Social Image Tagging With Diverse Semantics

Xueming Qian, *Member, IEEE,* Xian-Sheng Hua, *Senior Member, IEEE,* Yuan Yan Tang, *Fellow, IEEE,* and Tao Mei, *Senior Member, IEEE*

***Abstract*—We have witnessed the popularity of image-sharing websites for sharing personal experiences through photos on the Web. These websites allow users describing the content of their uploaded images with a set of tags. Those user-annotated tags are often noisy and biased. Social image tagging aims at removing noisy tags and suggests new relevant tags. However, most existing tag enrichment approaches predominantly focus on tag relevance and overlook tag diversity problem. How to make the top-ranked tags covering a wide range of semantic is still an opening, yet challenging, issue. In this paper, we propose an approach to retag social images with diverse semantics. Both the relevance of a tag to image as well as its semantic compensations to the already determined tags are fused to determine the final tag list for a given image. Different from existing image tagging approaches, the top-ranked tags are not only highly relevant to the image but also have significant semantic compensations with each other. Experiments show the effectiveness of the proposed approach.**

***Index Terms*—Image tagging, semantic, social media, tag diversity, tag enrichment, tag relevance.**

I. Introduction

MULTIMEDIA retrieval is very challenging due to the well-known semantic gaps [66]–[68]. The semantic gaps are caused by the low-level features that are insufficient to express the high-level semantics for the multimedia content including image, music, and video [63]–[75]. Multimedia content tagging is an effective way to minimize the semantic gaps in multimedia retrieval. Shen *et al.* [65] proposed a music tagging approach by modeling music information with hierarchical structure and uncovering the relationship between tags and concepts. The approach combines both multimodal and temporal information in music feature extraction and high-level semantic concept modeling for effective annotation of music documents into a joint framework. It maps sound documents to a representation in latent musical concept space to get more accurate description for the relevance music documents and tags. Image and video tagging is recently very hot due to the popularity of social network and social media-sharing websites.

Fig. 1. Flickr image with raw tags labeled by user. Only the three tags *Bird*, *Kingfisher*, and *Alcedoatthis* are relevant to the appearance of the image. The other tags are irrelevant to the image. The other semantics of this image (e.g., *Fish* and *Water*) are not disclosed by the raw tags.

Social image sharing websites (e.g., Picasa and Flickr) allow users to give several tags to describe the content of their shared photos [1], [19], [39]. The user-annotated tags are very useful. They make the social images better accessible to the public [2], [70]. The quality of tags is important for improving the performance of the tag-based applications [3]–[7], [19], such as tag-based image retrieval. The performances of the tag-based applications are inevitably influenced by the tag quality [7], [66], [68]. For example, in the tag-based image retrieval, some highly relevant images are not at the top-ranked search results if the content-related tags are not in the user-annotated tag list.

Some of the user-annotated tags are relevant to image and some are irrelevant. User-annotated tags usually only uncover a small part of semantics for a social image. We cannot expect user-annotated tags covering all the semantics for their shared images, especially using a limited number of tags. Thus, automatic image retagging and processing are aiming at improving tag qualities from social user-annotated tags [1]–[3], [6]–[8], [18], [23], [28]–[40], [64]–[73]. Measuring the relevance of tag to the image and the semantics coverage of the top-ranked tags are valuable for tag-based application.

Fig. 1 illustrates a social user-shared image and the annotated tags in Flickr. We find that only three tags *Bird*, *Kingfisher*, and *Alcedoatthis* are relevant to the image. The tags

Manuscript received September 2, 2012; revised September 16, 2013; accepted September 22, 2013. Date of publication March 19, 2014; date of current version November 13, 2014. This work was supported in part by the National Natural Science Foundation of China under Projects 60903121 and 61173109, and in part by Microsoft Research Asia and the Foundations of Macau University under Grant SRG010-FST11-TYY, Grant MYRG187(Y1-L3)-FST11-TYY, and Grant MYRG205 (Y1-L4)-FST11-TYY. This work was performed in part when the first author was visiting the University of Macau, Macau, China. This paper was recommended by Associate Editor L. Shao.

X. Qian is with the SMILES Laboratory, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

X.-S. Hua is with Microsoft (e-mail: xshua@microsoft.com).

Y. Y. Tang is with the Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yytang@umac.edu.cn).

T. Mei is with Microsoft Research Asia (e-mail: tmei@microsoft.com).

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

Digital Object Identifier 10.1109/TCYB.2014.2309593

Cannon and *40-D* are related to the photo taken by the digital camera. The tag *Dorset* is about the place the photo is taken at. The other tags are too subjective to provide semantics for the image and can be viewed as image content irrelevant tags. For the image, its content is disclosed well by the tags *Bird*, *Kingfisher*, *Water*, *Fish*, etc. However, the three relevant tags provided by the user (i.e., *Bird*, *Kingfisher*, and *Alcedoatthis*) only disclose one aspect of the main semantics *Bird* for this image. The other semantics of this image (e.g., *Water* and *Fish*) are not disclosed by the user-labeled tags. From this image, if the tags *Water* and *Fish* appear in the top-ranked tag list, then the semantics disclosed by the tag list is far wider than that by the social user-annotated tags. Thus, it is better that the top-ranked tags covering wide range of semantics should satisfy the following two conditions: 1) relevant to image and 2) large semantic compensations to each other.

How to make the top-ranked tags covering wide range of semantics is very important for social image content understanding and textual-based image retrieval. To our knowledge, this problem has somewhat overlooked by researchers in this area. The problem of tag diversity is not well addressed in the existing tagging approaches [1]–[3], [6]–[8], [18], [23], [28]–[40], [64]–[73]. However, how to make the top-ranked search results covering wide topics has been paid much attention for a long time in information retrieval [41]–[52] and image search [53]–[55]. For example, Goffman [43] ranked a document with respect to the documents appearing before it. Carbonell and Goldstein [42] proposed a maximal marginal relevance-based document ranking approach. This approach aims at maximizing the relevance while minimizing similarity of the document to the documents ranked ahead of it [42]. Zhai *et al.* [51] proposed a subtopic-based search approach. The goal of their approach is to return retrieval results that can cover wide range of subtopics [51], [52]. Now, we summarize image search with diversity. In image search, to make the search results diversified, the approaches can be classified into the following categories: 1) retrieval results clustering and representative image selection from each cluster [76]–[79]; 2) removing near duplicate images from the final ranked image list [44], [48], [50], [53], [54]; 3) textual searching with visual reranking to ensure result diversity [80]–[85]; 4) jointly optimizing the diversity and the relevance in a unified framework [55], [86]; and 5) improving return images with diversity by exploring geographical information from community-contributed photos [87]–[90].

If we can suggest the image content relevant and semantic compensative tags for social images, then we can know the main content of the image very well. This can be utilized in improving the diversity social image retrieval results by ranking the image according to the suggested tag list. Motivated by the search result diversity measurement approach in information retrieval, we propose a social image tagging with diverse semantics approach. Our algorithm can make the top-ranked tags are highly relevant to the image and cover wide range of semantics. First, we determine the relevance of each tag to image. Then we give the top-ranked tag of the image according to the relevance of the tag to the image. Finally, the ranks of the following tags are determined by their semantic compensations to the already ranked tag list. The main contributions of this paper are as follows: 1) we summarize the problem of relevance and diversity of a tag to image and give their measurement approaches; 2) we propose a social image tagging approach by fusing both the relevance and the diversity of a tag to the image into a unified framework; and 3) we propose an iterative greedy searching-based tagging approach to determine the optimal tag list.

The remainder of this paper is organized as follows. In Section II, related work on image tagging and image search with diversity is reviewed. In Section III, the diversity of a tag to the image is studied. In Section IV, the proposed image tagging approach with diverse semantics is given. In Section V, experimental results and discussions are given. Conclusions are drawn in Section VI.

II. Related Work

In this paper, we focus on image tagging with diverse semantic. In this section, we review the papers on image tagging in Section II-A and the image search with diversity in Section II-B, respectively.

A. Review of Image Tagging

Social image tagging approaches improve tag quality by ranking the initial tags according to their relevance to the image content, removing noisy tags from the initial tags and recommending some new relevant tags. Wang *et al.* [70] systematically reviewed the existing tagging-based applications. They categorize existing assistive tagging into three paradigms: tagging with data selection and organization, tag recommendation, and tag processing from the point of combining human's intelligence and computer's computation power. From [70], we find that image tagging can be carried out by utilizing visual information and textual information [5], [16]–[18], [29]–[40], [64]–[73]. Model-based and model-free approaches can be adopted to fuse multimodal features in tagging. The model-based approaches need to build models for each tag [8]–[14], [36], [38], [66], [68]. The model-free approaches predict relevant tags for an image by utilizing statistical properties of tags and the low-level visual features [5], [16], [20], [39], [40].

Wang *et al.* [36] utilized Gaussian mixture models to model each tag. The Gaussian mixture number is adaptively determined for each tag [36]. Xu *et al.* [30] utilized regularized latent Dirichlet allocation to model the tag similarity and tag relevance to refine tagging quality. Wu *et al.* [31] proposed a learning-based image tagging approach. The concepts that are harder to be predicated are modeled to improve tagging performances. Chen *et al.* [25] proposed a tag recommendation approach that predicts the possible tags using pretrained support vector machines predicators. Jin *et al.* [23] utilized WordNet to estimate the semantic correlation among the annotated tags. The weakly correlated tags are removed from the recommended tag list. Due to the fact that visual information is not utilized in tag enrichment, the recommended tags are the same for the images with the same initial tags [20].

Feng *et al.* [59] proposed a visual attention model-based tag ranking approach. The tags that correspond to the salient regions of an image are ranked ahead. The approach combines the extracted visual attention model and multiinstance learning and propagated to the salient region-based image annotation. However, this approach is subject to the salient region detection performances. How to select effective features for image annotation is also studied recently by exploring the sparsity of semantics and the low-level features [60]–[62]. Ma *et al.* [61] proposed a collaborative feature selection-based subspace sparsity representation approach for Web image annotation.

