Socializing the Semantic Gap: A Comparative Survey on Image Tag Assignment, Refinement, and Retrieval

XIRONG LI, Renmin University of China
TIBERIO URICCHIO, University of Florence
LAMBERTO BALLAN, University of Florence, Stanford University
MARCO BERTINI, University of Florence
CEES G. M. SNOEK, University of Amsterdam, Qualcomm Research Netherlands
ALBERTO DEL BIMBO, University of Florence

Where previous reviews on content-based image retrieval emphasize what can be seen in an image to bridge the semantic gap, this survey considers what people tag about an image. A comprehensive treatise of three closely linked problems (i.e., image tag assignment, refinement, and tag-based image retrieval) is presented. While existing works vary in terms of their targeted tasks and methodology, they rely on the key functionality of tag relevance, that is, estimating the relevance of a specific tag with respect to the visual content of a given image and its social context. By analyzing what information a specific method exploits to construct its tag relevance function and how such information is exploited, this article introduces a two-dimensional taxonomy to structure the growing literature, understand the ingredients of the main works, clarify their connections and difference, and recognize their merits and limitations. For a head-to-head comparison with the state of the art, a new experimental protocol is presented, with training sets containing 10,000, 100,000, and 1 million images, and an evaluation on three test sets, contributed by various research groups. Eleven representative works are implemented and evaluated. Putting all this together, the survey aims to provide an overview of the past and foster progress for the near future.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms: Algorithms, Documentation, Performance

Additional Key Words and Phrases: Social media, social tagging, tag relevance, content-based image retrieval, tag assignment, tag refinement, tag retrieval

14

X. Li and T. Uricchio both contributed equally and are corresponding authors.

This research was supported by NSFC (No. 61303184), SRFDP (No. 20130004120006), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01, No. 16XNQ013), SRF for ROCS, SEM, the Dutch national program COMMIT, the STW STORY project, Telecom Italia PhD grant funds, and the AQUIS-CH project granted by the Tuscany Region (Italy). L. Ballan also acknowledges the support of the EC's FP7 under grant agreement No. 623930 (Marie Curie IOF).

Authors' addresses: X. Li, Key Lab of Data Engineering and Knowledge Engineering, School of Information, Renmin University of China, No. 59 Zhongguancun Street, 100872 Beijing, China; email: xirong@ruc.edu.cn; C. G. M. Snoek, Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 904, Netherlands; email: cgmsnoek@uva.nl; T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo, Media Integration and Communication Center, University of Florence, Viale Morgagni 65, 50139 Firenze, Italy; emails: tiberio.uricchio@unifi.it, lballan@cs.stanford.edu, marco.bertini@unifi.it, alberto.delbimbo@unifi.it.

DOI: http://dx.doi.org/10.1145/2906152

ACM Reference Format:
Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Comput. Surv. 49, 1, Article 14 (June 2016), 39 pages.
DOI: http://dx.doi.org/10.1145/2906152

1. INTRODUCTION

Images want to be shared. Be it a drawing carved in rock, a painting exposed in a museum, or a photo capturing a special moment, it is the sharing that relives the experience stored in the image. Nowadays, several technological developments have spurred the sharing of images in unprecedented volumes. The first is the ease with which images can be captured in a digital format by cameras, cell phones, and other wearable sensory devices. The second is the Internet, which allows transfer of digital image content to anyone, anywhere in the world. Finally, and most recently, the sharing of digital imagery has reached new heights by the massive adoption of social network platforms. All of a sudden images come with tags. Tagging, commenting, and rating of any digital image have become a common habit. As a result, we observe a downpour of personally annotated user-generated visual content and associated metadata. The problem of image retrieval has been dilated with the problem of searching images generated within social platforms and improving social media annotations in order to permit effective retrieval.

Excellent surveys on content-based image retrieval have been published in the past. In their seminal work, Smeulders et al. review the early years up to the year 2000 by focusing on what can be seen in an image and introducing the main scientific problem of the field: the semantic gap as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [Smeulders et al. 2000]. Datta et al. continue along this line and describe the coming of age of the field, highlighting the key theoretical and empirical contributions of recent years [Datta et al. 2008]. These reviews completely ignore social platforms and socially generated images, which is not surprising as the phenomenon only became apparent after these reviews were published.

In this article, we survey the state of the art of content-based image retrieval in the context of social image platforms and tagging, with a comprehensive treatise of the closely linked problems of image tag assignment, image tag refinement, and tag-based image retrieval. Similar to Smeulders et al. [2000] and Datta et al. [2008], the focus of our survey is on visual information, but we explicitly take into account *and* quantify the value of social tagging.

1.1. Problems and Tasks

Social image tags are provided by common users. They often cannot meet high-quality standards related to content association, in particular for accurately describing objective aspects of the visual content according to some expert's opinion [Dodge et al. 2012]. Social tags tend to follow context, trends, and events in the real world. They are often used to describe both the situation and the entity represented in the visual content. In such a context, there are distinct problems to solve. On the one hand, social tags tend to be imprecise, ambiguous, and incomplete. On the other hand, they are biased toward personal perspectives. So tagging deviations due to spatial and temporal correlation to external factors are common phenomena [Golder and Huberman 2006; Sen et al. 2006; Sigurbjörnsson and Van Zwol 2008; Kennedy et al. 2006]. The focus of interests and motivations of an image retriever could be different from those of an image uploader.

Quite a few researchers have proposed solutions for image annotation and retrieval in social frameworks, although the peculiarities of this domain have been only

partially addressed. Concerning the role of visual content in social image tagging, several studies have shown that people are willing to tag objects and scenes presented in the visual content to favor image retrieval for the general audience [Ames and Naaman 2007; Sigurbjörnsson and Van Zwol 2008; Nov and Ye 2010]. It would be relevant to survey why people search images on social media platforms and what query terms they actually use. Although some query log data of a generic web image search have been made publicly accessible [Hua et al. 2013], its social media counterpart remains to be established. Most of the existing works have rather investigated the technological possibilities to automatically assign, refine, and enrich image tags. They mainly concentrated on how to expand the set of tags provided by the uploader by looking at tags that others have associated to similar content, and so expecting to include tags suited to the retriever's motivations. Consequently, images will become findable and potentially appreciated by a wider range of audiences beyond the relatively small social circle of the image uploader. We categorize these existing works into three different main tasks and structure our survey along these tasks:

—**Tag Assignment.** Given an unlabeled image, tag assignment strives to assign a (fixed) number of tags related to the image content [Makadia et al. 2010; Guillaumin et al. 2009; Verbeek et al. 2010; Tang et al. 2011].
—**Tag Refinement.** Given an image associated with some initial tags, tag refinement aims to remove irrelevant tags from the initial tag list and enrich it with novel, yet relevant, tags [Li et al. 2010; Wu et al. 2013; Znaidia et al. 2013; Lin et al. 2013; Feng et al. 2014].
—**Tag Retrieval.** Given a tag and a collection of images labeled with the tag (and possibly other tags), the goal of tag retrieval is to retrieve images relevant with respect to the tag of interest [Li et al. 2009b; Duan et al. 2011; Sun et al. 2011; Gao et al. 2013; Wu et al. 2013].

Other related tasks such as tag filtering [Zhu et al. 2010; Liu et al. 2011b; Zhu et al. 2012] and tag suggestion [Sigurbjörnsson and Van Zwol 2008; Li et al. 2009b; Wu et al. 2009] have also been studied. We view them as variants of tag refinement.

