Delft University of Technology

Geo-Distinctive Visual Element Matching for Location Estimation of Images

Li, Xinchao; Larson, Martha; Hanjalic, Alan

DOI
10.1109/TMM.2017.2763323

Publication date
2018

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Multimedia

Citation (APA)
Li, X., Larson, M., & Hanjalic, A. (2018). Geo-Distinctive Visual Element Matching for Location Estimation of Images. *IEEE Transactions on Multimedia*, *20*(5), 1179-1194. https://doi.org/10.1109/TMM.2017.2763323

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.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination and volume/issue indication.

Geo-Distinctive Visual Element Matching for Location Estimation of Images

Xinchao Li, Martha Larson, and Alan Hanjalic, *Fellow, IEEE*

***Abstract*—We propose an image representation and matching approach that substantially improves visual-based location estimation for images. The main novelty of the approach, called distinctive visual element matching (DVEM), is its use of 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 and analysis for both geo-constrained and geo-unconstrained location estimation cases using two large-scale, publicly available datasets: the San Francisco Landmark dataset with 1.06 million street-view images and the MediaEval'15 Placing Task dataset with 5.6 million geo-tagged images from Flickr. We present examples that illustrate the highly transparent mechanics of the approach, which are based on commonsense observations about the visual patterns in image collections. Our results show that the proposed method delivers a considerable performance improvement compared to the state-of-the-art.**

***Index Terms*—Geo-location Estimation, information retrieval, large scale image retrieval.**

I. 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 [35], [39], in recommending travel tours and venues [6], [50], in discovering areas of interest [34], in photo stream alignment [53], and in event mining from media collections [8], [54].

While many modern mobile devices can automatically assign geo-coordinates to images during capture, a great number of images lack this information [44]. Techniques that automatically estimate the location of an image [4], [11], [17], [18], [26], [44] 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 at distinct locations may be unexpectedly similar.

Fig. 1. Colored boxes indicate potential visual matches 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).

The core idea underlying our approach to this challenge is depicted in Fig. 1, which illustrates the pattern of visual matching that we will exploit in this paper. 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. Note that Fig. 1 does not represent the output of any specific visual matching system, but is rather intended to represent the commonsense observation about patterns of visual matching that motivates our approach.

Manuscript received January 28, 2016; revised December 30, 2016, April 1, 2017, and August 5, 2017; accepted August 27, 2017. *(Corresponding author: Xinchao Li.)*

X. Li and A. Hanjalic are with the Multimedia Computing Group, Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: x.li-3@tudelft.nl; a.hanjalic@tudelft.nl).

M. Larson is with the Multimedia Computing Group, Delft University of Technology, Delft 2628 CD, The Netherlands, and also with Radboud University, Nijmegen 6525 HP, The Netherlands (e-mail: m.a.larson@tudelft.nl).

Color versions of one or more of the figures in this paper will be available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2763323

© 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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, the location in the third column is the *same* as the query location. The matching areas contain visual elements corresponding 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 paper 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 [31] for image location estimation.

As will be explained in detail in Sections II and III, DVEM represents a considerable 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 'complete' in that it is aggregated per location and is 'contextual' in that it is specific to the query image. This representation is computed in the final stage of search-based geo-location estimation, during which top-ranked results are processed. Because the representation is calculated at prediction time, it can change as necessary for different queries. As discussed in Section III, existing approaches involve steps that rely on image-level representations. In contrast, our approach aggregates directly from the visual-element level to the location-level. This fact allows for highly effective integration of the geo-distinctiveness information. The experimental results we present in this paper 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 paper is organized as follows. In Section II, we present the rationale underlying our proposed approach, DVEM, and describe its novel contribution in more detail. Then, in Section III, we provide an overview of the related work in the domain of image location estimation and position our contribution with respect to it. Section IV describes the DVEM approach in detail. Our experimental setup is explained in Sections V and VI reports our experimental results. Section VII concludes the paper and provides an outlook towards future work.

II. 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. 1. In Section I, 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. 1 left and middle) capture an underlying reality that is different from the reality captured by the query. The figure shows two typical sources of confounding visual elements: (a) elements corresponding to real-world objects that are able to move from one location to the other, such as a FedEx truck. (b) 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 bases its prediction on visual elements that are discriminative for a specific location.

Location representation Images taken at the true location (Fig. 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. 1, the variation in the perspective is significant between the query and the 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 across the whole location, rather than merely counting visual elements common between a query and

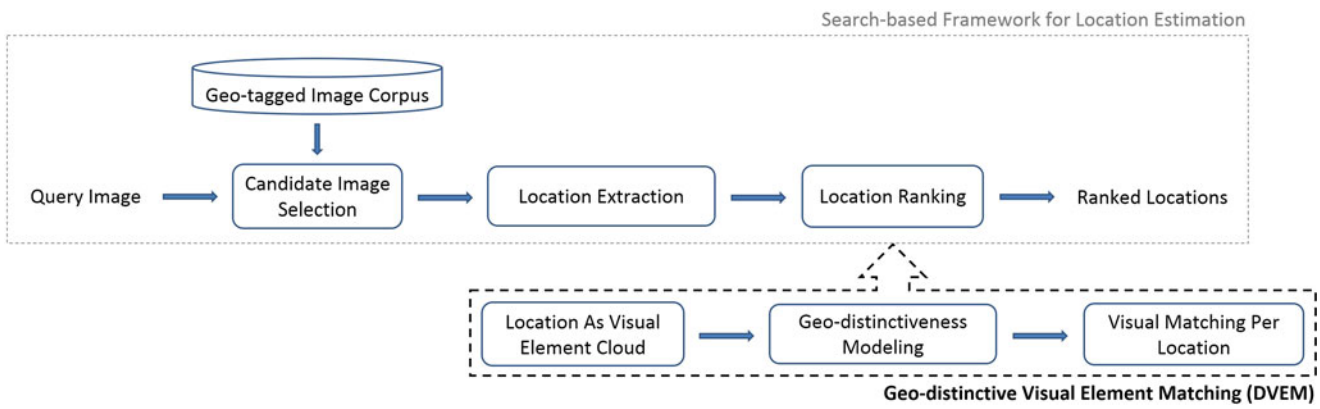


Fig. 2. Our proposed *Geo-distinctive Visual Element Matching* (DVEM) approach, depicted with its integration as the location ranking step of the generic search-based location estimation framework [31].

