Tag Features for Geo-Aware Image Classification

Shuai Liao, Xirong Li*, Heng Tao Shen, Yang Yang, Xiaoyong Du

***Abstract*—As geo tags recording where a picture was taken are becoming part of image metadata, studying image classification approaches that can favorably exploit geo tags and the underlying geo context is an emerging topic. This paper contributes to geo-aware image classification by studying how to encode geo information into image representation. Given a geo-tagged image, we propose to extract geo-aware tag features by tag propagation from the geo and visual neighbors of the given image. Depending on what neighbors are used and how they are weighted, we present and compare eight variants of geo-aware tag features. Using millions of Flickr images as source data for tag feature extraction, experiments on a popular benchmark set justify the effectiveness and robustness of the proposed tag features for geo-aware image classification.**

***Index Terms*—Geo-aware image classification, geo tags, tag features**

I. Introduction

GEO tags recording where a photo was taken are becoming part of image metadata, thanks to the widespread use of GPS-enabled cameras and smart phones. Social media platforms such as Flickr allow a user to manually assign geo tags to a specific image by pinpointing the image onto a digital world map [1], [2]. Moreover, research on automatic geo-tagging is making a good progress [3]–[5]. All this results in the ever-increasing amounts of geo-tagged images. As in early 2009, Flickr claimed to host over 100 million geo-tagged images[1]. Intuitively, knowing where an image was photographed may help predict what objects and scenes are present in the visual content, and consequently improve image classification. For instance, images taken in a zoo area have a higher chance to contain imagery of animals when compared to images taken in an urban area. It is thus valuable to study image classification approaches that can favorably exploit geo tags and the underling geo context.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was partly supported by NSFC (No. 61303184), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01) SRFDP (No. 20130004120006), SRF for ROCS, SEM, and Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2014-002).

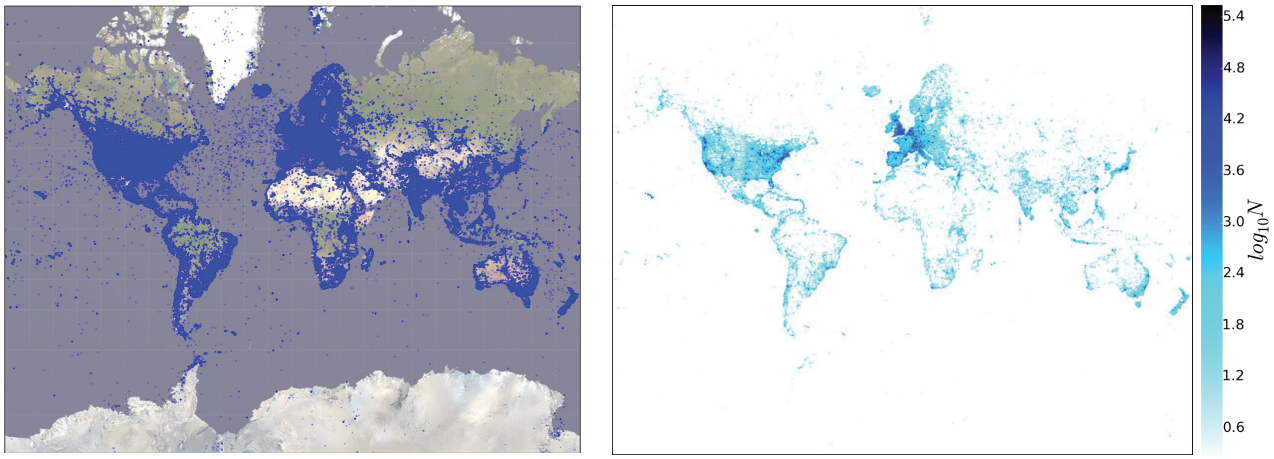
*Corresponding author. S. Liao, X. Li, and X. Du are with the Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China. X. Li is also with Shanghai Key Laboratory of Intelligent Information Processing, 200443 China. H. T. Shen is with the School of Information Technology and Electrical Engineering, University of Queensland, QLD 4072, Australia. Y. Yang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. Tel: +86-(0)10-62514562, e-mail: xirong.li@gmail.com

[1]http://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/

Towards geo-aware image classification, a number of methods have been introduced [6]–[14]. In [6], for instance, Moxley *et al.* build a geo-aware tag suggestion tool which recommend new tags to a given image based on tag statistics of images which were taken nearby the location of the given image. In Cao *et al.* [9], images taken in the same place are constrained to have the same labels. In a more recent work by Liu *et al.* [13], a geo-aware tag recommendation system is built by considering geo tags at a city level. All these works have demonstrated the potential of geo-aware image classification.

It is well recognized that features are crucial for image classification. Given powerful features, one can build an effective image classification system even with relatively simple classifiers. Nevertheless, few works are conducted to encode geo tags at the feature level. An initial work towards this direction is from Joshi and Luo [7]. Given a geo-tagged image, they first retrieve place entities that are closest to the location of the given image by querying a geographical information system database. Textual descriptions of these places are merged to form a bag-of-tags feature. Notice that visual content is not considered in this process, and consequently, two distinct images would have similar tags features as long as they are spatially close. So such a geo-aware feature is questionable. How to encode geo information at the feature level has not been well explored.

In the traditional content based scenario, features refer to low-level features directly extracted from pixels [15]–[17]. Among them, the bag of visual words (BovW) feature obtained by quantizing SIFT like local descriptors has been a *de facto* choice [18]. Nevertheless, it is difficult to directly encode geo information into the BovW feature. Wang *et al.* propose a novel tag-based image feature [19]. Given an unlabeled image, the tag feature is constructed by first retrieving visual neighbors from many user-tagged images, and then propagating tags associated with the neighbors to the given image. In order to make the tag feature geo-aware, one might consider replacing the visual neighbors by the image's geo neighbors for tag propagation. However, this straightforward solution is problematic due to the following reasons. First, the distribution of user-contributed images are inherently unbalanced, as Fig. 1 shows. This property requires a solution to be aware of the varied image density. Second, the effectiveness of geo clues for image classification is concept dependent [11]. Comparing two concepts 'dog' and 'panda', for instance, while the former can be seen almost everywhere, the latter can only be observed in specific zoos for common users. Finally, for varied reasons including measuring error of GPS sensors and unreliability of manual annotation, geo tags in social media are not meant to be perfect. By comparing geo tags



(a) Coverage diagram (b) Heat map of distribution

Fig. 1. **Placing a collection of 10 million Flickr images on the world map according to their geo tags**: (a) the geographical distribution of the images, and (b) the density visualized by a heat map in log scale (best viewed in color).

of Flickr images with manually corrected locations, a recent study suggests that deviation of geo tagging can range from hundreds of meters to several kilometers [20]. Solely exploring geo context is unreliable. Therefore, the feature in demand has to be both content-aware and geo-aware.