As a common factor in all the works for tag assignment, refinement, and retrieval, we reckon that the way in which the tag set expansion is performed relies on the key functionality of *tag relevance*, that is, estimating the relevance of a tag with respect to the visual content of a given image and its social context.

1.2. Scope, Aims, and Organization

We survey papers that learn tag relevance from images tagged in social contexts. While it would have been important to consider the complementarity of tags, only a few methods have considered multitag retrieval [Li et al. 2012; Nie et al. 2012; Borth et al. 2013]. Hence, we focus on methods that implement the unique-tag relevance model. We do not cover traditional image classification that is grounded on carefully labeled data. For a state-of-the-art overview in that direction, we refer the interested reader to Everingham et al. [2015] and Russakovsky et al. [2015]. Nonetheless, one may question the necessity of using socially tagged examples as training data, given that a number of labeled resources are already publicly accessible. An exemplar of such resources is ImageNet [Deng et al. 2009], providing crowdsourced positive examples for over 20,000 classes. Since ImageNet employs several web image search engines to obtain candidate images, its positive examples tend to be biased by the search results. As observed by Vreeswijk et al. [2012], the positive set of vehicles mainly consists of cars and buses, although vehicles can be tracks, watercraft, and aircraft. Moreover, controversial images are discarded upon vote disagreement during the crowdsourcing.

All this reduces diversity in visual appearance. We empirically show in Section 5.4 the advantage of socially tagged examples against ImageNet for tag relevance learning.

Reviews on social tagging exist. The work by Gupta et al. [2010] discusses papers on why people tag, what influences the choice of tags, and how to model the tagging process, but its discussion on content-based image tagging is limited. The focus of Jabeen et al. [2016] is on papers about adding semantics to tags by exploiting varied knowledge sources such as Wikipedia, DBpedia, and WordNet. Again, it leaves the visual information untouched.

Several reviews that consider socially tagged images have appeared recently. In Liu et al. [2011], technical achievements in content-based tag processing for social images are briefly surveyed. Sawant et al. [2011], Wang et al. [2012], and Mei et al. [2014] present extended reviews of particular aspects, that is, collaborative media annotation, assistive tagging, and visual search reranking, respectively. In Sawant et al. [2011], papers that propose collaborative image labeling games and tagging in social media networks are reviewed. Wang et al. [2012] survey papers where computers assist humans in tagging either by organizing data for manual labeling, by improving the quality of human-provided tags, or by recommending tags for manual selection, instead of applying purely automatic tagging. Mei et al. [2014] review techniques that aim to improve initial search results, typically returned by a text-based visual search engine, by visual search reranking. These reviews offer resumes of the methods and interesting insights on particular aspects of the domain, without giving an experimental comparison between the varied methods.

We notice efforts in empirical evaluations of social media annotation and retrieval [Sun et al. 2011; Uricchio et al. 2013; Ballan et al. 2015]. Sun et al. [2011] analyze different dimensions to compute the relevance score between a tagged image and a tag. They evaluate varied combinations of these dimensions for tag-based image retrieval on NUS-WIDE, a leading benchmark set for social image retrieval [Chua et al. 2009]. However, their evaluation focuses only on tag-based image ranking features, without comparing content-based methods. Moreover, tag assignment and refinement are not covered. Uricchio et al. [2013] and Ballan et al. [2015] compared three algorithms for tag refinement on the NUS-WIDE and MIRFlickr, a popular benchmark set for tag assignment and refinement [Huiskes et al. 2010]. However, the two reviews lack a thorough comparison between different methods under the umbrella of a common experimental protocol. Moreover, they fail to assess the high-level connection between image tag assignment, refinement, and retrieval.

The aims of this survey are twofold. First, we organize the rich literature in a taxonomy to highlight the ingredients of the main works in the literature and recognize their advantages and limitations. In particular, we structure our survey along the line of understanding how a specific method constructs the underlying tag relevance function. Witnessing the absence of a thorough empirical comparison in the literature, our second goal is to establish a common experimental protocol and successively exert it in the evaluation of key methods. Our proposed protocol contains training data of varied scales extracted from social frameworks. This permits us to evaluate the methods under analysis with data that reflect the specificity of the social domain. We have made the data and source code public[1] so that new proposals for tag assignment, tag refinement, and tag retrieval can be evaluated rigorously and easily. Taken together, these efforts should provide an overview of the field's past and foster progress for the near future.

The rest of the survey is organized as follows. Section 2 introduces a taxonomy to structure the literature on tag relevance learning. Section 3 proposes a new

[1]https://github.com/li-xirong/jingwei.

experimental protocol for evaluating the three tasks. A selected set of 11 representative works, described in Section 4, is compared extensively using this protocol, with results and analysis provided in Section 5. We provide concluding remarks and our vision about future directions in Section 6.

2. TAXONOMY AND REVIEW

2.1. Foundations

Our key observation is that the essential component, which measures the relevance between a given image and a specific tag, stands at the heart of the three tasks. In order to describe this component in a more formal way, we first introduce some notation.

We use x, t, and u to represent three basic elements in social images, namely, image, tag, and user. An image x is shared on social media by its user u. A user u can choose a specific tag t to label x. By sharing and tagging images, a set of users $\mathcal{U}$ contribute a set of n socially tagged images $\mathcal{X}$, wherein $\mathcal{X}_t$ denotes the set of images tagged with t. Tags used to describe the image set form a vocabulary of m tags $\mathcal{V}$. The relationship between images and tags can be represented by an image-tag association matrix $D \in \{0, 1\}^{n \times m}$, where $D_{ij} = 1$ means the ith image is labeled with the jth tag, and 0 otherwise.

Given an image and a tag, we introduce a real-valued function that computes the relevance between x and t based on the visual content and an optional set of user information Θ associated with the image:

$$f_\Phi(x, t; \Theta).$$

We use Θ in a broad sense, making it refer to any type of social context provided by or referring to the user like associated tags, where and when the image was taken, personal profile, and contacts. The subscript Φ specifies how the tag relevance function is constructed.

Having $f_\Phi(x, t; \Theta)$ defined, we can easily interpret each of the three tasks. Assignment and refinement can be done by sorting $\mathcal{V}$ in descending order by $f_\Phi(x, t; \Theta)$, while retrieval can be achieved by sorting the labeled image set $\mathcal{X}_t$ in descending order in terms of $f_\Phi(x, t; \Theta)$. Note that this formalization does not necessarily imply that the same implementation of tag relevance is applied for all three tasks. For example, for retrieval, relevance is intended to obtain image ranking [Li 2016], while tag ranking for each single image is the goal of assignment [Wu et al. 2009] and refinement [Qian et al. 2014].

Figure 1 presents a unified framework, illustrating the main data flow of varied approaches to tag relevance learning. Compared to traditional methods that rely on expert-labeled examples, a novel characteristic of a social-media-based method is its capability to learn from socially tagged examples with unreliable and personalized annotations. Such a training media is marked as $\mathcal{S}$ in the framework and includes tags, images, or user-related information. Optionally, in order to obtain a refined training media $\hat{\mathcal{S}}$, one might consider designing a filter to remove unwanted data. In addition, prior information such as tag statistics, tag correlations, and image affinities in the training media are independent of a specific image-tag pair. They can be precomputed for the sake of efficiency. As the filter and the precomputation appear to be a choice of implementation, they are positioned as auxiliary components in Figure 1.