an image in the background collection. In particular, we aim to integrate two tendencies, which are illustrated by the right column of Fig. 1. Here, it can be see that the match between query image and true location typically 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 query image. These tendencies can be considered to be reflections of the commonsense 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 [31]. As previously mentioned, DVEM contributes to the processing step in a search-based geo-location estimation pipeline. Fig. 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 shows that it replaces the Location Ranking step of GVR.

Here, we provide an introduction to 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:

1) *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.
2) *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.
3) *Visual Matching per Location* calculates the ranking score for each candidate location with respect to the query to incorporate both the distinctiveness of visual elements and the matching strength between visual elements and the location.

These steps are explained in detail in Section IV, 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. These representations aggregate evidence for ranking a location with respect to its specific competitors for each query. The aggregation moves from the visual-element level to the location level.

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 the 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 [3]. DVEM uses the same 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 differentiates DVEM from the ways in which TF-IDF has been deployed for bag-of-feature-based image retrieval in the past.

1) DVEM does not match at the level of the item (i.e., individual image) but rather matches at 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.
2) 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 has 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 compared to the total number of images in the collection 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 theoretically infinite in the case of geo-location estimation at global scale.

As we will show by experimental results in Section VI, these advantages deliver an overall significant improvement of the location estimation performance compared to state-of-the-art methods. For completeness, we mention another connection to classic information retrieval techniques. Query expansion [3] refines the query using the initial result list. The fact that our 'contextual' representation is related to the query, means that our approach can be conceptually considered to be related to query expansion, which has also proven effective in the visual domain. An example from the area of object-based image retrieval is [10], which builds a model of the query. DVEM goes beyond query expansion, as it builds multiple models, one for each member of a set of candidate locations represented in the initial results list.

III. 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 [5], [23], [42], [48] or a finite set of predetermined locations [21], [27], [33], [37], [43]. The second is *geo-unconstrained* approaches, which estimate geo-location at a global scale [18], [31]. The difference lies in the specification of the geo-location estimation task. Under geo-constrained approaches, the locations to be predicted are defined in advance. Under geo-unconstrained approaches, the locations to be predicted are defined by the data that is available at the moment of prediction. Note that if off-planet images are included in the collection, geo-unconstrained location would also cover locations beyond global scale. The challenge of geo-unconstrained geo-location estimation is daunting: a recent survey [25] 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.

A. Geo-Constrained Content-Based Location Estimation

City-scale location estimation. Chen *et al.* [5] 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 referred to as the *San Francisco Landmark dataset*, 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 [36] was employed to predict the image's location by propagating the location information from the top-returned image.

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 [15], [40], [47], [48], [55]. Gopalan [15], 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. Tolias *et al.* [47] proposed an aggregated selective matching kernel to enforce selective feature matching to improve retrieval on urban sceneries and building photos. Sattler *et al.* [40] exploited implicit feature matching in their structure-based localization strategy, which builds a 3D model for each geo-tagged landmark/object, and then compares the query image against these 3D models to find its location. Torii *et al.* [48] described a representation of the repeated structures present in images, which is shown to be a distinguishing feature for place recognition. Zhang *et al.* [55] proposed a graph-based query specific fusion approach where multiple retrieval sets are merged and reranked to further enhance the retrieval precision. The experiments in Section VI-D make a comparison with all these approaches. In addition to these approaches, Cummins and Newman [12] focused on recognizing places in the context of detecting loop closure in SLAM (simultaneous localization and mapping) system. A probabilistic approach is proposed that incorporates information regarding visual words that co-occur. The approach is able to explicitly account for distortion in the visual environment.

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.* [16] 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 costs of both the offline learning and online querying phases, the experiment was conducted on a limited dataset of 25 k photos from Google Streetview taken in Pittsburgh, U.S., covering roughly an area of 1.2×1.2 km^2. *Beyond city scale.* Authors that go beyond city scale, may still address only a constrained number of locations. Kalantidis *et al.* [21] investigated 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. Our approach

resembles [21] 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. Li *et al.* [27] 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 an image's location by matching its visual content against this hierarchical structure. Subsequent to the initial submission of this paper for review, Weyand *et al.* [51] proposed a deep learning approach to image geo-location estimation. This approach divides the world into regions based on photo density, and formulates the location estimation task as image classification. Then, it employs convolutional neural networks to classify query image into one of the regions. We have reimplemented [51] in [52] for the purpose of studying geo-privacy. In our experiments, DVEM outperformed [51].

B. Geo-Unconstrained Content-Based Location Estimation

Estimating location from image content on a global scale faces serious challenges. First, there are effectively an infinite number of locations in the world. Geo-unconstrained location estimation must strive to be able to make predictions for as many of these locations as possible. 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 on the surface of the Earth. Finally, the visual variability of images taken at 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 [18]. They proposed to use visual scene similarity between images to support location estimation with the assumption that images with higher visual scene similarity are more likely to have been taken at the same location. In recent years, research on geo-unconstrained location prediction has been driven forward by the MediaEval Placing Task [25]. The Placing Task result most relevant to DVEM is our submission to the 2013 Placing Task [29]. This submission deployed a combination of local and global visual representations within the GVR system [30], [31], and out-performed other visual-content-based approaches that year. In this paper, we adopt [31] as a baseline, which represents our 2013 Placing Task submission.

