

# Learning Social Tag Relevance by Neighbor Voting

Xirong Li, *Student Member, IEEE*, Cees G. M. Snoek, *Member, IEEE*, and Marcel Worring, *Member, IEEE*

**Abstract**—Social image analysis and retrieval is important for helping people organize and access the increasing amount of user-tagged multimedia. Since user tagging is known to be uncontrolled, ambiguous, and overly personalized, a fundamental problem is how to interpret the relevance of a user-contributed tag with respect to the visual content the tag is describing. Intuitively, if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. Starting from this intuition, we propose in this paper a neighbor voting algorithm which accurately and efficiently learns tag relevance by accumulating votes from visual neighbors. Under a set of well-defined and realistic assumptions, we prove that our algorithm is a good tag relevance measurement for both image ranking and tag ranking. Three experiments on 3.5 million Flickr photos demonstrate the general applicability of our algorithm in both social image retrieval and image tag suggestion. Our tag relevance learning algorithm substantially improves upon baselines for all the experiments. The results suggest that the proposed algorithm is promising for real-world applications.

**Index Terms**—Multimedia indexing and retrieval, neighbor voting, social tagging, tag relevance learning.

## I. INTRODUCTION

THE advent of social multimedia tagging—assigning tags or keywords to images, music, or video clips by common users—is significantly reshaping the way people generate, manage, and search multimedia resources. Good examples are Flickr, which hosts more than 2 billion images with around 3 million new uploaded photos per day [1], and YouTube, which serves 100 million videos and 65 000 uploads daily [2]. Apart from their usage for general-purpose search, these rich multimedia databases are triggering many innovative research scenarios in areas as diverse as personalized information delivery [3], landmark recognition [4], concept similarity measurement [5], tag recommendation [6], and automatic image tagging [7], [8]. One would expect user-contributed tags to be a good starting point for all these applications.

Despite the success of social tagging, however, tags contributed by common users are known to be ambiguous, limited in terms of completeness, and overly personalized [9], [10]. This is not surprising because of the uncontrolled nature of social tagging and the diversity of knowledge and cultural background of its users. Although the relevance of a tag given

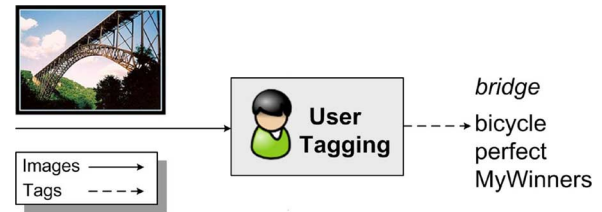


Fig. 1. **Dataflow of user tagging.** According to whether a tag is relevant with respect to a given image, we divide user-contributed tags into two types, namely objective and subjective tags. The objective tags are marked by an *italic* font. In this example, tag *bridge* is objective, while the other three tags are subjective. We aim for automated approaches to learning tag relevance.

the visual content can be subjective for a specific user, an objective criterion is desirable for general-purpose search and visual content understanding. We consider a tag relevant to an image if the tag accurately describes objective aspects of the visual content, or in other words, users with common knowledge relate the tag to the visual content easily and consistently. Other tags are subjective or overly personalized and thus we consider those irrelevant, as illustrated in Fig. 1. Apart from the fact that tags can be subjective, individual tags are mostly used once per image. This tagging behavior implies that given an image, relevant tags and irrelevant ones are not distinguishable by their occurrence frequency [11]. Hence, a fundamental problem in social image analysis and retrieval is how to accurately and efficiently learn the relevance of a tag with respect to the visual content the tag is describing.

Existing methods to automatically predict tag relevance with respect to the visual content often heavily rely on supervised machine learning methods [12]–[14]. In general, the methods boil down to learning a mapping between low-level visual features, e.g., color and local descriptors, and high-level semantic concepts, e.g., airplane and classroom. Since the number of training examples is limited for the supervised methods, the methods are not scalable to cover the potentially unlimited array of concepts existing in social tagging. Moreover, uncontrolled visual content contributed by users creates a broad domain environment having significant diversity in visual appearance, even for the same concept [15]. The scarcity of training examples and the significant diversity in visual appearance might make the learned models unreliable and difficult to generalize. Therefore, in a social tagging environment with large and diverse visual content, a light-weight or unsupervised learning method which effectively and efficiently estimates tag relevance is required.

Intuitively, if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. The intuition implies that the relevance of a tag with respect to an image might be inferred from tagging behavior of visual neighbors of that image. Starting from this intuition, we propose a novel neighbor voting algorithm for tag relevance learning. The key idea is, by propagating

Manuscript received January 05, 2009; April 13, 2009. First published August 18, 2009; current version published October 16, 2009. This work was supported in part by the EC-FP6 VIDI-Video project and in part by the STW SEARCHER project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu.

The authors are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, 1098 XG Amsterdam, The Netherlands (e-mail: x.li@uva.nl; cgmsnoek@uva.nl; m.worring@uva.nl).

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/TMM.2009.2030598

common tags through visual links introduced by visual similarity, each tag accumulates its relevance credit by receiving neighbor votes. Under a set of well-defined and realistic assumptions, we prove that our tag relevance learning algorithm is a good measure for both image ranking and tag ranking. To demonstrate the viability of the proposed algorithm, we provide a systematic evaluation on 3.5 million Flickr images for both social image retrieval and image tag suggestion.

The rest of the paper is organized as follows. We review related work in Section II. We then describe in detail tag relevance learning in Section III. We setup experiments in Section IV. Experimental results are presented in Section V. We conclude the paper in Section VI.

## II. RELATED WORK

We review work closely related to our motivation for tag relevance learning in the following two directions, that is, improving image tagging and improving image retrieval.

### A. Improving Image Tagging

Depending on whether a target image is labeled, we categorize existing methods into two main scenarios, namely improving image tagging for labeled images and automated image tagging for unlabeled images.

In the first scenario, given an image labeled with some tags, one tries to improve image tagging by removing noisy tags [16], recommending new tags relevant to existing ones [6], or reducing tag ambiguity [17]. In [16] for instance, the authors assume that the majority of existing tags are relevant with respect to the image. They then measure the relevance of a tag by computing word similarity between the tag and other tags, while in [6], the authors find new tags relevant with respect to the original ones by exploiting tag co-occurrence in a large user-tagged image database. To be precise, by using each of the original tags as a seed, they find a list of candidate tags having the largest co-occurrence with the seed tag. These lists are later aggregated into a single list and the top ranked tags are selected as the final recommendation. Since new tags are suggested purely using the initial tags, images with the same starting tags will end with the same new tags, regardless of the visual content. Hence, methods addressing both textual and visual clues are required.

Methods in the second scenario try to predict relevant tags for unlabeled images. We divide these methods according to their model-dependence into model-based and model-free approaches. The model-based approaches, often conducted in a supervised learning framework, focus on learning a mapping or projection between low-level visual features and high-level semantic concepts given a number of training examples [12]–[14], [18], [19]. Due to the expense of manual labeling, however, currently only a limited number of visual concepts can be modeled effectively. Besides, the approaches are often computationally expensive, making them difficult to scale up. Furthermore, the rapid growth of new multimedia data makes the trained models outdated quickly. To tackle these difficulties, a lightweight meta-learning algorithm is proposed in [20]. The gist of the algorithm is to progressively improve tagging accuracy by taking into account both the tags automatically predicted by an existing model and the tags provided by a user as implicit

relevance feedback. In contrast to the model-based approaches, the model-free approaches attempt to predict relevant tags for an image by utilizing images on the Internet [7], [8], [21], [22]. These approaches assume there exists a large well-labeled database such that one can find a visual duplicate for the unlabeled image. Then, automatic tagging is done by simply propagating tags from the duplicate to that image. In reality, however, the database is of limited-scale with noisy annotations. Hence, neighbor search is first conducted to find visual neighbors. Disambiguation methods are then used to select relevant tags out of the raw annotations of the neighbors. In [7], for instance, the authors rank tags in terms of their frequency in the neighbor set. However, tags occurring frequently in the entire collection may dominate the results. To restrain such effects, the authors in [8] re-weight the frequency of a tag by multiplying this frequency by its inverse document frequency (idf). The idf value of a tag is inversely and logarithmically proportional to the occurrence frequency of the tag in the entire collection. Nonetheless, the idf scheme tends to over-weight rare tags.

To summarize, the existing methods for image tagging try to rank relevant tags ahead of irrelevant ones in terms of the tags' relevance value with respect to an image. However, since the tag ranking criterion is not directly related to the performance of image retrieval using the tagging results, optimizing image tagging does not necessarily yield good image rankings [23].

### B. Improving Image Retrieval

Given unsatisfactory image tagging results, one might expect to improve image retrieval directly. Quite a few methods follow this research line, either by reranking search results in light of visual consistency [24]–[29] or by expanding the original queries [30]–[33]. We briefly review these methods in the following two paragraphs. For a more comprehensive survey, we refer to [15] and [34].