A number of implementations of the relevance function have been proposed that utilize different modes to expand the tag set by learning within the social context. They may exploit different media, such as tags only, tags and related image content, or tags, image content, and user-related information. Depending on how $f_\Phi(x, t; \Theta)$ is composed internally, we propose a taxonomy that organizes existing works along two dimensions, namely, *media* and *learning*. The media dimension characterizes *what*

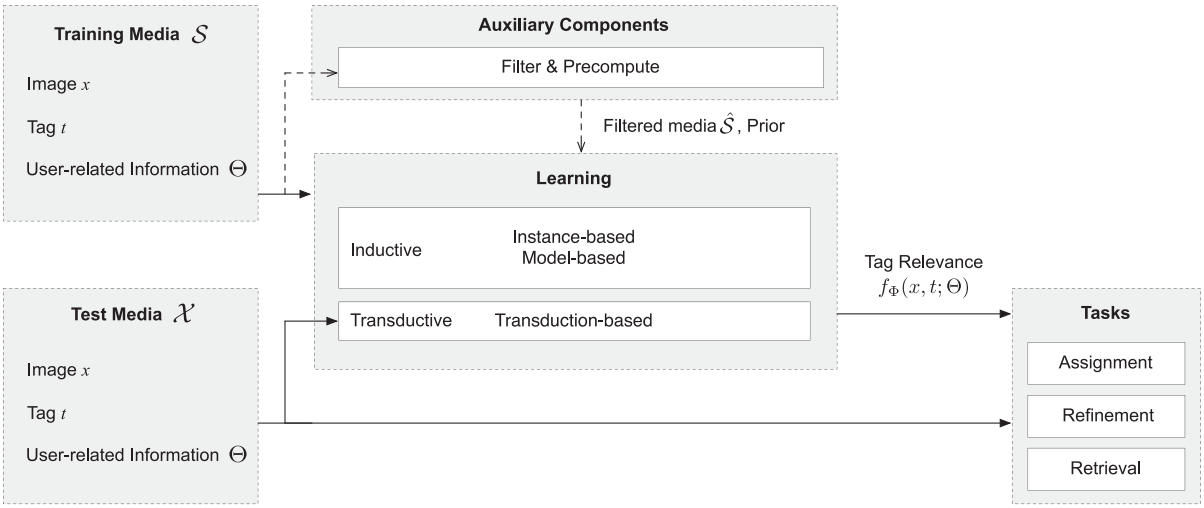


Fig. 1. **Unified framework of tag relevance learning for image tag assignment, refinement, and retrieval.** We follow the input data as it flows through the process of learning the tag relevance function $f_\Phi(x, t; \Theta)$ to higher-level tasks. Dashed lines indicate optional data flow. The framework jointly classifies existing works on assignment, refinement, and retrieval while at the same determining their main components.

essential information $f_\Phi(x, t; \Theta)$ exploits, while the learning dimension depicts *how* such information is exploited. Table I presents a list of the most significant contributions organized along these two dimensions. For a specific work, while Figure 1 helps illustrate the main data flow of the method, its position in the two-dimensional taxonomy is pinpointed via Table I. We believe such a context provides a good starting point for an in-depth understanding of the work. We explore the taxonomy along the media dimension in Section 2.2 and the learning dimension in Section 2.3. Auxiliary components are addressed in Section 2.4. A comparative evaluation of a few representative methods is presented in Section 4.

2.2. Media for Tag Relevance

Different sources of information may play a role in determining the relevance between an image and a social tag. For instance, the position of a tag appearing in the tag list might reflect a user's tagging priority to some extent [Sun et al. 2011]. Knowing what other tags are assigned to the image [Zhu et al. 2012] or what other users label about similar images [Li et al. 2009b; Kennedy et al. 2009] can also be helpful for judging whether the tag under examination is appropriate or not. Depending on what modalities in $\mathcal{S}$ are utilized, we divide existing works into the following three groups: (1) tag based, (2) tag + image based, and (3) tag + image + user information based, ordered in light of the amount of information they utilize. Table I shows this classification for several papers that appeared in the literature on the subject.

2.2.1. Tag Based. These methods build $f_\Phi(x, t; \Theta)$ purely based on tag information. The basic idea is to assign higher relevance scores to tags that are semantically close to the majority of the tags associated with the test image. To that end, in Sigurbjörnsson and Van Zwol [2008] and Zhu et al. [2012], relevant tags are suggested based on tag co-occurrence statistics mined from large-scale collections, while topic modeling is employed in Xu et al. [2009]. As the tag-based methods presume that the test image has been labeled with some initial tags (i.e., the initial tags are taken as the user information Θ), they are inapplicable for tag assignment.

2.2.2. Tag + Image Based. Works in this group develop $f_\Phi(x, t; \Theta)$ on the base of visual information and associated tags. The main rationale behind them is visual consistency; that is, visually similar images shall be labeled with similar tags. Implementations of this intuition can be grouped in three conducts. One, leverage images visually close to the test image [Li et al. 2009b, 2010; Verbeek et al. 2010; Ma et al. 2010; Wu et al. 2011; Feng et al. 2012]. Two, exploit relationships between images labeled with the same tag [Liu et al. 2009; Richter et al. 2012; Liu et al. 2011b; Kuo et al. 2012; Gao et al. 2013]. Three, learn visual classifiers from socially tagged examples [Wang et al. 2009; Chen et al. 2012; Li and Snoek 2013; Yang et al. 2014]. By propagating tags based on the visual evidence, the aforementioned works exploit the image modality and the tag modality in a sequential way. By contrast, there are works that concurrently exploit the two modalities. This can be approached by generating a common latent space upon the image-tag association [Srivastava and Salakhutdinov 2014; Niu et al. 2014; Duan et al. 2014], so that a cross-media similarity can be computed between images and tags [Zhuang and Hoi 2011; Qi et al. 2012; Liu et al. 2013]. In Pereira et al. [2014], the latent space is constructed by Canonical Correlation Analysis, finding two matrices that separately project feature vectors of image and tag into the same subspace. In Ma et al. [2010], a random walk model is used on a unified graph composed from the fusion of an image similarity graph with an image-tag connection graph. In Wu et al. [2013], Xu et al. [2014], and Zhu et al. [2010], predefined image similarity and tag similarity are used as two constraint terms to enforce that similarities induced from the recovered image-tag association matrix will be consistent with the two predefined similarities.

Although late fusion has been actively studied for multimedia data analysis [Atrey et al. 2010], improving tag relevance estimation by late fusion is not much explored. There are some efforts in that direction, among which interesting performance has been reported in Qian et al. [2014] and more recently in Li [2016].