Further, we focus on 2015, the most recent edition of the Placing Task [7], which received three submissions using visual-content-based approaches. Kelm *et al.* [22] exploited densely sampled local features (pairwise averaged DCT coefficients) for location estimation. Since this submission is not yet a functional, mature result [22], it is not considered further here. Li *et al.* [28] 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.* [24] 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 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 are dependent 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.

C. Geo-Distinctive Visual Element Modeling

As discussed in Section II, in a classical information retrieval system, document (item) distinctiveness is traditionally computed offline during the indexing phase at the level of the entire collection [3]. This approach is also used in the classical bag-of-feature-based image retrieval system. For example, the distinctiveness of each visual word is generated from its distribution in the image database, either as individual visual word [45], or as set of co-occurring visual words [9]. Note that our system uses the approach of [45] in the Candidate Image Selection step (first block in Fig. 2), as a standard best practice. Our novel use of geo-distinctiveness goes above and beyond this step, as we describe next.

The key example of the use of distinctiveness for content-based geo-location estimation is the work of Arandjelović and Zisserman [2], 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. Second, we use spatially verified salient points, rather than relying on the visual appearance of the descriptors of the salient points. As we will show with experimental results in Section VI, which uses Arandjelović and Zisserman [2] as one of the baselines, this added step of geo-distinctive visual elements matching significantly improves location estimation.

Where geo-distinctiveness has been used in the literature, it has been pre-computed with respect to the background collection (i.e., the database). Schindler *et al.* [42] mined geo-distinctive features from the database images, and used Information Gain to select discriminative features. Similarly, Knopp *et al.* [23] focused on the geo-distinctiveness of regions in the database images rather than of individual feature points. Turcot and Lowe [49] moved one step further to only select

features in the database images that are not only location-wise distinctive, but also geometrically robust. Doersch *et al.* [13] 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.* [14] incorporated 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.

Pre-computing distinctive features on the background collection is effective. However, it has drawbacks. First, the query image is not taken into account, as pointed out by Knopp *et al.* [23]. Second, for a given target image, it is not necessary to distinguish between every possible location, but rather only from the locations that are the closest candidates. As the collection grows larger, arguably generic geo-discriminative features will become less effective, and it will become more important to choose geo-discriminative features wisely in a query-specific manner. This consideration motivates the way in which we exploit geo-distinctiveness in our approach. We describe our use of geo-discriminativeness next.

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. We focus on the visual elements that we extract from the query, and model their geo-distinctiveness on the basis of the candidate locations for this particular query at prediction time. We calculate geo-distinctiveness with respect to the locations represented by the candidate images selected by the Candidate Image Selection step, and not with respect to the entire Geo-Tagged Image Corpus (cf. Fig. 2).

Independently, and in parallel with us developing our approach, Sattler *et al.* [41] also proposed to calculate geo-discriminative features using only the initial list of retrieved images. However, our approach differs from that of Sattler *et al.* [41] in a key aspect. As mentioned above, and explained in detail in Section IV, our approach works with locations with a 'contextual' representation we refer to as a visual element cloud. We use this representation through all the steps of calculating the score of each candidate location. Specifically, the contribution of geo-discriminativeness is calculated on the level of individual elements within the cloud. In contrast, Sattler *et al.* [41] made geo-discriminativeness contribute to the final ranking at the level of the image, rather than at the level of the visual element. This decision is not surprising, given that the focus of our paper is specifically geo-location estimation, whereas Sattler *et al.* [41] approach geo-location estimation first as an image-ranking problem. The superior performance of our approach, as shown by the experimental results in Section VI, demonstrates the effectiveness of our approach.

IV. Geo-Distinctive Visual Element Matching

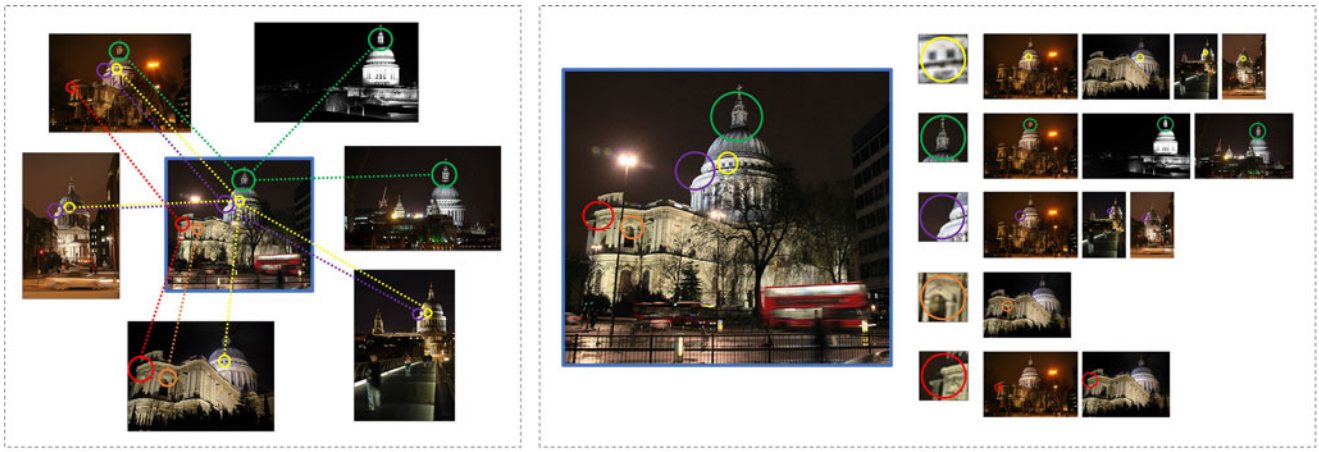
In this section, we present DVEM in depth, providing a detailed description of the components depicted in Fig. 2. We start with the GVR framework [31] (Fig. 2, top row), the generic search-based location estimation pipeline upon which DVEM builds. The framework was introduced in Section II. Here, we 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 pairwise geometric matching approach is used for this step [32]. 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 [31]. In the second step, Location Extraction, a set G of candidate locations is created by applying an interactive geo-clustering process using the geo-coordinates found by moving down the top ranked images [31]. 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 as to meet the prediction resolution of the system. 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. 2) in turn.