Some tagging approaches are carried out by finding visual near duplications for the input image from large-scale websites and propagating tags of the near duplicates to the image [5], [8], [16]–[18], [24]. Li *et al.* [8] proposed a neighbor voting algorithm for tag relevance learning. This approach is based on the fact that social users are likely to label visually similar images by the same tags. The relevance of a tag with respect to an image can be inferred from tags of visual similar neighbors [24]. Liu *et al.* [1] proposed an approach to rank and enrich tags using random walk model. The random walk model promotes the tags with many visual similar neighbors and weakens the tags with fewer neighbors. In [26] and [27], the tag refinement is carried out by utilizing the random walk with restart framework. The visual similarity (VS) and tag co-occurrence conditioned on visual similarities are utilized.

Moxley *et al.* [72] presented a Spirit-tagger tool that mines tags from geographical and visual information. These annotations are derived from image similarities constrained to a geographical radius, and a comparison of the local frequency of terms to their global frequency is used to weigh terms that occur frequently in a local area.

Liu *et al.* [64] fused the visual information, raw tags provided by social users, and the semantic correlations of the tags for image retagging. They modeled the image retagging process as a multiple graph-based multilabel learning problem [91]. Their algorithm propagates the information of each tag along a tag-specific graph and determines the tag-specific visual sub-vocabulary from a collection of social images with noisy tags [64].

Gao *et al.* [67] proposed a visual and textual joint relevance learning approach for tag-based social image search. Both visual and textual information are utilized to evaluate the relevance of social user-tagged images. They built a social image hypergraph with its vertices as images and hyperedges as visual or textural terms. The weights of hyperedges are updated throughout the learning process. Finally, the relevance is determined from the modulated hypergraph learning-based approach.

Yang *et al.* [69] proposed an inductive algorithm for image annotation by integrating label correlation and VS. They first construct a graph model according to image visual features to exploit the unlabeled images. Then a multilabel classifier is trained by integrating structure learning and graph-based transductive classification for providing image content-relevant labels during image annotation.

Li *et al.* [71] presented a solution to the annotation of specific products in videos by mining information from the Web. They used visual signatures to annotate video frames that are built on the bag-of-visual-words representation of the training data collected from Amazon and Google image search engine.

To enhance the descriptive ability of the existing tags and facilitate image retrieval, Yang *et al.* [73] proposed a tagging approach that aims at mining properties of tags such as shape, location, texture pattern, and color.

Jia *et al.* [2] fused the textual similarities of tags and visual similarities of images in a multigraph reinforcement framework to improve tag enrichment performances. Zhu *et al.* [29] took the low-rank, content consistency, tag correlation, and tag noise sparseness into account during tag enrichment. Zhou *et al.* [37] proposed a hybrid probabilistic model (HPM)-based image tagging approach. This approach integrates both low-level visual features and high-level textual features of an image when the image is with user-labeled initial tags. For the image without initial tags, they resort to the visual features to carry out image tagging. They track the sparse probability of the tag to image association matrix. A collaborative filtering approach is further utilized to improve tag enrichment performance. HPM jointly exploits both low-level visual features and the textual features of social user-provided tags in a unified probabilistic framework to recommend more content-relevant tags for the unlabeled image.

Qian *et al.* [38], [91] modeled all the tags by a fully connected graph. They view tag enrichment as a combinational optimization problem. Graph cut-based tag enrichment approach is proposed to determine the relevant tags. Moreover, in [40], they carried out tag filtering by using similar compatible principles. This approach determines the ranks of user-annotated tags by maximizing the compatible value of changing the labels of the tags from irrelevant to relevant at each step. Recently, we proposed [39] a social image tagging approach using users' own vocabularies. We recommend tags preferred by users to annotate their newly uploaded images by mining their tagging behavior from the history information in their social communities. The visual information, time taken by the image, and geographical location information of the photo are all contributive to annotate images with users own vocabularies.

The existing tag enrichment approaches are concentrated on improving tag qualities by enhancing the relevance of the top-ranked tags [1]–[3], [6]–[8], [18], [19], [23], [28]–[38], [40], [64]–[73]. In this paper, both the tag relevance and semantic diversity are taking into account for social image tagging. Our aim is to suggest image content relevant and semantic compensative tags at the top-ranked tag list.

B. Review of Image Search With Diversity

Usually, in image search, both the visual information of ranked image list and the textual descriptions are utilized to measure the relevance and diversity of the ranked images. Moreover, for social image search, the supplementary spatial information of images (such as the latitude and longitude) is also taken into account to make the final ranked images

with diversity. Now we give the corresponding review of the existing works on image search with diversity.

Due to the fact that the reliance on the textual information is associated with an image, existing image search engines lack the discriminative power to deliver visually diverse search results. The textual descriptions (including tags and comments) are important to find relevant images for a given query (in textual). But they provide little information about the rich visual content of images. In image retrieval, the final results are usually displayed in a ranked list. The ranks are related to the similarity of the images' metadata to the query.

Image clustering-based approaches are often utilized to group the ranked image list at first, and then the image search results can be diversified by selecting a representative image from each group [76]–[79], [92], [93]. For example, [76] used both image visual features and tags to carry out clustering while Cai *et al.* [77] proposed a hierarchical clustering-based approach to organize the searched images into different semantic clusters. Leuken *et al.* [78] deployed lightweight clustering techniques in combination with a dynamic weighting function of the visual features. Radu *et al.* [79] aimed to improve retrieval relevance by selecting a set of representative and diverse images from a candidate image set. To ensure representativeness, images are reranked according to the relevance of images to the query. At the same time, to ensure the diversity of the search results, the returned images are clustered and the best ranked images among the most representative in each cluster are retained. There are also other approaches that have taken full use of content-based image retrieval to ensure the diversified search result. At first, relevant images are obtained, then near-duplicate images are detected and removed from the final image list [44], [48], [50], [53], [54].

Weinberger *et al.* [80] presented a method for detecting the ambiguity of a query based on the textual features of the image dataset. If the query has ambiguity, then the ambiguity is reflected in the diversity of the top-ranked images. In [81], an adaptive model selection-based diversity of images search results is proposed by analyzing the topical diversity of image search results for textual features. Song *et al.* [82] proposed a reranking method based on topic richness analysis to enrich topic coverage in image retrieval results with diversity.

In [83], diversity of image search results is examined in the context of Web image search. The diversity of image search is achieved by image ranking and reranking through optimizing the diversity and the information richness. Zeigler *et al.* [84] made a balance between topic diversification and diversify personalized recommendation lists in final image ranking. Paramita *et al.* [85] fused the general diversity and the proposed spatial diversity into final image ranking.

Most approaches addressing diversity in information retrieval and image search consist of the following steps: 1) determining a set of potentially relevant images and 2) reranking the images to be diverse among the first step. In contrast to the general diverse approaches, Deselaers *et al.* [86] directly addressed the problem and jointly optimized the diversity and the relevance of the images in the retrieval ranking using techniques inspired by dynamic programming algorithms. The system is similar to [55]. The difference is that [86] use DP while [55] uses greedy search.

In contrast to the general diverse approaches as mentioned earlier, geographical information of community contributed photos is also made full use of to improve the results for queries with place names [87]–[90]. Rudinac *et al.* [87] used community contributed images to create representative and diverse visual summaries of specific geographic areas. They build a multimodal graph by fusing the relations between images, extracted visual features, text associated with the images, as well as users and their social network. The multimodal graph makes the final image list with diverse and representative. Popescu and Kanellos [88] presented a method for generating relevant and diversified visual summaries of places by using a geographical name. The geographic name is built from social community user-contributed data (such as Flickr and Panoramio) and its content reflects a community-based perception of places. In our previous papers, we also generate visual representative image by visual modeling images of landmarks in a generated topic album. Each topic album consists of images coming from the same place with various viewpoints. Based on viewpoint modeling, we can provide the ranked results with diverse viewpoints [89]. Moreover, we generate representative images for landmarks by incorporating the shooting locations of photos. This paper presents a representative images generation system by discovering high-frequency shooting locations from geo-tagged community-contributed photos [90].

III. Relevance and Diversity of Tag to Image

The diversity of a tag τ to image I cannot exist alone. It is a relative quantity. It relates to the relevance of the tag to the image and the semantic compensation of the tag to the tags ranked ahead of it. In this section, the relevance and diversity of tag to image are discussed, and their measurements are given in further sections.

A. Relevance of Tag to Image

Tag and image are two different media. Tag is in high-level semantic space (i.e., textual space). Image is in low-level visual feature space. The problem of measuring the relevance of a tag to image belongs to cross-media similarity measurement problem [74]. Intuitively, tag-to-tag similarity can be measured by textual similarity (TS) in semantic space. Image-to-image similarity is measured by the VS in low-level feature spaces. The relevance of a tag to image has been addressed in [1] and [58], respectively. To measure the relevance of a tag to image, we can resort to the following three approaches: 1) measuring the tag-to-image similarity in high-level semantic spaces; 2) measuring the tag-to-image similarity in low-level visual feature spaces; and 3) measuring the tag-to-image similarity in both high-level semantic spaces (i.e., semantic relevance) and low-level visual feature spaces (i.e., visual relevance). Now we give the measurements of visual relevance and semantic relevance in Measurement 1 and Measurement 2, respectively. A generalized relevance of a tag to image is given in Measurement 3 and

based on which the diversity of a tag to image is given in Measurement 4.

Measurement 1: The visual relevance of a tag τ to an image I is measured by the normalized VS of the image I to the images with content descriptive tags including the tag τ.

Let $VS(\tau, I)$ denote the VS of the tag τ to the image I. Similar to [1], the VS can be directly computed based on a Gaussian function with a radius parameterσ, that is

$$VS(\tau, I) = \frac{1}{|\Theta_\tau|} \sum_{x \in \Theta_\tau} \exp\left(-\|F_I - F_x\|^2 / \sigma^2\right) \tag{1}$$

where Θ_τ denotes the image set having the descriptive tag τ, the image number in Θ_τ is $|\Theta_\tau|$. σ^2 is the set to be the median value of all the pairwise Euclidean distances between images [1]. F_I and F_x are the visual features of the images I and x. $||*||^2$ is the l_2-norm of vector *. Measurement 1 actually maps the tag from textual space to low-level feature space by representing the tag with a set of images containing the tag. Then, the visual relevance of a tag to image can be measured by the image–image similarity in low-level visual feature space.

Measurement 2: The semantic relevance of a tag τ to the image I is measured by the normalized TS of tag τ to the content-related tags of the image.