Reranking methods assume that the majority of search results are relevant with respect to the query and relevant examples tend to have similar visual patterns such as color and texture. To find the dominant visual patterns, density estimation methods are often used, typically in the form of clustering [25], [26] and random walk [28]. In [28] for instance, the authors leverage a random walk model to find visually representative images in a search result list obtained by text-based retrieval. To be precise, first an adjacent graph is constructed wherein each node corresponds to a result image and the edge between two nodes are weighted in terms of the visual similarity between the two corresponding images. A random walk is then simulated on the graph to estimate the probability that each node is visited. Since images in dense regions are more likely to be visited, the above probability is used to measure the representativeness of an image in the visual feature space and accordingly rerank the search results. However, density estimation is inaccurate when feature dimensionality is high and samples are insufficient for computing the density [35]. Besides, density estimation is computationally expensive. In [26] for example, the authors report an execution time of 18 s per search round, while a study on web users [36] shows the tolerable waiting time for web information retrieval is only 2 s, approximately. The difficulty in density estimation and the associated computational expense put the utility of reranking methods for social image retrieval into question.

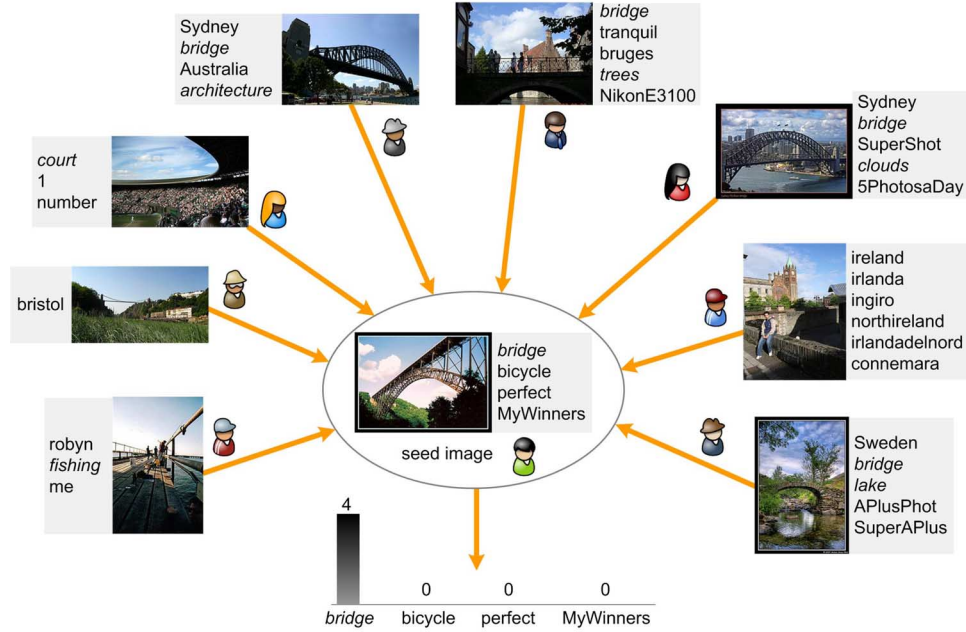


Fig. 2. **Learning tag relevance by neighbor voting.** The tag relevance value of each tag is estimated by accumulating the neighbor votes it receives from visually similar images of the seed image. In this example, since four neighbor images are labeled with *bridge*, the tag relevance value of *bridge* with respect to the seed image is 4. Hence, we update the tag frequency of *bridge* from 1 to 4.

Query expansion methods augment the original query by automatically adding relevant terms [30]–[32]. In [31], for instance, the authors use synonyms from a dictionary, whereas in [30], the authors select strongly related terms from text snippets returned by web search engines. Another example is [32], where the authors use clustering methods to find correlated tags. Though adding more query terms may retrieve more relevant results, how to choose appropriate expansion terms requires further research [37].

In summary, the reranking and query expansion methods try to rank images relevant with respect to a query ahead of irrelevant images. However, the methods leave the fundamental problem of subjective user tagging unaddressed.

Though we have witnessed great efforts devoted into improving both image tagging and image retrieval, the efforts are almost disconnected. Recent research, e.g., [38]–[41], investigates the potential of leveraging automatic tagging results for image and video retrieval. To the best of our knowledge, however, up until now, the solutions to the two problems are still separated, including our previous works [11], [22] which deal with social image retrieval and social image tagging, respectively. This work is an attempt to solve image ranking and tag ranking in a unified tag relevance learning framework. In contrast to approaches for image ranking which are query-dependent, e.g., [25] and [28], our algorithm is query-independent. This advantage allows us to run the algorithm offline without imposing extra waiting time on users. Further, by updating tag frequency with the learned tag frequency, we seamlessly embed visual information into current tag-based social image retrieval paradigms. For automatic image tagging, our algorithm shares similarities with the model-free approaches, e.g., [7], [8], and [21], since they can be regarded as propagating tags between neighbor images. Note, however, that our algorithm is more

general as it is applicable to both image retrieval and tagging. Moreover, we provide a formal analysis which is missing in previous studies.

### III. LEARNING TAG RELEVANCE BY NEIGHBOR VOTING

In order to fulfill image retrieval, we seek a tag relevance measurement such that images relevant with respect to a tag are ranked ahead of images irrelevant with respect to the tag. Meanwhile, to fulfill image tagging, the measurement should rank tags relevant with respect to an image ahead of tags irrelevant with respect to the image. Recall the intuition that if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. This intuition suggests that the relevance of a tag given an image might be inferred from how visual neighbors of that image are tagged: the more frequent the tag occurs in the neighbor set, the more relevant it might be, as illustrated in Fig. 2. However, some frequently occurring tags, such as “2007” and “2008”, are unlikely to be relevant to the majority of images. Hence, a good tag relevance measurement should take into account the distribution of a tag in the neighbor set and in the entire collection, simultaneously. Motivated by the informal analysis above, we propose a neighbor voting algorithm for learning tag relevance, as depicted in Fig. 2. Though the proposed algorithm is simple, we deem it important to gain insight into the rationale for the algorithm. The following two subsections serve for this purpose. Concretely, we first define in Section III-A two criteria to describe the general objective of tag relevance learning. Then, in Section III-B, we provide a formal analysis of user tagging and content-based nearest neighbor search. We see how our algorithm is naturally derived from the analysis. Finally, we describe in detail the algorithm in Section III-C.

### A. Objective of Tag Relevance Learning

We first introduce some notation for the ease of explanation. We denote a collection of user-tagged images as  $\Phi$  and a vocabulary of tags used in  $\Phi$  as  $W$ . For an image  $I \in \Phi$  and a tag  $w \in W$ , let  $r^*(w, I) : \{W, \Phi\} \mapsto \mathbf{R}$  be a tag relevance measurement. We call  $r^*(w, I)$  an ideal measurement for image and tag ranking if it satisfies the following two criteria:

**Criterion 1: Image ranking.** Given two images  $I_1, I_2 \in \Phi$  and tag  $w \in W$ , if  $w$  is relevant to  $I_1$  but irrelevant to  $I_2$ , then

$$r^*(w, I_1) > r^*(w, I_2). \quad (1)$$

**Criterion 2: Tag ranking.** Given two tags  $w_1, w_2 \in W$  and image  $I \in \Phi$ , if  $I$  is relevant to  $w_1$  but irrelevant to  $w_2$ , then

$$r^*(w_1, I) > r^*(w_2, I). \quad (2)$$

Our goal is to find a tag relevance measurement satisfying the two criteria.

### B. Learning Tag Relevance From Visual Neighbors

As aforementioned, given an image  $I$  labeled with a tag  $w$ , the occurrence frequency of  $w$  in visual neighbors of  $I$  to some extent reflects the relevance of  $w$  with respect to  $I$ . Note that the neighbors can be decomposed into two parts according to their relevance to  $w$ , i.e., images relevant and irrelevant to  $w$ . If we know how relevant and irrelevant images are labeled with  $w$  and how they are distributed in the neighbor set, we can estimate the tag's distribution in the neighbors.

To formalize the above notions, we first define a few notations as listed in Table I. We now study how images relevant and irrelevant to a tag are labeled with that tag. In a large user-tagged image database, it is plausible that for a specific tag  $w$ , the number of images irrelevant to the tag is significantly larger than the number of relevant images, i.e.,  $|R_w^c| \gg |R_w|$ , where  $|\bullet|$  is the cardinality operator on image sets. Moreover, one might expect that user tagging is better than tagging at random such that relevant images are more likely to be labeled, meaning  $|L_w \cap R_w| > |L_w \cap R_w^c|$ .

By approximating the probability of correct tagging  $P(w|R_w)$  using  $|L_w \cap R_w|/|R_w|$  and the probability of incorrect tagging  $P(w|R_w^c)$  using  $|L_w \cap R_w^c|/|R_w^c|$ , we have  $P(w|R_w) > P(w|R_w^c)$ . Hence, we make an assumption on user tagging behavior, that is:

**Assumption 1: User tagging.** In a large user-tagged image database, the probability of correct tagging is larger than the probability of incorrect tagging.

