representation for each location with respect to the query image by aggregating the visual evidence (the matched visual elements) from the images taken at the same location. In the end, we searched for a model to combine the geo-distinctiveness with aggregated visual evidence to significantly improve the location estimation. The resulting location estimation system exploits geo-distinctiveness of visual elements found in the query image and further strengthens the support for finding the true location by devising an aggregated visual representation of a location that combines all visual elements from the query found in the images of that location. The evaluation conducted on two publicly available datasets demonstrates that the proposed approach redefines the state-of-the-art in both geo-constrained and geo-unconstrained location estimation. The superiority of the proposed approach compared to the state-of-the-art solutions becomes evident in a context in which a large number of common visual elements appearing at multiple locations and a high degree of visual duplication among photos taken in the same location bias the location estimation. The proposed approach handles this aspect by differentiating the visual elements based on their geo-distinctiveness and conducting visual matching per location to remove the redundancy and aggregate the visual evidence from multiple view angles.

5.2. DIRECTIONS FOR FUTURE RESEARCH

Based on the findings presented in this thesis, we would like to make the following recommendations for future work which we think are substantial and promising for large scale image retrieval and geo-location estimation.

1. Encoding pairwise geometric relations into visual representations

In Chapter 3, we presented a geometric verification method for large-scale object retrieval. This method exploits the geometric relations between matched salient points in order to improve the initial ranked list of images generated solely on the basis of the appearance of their salient points without geometric constraints. In order to improve the overall robustness and speed up the retrieval process, one could consider encoding the pairwise geometric relation together with the typical TF-IDF weighted BOF visual representation already in the initial ranking step. While there have been initial efforts in the direction of encoding geometric information into the visual representation [1, 2], the improvements gained from these efforts were rather limited, mainly due to the encoded geometric information being too weak. In contrast, as pairwise geometric relations were shown to be more effective for geometric verification, incorporating these relations into the initial ranking step could be more effective. The challenge related to this is how to encode the geometric information into the visual representation in the way that pairwise geometric relations can be exploited.

2. Incorporating pixel-level object class knowledge into object retrieval

Building on recent breakthroughs in semantic segmentation and fine-grained localization using deep learning techniques, attempts have been made to simultaneously detect and segment objects contained in an image [3, 4], from which the pixel-level object class knowledge is available. This knowledge can then serve as the context information in object retrieval and provide guidance for making geometric constraints when building correspondences between images. For example, salient points could be marked with the class labels based on their position in the image, and one geometric constraint could be that two salient points from two different classes should not match each other. We believe that it would be promising to investigate the ways to incorporate pixel-level object class knowledge into object retrieval, and then specifically within the application domain of location estimation. The challenge in this direction is how to master the pixel-level object class knowledge which usually contains noise as a result of the immature semantic segmentation, and maximize its contribution to object retrieval.

3. Learning visual representations for object retrieval

Thanks to the recent breakthrough in deep learning and image recognition, learned visual representations from Convolutional Neural Networks (CNNs) is becoming a successful alternative to hand-crafted descriptors in the context of image classification and object detection. On the other hand, in object-based image retrieval, the current approaches exploiting CNN learned features [5, 6] still struggles to outperform the start-of-the-art which relies on precise image region matching using hand-crafted descriptors. It is for this reason that we rely on conventional descriptors to build our current system. However, along with the progress of research in learning visual representations, we expect a success of learned visual representations in object retrieval. The reason is that such learned features (typically multilayer) not only include low-level descriptions on local, small image regions, e.g., corners, but also high-level abstraction about the global image content, which is usually hard to for a human observer to design. As visual representations are building blocks in our system, we expect an additional boost from every advance in visual representation. Future research in this direction could be conducted in several ways: investigating the contributions of the different level image representations learned from CNN, adapting the network structure, and customizing the learning targets of the network, all in the context of object retrieval.

4. Coverage versus Scalability

Our choice for a search-based approach to location estimation requires that there is at least one image in the geo-tagged image collection that is taken at the query's location and that has sufficient overlap with the query in terms of its visual content. As discussed in the experimental results section of Chapter 2, about 80% of queries do not meet this requirement in the case of social image collections. In other words, for 80% of queries there is no collocated visual counterpart in the collection of geo-annotated images. Therefore, in order to create context for the presented framework to work to the best of its ability, the coverage of the collection of geo-annotated images serving as references needs to be improved. A straightforward way to handle this issue is to add as many images per location as possible. This strategy will, however, make the size of the image collection grow even further and introduce new challenges related to the scalability of the approach. Alternatively, future research in this direction could aim at generating compact but comprehensive visual representations of locations based on the available images, which would speed up the computation of matching with the query. Note that this idea of generating a compact but comprehensive visual representation of location is different from the current model-based approaches discussed in Section 1.4. Model-based approaches aggregate all images in one location into one class, and try to learn a discriminative visual representation for this location. However, this way of learning visual representations tends to focus on frequently occurring visual scenes, and thus the visual representation can hardly be comprehensive.