The semantic relevance measurement approach actually converts the image from low-level visual feature space to textual space. Then, tag-to-image similarity can be measured using tag–tag similarity in textual space. For a social image I, we only have the user-labeled initial tags. Parts of user-annotated tags are highly relevant to image content [1], [22], [32], [38], [39], [64]. This can be shown by the initial tags of the exemplar images in Fig. 1 and Table I. Therefore, we utilize user-annotated tags as the image I-related tags to measure the semantic relevance of a tag to image. Thus, the semantic relevance of a tag τ to image I can be measured by the TS of tag τ to the user-annotated tag set $\varphi = \{\tau_1, \cdots, \tau_{|\varphi|}\}$. Let $SS(\tau, I)$ denote the semantic similarity of tag τ to image I, we can represent it as follows:

$$SS(\tau, I) = TS(\tau, \varphi) = \frac{1}{|\varphi|} \sum_{\tau_i \in \varphi} \exp\left(-d\left(\tau, \tau_i\right)\right) \tag{2}$$

where $d(\tau, \tau_i)$ denotes the textual (or semantic) distance of tags τ and τ_i. The valid initial tag number is $|\varphi|$. $TS(\tau, \varphi)$ denotes the normalized TS of tag τ to the initial tag set $\varphi = \{\tau_1, \cdots, \tau_{|\varphi|}\}$ of the image. Thus, in this paper, we utilize Google distance to measure the distance of two tags [57]. It is expressed as follows:

$$d(p, q) = \frac{\max(\log f(p), \log f(q)) - \log f(p, q)}{\log W - \min(\log f(p), \log f(q))} \tag{3}$$

where $f(p)$ and $f(q)$ are the numbers of images containing tag p and tag q on Flickr, respectively. $f(p, q)$ is the number of images containing both the tags p and q on Flickr. These numbers can be obtained by performing search by tag on Flickr website using the tags as query terms. W is the total number of images on Flickr.

Measurement 3: The relevance of a tag τ to the image I is related to both the visual relevance in low-level visual spaces and the semantic relevance in high-level textual spaces.

Let $RS(\tau, I)$ denote the relevance of the tag τ to the image I with initial tag set $\varphi = \{\tau_1, \cdots, \tau_{|\varphi|}\}$, in short, $r(\tau)$. It is expressed as follows:

$$\begin{aligned} r(\tau) = RS(\tau, I) &= \alpha SS(\tau, I) + (1 - \alpha) VS(\tau, I), \alpha \in [0, 1] \\ &= \alpha TS(\tau, \varphi) + (1 - \alpha) VS(\tau, I), \alpha \in [0, 1] \end{aligned} \tag{4}$$

where α is a ratio with its range in [0, 1]. The smaller α means the stronger the weight of visual relevance. $\alpha = 0$ means only VS is utilized in tag-to-image relevance measurement [i.e., (1)]. $\alpha = 1$ means only TS is utilized in tag relevance measurement [i.e., (2)]. Equation (4) is a general case of the tag-to-image relevance measurement approach. It fuses both the visual relevance and the semantic relevance in measuring the relevance of a tag to image. For social image without user-labeled initial tags, only the visual relevance is utilized. The impact of the parameter α on tag enrichment performance is discussed in experiments.

B. Diversity of Tag to Image

The diversity of a tag τ to the image I with already ranked tag set Γ is related to the following two aspects: 1) similarities of tag τ to the tags in the tag set Γ and 2) relevance of tag τ to the visual content of image I. If the semantic score of a tag τ to the tags in the tag set Γ is very high, then the improvement of semantic coverage of this tag over the tags in Γ is very limited. If the relevance score of a tag to the image is very low, then the diverse score of this tag to the image is also very small.

Assuming that tag *Kingfisher* is the first-ranked tag for the image as shown in Fig. 1, let us analyze the diversities of the following three tags: 1) *Fire*; 2) *Water*; and 3) *Bird* to the image. As the tag *Fire* is irrelevant to image content, its diverse semantic to image is zero ideally, even though it has large semantic compensation to the image content relevant tag *Kingfisher* in textual space. The tags *Water* and *Bird* are both highly relevant to image content. The semantic compensation of *Water* to *Kingfisher* is far larger than that of *Bird* to *Kingfisher*. From the above comparison, it is reasonable that the diversity of a tag to image is proportional to its relevance to image content and its semantic compensation to the tags ranked ahead of it.

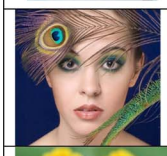
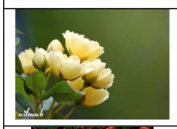
Measurement 4: The diversity of a tag τ to an image I with already determined tag set Γ is proportional to the relevance of the tag to image and the semantic compensation to the tags in Γ.

Let $D(\tau)$ denote the diversity of the tag τ to the image I with already determined tag list $\Gamma = \{\tau_1, \cdots, \tau_{|\Gamma|}\}$. In terms of Measurement 4, the diversity of the tag τ to the image I is measured by the product of the relevance of this tag to image and its semantic compensation to the tags ranked ahead of it. More generally, the diversity of the tag can be represented as follows:

$$D(\tau) = r(\tau) \times C(\tau)^l \tag{5}$$

TABLE I

EXEMPLAR IMAGES FOR SHOWING THE INITIAL TAG RANKING AND IMAGE RETAGGING PERFORMANCES OF DIFFERENT APPROACHES

	FLICKR	Initial Tag Ranking				Tag Enrichment			
photo	INIT	RANK	RLVT	NBVT	DIVS	RANK	RLVT	NBVT	DIVS
	sea portrait beach smile reflex spring uro	sea beach smile portrait	beach sea smile portrait	sea beach portrait smile	beach portrait smile sea	beach sand ocean sea water waves woman man holiday shore	beach sea portrait smile ocean sand water girl woman waves	sea beach ocean water blue sky nature sand nikon landscape	beach portrait smile water waves girl ocean sand sea woman
	bridge sun lake set louisiana causeway	set louisiana lake sun bridge	set sun lake bridge louisiana	sun lake set bridge louisiana	set lake bridge louisiana sun	sky cloud landscape ocean water sunset coast sun travel set	set lake sun bridge louisiana midnight water day silhouette cross	sun sunset sky nature sea blue water beach nikon landscape	set lake bridge midnight cross day water louisiana silhouette sun
	sunset sea sun black love nature evening natural quality latvia	sunset sun sea natural evening love nature quality black	sunset sun sea evening natural nature love black quality	sun sunset nature sea evening black love natural quality	sunset black love evening natural quality sun sea nature	sunset sun sea ocean beach natural water evening landscape love	sunset sun sea ocean beach natural water evening landscape love	sun sunset sky nature sea water beach sunrise landscape nikon	sunset love beach ocean sea sun natural evening water landscape
	river cross tiger	cross river tiger	river cross tiger	tiger river cross	river tiger cross	cross river travel water nikon tiger digital asia eos portrait	river cross tiger year drinking god bicycle travel sitting atmosphere	tiger zoo animal cat wildlife nature nikon india tigris mammal	river tiger cross drinking god bicycle year atmosphere travel sitting
	tower landscape unesco sunburst sari malacca	unesco sunburst tower landscape	sunburst unesco tower landscape	tower landscape unesco sunburst	sunburst tower unesco landscape	sunburst unesco landscape tower architecture wide travel sky cloud coast	sunburst unesco landscape tower architecture wide travel sky cloud coast	tower architecture blue sky building night nikon landscape paris europe	sunburst tower unesco landscape sky wide coast travel architecture cloud
	portrait woman girl beauty glamour	woman girl portrait glamour beauty	woman girl portrait glamour beauty	portrait beauty girl woman glamour	woman beauty girl portrait glamour	glamour woman girl portrait fashion beauty art face photography color	woman girl portrait glamour beauty fashion face lips hair lady	bird blue nature green animal butterfly feather zoo wildlife color	woman beauty girl face glamour hair lady portrait fashion lips
	flower butterfly bokeh pollen	bokeh flower pollen butterfly	bokeh flower butterfly pollen	butterfly flower bokeh pollen	bokeh pollen flower butterfly	pollen plant garden flower bokeh flora yellow insect leaf green	bokeh flower butterfly pollen insect stamen nectar plant garden yellow	butterfly nature bird blue insect green wildlife animal zoo lepidoptera	bokeh pollen flower butterfly yellow stamen nectar plant insect garden
	flower yellow japan little	flower yellow japan little	flower yellow japan little	flower yellow japan little	flower little yellow japan	beauty little flower yellow green plant garden color bokeh flora	flower yellow japan little plant small sunflower blossom green tokyo	rose flower red nature garden nikon yellow excellence green love	flower little yellow japan blossom sunflower small tokyo plant green
	tree green apple garden branch	branch tree green garden apple	green branch tree apple garden	apple green tree garden branch	green apple tree garden branch	branch tree color green leaf flora nature fauna apple natural	green tree branch apple ali garden color natura fauna leaf	apple fruit red food green tree nature nikon macintosh iphone	green apple tree garden leaf color branch ali natura fauna
	woman art apple	woman art apple	woman art apple	apple woman art	woman apple art	woman art girl apple portrait photography man painting fashion fantasy	woman art apple girl seul lady agua photography fashion portrait	apple iphone macintosh red fruit nikon ipod green food computer	woman apple art lady agua portrait fashion photography girl seul

where $r(\tau)$ is the relevance of tag τ to image and $C(\tau)$ is the semantic compensations of tag τ to the tag set Γ and l is a positive real number, which can be varied from $(0, +\infty)$. More detailed discussions on different diversity measurement approaches to the tagging performances are discussed in detail in Section V-E. For our baseline approach, we set $l = 1$. Thus, diversity can also be represented as follows:

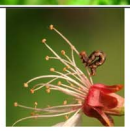
$$D(\tau) = r(\tau) \times C(\tau). \tag{6}$$

