Advances in Computer Vision and Pattern Recognition

Amir R. Zamir
Asaad Hakeem
Luc Van Gool
Mubarak Shah
Richard Szeliski *Editors*

Large-Scale Visual Geo-Localization

Springer

Advances in Computer Vision and Pattern Recognition

More information about this series at http://www.springer.com/series/4205

Amir R. Zamir · Asaad Hakeem
Luc Van Gool · Mubarak Shah
Richard Szeliski
Editors

Large-Scale Visual Geo-Localization

Springer

Editors
Amir R. Zamir
Department of Computer Science
Stanford University
Stanford, CA
USA

Mubarak Shah
Department of Computer Science
University of Central Florida
Orlando, FL
USA

Asaad Hakeem
Machine Learning Division
Decisive Analytics Corporation
Arlington, VA
USA

Richard Szeliski
Facebook
Seattle, WA
USA

Luc Van Gool
Computer Vision Laboratory
ETH Zürich
Zürich
Switzerland

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-3-319-25779-2 ISBN 978-3-319-25781-5 (eBook)
DOI 10.1007/978-3-319-25781-5

Library of Congress Control Number: 2016932324

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Preface

One of the greatest inventions of all times is *camera*, a device for capturing images of the world. Computer vision emerged as the scientific field of understanding the images that cameras produce. Another influential invention of modern technology is *positioning systems*, instruments (namely GPS) for identifying one's location. Even though these two devices are often viewed as belonging to completely different realms, their products (images and locations) have a close relationship. For instance, by just looking at an image, one can sometimes guess its location, e.g., when a rain forest is in view or looking at Eiffel Tower. The opposite is also often true: knowing the location of an image can provide a rich context and significantly assist the process of understanding the visual content. The goal of this book is to explore this bidirectional relationship between images and locations.

In this book, we provide a comprehensive map of the state of the art in large scale visual geo-localization and discuss the emerging trends in this area. Visual geo-localization is defined as the problem of identifying the location of the camera that captured an image and/or the content of the image. Geo-localization finds numerous applications in organizing photo collections, law enforcement, or statistical analysis of commercial market trends.

The book is divided into four main parts: Data-Driven Geo-localization, Semantic Geo-localization, Geometric Matching based Geo-localization, and Real-World Applications. The first part (Data-Driven Geo-localization) discusses recent methods that exploit internet-scale image databases for devising geographically rich features and geo-localizing query images at a city, country, or global scale. The second part (Semantic Geo-localization) presents geo-localization techniques that are built upon high-level and semantic cues. The third part (Geometric Matching based Geo-localization) focuses on the methods that perform localization by geometrically aligning the query image against a 3D model (namely digital elevation model or structure-from-motion reconstructed cities).

In the fourth part of the book (Real-World Applications), methods that use the geo-location of the image, whether extracted automatically or acquired from a GPS-chip, for understanding the image content are presented. Such frameworks are of great importance as cell phones and cameras are now being equipped with built-in localization devices. Therefore, it is of particular interest to develop techniques that accomplish image analysis *assisted by geo-location*. We also discuss several approaches for geo-localization under more practical settings, e.g., when the supervision of an expert in the form of a user-in-the-loop system is available.

The editors sincerely appreciate everyone who assisted in the process of preparing this book. In particular, we thank the authors, who are among the leading and most active researchers in the field of computer vision. Without their contribution in writing and peer reviewing the chapters, this book would not have been possible. We are also thankful to Springer for providing us with excellent support. Lastly, the editors are truly grateful to their families and loved ones for their continual support and encouragement throughout the process of writing this book.

Stanford, CA, USA — Amir R. Zamir
Arlington, VA, USA — Asaad Hakeem
Zürich, Switzerland — Luc Van Gool
Orlando, FL, USA — Mubarak Shah
Seattle, WA, USA — Richard Szeliski
May 2015

Contents

Contributors

Mathieu Aubry LIGM (UMR CNRS 8049), ENPC/Université Paris-Est, Marne-la-Vallée, France

Georges Baatz Google Inc., Zürich, Switzerland

Mayank Bansal Vision Technologies Lab, SRI International, Princeton, NJ, USA; Center for Vision Technologies, SRI International, Princeton, NJ, USA

Serge Belongie Cornell University, Ithaca, NY, USA

Liangliang Cao IBM Watson Research, New York, USA

Mei Chen Intel Corporation, Hillsboro, OR, USA

Yi Chen Object Video, Inc., Reston, VA, USA

Hui Cheng Vision Technologies Lab, SRI International, Princeton, NJ, USA

Matthew Clements UC Berkeley, Berkeley, CA, USA

Dave Conger Object Video, Inc., Reston, VA, USA

David J. Crandall Indiana University, Bloomington, IN, USA

Kostas Daniilidis Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

Alexei A. Efros Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA, USA

Pascal Fua EPFL, Lausanne, Switzerland

Kiran Gunda Object Video, Inc., Reston, VA, USA

Himaanshu Gupta Object Video, Inc., Reston, VA, USA

Asaad Hakeem Decisive Analytics Corporation, Arlington, VA, USA

James Hays Brown University, Providence, RI, USA

Martial Hebert Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Varsha Hedau Apple, Cupertino, CA, USA

James C. Hoe Carnegie Mellon University, Pittsburgh, PA, USA

Daniel P. Huttenlocher Cornell University, Ithaca, NY, USA

Rahul Kumar Jha University of Michigan, Ann Arbor, USA

Leif Kobbelt RWTH Aachen University, Aachen, Germany

Jana Košecká George Mason University, Fairfax, VA, USA

Kevin Köser GEOMAR, Kiel, Germany

Neeraj Kumar Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

L'ubor Ladický ETH Zürich, Zürich, Switzerland

Stefan Lee Indiana University, Bloomington, IN, USA

Yong Jae Lee Department of Computer Science, UC Davis, Davis, CA, USA

Bastian Leibe RWTH Aachen University, Aachen, Germany

Li-Jia Li Yahoo! Research, Sunnyvale, USA

Yunpeng Li École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Tsung-Yi Lin Cornell University, Ithaca, NY, USA

