BUILDING GEO-AWARE TAG FEATURES FOR IMAGE CLASSIFICATION

Shuai Liao, Xirong Li, Xiaoxu Wang, Xiaoyong Du*

Multimedia Computing Lab, Renmin University of China, Beijing 100872, China
Key Lab of Data Engineering and Knowledge Engineering, Beijing 100872, China
State Key Lab of Software Development Environment, Beijing 100191, China
{leoshine,xirong,wangxiaoxuinfo,duyong}@ruc.edu.cn

ABSTRACT

Given the proliferation of geo-tagged images, geo-aware image classification is an emerging topic. To derive a better image representation, tag features which represents an image as a histogram of tags are recently introduced. However, it is unclear whether geo tags can improve the tag features. To resolve the uncertainty, this paper studies geo-aware tag features. Our work is based on previous work which builds tag features by propagating tags from visual neighbors retrieved from many user-tagged images. What is different is that we build tag features by tag propagation from the union of visual and geo neighbors. This simple modification makes the new tag feature both content-aware and geo-aware. Using 1M Flickr images as a source set to construct the tag feature, experiments on the public NUS-WIDE set justify our proposal. The geo-aware tag feature outperforms the previous tag feature and a standard bag of visual words feature. Our geo-aware image classification system beats a recent alternative. For its simplicity and effectiveness, we consider the proposed tag feature promising for geo-aware image classification.

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

1. INTRODUCTION

Billions of images are shared on social media platforms such as Facebook and Flickr. So the quest for automated image classification is naturally on. To build an image classification system, features that can effectively represent the visual content are essential. Traditionally, features refer to low-level features directly extracted from pixels [1]. Among them, the bag of visual words feature obtained by quantizing SIFT like local descriptors has been a *de facto* choice [2]. Nevertheless, due to the semantic gap, low-level features remain a limited representation for categorizing images.

*Corresponding author. This work was partially supported by NSFC (61303184), State Key Laboratory of Software Development Environment Open Fund (SKLSDE-2012KF-09), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (14XNLQ01), SRFDP (20130004120006), and SRF for ROCS, SEM.

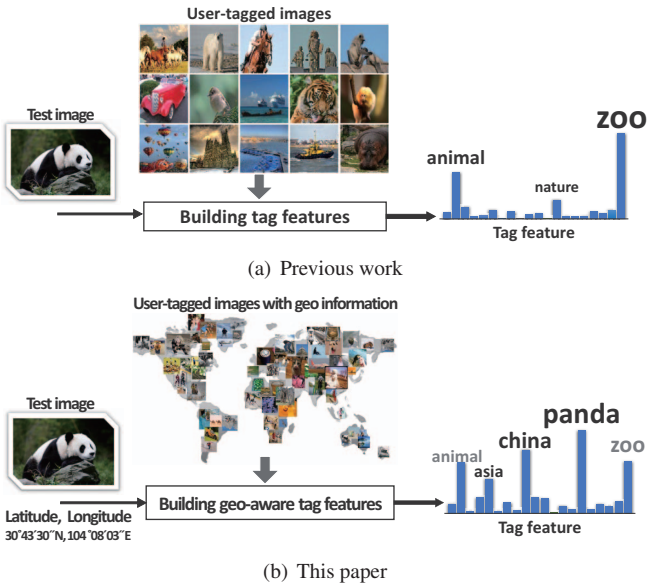

Fig. 1. **Exploiting many user-tagged images for building tag features for image classification**. Given the increasing availability of geo tags, this paper improves previous work by developing a novel geo-aware tag feature.

Towards representing images at a higher level, a novel tag feature is proposed [3]. 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. As exemplified in Fig. 1(a), the tag feature for an image of 'panda' contains salient dimensions such as 'zoo', 'animal', and 'nature'. Since tags are more interpretable than visual words, the tag feature describes images at a higher level compared to the bag of visual words feature.