Now we turn to measure the semantic compensation $C(\tau)$ of a tag τ to a tag set. From the above analysis, we find that the semantic compensations of a tag to the tag set can be uncovered by analyzing the textual distance of tag τ to the tags in Γ. The larger the distance of a tag to the tag set, the larger the semantic compensation. Intuitively, the minimum distance and the average distance can be utilized directly to measure the semantic compensation. In the form of using the minimum distance (denoted as MIN) of tag τ to the tags in Γ to measure the semantic compensations, $C(\tau)$ can be expressed as follows:

$$C(\tau) = \min_{\tau_i \in \Gamma} (1 - s(\tau, \tau_i)) \tag{7}$$

where $s(\tau, \tau_i) = \exp(-d(\tau, \tau_i))$. In the form of using the average distance (denoted AVR) of a tag τ to the tags in Γ to

TABLE I
CONTINUED

	summer flower butterfly purple prairie wildflower	wildflower flower purple prairie butterfly	wildflower prairie flower purple butterfly	butterfly flower purple prairie wildflower	wildflower butterfly flower purple prairie	wildflower prairie yellow meadow farm flora green flower nature plant	wildflower flower purple prairie butterfly bokeh insect yellow flora nature	butterfly nature insect flower green bokeh yellow blue nikon monarch	wildflower butterfly purple nature insect yellow flora flower bokeh prairie
	insect cherry spring blossom	blossom cherry insect	blossom cherry insect	cherry blossom insect	blossom insect cherry	blossom plant flower leaf flora garden green bokeh petal nature	blossom cherry insect bokeh makro plant branch leaf flower garden	cherry red fruit blossom nature green tree food flower nikon	blossom insect bokeh leaf garden plant makro cherry branch flower
	bird forest landscape eos taiwan q alishan	landscape eos taiwan forest bird	landscape forest eos taiwan bird	forest landscape eos bird taiwan	landscape bird taiwan eos forest	landscape morning sunrise color mountain bird water scenic cloud tree	landscape eos taiwan forest tree bird color asia nature steam	forest nature tree landscape green water wald nikon germany snow	landscape bird eos forest tree taiwan nature steam color asia
	park painting spring nice artist	artist painting nice park	artist painting nice park	park painting nice artist	artist park painting nice	nice art artist painting photography design color life beauty natural	artist painting nice art park lighting lady stunning world best	cherry blossom tree sakura nature flower red japan green fruit	artist park painting nice art lighting stunning world best lady
	art japan kyoto geisha	japan kyoto art	japan kyoto art	japan art kyoto	japan art kyoto	kyoto asia japan travel sakura art culture world photography color	japan kyoto art bamboo culture asia sakura kimono tokyo roppongi	cherry blossom sakura flower tree nature japan red bokeh nikon	japan art kyoto kimono culture asia sakura tokyo bamboo roppongi
	africa forest aircraft sudan infrastructure cessna yambio	aircraft africa forest	aircraft africa forest	forest africa aircraft	aircraft forest africa	aircraft holiday travel photography island art dusk best tourism aviation	aircraft africa forest heat valley wales international transport travel minolta	forest nature tree landscape green wald nikon water sun germany	aircraft forest africa travel minolta wales international valley heat transport
	tree pinetree pine oregon forest truck logging lumber pinaceae	pine forest tree oregon truck	pine forest tree oregon truck	forest tree pine oregon truck	pine truck tree forest oregon	pine forest tree oregon truck moss trail mist fog landscape	pine forest tree oregon truck moss trail mist fog landscape	forest nature green tree landscape wood nikon water sun germany	pine truck tree forest oregon landscape mist moss trail fog
	road winter sky landscape highway horizon nevada	road highway landscape nevada horizon sky	road highway nevada landscape horizon sky	highway road sky landscape nevada horizon	road sky horizon landscape highway nevada	scenic nevada horizon landscape mountain cloud travel sky road sunset	road landscape highway nevada sky horizon scenic mountain cloud travel	road highway california sky landscape police canada travel blue patrol	road sky horizon travel scenic nevada mountain highway landscape cloud
	flower insect lavender bee lamiaceae	bee insect lamiaceae lavender flower	bee insect lamiaceae flower lavender	lavender flower bee insect lamiaceae	bee flower lavender insect lamiaceae	plant bee garden flower purple insect bokeh green lavender flora	bee insect flower lamiaceae lavender bumblebee purple garden plant bokeh	lavender purple flower garden bee bokeh nature blue green insect	bee flower lavender lamiaceae bokeh purple bumblebee insect garden plant
	woman hand purple arm lavender 365	arm hand purple lavender woman	arm hand woman purple lavender	lavender purple woman hand arm	arm lavender woman hand purple	arm hand woman girl colorful photography hair green black blue	arm hand purple lavender woman lady shirt bracelet girl scarf	lavender purple flower sachet etsy nature soap garden vintage bokeh	arm lavender woman bracelet shirt lady scarf hand purple girl

measure the semantic compensations, $C(\tau)$ can be expressed as follows:

$$C(\tau) = 1 - \frac{1}{|\Gamma|} \sum_{t \in \Gamma} s(\tau, t). \tag{8}$$

The impact of using MIN and AVR to tagging performances is discussed in Section V.

IV. IMAGE TAGGING WITH DIVERSE SEMANTICS

Our approach selects an optimal tag from the unranked tag list at each step with respect to the diverse semantic criteria. By adding the optimal tag to the already determined tag list, the semantic coverage can be maximized. In this section, we first introduce the low-level features utilized in this paper and then give the details of tagging process.

A. Low-Level Feature Representation

For each image, we extract 470-D low-level visual features, including 225-D blockwise color moment features generated from 5-by-5 fixed partition of the image, 170-D hierarchical wavelet packet features [56], and 75-D edge distribution histogram features. The reason that we use these features

is mainly taking the following facts into account: 1) color and texture information are important features to represent the visual content of nature image; 2) the blockwise color moment is efficient to capture the localized color feature of image, 3) the hierarchical wavelet packet descriptor makes full use of the multiresolution characteristics of wavelet packet transform to represent the salient texture information; and 4) the edge histogram is a simplified local feature representation approach that is robust to location and illumination variations. Note that for NUS-WIDE dataset, we use the low-level features shared by Chua *et al.* [19] rather than the features mentioned earlier.

B. Proposed Tagging Approach

Given a tag set $\Gamma = \{\tau_1, \cdots, \tau_{|\Gamma|}\}$, let $\Gamma^{'}$ denote the set of an ordering of tags τ_i ($\tau_i \in \Gamma, i \in \{1, \cdots, |\Gamma|\}$, and $\Gamma^{'} = \{\tau_1^{'}, \cdots, \tau_{|\Gamma|}^{'}\}$). The proposed social image tagging approach is aiming at making the top-ranked tags covering diverse semantics. The tags with high relevance to image have significant semantic compensations to their previous and are ranked ahead. From the tag diversity measurement given in (6), we find that our approach is actually a greedy search-based tag order determination approach. It aims at maximizing the diversities of the top-ranked tags by looking for best candidate tags at each step. Thus, the optimal tag list for the social image is determined iteratively. Based on the top-ranked $y-1$ tags, our algorithm determines an optimal tag τ_y^* at the yth step as follows:

$$\tau_y^* = \arg \max_{\tau_k \in \Gamma - \Gamma_y^*} D(\tau_k) = \arg \max_{\tau_k \in \Gamma - \Gamma_y^*} \omega_{\tau_k} \times C_{\tau_k} \tag{9}$$

where $\Gamma_y^* = \{\tau_1^*, \cdots, \tau_{y-1}^*\}$ is the ranked tag list at the previous $y-1$ steps.

We take Fig. 1 as an example to show the process of our greedy search-based optimal tag selection at each step. For all the tags in $\Gamma = \{\tau_1, \cdots, \tau_{|\Gamma|}\}$, we need to calculate their relevance to image. Actually, in this step, removing the tags with very low relevant scores from the candidate, tag list does not influence the tagging performances. From example, *Kingfisher* is the highest relevant tag to the image, and then we determine it as the first-ranked tag. For the other image content relevant tags, we further determine its semantic compensations to the already determined tag list $\Gamma^{'} = \{$'*Kingfisher*'$\}$ in the next step. We find that *Water* is highly relevant to the image and have largest semantic compensations to the top-ranked tag *Kingfisher*. According to the diverse semantic criteria as shown in (6), its diverse score is the highest among the remaining tags. Thus, *Water* is assigned as the second-ranked tag. At this time, the top two tags are *Kingfisher* and *Water*, that is, the top-ranked tag list at the second step is $\Gamma^{'} = \{$*Kingfisher*, *Water*$\}$. Based on the tag list $\Gamma^{'}$, the third tag can also be determined by maximizing the semantic coverage of these tags. The other tags can be determined iteratively, vice versa.

The corresponding algorithm is shown in Fig. 2. Let $\Omega = \{v_1, \cdots, v_{|\Omega|}\}$ denote the set of the whole tag vocabulary, which is obtained from the crawled Flickr images. The tag number in Ω is $|\Omega|$. Actually, if we set $\Omega = \varphi = \{t_1, \cdots, t_{|\varphi|}\}$, then our approach is to rank user-annotated tags with diverse semantics.

Now, we briefly analyze the computational complexity of the proposed approach. In the tag-to-image relevance calculation, we need to carry out low-level feature extraction and determine the VS of input image with the tag-related images. This process takes almost more than 99.99 of the whole computational cost of the whole social image tagging process, especially, for the tag vocabulary $\Omega = \{v_1, \cdots, v_{|\Omega|}\}$ containing thousands of tags. The greedy search-based optimal tag determination is comparatively computational efficient. As the relevant scores of tags to image are determined and the tag-to-tag similarity can be determined offline, we only need to update tag diversity in each step. Removing image content irrelevant tags by setting minimum tag-to-image relevant threshold can speed up the greedy search algorithm. However, compared with the computational costs utilized in measuring tag to image relevance, the saved computational cost is not too significant.

V. Experiments and Discussion

We compare DIVS with the raw tags by Flickr users (denoted INIT), tag concurrence-based approach (denoted COCR), random walk-based tag ranking approach [1] (denoted RANK), and visual neighbor voting (denoted NBVT) [8]. Our experiments consist of the raw tag ranking, tag enrichment, and tag-based image retrieval. We conduct experiments on our crawled dataset [38], [40] to show the effectiveness of the proposed approach by providing both subjective and objective tagging performances and tag-based image retrieval performance. Moreover, experiment on NUS-WIDE [19] is also given. This section is organized as follows: 1) our dataset is briefly reviewed in Section V-A; 2) performances evaluation criteria is illustrated in Section V-B; 3) subjective tagging performance is provided in Section V-C; 4) exemplar tagging performances and discussions are given in Sections V-D and V-E; 5) experiment on NUS-WIDEis provided in Section V-F; and 6) tag-based image search is evaluated in Section V-G.