2.2.3. Tag + Image + User-Related Information Based. In addition to tags and images, this group of works exploits user information, motivated from varied perspectives. User information ranges from the simplest user identities [Li et al. 2009b], to tagging preferences [Sawant et al. 2010], to user reliability [Ginsca et al. 2014], to image group memberships [Johnson et al. 2015]. With the hypothesis that a specific tag chosen by many users to label visually similar images is more likely to be relevant with respect to the visual content, Li et al. [2009b] utilize user identity to ensure that learning examples come from distinct users. A similar idea is reported in Kennedy et al. [2009], finding visually similar image pairs with matching tags from different users. Ginsca et al. [2014] improve image retrieval by favoring images uploaded by users with good credibility estimates. The reliability of an image uploader is inferred by counting matches between the user-provided tags and machine tags predicted by visual concept detectors. In Sawant et al. [2010] and Li et al. [2011b], personal tagging preference is considered in the form of tag statistics computed from images a user has uploaded in the past. These past images are used in Liu et al. [2014] to learn a user-specific embedding space. In Sang et al. [2012a], user affinities, measured in terms of the number of common groups users are sharing, is considered in a tensor analysis framework. Similarly, tensor-based low-rank data reconstruction is employed in Qian et al. [2015] to discover latent associations between users, images, and tags. Photo timestamps are exploited for time-sensitive image retrieval [Kim and Xing 2013], where the connection between image occurrence and various temporal factors is modeled. In McParlane et al. [2013b], time-constrained tag co-occurrence statistics are considered to refine the output of visual classifiers for tag assignment. In their follow-up work [McParlane et al. 2013a], location-constrained tag co-occurrence computed from images

Table I. The Taxonomy of Methods for Tag Relevance Learning, Organized Along the *Media* and *Learning* Dimensions of Figure 1 (Methods for Which This Survey Provides an Experimental Evaluation Are Indicated in **Bold Font)**

	Learning		
Media	*Instance Based*	*Model Based*	*Transduction Based*
tag	**Sigurbjörnsson and Van Zwol [2008]** **Zhu et al. [2012]**	Xu et al. [2009] Wu et al. [2009]	–
tag + image	**Liu et al. [2009]** **Makadia et al. [2010]** Tang et al. [2011] Wu et al. [2011] Yang et al. [2011] Truong et al. [2012] Qi et al. [2012] Lin et al. [2013] Lee et al. [2013] Uricchio et al. [2013] Zhu et al. [2014] Ballan et al. [2014] Pereira et al. [2014]	**Guillaumin et al. [2009]** Verbeek et al. [2010] Li et al. [2010] Ma et al. [2010] Liu et al. [2011b] Duan et al. [2011] Feng et al. [2012] Srivastava and Salakhutdinov [2014] **Chen et al. [2012]** Lan and Mori [2013] **Li and Snoek [2013]** Li et al. [2013] Wang et al. [2014] Niu et al. [2014]	**Zhu et al. [2010]** Wang et al. [2010] Li et al. [2010] Zhuang and Hoi [2011] Richter et al. [2012] Kuo et al. [2012] Liu et al. [2013] Gao et al. [2013] Wu et al. [2013] Yang et al. [2014] Feng et al. [2014] Xu et al. [2014]
tag + image + user	**Li et al. [2009b]** Kennedy et al. [2009] Li et al. [2010] Znaidia et al. [2013] Liu et al. [2014]	Sawant et al. [2010] Li et al. [2011b] McAuley and Leskovec [2012] Kim and Xing [2013] McParlane et al. [2013a] Ginsca et al. [2014] Johnson et al. [2015]	**Sang et al. [2012a]** Sang et al. [2012b] Qian et al. [2015]

taken in a specific continent is further included. User interactions in social networks are exploited in Sawant et al. [2010], computing local interaction networks from the comments left by other users. In McAuley and Leskovec [2012] and Johnson et al. [2015], social network metadata such as image groups membership or contacts of users is employed to resolve ambiguity in visual appearance.

Comparing the three groups, tag + image appears to be the mainstream, as evidenced by the imbalanced distribution in Table I. Intuitively, using more media from $\mathcal{S}$ would typically improve tag relevance estimation. We attribute the imbalance among the groups, in particular the relatively few works in the third group, to the following two reasons. First, no publicly available dataset with expert annotations was built to gather representative and adequate user information; for example, MIRFlickr has nearly 10,000 users for 25,000 images, while in NUS-WIDE, only 6% of the users have at least 15 images. As a consequence, current works that leverage user information are forced to use a minimal subset to alleviate sample insufficiency [Sang et al. 2012a, 2012b] or homemade collections with social tags as ground truth instead of benchmark sets [Sawant et al. 2010; Li et al. 2011b]. Second, adding more media often results in a substantial increase in terms of both computation and memory, for example, the cubic complexity for tensor factorization in Sang et al. [2012a]. As a tradeoff, one has to use $\mathcal{S}$ of a much smaller scale. The dilemma is whether one should use large data with less media or more media but less data.

It is worth noting that the aforementioned groups are not exclusive. The output of some methods can be used as a refined input of some other methods. In particular, we observe a frequent usage of tag-based methods by others for their computational efficiency. For instance, tag relevance measured in terms of tag similarity is used in

Zhuang and Hoi [2011], Gao et al. [2013], and Li and Snoek [2013] before applying more advanced analysis, while nearest neighbor tag propagation is a preprocess used in Zhu et al. [2010]. The number of tags per image is embedded into image retrieval functions in Liu et al. [2009], Xu et al. [2009], Zhuang and Hoi [2011], and Chen et al. [2012].

Given the varied sources of information one could leverage, the subsequent question is how the information is exactly utilized, which will be made clear next.

2.3. Learning for Tag Relevance

This section presents the second dimension of the taxonomy, elaborating on various algorithms that implement the computation of tag relevance. Ideally, given the large-scale nature of social images, a desirable algorithm shall maintain good scalability as the data grows. The algorithm shall also provide a flexible mechanism to effectively integrate various types of information including tags, images, social metadata, and so forth while at the same time being robust when not all the information is available. In what follows, we review existing algorithms on their efforts to meet the requirements.

Depending on whether the tag relevance learning process is transductive (i.e., producing tag relevance scores without distinction as training and testing), we divide existing works into transduction based and induction based. Since the latter produces rules or models that are directly applicable to a novel instance [Michalski 1993], it has a better scalability for large-scale data compared to its transductive counterpart. Depending on whether an explicit model, let it be discriminative or generative, is built, a further division for the induction-based methods can be made: instance-based algorithms and model-based algorithms. Consequently, we divide existing works into the following three exclusive groups: (1) instance based, (2) model based, and (3) transduction based.

2.3.1. Instance Based. This class of methods does not perform explicit generalization but instead compares new test images with training instances. It is called instance based because it constructs hypotheses directly from the training instances themselves. These methods are nonparametric, and the complexity of the learned hypotheses grows as the amount of training data increases. The neighbor voting algorithm [Li et al. 2009b] and its variants [Kennedy et al. 2009; Li et al. 2010; Truong et al. 2012; Lee et al. 2013; Zhu et al. 2014] estimate the relevance of a tag t with respect to an image x by counting the occurrence of t in annotations of the visual neighbors of x. The visual neighborhood is created using features obtained from early fusion of global features [Li et al. 2009b], distance metric learning to combine local and global features [Verbeek et al. 2010; Wu et al. 2011], cross-modal learning of tags and image features [Qi et al. 2012; Ballan et al. 2014; Pereira et al. 2014], and fusion of multiple single-feature learners [Li et al. 2010; Li 2016]. While the standard neighbor voting algorithm [Li et al. 2009b] simply lets the neighbors vote equally, efforts have been made to (heuristically) weight neighbors in terms of their importance. For instance, in Truong et al. [2012] and Lee et al. [2013], the visual similarity is used as the weights. As an alternative to such a heuristic strategy, Zhu et al. [2014] model the relationships among the neighbors by constructing a directed voting graph, wherein there is a directed edge from image x_i to image x_j if x_i is in the k nearest neighbors of x_j. Subsequently, an adaptive random walk is conducted over the voting graph to estimate the tag relevance. However, the performance gain obtained by these weighting strategies appears to be limited [Zhu et al. 2014]. The kernel density estimation technique used in Liu et al. [2009] can be viewed as another form of weighted voting, but the votes come from images labeled with t instead of the visual neighbors. Yang et al. [2011] further consider the distance of the test image to images not labeled with t. In order to eliminate semantically unrelated