Eriko Nurvitadhi Intel Corporation, Hillsboro, OR, USA

Hyun Soo Park Carnegie Mellon University, Pittsburgh, PA, USA

Minwoo Park Object Video, Inc., Reston, VA, USA

Marc Pollefeys ETH Zürich, Zürich, Switzerland

Gang Qian Object Video, Inc., Reston, VA, USA

Bryan Russell Adobe Research, Lexington, KY, USA

Torsten Sattler Department of Computer Science, ETH Zürich, Zürich, Switzerland

Olivier Saurer ETH Zürich, Zürich, Switzerland

Harpreet Sawhney Center for Vision Technologies, SRI International, Princeton, NJ, USA

Steven Seitz Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Khurram Shafique Object Video, Inc., Reston, VA, USA

Mubarak Shah University of Central Florida, Orlando, FL, USA

David Ayman Shamma Yahoo! Research, San Francisco, USA

Yaser Sheikh Carnegie Mellon University, Pittsburgh, PA, USA

Gautam Singh George Mason University, Fairfax, VA, USA

Sudipta N. Sinha Microsoft Research, Redmond, WA, USA

Josef Sivic Inria, WILLOW Project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR, Paris, France

Noah Snavely Cornell University, Ithaca, NY, USA

Richard Szeliski Facebook, Seattle, WA, USA

Bart Thomee Yahoo! Research, San Francisco, USA

Raphael Townshend Stanford University, Stanford, CA, USA

Eric Tzeng UC Berkeley, Berkeley, CA, USA

Luc Van Gool ETH Zurich, Zurich, Switzerland

Nick Vander Valk Vision Technologies Lab, SRI International, Princeton, NJ, USA

Yang Wang University of Manitoba, Winnipeg, Canada

Yu Wang Carnegie Mellon University, Pittsburgh, PA, USA

Avideh Zakhor UC Berkeley, Berkeley, CA, USA

Amir R. Zamir Stanford University, Stanford, CA, USA

Andrew Zhai UC Berkeley, Berkeley, CA, USA

Jiejie Zhu Vision Technologies Lab, SRI International, Princeton, NJ, USA

C. Lawrence Zitnick Facebook AI Research, Palo Alto, CA, USA

Chapter 1
Introduction to Large-Scale Visual Geo-localization

Amir R. Zamir, Asaad Hakeem, Luc Van Gool, Mubarak Shah and Richard Szeliski

Abstract Despite recent advances in computer vision and large-scale indexing techniques, automatic geo-localization of images and videos remains a challenging task. The majority of existing computer vision solutions for geo-localization are limited to highly-visited urban regions for which a significant amount of geo-tagged imagery is available, and therefore, do not scale well to large and ordinary geo-spatial regions. In this chapter, we provide an overview of the major research themes in visual geo-localization, investigate the challenges, and point to problem areas that will benefit from common synthesis of perspectives from these research themes. In particular, we discuss how the availability of web-scale geo-referenced data affects visual geo-localization, what role semantic information plays in this problem, and how precise localization can be achieved using large-scale textured (RGB) and untextured (non-RGB) 3D models. We also introduce a few real-world applications which became feasible as a result of the capability of estimating an image's geo-location. We conclude this chapter by providing an overview of the emerging trends in visual geo-localization and a summary of the rest of the chapters of the book.

A.R. Zamir (✉)
Stanford University, Stanford, CA, USA
e-mail: zamir@cs.stanford.edu

A. Hakeem
Decisive Analytics Corporation, Arlington, VA, USA
e-mail: asaad.hakeem@dac.us

L. Van Gool
ETH Zurich, Zurich, Switzerland
e-mail: vangool@vision.ee.ethz.ch

M. Shah
University of Central Florida, Orlando, FL, USA
e-mail: shah@crcv.ucf.edu

R. Szeliski
Facebook, Seattle, WA, USA
e-mail: szeliski@fb.com

A.R. Zamir et al. (eds.), *Large-Scale Visual Geo-Localization*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-3-319-25781-5_1

1.1 Introduction

Geo-localization is the problem of discovering the location where an image or video was captured. This task arises in a variety of real-world applications. Consumers of imagery may be interested in determining where and when an image/video was taken, who is in the image, what the different landmarks and objects in the depicted scene are, and how they are related to each other. Local government agencies may be interested in using large-scale imagery to automatically obtain and index useful geographic and geological features and their distributions in a region of interest. Similarly, local businesses may utilize content statistics to target their marketing based on 'where', 'what', and 'when' that may automatically be extracted using visual analysis and geo-localization of large-scale imagery. Law enforcement agencies are often interested in finding the location of an incident captured in an image, for forensic purposes. Many of these applications require establishing a relationship between the visual *content* and the location; a task that demands more than simply a GPS-location and requires careful content analysis.

Despite the enormous recent progresses in the field of computer vision, automatic geo-localization is still a difficult problem. It involves identifying, extracting, and indexing geo-informative features from a large variety of datasets, discovering subtle geo-location cues from the imagery, and searching in massive reference visual and nonvisual databases, such as Geographic Information System (GIS). To solve these problems, significant technological advancements are needed in the areas of data-driven discovery of geo-informative features, geometric modeling and reasoning, semantic scene understanding, context-based reasoning, and exploitation of cross-view aerial imagery and cross-modality GIS data. In addition, complementary viewpoints and techniques from these diverse areas will provide additional insight into the problem domain and may spur new research directions, which will be likely to not remain limited to the geo-localization topic.

The central aim of this monograph is to facilitate the exchange of ideas on how to develop visual geo-localization and geo-spatial capabilities. The following are some of the scientific questions and challenges we hope to address:

- What are the general principles that help in characterization and geo-localization of visual data? What features, ranging from low-level and data-driven to high-level and semantic, are geo-spatially informative?
- How can verifiable mathematical models of visual analysis and geo-localization be developed based on these principles? What are the proper search techniques for matching the extracted informative features?
- How can these principles be used to enhance the performance of generic mid- to high-level vision tasks, such as object recognition, scene segmentation/understanding, geometric modeling, alignment, and matching, or empower new applications?

To address these challenges, a number of peer-reviewed chapters from leading researchers in the fields of computer vision, multimedia, and machine learning are compiled. These chapters provide an understanding of the state of the art, open problems, and future directions related to visual geo-spatial analysis and geo-localization of large-scale imagery as a scientific discipline.

The rest of the chapter is organized as follows: In Sect. 1.2, we introduce the central themes and topics in the book. In Sect. 1.3, a brief overview of the emerging trends in the field of geo-localization is provided. In Sect. 1.4, we summarize of the organization of the rest of the book, giving a brief introduction to each chapter.