A. Data Collection

We randomly select 25 queries, including *Alcedoatthis*, *Apple*, *Beach*, *Bear*, *Butterfly*, *Cherry*, *Deer*, *Eagle*, *Forest*, *Highway*, *Jeep*, *Lavender*, *Lotus*, *Orange*, *Peacock*, *Rose*, *Sailship*, *Sea*, *Sky*, *Strawberry*, *Sun*, *Sunflower*, *Tiger*, *Tower*, and *Zebra*, and then perform tag-based image search with "ranking by interestingness" option on Flickr. The representative image of each query is shown in Fig. 3. The images are of medium sizes with maximum width or height fixed to 500 pixels. The top 5000 returned images for each query are collected together with their associated information, including tags, uploading time, user identifier, and others. In this way, we obtain an image collection with 52 418 images. In total, there are 887 353 raw tags. The maximum raw tag number of an image is 256 and the average raw tag number per image is 16.93.

Many of the initial tags are noisy and meaningless. Hence, we adopt a prefiltering process for these tags. We match each tag with the entries in a Wikipedia thesaurus and only the tags that have a coordinate in Wikipedia are kept. Moreover,

---

Input: Given image I and its valid initial tag set $\varphi = \{\tau_1, \mathrm{L}\ , \tau_{|\varphi|}\}$. The valid initial tag number is $|\varphi|$.

Initialize $\Gamma_y^* = \varnothing$ (i.e. $|\Gamma_y^*| = 0$) and y=1.

1) Calculate the relevance of tag to image $r(\tau), \tau \in \varphi$ according to Eq. (4).

2) Select a tag from $\varphi = \{\tau_1, \mathrm{L}\ , \tau_{|\varphi|}\}$ and assign it with the first rank as follows.

$$\tau_y^* = \arg\max_{\tau_k \in \varphi} r(\tau_k)$$

3) Update: $\Gamma_y^* \leftarrow \Gamma_y^* + \tau_y^*$, $\Gamma_c \leftarrow \varphi - \Gamma_y^*$, and $y \leftarrow y+1$.

4) Calculate the diverse score of the tag τ ($\tau \in \Gamma_c$) to the image I given the already ranked tag set Γ_y^* according to Eq.(8).

5) Select an optimal tag τ_y^* from Γ_c and assign it with the y-th rank as follows

$$\tau_y^* = \arg\max_{\tau_k \in \Gamma_c} r(\tau_k) \times C(\tau_k)$$

where $C(\tau_k) = \min_{\tau \in \Gamma_y^*}(1 - s(\tau, \tau_k))$.

6) Repeat Step 3) and Step 5) until $\Gamma_c = \varnothing$.

Output: The ranked tag list $\varphi^* = \Gamma_y^*$.

Fig. 2. Algorithm of image tagging with diverse semantic.

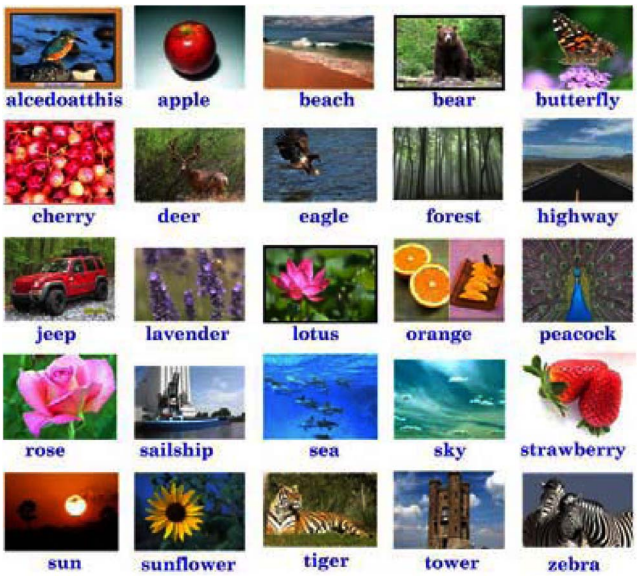


Fig. 3. Representative images of the 25 queries crawled from Flickr.

the tags with their appearing times less than 20 in the dataset are removed. There are also some high-frequency stop words, such as picture, photo, and others. These tags contain little information. We also consider that these tags are irrelevant to the image. They can be suppressed by removing the tags that appear more than 5000 times in the dataset. The vocabulary size of tag set is 1436. After preprocessing, the valid tag number is 312 652. The maximum valid tag number of an image is 65. The average valid tag number per image is 5.96.

B. Performance Evaluation Criteria

Five hundred images randomly selected from our Flickr dataset are utilized for initial tag ranking and tag enrichment performances evaluation. Five volunteers are involved in tag enrichment performances evaluation. They all claim that they are familiar with image share websites and work in the social image tagging, image retrieval related area. In order to show whether the top-ranked tags cover wide range of semantics, we ask them to evaluate the top-ranked tags for the test images of different algorithms with respect to tag diversity. In tag diversity, a tag is evaluated with respect to its relevance to image content and the diverse semantics of it is provided to cover the semantics over the tags ranked ahead of it. Thus, the volunteers need to assign the relevance score and the diverse score for each ranked tag.

1) *Assign Tag Relevance Score:* For each image, each of the enriched tags is assigned with one of the five scores according to its relevance to image: most relevant (score 4), relevant (score 3), partial relevant (score 2), weak relevant (score 1), and irrelevant (score 0). We ask the five annotators to assign the relevant score for each tag. We assign the tag most relevant to the image content by taking into the following aspects into account.

1) The tag is with most relevant to the image if most important content of the image is disclosed by it.
2) The tag is with relevant to the image if important content but not the most important content is disclosed by it.
3) The tag is with partial relevant to the image if some parts of image content is disclosed by it.
4) The tag is with weak relevant if a small part of image content is disclosed by it.
5) The tag is with irrelevant to the image if no content of the image is disclosed by it.

For example, in Fig. 1, the tags *Kingfisher* can be viewed as the most relevant to the image content while *Eagle* is irrelevant to image. *River* is a relevant tag to the image while *Mountain* does not. *Blue* is partial relevant to the image while *Red* is not. *Water* is relevant to the image, *Water drop* is viewed as weak relevant to the image content, and *Fire* is irrelevant to the image content.

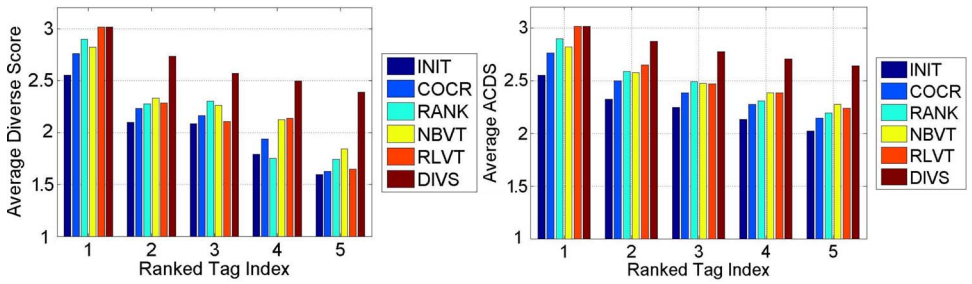


Fig. 4. Tag ranking performances of INIT, COCR, RANK, NBVT, RLVT, and DIVS. (a) ADSs. (b) Average ACDS.

2) *Assign Tag Diverse Score:* For each tag recommended by different tag enrichment approaches, we ask the annotators to classify the diverse score of each tag into one of the following five levels: very large (score 4), large (score 3), medium (score 2), weak (score 1), and none (score 0). The diverse score of a tag is assigned according to the following rules.

1) If and only if the tag is most relevant or relevant to the image and it has large semantic compensations to the tags ranked ahead of it for descripting the content of the input image, then we assign the diverse score of this tag as very large.
2) If and only if the tag is partial relevant to the image and it has large semantic compensation to the tags ranked ahead of it, then we assign the diverse score of this tag as large.
3) If the tag is partial relevant to the image content and it has partial semantic compensation to the tags ranked ahead of it, then we assign the diverse score of this tag as medium.
4) If the tag is weak relevant to the image content and it has weak semantic compensation to the tags ranked ahead of it, then we assign the diverse score of this tag as weak.
5) If the tag is irrelevant to the image content, then we assign the diverse score of this tag as *none*. For the first tag, its diverse score is set to be the relevance score.

In order to make the evaluation valid, the following rules are set: 1) the test image is shown and the corresponding tags lists by different approaches are shown; 2) then the five annotators are asked to provide their scores for the ranked tags; and 3) finally, we use the major voting-based approach for determining the final scores of the tags. We encourage discussions for the annotation results when the majority of them do not agree with each other. After discussion, they give their scores again until they agree with the final scores.

C. Tag Enrichment Performances Evaluation

In this paper, six approaches are evaluated to show the effectiveness of the proposed tag enrichment approach. They are the initial tags annotated by Flickr users (denoted INIT), random walk-based approach [1] (denoted RANK), tag co-occurrence-based approach [20] (denoted COCR), neighbor voting-based approach [8] (denoted NBVT), the relevance-based approach (denoted RLVT; in this approach, the diversity of each tag is viewed identical, it is a special case of the proposed approach), and the proposed diverse semantic-based approach (denoted DIVS). In NBVT, we set the visual neighbors to be 500 according to the suggestions given in [8] and [35]. In RANK, NBVT, RLVT, and DIVS, the low-level features are all the same. Thus, the comparisons are fair. Our experiments consist of two parts: initial tag ranking and tag enrichment.