Next, we analyze the distribution of images relevant and irrelevant with respect to tag  $w$  in the  $k$  nearest neighbor set of image  $I$ . Compared to random sampling, a content-based visual search defined by a similarity function  $f$  can be viewed as a sampling process biased by the query image. We consider two situations with respect to the visual search accuracy, that is, equal to and better than random sampling. In the first situation where the visual search is equal to random sampling, the number of relevant images in the neighbor set is the same as the number of relevant

TABLE I  
MAIN NOTATIONS DEFINED IN THIS WORK

Notation	Definition
$\Phi$	a collection of user-tagged images.
$L_w$	$L_w \subset \Phi$ , all images labeled with tag $w$ in the collection.
$R_w$	$R_w \subset \Phi$ , all images relevant with respect to tag $w$ in the collection.
$R_w^c$	$R_w^c = \Phi \setminus R_w$ , all images irrelevant with respect to tag $w$ in the collection.
$P(w R_w)$	probability of correct tagging, i.e., an image randomly selected from $R_w$ is labeled with tag $w$ .
$P(w R_w^c)$	probability of incorrect tagging, i.e., an image randomly selected from $R_w^c$ is labeled with tag $w$ .
$P(R_w)$	probability that an image randomly selected from the entire collection is relevant to tag $w$ .
$P(R_w^c)$	probability that an image randomly selected from the entire collection is irrelevant to tag $w$ .
$f$	a similarity function between two images, measured on low-level visual features.
$N_f(I, k)$	$N_f(I, k) \subset \Phi$ , $k$ nearest neighbors ( $k$ -nn) of an image $I$ found in the collection by $f$ .
$N_{rand}(k)$	$N_{rand}(k) \subset \Phi$ , $k$ images randomly selected from the collection.
$n_w[\bullet]$	an operator counting the number of tag $w$ in any subset of the collection.

images in a set of  $k$  images randomly selected from the collection. While in the second situation where the visual search is better than random sampling, given two images  $I_1$  relevant to tag  $w$  and  $I_2$  irrelevant to  $w$ , we expect to have

$$|N_f(I_1, k) \cap R_w| > |N_{rand}(k) \cap R_w| > |N_f(I_2, k) \cap R_w|.$$

For instance, consider  $w$  to be “bridge”,  $I_1$  a bridge image, and  $I_2$  a non-bridge image. In this example,  $N_f(I_1, k)$  should contain more bridge images than  $N_{rand}(k)$ , while  $N_f(I_2, k)$  should contain less bridge images than  $N_{rand}(k)$ . Viewing random sampling as a baseline, we introduce an offset variable  $\varepsilon_{I,w}$  to indicate the visual search accuracy. In particular, we use  $(P(R_w) + \varepsilon_{I,w})$  to represent the probability that an image randomly selected from the neighbor set  $NN_f(I, k)$  is relevant with respect to  $w$ . Since an image is either relevant or irrelevant to  $w$ , we use  $(1 - (P(R_w) + \varepsilon_{I,w}))$ , namely  $(P(R_w^c) - \varepsilon_{I,w})$ , to represent the probability that an image randomly selected from  $NN_f(I, k)$  is irrelevant with respect to  $w$ . Then, the number of relevant images in the neighbor set is expressed as

$$|N_f(I, k) \cap R_w| = k \cdot (P(R_w) + \varepsilon_{I,w}) \quad (3)$$

and the number of irrelevant images in the neighbor set as

$$|N_f(I, k) \cap R_w^c| = k \cdot (P(R_w^c) - \varepsilon_{I,w}). \quad (4)$$

It is worth mentioning that the variable  $\varepsilon_{I,w}$  is introduced to help us derive important properties of the proposed algorithm. We do not rely on  $\varepsilon_{I,w}$  for implementing the algorithm.

Based on the above discussion, if the visual search is equal to random sampling, we have  $\varepsilon_{I,w} = 0$ . If the visual is better than random sampling, we have

$$\varepsilon_{I_1,w} > 0 > \varepsilon_{I_2,w}, \text{ for } I_1 \in R_w \text{ and } I_2 \in R_w^c. \quad (5)$$

We then make our second assumption as

**Assumption 2: Visual search.** A content-based visual search is better than random sampling.

Bearing the analysis of user tagging and visual search in mind, we now consider the distribution of tag  $w$  within the neighbor set of image  $I$ . Since we can divide the neighbor set into two distinct subsets  $N_f(I, k) \cap R_w$  and  $N_f(I, k) \cap R_w^c$ , we count the number of  $w$  in the two subsets, separately. That is,

$$\begin{aligned} n_w[N_f(I, k)] &= n_w[N_f(I, k) \cap R_w] + n_w[N_f(I, k) \cap R_w^c] \\ &= k \cdot (P(R_w) + \varepsilon_{I,w}) P(w|R_w) \\ &\quad + k \cdot (P(R_w^c) - \varepsilon_{I,w}) P(w|R_w^c). \end{aligned} \quad (6)$$

In a similar fashion, we derive

$$n_w[N_{rand}(k)] = k \cdot (P(R_w)P(w|R_w) + P(R_w^c)P(w|R_w^c)). \quad (7)$$

Since  $n_w[N_{rand}(k)]$  reflects the occurrence frequency of  $w$  in the entire collection, we denote it as  $Prior(w, k)$ . By substituting (7) into (6), we obtain

$$n_w[N_f(I, k)] - Prior(w, k) = k \cdot (P(w|R_w) - P(w|R_w^c)) \varepsilon_{I,w}. \quad (8)$$

Further, by defining

$$tagRelevance(w, I, k) := n_w[N_f(I, k)] - Prior(w, k) \quad (9)$$

we arrive at the following two theorems:

**Theorem 1: Image ranking.** Given assumption 1 and assumption 2,  $tagRelevance$  yields an ideal image ranking for tag  $w$ , that is, for  $I_1 \in R_w$  and  $I_2 \in R_w^c$ , we have  $tagRelevance(w, I_1) > tagRelevance(w, I_2)$ .

**Theorem 2: Tag ranking.** Given assumption 1 and assumption 2,  $tagRelevance$  yields an ideal tag ranking for image  $I$ , that is, for two tags  $w_1$  and  $w_2$ , if  $I \in R_{w_1}$  and  $I \in R_{w_2}^c$ , we have  $tagRelevance(w_1, I) > tagRelevance(w_2, I)$ .

We refer to the Appendix for detailed proofs of the two theorems. Note that in the proof of theorem 1, assumption 2 (5) can be relaxed as  $(\varepsilon_{I_1,w} > \varepsilon_{I_2,w})$  which we call relaxed assumption 2. Since the relaxed assumption is more likely to hold than its origin, this observation indicates that image ranking is relatively easier than tag ranking.

Our tag relevance function in (9) consists of two components which represents the distribution of the tag in the local neighborhood and in the entire collection, respectively. This observation confirms our conjecture made in the beginning of Section III that a good tag relevance measurement should take both distribution into account.

### C. Neighbor Voting Algorithm

We have argued in Section III-B that learning tag relevance boils down to computing  $(n_w[N_f(I, k)] - Prior(w, k))$ , i.e.,

the count of tag  $w$  in the  $k$  nearest neighbors of image  $I$  minus the prior frequency of  $w$ . Consider that each neighbor votes on  $w$  if it is labeled with  $w$  itself,  $n_w[N_f(I, k)]$  is then the count of neighbor votes on  $w$ . Thereby, we introduce a neighbor voting algorithm: given a user-tagged image, we first perform content-based  $k$ -nn search to find its visual neighbors, and then for each neighbor image, we use its tags to vote on tags of the given image. We approximate the prior frequency of tag  $w$  as

$$Prior(w, k) \approx k \frac{|L_w|}{|\Phi|} \quad (10)$$

where  $k$  is the number of visual neighbors,  $|L_w|$  the number of images labeled with  $w$ , and  $|\Phi|$  the size of the entire collection. Note that the function  $tagRelevance$  in (9) does not necessarily obtain positive results. We set the minimum value of  $tagRelevance$  to 1. In other words, if the learned tag relevance value of a user-contributed tag is less than its original frequency in an image, we reject the tag relevance learning result for that image. In addition, we observe that the voting result might be biased by individual users who have a number of visually similar images, as shown in Fig. 3(a). In order to make the voting decision more objective (which we target at), we introduce a unique-user constraint on the neighbor set. That is, each user has at most one image in the neighbor set per voting round. As shown in Fig. 3(b), with the unique-user constraint, we effectively reduce the voting bias. We finally summarize the procedure for learning tag relevance by neighbor voting in Algorithm 1.

---

#### Algorithm 1 Learning tag relevance by neighbor voting

---

**Input:** A user-tagged image  $I$ .

**Output:**  $tagRelevance(w, I, k)$ , i.e., the tag relevance value of each tag  $w$  in  $I$ .

Find  $k$  nearest visual neighbors of  $I$  from the collection with the unique-user constraint, i.e., a user has at most one image in the neighbor set.

**for** tag  $w$  in tags of  $I$  **do**

$$tagRelevance(w, I, k) = 0$$

**end for**

**for** image  $J$  in the neighbor set of  $I$  **do**

**for** tag  $w$  in  $(tags\_of\_J \cap tags\_of\_I)$  **do**

$$\begin{aligned} tagRelevance(w, I, k) &= \\ tagRelevance(w, I, k) &+ 1 \end{aligned}$$

**end for**

**end for**

$$tagRelevance(w, I, k) = tagRelevance(w, I, k) - Prior(w, k)$$

$$tagRelevance(w, I, k) = \max(tagRelevance(w, I, k), 1)$$


---





Fig. 3. **Tag relevance learning with the unique-user constraint.** The query example is the biggest image in the center of (a) and (b). The query is labeled with tag “tiger” by a user. Figure (a) shows visual neighbors without the unique-user constraint, namely standard content-based search. Since the neighbor set is dominated by images from few users, the tag relevance value of “tiger” voted by 1000 neighbors is 557. While in Figure (b), with the unique-user constraint, each user has at most one image in the neighbor set per voting round. The tag relevance value of “tiger” voted by 1000 neighbors is thus reduced to 6. The unique-user constraint makes the voting result more objective.