samples in the neighborhood, sparse reconstruction from a k-nearest neighborhood is used in Tang et al. [2009, 2011]. In Lin et al. [2013], with the intention of recovering missing tags by matrix reconstruction, the image and tag modalities are separately exploited in parallel to produce a new candidate image-tag association matrix each. Then, the two resultant tag relevance scores are linearly combined to produce the final tag relevance scores. To address the incompleteness of tags associated with the visual neighbors, Znaidia et al. [2013] propose to enrich these tags by exploiting tag co-occurrence in advance of neighbor voting.

2.3.2. Model Based. This class of tag relevance learning algorithms puts their foundations on parameterized models learned from the training media. Notice that the models can be tag specific or holistic for all tags. As an example of holistic modeling, a topic model approach is presented in Wang et al. [2014] for tag refinement, where a hidden topic layer is introduced between images and tags. Consequently, the tag relevance function is implemented as the dot product between the topic vector of the image and the topic vector of the tag. In particular, the authors extend the Latent Dirichlet Allocation model [Blei et al. 2003] to force images with similar visual content to have similar topic distribution. According to their experiments [Wang et al. 2014], however, the gain of such a regularization appears to be marginal compared to the standard Latent Dirichlet Allocation model. Li et al. [2013] first find embedding vectors of training images and tags using the image-tag association matrix of $\mathcal{S}$. The embedding vector of a test image is obtained by a convex combination of the embedding vectors of its neighbors retrieved in the original visual feature space. Consequently, the relevance score is computed in terms of the Euclidean distance between the embedding vectors of the test image and the tag.

For tag-specific modeling, linear SVM classifiers trained on features augmented by pretrained classifiers of popular tags are used in Chen et al. [2012] for tag retrieval. Fast intersection kernel SVMs trained on selected relevant positive and negative examples are used in Li and Snoek [2013]. A bag-based image reranking framework is introduced in Duan et al. [2011], where pseudo-relevant images retrieved by tag matching are partitioned into clusters using visual and textual features. Then, by treating each cluster as a bag and images within the cluster as its instances, multiple-instance learning [Andrews et al. 2003] is employed to learn multiple-instance SVMs per tag. Viewing the social tags of a test image as ground truth, a multimodal tag suggestion method based on both tags and visual correlation is introduced in Wu et al. [2009]. Each modality is used to generate a ranking feature, and the tag relevance function is a combination of these ranking features, with the combination weights learned online by the RankBoost algorithm [Freund et al. 2003]. In Guillaumin et al. [2009] and Verbeek et al. [2010], logistic regression models are built per tag to promote rare tags. In a similar spirit to Li and Snoek [2013], Zhou et al. [2015] learn an ensemble of SVMs by treating tagged images as positive training examples and untagged images as candidate negative training examples. Using the ensemble to classify image regions generated by automated image segmentation, the authors assign tags at the image level and the region level simultaneously.

2.3.3. Transduction Based. This class of methods consists of procedures that evaluate tag relevance for all image-tag pairs by minimizing a specific cost function. Given the initial image-tag association matrix D, the output of the procedures is a new matrix $\hat{D}$, the elements of which are taken as tag relevance scores. Due to this formulation, no explicit form of the tag relevance function exists, nor does any distinction between training and test sets [Joachims 1999]. If novel images are added to the initial set, minimization of the cost function needs to be recomputed.

The majority of transduction-based approaches are founded on matrix factorization [Zhu et al. 2010; Sang et al. 2012a; Liu et al. 2013; Wu et al. 2013; Kalayeh et al. 2014; Feng et al. 2014; Xu et al. 2014]. In Zhuang and Hoi [2011], the objective function is a linear combination of the difference between $\hat{D}$ and the matrix of image similarity, the distortion between $\hat{D}$ and the matrix of tag similarity, and the difference between $\hat{D}$ and D. A stochastic coordinate descent optimization is applied to a randomly chosen row of $\hat{D}$ per iteration. In Zhu et al. [2010], considering the fact that D is corrupted with noise derived by missing or overpersonalized tags, robust principal component analysis with Laplacian regularization is applied to recover $\hat{D}$ as a low-rank matrix. In Wu et al. [2013], $\hat{D}$ is regularized such that the image similarity induced from $\hat{D}$ is consistent with the image similarity computed in terms of low-level visual features, and the tag similarity induced from $\hat{D}$ is consistent with the tag correlation score computed in terms of tag co-occurrence. Xu et al. [2014] propose to reweight the penalty term of each image-tag pair by their relevance score, which is estimated by a linear fusion of tag-based and content-based relevance scores. To incorporate the user element, Sang et al. [2012a] extend D to a three-way tensor with tag, image, and user as each of the ways. A core tensor and three matrices representing the three media, obtained by Tucker decomposition [Tucker 1966], are multiplied to construct $\hat{D}$.

As an alternative approach, in Feng et al. [2014], it is assumed that the tags of an image are drawn independently from a fixed but unknown multinomial distribution. Estimation of this distribution is implemented by maximum likelihood with low-rank matrix recovery and Laplacian regularization like Zhu et al. [2010].

Graph-based label propagation is another type of transduction-based method. In Richter et al. [2012], Wang et al. [2010], and Kuo et al. [2012], the image-tag pairs are represented as a graph in which each node corresponds to a specific image and the edges are weighted according to a multimodal similarity measure. Viewing the top-ranked examples in the initial search results as positive instances, tag refinement is implemented as a semisupervised labeling process by propagating labels from the positive instances to the remaining examples using random walk. While the edge weights are fixed in the aforementioned works, Gao et al. [2013] argue that fixing the weights could be problematic, because tags found to be discriminative in the learning process should adaptively contribute more to the edge weights. In that regard, the hypergraph learning algorithm [Zhou et al. 2006] is exploited and weights are optimized by minimizing a joint loss function that considers both the graph structure and the divergence between the initial labels and the learned labels. In Liu et al. [2011a], the hypergraph is embedded into a lower-dimension space by hypergraph Laplacian.

Comparing the three groups of methods for learning tag relevance, an advantage of instance-based methods against the other two groups is their flexibility to adapt to previously unseen images and tags. They may simply add new training images into $\mathcal{S}$ or remove outdated ones. The advantage, however, comes with a price that $\mathcal{S}$ has to be maintained, a nontrivial task given the increasing amount of training data available. Also, the computational complexity and memory footprint grow linearly with respect to the size of $\mathcal{S}$. In contrast, model-based methods could be more swift, especially when linear classifiers are used, as the training data is compactly represented by a fixed number of models. As the imagery of a given tag may evolve, retraining is required to keep the models up to date.