1.2 Central Themes and Topics

Until the early 2000s, the majority of automatic visual geo-localization methods were targeted toward airborne and satellite imagery. That is, the furnished datasets were mainly composed of images with nadir or limb views, and the query data was captured from either satellites or aircraft. Toward this end, many successful methods [1–3] were developed which were primarily based on planar-scene registration techniques and often utilized subsidiary models, such as Digital Elevation Model (DEM) Digital Elevation Map (DEM), and Digital Terrain Model (DTM), along with regular aerial imagery and Digital Orthophoto Quadrangles (DOQ).

However, during the past decade, the production of visual data has undergone a major shift with the sudden surge of consumer imagery, which mainly retains a ground-level viewpoint. The shift was mostly due to the plummeting cost of the photographic devices as well as the increasing convenience of sharing the multimedia material, including pictures. Currently, ground-level images and videos are primarily produced either through systematic efforts by governments and the private sector (e.g., Google Street View [4]) or directly by consumers (crowdsourcing [5–7]). Google Street View, which provides dense spherical views of public roadways, and Panoramio Collection [5], which is composed of crowdsourced images, are two notable examples of such structured and unstructured databases, respectively.

This considerable shift poses new challenges in the context of automatic geo-localization and demands novel techniques capable of coping with this substantial change. For instance, unlike the traditional geo-registration problem where the scene was mainly assumed to be planar, the ground-level images show heavy nonplanarity. In addition, in traditional aerial images, each pixel is associated with a GPS-location while in the ground-level user images, the entire image is often associated with one GPS-tag which is the location of camera. The following are some of the primary challenges that have emerged as a result of this shift in the production of visual data:

- **Large-scale data handling**: The amount of produced data is prohibitively massive and is increasing sharply. This makes devising techniques that are efficient in preprocessing (leveraging the large amount of reference data) and query processing (sifting through the preprocessed data) essential.
- **Necessity of an accurate geo-location**: Many of the procedures that use geo-tags as their input require a precise geo-location, particularly in the urban areas.

In general, extracting a coarse geographical location, e.g., which continent the image was captured at or distinguishing between desert and coast, has limited applications, and performing the geo-localization with an accuracy comparable to, or better than, handheld GPS devices is desirable.

- **Ambiguity and excessive similarity of visual features**: Unless the data include distinctive objects, such as landmarks, discovering the location merely based on visual information is often challenging due to the significant similarity between man-made structures. This issue becomes critical when city or country scale geo-localization is of interest.
- **Undesirable photography effects**: Unwanted effects, such as suboptimal lighting, frequent occlusions by moving objects, lens distortions, or stitching artifacts, often introduce additional complexities.
- **Lack of unified reference data**: Unlike the conventional geo-registration problem where the reference data (e.g., DOQ) was often unified and enjoyed a homogeneous format, the reference resources for geo-localization of ground-level images are commonly diverse, cross-modality (e.g., GIS which is semantic), and cross-view (e.g., bird's eye imagery which is oblique).

We investigate these challenges in-depth and provide a set of solutions and directions for the future research for each. The chapters of the book are distributed over four major themes and topics described next.

1.2.1 Part I: Data-Driven Geo-localization

Methods for exploiting web-scale datasets for geo-localization and extraction of geographically informative features.

The recent availability of massive repositories of public geo-tagged imagery from around the world has paved the way for geo-localization methods that adopt a data-driven approach. One example of such strictly data-driven methods are those that employ an *image retrieval*-based strategy. That is, match the content of the query image against a dataset of geo-tagged images, identify the reference images with high similarity to the query, and estimate the location of the query based on the geo-tags of the retrieved reference images [8–16]. This approach has been observed to yield promising results as image retrieval and image geo-localization share a great deal of similarity in terms of the problem definition and the challenges faced. The requirement of handling a massive volume of data, or the necessity of dealing with undesirable photography effects, nonplanarity of the scene, and frequent occlusions by irrelevant objects are some of these common challenges.

However, image geo-localization and image retrieval differ in a few fundamental aspects: In image retrieval, the ultimate goal is to find all of the images that match the query with different amounts of similarity. On the contrary, in image geo-localization, the goal is to propose the best location for the query that does not necessarily require finding a large number of matching images. For instance, estimating the location of the query is deemed easier having a few geo-tagged images with relatively similar viewpoints (and probably substantially similar in content) as compared to a large number of not-so-similar images. Additionally, all forms of resemblance, such as semantic similarity (e.g., sharing generic objects), is typically in the interest of image retrieval. In contrast, the primary objective in image geo-localization is to find the images that indeed show the *same* location, and not just a similar one. Such divergences signify that adopting image retrieval and matching techniques off-the-shelf is not an adequate solution for geo-localization, and devising methods specifically intended for the task of localization, even if inspired by image retrieval, is vital. The primary challenges and opportunities in this context are: expanding the applicability area of geo-localization techniques, and identifying geographically informative features. Each of these are elaborated below:

- **Expanding the localization coverage using cross-view imagery and web-scale repositories of user-shared images**: One of the most critical characteristics of any geo-localization system is the size of its covered area. The techniques that are applicable to a larger region are significantly more desirable and practically useful [14, 17, 18]. Databases that are collected in a systematical and controlled manner, such as Google Street View [4], have a limited coverage when it comes to arbitrary regions (e.g., Sahara desert or unpaved roads). This signifies the necessity of leveraging new resources of reference data that do not suffer from this limitation. One remedy to this issue is utilizing aerial imagery (satellite or bird's eye), which is known to be typically available for any arbitrary location. However, this requires performing the matching in a challenging cross-view manner which will be discussed in detail in Chaps. 4 and 5. Adopting other reference modalities with a ubiquitous coverage, such as DEM, GIS, or DTM, are also additional solutions that we will discuss in Chaps. 6, 11–13. Moreover, another possible approach is to utilize the (rapidly expanding) ground-level user-shared image repositories as they typically include images from any point that could be visited by humans. This is discussed in Chap. 3.
- **Devising geographically informative features:** Employing features that are rich and distinctive in capturing geo-spatial information is essential for successful and efficient geo-localization [19, 20]. This is of particular interest as the image features that are devised for generic content matching are not necessarily a good fit for capturing location-dependent information. For instance, the subtle architectural styles that indeed encode a lot of information about the location are typically lost in feature extraction, e.g., SIFT, and quantization, e.g., Bag of Words modeling, steps of existing techniques. Luckily, availability of large-scale geo-tagged data empowers adopting a bottom-up approach to this problem. In Chap. 2, we will discuss how a set of *data-driven* mid-level features that are specifically devised to be geo-spatially discriminative can be extracted from a web-scale dataset of geo-tagged images. In addition, leveraging the massive amount of available data, such features can be simultaneously learned to be invariant with respect to undesirable effects, such as lighting or viewpoint changes.