In this paper, we contribute to geo-aware image classification by encoding geo information at the feature level. Although a considerable amount of works have been done to exploit many geo-tagged images shared on social media, most of them focus on leveraging the geo clue to recommend novel tags [6], [8], [12], [21], without an explicit recognition of the tags as feature vectors for classification. Because of this, the output of a geo-aware tag recommendation system, e.g., [8], is chosen to optimize tag ranks, rather than for representation. Consequently, simply transforming the output to generate a tag feature vector is questionable. Our work bears some technical resemblance to previous works, as they perform tag propagation. However, we take a one step further by explicitly defining a new tag feature which can simultaneously describe the image content and the associated geo context. As for many images their geo tags are known to be inaccurate, we study the robustness of the proposed tag feature given noise geo tagging. To the best of our knowledge, such a quantitative evaluation has not been done before in the public literature of geo-aware image classification. A preliminary version of this work appears in [22]. We extend [22] as follows. Towards finding the optimal strategy for extracting geo-aware tag features, we study what neighbors to use and how the neighbors shall be weighted, increasing the number of geo-aware tag features from one in [22] to eight. Besides the newly added robust analysis, we enlarge the scale of the geo-tagged image collection from one million to ten million to reveal the influence of data scale on geo-aware tag features.

The rest of the paper is organized as follows. Related work is reviewed in Section II. Our approach to building geo-aware tag features is presented in in Section III. Experiments are setup in Section IV, followed by result analysis in Section V. Conclusions are given in Section VI.

II. Related work

This paper is to develop a semantic feature for geo-aware image classification, so it is in connection with research on semantic features and research on geo-aware image classification. In what follows, we first shortly review some representative works on the two topics.

A. Building semantic features

In the context of image retrieval, Rasiwasia *et al.* are probably the first to build a semantic feature as an alternative to low-level visual features [23]. They construct a semantic space, wherein each dimension corresponds to a specific concept classifier. For a given image, by classifying it using these classifiers, the authors map the given image into the semantic space. By applying visual concept classifiers to videos, Merler *et al.* build a semantic feature for video event recognition [24].

For image localization, Fang *et al.* describe visual patterns of a given place by a set of geographically informative attributes such as red wall and yellow roof of ancient architectures in Beijing, the mosques' rounded roof in Cairo, the windows with railing and the stylistic balcony in London [25]. As mainly designed for landmark recognition, these attributes are unsuitable for classifying generic images. Sergej proposes to describe an image with a probabilistic distribution over latent topics jointly learned by LDA model on many geo-tagged images [26], but the method requires test images to be associated with user tags. For tag suggestion, Liu *et al.* represent both images and tags in a common subspace, learned from many user-tagged images [13]. By restricting the images to be taken within the same city, the subspace carries geo information to some extent. However, the city level resolution is too coarse to precisely describe the geo context of a specific image.

Wang *et al.* propose a tag-based image feature [19]. For a given image, a tag histogram is built by retrieving images

visually similar to the given image from a large collection of user-tagged images, and counting tags associated with the similar images. In contrast to the predefined concepts as used in [23], [24], [27], user tags provide a more lively description of images. Furthermore, constructing the tag feature is lightweight as it requires no classifier training. A recent survey shows that tags propagated from visual neighbors as as effective as tags predicted by some heavily trained models for describing image content [16]. Thus, we opt to develop our solution on the basis of [19].

B. Geo-aware image classification

While quite a few papers have been published to exploit geo tags in varied ways, most of them do not consider building geo-aware features for image classification. In the work by Moxley *et al.* [6], for instance, a novel image is automatically annotated by a geo k nearest neighbor classifier. In that work, geo tags are used to compute the geographical distance between images. In a follow-up work [8], Kleban *et al.* propose to weight the geo neighbors in terms of their visual similarity to the novel image. Focusing on personal photo albums within the same context in terms of location and date, Cao *et al.* [21] propose a batch-tagging method. They utilize geo tags as a constraint to enforce that photos which are geographically close shall share the same labels. Instead of devising a new geo-based classifier, Li *et al.* build a geo-aware image classification system by combining multiple meta classifiers which are either content-based or geo-based [11]. Qian *et al.*exploit visual, geo, and temporal neighbors for personalized image tagging, without considering tag features [12] . Zhang *et al.* [28] propose to measure the semantic similarity of tags by comparing geo-spatial, temporal, and geo-temporal patterns of use in a large collection of geo-tagged Flickr images. According to their study, tags that have strong geo-temporal patterns are related to periodic events such as festivals and seasons. We are interested in generic visual concepts which may lack geo-temporal patterns.

Different from the above works, Joshi and Luo introduce a tag feature by a geo-decoding using a geographical information system [7]. The system contains over 8 million place names all over the world with manually edited descriptions. Given a test image, its geo tags are used to localize places nearby. The descriptions of these places are combined to form the tag feature. As aforementioned, this tag feature fully ignores the visual content. By contrast, our tag features are both content-aware and geo-aware.

III. Our Approach

By exploiting many user-tagged images and associated geographical information, we aim to build a novel tag feature for geo-aware image classification. For the ease of consistent description, we first introduce some notations. Let x be a specific image, which is associated with geo tags showing the location where the image was taken. We use t to denote a user tag, and $\mathcal{V} = \{t_1, \ldots, t_m\}$ as a vocabulary of m tags. Representing the image by these tags results in a tag feature of m dimensions, denoted as $T(x)$. We will build image classifiers upon the tag feature. In order to construct a geo-aware version of $T(x)$, we need a source set of many images which contain both user tags and geo tags. While previous work exploits only visual affinity of such a source set [19], we exploit both visual and geo affinity for deriving tag features that are both content-aware and geo-aware. To clarify the difference, we use $\mathcal{S}_{vis}$ to denote a source set from which visually similar images are found, and $\mathcal{S}_{geo}$ for another source set wherein geographically close images are found. While the two sets can be the same in theory, computing geo affinity is much more light than computing visual affinity. Consequently, the scale of $\mathcal{S}_{geo}$ can be much larger compared to $\mathcal{S}_{vis}$.

Next, we depict how to build geo-aware tag features in Section III-A, followed by their usage for image classification in Section III-B.

A. Strategies for building geo-aware tag features

For a given image, we encode the geo context of this image into its tag feature by propagating tags from the image's geo neighbors. As noted in Section I, fully relying on the geo neighbors could result in suboptimal tag features. Exploiting both geo and visual affinities is necessary. Depending on what neighbors are considered and how the neighbors are weighted, we propose multiple strategies, resulting in eight variants of geo-aware tag features.

Strategy 1: Tag propagation from geo neighbors. Given an image x, we first retrieve its k nearest geo neighbors from the source set $\mathcal{S}_{geo}$ in terms of the geographical distance. For each tag $t \in \mathcal{V}$, we count the number of the geo neighbors labeled with t. As illustrated in left part of Fig. 2, the resultant tag histogram, after l_1 normalization, forms the tag feature. We use $T_{gnn}(x)$ to indicate the tag feature generated from the geo neighbors.

As aforementioned, the density of user-contributed images varies over locations, and consequently yields geo neighbors of varied quality. In particular, in areas of low density, geo neighbors in distance should not be treated equally as those located nearby. The idea is to differentiate them by assigning larger weight to the nearer neighbors and smaller weights to more distant ones. For each tag $t \in \mathcal{V}$, we accumulate the weights of geo neighbors labeled with t, instead of the binary count as for $T_{gnn}(x)$.