Different from instance-based and model-based learning, where individual tags are considered independently, transduction-based learning methods via matrix factorization can favorably exploit intertag and interimage relationships. However, their ability to deal with the extremely large number of social images is a concern. For instance,

the use of Laplacian graphs results in a memory complexity of $O(|\mathcal{S}|^2)$. The accelerated proximal gradient algorithm used in Zhu et al. [2010] requires Singular Value Decomposition, which is known to be an expensive operation. The Tucker decomposition used in Sang et al. [2012a] has a cubic computational complexity with respect to the number of training samples. We notice that some engineering tricks have been considered in these works, which alleviate the scalability issue to some extent. In Zhuang and Hoi [2011], for instance, clustering is conducted in advance to divide $\mathcal{S}$ into much smaller subsets, and the algorithm is applied to these subsets, separately. By making the Laplacian more sparse by retaining only the k-nearest neighbors [Zhu et al. 2010; Sang et al. 2012a], the memory footprint can be reduced to $O(k \cdot |\mathcal{S}|)$, with the cost of performance degeneration. Perhaps due to the scalability concern, works resorting to matrix factorization tend to experiment with a dataset of relatively small scale.

In summary, instance-based learning, in particular neighbor voting, is the first choice to try for its simplicity and decent performance. When the test tags are well defined (in the sense of relevant learning examples that can be collected automatically), model-based learning is more attractive. When the test images share similar social context (e.g., images shared by a group of specific interest), they tend to be on similar topics. In such a scenario, transduction-based learning that exploits the interimage relationship is more suited.

2.4. Auxiliary Components

The *Filter* and the *Precompute* component are auxiliary components that may sustain and improve tag relevance learning.

Filter. As social tags are known to be subjective and overly personalized, removing personalized tags appears to be a natural and simple way to improve the tagging quality. This is usually the first step performed in the framework for tag relevance learning. Although there is a lack of golden criteria to determine which tags are personalized, a popular strategy is to exclude tags that cannot be found in the WordNet ontology [Zhu et al. 2010; Li et al. 2011b; Chen et al. 2012; Zhu et al. 2012] or a Wikipedia thesaurus [Liu et al. 2009]. Tags with rare occurrence, say, appearing less than 50 times, are discarded in Verbeek et al. [2010] and Zhu et al. [2010]. For methods that directly work on the image-tag association matrix [Zhu et al. 2010; Sang et al. 2012a; Wu et al. 2013; Lin et al. 2013], reducing the size of the vocabulary in terms of tag occurrence is an important prerequisite to keep the matrix in a manageable scale. Observing that images tagged in a batch manner are often nearly duplicate and of low tagging quality, batch-tagged images are excluded in Li et al. [2012]. Since relevant tags may be missing from user annotations, the negative tags that are semantically similar or co-occurring with positive ones are discarded in Sang et al. [2012a]. As the previous strategies do not take the visual content into account, they cannot handle situations where an image is incorrectly labeled with a valid and frequently used tag, say, "dog." In Li et al. [2009a], tag relevance scores are assigned to each image in $\mathcal{S}$ by running the neighbor voting algorithm [Li et al. 2009b], while in Li and Snoek [2013], the semantic field algorithm [Zhu et al. 2012] is further added to select relevant training examples. In Qian et al. [2015], the annotation of the training media is enriched by a random walk.

Precompute. The precompute component is responsible for the generation of the prior information that is jointly used with the refined training media $\hat{\mathcal{S}}$ in learning. For instance, global statistics and external resources can be used to synthesize new prior knowledge useful in learning. The prior information commonly used is tag statistics in $\mathcal{S}$, including tag occurrence and tag co-occurrence. Tag occurrence is used in Li et al. [2009b] as a penalty to suppress overly frequent tags. Measuring the semantic similarity between two tags is important for tag relevance learning algorithms that exploit

tag correlations. While linguistic metrics such as those derived from WordNet were used before the proliferation of social media [Jin et al. 2005; Wang et al. 2006], they do not directly reflect how people tag images. For instance, the tags "sunset" and "sea" are weakly related according to the WordNet ontology, but they often appear together in social tagging as many of the sunset photos are shot around seasides. Therefore, similarity measures that are based on tag statistics computed from many socially tagged images are in dominant use. Sigurbjörnsson and van Zwol [2008] utilized the Jaccard coefficient and a conditional tag probability in their tag suggestion system, while Liu et al. [2013] used normalized tag co-occurrence. To better capture the visual relationship between two tags, Wu et al. [2008] proposed the Flickr distance. The authors represent each tag by a visual language model, trained on bag-of-visual-words features of images labeled with this tag. The Flickr distance between two tags is computed as the Jensen-Shannon divergence between the corresponding models. Later, Jiang et al. [2009] introduced the Flickr context similarity, which also captures the visual relationship between two tags, but without the need of the expensive visual modeling. The trick is to compute the Normalized Google Distance [Cilibrasi and Vitanyi 2007] between two tags, but with tag statistics acquired from Flickr image collections instead of Google indexed web pages. For its simplicity and effectiveness, we observe a prevalent use of the Flickr context similarity in the literature [Liu et al. 2009; Zhu et al. 2010; Wang et al. 2010; Zhuang and Hoi 2011; Zhu et al. 2012; Gao et al. 2013; Li and Snoek 2013; Qian et al. 2014].

3. A NEW EXPERIMENTAL PROTOCOL

In spite of the expanding literature, there is a lack of consensus on the performance of the individual methods. This is largely due to the fact that existing works either use homemade data (see Liu et al. [2009], Wang et al. [2010], Chen et al. [2012], and Gao et al. [2013]), which are not publicly accessible, or use selected subsets of benchmark data (e.g., as in Zhu et al. [2010], Sang et al. [2012a], and Feng et al. [2014]). As a consequence, the performance scores reported in the literature are not comparable across the papers.

Benchmark data with manually verified labels is crucial for an objective evaluation. As Flickr has been well recognized as a profound manifestation of social image tagging, Flickr images act as a main source for benchmark construction. MIRFlickr from the Leiden University [Huiskes et al. 2010] and NUS-WIDE from the National University of Singapore [Chua et al. 2009] are the two most popular Flickr-based benchmark sets for social image tagging and retrieval, as demonstrated by the number of citations. On the use of the benchmarks, one typically follows a single-set protocol, that is, learning the underlying tag relevance function from the training part of a chosen benchmark set and evaluating it on the test part. Such a protocol is inadequate given the dynamic nature of social media, which could easily make an existing benchmark set outdated. For any method targeting social images, a cross-set evaluation is necessary to test its generalization ability, which is, however, overlooked in the literature.

Another desirable property is the capability to learn from the increasing amounts of socially tagged images. Since existing works mostly use training data of a fixed scale, this property has not been well evaluated.