Intuitively, knowing where an image was photographed may reduce the uncertainty in interpreting the visual content, and consequently improve image classification. Comparing two visual concepts 'dog' and 'panda', for instance, while the former can be seen almost everywhere, the latter can only be observed in specific zoos by common users. Thanks to

the wide use of GPS-enabled camera phones and online geo-tagging services, geo tags as latitude and longitude become part of image metadata. Moreover, research on automatic geo-tagging is making a good progress [4]. All these make geo tags increasingly available. Joshi and Luo were among the first to build tag features for geo-tagged images [5]. By querying a geographical information system database, they find place entities that are closest to the location of a geo-tagged image, and merge textual descriptions of these places to form the tag feature. Notice that in the above process the visual content is not taken into account, meaning two distinct images would have similar tag features as long as they are geographically close. As reported in Li *et al.* [6], this tag feature is not comparable to bag of visual words for image classification.

Given the power of tag features for representing an image at a higher level and the potential of geo tags for revealing the image's geo context, we are curious to know *whether geo tags improve tag features*. A desirable tag feature shall be both content-aware and geo-aware. To that end, we propose in this paper the joint use of visual and geo clues for building tag features. To the best of our knowledge, such a study has not been done before. As exemplified in Fig. 1(b), compared to the tag feature built using the visual clue alone, the new tag feature is more discriminative.

2. RELATED WORK

This paper is to develop a semantic representation of geo-tagged images for image classification, so it is in connection with work on building semantic features and work on utilizing geo tags for image classification. In what follows, we shortly review some representative papers in these two directions.

2.1. 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 [7]. The authors construct a semantic space, wherein each dimension corresponds to a specific concept which is associated with a pre-trained visual 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 [8]. Different from [7, 8], Wang *et al.* take a nonparametric approach [3], building a tag feature by collecting tags from user-tagged images which are visual neighbors of the given image. In contrast to the fixed concepts, user tags provide a more lively description of images. Furthermore, constructing the tag feature is lightweight as it requires no classifier training. Thus, we opt to develop our solution on the basis of [3].

2.2. 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.* [9], 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. Focusing on personal photo albums within the same context in terms of location and date, Cao *et al.* [10] 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 [6]. Qian *et al.* [11] exploit visual, geo, and temporal neighbors for personalized image tagging, without considering tag features. In contrast to [6, 9–11], Joshi and Luo introduce a tag feature by an inverse geo-encoding using a geographical information system [5]. The system contains over 8 million place names all over the world with manually edited descriptions. Given a test image, the authors use its geo tags to localize places nearby, and then combine the descriptions of these places to form the tag feature. As discussed in Section 1, this tag feature fully ignores the visual content. By contrast, we will build a tag feature that is both content-aware and geo-aware.

3. OUR APPROACH

By exploiting many user-tagged images and associated geographical information, we aim to build a geo-aware tag feature for better image classification. For the ease of consistent description, we introduce the following 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 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 classifiers upon the tag feature. In order to construct a geo-aware version of $T(x)$, we need many images which contain both user tags and geo tags. We term such a set of images as a source set $\mathcal{S}$.

Next, we present our approach to the geo-aware tag feature in Section 3.1, and its use for image classification in Section 3.2.

3.1. Building tag features by tag propagation

As we have noted in Section 1, for a given image, previous work constructs a tag feature by propagating tags from the image's visual neighbors [3], without taking the geo information into account. So a straightforward alternative is to replace the visual neighbors by the geo neighbors. However, for a considerable amount of images, the sparseness of their geo