5. Generating confidence scores for location estimation

The uncertainty of the presence of collocated visual counterparts of the query in the geo-annotated reference collection may also call for research on models that could generate a confidence score indicating to which extent submitting a particular query image to the system is likely to result in a successful location estimation. This confidence score can be interpreted as the probability that an estimated location falls in a certain geographical radius of the ground truth, or the score can be interpreted as the error in distance of the location estimated [7]. To this end, it would be useful to investigate how to estimate the confidence/difficulty for a given query with respect to the background image collection used. As this application level problem is related to the research topic of query performance prediction in the classical information retrieval [8, 9], further research tackling this problem can get started by investigating how to extend the existing knowledge and best practices in query performance prediction to the particular application context of location estimation.

6. Location-oriented evaluation metrics

In this thesis, we solely focus on the geo-graphical characteristic of one location which is the geo-coordinates without further linking it with other socially or politically determined territorial entity, e.g., street, city, country. Depending on the application scenario, users may care more about whether the location estimation falls in the correct street or city, rather than within a certain distance radius. In this case, this user need should be reflected in the evaluation metrics used to measure the performance of a geo-location estimation algorithm. In the MediaEval Placing Task 2015 [10], a new evaluation metrics was introduced, which measures performance based on whether the estimation falls in the same entity with the true location, the entity can be street, city, state, and country. Note that solely rely on location entity can be problematic on the boundary of the entity. For example, suppose estimation error in distance is 1km, and query's location is at the border of one city, the prediction is only 1km away from the true location, but it falls in the boundary of another city. So the geographic distance is small, but the distance in terms of entity is large. Again, depending on the application scenario, a proper location-oriented evaluation metrics needs to be defined which should include distance-based and/or entity-based measurement.

5.3. CONCLUSION

In this chapter, we have discussed the algorithmic solutions developed in this thesis, our general findings, and the challenges and opportunities that we think are substantial and promising for future research. With the continued development in geo-aware social media, we expect that there will be more research effort dedicated to tackle these challenges and make significant progress in large-scale image retrieval and geo-location estimation.

REFERENCES

[1] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Spatial-bag-of-features. In *Proc. CVPR '10*, 2010.

[2] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[3] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proc. CVPR '15*, 2015.

[4] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *Proc. ICCV '15*, 2015.

[5] Konda Reddy Mopuri and R. Venkatesh Babu. Object level deep feature pooling for compact image representation. In *Proc. CVPR '15*, 2015.

[6] Tolias Giorgos, Sicre Ronan, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *Proc. ICLR '16*, 2016.

[7] Claudia Hauff, Bart Thomee, and Michele Trevisiol. Working Notes for the Placing Task at MediaEval 2013. 2013.

[8] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A survey of pre-retrieval query performance predictors. In *Proc. CIKM '08*, 2008.

[9] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. Predicting query performance. In *Proc. SIGIR '02*, 2002.

[10] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. The placing task at mediaeval 2015. In *MediaEval 2015 Workshop*, 2015.

ACKNOWLEDGEMENTS

THE four years of Ph.D. life is a vital period in becoming a professional researcher. This journey is hard and challenging in nature, as it is meant to sharpen one's research skills. The journey involves a huge amount of motivation, determination and hard work. Fortunately, during my journey pursuing the Ph.D, I have had so many prominent and kind people to be with me as my supervisors, labmates, friends and family. I would like to acknowledge all of them for their great help.

First of all, I would like to express my highest gratitude to my supervisors, Alan and Martha, for your step by step guidance, essential support and warm encouragement during the whole progress of my Ph.D. research. I enjoyed the strategic discussions with you, which helped me to clarify my research questions and illuminated my way towards convergence. I appreciated your great help on my scientific writing, which always allows the paper itself to shine. Your passion dedicated to research motivated me to move forward, and it was a great pleasure to work with you.

I am very glad to be a member of Multimedia Computing Group, where I have had excellent colleagues to work with. Yue, Steven, Peng and Christoph: Thanks for your inspiring discussions and suggestions about my research topic, and I always enjoyed the brainstorms with you. To Wen, Raynor and Jeroen: Thanks for always being ready to help me with my technical questions. It was such wonderful brain exercise to discuss with you the optimization of design pattern for large-scale data processing. Jaeyoung and Michael: Thanks for your efficient teamwork. I appreciated the cooperation with you. Claudia: Thanks for introducing me to the world of MapReduce for big data. It always makes me feel excited when I master large scale data. Juan, Cynthia, Judith, Bo, Alessio, Christina, Yi, Babak, Karthik, Ernestasia and Soude: It was truly enjoyable to have all of you as my colleague and friends, and I enjoyed a lot of fun while being with you in our group. Thanks for all your time and support.