Following these considerations, we present a new experimental protocol, wherein training and test data from distinct research groups are chosen for evaluating a number of representative works in the cross-set scenario. Training sets with their size ranging from 10,000 to 1 million images are constructed to evaluate methods of varied complexity. To the best of our knowledge, such a comparison between many methods on varied scale datasets with a common experimental setup has not been conducted

Table II. Our Proposed Experimental Protocol Instantiates the *Media* and *Tasks* Dimensions of Figure 1 with Three Training Sets and Three Test Sets for Tag Assignment, Refinement, and Retrieval (Note That the Training Sets Are Socially Tagged; They Have No Ground Truth Available for Any Tag)

	Media Characteristics				Tasks		
Media	**# Images**	**# Tags**	**# Users**	**# Test Tags**	**Assignment**	**Refinement**	**Retrieval**
Training media $\mathcal{S}$:							
Train10k	10,000	41,253	9,249	–	✓	✓	✓
Train100k	100,000	214,666	68,215	–	✓	✓	✓
Train1m [Li et al. 2012]	1,198,818	1,127,139	347,369	–	✓	✓	✓
Test media $\mathcal{X}$:							
MIRFlickr [Huiskes et al. 2010]	25,000	67,389	9,862	14	✓	✓	–
Flickr51 [Wang et al. 2010]	81,541	66,900	20,886	51	–	–	✓
NUS-WIDE [Chua et al. 2009]	259,233	355,913	51,645	81	✓	✓	✓

before. For the sake of experimental reproducibility, all data and code are available online.[1]

3.1. Datasets

We describe the training media $\mathcal{S}$ and the test media $\mathcal{X}$ as follows, with basic data characteristics and their usage summarized in Table II.

Training media $\mathcal{S}$. We use a set of 1.2 million Flickr images collected by the University of Amsterdam [Li et al. 2012] by using over 25,000 nouns in WordNet as queries to uniformly sample images uploaded between 2006 and 2010. Based on our observation that batch-tagged images, namely, those labeled with the same tags by the same user, tend to be near duplicate, we have excluded these images beforehand. Other than this, we do not perform near-duplicate image removal. To meet with methods that cannot handle large data, we created two random subsets from the entire training sets, resulting in three training sets of varied sizes, termed as Train10k, Train100k, and Train1m, respectively.

Test media $\mathcal{X}$. We use MIRFlickr [Huiskes et al. 2010] as in Verbeek et al. [2010], Zhu et al. [2010], and Uricchio et al. [2013] and NUS-WIDE [Chua et al. 2009] as in Tang et al. [2011], McAuley and Leskovec [2012], Zhu et al. [2010], and Uricchio et al. [2013] for tag assignment and refinement. We use NUS-WIDE for evaluating tag retrieval as in Sun et al. [2011] and Li et al. [2011a]. In addition, for retrieval, we collected another test set, namely, Flickr51, contributed by Microsoft Research Asia [Wang et al. 2010; Gao et al. 2013]. The MIRFlickr set contains 25,000 images with ground truth available for 14 tags. The NUS-WIDE set contains 259,233 images with ground truth available for 81 tags. The Flickr51 set consists of 81,541 Flickr images with partial ground truth provided for 55 test tags. Among the 55 tags, there are four tags that either have zero occurrence in our training data or have no correspondence in WordNet, so we ignore them. Differently from the binary judgments in NUS-WIDE, Flickr51 provides graded relevance, with 0, 1, and 2 to indicate irrelevant, relevant, and very relevant, respectively. Moreover, the set contains several ambiguous tags such as "apple" and "jaguar," where relevant instances could exhibit completely different imagery (e.g., Apple computers versus fruit apples). Following the original intention of the datasets, we use MIRFlickr and NUS-WIDE for evaluating tag assignment and tag refinement, and Flickr51 and NUS-WIDE for tag retrieval. For all three test sets, we use the full dataset for testing.

Although the training and test media are all from Flickr, they were collected independently, and consequently they have a relatively small amount of images overlapped with each other, as shown in Table III.

Table III. Data Overlap Between Train1M and the Three Test Sets, Measured in Terms of the Number of Shared Images, Tags, and Users, Respectively

	Overlap with Train1M		
Test Media	**# Images**	**# Tags**	**# Users**
MIRFlickr	–	693	6,515
Flickr51	730	538	14,211
NUS-WIDE	7,975	718	38,481

Tag overlap is counted on the top 1,000 most frequent tags. As the original photo ids of MIRFlickr have been anonymized, we cannot check image overlap between this dataset and Train1M.

3.2. Implementation

This section describes common implementations applicable to all three tasks, including the choice of visual features and tag preprocessing. Implementations that are applied uniquely to single tasks will be described in the coming sections.

Visual features. Two types of features are extracted to provide insights of the performance improvement achievable by appropriate feature selection: the classical bag of visual words (BoVW) and the current state-of-the-art deep learning-based features extracted from Convolutional Neural Networks (CNNs). The BoVW feature is extracted by the color descriptor software [Van De Sande et al. 2010]. SIFT descriptors are computed at dense sampled points, at every 6 pixels for two scales. A codebook of size 1,024 is created by K-means clustering. The SIFTs are quantized by the codebook using hard assignment and aggregated by sum pooling. In addition, we extract a compact 64D global feature [Li 2007], combining a 44D color correlogram, a 14D texture moment, and a 6D RGB color moment, to compensate the BoVW feature. The CNN feature is extracted by the pretrained VGGNet [Simonyan and Zisserman 2015]. In particular, we adopt the 16-layer VGGNet and take as feature vectors the last fully connected layer of ReLU activation, resulting in a feature vector of 4,096 dimensions per image. The BoVW feature is used with the l_1 distance and the CNN feature is used with the cosine distance for their good performance.

Vocabulary $\mathcal{V}$. As what tags a person may use is meant to be open, the need for specifying a tag vocabulary is merely an engineering convenience. For a tag to be meaningfully modeled, there has to be a reasonable amount of training images with respect to that tag. For methods where tags are processed independently from the others, the size of the vocabulary has no impact on the performance. In the other cases, in particular, for transductive methods that rely on the image-tag association matrix, the tag dimension has to be constrained to make the methods runnable. In our case, for these methods a three-step automatic cleaning procedure is performed on the training datasets. First, all the tags are lemmatized to their base forms by the NLTK software [Bird et al. 2009]. Second, tags not defined in WordNet are removed. Finally, in order to avoid insufficient sampling, we remove tags that cannot meet a threshold on tag occurrence. The thresholds are empirically set as 50, 250, and 750 for Train10k, Train100k, and Train1m, respectively, in order to have a linear increase in vocabulary size versus a logarithmic increase in the number of labeled images. This results in a final vocabulary of 237, 419, and 1,549 tags, respectively, with all the test tags included. Note that these numbers of tags are larger than the number of tags that can be actually evaluated. This allows us to build a unified evaluation framework that is more handy for cross-dataset evaluation.

3.3. Evaluating Tag Assignment

Evaluation criteria. A good method for tag assignment shall rank relevant tags before irrelevant tags for a given test image. Moreover, with the assigned tags, relevant

images shall be ranked before irrelevant images for a given test tag. We therefore use the image-centric Mean image Average Precision (MiAP) to measure the quality of tag ranking, and the tag-centric Mean Average Precision (MAP) to measure the quality of image ranking. Let m_{gt} be the number of ground-truth test tags, which is 14 for MIRFlickr and 81 for NUS-WIDE. The image-centric Average Precision of a given test image x is computed as

$$iAP(x) := \frac{1}{R}\sum_{j=1}^{m_{gt}} \frac{r_j}{j}\delta(x, t_j), \tag{1}$$

where R is the number of relevant tags of the given image, r_j is the number of relevant tags in the top j ranked tags, and $\delta(x_i, t_j) = 1$ if tag t_j is relevant and 0 otherwise. MiAP is obtained by averaging $iAP(x)$ over the test images.