#### IV. EXPERIMENTAL SETUP

##### A. Experiments

We evaluate our tag relevance learning algorithm in both an image ranking scenario and a tag ranking scenario. For image ranking, we compare three tag-based image retrieval methods with and without tag relevance learning. For tag ranking, we demonstrate the potential of our algorithm in helping user tagging in two settings, namely, tag suggestion for labeled images and tag suggestion for unlabeled images. Specifically, we design the following three experiments.

*Experiment 1: Tag-Based Image Retrieval:* We employ a general tag-based retrieval framework widely used in existing systems such as Flickr and YouTube. We adopt OKAPI-BM25, a well-founded ranking function for text retrieval [42], as a baseline. Given a query  $q$  containing keywords  $\{w_1, \dots, w_n\}$ , the relevance score of an image  $I$  is computed as

$$score(q, I) = \sum_{w \in q} qtf(w)idf(w) \frac{tf(w) \cdot (k_1 + 1)}{tf(w) + k_1 \cdot \left(1 - b + b \frac{l_I}{l_{avg}}\right)} \quad (11)$$

where  $qtf(w)$  is the frequency of tag  $w$  in  $q$ ,  $tf(w)$  the frequency of  $w$  in the tags of  $I$ ,  $l_I$  the total number of tags of  $I$ , and  $l_{avg}$  the average value of  $l_I$  over the entire collection. The function  $idf(w)$  is calculated as  $\log(N - |L_w| + 0.5) / (|L_w| + 0.5)$ , where  $N$  is the number of images in the collection and  $|L_w|$  is the number of images labeled with  $w$ . By using learned tag relevance value as updated tag frequency in the ranking function, namely substituting  $tagRelevance(w, I, k)$  for  $tf(w)$  in (11), we investigate how our algorithm improves upon the baseline. We study the performance of the baseline method and our method, given various combinations of parameters. In total, there are three parameters to optimize. One is  $k$ , the number of neighbors for learning tag relevance. We choose  $k$  from  $\{100; 200; 500; 1000; 2000; 5000; 10\,000; 15\,000; 20\,000\}$ . The other two are  $b$  and  $k_1$  in OKAPI-BM25. The parameter  $b$  ( $0 \leq b \leq 1$ )

controls the normalization effect of document length. Here, document length is the number of tags in a labeled image. We let  $b$  range from 0 to 1 with interval 0.1. The variable  $k_1$  is a positive parameter for regularizing the impact of tag frequency. Since  $k_1$  does not affect ranking for single-word queries, we set  $k_1$  to 2, a common choice in text retrieval [42].

Considering that the OKAPI-BM25 ranking function originally aims for text retrieval and hence might not be optimal for tag-based image retrieval, we further compare with a recent achievement in web image retrieval by Jing and Baluja [28] (see details in Section II-B). As depicted in [28], there are two parameters to optimize: a dump factor  $d$  ( $d > 0.8$ ) controlling the restart probability of random walk and  $m$  the number of top ranked results in an initial list to calculate the prior probability. We try various parameter combinations, i.e.,  $d \in \{0.85; 0.90; 0.95\}$  and  $m \in \{5; 10; 20; 100; 1000\}$ .

*Experiment 2: Tag Suggestion for Labeled Images:* Given an image labeled with some tags, we aim for automated methods that accurately suggest new tags relevant to the image. We investigate how our algorithm improves upon a recent method by Sigurbjörnsson and Van Zwol [6] by introducing visual content information into the tag suggestion process. Similar to [6], we first find  $x$  candidate tags having the highest co-occurrence with the initial tags. For each candidate tag, we then compute its relevance score with respect to the image as follows:

$$score(c, I) = score(c, \mathbf{w}_I) \cdot \frac{\lambda}{\lambda + (rank_c - 1)} \quad (12)$$

where  $c$  is the candidate tag,  $I$  the image, and  $\mathbf{w}_I$  the set of initial tags. The function  $score(c, \mathbf{w}_I)$  computes a relevance score between the candidate tag and the initial tags. We adopt  $Vote^+$ , the best method in [6], as an implementation of the  $score$  function. The input  $rank_c$  is the position of tag  $c$  in the candidate tag list ranked by  $tagRelevance$  in descending order. The variable  $\lambda$  is a positive parameter for regularizing the effect of tag relevance learning. By optimizing the algorithm on the same training set

as used in [6], we determine the optimized setting of the two parameters  $x$  and  $\lambda$  as 17 and 20, respectively.

**Experiment 3: Tag Suggestion for Unlabeled Images:** We compare with two model-free approaches: a tag frequency (tf) approach by Torralba *et al.* [7] and an approach by Wang *et al.* [8] which re-weights the frequency of a tag by its inverse document frequency (tf-idf). For our algorithm, since no user-defined tags are available, we consider all tags in the vocabulary as candidates. We estimate *tagRelevance* for each candidate tag with respect to the unlabeled image, and then rank the tags in descending order by *tagRelevance*. We take care to make the comparison fair. First, since the baselines do not consider user information, we remove the unique-user constraint from our algorithm. Second, for all methods, we fix the number of the visual neighbors to 500, as suggested in [8]. Finally, for each method, we select the top five tags as a final suggestion for each test image.

In all the three experiments, we use *baseline* to represent the baseline methods, and *tagRelevance* for our method.

### B. Data Collections

We choose Flickr as a test case of user tagging. We downloaded images from Flickr by randomly generating photo ids as query seeds. By removing images having no tags and those failed to extract visual features, we obtain 3.5 million labeled images in total. The images are of medium size with maximum width or height fixed to 500 pixels. After Porter stemming, the number of distinct tags per image varies from 1 to 1230, with an average value of 5.4. The collection has 573 115 unique tags and 272 368 user ids.

Note that the image retrieval experiment studies how well images are ranked, while the two tag suggestion experiments focus on how well tags are ranked. Different targets result in two different evaluation sets, one for image retrieval and the other for tag suggestion.

**Evaluation Set for Image Retrieval:** We create a ground truth set as follows. We select 20 diverse visual concepts as queries. The queries are listed in Table II with visual examples in Fig. 4. As defined earlier, we consider a query concept and an image relevant if the concept is clearly visible in the image and we shall relate the concept to the visual content easily and consistently with common knowledge. Therefore, toys, cartoons, painting, and statues of the concept are treated as irrelevant. For each query, we randomly select 1000 examples from images labeled with the query in our 3.5 million Flickr collection, and relabel them according to our labeling criterion. We report user tagging accuracy of all 20 queries in Table II. For each query, we score its 1000 test images with the two baseline methods and the proposed algorithm, respectively. The images are then ranked in light of their relevance scores. If two images have the same score, they are ranked according to photo ids in descending order so that latest uploaded images are ranked ahead.

**Evaluation Set for Tag Suggestion:** To evaluate the performance of tag suggestion for labeled and unlabeled images, we adopt a ground truth set from [6], which is created by manually assessing the relevance of tags with respect to images. The set consists of 331 Flickr images, having no overlap with the 3.5 million collection. Since the relevance of tags “2005”, “2006”, and “2007” with respect to an image is quite subjective, we re-

TABLE II  
GROUND TRUTH STATISTICS FOR OUR IMAGE RETRIEVAL EXPERIMENT.  
EACH QUERY HAS 1000 MANUALLY LABELED EXAMPLES. USER TAGGING  
ACCURACY IS THE NUMBER OF RELEVANT IMAGES DIVIDED BY 1000

Query	3.5 million user-tagged images	
	Tag frequency	User tagging accuracy
airplane	15,231	0.447
beach	64,348	0.331
boat	25,385	0.424
bridge	25,197	0.762
bus	14,296	0.641
butterfly	8,476	0.701
car	37,614	0.548
cityscape	11,063	0.657
classroom	7,763	0.388
dog	52,981	0.764
flower	71,699	0.829
harbor	8,420	0.503
horse	27,008	0.736
kitchen	11,464	0.389
lion	8,509	0.326
mountain	36,844	0.502
rhino	4,929	0.346
sheep	3,603	0.525
street	40,772	0.426
tiger	8,214	0.224



Fig. 4. Visual examples of 20 queries in our image retrieval experiment.

move the three tags from the ground truth beforehand. Note that these tags might be predicted by tag suggestion methods. In that case, we consider the tags irrelevant. The number of tags per image in the evaluation set varies from 1 to 14, with an average value of 5.5. Examples of the ground truth are shown in Fig. 5. For experiment 2, we follow the same data partition as [6], that is, 131 images for training and the remaining 200 for testing. Since no training is required for all the three methods in experiment 3, we take the entire ground truth set (331 images in total) for testing.

### C. Evaluation Criteria

For image retrieval, images relevant with respect to user queries should be ranked as high as possible. Meanwhile,



Fig. 5. Multimedia examples of the ground truth for our tag suggestion experiments.

ranking quality of the whole list is important not only for user browsing, but also for applications using search results as a starting point. For tag suggestion, tags relevant with respect to user images should be ranked as high as possible. Moreover, the candidate tag list should be short such that users pick out relevant tags easily and efficiently. Therefore, we adopt the following two standard criteria to measure the different aspects of the performance. Given a ranked list of  $l$  instances where an instance is an image for image retrieval and a tag for tag suggestion, we measure.