To compute the weights, we consider the geo affinity and the visual affinity, respectively. For the geo affinity, we compute the geo distance $d_{geo}(x, x')$ between a specific geo neighbor x' and the given image x. The distance is converted to a weight by $\exp(-\frac{d_{geo}(x,x')}{h_{geo}})$, where h_{geo} is an attenuation coefficient. The resultant tag feature is denoted as $T_{gnn_{gw}}(x)$.

In a similar vein, we compute the visual affinity based weight as $\exp(-\frac{d_{vis}(x,x')}{h_{vis}})$, where $d_{vis}(x, x')$ is an image-wise visual distance function, and h_{vis} is an attenuation coefficient. The corresponding tag feature is denoted as $T_{gnn_{vw}}(x)$. Notice that the way this feature extracted is

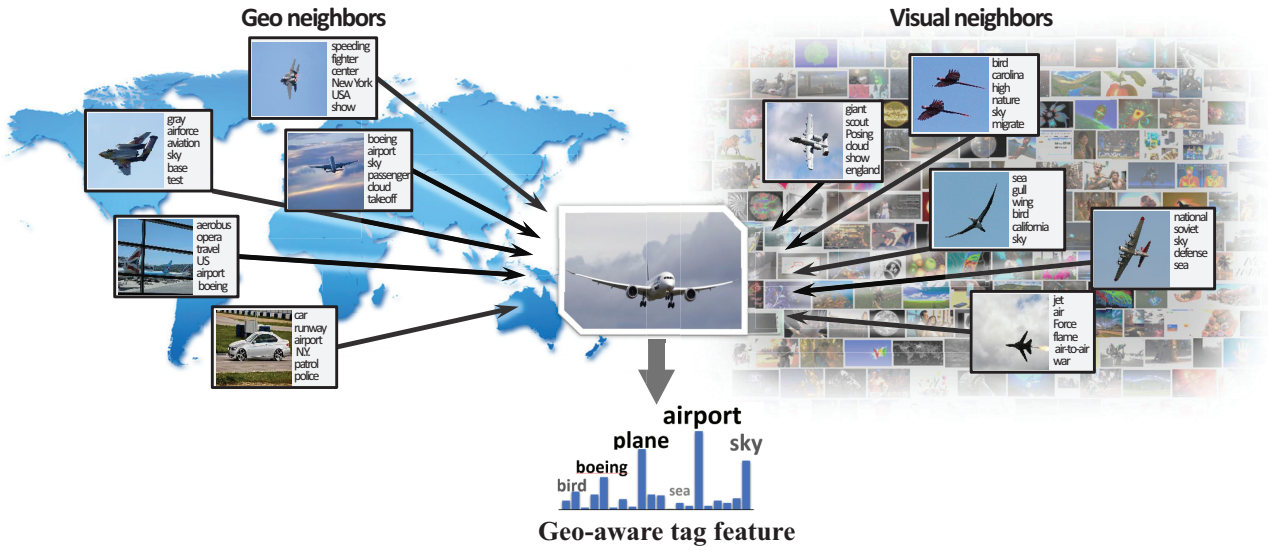


Fig. 2. **A conceptual illustration of the proposed tag features for geo-aware image classification**. Given a novel geo-tagged image, we first retrieve its geo neighbors and visual neighbors from many socially tagged images. User tags of these geo/visual neighbors are then propagated to the given image to form a histogram of tags. The joint use of the geo and visual neighbors make the tag feature both content-aware and geo-aware.

technically close to the image annotation model by Kleban *et al.* [8], when viewing the output of their model as individual dimensions of the tag feature.

Strategy 2: Tag propagation from visual-reranked geo neighbors. We are given with geo tags not meant to be perfect. The deviation of geo tags from the true location where an image was taken may give an erroneous profile of the geo context. Though the neighbors are weighted in Strategy 1, their order remains unchanged. So we consider directly manipulating the order by visual reranking. In particular, we first retrieve a relatively large geo neighbors candidate set and then re-rank them according to their visual affinity to the given image. We use $T_{gnn_{vr}(x)}$ to denote the tag feature produced by this strategy.

Strategy 3: Tag propagation from visual/geo neighbors. As the effectiveness of geo clues for image classification is concept dependent, constructing tag features solely from geo context is suboptimal. Next, we investigate strategies that deriving tag features by propagating tags from both visual and geo neighbors simultaneously. This idea is illustrated in Fig. 2, wherein on the left side we first retrieve the k geo neighbors from $\mathcal{S}_{geo}$ and then on the right side we retrieve another set of k visual neighbors from $\mathcal{S}_{vis}$. Tag propagation is performed from the union of visual and geo neighbors. This simple strategy makes the resultant tag features aware of visual content and the underlying geo context.

According to the use of unweighted/weighted geo neighbors in Strategy 1, we denote these new variants of the tag feature as $T_{vnn\cup gnn}(x)$, $T_{vnn\cup gnn_{gw}}(x)$, $T_{vnn\cup gnn_{vw}}(x)$ and $T_{vnn\cup gnn_{vr}}(x)$.

Notice that the visual neighbors are treated equally in the tag propagation process. Our consideration of not weighting them in terms of their geo or visual distance to the given image is as follows. First, except for a few landmarks, visual patterns shared by the given image and its visual neighbors could appear in many places of the world. Hence, it does not make sense to re-weight visual neighbors in terms of the geo distance. Second, given high-dimensional visual features, the distance variance in the neighborhood is relatively small. Consequently, the difference made by the visual distance based weights is marginal.

For the ease of reference, we summarize the eight features in Table I. For a more comprehensive picture, we also analyze the efficiency in tag feature extraction. The amount of computational work is mainly on finding visual/geo neighbors and extracting a tag histogram from the neighbors. Suppose the dimensionality of the visual feature vector is p. Finding k visual neighbors from $\mathcal{S}_{vis}$ has a computational complexity of $O(p\cdot|\mathcal{S}_{vis}|+k\cdot\log|\mathcal{S}_{vis}|)$. The complexity of computing the geographical distance between two points, as the great-circle distance we use in this paper, is constant and small. Hence, retrieving k geo neighbors from $\mathcal{S}_{geo}$ has a complexity of $O(|\mathcal{S}_{geo}|+k\cdot\log|\mathcal{S}_{geo}|)$. In addition, the neighbor weighting schemes also introduce some overhead, e.g., $O(k)$ for $T_{gnn_{gw}}(x)$ and $O(p\cdot k)$ for $T_{gnn_{vw}}(x)$. The complexity of extracting the tag histogram is $O(k\cdot l+m)$, where l is the average number of tags per image. The overall complexity is given in Table I.

B. Image classification using tag features

Since the tag features are designed to capture higher level semantics, they are complementary to visual features to some extent. Hence, geo-aware image classification by multi-modal fusion is meaningful.

There are a large body of research works for multimedia fusion [11], [29], [30]. In the context of web image search, Wang *et al.* [29] propose to integrate multiple modalities in a graph-based learning framework for re-ranking, achieved by linearly combining normalized graph Laplacian regularizers generated from different modalities. As noted in

TABLE I
VARIANTS OF GEO-AWARE TAG FEATURES STUDIED IN THIS PAPER.