1.2.2 Part II: Semantic Geo-localization

Identifying the geo-location based on high-level and semantic cues.

In Part I, the methods that adopt a data-driven approach to geo-localization, and consequently, do not make use of cues that are necessarily semantically meaningful were discussed. However, high-level and semantic features of an image carry a significant amount of location-dependent information. Such cues range from human-related characteristics, such as text, architectural style, vehicle types, or urban structures, to natural properties, such as foliage type or weather. Each of these are discriminative geo-location cues: the language of a sign can specify the country/region the picture was captured at; the architectural style of buildings can narrow down the search to certain countries or even cities; observing a particular type of foliage will significantly reduce the search to certain geographical areas. The *aggregation* of such diverse cues, especially when coupled with their geometric arrangement [21], indeed leads to a surprisingly discriminative signature for geo-location [21, 22].

Even though we, as humans, often perform visual geo-localization in such semantic manner and a similar approach for aerial geo-registration has been developed before [23, 24], automatic geo-localization of ground-level imagery based on semantic cues is relatively under-explored [25]. Fortunately, there are reference databases where such features can be extracted from; Geographical Information System (GIS) which contains the location of a large and detailed set of semantic object/structures/land covers, Wikipedia and Wikimapia that provide almost every type of location specific information about events and landmarks, and hashtags of GPS-tagged images in online photo repositories which establish a link between textual tags and geo-locations are some of the nominal examples.

The three main challenges of semantic geo-localization can be summarized as: what features to employ, how to match them, and how to consolidate the diverse semantic cues in a unified system. In Chap. 6, we discuss how a set of semantic features based on scene segments, e.g., road, building facade, etc., along with a histogram representation can significantly narrow down the search area. Also, in Chap. 2, we investigate a set of mid-level features that are found in a data-driven manner but encode beyond low-level pixel information by capturing architectural and temporal information. In addition, in Chap. 6, it will be shown that textual tags and temporal orders of images can considerably improve geo-localization, especially for landmarks.

1.2.3 Part III: Geometric Matching-Based Geo-localization

Geometric alignment of query image to geo-referenced 3D and elevation models.

Data-driven and semantic matching techniques typically provide a rather coarse estimation of the geo-location. Geometric matching-based methods, on the other hand, match the query image against a precise scene model and provide a means for acquiring a more accurate geo-location. Such methods typically involve, first, constructing a geo-referenced 3D scene model (commonly acquired from Structure from Motion) [26–29], and then, performing 2D-to-3D matching between the features extracted from the query image and the reference scene model [30–32]. This approach often comes with additional byproducts of a full 6 DoF camera parameters estimation (i.e., camera rotation as well as the location) and an estimation of the location of the image *content* (besides the location of the camera). Matching against a textured (RGB) 3D point cloud or an untextured (non-RGB) model, such as DEM, constitute the two main classes of approaches in this area; the latter is specifically a good fit for the areas where dense image coverage is not available. The challenges of each are elaborated below:

- **Constructing high fidelity 3D textured (RGB) scene models and performing 2D-to-3D matching against them**: A significant portion of the geometric matching-based geo-localization methods require a 3D point cloud associated with RGB information as their reference input. The sheer size of image datasets from which such 3D models are constructed, and the large variations in imaging and scene parameters pose unique challenges for a wide-area reconstruction. Recent work in Structure from Motion (SfM) has successfully built 3D models from city-scale unstructured collections of images from publicly available photo repositories [26, 27, 33]. This formulation first uses a hybrid discrete-continuous optimization for finding a coarse initial solution and then improves it using bundle adjustment. It naturally incorporates various sources of information about both the cameras and the points, including noisy geo-tags and vanishing point estimates. By using all of the available data at once (rather than incrementally), and by allowing additional forms of constraints, the approach has been shown to be quite robust under realistic settings.
  Once the reconstruction of the reference model is done, the query image features need to be matched against it in a 2D-to-3D manner. The primary challenges of this step are: coping with the massive size of the reference 3D model, effective utilization of the geometric information of the 3D model during matching (besides the RGB information), and performing the matching in a robust way despite the large variation in imaging conditions. In Chaps. 8 and 10, we discuss how an accurate large-scale point cloud can be formed and its size can be significantly reduced without hurting the overall geo-localization performance. Also, in Chap. 9, we will discuss how the accuracy of 2D-to-3D matching can be improved at no or little computational cost, by utilizing the regularities (specifically, co-visibility, and spatial constraints) inherent in point clouds. In Chap. 14, it will be demonstrated that geometric matching-based geo-localization can be extended to aligning even different modalities, e.g., paintings, against image-based 3D models. The techniques developed for this robust extension can be adopted for performing image alignment in presence of severe lighting, viewpoint, and imaging condition changes.
- **Geometric alignment of images to untextured (non-RGB) 3D models**: The aforementioned point clouds are typically formed using SfM techniques which

require a significant number of images for reconstruction. That is why the best performance of such methods are reported for heavily visited areas where a large number of user-shared images are available. However, a significant portion of the Earth does not have such dense image coverage that would yield a 3D point cloud. On the other hand, Digital Elevation Models (DEM), Digital Terrain Model (DTM), and similar non-RGB 3D models that are typically acquired from satellite multispectral analysis and remote sensing are commonly available for any spot on Earth. Therefore, several approaches for utilizing this type of reference data for performing geo-localization in sparsely visited/populated locations have been developed [34–36]. Such methods typically involve extracting a set of relevant features (e.g., mountain folds and creases) from the query and matching them against DEM. The main challenges in this area lie in: what image features would allow performing such cross-modality matching, and how to cope with the large-scale search process. In Chaps. 11, 12, and 13, three methods adopting this approach are presented, which are applicable to mountainous areas (mountain contour matching) as well as desert regions (terrain contour matching).

1.2.4 Part IV: Real-World Applications

Practical tools for geo-localization of web-scale databases and utilizing the extracted geo-location.