A. 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. Note that a separate visual element cloud is created for each location g in G. The cloud, illustrated in Fig. 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 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 pairwise geometric matching [32], as applied in the Candidate Image Selection step, but another geometric verification approach could also be chosen. The result of geometric matching



(a) (b)

Fig. 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 that these elements appear in.

is a set of one-to-one correspondences c between salient points in the query and in the individual images I_g [cf. Fig. 3(a)], 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 explained 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 $Score(g, q)$ 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. 3(b)]. Formally expressed, the visual element cloud $\mathbf{S}_g$ for location g is calculated as:

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

Here, $\mathbf{E}_g$ is the set of visual elements that occur in the query and link the query 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 jth 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 jth appearance of the visual element e is involved, and is defined by

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

Here, δ controls the smoothing speed as shown in Fig. 4. The choice of smoothing function is motivated by the need to take advantage of the information in the middle range of the matching score, and prevent values in the high range from dominating. Note that any smoothing function with the general form of the curve in Fig. 4 is potentially a viable smoothing function. Here, we investigate only (2).

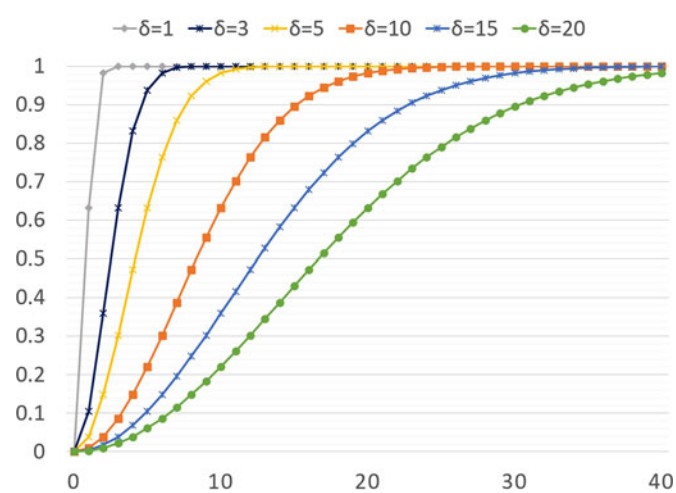


Fig. 4. Matching score smoothing function $w(e)_j$ vs. $IniScore(c)$ for various δ.

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.

B. Geo-Distinctiveness Modeling

We start our explanation of geo-distinctiveness modeling with an illustration of the basic mechanism. Fig. 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 (right-hand 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. 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 resolutions is shown in Figure (b), with the color changing from black to red to represent the increase of geo-distinctiveness.

Fig. 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. 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.

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. 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 thenxbrk defined as

$$W_{Geo}(e) = \begin{cases} \log(|G|/n(r(e))), & \text{if } n(r(e)) < \vartheta \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $|G|$ is the total number of location candidates. Further, the rest of the notation used in the formula is defined as follows: $r(e)$ 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. We note that (3) takes the general form of a standard IDF weight in information retrieval. We anticipate that any conventional variation on the formulation of IDF would be a viable weight. However, here we investigate (3).

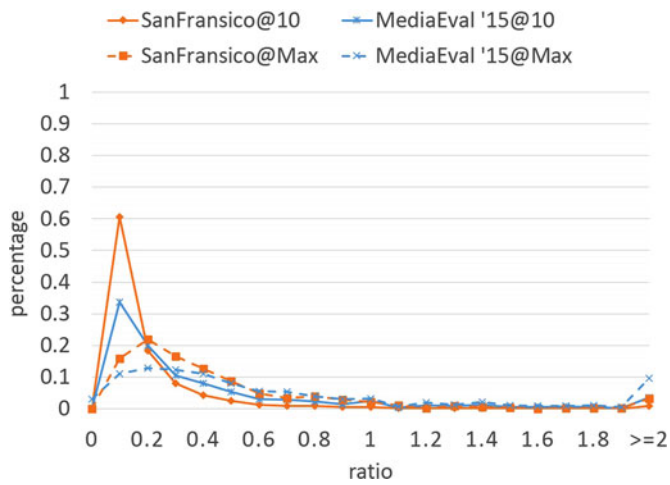


Fig. 6. Distribution of the ratio of number of unique visual elements between wrong locations and true locations averaged over all queries in the dataset. The scheme with @10 means the results are calculated based on the top-10 wrong locations in the initial ranked list for each query. The scheme with @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.

Fig. 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.

C. Visual Matching Per Location

We start our discussion of visual matching by considering the patterns of visual elements associated with a true match between a query image and a location. First, we investigate whether or not we can indeed expect relatively 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. These are the same geo-location estimation image sets used in our experiments and will be described in detail in Section V. Results are shown in Fig. 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 fact that there are relatively more unique visual elements present in the true location compared with a wrong location. This fact motivates us to include aggregation of visual elements as part of our visual matching model.

Next, we return to our earlier statement (see Section II) that we expect the match between queries and a true location to display (a) a 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. 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 select 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 and 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 score between location g and the 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 (3): $\mathbf{r}_q = (W_{Geo}(0), W_{Geo}(1), ..., W_{Geo}(k))$. 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,

$$Score(g, q) = \mathbf{r}_q \cdot \mathbf{r}_g = \sum_{e \in \mathbf{E}_g} \tilde{w}_e W_{Geo}(e)^2 \tag{4}$$

Further experiments, not reported here, show that the weights of $\mathbf{r}_q$ provide an essential contribution to the calculation. 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.