**Precision at  $n$  ( $P@n$ ):** The proportion of relevant instances in the top  $n$  retrieved results, where  $n \leq l$ . For image retrieval, we report  $P@10$ ,  $P@20$ , and  $P@100$  for each query. For tag suggestion, we report  $P@1$  and  $P@5$ , averaged over all test images, as used in [6]. We consider a predicted tag relevant with respect to a test image if the tag is from the ground truth tags of the image. The Porter stemming is done before tag matching. Since we always predict five tags for each image, for those images having less than five ground truth tags, their  $P@5$  will be smaller than 1.

**Average Precision (AP):** AP measures ranking quality of the whole list. Since it is an approximation of the area under the precision-recall curve [43], AP is commonly considered as a good combination of precision and recall, e.g., [23], [26], and [33]. The AP value is calculated as  $(1/R) \sum_{i=1}^l (R_i/i) \delta_i$ , where  $R$  is the number of relevant instances in the list,  $R_i$  the number of relevant instances in the top  $i$  ranked instances,  $\delta_i = 1$  if the  $i$ th instance is relevant and 0 otherwise. To evaluate the overall performance, we use mean average precision (MAP), a common measurement in information retrieval. MAP is the mean value of the AP over all queries in the image retrieval experiment and all test images in the tag suggestion experiments.

#### D. Large-Scale Content-Based Visual Search

To implement the neighbor voting algorithm, we need to define visual similarity between images and then search visual neighbors in our 3.5 million Flickr photo database. Visual similarity between two images is measured using corresponding visual features. Since we need features relatively stable for search and efficient to compute to cope with millions of images, we adopt a combined 64-D global feature as a tradeoff between effectiveness and efficiency. The feature is calculated as follows. For each image, we extract 44-D color correlogram [44], 14-D color texture moment [45], and 6-D RGB color moment. We separately normalize the three features into unit length and concatenate them into a single vector. We use the Euclidean distance as a dissimilarity measurement. The feature is used throughout all the three experiments.

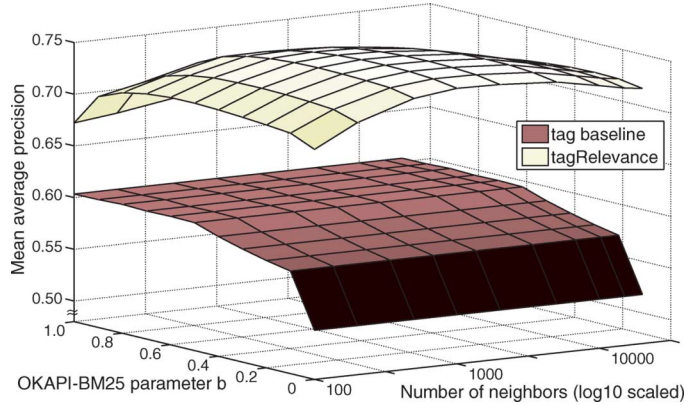


Fig. 6. **Experiment 1: An overall comparison between image retrieval methods with and without tag relevance learning.** The *tag baseline* method uses original tags, while our *tagRelevance* method uses learned tag relevance as updated tag frequency. We study the retrieval performance given various combinations of the OKAPI-BM25 parameter  $b$  and the number of neighbors for tag relevance learning. We measure the overall performance using mean average precision of the 20 queries from Fig. 4. The *tagRelevance* consistently outperforms the *tag baseline* for all parameter settings.

To search millions of images by content, efficient indexing methods are imperative for speed up. We adopt a  $K$ -means clustering-based method for its empirical success in large-scale content-based image retrieval [46]. First for indexing, we divide the whole dataset into smaller subsets by the  $K$ -means clustering. Each subset is indexed by a cluster center. Then for a query image, we find neighbors within fewer subsets whose centers are the closest to the query. The search space is thus reduced. Since the search operation in individual subsets can be executed in parallel, we execute neighbor search in a distributed super computer.

## V. RESULTS

### A. Experiment 1: Tag-Based Image Retrieval

As shown in Fig. 6, our *tagRelevance* substantially outperforms the *tag baseline* for all parameter settings. Recall that the OKAPI-BM25 parameter  $b$  controls the impact of normalizing scores by the total number of tags within an image. Hence, we observe different behavior of  $b$  in the two methods: the *tag baseline* tends to perform well when  $b$  approaches 1; in contrast, the *tagRelevance* improves as  $b$  approaches 0. Since tag frequency is not discriminative in original tagging, the baseline method heavily relies on the normalization factor. While in the new method, tag frequency becomes more discriminative after tag relevance learning.

The proposed algorithm is also robust to the number of neighbors used for voting. To show this property, we first run leave-one-out cross validation on the 20 queries to determine the optimized OKAPI-BM25 parameter  $b$  for *tag baseline* and our method, which is 0.8 and 0.3, respectively. As shown in Fig. 7, *tagRelevance* consistently outperforms *tag baseline*. More precisely, we reach at least 20% relative improvement in terms of MAP when the number of neighbors is between 200 and 20 000.

We conclude experiment 1 with a per-query comparison between three methods, namely *tag baseline*, *baseline* [28], and our *tagRelevance*. We again use the optimized parameters for



TABLE III

**EXPERIMENT 1: PER-QUERY COMPARISON BETWEEN IMAGE RETRIEVAL METHODS WITH AND WITHOUT TAG RELEVANCE LEARNING. BOLD NUMBERS INDICATE THE TOP PERFORMERS. FOR MOST OF THE 20 QUERIES, WE IMPROVE UPON THE BASELINE METHODS BY USING LEARNED TAG FREQUENCY AS UPDATED TAG FREQUENCY. ON AVERAGE, COMPARED WITH THE TAG BASELINE, WE OBTAIN A RELATIVE IMPROVEMENT IN TERMS OF P@20 BY 28.8% AND 24.3% IN TERMS OF MAP. COMPARED WITH THE BASELINE [28], WE OBTAIN A RELATIVE IMPROVEMENT IN TERMS OF P@20 BY 15.3% AND 19.9% IN TERMS OF MAP**

Query	Precision at 5			Precision at 20			Precision at 100			Average precision		
	tag baseline	baseline [28]	tagRelevance	tag baseline	baseline [28]	tagRelevance	tag baseline	baseline [28]	tagRelevance	tag baseline	baseline [28]	tagRelevance
airplane	0.400	<b>0.800</b>	0.600	0.500	<b>0.750</b>	0.400	<b>0.520</b>	<b>0.520</b>	0.510	0.446	0.513	<b>0.531</b>
beach	0.400	0.400	<b>1.000</b>	0.500	0.350	<b>0.900</b>	0.370	0.370	<b>0.710</b>	0.383	0.356	<b>0.666</b>
boat	0.400	0.200	<b>1.000</b>	0.600	0.550	<b>0.950</b>	0.520	0.520	<b>0.720</b>	0.477	0.487	<b>0.619</b>
bridge	<b>1.000</b>	0.800	0.800	<b>0.950</b>	0.900	0.900	0.880	0.880	<b>0.900</b>	0.802	0.806	<b>0.830</b>
bus	<b>1.000</b>	<b>1.000</b>	0.600	0.700	<b>0.850</b>	<b>0.850</b>	0.740	0.740	<b>0.870</b>	0.684	0.792	<b>0.836</b>
butterfly	0.800	0.800	<b>1.000</b>	0.800	0.900	<b>0.950</b>	0.940	0.940	<b>0.990</b>	0.816	0.838	<b>0.932</b>
car	<b>1.000</b>	<b>1.000</b>	0.800	0.650	0.800	<b>0.900</b>	0.660	0.660	<b>0.800</b>	0.610	0.674	<b>0.730</b>
cityscape	0.000	<b>1.000</b>	<b>1.000</b>	0.500	<b>0.950</b>	<b>0.950</b>	0.690	0.690	<b>0.980</b>	0.698	0.683	<b>0.907</b>
classroom	0.800	<b>1.000</b>	0.600	0.500	<b>0.900</b>	0.750	0.500	0.500	<b>0.600</b>	0.482	<b>0.551</b>	0.532
dog	<b>1.000</b>	0.800	<b>1.000</b>	0.950	0.950	0.950	0.830	0.830	<b>0.930</b>	0.806	0.820	<b>0.869</b>
flower	1.000	1.000	1.000	0.900	0.950	<b>1.000</b>	0.910	0.910	<b>0.980</b>	0.889	0.891	<b>0.963</b>
harbor	<b>0.800</b>	0.600	<b>0.800</b>	0.700	0.650	<b>0.950</b>	0.600	0.600	<b>0.900</b>	0.582	0.614	<b>0.768</b>
horse	0.800	<b>1.000</b>	<b>1.000</b>	0.550	0.950	<b>1.000</b>	0.700	0.700	<b>0.890</b>	0.718	0.774	<b>0.829</b>
kitchen	0.800	<b>1.000</b>	<b>1.000</b>	0.800	<b>0.900</b>	<b>0.900</b>	0.600	0.600	<b>0.900</b>	0.518	0.642	<b>0.742</b>
lion	0.800	0.800	<b>1.000</b>	0.950	0.450	<b>1.000</b>	0.420	0.420	<b>0.930</b>	0.476	0.393	<b>0.774</b>
mountain	0.600	0.400	<b>1.000</b>	0.500	0.650	<b>0.900</b>	0.500	0.500	<b>0.840</b>	0.517	0.550	<b>0.769</b>
rhino	1.000	1.000	1.000	0.950	<b>1.000</b>	0.950	0.820	0.820	<b>0.860</b>	0.697	0.659	<b>0.746</b>
sheep	<b>1.000</b>	<b>1.000</b>	0.800	0.850	0.900	<b>0.950</b>	0.790	0.790	<b>0.890</b>	0.638	0.677	<b>0.748</b>
street	0.400	0.400	<b>0.600</b>	0.300	0.500	<b>0.600</b>	0.390	0.390	<b>0.680</b>	0.412	0.477	<b>0.578</b>
tiger	0.400	0.800	<b>1.000</b>	0.550	0.450	<b>0.900</b>	0.610	0.610	<b>0.780</b>	0.442	0.338	<b>0.673</b>
average	0.720	0.790	<b>0.880</b>	0.685	0.765	<b>0.882</b>	0.649	0.649	<b>0.833</b>	0.605	0.627	<b>0.752</b>