To evaluate the performances of different tagging approaches, we use average diverse score (ADS) and average accumulating diverse score (AAD) to evaluate the tag enrichment and ranking performances. The ADS at the kth-ranked tag index is calculated as follows:

$$ADS(k) = \sum_{i=1}^{K} D_i^k \big/ K \tag{10}$$

where D_i^k is the diverse score of the ith test image under ranked tag index k and K is the total test image number. The AAD is as follows:

$$AAD(k) = \sum_{i=1}^{K} ACDS_i^k \big/ K \tag{11}$$

where $ACDS_i^k$ is the diverse score of the ith test image under ranked tag index k

$$ACDS_i^k = \left(ACDS_i^{k-1} + D_i^k\right) / k. \tag{12}$$

The ADS is to measure the semantic compensation of current tag to its previously ranked tags. The AAD is to evaluate the semantic coverage of already determined tags. The ADS measures the semantic coverage of top-ranked tags from a local point of view. The AAD measures the semantic coverage of top-ranked tags from a global point of view.

1) *Initial Tag Ranking Performances:* In Fig. 4(a) and (b), the ADSs and the average ACDS of INIT, RANK, COCR, NBVT, RLVT, and DIVS for the test images for the top five ranked tags are plotted. From Fig. 4, the scores of different approaches are all in descending with the increase of the ranked tag index. The ADSs and ACDS of the first five tags of DIVS are larger than these of the others. This shows that our approach is effective to determine semantic compensative tags and ranks them ahead.

2) *Tag Enrichment Performances Evaluation:* Fig. 5 shows the ADSs and average ACDS of the corresponding tag enrichment performances of INIT, COCR, RANK, NBVT, RLVT, and DIVS for the top ten ranked tags, respectively. For the

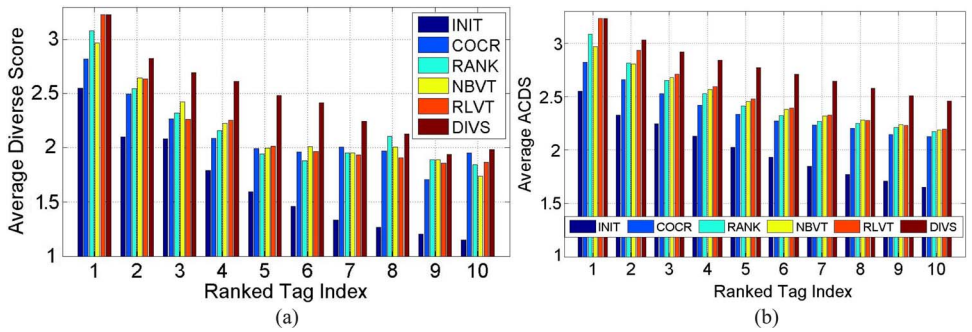


Fig. 5. Tag enrichment performances comparisons for INIT, COCR, RANK, NBVT, RLVT, and DIVS. (a) ADSs. (b) Average ACDS.

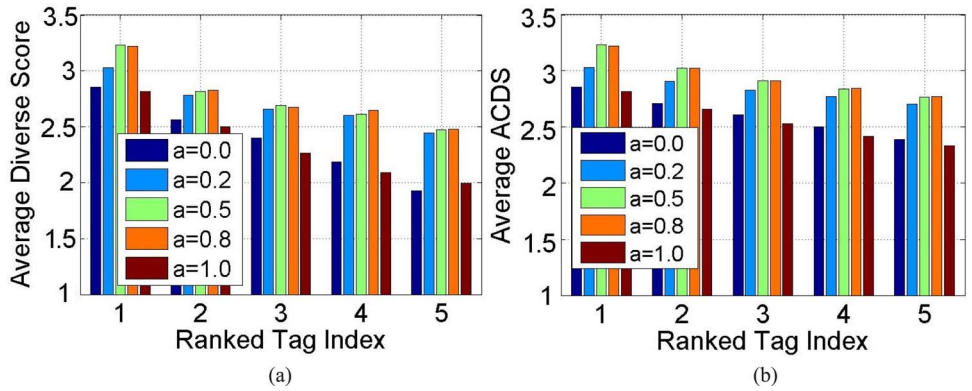


Fig. 6. Impacts of parameter α to tagging performances with the top-ranked tag index in the range [1], [5]. (a) ADSs. (b) Average ACDS.

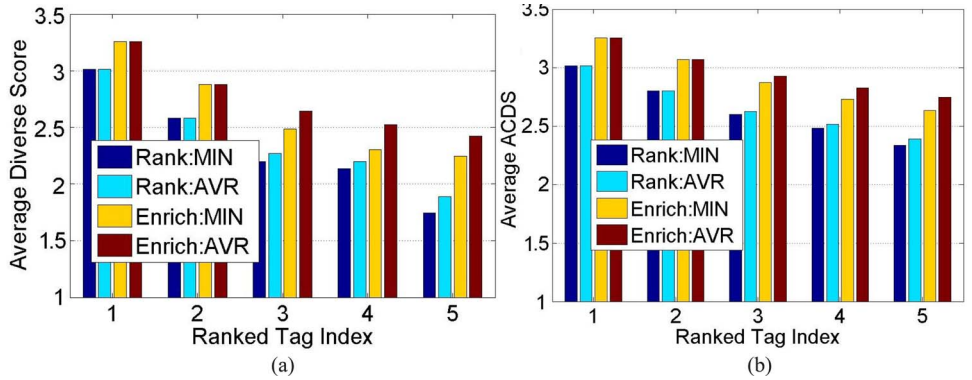


Fig. 7. Average relevant and diverse scores of MIN and AVR with the top-ranked tag index in the range [1], [5] for the 100 test images. (a) Average relevant scores. (b) ADSs.

first-ranked tags, the ADSs of INIT, COCR, RANK, NBVT, RLVT, and DIVS are 2.55, 2.82, 3.08, 2.97, 3.23, and 3.23, respectively. And for the fifth-ranked tags, the corresponding ACDS of COCR, RANK, NBVT, RLVT, and DIVS are 2.3338, 2.4106, 2.4535, 2.4790, and 2.7696, respectively. The corresponding accumulating diverse scores and ACDS are larger than these of tag ranking approaches. Moreover, the accumulating diverse scores of COCR, RANK, NBVT, and RLVT are very close. Comparatively, better performances are achieved by DIVS.

D. Exemplar Results of Tag Ranking and Tag Enrichment

Table I shows several exemplar images (as shown in the first column) by providing their initial tags (i.e., INIT as shown in the second column), tag ranking results of RANK, NBVT, RLVT, and DIVS, and tag enrichment results of RANK, NBVT, RLVT, and DIVS. Please turn to the last two pages for details. Only the top-ranked ten tags of the tag enrichment approaches are listed. Due to page size limits, the results of COCR are not provided in Table I.

In Table I, some initial tags appeared in INIT but did not appeared in the ranked tag lists of RANK, NBVT, RLVT, and DIVS are due to the fact that their appearing times are less than 20. They are removed from the valid tag list before carrying out tag ranking and new tag enrichment.

From the enriched results, it is clear that the top-ranked tags of DIVS are highly relevant to the image content and have significant semantic compensations to the tags before them. For example, for the second image, the first three tags enriched by DIVS are *Set*, *Lake*, and *Bridge*; by RLVT are

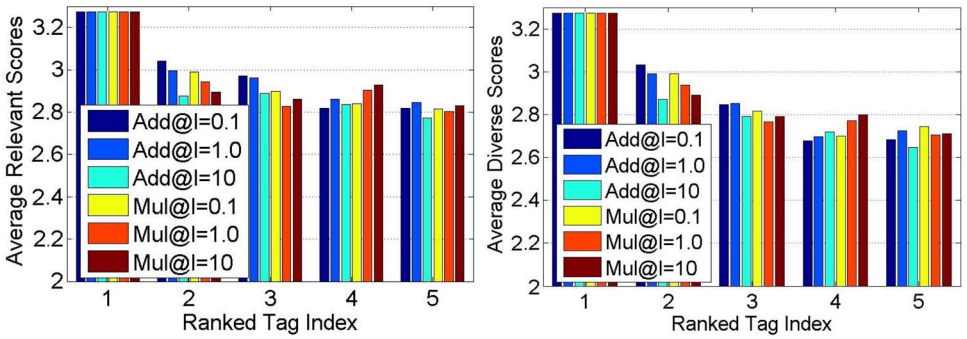


Fig. 8. Average relevant scores and diverse scores of different tag diversity measurement approaches under different similarity measurement approaches for the top five ranked tags for the randomly evaluated 472 test images. (a) Average relevant scores. (b) ADSs.

TABLE II

RECALL, PRECISION, AND F1 VALUES OF THE TOP 5 RANKED TAGS AND TOP 10 RANKED TAGS RECOMMENDED BY RANK, NBVT, RLVT, AND DIVS

	Top 5 ranked tags			Top 10 ranked tags		
	Recall	Precision	F1	Recall	Precision	F1
RANK	30.10	15.79	20.71	43.06	13.19	20.19
NBVT	31.17	18.19	22.97	36.10	19.31	25.16
RLVT	40.63	19.17	26.05	55.43	15.91	24.72
DIVS	40.63	19.17	26.05	55.43	15.91	24.72

Set, *Lake*, and *Sun*; and by NBVT are *Sun*, *Sunset*, and *Sky*. The three tags recommended by DIVS cover wide range of semantics than the tags enriched by the other approaches. For the last image, the top three tags enriched by DIVS are *Arm*, *Lavender*, and *Woman*; by RLVT are *Arm*, *Hand*, and *Purple*; and by NBVT are *Lavender*, *Purple*, and *Flower*. The three tags enriched by DIVS also cover wide range of semantics than the tags enriched by the other approaches. From the exemplar images, we find that the performances of the ranked tags and enriched tags by DIVS are better than the others.

E. Discussion on Parameters

In this section, we discuss the impacts of the parameter α and tag diversity measurement approaches to the tag enrichment performances of DIVS. Experiments are carried out on randomly selected 100 images by returning the top-ranked five tags.

1) *Impact of Parameter α:* In the above experiments, the parameter α, as shown in (4), is set to be 0.5. The impact of this parameter to the tagging performance is shown in Fig. 6. The experiments are carried out on the randomly selected 100 test images with $\alpha \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$. $\alpha = 0.0$ means that only visual features are utilized in tag enrichment. This case is identical to tagging images without user-labeled initial tags. In this circumstance, its performance is similar to that of NBVT. Comparatively, the recommended tags are with low relevance and low diversity. α=1.0 means that only high-level TSs are utilized in tag enrichment; thus, some relevant tags cannot be inferred effectively without using the visual information of the image. The corresponding performance is similar to that of COCR. Better performances are achieved when both visual and textual relevance are taken into account.