V. 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 (see Section VI-A), geo-contrained location estimation (see Section VI-B), geo-unconstrained location estimation (see Section VI-C), and our comparison with the state of the art (see Section VI-D).

A. Datasets

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 [5]: 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 1.06 M 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 VI-A and VI-B we use 10% of the test images for development, and report results on the other 90% of the test images. For the San Francisco landmark data set, we follow the common practice of considering the location of both the database photos and the query images to be the building ID, and not the geo-coordinates. The geo-location of an image is considered correctly predicted if the building ID is correctly predicted. Note that one query image can have multiple associated building IDs. For evaluation, a prediction is considered successful if the estimated location (building ID) is one of the query's building IDs.

MediaEval'15 Placing Task dataset [7]: This dataset is designed for global scale location estimation, i.e., geo-unconstrained location estimation. It is a subset of the YFCC100M collection [46], 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 4.6 M training and 1 M test images. Here again for our detailed experiments in Sections VI-A and VI-C 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. These geo-coordinates define the location of an image. 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.

B. 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 [32]. 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 [1], [19], [45]. To further speed up retrieval and improve accuracy, we use pairwise geometric matching in the following pipeline of state-of-the-art solutions:

1) Features & Vocabularies: Since up-right Hessian-Affine detector [38] and Root-SIFT [1] 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 word vocabulary with a size of 65,536 based on the features from the training images.
2) Multiple Assignment: To address the quantization noise introduced by visual word assignment, we adopt the strategy used in [20], 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 (on average each descriptor is assigned to four visual words), we only apply this at the query side.
3) Initial ranking: We adopt the Hamming Embedding technique combined with burstiness weighting proposed in [19] in the initial ranking phase.
4) Geometric verification: To find the reliable correspondences for DVEM, the pairwise geometric matching technology [32] 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 an experiment conducted on the development data, we established the importance of pruning. Specifically, we found that due to a high inter-similarity of the street view images taken in downtown San Francisco, removing the correspondences with a 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. 2) and for two baselines, as discussed next.

C. 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 (see Sections VI-B and VI-C) assesses the ability of DVEM to out-

perform other search-based geo-location estimation approaches, represented by VisNN and GVR:

1) *VisNN*: Our implementation of the 1-NN approach [18], 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.
2) *GVR*: Method used in [31], which expands the candidate images by their locations and uses the summed 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 [29], [30].

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

D. Evaluation Metrics

Our main evaluation metric is Hit Rate at the top N results ($HR@N$). Recall that given a query, the system returns a ranked list of possible locations. $HR@N$ measures the proportion of queries whose correct location falls within the top N predicted locations. $HR@1$ represents the ability of the system to correctly predict the location of a query image when it is forced to make a single prediction (top 1 result) for every query image.

In order to compare our method directly to other work on the San Francisco Landmark Dataset, we also report results using a *Precision-Recall Curve*. Like $HR@1$, our Precision-Recall Curve reflects the performance of the system with respect to its top-1 (best) estimate. The curve is generated by imposing a confidence threshold on the prediction, and changing the threshold so as to generate precision scores at a range of specific recall rates. Note that because we are interested in only the top-1 prediction, the following holds: If the score of the query exceeds the threshold—which means that a prediction is made for the query—then the precision is either 1 (location is correctly predicted) or 0 (location is not correctly predicted).

VI. Experimental Results

We implemented our DVEM framework on top of the object-based image retrieval system [32] by constructing a Map-Reduce-based structure on a Hadoop-based cluster containing 1, 500 cores. The initial visual ranking (the Candidate Image Selection step in Fig. 2) takes about 105 mins for San Francisco Landmark dataset (803 queries on a collection of 1.06 M photos) and about 88 hours for the MediaEval'15 dataset (1 M queries on a collection of 4.6 M photos). The DVEM stage is executed after the initial visual ranking, and takes 2.4 ms per query.

In this section, we report the experimental results and compare our DVEM method with reference methods in both the area of geo-constrained and of geo-unconstrained location estimation. We use part of the test data (10% for San Francisco Landmark dataset and 2% for MediaEval'15 dataset) as development data 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 IV-B, the frequency threshold ϑ defined in (3) and the query region size b defined in Section IV-C. The parameter δ defined in (2) is set empirically to 5 based on the general observation that the initial correspondence score generated by pairwise geometric matching [32] 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.

TABLE I
HR@1 (%) Comparison of DVEM on Development Data for San Francisco (Fixed Ground Truth) and MediaEval'15 Datasets ($r_{eval} = 1$ km) With Different a, b, and ϑ

	San Francisco $\vartheta = 5$				MediaEval '15 $\vartheta = 6$			
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

A. 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 data. We explore the parameter space with grid search, as shown in Table I. For both a and b, we considered the values 0, 30, 20 and 10 (Table I, 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. With a or b going below 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 Landmark 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 I, bottom). We choose $\vartheta = 5$ for the San Francisco Landmark 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. A manual investigation of the difference revealed that it can be attributed to the difference in the respective capture conditions. During the investigation it was observed that the queries in the San Francisco Landmark dataset 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 result in the salient