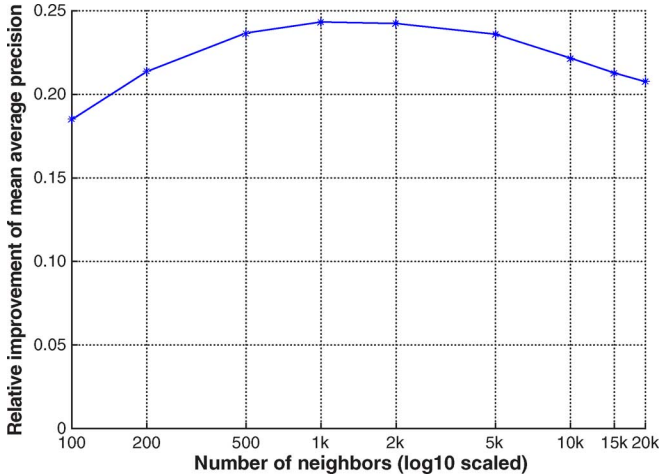


Fig. 7. Experiment 1: Relative improvement in terms of mean average precision (MAP) over the best tag baseline with respect to the number of neighbors for learning tag relevance. The best baseline is reached at  $b = 0.8$  with MAP 0.605. By using learned tag relevance value as updated tag frequency for retrieval, we obtain at least 20% relative improvement in terms of MAP when the number of neighbors is between 200 and 20 000.

tag baseline and tagRelevance. The number of neighbors is 1000. For baseline [28], we take the best output of tag baseline as initial search results and run leave-one-out cross validation to obtain an optimized parameter setting, i.e.,  $d = 0.85$  and  $m = 100$ . As shown in Table III, for some queries, baseline [28] is on a par with our tagRelevance, especially for the top ranked results. Nevertheless, for the majority of the queries and the evaluation metrics, the proposed algorithm compares favorably with the two baselines. On average, compared with the tag baseline, we obtain a relative improvement in terms of P@20 by 28.8% and 24.3% in terms of MAP. Compared with the baseline [28], we obtain a relative improvement in terms of P@20 by 15.3% and 19.9% in terms of MAP.

TABLE IV  
EXPERIMENT 2: TAG SUGGESTION FOR LABELED IMAGES.  
FOR EACH IMAGE, WE CHOOSE THE TOP FIVE RANKED TAGS.  
BOLD NUMBERS INDICATE THE TOP PERFORMERS

Evaluation criteria	Tag suggestion methods	
	baseline [6]	tagRelevance
Precision at 1	0.522	<b>0.555</b>
Precision at 5	0.359	<b>0.375</b>
Mean average precision	0.622	<b>0.663</b>

### B. Experiment 2: Tag Suggestion for Labeled Images

We report the performance of the two tag suggestion methods in Table IV. For all evaluation metrics, the tagRelevance improves upon the baseline. More precisely, we obtain an improvement of 6.3% in terms of P@1 and 6.6% in terms of MAP. While the improvement in terms of P@5 is 4.5%, which is relatively small. The reasons are two-fold. First, by measuring the relevance of a candidate tag with respect to an image at both textual and visual aspects, the tagRelevance is more likely to rank relevant tags ahead of irrelevant ones. Second, since we use the baseline as a starting point, if the method fails to retrieve relevant tags, it is unlikely to create a better ranked list. As shown in Table V, compared to the baseline, our method finds more relevant tags which describe visual aspects of an image.

### C. Experiment 3: Tag Suggestion for Unlabeled Images

As shown in Table VI, our tagRelevance method outperforms the two baseline methods for all evaluation criteria. Since the tf baseline [7] ranks tags in terms of tag frequency, it tends to suggest tags occurring frequently in the entire collection such as “2006”. By re-weighting tag frequency using the idf value, the tf-idf baseline [8] may restrain such effects to some extent. However, it risks over-weighting rare tags like “campcourtney”. By contrast, our tagRelevance uses the frequency of a tag minus

TABLE V  
EXPERIMENT 2: EXAMPLES OF TAG SUGGESTION FOR LABELED IMAGES BY DIFFERENT METHODS. THE *italic* FONT INDICATES RELEVANT TAGS AND THE **BOLD** FONT INDICATES UNIQUE RELEVANT TAGS PRODUCED BY OUR METHOD. WE IMPROVE UPON THE *baseline* BY ADDRESSING TAG RELEVANCE WITH RESPECT TO THE VISUAL CONTENT. COMPARED TO THE *baseline*, OUR METHOD FINDS MORE RELEVANT TAGS WHICH DESCRIBE VISUAL ASPECTS OF THE IMAGES



User-labeled images		New suggested tags	
Image	Tags	baseline [6]	tagRelevance
	lighthouse	beach sea ocean harbor 2005	sea beach ocean harbor <b><i>sunset</i></b>
	loch scotland lake waves	water castle <i>beach</i> katrine edinburgh	water <b><i>mountain</i></b> <i>beach</i> castle sea
	d40 london stonehenge uk bath	england sister nikon nikond40 stone	england sister water <b><i>street</i></b> stone
	mexico	2006 vacation new oaxaca honeymoon	2006 vacation <b><i>beach</i></b> new honeymoon




TABLE VI  
EXPERIMENT 3: TAG SUGGESTION FOR UNLABELED IMAGES. FOR EACH IMAGE, WE CHOOSE THE TOP FIVE RANKED TAGS. **BOLD NUMBERS** INDICATE THE TOP PERFORMERS

Evaluation criteria	Tag suggestion methods		
	baseline [7]	baseline [8]	tagRelevance
Precision at 1	0.061	0.068	<b>0.097</b>
Precision at 5	0.068	0.059	<b>0.074</b>
Mean average precision	0.126	0.120	<b>0.153</b>

its prior frequency to restrain high frequent tags. Meanwhile, since the prior frequency of the rare tags are small, these tags are not over-weighted. Hence, our method is more effective and robust.

Since all the three methods rely on the effectiveness of the visual search, we further study how the methods behave when the accuracy of the visual search is low ( $P@n < 0.05$ ), medium ( $0.05 \leq P@n \leq 0.20$ ), and high ( $P@n > 0.20$ ). As illustrated in Table VII, we select three test images, where the manually assessed accuracy of the 30 nearest neighbors is 0.77, 0.00, and 0.10, respectively. We observe that all methods succeed when the visual search is good. Obviously, all methods fail when no relevant images exist in the neighbor set. Interestingly, in an intermediate situation when the visual search is unsatisfactory with only a few relevant examples in the neighbor set, our method predicts more relevant tags than the two baseline methods. We make a further investigation on the entire test set. Since manually assessing the visual search accuracy for the test set is laborious, we estimate the accuracy as follows. For each test image with a number of ground truth tags, we consider a

TABLE VII  
EXPERIMENT 3: EXAMPLES OF TAG SUGGESTION FOR UNLABELED IMAGES BY DIFFERENT METHODS. THE *italic* FONT INDICATES RELEVANT TAGS AND THE **BOLD** FONT INDICATES UNIQUE RELEVANT TAGS PRODUCED BY OUR METHOD. WE ILLUSTRATE HOW THE THREE METHODS PERFORM WHEN THE ACCURACY OF THE VISUAL SEARCH IS HIGH (FOR THE IMAGE AT THE TOP), LOW (FOR THE IMAGE IN THE MIDDLE), OR MEDIUM (FOR THE IMAGE AT THE BOTTOM). COMPARED TO THE TWO BASELINE METHODS, OUR METHOD PREDICTS MORE RELEVANT TAGS EVEN WHEN THE VISUAL SEARCH IS UNSATISFACTORY

Visual Search		Suggested tags by different methods		
Image	Accuracy	baseline [7]	baseline [8]	tagRelevance
	0.77	<i>flower</i> <i>red</i> macro nature garden	<i>flower</i> <i>red</i> macro <i>rose</i> garden	<i>flower</i> <i>red</i> macro <i>rose</i> garden
	0.00	2006 family japan beach vacation	2006 cat family campcourtney august12006	icehockey hockey family hurricane cat
	0.10	2006 wedding japan park vacation	2006 pepperell wedding japan park	japan <b><i>bike</i></b> hiking park texas

neighbor image relevant if the tags of the neighbor image and the tags of the test image have at least one tag in common. It is in this way that we count relevant neighbors and subsequently compute the visual search accuracy. As shown in Fig. 8, our algorithm outperforms the baselines, given different visual search accuracy. In particular, our algorithm performs especially better when the visual search accuracy is medium or low. The evidence from both Table VII and Fig. 8 demonstrates the potential of our tag relevance learning algorithm. In addition, note that the majority of the test images have unsatisfactory visual search results (61.9% low and 35.5% medium), resulting in a relatively low performance for automatic image tagging. This observation implies that tag suggestion for unlabeled images can be improved further by including more advanced visual features.