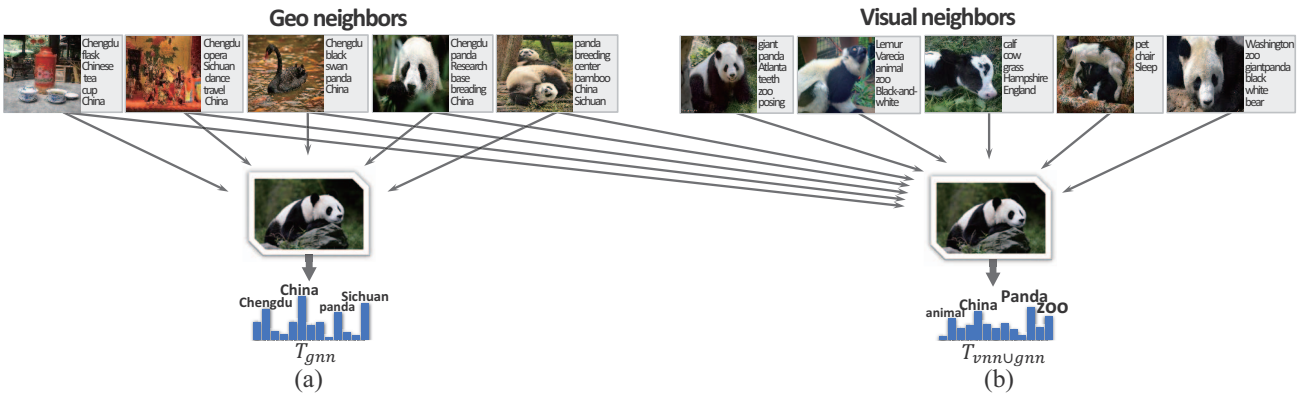

Fig. 2. **A conceptual illustration of the two proposed strategies for building geo-aware tag features**: (a) Tag propagation from geo neighbors, and (b) Tag propagation from both visual neighbors and geo neighbors.

neighbors may result in suboptimal tag features. In that regard, we investigate the following two strategies for building tag features, that is, 1) tag propagation from geo neighbors, and 2) tag propagation from visual/geo joint neighbors.

Strategy 1: Tag propagation from geo neighbors. Given an image x, we first retrieve its k nearest geo neighbors from the source set $\mathcal{S}$ 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 Fig. 2(a), 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.

Strategy 2: Tag propagation from visual/geo neighbors. Different from the strategy I, we retrieve from $\mathcal{S}$ both the k geo neighbors and another set of k neighbors measured in terms of a visual distance. We use a standard bag of visual words feature for computing the visual distance. As shown in Fig. 2(b), we extract the tag feature in a manner similar to the first strategy, but from the union of visual and geo neighbors. We denote this new variant of the tag feature as $T_{vnn\cup gnn}(x)$.

3.2. Combining textual and visual classifiers

Given the two tag features, namely T_{gnn} and $T_{vnn\cup gnn}$, 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.* [12] 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.

The textual and visual classifiers, working in distinct feature spaces, are compelmentary to each other to some extent. So we combine them to form the final classification system. On combining heterogeneous classifiers, a recent study by Li *et al.* [6] shows that linear combination with the combination weights optimized by coordinate ascent is 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 follow this good practice to combine the textual and visual classifiers.

4. EXPERIMENTAL SETUP

4.1. Data sets

Source set. As an instantiation of the source set $\mathcal{S}$, we use one million Flickr images collected by [6]. All the images are geo-tagged, associated with latitude and longitude. Covering photos taken in over 100 countries by 145K distinct users, the source set is quite diverse.

Ground-truth data. We use the geo-tagged part of the popular NUS-WIDE set [13], 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 'whale', 'valley', and 'protest'.

4.2. Experiments

In order to answer the question of whether geo tags improve tag features, we conduct the following two experiments.

Experiment 1: Do geo tags help when using the tag feature alone? In this experiment, we build image classifiers using the tag features alone. Following the notation in Section 3, we use T_{gnn} and $T_{vnn\cup gnn}$ to indicate the classifiers built on the two geo-aware tag features, respectively. In a similar style, we use T_{vnn} to indicate the classifier built on the previous tag feature [3]. A comparison between T_{vnn}, T_{gnn}, and

$T_{vnn\cup gnn}$ will show which tag feature is the best, and consequently allow us to justify if geo tags help for classification only using the tag feature.

For fair comparisons, whenever applicable we make sure that the three tag features share the same input and parameters. As an instantiation of $\mathcal{V}$, we construct a common tag vocabulary consisting of the top 2,000 most frequent tags in the source set, but with standard stop words, camera brands, digital numbers excluded beforehand. So the dimensionality of the tag features m is 2,000. We empirically set the number of visual/geo neighbors k to 150. For each concept and for each tag feature, we train a FikSVMs classifier using the software of [14], obtaining $75 \times 3 = 225$ classifiers in total.