Thus far, we focused on fully automated geo-localization frameworks and discussed how this could be accomplished using data-driven features, semantic cues, or geometric matching against 3D models. However, a fundamental follow-up question would be: *What can this extracted geo-location be used for?* This question is of particular importance as the modern cameras and camera phones are GPS enabled and generate photos that are automatically geo-tagged at the time of collection. This signifies the importance of developing methods that can *make use of the geo-location* of images in a principled manner, for instance for understanding the content. In addition, the cameras being GPS enabled means at least a coarse location for the image is known at the time of collection; therefore, there is a growing interest in visual geo-localization techniques that perform the localization *aided by the initial location*, leading to an accuracy far beyond GPS. In the final part of this monograph, we focus on methods for utilizing the geo-location in real-world, performing the visual localization aided by an initial GPS location, and human-in-the-loop localization frameworks that can leverage multimodal cues for a practical geo-localization.

It is of particular interest to leverage the geo-location of an image, whether found visually or using a GPS chip, for understanding its content. The majority of the current applications which utilize the geo-location of images, such as location-based retrieval [5] or large-scale 3D reconstruction [37], are not intended to provide a high-level understanding of the image content. With the exception of a few recent methods which use the geo-tags in connection with the image content (e.g., GIS-assisted object detection [21] or GPS-assisted visual business recognition [38]), the potential impact of the image geo-tags on understating the image content is largely unexplored. This is quite unproductive as knowing the location of an image immediately provides a strong context about the image content: an image taken in Manhattan is expected to show large buildings and maybe yellow cabs; an image taken in a national forest cannot include a skyscraper; knowing a picture was taken in East Asia would give a strong prior about what type of architecture to expect to see.

In Chap. 17, a framework that analyzes the image content in connection with its GPS-tag is presented. The aim of this method is to provide a convenient and automatic way for browsing personal photo collections based on user's textual queries (e.g., "the pictures I took at Radiohead concert"), but without the need for any manual annotation. This is accomplished by using the GPS-tag of the images in the personal photo collection along with their time/date stamp for searching in the web databases (e.g., Wikimapia). This search yields a set of hypotheses for the events that could have possibly happened at that particular time and location (e.g., a concert). These hypotheses are then pruned and verified by matching them against the image content.

As another real-world application, we discuss the situations where performing the visual geo-localization fully from scratch is inefficient and even unnecessary. An example is the cases where some level of human supervision or a prior knowledge about the location of the image is available. In Chap. 15, we will discuss a framework that effectively makes use of the location the GPS chip of the camera provides for performing fast landmark recognition on low-powered mobile device. In Chap. 16, a human-in-the-loop geo-localization framework in which an analyst's feedback is looped back into the system is introduced. The key idea behind this method is to combine a minimal intervention of an analyst with the power of automatically mined web-scaled databases to achieve a localization accuracy that outperforms both fully automated methods and purely manual geo-localization.

As discussed earlier, the end goal of the methods discussed in this part is not an entirely automated geo-localization, but either utilizing the geo-location for content understanding or performing the localization in an assisted manner under practical settings.

1.3 Emerging Trends

Geo-spatial localization is a fast evolving field, mostly due to availability of unprecedented data resources, new applications, and novel techniques. In this section, we briefly overview some of the emerging trends in this area each centered around one of the aforementioned aspects.

1.3.1 New Geo-referenced Data Resources

Geo-localization is constantly affected by the available resources of geo-referenced data [14, 39–42]. It experienced a major shift with the emergence of massive ground-level imagery from users and street view during the past decade. Similar changes in the geo-referenced data are expected to continue and define new opportunities and challenges. One of the new resources that is becoming extensively popular is personal drones. Such inexpensive, GPS-enabled, and widely available UAVs can provide an HD, and sometimes live, aerial imagery of an arbitrary location. This is considered a major change as it takes collection of aerial imagery to a personal level, while it used to be exclusive to large companies and governments until not too long ago. This is similar to what cell phones and user-shared imagery brought in ground-level geo-localization about a decade ago, when collection of large-area datasets was limited to governments and large corporations, such as Google.

Another shift in geo-referenced data resources is taking place as a result of significant improvements in the spatial and temporal resolution of satellite images. As of now, typical GSD (Ground Sampling Distance that is the metric distance represented by one pixel) values have reached the accuracy of <0.35 m, and the temporal resolution (satellite revisit time) has reduced to <1 day [43, 44]. In addition, impressive near-live HD videos from space have been recently offered in the commercial market [45] that provide new grounds for performing temporal modeling along with spatial geo-localization. Such data resources are all new and rapidly changing which require novel approaches to geo-spatial analysis in order to effectively accommodate and utilize them.

1.3.2 Temporal Geo-localization and New Applications

The majority of frameworks for geo-localization are targeted toward extracting the spatial location. With the exception of a few efforts [46–48], the temporal aspect of location is largely ignored, whereas providing information about "*when*" a particular event takes place is as important as "*where*." Appending time information to geo-spatial localization is expected to be of growing interest in the future. Government agencies for law enforcement purposes or local businesses for analyzing their target market are only some of the entities that are interested in the temporal aspect as much as they are interested in the spatial location. Performing the geo-localization in a timely manner and joint modeling of temporal and geo-spatial changes in visual data are some of the potential tasks in this area.

In particular, the temporal aspect of cross-view geo-localization is further brought to interest by the notable improvements in the temporal resolution of satellite imagery (currently as low as <1 day and decreasing) and the near-live HD videos from space, as elaborated in the previous subsection.

1.3.3 Deep Learning Based Geo-localization

Convolutional neural networks have remarkably improved the state of the art in several computer vision problems, namely object detection [49], scene recognition [50], and depth from monocular vision [51]. In essence, a convolutional neural network is a cascade of several layers of convolution operations followed by nonlinearity with fully learned parameters. It essentially provides a means for end-to-end and unified feature learning and classification/regression. With the notable abilities that deep learning-based methods have recently shown in solving basic computer vision problems, it is expected that they will bring substantial improvements to the field of geo-localization as well. Very recently, a few attempts have shown promising results for cross-view image geo-localization [52] and semantic feature learning [53] employing deep Learning-based techniques that demonstrate the potential merits in this approach. Considering the basic characteristics of deep learning, it is expected to be particularly beneficial in cross-view/cross-modality matching, end-to-end geo-spatial feature learning, and temporal geo-localization using recurrent neural networks.