#### D. Discussion

So far, we have verified the effectiveness of the proposed algorithm for tag-based image retrieval and automatic tag suggestion for labeled and unlabeled images. As discussed in Section III-B, since image ranking imposes a relatively looser requirement on content-based visual search than tag ranking, the former is easier than the latter. The empirical evidences from the three experiments confirm this conclusion. To better understand how our assumptions on visual search hold in practice, we introduce a validation experiment as follows. For each of the 20 queries used in the image retrieval experiment, we count the proportion of  $\langle \text{relevant image}, \text{irrelevant image} \rangle$  pairs that satisfy assumption 2 and relaxed assumption 2, respectively. Note that for other visual similarity functions in the literature, we can use this method to estimate how a particular visual similarity measurement meets the assumptions and consequently select proper features based on the estimation. As shown in the boxplot in Fig. 9, on average, 37.8% pairs satisfy

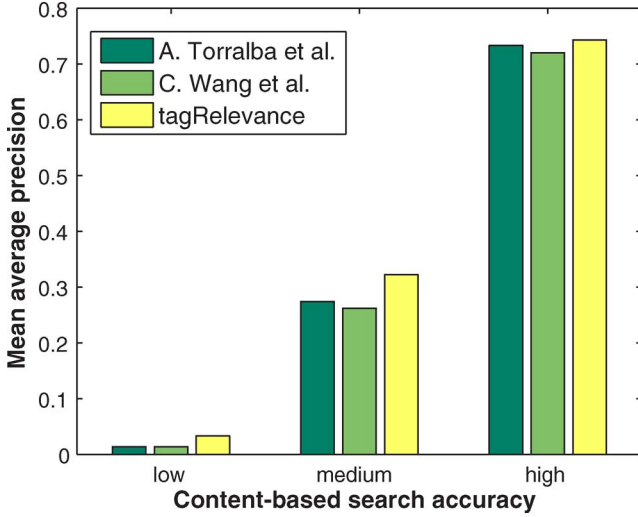


Fig. 8. **Experiment 3: The effect of content-based visual search on tag suggestion for unlabeled images.** We categorize the accuracy of a content-based visual search into three levels, that is, low (precision < 0.05), medium ( $0.05 \leq \text{precision} \leq 0.20$ ), and high (precision > 0.20). The proposed algorithm outperforms the baselines, given different levels of visual search accuracy. In particular, our algorithm performs especially better when the visual search accuracy is medium or low.

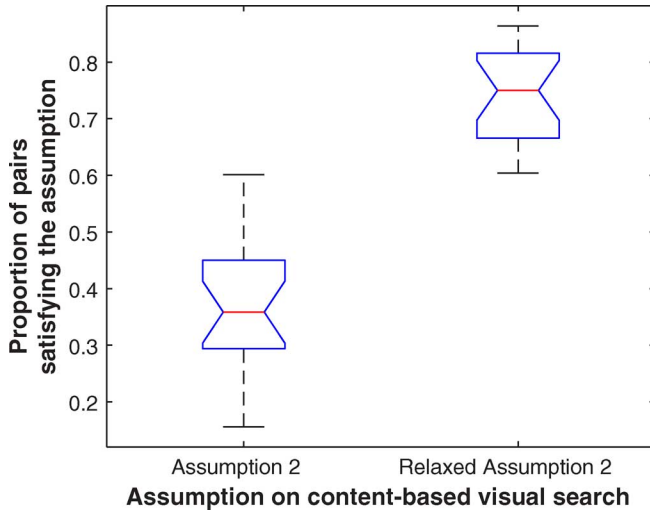


Fig. 9. **Validation of the two assumptions on content-based visual search.** We refer to Section III-B for the definitions of assumption 2 and relaxed assumption 2. For each of the 20 queries used in the image retrieval experiment, we count the proportion of  $\langle \text{relevantimage}, \text{irrelevantimage} \rangle$  pairs that satisfy assumption 2 and relaxed assumption 2, respectively. We use boxplot to visualize the results. On average, 37.8% pairs satisfy assumption 2 and 73.4% pairs satisfy relaxed assumption 2.

assumption 2 and 73.4% pairs satisfy relaxed assumption 2. The results again verify our conclusions that learned tag relevance is a good criterion for image ranking, and it can be improved further for tag ranking by leveraging more advanced visual features.

Up to now, we have successfully managed 3.5 million user-tagged images by executing our algorithm in parallel. Considering the heavy computation effort, however, it would be interesting to investigate in the future how to regularize the learning process, say from a Hill-climbing set, to ease the computation

for new user-submitted images. Though our evaluations are conducted on Flickr, the proposed algorithm is general. Hence, it is also applicable to other social photo sharing websites. Finally, we present in Fig. 10 some of the tag relevance learning results with updated tag frequency.

## VI. CONCLUSIONS

Since user tagging is known to be subjective and overly personalized, a fundamental problem in social image analysis and retrieval is how to accurately interpret the relevance of a tag with respect to the visual content the tag is describing. In this paper, we propose a neighbor voting algorithm as an initial step towards conquering the problem. Our key idea is to learn the relevance of a tag with respect to an image from tagging behaviors of visual neighbors of that image. In particular, our algorithm estimates tag relevance by counting neighbor votes on tags. We show that when 1) the probability of correct user tagging is larger than the probability of incorrect user tagging and 2) content-based visual search is better than random sampling, our algorithm produces a good tag relevance measurement for both image ranking and tag ranking. Moreover, since the proposed algorithm does not require any model training for any visual concept, it is efficient in handling large-scale image data sets.

To verify our algorithm, we conduct three experiments on 3.5 million Flickr photos: one image ranking experiment and two tag ranking experiments. For the image ranking experiment, we improve social image retrieval by using learned tag relevance as updated tag frequency in a general tag-based retrieval framework. Retrieval with tag relevance learning obtains a 24.3% relative improvement in terms of mean average precision, when compared to a tag-based retrieval baseline. For the tag ranking experiments, we consider two settings, i.e., tag suggestion for labeled images and tag suggestion for unlabeled images. In the tag suggestion experiment for labeled images, our algorithm finds more tags which describe visual aspects of an image, leading to a relative improvement of 6.3% in terms of mean average precision when compared to a text baseline. In the tag suggestion experiment for unlabeled images, our algorithm compares favorably against two baselines. Specifically, we effectively restrain high frequency tags without over-weighting rare tags. Our study demonstrates that the proposed algorithm predicts more relevant tags even when the visual search is unsatisfactory. In summary, all the three experiments show the general applicability of tag relevance learning for both image ranking and tag ranking. The results suggest a large potential of our algorithm for real-world applications.

## APPENDIX

In this section, we proof the two theorems introduced in Section III.

**Theorem 1: Image Ranking:** Given assumption 1 and assumption 2, *tagRelevance* yields an ideal image ranking for tag  $w$ , that is, for  $I_1 \in R_w$  and  $I_2 \in R_w^c$ , we have  $\text{tagRelevance}(w, I_1) > \text{tagRelevance}(w, I_2)$ .

**Proof:** Recall (8) and (9) that

$$\begin{aligned} \text{tagRelevance}(w, I_1) &= k \cdot (P(w|R_w) - P(w|R_w^c)) \varepsilon_{I_1, w} \\ \text{tagRelevance}(w, I_2) &= k \cdot (P(w|R_w) - P(w|R_w^c)) \varepsilon_{I_2, w} \end{aligned}$$





Fig. 10. **Results of learning tag relevance by neighbor voting.** The images are user-tagged photos from our 3.5 million Flickr collection. The texts on the right side of each image are user-contributed tags followed by estimated tag relevance value. The number of neighbors for tag relevance learning is 1000.

we have

$$\begin{aligned} & \text{tagRelevance}(w, I_1) - \text{tagRelevance}(w, I_2) \\ &= k \cdot (P(w|R_w) - P(w|R_w^c)) (\varepsilon_{I_1, w} - \varepsilon_{I_2, w}). \end{aligned}$$

Given assumption 1, we have

$$P(w|R_w) - P(w|R_w^c) > 0$$

and given assumption 2, we get

$$\varepsilon_{I_1, w} - \varepsilon_{I_2, w} > 0.$$

Hence,  $\text{tagRelevance}(w, I_1) > \text{tagRelevance}(w, I_2)$ . Note that we only require  $\varepsilon_{I_1, w} - \varepsilon_{I_2, w} > 0$ , thereby the assumption 2, namely  $\varepsilon_{I_1, w} > 0 > \varepsilon_{I_2, w}$ , can be relaxed as  $\varepsilon_{I_1, w} > \varepsilon_{I_2, w}$ . We call the latter relaxed assumption 2.  $\square$

**Theorem 2: Tag Ranking:** Given assumption 1 and assumption 2,  $\text{tagRelevance}$  yields an ideal tag ranking for image  $I$ , that is, for two tags  $w_1$  and  $w_2$ , if  $I \in R_{w_1}$  and  $I \in R_{w_2}^c$ , we have  $\text{tagRelevance}(w_1, I) > \text{tagRelevance}(w_2, I)$ .