Tag feature	What neighbors?	How to weight geo neighbors?	Computational complexity
$T_{gnn}(x)$	geo neighbors	Uniform	$O(\|\mathcal{S}_{geo}\| + k \cdot \log \|\mathcal{S}_{geo}\| + m)$
$T_{gnn_{gw}}(x)$	geo neighbors	$\exp(-\frac{d_{geo}(x,x')}{h_{geo}})$	$O(\|\mathcal{S}_{geo}\| + k \cdot \log \|\mathcal{S}_{geo}\| + m)$
$T_{gnn_{vw}}(x)$	geo neighbors	$\exp(-\frac{d_{vis}(x,x')}{h_{vis}})$	$O(\|\mathcal{S}_{geo}\| + k \cdot \log \|\mathcal{S}_{geo}\| + pk + m)$
$T_{gnn_{vr}}(x)$	visual-reranked geo neighbors	Uniform	$O(\|\mathcal{S}_{geo}\| + k \cdot \log \|\mathcal{S}_{geo}\| + (p + \log k)k + m)$
$T_{vnn \cup gnn}(x)$	union of geo and visual neighbors	Uniform	$O(p \cdot \|\mathcal{S}_{vis}\| + \|\mathcal{S}_{geo}\| + k \cdot (\log \|\mathcal{S}_{vis}\| + \log \|\mathcal{S}_{geo}\|) + m)$
$T_{vnn \cup gnn_{gw}}(x)$	union of geo and visual neighbors	$\exp(-\frac{d_{geo}(x,x')}{h_{geo}})$	$O(p \cdot \|\mathcal{S}_{vis}\| + \|\mathcal{S}_{geo}\| + k \cdot (\log \|\mathcal{S}_{vis}\| + \log \|\mathcal{S}_{geo}\|) + m)$
$T_{vnn \cup gnn_{vw}}(x)$	union of geo and visual neighbors	$\exp(-\frac{d_{vis}(x,x')}{h_{vis}})$	$O(p \cdot \|\mathcal{S}_{vis}\| + \|\mathcal{S}_{geo}\| + k \cdot (\log \|\mathcal{S}_{vis}\| + \log \|\mathcal{S}_{geo}\| + p) + m)$
$T_{vnn \cup gnn_{vr}}(x)$	union of visual-reranked geo neighbors and visual neighbors	Uniform	$O(p \cdot \|\mathcal{S}_{vis}\| + \|\mathcal{S}_{geo}\| + k \cdot (\log \|\mathcal{S}_{vis}\| + \log \|\mathcal{S}_{geo}\| + p) + m)$

[29], a prerequisite for the graph-based learning algorithm is that given a textual query, the images to be re-ranked are partially labelled, and the label information are used as a priori to regularize the learning process. Retrieved by a commercial search engine, the labels of the initial search results in [29] are relatively reliable. In contrast, no initial results are available for geo-aware image classification, making the graph-based learning unsuitable in our context. Therefore, per modality we build SVM classifiers which require no label information on the test data.

On combining heterogeneous SVM classifiers, Guillaumin *et al.* [30] use a Multiple Kernel Learning (MKL) framework, where kernels of different modalities are linearly combined. An attractive property of this framework is that both the classifier parameters and the combination weights are jointly optimized. However, for the consideration of efficiency, current MKL implementations, e.g., [31], require the full kernel matrices of every modality to be precomputed and cached into memory. This limits their scalability for large-scale classification. As an alternative, Li *et al.* [11], [32] suggest that combining trained classifiers by late fusion with the fusion weights optimized by coordinate ascent is also effective. By optimizing one weight per time, coordinate ascent is able to effectively maximize non-differentiable performance metrics, e.g., the popular average precision. We thus go for this simpler and more scalable strategy for multi-modal fusion. As discussed in Section I, the effectiveness of the geo clues is concept dependent. Hence, we learn the fusion weights per concept.

Given the tag features, we build textual classifiers for classifying unlabeled and unseen images. In particular, we adopt the Fast Intersection Kernel SVMs (FikSVMs) proposed by Maji *et al.* [33] for its good performance and high efficiency. By computing the decision function via linear interpolation on a fixed array of precomputed points, FikSVMs makes its prediction complexity irrelevant with respect to the number of support vectors. Similarly, we use FikSVMs to build visual classifiers based on the bag of visual words feature.

IV. EXPERIMENTAL SETUP

A. Data sets

Source sets. For extracting the proposed tag features, we need many images which are associated with both user tags and geo tags. To that end, we instantiate $\mathcal{S}_{geo}$ and $\mathcal{S}_{vis}$ with a set of one million Flickr images collected by Li *et al.* [11]. Covering photos taken in over 100 countries by 145K distinct users, this set is diverse in terms of both visual content and geographical distribution. Further, to study the impact of the source set size on the tag features, we enlarge $\mathcal{S}_{geo}$ by adding geo-tagged Flickr images from the MediaEval Placing tasks 2013 and 2014 [34], with batch-tagged images excluded. This results in a set of 10 million images, and consequently allows us to study the influence of the dataset scale on geo-aware tag features. Notice that we fix $\mathcal{S}_{vis}$ to be the 1M set, as the effect of data scale on visual neighbors has been discussed in the literature, c.f. [19], [35].

Ground-truth data. To evaluate a geo-aware image classification system, ground truth that contains both geo tags and expert labels are required. To fulfill this requirement, we use the geo-tagged part of the popular NUS-WIDE set [36], resulting in a training set of 41,173 images and a test set of 27,401 images. We make a random partition of the training set, where 70% of the data is used for training classifiers, and the remaining 30% is used for optimizing the combination of textual and visual classifiers. Ignoring rare concepts (with less than 10 positive examples), we test on 75 concepts covering an array of objects, scenes, and events such as 'dog', 'valley', and 'protest'.

B. Implementation

Parameters for tag features. Following [19], we construct the tag vocabulary $\mathcal{V}$ by selecting the top 2,000 most frequent tags in the 1M set, with stop words, camera brands, and non English words excluded beforehand. So all tag features have a dimensionality of 2,000.

BovW features are extracted for finding visual neighbors and for building visual classifiers. For each image, we use the color descriptor software [18] to extract many SIFT descriptors in a dense manner. With the descriptors

quantized by a precomputed codebook of size 1,024, the image is represented by a 1,024-dim BovW vector. The visual distance function $d_{vis}(x, x')$ is defined as the l_1 distance between the visual feature vectors, while the geographical distance $d_{geo}(x, x')$ is implemented by the great-circle distance between two points. The attenuation coefficients h_{vis} and h_{geo} are empirically set to be 0.681 and 10. For the number of neighbors k, by trying a range of values, we find $k = 150$ a solid choice.

For each concept and for each tag feature, we train FikSVM classifiers using the implementation of [37].

Evaluation criteria. We report Average Precision (AP) per concept, and mean Average Precision (mAP) to measure the overall performance.

V. Experiments and Results

Four experiments are designed to reveal how the tag features perform when used alone or in combination with visual features, how sensitive they are with with respect to extra noise artificially introduced into geo tagging, and how they react as the source set grows.