Living in a foreign country for four years can be very hard for a homesick person. I am so grateful to have a warm Chinese community to be with me in the Netherlands. Xialu, Qiushi, Meiru, Yunlong, Zhijie, Xinyu, Lu, Feifei, Yuanhao, Wenjie, Yazhou, Gunagliang, Sha, Kaihua, Changgong, Jingtang, Yunlong, Weiling and Xiangrong: Thanks all of you for so many things.

Just like what the sunshine means to the sunflower, the love from my family is always my source of energy and sense of goal in life. It wouldn't have even been possible for me to start this journey without all the support from my family. Yuan, I am so lucky to have you be with me to cherish the same ideals and enjoy every piece of flavor on this journey. Deepest thanks for all your warmest love. 爸爸妈妈，你们一直以来坚定的支持和无微不至的关怀支撑着我从人生的一个胜利走向另一个胜利，深深的感谢你们为我所付出的一切！

CURRICULUM VITAE

06-10-1985	Born in Jinan, China.

EDUCATION

2004–2008	Shandong University of Technology, China B.Sc. in Electronic Information Engineering (Excellent Graduate)
2008–2011	Shandong University, China M.Sc. in Information Science and Engineering (Outstanding Master's Thesis) Supervisor: Prof. Ju Liu Thesis: Step Projection-Based Spread Transform Dither Modulation for Digital Watermark
2011–2015	Delft University of Technology, the Netherlands Ph.D. in Computer Science Supervisors: Prof. Alan Hanjalic and Prof. Martha Larson Thesis: Large-scale Image Retrieval for Geo-location Estimation

PUBLICATIONS

1. Xinchao Li, Martha Larson and Alan Hanjalic, Geo-distinctive Visual Element Matching for Location Estimation of Images, submitted to *IEEE Transactions on Multimedia*, 2016.

2. Xinchao Li, Peng Xu, Yue Shi, Martha Larson and Alan Hanjalic, Simple Tag-based Subclass Representations for Visually-varied Image Classes, *Proc. International Workshop on Content-based Multimedia Indexing (CBMI '16)*, Bucharest, Romania, 2016.

3. Xinchao Li, Martha Larson and Alan Hanjalic, Global-Scale Location Prediction for Social Images using Geo-Visual Ranking, *IEEE Transactions on Multimedia*, 17(5): 674-686, 2015.

4. Xinchao Li, Martha Larson and Alan Hanjalic, Pairwise Geometric Matching for Large-scale Object Retrieval, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, Boston, US, 2015.

5. Jaeyoung Choi and Xinchao Li, The 2014 ICSI/TU Delft Location Estimation System, *Proc. MediaEval 2014 Workshop*, Barcelona, Spain, 2014.

6. Xinchao Li, Martha Larson and Alan Hanjalic, Geo-visual ranking for location prediction of social images, *Proc. International Conference on Multimedia Retrieval (ICMR '13)*, Dallas, US, 2013.
7. Xinchao Li, Michael Riegler, Martha Larson and Alan Hanjalic, Exploration of feature combination in geo-visual ranking for visual content-based location prediction, *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, 2013.
8. Xinchao Li, Claudia Hauff, Martha Larson and Alan Hanjalic, Preliminary Exploration of the Use of Geographical Information for Content-based Geo-tagging of Social Video, *Proc. MediaEval 2012 Workshop*, Pisa, Italy, 2012.
9. Xinchao Li, Ju Liu, Jiande Sun and Xiaohui Yang, Step Projection-Based Spread Transform Dither Modulation, *Journal of IET Information Security*, 5(3): 170-180, 2011.
10. Xiaohui Yang, Ju Liu, Jiande Sun, Xinchao Li, Wei Liu and Yuling Gao, DIBR based view synthesis for free-viewpoint television, *Proc. IEEE 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON '11)*, Antalya, Turkey, 2011.
11. Xinchao Li, Ju Liu, Jiande Sun, Xiaohui Yang and Wei Liu, Multiple Watermarking Algorithm Based on Spread Transform Dither Modulation, *arXiv:1601.04522*, 2011.
12. Xinchao Li, Ju Liu, Jiande Sun, Zhaowan Sun and Huibo Hu, Improved spread transform dither modulation-based watermark algorithm based on step projection, *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '10)*, Shanghai, China, 2010.
13. Jiande Sun, Niqing Yang, Ju Liu, Xiaohui Yang, Xinchao Li and Lei Zhang, Video watermarking scheme based on spatial relationship of DCT coefficients, *Proc. IEEE 8th World Congress on Intelligent Control and Automation (WCICA '10)*, Jinan, China, 2010.
14. Xiushan Nie, Jianping Qiao, Ju Liu, Jiande Sun, Xinchao Li and Wei Liu,LLE-based video hashing for video identification, *Proc. IEEE 10th International Conference on Signal Processing (ICSP '10)*, Hong Kong, China, 2010.