Experiment 2: Do geo tags help when combining the texutal and visual features? As mentioned in Section 3.2, the combination of image classifiers built on the tag and visual features yields the best performance. In this experiment, we investigate whether geo tags also help in such a combination scenario. To that end, we build visual classifiers based on bag of visual words features. For each image, we employ the color descriptor software [2] to extract many SIFT descriptors by dense sampling. With the descriptors quantized by a precomputed codebook of 1,204 visual words, the image is represented by a 1,024-dim bag of visual words feature. Note that we also use this feature and the l_1 distance to find visual neighbors for constructing the tag features.

In experiment 2, we implement the following systems:
1) $Visual$: Using the bag of visual words feature,
2) $Visual$ + T_{vnn}: Combining the visual classifier and the textual classifier built upon T_{vnn} [3],
3) $Visual + geoknn$: Combining the visual classifier and a geo KNN classifier [6],
4) *GeoVisualKNN*: A KNN classifier exploits both visual and geo neighbors[1] [11],
5) $Visual + T_{vnn\cup gnn}$ (this work): Combining the visual classifier and the textual classifier built upon $T_{vnn\cup gnn}$.

Note that we let the three combined systems use the same visual classifiers, and the weights optimized by the same technology on the validation set. This setting allows us to properly verify the effectiveness of the proposed geo-aware tag feature.

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

5. RESULTS

5.1. Experiment 1: Do geo tags help when using the tag feature alone?

As shown in Table 1, the proposed geo-aware tag feature $T_{vnn\cup gnn}$ is the best, scoring an mAP of 0.271. In contrast, the standard tag feature T_{vnn} has a lower mAP of 0.159.

[1]Unlike the other systems, [11] does not utilize the training set of NUS-WIDE.

Table 1. **Feature-level comparison**.

Feature	Method	mAP
T_{vnn}	Tag propagation from visual neighbors [3]	0.159
T_{gnn}	Tag propagation from geo neighbors	0.138
$T_{vnn\cup gnn}$	Tag propagation from visual/geo neighbors	**0.271**

The result also shows that building tag features using geo neighbors alone is not good: T_{gnn} with an mAP of 0.138 is worse than T_{vnn}. The effectiveness of T_{gnn} mainly depends on the quality of the 150 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}. Therefore, the joint use of both visual and geo neighbors is essential for building a good tag feature.

One might argue that the superior performance of $T_{vnn\cup gnn}$ is due 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 for a more comprehensive comparison, 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.

On the base of the above results, we conclude that geo tags are helpful for building an effective tag feature.

5.2. Experiment 2: Do geo tags help when combining the texutal and visual features?

As shown in the Table 2, the system $Visual + T_{vnn\cup gnn}$ which combines visual classifiers and textual classifiers built upon $T_{vnn\cup gnn}$ is the top performer, reaching the highest mAP of 0.325. The experimental result also confirms some findings from previous studies. That is, i) exploiting geo tags ($Visual + geoknn$) improves image classification ($Visual$) [6], and ii) adding tag features ($Visual + T_{vnn}$) improves image classification ($Visual$) [3]. As the systems with SVMs classifiers leverage the NUS-WIDE training set, they outperform the *GeoVisualKNN* system [11] which directly use neighbors retrieved from the source set. Furthermore, we find that the proposed geo-ware tag feature $T_{vnn\cup gnn}$ surpasses the bag of visual words feature $Visual$ (0.271 versus 0.226).

As shown 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

Table 2. **System-level comparison**.

System	mAP
GeoVisualKNN [11]	0.113
$Visual$	0.226
$Visual + geoknn$ [6]	0.251
$Visual + T_{vnn}$ [3]	0.236
$Visual + T_{vnn \cup gnn}$ (this work)	**0.325**

the top five ranked concepts are 'whale', 'horse', 'bear', 'temple', and 'coral', while the bottom five ranked 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.

5.3. Examples

For a more intuitive understanding of the geo-aware tag feature, we visualize in Fig. 6 the accumulated difference between the two tag features $T_{vnn \cup gnn}(x)$ and $T_{vnn}(x)$ of all the positive examples of a specific concept. For a better view, we only show tags that correspond to the largest difference values. As shown in Fig. 6(a), for concept '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 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. 6(b) and Fig. 6(c), where tags such as 'alaska' and 'zoo' are strengthened for concept 'bear', and 'india' for concept '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.