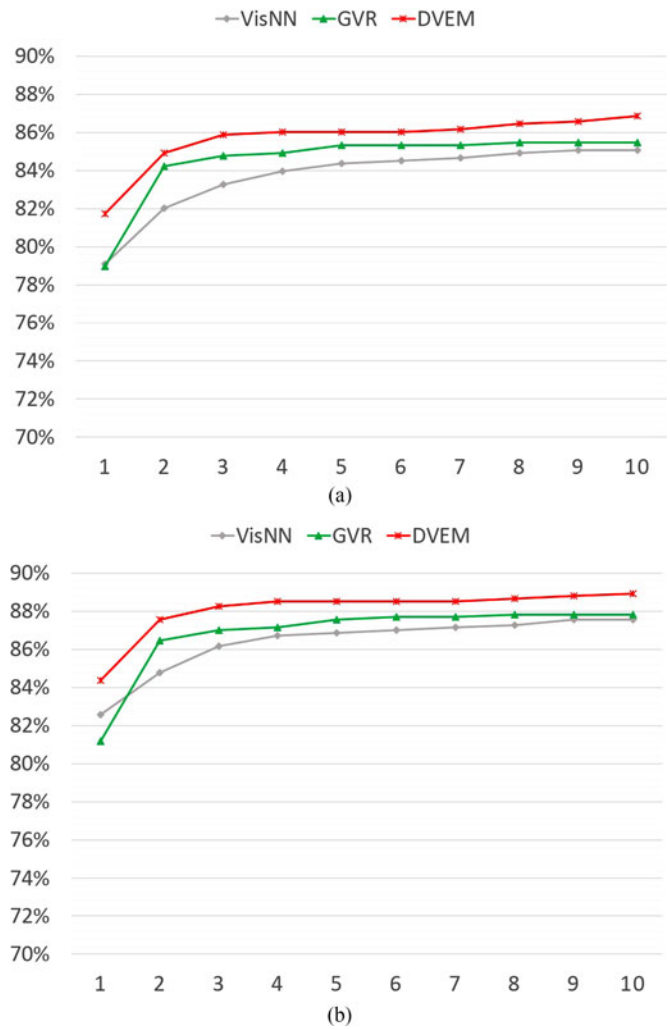


Fig. 8. HR@N performance (%) for varying N on the test set of the San Francisco Landmark dataset. (a) performance with respect to the original ground truth, (b) performance with respect to the fixed ground truth released in April 2014.

Fig. 9. Performance on the test set of the MediaEval'15 Placing Task dataset. (a) HR@1 performance (%) with respect to different evaluation radiuses, (b) HR@N performance (%) for varying N and at the evaluation radius of 1 km.

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 (specifically, $a = 10$) 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 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. Aggregating the salient points 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.

B. Geo-Constrained Location Estimation

The performance of different methods on the San Francisco Landmark dataset is illustrated in Fig. 8. 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 fixed ground truth released in April 2014 [see Fig. 8(b)].

GVR performs even worse than VisNN with respect to the fixed ground truth. This is due to the fact that in the street-view dataset the background collection images are captured by the survey vehicle, which can make multiple near-duplicate images per location. When a location contains the same visual elements as the query image, e.g., the fire escapes in Fig. 5(b), the summed visual similarity of images taken in this location will heavily influence the estimation. Recall from Section V-C that GVR expands the candidate images by their locations and uses the summed visual similarity of images located in one location as the ranking score. As such, GVR makes use of the summed visual similarity of images, and is impacted by this effect. In contrast, DVEM can handle this situation since it

Fig. 10. 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.

differentiates visual elements 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, 52 out of 803 (6.5%) query images do not correspond to any image in the database collection in terms of location. 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 14 k buildings based on the location of the camera [5]. This process 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%.

C. Geo-Unconstrained Location Estimation

Fig. 9 shows the performance of different methods with different values of r_{eval} [see Fig. 9(a)] and different Hit Rates [see Fig. 9(b)] 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. In contrast, DVEM is able to maintain robust performance in such cases. Fig. 10 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 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 that 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.

D. 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.

First, we look at geo-constrained location estimation using the San Francisco Landmark dataset. Results are reported in Figs. 11 and 12. We compare DVEM with the methods of Tolias *et al.* [47], Arandjelović and Zisserman [2], Torii *et al.* [48], Chen *et al.* [5], Zhang *et al.* [55], Gopalan [15], Sattler *et al.* [40], and Sattler *et al.* [41]. All results are calculated on the test set as defined in the San Francisco Landmark dataset releases, and are reported as they appear in each paper cited. For DVEM, we generate the Precision-Recall-Curve in Fig. 12 by thresholding

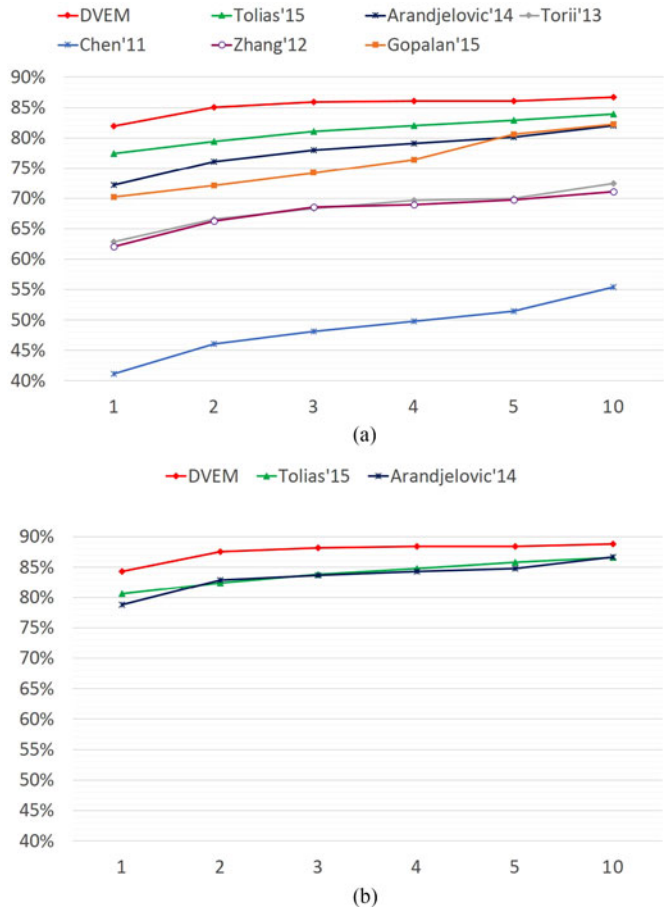


Fig. 11. HR@N performance (%) for varying N on the test set of the San Francisco Landmark dataset. (a) performance with respect to the original ground truth, (b) performance with respect to the fixed ground truth released in April 2014.