A. Experiment 1. Classification using Tag Features only

In this experiment, image classifiers are built with only tag features. As shown in Table II, building geo-aware tag features solely by the geo clue is not always better than the standard tag feature: T_{gnn} with an mAP of 0.138 is worse than T_{vnn} having an mAP of 0.159. The effectiveness of T_{gnn} fully depends on the quality of the geo neighbors. The more distant a neighbor is found, the less likely that the neighbor's tags are useful to describe the test image. For a better understanding of the relatively low performance of T_{gnn}, for each test image we analyze how its geo neighbors are geographically distributed. As shown in Fig. 3, for only 20% of the test images, their 150 geo neighbors are found within a radius of one kilometer. For over 50% of the test images, their 150 geo neighbors cannot be fully retrieved within a radius of ten kilometers. The lack of reliable geo neighbors affects the effectiveness of T_{gnn}. This issue can be tackled to some extent by assigning weights to the neighbors or reranking them in terms of their visual distance, as justified by the result that $T_{gnn_{vr}}$, $T_{gnn_{gw}}$, and $T_{gnn_{vw}}$ outperform T_{vnn}.

The top performers are $T_{vnn \cup gnn_{gw}}$, $T_{vnn \cup gnn_{vw}}$, and $T_{vnn \cup gnn}$, scoring mAP of 0.293, 0.273, and 0.271, respectively. The results show that joint use of both visual and geo neighbors is effective. Notice that while $T_{gnn_{vr}}$ is the best when using only geo neighbors, the union of the visual neighbors and the visually reranked geo neighbors yields less effective tag features, as $T_{vnn \cup gnn_{vr}}$ is worse than $T_{vnn \cup gnn}$. This is because the visual modality is emphasized by visual reranking, making the geo and visual neighbors less complementary to each other.

Concerning the superior performance of $T_{vnn \cup gnn}$ against T_{gnn}, one might attribute the result to the double use of the neighbors. Compared to T_{vnn} using k neighbors, $T_{vnn \cup gnn}$ employs $2 \times k$ neighbors in theory. So we report the performance given different values of k. As shown in Fig. 4, $T_{vnn \cup gnn}$ beats T_{vnn} under different values of k. Also we observe that the tag features are robust with respect to the choice of k.

TABLE II
Experiment 1. Image classification using tag features only. The features have been sorted in terms of their performance.

Tag Feature	mAP
$T_{vnn \cup gnn_{gw}}$	0.293
$T_{vnn \cup gnn_{vw}}$	0.273
$T_{vnn \cup gnn}$	0.271
$T_{vnn \cup gnn_{vr}}$	0.245
$T_{gnn_{vr}}$	0.194
$T_{gnn_{gw}}$	0.176
$T_{gnn_{vw}}$	0.169
T_{vnn}	0.159
$T_{GlobalAnno}$	0.142
T_{gnn}	0.138

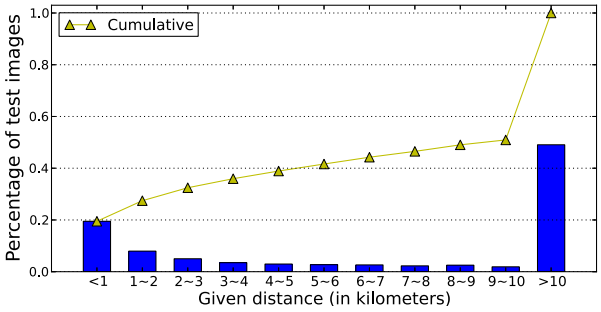


Fig. 3. **Percentage of the test images whose 150 geo neighbors can be fully retrieved within a given distance**. For more than 50% of the test images, their geo neighbors are not fully covered within a radius of 10 kilometers.

As aforementioned, while [8] is not meant for building tag features, the output of their model can technically be used to generate a tag feature, denoted as $T_{GlobalAnno}$. As shown in Table II, $T_{GlobalAnno}$ is less effective than $T_{gnn_{vw}}$. The difference is due to the fact that when computing the visual weights, $T_{GlobalAnno}$ uses the l_2 distance according to [8], while $T_{gnn_{vw}}$ uses the l_1 distance.

B. Experiment 2. Classification using Visual and Tag Features

To study whether the geo-aware tag features are helpful when used together with visual features, we implement and compare the following systems:

1) $Visual$: Classification using the visual feature only,
2) $Visual + T_{vnn}$: Combining the visual classifier and the textual classifier built upon T_{vnn} [19],
3) $Visual + T_{vnn \cup gnn}$: Combining the visual classifier and the textual classifier built upon $T_{vnn \cup gnn}$ [22],
4) $Visual + T_{vnn \cup gnn_{vw}}$: Combining the visual classifier and the textual classifier built upon $T_{vnn \cup gnn_{vw}}$,
5) $Visual + T_{vnn \cup gnn_{gw}}$: Combining the visual classifier and the textual classifier built upon $T_{vnn \cup gnn_{gw}}$,

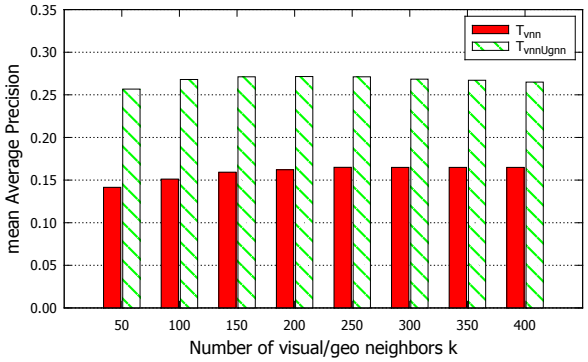


Fig. 4. **The influence of the number of neighbors k on the effectiveness of the tag features**. While both T_{vnn} and $T_{vnn\cup gnn}$ are robust with respect to the choice of k, $T_{vnn\cup gnn}$ consistently outperforms T_{vnn}.

TABLE III
SYSTEM-LEVEL COMPARISON.

System	mAP
GlobalAnno [8]	0.080
GeoVisualKNN [12]	0.113
$Visual$	0.226
$Visual + geoknn$ [11]	0.251
$Visual + T_{vnn}$ [19]	0.236
This work	
$Visual + T_{vnn\cup gnn}$	0.325
$Visual + T_{vnn\cup gnn_{vw}}$	0.327
$Visual + T_{vnn\cup gnn_{gw}}$	0.339
$Visual + T_{vnn\cup gnn_{vr}}$	0.298

6) $Visual + T_{vnn\cup gnn_{vr}}$: Combining the visual classifier and the textual classifier built upon $T_{vnn\cup gnn_{vr}}$.
In addition, we compare with the following geo-aware image classification systems, which do not consider geo information at the feature level:
7) $Visual + geoknn$: Combining the visual classifier and a geo KNN classifier [11],
8) *GeoVisualKNN*: A KNN classifier exploits both visual and geo neighbors [12].
9) *GlobalAnno*: A geo KNN classifier with geo neighbors weighted by their visual similarity to test images [8].