1.4 Organization of the Book

An overview of each of the chapters is provided below:

Part I: Data-Driven Geo-localization: *Methods for exploiting web-scale datasets for geo-localization and extraction of geographically informative features*.

Chapter 2: Discovering Mid-level Visual Connections in Space and Time

This chapter explores what a mid-level visual representation can bring to geo-spatial and longitudinal analyses. Specifically, the authors present a weakly-supervised visual data mining approach that discovers connections between recurring mid-level visual elements in historic (temporal) and geographic (spatial) image collections, and attempts to capture the underlying visual style. In contrast to existing discovery methods that mine for patterns that remain visually consistent throughout the dataset, the goal here is to discover visual elements whose appearance changes due to change in time or location; i.e., exhibit consistent stylistic variations across the label space (date or geo-location). To discover these elements, the authors first identify groups of patches that are style-sensitive. Then, they incrementally build correspondences to find the same element across the entire dataset. Finally, they train style-aware regressors that model each element's range of stylistic differences. The authors demonstrate the method's effectiveness on the related task of fine-grained classification.

Chapter 3: Where the Photos were Taken: Location Prediction by Learning from Flickr Photos

This chapter investigates the characteristics of geographically tagged Internet photos and determines their location based on the visual content. To build reliable geographical estimators, it is important to find distinguishable geographical clusters in the world. These clusters cover general geographical regions not limited to just landmarks. Geographical clusters provide more training samples and hence lead to better recognition accuracy. To solve this estimation problem, the authors built an efficient solver to find the clusters employing a set of latent variables. They demonstrate detailed qualitative results obtained from beach photos taken in different continents.

Chapter 4: Cross-View Image Geo-localization

In this chapter, a cross-view feature translation approach that greatly extends the reach of image geo-localization methods to a vast majority of the Earth's land area that has no ground level reference photos available is presented. The authors present a method that can often localize a query even if it has no corresponding ground-level images in the database. The key idea is to learn a mapping from ground-level appearance to overhead appearance and land cover attributes. This relationship is learned from sparsely available geo-tagged ground-level images and the corresponding aerial and land cover data at those locations. The authors demonstrate their method over a 1135 km^2 region containing a variety of scenes and land cover types. For each query, their algorithm produces a probability density over the region of interest.

Chapter 5: Ultrawide Baseline Facade Matching for Geo-localization

This chapter describes a technique that matches street-level images to a database of airborne images. This problem is hard because of extreme viewpoint and illumination differences. Color/gradient distributions or local descriptors fail to match under such setting forcing us to rely on the structure of self-similarity of patterns on facades. The authors propose to capture this structure with a novel "scale-selective self-similarity" (S^4) descriptor which is computed at each point on the facade at its inherent scale. To achieve this, the authors introduce a new method for scale selection which enables the extraction and segmentation of facades as well. They also introduce a novel geometric method that aligns satellite and bird's-eye-view imagery to extract building facade regions in a stereo graph-cuts framework. Matching of the query facade to the database facade regions is done with a Bayesian classification of the street-view query S^4 descriptors given all labeled descriptors in the bird's-eye view database. The authors also discuss geometric techniques for camera pose-estimation using correspondences between building corners in the query and the matched aerial imagery. They demonstrate the retrieval accuracy on a challenging set of publicly available imagery and compare with standard SIFT-based techniques.

Part II: Semantic Reasoning-based Geo-localization: *Identifying the geo-location based on high-level and semantic cues.*

Chapter 6: Semantically Guided Geo-localization and Modeling in Urban Environments

This chapter presents a technique that enables semantic labeling of both query views and the reference dataset through semantic segmentation that can aid (1) retrieval of views similar and possibly overlapping with the query and (2) guiding the recognition and discovery of commonly occurring scene layouts in the reference dataset.

The authors demonstrate the effectiveness of these semantic representations on examples of localization, semantic concept discovery, and intersection recognition in the images of urban scenes.

Chapter 7: Recognizing Landmarks in Large-Scale Social Image Collections

In this chapter, a technique that can recognizes popular landmarks in large-scale datasets of unconstrained consumer images from Flickr by formulating a classification problem involving nearly 2 million images and 500 categories is described. The dataset and categories are formed automatically from geo-tagged photos from Flickr by looking for peaks in the spatial geo-tag distribution corresponding to frequently photographed landmarks. The authors learn models for these landmarks with a multiclass support vector machine, using classic vector-quantized interest point descriptors as features. They also incorporate the semantic non-visual metadata available on modern photo-sharing sites, showing that *textual tags* and *temporal constraints* lead to significant improvements in classification rate. Finally, they apply recent breakthroughs in deep learning with Convolutional Neural Networks, finding that these models can dramatically outperform the traditional recognition approaches to this problem, and even beat human observers in some cases.

Part III: Geometric Matching Based Geo-localization: *Geometric alignment of query image to geo-referenced 3D and elevation models.*

Chapter 8: Worldwide Pose Estimation Using 3D Point Clouds

This chapter addresses the problem of determining where a photo was taken by estimating a full 6-DOF-plus-intrinsic camera pose with respect to a large geo-registered 3D point cloud, bringing together research on image localization, landmark recognition, and 3D pose estimation. The authors propose a method that scales to datasets with hundreds of thousands of images and tens of millions of 3D points through the use of two new techniques: a co-occurrence prior for RANSAC and bidirectional matching of image features with 3D points. The authors evaluate their method on several large data sets, and show state-of-the-art results on landmark recognition as well as the ability to locate cameras to within meters, requiring only seconds per query.

Chapter 9: Exploiting Spatial and Co-visibility Relations for Image-Based Localization

This chapter describes a technique that can increase the effectiveness of approaches based on prioritized 2D-to-3D matching at little to no additional run-time costs by exploiting both spatial and co-visibility relations between the 3D points in the model. Geometry matching based localization techniques aim to estimate the position and orientation from which a given query image was taken with respect to a 3D model of the scene. Recent advances in Structure-from-Motion, which allow us to reconstruct large scenes in little time, create a need for image-based localization approaches that handle large-scale models consisting of millions of 3D points both efficiently and effectively in order to localize as many query images as possible in as little time as possible. While multiple efficient localization methods based on prioritized feature matching have been proposed recently, they lack the effectiveness of slower approaches. The authors demonstrate that the resulting localization framework incorporates both 2D-to-3D and 3D-to-2D matching and achieves state-of-the-art efficiency and effectiveness.