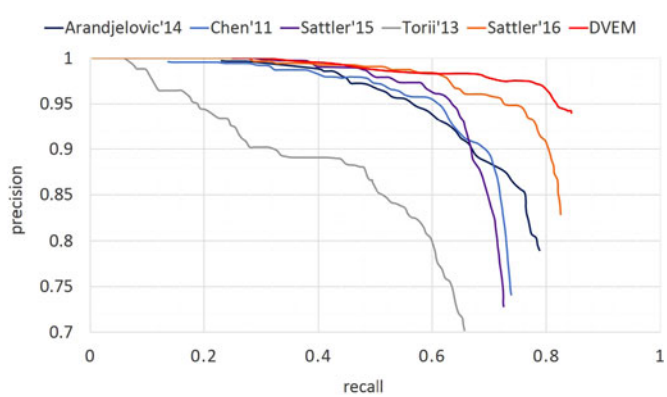


Fig. 12. Precision-Recall for the top-1 estimation on the test set of the San Francisco Landmark dataset with the fixed ground truth released in April 2014.

the score generated by (4). In all three graphs of Figs. 11 and 12, our proposed DVEM approach outperforms the state-of-the-art.

For completeness, we include additional discussion of our experimental design. The papers cited in Figs. 11 and 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 (see Section VI-A). Referring back to Table I, we can see that our

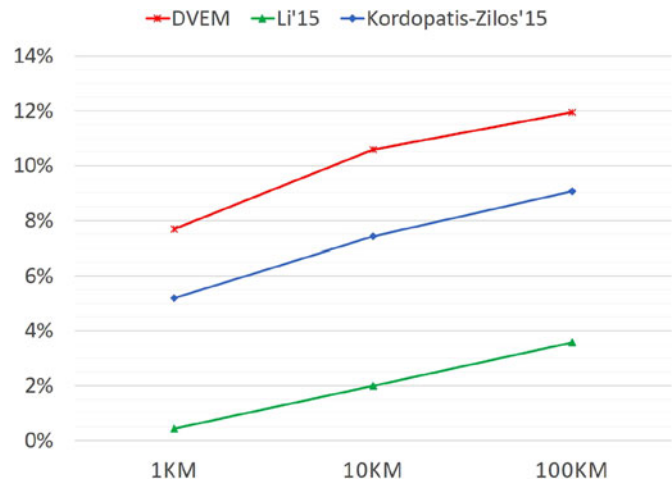


Fig. 13. HR@1 performance (%) with respect to different evaluation radiuses on the MediaEval'15 Placing Task dataset.

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 poorly tuned version of DVEM falls by this amount, it still remains competitive with well-tuned versions of the other approaches in Figs. 11 and 12. This assures us that the superiority of our approach does not lie in our choice of tuning.

Next, we look at geo-unconstrained location estimation. We compare our method to Li *et al.* [28], and the neural network-based representation-learning approach by Kordopatis-Zilos *et al.* [24]. 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. 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 leader board 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 I 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. 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.

VII. Conclusion

We have presented a visual-content-based approach for prediction of the geo-locations of images, based on commonsense

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 between a query and a location, and leveraging geo-distinctiveness. The 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 three additional observations about the value of the DVEM approach moving forward. First, 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 [24]. These can enhance DVEM, i.e., learned image representations would replace the bag-of-words pipelines in the initial ranking step (Candidate Image Selection in Fig. 2).

Second, note that as the background collection (i.e., the Geotagged Image Corpus in Fig. 2) grows, DVEM either needs to choose between a larger number of location candidates, or needs to sacrifice its performance at lower resolutions. How this trade-off is made in practice will depend both on the density distribution of images that are added to the background collection, and the impact of these images on patterns of visual similarity. Here, human behavior is critical. For example, if San Francisco suddenly starts building large numbers of identical buildings distributed throughout the city, this development will have an impact on DVEM. The interaction between how people design the environment around them, the patterns with which they take pictures, and automatic geo-location approaches such as DVEM will remain a fascinating direction of study as increasingly more images are captured, and become available for multimedia research.

Finally, independent of other developments, we believe that a key innovation of DVEM will remain important. Recall that 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. This ability to automatically adjust the precision of the prediction to the information (visual evidence) available in the data set is an important property of the algorithm, and that will deserve additional investigation in the future.

Acknowledgment

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

[1] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. 2012 IEEE Conf. Comput. Vision Pattern Recog.*, 2012, pp. 2911–2918.

[2] R. Arandjelović and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *Proc. Asian Conf. Comput. Vision*, 2014, pp. 188–204.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, vol. 463. New York, NY, USA: ACM Press, 1999.

[4] J. Cao, Z. Huang, and Y. Yang, "Spatial-aware multimodal location estimation for social images," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 119–128.

[5] D. M. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proc. 2011 IEEE Conf. Comput. Vision Pattern Recog.*, 2011, pp. 737–744.

[6] Y.-Y. Chen, A.-J. Cheng, and W. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1283–1295, Oct. 2013.

[7] J. Choi, C. Hauff, O. Van Laere, and B. Thomee, "The placing task at mediaeval," in *Proc. MediaEval 2015 Workshop*, vol. 1436, 2015.

[8] J. Choi, E. Kim, M. Larson, G. Friedland, and A. Hanjalic, "Evento 360: Social event discovery from web-scale multimedia collection," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 193–196.