For a fair comparison, we let all the combined systems use the same visual classifiers, and the weights optimized by the same technology on the validation set. Table III shows the results of Experiment 2. Unlike the other systems, GlobalAnno and GeoVisualKNN use no training data from NUS-WIDE, hence their performance are much lower. We confirm previous findings that the geo clue is helpful for improving image classification [11], and adding tag features helps also [19]. The systems using the visual features and the geo-aware tag features perform best, showing that the proposed geo-aware tag features are effective when used in combination with the visual feature.

We compare coordinate ascent based late fusion and Multiple Kernel Learning based fusion. Following [30], we use the SimpleMKL package [31]. As the package needs to load full kernel matrices of all modalities into main memory, we use 70% of the training set, the maximum data size our computation facility can handle. For Visual + $T_{vnn\cup gnn}$, the mAP of SimpleMKL is 0.311, lower than its late fusion counterpart (mAP of 0.325). Besides the use of less training data, another possible reason for the relatively lower performance of SimpleMKL is that its learning process does not directly optimize mAP.

A per-concept comparison is given in Fig. 5. For 42 out of the 75 concepts, $Visual + T_{vnn\cup gnn}$ improves over $Visual + T_{vnn}$ with an absolute gain larger than 0.05. Sorting the concepts in descending order by the absolute improvement, we observe that the top ranked five concepts are 'whale', 'horse', 'bear', 'temple', and 'coral', while the bottom ranked five concepts are 'sky', 'moon', 'cloud', 'toy', and 'garden'. The occurrence of the top ranked concepts is much geographically dependent than the bottom ranked concepts. The results suggest that the proposed tag feature is aware of an image's geo context.

C. Experiment 3. Robust Analysis with Extra Geo Noises

Although the geo tags are already known to be inaccurate, we study the robustness of the proposed tag features by artificially adding more noises. In particular, for each train/test image, we let its geo tags deviate from the original coordinate in a randomly picked direction, with the deviation distance chosen from {10 meters, 100 meters, 1 km, 2 km, 5 km, 10 km, 20 km}. We conduct robust analysis for $T_{gnn}(x)$, $T_{vnn\cup gnn}(x)$, and $Visual + T_{vnn\cup gnn}(x)$.

As shown in Fig. 6, the performance of the three systems degenerates as more noises are introduced. Notice that T_{gnn} is more sensitive to the noises, with a relative performance loss of 12.1% and 19.5% at the disturbance level of 2 km and 5 km, respectively. The visual modality improves the robustness, both in the early stage for building the tag feature $T_{vnn\cup gnn}$ and in the late stage for building the classification system $Visual + T_{vnn\cup gnn}$. The results show that our system is tolerable to a geo tagging error of 1 km.

Fig. 7 shows the influence of extra geo noises (with 20 km deviation) on individual concepts. The behavior of the concepts is largely consistent with what we have observed in Fig. 5, i.e., concepts that benefit more from the geo information are more sensitive to the geo noises, see 'tiger', 'airplane', and 'protest', for instance. Interestingly, we observe several exceptions such as 'horse', 'glacier', and 'whale'. These concepts are more likely to be observed in much wide areas, so a coarse localization already helps. The influence of geo noises varies over concepts.

D. Experiment 4. The Effect of Data Scale

To study the influence of data scale on building geo-aware tag features, we consider $\mathcal{S}_{geo}$ of varied scales, let it be $\{0.2, 0.6, 1, 5, 10\}$ millions, respectively. Similar to the previous experiment, we experiment with T_{gnn} , $T_{vnn\cup gnn}$, and $Visual + T_{vnn\cup gnn}$.

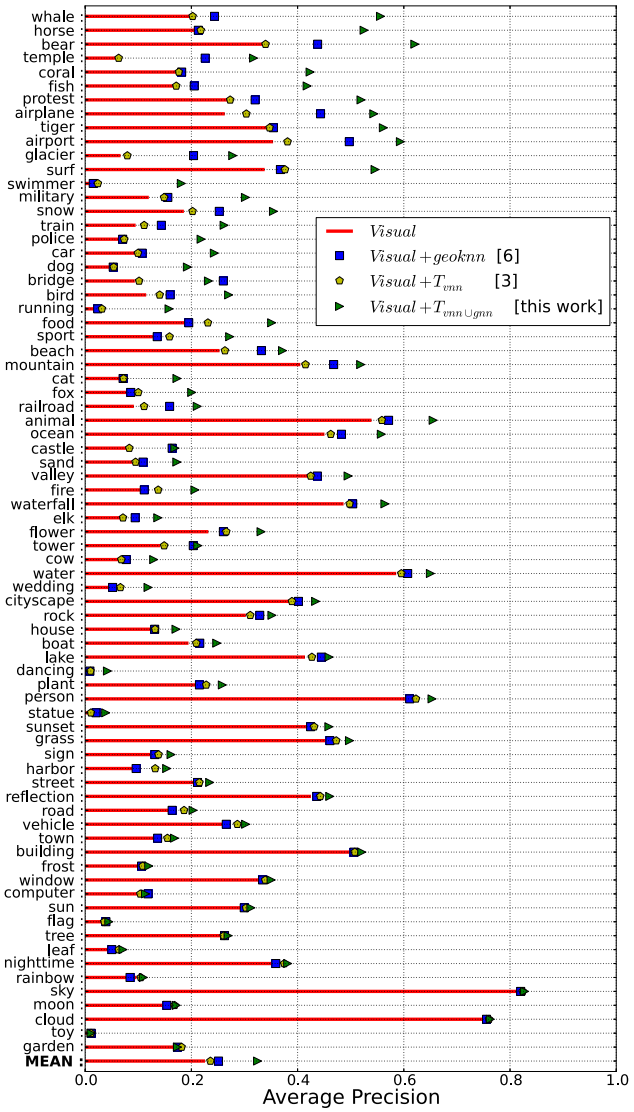


Fig. 5. **Per-concept comparison**. The concepts are sorted in descending order according to the absolute improvement of $Visual + T_{vnn \cup gnn}$ over $Visual + T_{vnn}$.

As shown in Table IV, the performance of the three systems improves as $\mathcal{S}_{geo}$ grows. As the data scale increases, more geo neighbors can be retrieved in a small region, making the geo context better modeled. As a consequence, we observe a relatively rapid improvement of T_{gnn}. Still, the joint use of the geo and visual neighbors is important. When 10M geo-tagged images are used, $Visual + T_{vnn \cup gnn}$ produces the highest mAP of 0.347.