**Proof:** Recall (8) and (9) that

$$\begin{aligned} \text{tagRelevance}(w_1, I) &= k \cdot (P(w_1|R_{w_1}) - P(w_1|R_{w_1}^c)) \varepsilon_{I, w_1} \\ \text{tagRelevance}(w_2, I) &= k \cdot (P(w_2|R_{w_2}) - P(w_2|R_{w_2}^c)) \varepsilon_{I, w_2}. \end{aligned}$$

Given assumption 1, we have

$$\begin{aligned} P(w_1|R_{w_1}) - P(w_1|R_{w_1}^c) &> 0 \\ P(w_2|R_{w_2}) - P(w_2|R_{w_2}^c) &> 0 \end{aligned}$$

and given assumption 2, we get

$$\varepsilon_{I, w_1} > 0 > \varepsilon_{I, w_2}.$$

Note that multiplying positive factors does not change the direction of an inequation. Therefore, by multiplying the left side and the right side of the above inequation by  $k(P(w_1|R_{w_1}) - P(w_1|R_{w_1}^c))$  and  $k(P(w_2|R_{w_2}) - P(w_2|R_{w_2}^c))$ , respectively, we obtain

$$\begin{aligned} k \cdot (P(w_1|R_{w_1}) - P(w_1|R_{w_1}^c)) \varepsilon_{I, w_1} &> 0 > \\ k \cdot (P(w_2|R_{w_2}) - P(w_2|R_{w_2}^c)) \varepsilon_{I, w_2}. \end{aligned}$$

Hence,  $\text{tagRelevance}(w_1, I) > \text{tagRelevance}(w_2, I)$ .  $\square$

## ACKNOWLEDGMENT

The authors would like to thank B. Sigurbjörnsson and R. van Zwol for their ground truth used in our tag suggestion experiments. The authors also would like to thank A. Setz for his contributions in creating the ground truth for our image retrieval experiment.

## REFERENCES

- [1] E. Auchard, "Flickr to map the world's latest photo hotspots," Reuters, Nov. 2007. [Online]. Available: <http://www.reuters.com/article/technologyNews/idUSHO94233920071119?sp=true>.
- [2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Internet Measurement*, 2007, pp. 1–14.
- [3] D. A. Shamma, R. Shaw, P. L. Shafon, and Y. Liu, "Watch what I watch: Using community activity to understand content," in *Proc. ACM MIR*, 2007, pp. 275–284.
- [4] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," in *Proc. ACM Multimedia*, 2007, pp. 631–640.
- [5] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," in *Proc. ACM Multimedia*, 2008, pp. 31–40.
- [6] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. WWW*, 2008, pp. 327–336.
- [7] 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, 2008.
- [8] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Scalable search-based image annotation," *Multimedia Syst.*, vol. 14, no. 4, pp. 205–220, 2008.
- [9] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Inf. Sci.*, vol. 32, no. 2, pp. 198–208, 2006.
- [10] K. K. Matusiak, "Towards user-centered indexing in digital image collections," *OLC Syst. Serv.*, vol. 22, no. 4, pp. 283–298, 2006.
- [11] 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.
- [12] 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.
- [13] 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, 2003.
- [14] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, 2008.



- [15] A. W. M. 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, 2000.
- [16] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & Wordnet," in *Proc. ACM Multimedia*, 2005, pp. 706–715.
- [17] K. Weinberger, M. Slaney, and R. van Zwol, "Resolving tag ambiguity," in *Proc. ACM Multimedia*, 2008, pp. 111–119.
- [18] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using SVM," in *Proc. SPIE*, 2004, pp. 330–338.
- [19] 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, 2007.
- [20] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Tagging over time: Real-world image annotation by lightweight meta-learning," in *Proc. ACM Multimedia*, 2007, pp. 393–402.
- [21] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, 2008.
- [22] X. Li, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proc. ICASSP*, 2009, pp. 3717–3720.
- [23] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, 2008.
- [24] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. CIVR*, 2003, pp. 649–654.
- [25] G. Park, Y. Baek, and H.-K. Lee, "Majority based ranking approach in web image retrieval," in *Proc. CIVR*, 2003, pp. 499–504.
- [26] W. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Multimedia*, 2006, pp. 35–44.
- [27] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for Google images," in *Proc. ECCV*, 2004, pp. 242–256.
- [28] Y. Jing and S. Baluja, "VisualRank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, 2008.
- [29] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proc. CIVR*, 2007, pp. 17–24.
- [30] W.-H. Lin and A. Hauptmann, "Web image retrieval re-ranking with relevance model," in *Proc. Web Intelligence*, 2003, pp. 242–248.
- [31] T.-S. Chua, S.-Y. Neo, K.-Y. Li, G. Wang, R. Shi, M. Zhao, and H. Xu, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *Proc. TRECVID Workshop*, 2004.
- [32] G. Begelman, P. Keller, and F. Smadja, "Automated tag clustering: Improving search and exploration in the tag space," in *Proc. WWW Collaborative Web Tagging Workshop*, 2006.
- [33] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proc. ACM Multimedia*, 2007, pp. 991–1000.
- [34] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, 2008.
- [35] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley, 1992.
- [36] F. F.-H. Nah, "A study on tolerable waiting time: How long are Web users willing to wait?," *J. Beh. Inf. Technol.*, vol. 23, no. 3, pp. 153–163, 2004.
- [37] B. Billerbeck and J. Zobel, "Questioning query expansion: An examination of behaviour and parameters," in *Proc. Australasian Database Conf.*, 2004, pp. 69–76.
- [38] R. Datta, W. Ge, J. Li, and J. Z. Wang, "Toward bridging the annotation-retrieval gap in image search," *IEEE Multimedia*, vol. 14, no. 3, pp. 24–35, 2007.
- [39] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval?," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 958–966, 2007.
- [40] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.
- [41] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, 2007.
- [42] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments—Part 2," *J. Inf. Process. Manage.*, vol. 36, no. 6, pp. 809–840, 2000.
- [43] M. Zhu, Recall, Precision and Average Precision, University of Waterloo, Waterloo, ON, Canada, working Paper 2004-09, 2004, Tech. Rep.

- [44] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. CVPR*, 1997, pp. 762–768.
- [45] H. Yu, M. Li, H. Zhang, and J. Feng, "Color texture moment for content-based image retrieval," in *Proc. ICIP*, 2002, pp. 929–932.
- [46] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, "Image annotation by large-scale content-based image retrieval," in *Proc. ACM Multimedia*, 2006, pp. 607–610.



**Xirong Li** (S'09) received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University, Beijing, China, in 2005 and 2007, respectively. Since August 2007, he has been pursuing the Ph.D. degree at the Intelligent Systems Lab Amsterdam, University of Amsterdam, Amsterdam, The Netherlands.

His research focuses on social multimedia analysis and retrieval.



**Cees G. M. Snoek** (S'01–M'06) received the M.Sc. degree in business information systems and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2000 and 2005, respectively.

He is currently a Senior Researcher at the Intelligent Systems Lab Amsterdam. He was a Visiting Scientist at Informedia, Carnegie Mellon University, Pittsburgh, PA, in 2003. His research interests focus on multimedia signal processing and analysis, statistical pattern recognition, content-based information retrieval, social media retrieval, and large-scale benchmark evaluations, especially when applied in combination for video retrieval. He has published over 70 refereed book chapters, journal, and conference papers in these fields, and he serves on the program committee of several conferences. He is a lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is co-initiator and co-organizer of the annual VideOlympics, and he was the local chair of the 2007 ACM International Conference on Image and Video Retrieval. He is a lecturer of post-doctoral courses given at international conferences and European summer schools.

Dr. Snoek received a young talent (VENI) grant from the Netherlands Organization for Scientific Research in 2008.



**Marcel Worring** (M'03) received the M.Sc. degree (honors) in computer science from the Vrije Universiteit, Amsterdam, The Netherlands, in 1988 and the Ph.D. degree in computer science from the University of Amsterdam in 1993.

He is currently an Associate Professor in the Intelligent Systems Lab Amsterdam of the University of Amsterdam. His interests are in multimedia search and systems. With the MediaMill team, he has been developing techniques for semantic video indexing as well as systems for interactively searching large video archives which have been successful over the last years in the TRECVID benchmark, the de-facto standard on the topic. The methodologies developed are now being applied to visual search in large video archives of news and other broadcast data as well as in the field of forensic intelligence in particular for fighting child abuse and surveillance. He has published over 100 scientific papers covering a broad range of topics from low-level image and video analysis up to applied papers in interactive search. He serves on the program committee of the major conferences in the field.

Dr. Worring was the chair of the IAPR TC12 on Multimedia and Visual Information Systems and general co-chair of the 2007 ACM International Conference on Image and Video Retrieval in Amsterdam, co-organizer of the first and second VideOlympics, a real-time evaluation of video retrieval systems, and short program chair for the ACM Multimedia 2009. He is an associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA and of *Pattern Analysis and Applications*.