[9] O. Chum and J. Matas, "Unsupervised discovery of co-occurrence in sparse high dimensional data," in *Proc. 2010 IEEE Conf. Comput. Vision Pattern Recog.*, 2010, pp. 3416–3423.

[10] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. 2011 IEEE Conf. Comput. Vision Pattern Recog.*, 2011, pp. 889–896.

[11] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 761–770 .

[12] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[13] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes Paris look like Paris?" *ACM Trans. Graph.*, vol. 31, no. 4, 2012, Art. no. 101.

[14] Q. Fang, J. Sang, and C. Xu, "Giant: Geo-informative attributes for location recognition and exploration," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 13–22.

[15] R. Gopalan, "Hierarchical sparse coding with geometric prior for visual geo-location," in *Proc. 2015 IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 2432–2439.

[16] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proc. 2013 IEEE Conf. Comput. Vision Pattern Recog.*, 2013, pp. 907–914.

[17] T. Guan, Y. He, J. Gao, J. Yang, and J. Yu, "On-device mobile visual location recognition by integrating vision and inertial sensors," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1688–1699, Nov. 2013.

[18] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. 2008 IEEE Conf. Comput. Vision Pattern Recog.*, 2008, pp. 1–8.

[19] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. 2009 IEEE Conf. Comput. Vision Pattern Recog.*, 2009, pp. 1169–1176.

[20] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[21] Y. Kalantidis, G. Tolias, E. Spyrou, P. Mylonas, and Y. Avrithis, "VIRaL: Visual image retrieval and localization," *Multimedia Tools Appl.*, vol. 51, pp. 555–592, 2011.

[22] P. Kelm, S. Schmiedeke, and L. Goldmann, "Imcube @ mediaEval 2015 placing task: A hierarchical approach for geo-referencing large-scale datasets," in *Proc. MediaEval*, vol. 1436, 2015.

[23] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *Proc. 11th Eur. Conf. Comput. Vision I*, 2010, pp. 748–761.

[24] G. Kordopatis-Zilos, A. Popescu, S. Papadopoulos, and Y. Kompatsiaris, "CERTH/CEA LIST at mediaEval placing task 2015," in *Proc. MediaEval*, vol. 1436, 2015.

[25] M. 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*. New York, NY, USA: Springer, 2015.

[26] M. Larson *et al.*, "Automatic tagging and geotagging in video collections and communities," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, Paper 51.

[27] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[28] L. T. Li *et al.*, "RECOD @ Placing task of mediaEval 2015," in *Proc. MediaEval*, vol. 1436, 2015.

[29] X. Li *et al.*, "Exploration of feature combination in geo-visual ranking for visual content-based location prediction," presented at the MediaEval Workshop, Barcelona, Spain, Oct. 18-19, 2013.

[30] X. Li, M. Larson, and A. Hanjalic, "Geo-visual ranking for location prediction of social images," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retrieval*, 2013, pp. 81–88.

[31] X. Li, M. Larson, and A. Hanjalic, "Global-scale location prediction for social images using geo-visual ranking," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 674–686, May 2015.

[32] X. Li, M. Larson, and A. Hanjalic, "Pairwise geometric matching for large-scale object retrieval," in *Proc. 2015 IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 5153–5161.

[33] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. 2009 IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 1957–1964.

[34] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 589–598.

[35] J. Liu *et al.*, "Presenting diverse location views with real-time near-duplicate photo elimination," in *Proc. 2013 IEEE 29th Int. Conf. Data Eng.*, 2013, pp. 505–516.

[36] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[37] W. Min, C. Xu, M. Xu, X. Xiao, and B.-K. Bao, "Mobile landmark search with 3D models," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 623–636, Apr. 2014.

[38] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. 2009 IEEE Conf. Comput. Vision Pattern Recog.*, 2009, pp. 9–16.

[39] S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community-contributed images," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 921–932, Jun. 2013.

[40] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *Proc. 2015 IEEE Int. Conf. Comput. Vision*, 2015, pp. 2102–2110.

[41] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proc. 2016 IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1582–1590.

[42] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. 2007 IEEE Conf. Comput. Vision Pattern Recog.*, 2007, pp. 1–7.

[43] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert, "Detecting and matching repeated patterns for automatic geo-tagging in urban environments," in *Proc. 2008 IEEE Conf. Comput. Vision Pattern Recog.*, 2008, pp. 1–7.

[44] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing Flickr photos on a map," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 484–491.

[45] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vision*, 2003, pp. 1470–1477.

[46] B. Thomee *et al.*, "The new data and new challenges in multimedia research," arXiv:1503.01817.

[47] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vision*, vol. 116, pp. 247–261, 2015.

[48] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. 2013 IEEE Conf. Comput. Vision Pattern Recog.*, 2013, pp. 883–890.

[49] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *Proc. 2009 IEEE 12th Int. Conf. Comput. Vision Workshops*, 2009, pp. 2109–2116.

[50] X. Wang *et al.*, "Semantic-based location recommendation with multimodal venue semantics," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 409–419, Mar. 2015.

[51] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 37–55.

[52] J. Choi, M. Larson, X. Li, K. Li, G. Friedland, and A. Hanjalic, "The geoprivacy bonus of popular photo enhancements," *Proc. ICMR*, 2017.

[53] J. Yang, J. Luo, J. Yu, and T. S. Huang, "Photo stream alignment and summarization for collaborative photo collection and sharing," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1642–1651, Dec. 2012.

[54] J. Yuan, J. Luo, and Y. Wu, "Mining compositional features from GPS and visual cues for event recognition in photo collections," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 705–716, Nov. 2010.

[55] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Proc. 12th Eur. Conf. Comput. Vision*, 2012, pp. 660–673.