Chapter 10: 3D Point Cloud Reduction using Mixed-Integer Quadratic Programming

A method to accelerate the matching process and reduce the memory footprint by analyzing the view-statistics of points in a training corpus is provided in this chapter. Given a training image set that is representative of common views of a scene, the authors propose an approach that identifies a compact subset of the 3D point cloud for efficient localization, while achieving comparable localization performance to using the full 3D point cloud. The authors demonstrate that the problem can be precisely formulated as a mixed-integer quadratic program and present a point-wise descriptor calibration process to improve matching. They also show that their algorithm outperforms the state-of-the-art greedy algorithm on standard datasets, on measures of both point-cloud compression and localization accuracy.

Chapter 11: Image Based Large-Scale Geo-localization in Mountainous Regions

This chapter describes a technique that can geo-localize images in mountainous terrain and use digital elevation models to extract representations for fast visual database lookup. The authors propose an automated approach for very large-scale visual localization that can efficiently exploit visual information (contours) and geometric constraints (consistent orientation) at the same time. They demonstrate and validate their approach at the scale of Switzerland (40,000 km^2) using over 1000 landscape query images.

Chapter 12: Adaptive Rendering for Large-Scale Skyline Characterization and Matching

This chapter explores an adaptive rendering approach for large-scale skyline characterization and matching. Given an image, the authors propose a system that automatically extracts the skyline and then matches it to a database of reference skylines extracted from rendered images using digital elevation data (DEM). The sampling density of these rendering locations determines both the accuracy and the speed of skyline matching. The proposed approach successfully combines global planning and local greedy search strategies to select new rendering locations incrementally. The authors report quantitative and qualitative results from synthesized and real experiments, where a $4\times$ computational speedup is achieved.

Chapter 13: User-Aided Geo-localization of Untagged Desert Imagery

This chapter presents a system for user-aided visual localization of desert imagery without the use of any metadata such as GPS readings, camera focal length, or field-of-view. The system makes use only of publicly available datasets—in particular, digital elevation models (DEMs)—to rapidly and accurately locate photographs in nonurban environments such as deserts. The authors propose a system that generates synthetic skyline views from a DEM and extracts stable concavity-based features from these skylines to form a database. To localize queries, a user manually traces the skyline on an input photograph. The skyline is automatically refined based on this estimate, and the same concavity-based features are extracted. They then apply a variety of geometrically constrained matching techniques to efficiently and accurately match the query skyline to a database skyline, thereby localizing the query image. The authors evaluate their system using a test set of 44 ground-truthed images over a 10,000 km^2 region of interest in a desert and show that in many cases, queries can be localized with precision as fine as 100 m^2.

Chapter 14: Visual Geo-localization of Non-photographic Depictions via 2D-3D Alignment

In this chapter, a technique that can geo-localize arbitrary 2D depictions of architectural sites, including drawings, paintings, and historical photographs is described. This goal is achieved by aligning the input depiction with a 3D model of the corresponding site. The task is difficult as the appearance and scene structure in the 2D depictions can be very different from the appearance and geometry of the 3D model, e.g., due to the specific rendering style, drawing error, age, lighting, or change of seasons. In addition, it is a hard search problem: the number of possible alignments of the painting to a set of 3D models from different architectural sites is huge. To address these issues, the authors have developed a compact representation of complex 3D scenes. It is demonstrated that the proposed approach can automatically identify the correct architectural site as well as recover an approximate viewpoint of historical photographs and paintings with respect to the 3D model of the site.

Part IV: Real-World Applications: *Practical Tools for Geo-Localization of Web-Scale Databases and Utilizing the Extracted Geo-location.*

Chapter 15: A Memory Efficient Discriminative Approach for Location-Aided Recognition

In this chapter, a GPS-assisted visual recognition technique for fast recognition of urban landmarks on a GPS-enabled mobile device is presented. Most of the existing similar methods offload their computation to a server by uploading the query image. Over a slow network, this can cause a latency of several seconds. In contrast, the authors present an approach that requires uploading only the approximate GPS location to a server after which a compact, location-specific classifier is downloaded to the device and all subsequent computation is performed onboard. Their approach is supervised and involves training compact random forest classifiers (RDF) on a database of geo-tagged images. The feature vector for the RDF is computed by densely searching the image for the presence of selective discriminative local image patches extracted from the training images. The images are rectified using detected vanishing points and binary descriptors allow for an efficient search for the discriminative patches, a step that is further accelerated using min-hash. The authors have evaluated the performance of their approach on representative urban datasets where it outperforms traditional methods based on bag of visual words representation or direct matching of local feature descriptors, neither of which are feasible when processing must occur on a low-power mobile device.

Chapter 16: A Real-World System for Image/Video Geo-localization
This chapter presents WALDO (Wide Area Localization of Depicted Objects), a system that solves the challenging problem of image geo-localization by combining the insight of analysts with the power of automated analysis for Internet-scale, geo-location-driven data mining. WALDO's goal-driven constrained resource management leverages a full spectrum of data-driven, semantic, and geometric geo-localization experts and user tools.

Chapter 17: Photo Recall: Using the Internet to Label Your Photos

In this chapter, a system for searching your personal photos using an extremely wide range of text queries is proposed. The authors accomplish this by finding the correlations between the information in the photos—the timestamps, GPS locations, and image pixels—and the information mined from the Internet. This includes matching dates to holidays listed on Wikipedia, GPS coordinates to places listed on Wikimapia, places, and dates to find named events using Google, visual categories using classifiers either pre-trained on ImageNet or trained on-the-fly using results from Google Image Search, and object instances using interest point-based matching, again using results from Google Images. The authors tie all of these disparate sources of information together in a unified way, allowing for fast and accurate searches using whatever information the user remembers about a photo. They represent all of this information in a layered graph which prevents duplication of effort and data storage, while simultaneously allowing for fast searches, generating meaningful descriptions of search results, and even suggesting query completions to the user as he/she types, via auto-complete.