E. Examples

For a more intuitive understanding of the geo-aware tag feature, we visualize in Fig. 8 the accumulated difference between $T_{vnn \cup gnn}(x)$ and $T_{vnn}(x)$ of all the positive examples of a specific concept. For 'airplane', related tags such as 'airport', 'airplane', and 'boeing' are enhanced in $T_{vnn \cup gnn}(x)$. As the tag features are l_1 normalized, the increase of a specific tag in the retrieved neighbors will decrease the values of other tags. A tag gets a negative

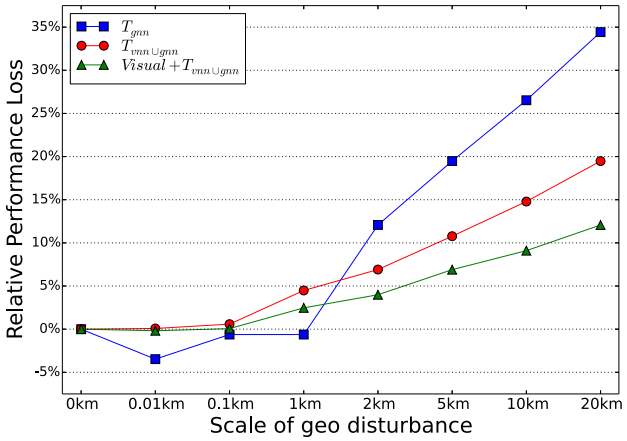


Fig. 6. **Experiment 3. Robust Analysis with Extra Geo Noises**. The different scales of geo disturbance is manually added by randomly choosing a direction and deviating the original location at in a certain distance. Fusion makes the geo-aware tag feature more robust with respect to geo tagging noises.

TABLE IV
EXPERIMENT 4. THE EFFECT OF DATA SCALE ON GEO-AWARE TAG FEATURES. THE PERFORMANCE IMPROVES AS THE DATA SCALE INCREASES.

	$\lvert\mathcal{S}_{geo}\rvert$				
System	**0.2 m**	**0.6 m**	**1 m**	**5 m**	**10 m**
T_{gnn}	0.109	0.129	0.138	0.175	0.183
$T_{vnn \cup gnn}$	0.254	0.261	0.271	0.290	0.293
$Visual + T_{vnn \cup gnn}$	0.311	0.321	0.325	0.343	0.347

difference score means its value in the geo-aware feature is smaller than its counterpart in T_{vnn}, and consequently this dimension becomes less important for classification. Similar results are observed in Fig. 8(b) and Fig. 8(c), where tags such as 'alaska' and 'zoo' are strengthened for 'bear', and 'india' for 'temple'. Moreover, unwanted tags propagated from semantically irrelevant visual neighbors such as 'bird' and 'dog' are suppressed for concept 'airplane' and 'bear', respectively. These qualitative results again show the effectiveness of the proposed geo-aware tag feature for representing images at a higher level.

VI. SUMMARY AND CONCLUSIONS

For classifying the increasing amounts of geo-tagged images, this paper proposes geo-aware tag features. The features are extracted by exploiting millions of online images associated with both user tags and geo tags. Different from the traditional tag feature obtained by tag propagation from visual neighbors, the new tag feature is built by tag propagation from the union of visual and geo neighbors. This small revision, however, leads to powerful representations for geo-aware image classification. Classification experiments on the popular NUS-WIDE set support the following conclusions.

1) Using geo neighbors alone is problematic due to their sparseness. The joint use of visual and geo neighbors is

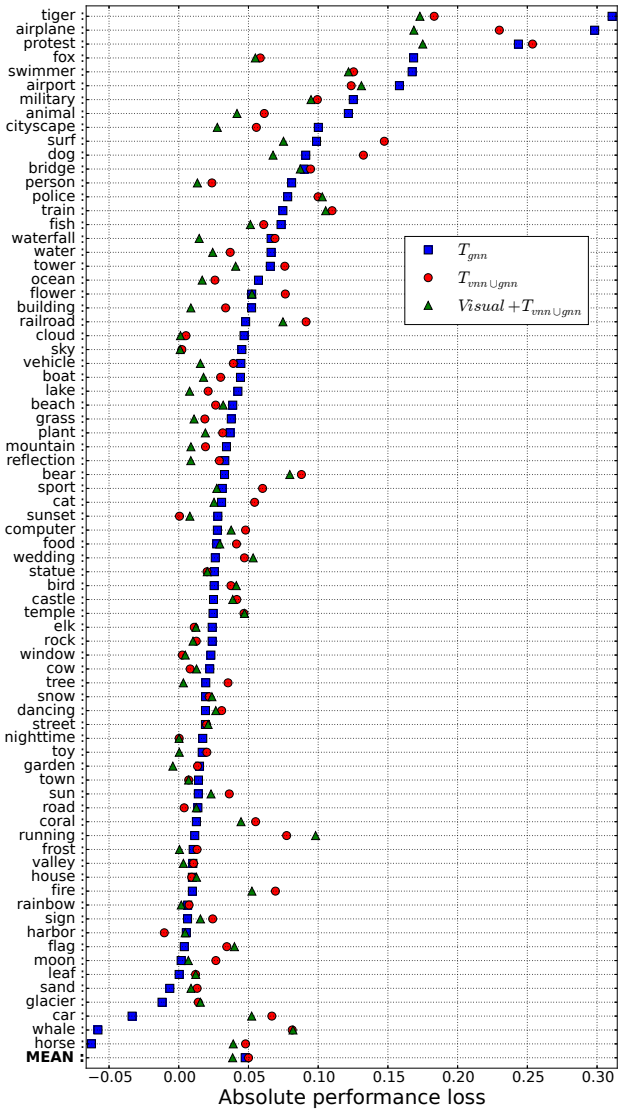


Fig. 7. **The influence of extra geo noises on individual concepts**. The scale of geo disturbance is 20 km. Concepts are sorted in descending order according to the absolute performance loss of T_{gnn}.

a must, resulting in a tag feature with mAP of 0.271, outperforming the traditional tag feature (mAP of 0.159) and the standard BovW feature (mAP of 0.226).

2) The tag features can be further improved by two means, either by weighting the geo neighbors in terms of their geo distance or by using a larger source set for tag propagation. These means give a relative gain of approximately 8%.

3) The tag features also work well when used in combination with the BovW feature, scoring mAP of 0.347. The system is quite robust as it can maintain its performance even though the geo tags deviate one kilometer from their original coordinates.

To conclude, the proposed tag features are a simple and good representation for geo-aware image classification.

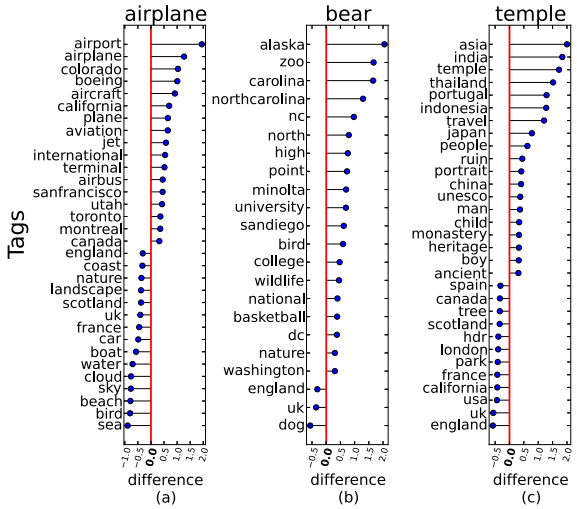


Fig. 8. **Visualizing the difference between the two tag features $T_{vnn\cup gnn}(x)$ and $T_{vnn}(x)$**. Falling on the right-hand side of the vertical red line means a tag is enhanced in $T_{vnn\cup gnn}(x)$, while falling on the left-hand side of the line means the tag is suppressed in $T_{vnn\cup gnn}(x)$. For a better view, only tags having the largest values are shown.