2) *Impact of Tag Diversity Measurement Approaches:* In the above experiments, tag diversity is measured by the minimum score (denoted MIN) as shown in (7). Moreover, another tag diversity measurement approach using the average score (denoted AVR) is also given in (8). In this section, we discuss the impacts of tag diversity measurement approaches to tag enrichment performances. The average relevant scores and diverse scores of MIN and AVR with the top-ranked tag index in the range [1, 5] for the 100 test images are shown in Fig. 7(a) and (b), respectively. From this comparison, we find that AVR-based tag diversity measurement approach is comparatively better than that of the MIN-based approach. Compared to MIN, AVR is robust to noise.

Moreover, in (5), the tag diversity measurement approach is given by the product of two terms $r(\tau)$ and $C(\tau)^l$. In this section, the performances of our tagging approach under different diversity measurement approaches are given. The performances of using the multiple of relevance and diversity as shown in (5) by setting $l = \{0.1, 1.0, 10\}$, the corresponding approaches are denoted by Mul@$l = 0.1$, Mul@$l = 1.0$, and Mul@$l = 10$. Moreover, we also use the summarization-based diversity measurement approach, where the diversity is expressed as follows:

$$D(\tau) = r(\tau) + C(\tau)^l. \tag{13}$$

By also setting $l = \{0.1, 1.0, 10\}$, the corresponding approaches are denoted by Add @$l = 0.1$, Add @$l = 1.0$, and Add @$l = 10$. The averaged relevant scores and the ADSs of

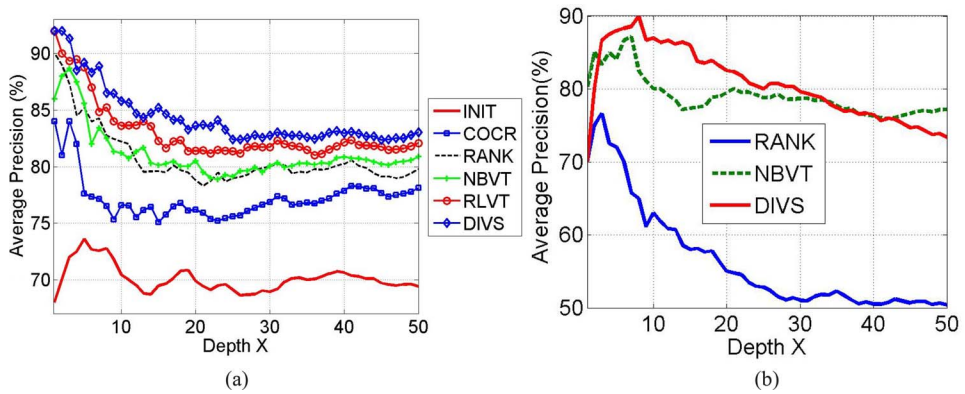


Fig. 9. Tag-based image search results of different approaches with depth x in the range [1], [50]. (a) On our dataset. (b) On NUS-WIDE.

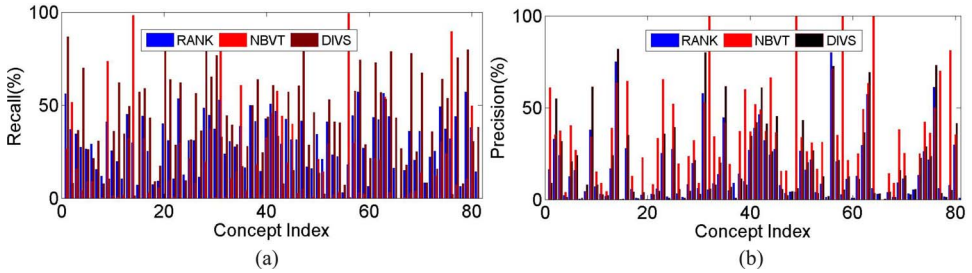


Fig. 10. (a) Recall and (b) precision of RANK, NBVT, and DIVS for the top recommended tags for the 81 concepts of NUS-WIDE.

the top-ranked five tags of randomly selected 472 test images are shown in Fig. 8(a) and (b), respectively, except that the performances of Add@$l=10$ are comparatively lower than the others. From Fig. 8, we find that under the criteria Add, the top-ranked three tags are with high average scores. This shows that tags with high relevance and with significant semantic compensations are ranked ahead by using the summarization of the relevance and diversity.

F. Experiments on NUS-WIDE

The NUS-WIDE dataset is a large-scaled real-world dataset crawled from Flickr. The data is composed with two parts: the training part, which contains 27 807 images, and the testing part, which contains 27808 images. All images are manually annotated with the concepts from 81 ground truth. Thus, in this paper, recall, precision, and F1 are used to measure the performance of different image tagging approach. Except the images and the ground truth labels for each image, the low-level features extracted from the image including color histogram (64-D), color correlation histogram (73-D), edge-detection histogram (73-D), block-wised color moments (256-D), and wavelet texture (128-D) are also provided [19]. We use the features provided by NUS-WIDE, rather than those of ourselves.

There are no initial tags for the test and training image; thus, in this section, we only use the visual features for tag recommendation. Correspondingly, we only use the visual features for determining the relevant tags. Thus, we can only provide the performances of RANK, RLVT, NBVT, and DIVS. In order to make fair comparisons, the features utilized by RANK, RLVT, NBVT, and DIVS are all the same. All the five low-level features are utilized; the total dimension of the feature of each image is 594. The corresponding average recall, precision, and F1 values of the 81 concepts by recommending 5 and 10 tags are shown in Table II, respectively. From Table II, we find that the performances of RLVT and DIVS are the same. This is caused by the fact that the 81 concepts in NUS-WIDE dataset are manually labeled. They are independent and have high compensation in semantic space. For more detail, the recall and precision values of RANK, NBVT, RLVT (the same as DIVS), and DIVS of the 81 concepts under top five ranked tags are plotted in Fig. 10(a) and (b), respectively.

G. Tag-Based Image Search

We conduct image search to verify the effectiveness of the proposed tagging approach of our crawled dataset. We first select the 25 queries as described in Fig. 3 to carry out tag-based image search. Then we compare the tag-based image search results based on the enriched tags by the following methods: 1) image search with initial tags labeled by user (INIT), that is, index the images with original tags; 2) image search with tags enriched by COCR, that is, index the images by the tags enriched by tag concurrence-based approach; 3) image search with tags enriched by RANK, that is, index the images by the tags enriched by random walk-based ranking; 4) image search with tags enriched by NBVT, that is, index the images by the tags enriched by the visual neighbor voting; and 5) image search with tags enriched by DIVS, that is, index the images by the tags enriched by our baseline method.

To quantitatively compare the image search results, we obtain the ranked image lists of different approaches for each query. We manually label the relevance of the top 50 images

of each query. For each ranking list, the images are decided as relevant or irrelevant with respect to the query terms. We use precision as image search evaluation metric. Given a ranked image list, the precision at depth n is defined as follows:

$$P_n = \frac{1}{n}\sum_{j=1}^{n} R_j \tag{14}$$

where R_j measures the relevance of the jth instance to the query. $R_j = 1$ if the jth instance is relevant and 0 otherwise. To evaluate the overall performance, we use average precision (AP) of the 25 queries. Note that during the textual-based indexing, the ranks of tags for the image are taken into account. We rank the image according to the scores of each tag in the ranks. We first return the images with tags matched with query with the first rank, then the second rank, and so on. For the images with the same ranks, the final-ranked images are ranked by the descending order of relevant scores.

Fig. 9(a) illustrates the AP at different return depths on our crawled dataset. We can see that the search results on the enriched tags by COCR, RANK, NBVT, RLVT, and DIVS are better than the INIT. Moreover, our approach DIVS outperforms the RANK, NBVT, and RLVT. This shows that our approach can assign the tags highly relevant to image content and with diverse semantics ahead.

Moreover, we also carry out tag-based image retrieval on NUS-WIDEtest dataset (with 27 808 images) by utilizing the following 10 concepts as queries: *Animal*, *Buildings*, *Clouds*, *Grass*, *Lake*, *Nighttime*, *Ocean*, *Sky*, *Soccer*, and *Sports*. The corresponding AP of the top-ranked 50 images of RANK, NBVT, and DIVS are shown in Fig. 9(b).

VI. Conclusion

In this paper, we address the tag diversity problem in social image tagging and give the corresponding measurements. The proposed tagging with diverse semantic approach improves the semantic coverage for an image from the top-ranked tags. Tag diversity is proportional to its relevance to image and sematic compensations to the tags ranked ahead of it. Making sure the tag is relevant to image content is important in the proposed diverse semantic-based tag enrichment. If some irrelevant tags are falsely determined as relevant tags, then these tags will give negative impacts on selecting optimal tags for improving the semantic coverage for the image. Two different tag diversity measurement approaches are evaluated.

Our approach can be utilized for reranking the enriched tags by the other image tagging or annotation algorithms to improve the semantic coverage of the top-ranked tags. The proposed image tagging with diverse semantics can be utilized to improve textual-based image retrieval because the top-ranked tags are highly relevant to the image and have large semantic compensation. The images containing the query textual terms that appear in the top-ranked tag list can be viewed as more relevant to the query. Moreover, the proposed approach can be utilized in selecting positive training samples and filtering noise samples from a large-scale weak-labeled image set in active learning.

References

[1] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang, "Tag ranking," in *Proc. WWW*, 2009, pp. 351–360.

[2] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 971–980.

[3] X. Li, L. Chen, L. Zhang, W. Ma, and F. Lin, "Image annotation by large-scale content-based image retrieval," in *Proc. ACM MM*, 2006, pp. 2057–2063.

[4] X. Rui, M. Li, Z. Li, W. Ma, and N. Yu, "Bipartite graph reinforcement model for web image annotation," in *Proc. ACM MM*, 2007, pp. 585–594.

[5] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Scalable search-based image annotation," *Multimedia Syst.*, vol. 14, no. 4, pp. 205–220, 2008.

[6] L. Kennedy, S. Chang, and I. Kozintsev, "To search or to label? Predicting the performance of search-based automatic image classifiers," in *Proc. ACM MIR*, 2006, pp. 249–258.