References

1. Sheikh Y, Khan S, Shah M (2004) Feature-based georegistration of aerial images. GeoSensor Netw 4
2. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21(11): 977–1000
3. Kumar R, Sawhney H, Asmuth J, Pope A, Hsu S (1998) Registration of video to georeferenced imagery. In: Proceedings of fourteenth international conference on pattern recognition, vol 2, pp 1393–1400
4. https://www.google.com/maps/streetview/
5. http://www.panoramio.com/
6. https://www.flickr.com/
7. http://picasa.google.com/
8. Grant S, Brown M, Szeliski R (2007) City-scale location recognition. In: IEEE conference on computer vision and pattern recognition, CVPR'07
9. Zamir AR, Shah M (2014) Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. In: T-PAMI
10. Torii A, Sivic J, Pajdla T, Okutomi M (2013) Visual place recognition with repetitive structures. In: IEEE conference on computer vision and pattern recognition (CVPR)

11. Gronat P et al (2013) Learning and calibrating per-location classifiers for visual place recognition. In: IEEE conference on computer vision and pattern recognition (CVPR)
12. Knopp J, Sivic J, Pajdla T (2010) Avoiding confusing features in place recognition. Computer vision-ECCV. Springer, Heidelberg, pp 748–761.
13. Chen DM et al (2011) City-scale landmark identification on mobile devices. In: IEEE conference on computer vision and pattern recognition (CVPR)
14. Lin T-Y, Belongie S, Hays J (2013) Cross-view image geolocalization. In: IEEE conference on computer vision and pattern recognition (CVPR)
15. Hays J, Efros AA (2008) IM2GPS: estimating geographic information from a single image. In: IEEE conference on computer vision and pattern recognition (CVPR)
16. Vaca-Castano G, Zamir AR, Shah M (2012) City scale geo-spatial trajectory estimation of a moving camera. In: Computer vision and pattern recognition (CVPR)
17. Bansal M et al (2011) Geo-localization of street views with aerial image databases. In: Proceedings of the 19th ACM international conference on multimedia
18. Bansal M, Daniilidis K, Sawhney H (2012) Ultra-wide baseline facade matching for geo-localization computer vision-ECCV 2012. In: Workshops and demonstrations. Springer, Berlin
19. Doersch C et al (2012) What makes Paris look like Paris? ACM Trans Graph (TOG) 31(4):101
20. Lee YJ, Efros AA, Hebert M (2013) Style-aware mid-level representation for discovering visual connections in space and time. In: ICCV
21. Ardeshir S, Zamir AR, Shah M (2014) GIS-assisted object detection and geospatial localization. In: European conference on computer vision (ECCV)
22. Sullivan A. The view from your window contest. http://dish.andrewsullivan.com/vfyw-contest/
23. Wang C, Croitoru A, Stefanidis A, Agouris P (2007) Image-to-X registration using linear features and networks. In: FUZZ-IEEE
24. Wang C, Stefanidis A, Agouris P (2007) Relaxation matching for georegistration of aerial and satellite imagery. In: IEEE ICIP
25. Castaldo F, Zamir A, Angst R, Palmieri F, Savarese S (2015) The IEEE international conference on computer vision (ICCV) workshops, pp 9–17
26. Agarwal S, Snavely N, Simon I, Seitz SM, Szeliski R (2009) Building rome in a day. In: ICCV
27. Crandall D, Owens A, Snavely N, Huttenlocher D (2011) Discrete-continuous optimization for large-scale structure from motion. In: CVPR 2011. Best paper award runner-up
28. Snavely N, Simon I, Goesele M, Szeliski R, Seitz SM (2010) Scene reconstruction and visualization from community photo collections. In: Proceedings of the IEEE
29. Grzeszczuk R, Kosecka J, Hile H, Vedantham R (2009) Creating compact architectural models by geo-registering image collections. In: IEEE international workshop on 3-D digital imaging and modeling, ICCV
30. Li Y et al (2012) Worldwide pose estimation using 3d point clouds. Computer vision-ECCV 2012. Springer, Berlin, pp 15–29
31. Snavely N, Garg R, Seitz SM, Szeliski R (2008) Finding paths through the world's photos, SIGGRAPH
32. Kosecka J, Zhang W (2007) Image based localization. IEEE Trans Rob
33. Li, Y, Snavely N, Huttenlocher DP (2010) Location recognition using prioritized feature matching. Computer vision-ECCV 2010. Springer, Berlin, pp 791–804
34. Baatz G et al (2012) Large scale visual geo-localization of images in mountainous terrain. Computer vision-ECCV 2012. Springer, Berlin, pp 517–530
35. Zakhor A et al (2013) User-driven geolocation of untagged desert imagery using digital elevation models
36. Baboud L et al (2011) Automatic photo-to-terrain alignment for the annotation of mountain pictures. In: IEEE conference on computer vision and pattern recognition (CVPR)
37. Snavely N, Seitz SM, Szeliski R (2006) Photo tourism: exploring photo collections in 3D. ACM Trans Graph (TOG) 25(3):835–846
38. Zamir AR, Dehghan A, Shah M (2013) Visual business recognition: a multimodal approach. In: Proceedings of the 21st ACM international conference on multimedia

39. Jacobs N, Miskell K, Pless R (2011) Webcam geo-localization using aggregate light levels. In: Applications of computer vision (WACV)
40. Jacobs N, Satkin S, Roman N, Speyer R, Pless R (2007) Geolocating static cameras. In: ICCV
41. Lalonde J-F, Narasimhan SG, Efros AA (2010) What do the sun and the sky tell us about the camera. IJCV 88(1):24–51
42. Zamir A, Shah M (2010) Accurate image localization based on google maps street view. Computer vision–ECCV 2010. Springer, Berlin, pp 255–268
43. http://www.satimagingcorp.com/
44. https://www.digitalglobe.com/
45. http://www.skyboximaging.com/
46. Schindler G, Dellaert F (2012) 4D cities: analyzing, visualizing, and interacting with historical urban photo collections. J Multimed 7(2):124–131
47. Matzen K, Snavely N (2014) Scene chronology. Computer vision-ECCV 2014. Springer International Publishing, pp 615–630
48. Torii A et al (2015) 24/7 place recognition by view synthesis. In: CVPR 2015–28th IEEE conference on computer vision and pattern recognition
49. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems
50. Zhou B et al (2014) Learning deep features for scene recognition using places database. In: Advances in neural information processing systems
51. Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems
52. Lin T-Y et al (2015) Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition
53. Workman S, Souvenir R, Jacobs N (2015) Wide-area image geolocalization with aerial reference imagery. arXiv:1510.03743