REFERENCES

[1] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: context and content in community-contributed media collections," in *Proc. of ACM MM*, 2007, pp. 631–640.

[2] P. Serdyukov, V. Murdock, and R. Van Zwol, "Placing Flickr photos on a map," in *Proc. of SIGIR*, 2009, pp. 484–491.

[3] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. of CVPR*, 2008, pp. 1–8.

[4] J. Li, X. Qian, 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.

[5] S. Jiang, X. Qian, Y. Xue, F. Li, and X. Hou, "Generating representative images for landmark by discovering high frequency shooting locations from community-contributed photos," in *Proc. of ICME*, 2013, pp. 1–6.

[6] E. Moxley, J. Kleban, and B.S Manjunath, "SpiritTagger: a geo-aware tag suggestion tool mined from Flickr," in *Proc. of MIR*, 2008, pp. 24–30.

[7] D. Joshi and J. Luo, "Inferring generic activities and events from image content and bags of geo-tags," in *Proc. of CIVR*, 2008, pp. 37–46.

[8] J. Kleban, E. Moxley, J. Xu, and B. Manjunath, "Global annotation on georeferenced phototgraphs," in *Proc. of CIVR*, 2009, p. 12.

[9] L. Cao, J. Yu, J. Luo, and T. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in *Proc. of ACM MM*, 2008, pp. 1–8.

[10] K. Yaegashi and K. Yanai, "Geotagged image recognition by combining three different kinds of geolocation features," in *Proc. of ACCV*, 2010, pp. 360–373.

[11] X. Li, C. Snoek, M. Worring, and A. Smeulders, "Fusing concept detection and geo context for visual search," in *Proc. of ICMR*, 2012, p. 4.

[12] X. Qian, X. Liu, C. Zheng, Y. Du, and X. Hou, "Tagging photos using users' vocabularies," *Neurocomputing*, vol. 111, pp. 144–153, Jul. 2013.

[13] J. Liu, Z. Li, J. Tang, Y. Jiang, and H. Lu, "Personalized geo-specific tag recommendation for photos on social websites.," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 588–600, Apr. 2014.

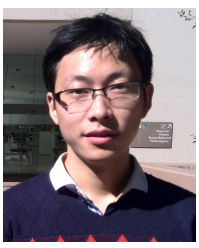
[14] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE Trans. Knowl. and Data Eng.*, vol. 26, no. 7, pp. 1763–1777, Jul. 2014.

[15] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 5, Apr. 2008.

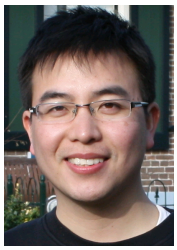
[16] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. Snoek, and A. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval," *arXiv:1503.08248*, 2015.

[17] Y. Yang, Y. Yang, and H. Shen, "Effective transfer tagging from

image to video," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 2, pp. 14:1–14:20, May 2013.
[18] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[19] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *Proc. of CVPR*, 2009, pp. 1367–1374.
[20] D. Zielstra and H. Hochmair, "Positional accuracy analysis of flickr and panoramio images for selected world regions," *Journal of Spatial Science*, vol. 58, no. 2, pp. 251–273, Jun. 2013.
[21] L. Cao, J. Luo, H. Kautz, and T. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 208–219, Feb. 2009.
[22] S. Liao, X. Li, X. Wang, and X. Du, "Building geo-aware tag features for image classification," in *Proc. of ICME*, 2014, pp. 1–6.
[23] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.
[24] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 88–101, Feb. 2012.
[25] Q. Fang, J. Sang, and C. Xu, "GIANT: Geo-informative attributes for location recognition and exploration," in *Proc. of ACM MM*, 2013, pp. 13–22.
[26] Sergej Sizov, "Geofolk: latent spatial semantics in web 2.0 social media," in *Proc. of WSDM*, 2010, pp. 281–290.
[27] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Exploiting web images for robust semantic video indexing via sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.
[28] H. Zhang, M. Korayem, E. You, and D. Crandall, "Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities," in *Proc. of WSDM*, 2012, pp. 33–42.
[29] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Processing*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.
[30] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. of CVPR*, 2010, pp. 902–909.
[31] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.
[32] X. Li, S. Liao, B. Liu, G. Yang, Q. Jin, J. Xu, and X. Du, "Renmin University of China at ImageCLEF 2013 scalable concept image annotation," in *CLEF Working Notes*, 2013.
[33] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. of CVPR*, 2008, pp. 1–8.
[34] C. Hauff, B. Thomee, and M. Trevisiol, "Working notes for the placing task at MediaEval 2013," in *MediaEval*, 2013.
[35] X-J Wang, Lei Zhang, Ming Liu, Yi Li, and Wei-Ying Ma, "Arista-image search to annotation on billions of web photos," in *Proc. of CVPR*, 2010, pp. 2987–2994.
[36] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proc. of CIVR*, 2009, p. 48.
[37] X. Li, C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "Bootstrapping visual categorization with relevant negatives," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 933–945, Jun. 2013.

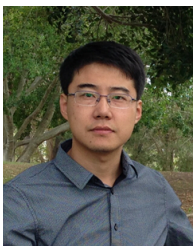


Shuai Liao received B.E. degree in Software Engineering from University of Electronic Science and Technology of China in 2012. He is currently a master student in the School of Information, Renmin University of China. His research interest is multimedia retrieval and computer vision.



Xirong Li received the B.S. and M.E. degrees from the Tsinghua University in 2005 and 2007, respectively, and the Ph.D. degree from the University of Amsterdam in 2012, all in computer science. He is currently an Assistant Professor in the Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China. His research focuses on image and video retrieval. He has been awarded the ACM SIGMM Best PhD Thesis Award 2013, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award 2012, the Best Paper Award of the ACM CIVR 2010, and PCM 2014 Outstanding Reviewer Award.

Heng Tao Shen obtained his BSc with 1st class Honours and PhD from Department of Computer Science, National University of Singapore in 2000 and 2004 respectively. He is currently a Professor of Computer Science and an ARC Future Fellow in the School of Information Technology and Electrical Engineering at the University of Queensland. His research interests mainly include Multimedia/Mobile/Web Search, Multimedia Content Analysis, and Big Data Management on spatial, temporal, multimedia and social media databases. Heng Tao has extensively published and served on program committees in most prestigious international publication venues of interests. He received the Chris Wallace Award for outstanding Research Contribution in 2010 conferred by Computing Research and Education Association, Australasia. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering and a TPC Co-Chair for ACM Multimedia 2015. He holds a position of Visiting Professor in Nagoya University and National University of Singapore.



Yang Yang received the B.S. degree from Jilin University in 2006, the M.E. degree from Peking University in 2009, and the Ph.D. degree from the University of Queensland in 2013. He was a Research Fellow with the School of Computing, National University of Singapore. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia information retrieval, social media analysis, and machine learning.

Xiaoyong Du obtained his B.S. degree from Hangzhou University in 1983, the M.E. degree from Renmin University of China in 1988, and PhD from Nagoya Institute of Technology, Japan in 1997. He is currrently a Professor in the School of Information, Renmin University of China. His current research interests include databases and intelligent information retrieval. He has published more than 100 papers on related areas.