[7] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[8] X. Li, C. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[9] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, no. 6, pp. 1107–1135, 2003.

[10] E. Chang, G. Kingshy, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 26–38, Jan. 2003.

[11] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.

[12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Tagging over time: Realworld image annotation by lightweight meta-learning," in *Proc. ACM MM*, 2007, pp. 393–402.

[13] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," in *Proc. SPIE*, 2004, pp. 330–338.

[14] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1575–1589, Sep. 2007.

[15] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[16] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[17] X. Wang, L. Zhang, X. Li, and W. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[18] X. Li, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proc. ICASSP*, 2009, pp. 3717–3720.

[19] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. CIVR*, 2009, pp. 1–9.

[20] B. Sigurbjörnsson and R. Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. WWW*, 2008, pp. 327–336.

[21] K. Weinberger, M. Slaney, and R. Zwol, "Resolving tag ambiguity," in *Proc. ACM MM*, 2008, pp. 111–119.

[22] D. Liu, X. Hua, M. Wang, and H. Zhang, "Retagging social images based on visual and semantic consistency," in *Proc. ACM WWW*, 2010, pp. 1149–1150.

[23] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence and wordnet," in *Proc. ACM MM*, 2005, pp. 706–715.

[24] X. Wang, L. Zhang, M. Liu, Y. Li, and W. Ma, "ARISTA: Image search to annotation on billions of web photos," in *Proc. CVPR*, 2010, pp. 2987–2994.

[25] H. Chen, M. Chang, P. Chang, M. Tien, W. Hsu, and J. Wu, "SheepDog: Group and tag recommendation for Flickr photos by automatic search-based learning," in *Proc. ACM MM*, 2008, pp. 1155–1156.

[26] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Scalable search-based image annotation," *Multimedia Syst.*, vol. 14, no. 4, pp. 205–220, 2008.

[27] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Content-based image annotation refinement," in *Proc. CVPR*, Jun. 2007, pp. 1–8.

[28] J. Jia, N. Yu, X. Rui, and M. Li, "Multigraph similarity reinforcement for image annotation refinement," in *Proc. ICIP*, 2008, pp. 993–996.

[29] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM MM*, 2010, pp. 461–470.

[30] H. Xu, J. Wang, X. Hua, and S. Li, "Tag refinement by regularized LDA," in *Proc. ACM MM*, 2009, pp. 573–576.

[31] L. Wu, L. Yang, N. H. Yu, and X. Hua, "Learning to tag," in *Proc. ACM WWW*, 2009, pp. 361–370.

[32] D. Liu, X. Hua, M. Wang, and H. Zhang, "Image retagging," in *Proc. ACM MM*, 2010.

[33] M. Huiskes and M. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM MIR*, 2008.

[34] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," in *Proc. ACM MM*, 2008, pp. 31–40.

[35] X. Li, C. G. M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proc. ACM MIR*, 2008, pp. 180–187.

[36] M. Wang, K. Yang, X. Hua, and H. Zhang, "Visual tag dictionary: Interpreting tags with visual words," in *Proc. WSMC*, 2009.

[37] N. Zhou, W. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1281–1294, Jul. 2011.

[38] X. Qian and X. Hua, "Graph-cut based tag enrichment," in *Proc. SIGIR*, 2011.

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

[40] X. Qian, X. Hua, and X. Hou, "Tag filtering based on similar compatible principle," in *Proc. ICIP*, 2012, pp. 2349–2352.

[41] D. Cai, X. He, Z. Li, W. C. Ma, and J. C. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *Proc. ACM MM Conf.*, 2004, pp. 952–959.

[42] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. SIGIR Conf.*, 1998, pp. 335–336.

[43] W. Goffman, "A searching procedure for information retrieval," *Inform. Storage Retrieval*, vol. 2, pp. 73–78, 1964.

[44] A. Jaimes, S. C. Chang, and A. C. Loui, "Detection of non-identical duplicate consumer photographs," in *Proc. ACM MM Conf.*, 2003, pp. 16–20.

[45] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W. Ma, "IGroup: Web image search results clustering," in *Proc. ACM MM*, 2006, pp. 587–596.

[46] R. H. V. Leuken, L. Garcia, X. Olivares, and R. Zwol, "Visual diversification of image search results," in *Proc. WWW*, 2009, pp. 341–350.

[47] K. Song, Y. Tian, T. Huang, and W. Gao, "Diversifying the image retrieval results," in *Proc. ACM MM*, 2006, pp. 707–710.

[48] S. Srinivasan and N. Sawant, "Finding near-duplicate images on the web using fingerprints," in *Proc. ACM MM*, 2008, pp. 881–884.

[49] A. Sun and S. S. Bhowmick, "Image tag clarity: In search of visual representative tags for social images," in *Proc. 1st SIGMM WSM*, 2009, pp. 19–26.

[50] B. Wang, Z. Li, M. Li, and W. Ma, "Large-scale duplicate detection for web image search," in *Proc. ICME*, 2006, pp. 353–356.

[51] C. Zhai, W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," *Inform. Process. Manag.*, pp. 10–17, 2006.

[52] C. Zhai and J. Lafferty, "A risk minimization framework for information retrieval," *Inform. Process. Manag.*, pp. 31–55, 2006.

[53] W. Zhao and C. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 412–423, Feb. 2009.

[54] J. Zhu, S. Hoi, M. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *Proc. ACM MM*, 2008, pp. 41–50.

[55] M. Wang, K. Yang, X. Hua, and H. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, Dec. 2010.

[56] X. Qian, G. Liu, D. Guo, Z. Li, Z. Wang, and H. Wang, "Object categorization using hierarchical wavelet packet texture descriptors," in *Proc. ISM*, 2009, pp. 44–51.

[57] R. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[58] S. Zhu, G. Wang, C. Ngo, and Y. Jiang, "On the sampling of web images for learning visual concept classifiers," in *Proc. CIVR*, 2010.

[59] S. Feng, C. Lang, and D. Xu, "Beyond tag relevance: Integrating visual attention model and multiinstance learning for tag saliency ranking," in *Proc. CIVR*, 2010, pp. 288–295.

[60] S. Zhang, J. Huang, H. Li, and D. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012.

[61] Z. Ma, F. Nie, Y. Yang, J. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[62] Y. Han, F. Wu, Q. Tian, and Y. Zhuang, "Image annotation by input-output structural grouping sparsity," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 3066–3079, Jun. 2012.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[64] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug. 2011.

[65] J. Shen, W. Meng, S. Yan, H. Pang, and X. Hua, "Effective music tagging through advanced statistical modeling," in *Proc. SIGIR*, 2010.

[66] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[67] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[68] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[69] Y. Yang, F. Wu, F. Nie, H. Shen, Y. Zhuang, and A. Hauptmann, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[70] M. Wang, B. Ni, X. Hua, and T. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," *ACM Comput. Surveys*, vol. 44, no. 4, Aug. 2012.

[71] G. Li, M. Wang, Z. Lu, R. Hong, and T. Chua, "In-video product annotation with web information mining," *ACM Trans. Multimedia Comput., Commun. Applic.*, vol. 8, no. 4, pp. 55:1–55:19, 2012.

[72] E. Moxley, J. Kleban, and B. S. Manjunath, "SpiritTagger: A geo-aware tag suggestion tool mined from Flickr," in *Proc. ACM MIR.*, 2008.

[73] K. Yang, X. Hua, M. Wang, and H. Zhang, "Tag tagging: Towards more descriptive keywords of image content," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 662–673, Aug. 2011.

[74] Y. Yang, F. Wu, D. Xu, Y. Zhuang, and L. Chia, "Cross-media retrieval using query dependent search methods," *Pattern Recognit.*, vol. 43, no. 8, pp. 2927–2936, 2010.

[75] M. Wang, R. Hong, G. Li, Z. Zha, S. Yan, and T. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 975–985, Aug. 2012.

[76] P. Moëllic, J. Haugeard, and G. Pitel, "Image clustering based on a shared nearest neighbors approach for tagged collections," in *Proc. CIVR*, 2008, pp. 269–278.

[77] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, "Hierarchical clustering of www image search results using visual, textual and link information," in *Proc. ACM MM*, 2004.

[78] R. Leuken, L. Garcia, X. Olivares, and R. Zwol, "Visual diversification of image search results," in *Proc. WWW*, 2009.

[79] A. Radu, J. Stöttinger, B. Ionescu, M. Menéndez, and F. Giunchiglia, "Representativeness and diversity in photos via crowd-sourced media analysis," in *Proc. 10th Int. Workshop-AMR*, 2012.

[80] K. Weinberger, M. Slaney, and R. van Zwol, "Resolving tag ambiguity," in *Proc. ACM MM*, 2008.

[81] R. Zwol, V. Murdock, L. Garcia, and G. Ramirez, "Diversifying image search with user generated content," in *Proc. ACM MIR*, 2008.

[82] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proc. ACM MM*, 2006.

[83] B. Zhang *et al.*, "Improving web search results using affinity graph," in *Proc. SIGIR*, 2005.

[84] C. Ziegler, S. McNee, J. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proc. WWW*, 2005.

[85] M. Paramita, J. Tang, and M. Sanderson, "Generic and spatial approaches to image search results diversification," in *Proc. ECIR*, 2009.

[86] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *Proc. ACM CIVR*, 2009, pp. 1–8.
[87] S. Rudinac, A. Hanjalic, and M. Larson, "Finding representative and diverse community contributed images to create visual summaries of geographic areas," in *Proc. ACM MM*, 2011, pp. 1109–1112.
[88] A. Popescu and I. Kanellos, "Creating visual summaries for geographic regions," in *Proc. 10th Int. Workshop AMR*, 2012.
[89] Y. Xue and X. Qian, "Visual summarization of landmarks via viewpoint modeling," in *Proc. ICIP*, 2012, pp. 2873–2876.
[90] 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. ICME*, 2013.
[91] Q. Li, Y. Gu, and X. Qian, "LCMKL: Latent-community and multikernel learning based image annotation," in *Proc. ACM CIKM*, 2013.
[92] 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.
[93] J. Li, X. Qian, Y. Tang, L. Yang, and C. Liu, "GPS estimation from users' photos," in *Proc. MMM*, 2013, pp. 118–129.

Authors' biographies and photographs are not available at the time of publication.