6. CONCLUSIONS

For classifying the ever-growing amounts of geo-tagged images, we proposed in this paper a *geo-aware* tag feature. This feature is extracted by exploiting many online images associated with both user tags and geo tags. Different from the existing 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 a powerful feature for geo-aware image classification. Using one million Flickr images as a source set, two experiments on the public NUS-WIDE data support the following conclusions. While geo tags are helpful for building tag features, we find that using geo neighbors alone is problematic due to their sparseness. Propagating tags from both visual and geo neighbors yields the best tag feature. With an mAP of 0.271, the proposed tag feature outperforms the existing tag feature (mAP of 0.159) and a standard bag of visual words feature (mAP of 0.226). The tag feature also works well when used in combination with bag of visual words. With an mAP of 0.325, the combined system beats its counterpart that uses the previous tag feature (mAP of 0.236) and a recent geo-aware image classification system (mAP of 0.251). The geo-aware tag feature is simple and effective, making it attractive for classifying geo-tagged images.

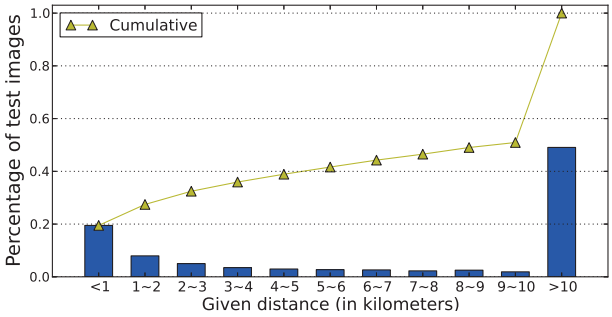

Fig. 3. **Percentage of the test images whose 150 geo neighbors can be fully retrieved within a given distance**. For more than 50% of the test images, their geo neighbors are not fully covered within a radius of 10 kilometers.

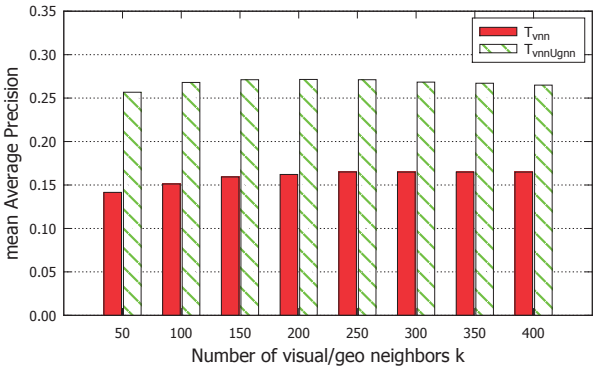

Fig. 4. **The influence of the number of neighbors k on the effectiveness of the tag features**. While both T_{vnn} and $T_{vnn \cup gnn}$ are robust with respect to the choice of k, $T_{vnn \cup gnn}$ outperforms T_{vnn} under different settings of k.

7. REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Survey*, 2008.

[2] K. van de Sande, T. Gevers, and C. Snoek, "Evaluat-

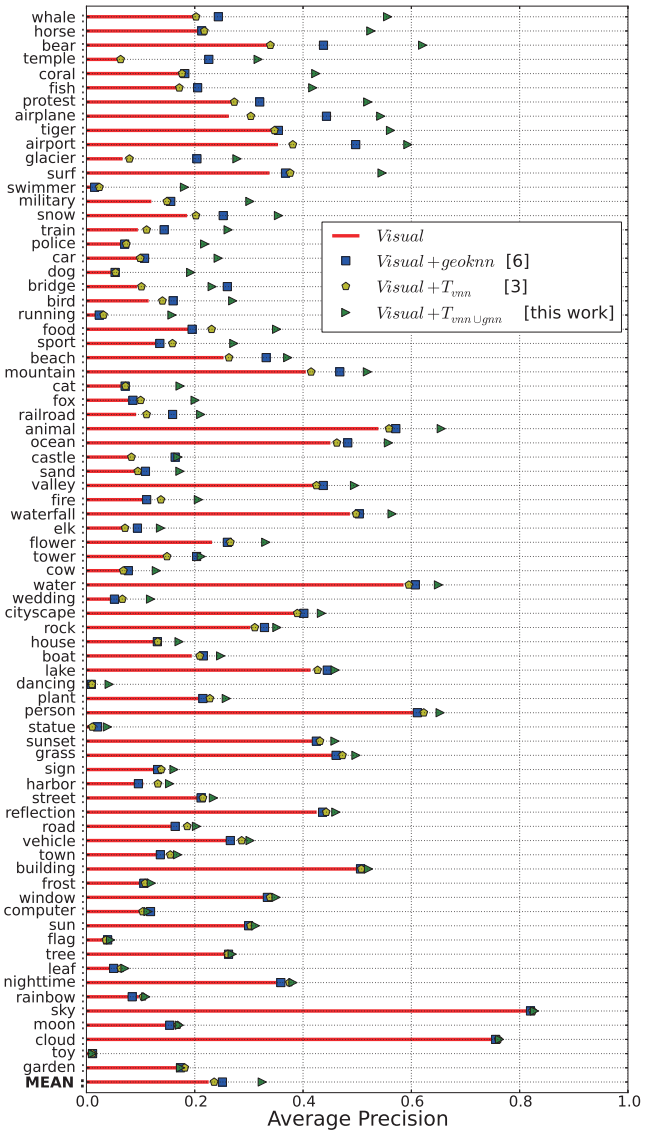

Fig. 5. **Per-concept comparison**. The concepts are sorted in descending order according to the absolue improvement of $Visual + T_{vnn \cup gnn}$ over $Visual + T_{vnn}$.

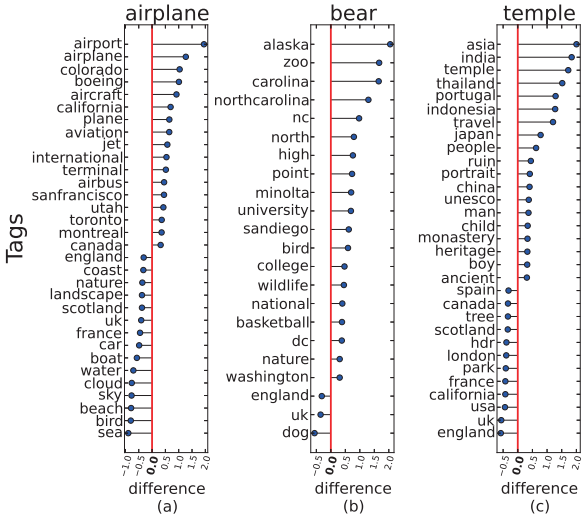

Fig. 6. **Visualizing the difference between the two tag features $T_{vnn \cup gnn}(x)$ and $T_{vnn}(x)$.** Falling on the right-hand side of the vertical red line means a tag is enhanced in $T_{vnn \cup gnn}(x)$, while falling on the left-hand side of the line means the tag is suppressed in $T_{vnn \cup gnn}(x)$.

ing color descriptors for object and scene recognition," *TPAMI*, 2010.

[3] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *CVPR*, 2009.

[4] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *CVPR*, 2008.

[5] D. Joshi and J. Luo, "Inferring generic activities and events from image content and bags of geo-tags," in *CIVR*, 2008.

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

[7] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *TMM*, 2007.

[8] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *TMM*, 2012.

[9] E. Moxley, J. Kleban, and B.S Manjunath, "SpiritTagger: a geo-aware tag suggestion tool mined from flickr," in *MIR*, 2008.

[10] L. Cao, J. Luo, H. Kautz, and T. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," *TMM*, 2009.

[11] X. Qian, X. Liu, C. Zheng, Y. Du, and X. Hou, "Tagging photos using users' vocabularies," *Neurocomputing*, 2013.

[12] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*, 2008.

[13] 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 *CIVR*, 2009.

[14] X. Li, C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "Bootstrapping visual categorization with relevant negatives," *TMM*, 2013.