The tag-centric Average Precision of a given test tag t is computed as

$$AP(t) := \frac{1}{R}\sum_{i=1}^{n} \frac{r_i}{i}\delta(x_i, t), \tag{2}$$

where R is the number of relevant images for the given tag and r_i is the number of relevant images in the top i ranked images. MAP is obtained by averaging $AP(t)$ over the test tags.

The two metrics are complementary to some extent. Since MiAP is averaged over images, each test image contributes equally to MiAP, as opposed to MAP, where each tag contributes equally. Consequently, MiAP is biased toward frequent tags, while MAP can be easily affected by the performance of rare tags, especially when m_{gt} is relatively small.

Baseline. Any method targeting at tag assignment shall be better than a random guess, which simply returns a random set of tags. The RandomGuess baseline is obtained by computing MiAP and MAP given the random prediction, which is run 100 times with the resulting scores averaged.

3.4. Evaluating Tag Refinement

Evaluation criteria. As tag refinement is also meant for improving tag ranking and image ranking, it is evaluated by the same criteria (i.e., MiAP and MAP) as used for tag assignment.

Baseline. A natural baseline for tag refinement is the original user tags assigned to an image, which we term as UserTags.

3.5. Evaluating Tag Retrieval

Evaluation criteria. To compare methods for tag retrieval, for each test tag we first conduct a tag-based image search to retrieve images labeled with that tag, and then sort the images by the tag relevance scores. We use MAP to measure the quality of the entire image ranking. As users often look at the top-ranked results and hardly go through the entire list, we also report Normalized Discounted Cumulative Gain (NDCG), commonly used to evaluate the top few ranked results of an information retrieval system [Järvelin and Kekäläinen 2002]. Given a test tag t, its NDCG at a particular rank position h is defined as

$$NDCG_h(t) := \frac{DCG_h(t)}{IDCG_h(t)}, \tag{3}$$

where $DCG_h(t) = \sum_{i=1}^{h} \frac{2^{rel_i}-1}{\log_2(i+1)}$, rel_i is the graded relevance of the result at position i, and $IDCG_h$ is the maximum possible DCG till position h. We set h to be 20, which

corresponds to a typical number of search results presented on the first two pages of a web search engine. Similar to MAP, $NDCG_{20}$ of a specific method on a specific test set is averaged over the test tags of that test set.

Baselines. When searching for relevant images for a given tag, it is natural to ask how much a specific method gains compared to a baseline system that simply returns a random subset of images labeled with that tag. Similar to the refinement baseline, we also denote this baseline as UserTags, as both of them purely use the original user tags. For each test tag, the test images labeled with this tag are sorted at random, and MAP and $NDCG_{20}$ are computed accordingly. The process is executed 100 times, and the average score over the 100 runs is reported.

The number of tags per image is often included for image ranking in previous works [Liu et al. 2009; Xu et al. 2009]. Hence, we build another baseline system, denoted as TagNum, which sorts images in ascending order by the number of tags per image. The third baseline, denoted as TagPosition, is from Sun et al. [2011], where the relevance score of a tag is determined by its position in the original tag list uploaded by the user. More precisely, the score is computed as $1 - position(t)/l$, where l is the number of tags.

4. METHODS SELECTED FOR COMPARISON

Despite the rich literature, most works do not provide code. An exhaustive evaluation covering all published methods is impractical. We have to leave out methods that do not show significant improvements or novelties w.r.t. the seminal papers in the field, and methods that are difficult to replicate with the same mathematical preciseness as intended by their developers. We drive our choice by the intention to cover methods that aim for each of the three tasks, exploiting varied modalities by distinct learning mechanisms. Eventually we evaluate 11 representative methods. For each method, we analyze its scalability in terms of both computation and memory. Our analysis leaves out operations that are independent of specific tags and thus only need to be executed once in an offline manner, such as visual feature extraction, tag preprocessing, prior information precomputing, and filtering. The main properties of the methods are summarized in Table IV. Concerning the choices of parameters, we adopt what the original papers recommend. When no recommendation is given for a specific method, we try a range of values to our best understanding and choose the parameters that yield the best overall performance.

4.1. Methods Under Analysis

1. SemanticField [Zhu et al. 2012]. This method measures tag relevance in terms of an averaged semantic similarity between the tag and the other tags assigned to the image:

$$f_{SemField}(x, t) := \frac{1}{l_x} \sum_{i=1}^{l_x} sim(t, t_i), \tag{4}$$

where $\{t_1, \ldots, t_{l_x}\}$ is a list of l_x social tags assigned to the image x, and $sim(t, t_i)$ denotes a semantic similarity between two tags. SemanticField explicitly assumes that several tags are associated to visual data and their coexistence is accounted for in the evaluation of tag relevance. Following Zhu et al. [2012], the similarity is computed by combining the Flickr context similarity and the WordNet Wu-Palmer similarity [Wu and Palmer 1994]. The WordNet-based similarity exploits path length in the WordNet hierarchy to infer tag relatedness. We make a small revision of Zhu et al. [2012] (i.e., combining the two similarities by averaging instead of multiplication), because the former strategy produces slightly better results. SemanticField requires no training except for computing tag-wise similarity, which can be computed offline and is thus

Table IV. Main Properties of the 11 Methods Evaluated in This Survey Following the Dimensions of Figure 1 (The Computational and Memory Complexity of Each Method Is Based on Processing n Test Images and m Test Tags by Exploiting the Training Set $\mathcal{S}$)

			Auxiliary Component		Learning			
Methods	**Test Media**	**Task**	*Filter*	*Precompute*	*Train Computation*	*Test Computation*	*Train Memory*	*Test Memory*
Instance based:								
SemanticField	tag	Retrieval	WordNet	$sim(t,t')$	–	$O(nml_x)$	–	$O(m^2)$
TagCooccur	tag	Refinement Retrieval	–	Tag prior Co-occurrence	–	$O(nml_x)$	–	$O(m^2)$
TagRanking	tag + image	Retrieval	–	$sim(t,t')$	–	$O(n(md\bar{n}+Lm^2))$	–	$O(\max(d\bar{n}, m^2))$
KNN	tag + image	Assignment Retrieval	–	–	–	$O(n(d\|\mathcal{S}\|+k\log\|\mathcal{S}\|))$	–	$O(d\|\mathcal{S}\|)$
TagVote	tag + image	Assignment Retrieval	–	Tag prior	–	$O(n(d\|\mathcal{S}\|+k\log\|\mathcal{S}\|))$	–	$O(d\|\mathcal{S}\|)$
TagCooccur+	tag + image	Refinement Retrieval	–	Tag prior Co-occurrence	–	$O(n(d\|\mathcal{S}\|+k\log\|\mathcal{S}\|))$	–	$O(d\|\mathcal{S}\|)$
Model based:								
TagProp	tag + image	Assignment Retrieval	–	–	$O(l\cdot m\cdot k)$	$O(n(d\|\mathcal{S}\|+k\log\|\mathcal{S}\|))$	$O(d\|\mathcal{S}\|+2m)$	$O(d\|\mathcal{S}\|+2m)$
TagFeature	tag + image	Assignment Retrieval	–	Tag classifiers	$O(m(d+d')p)$	$O(nm(d+d'))$	$O((d+d')p)$	$O(m(d+d'))$
RelExample	tag + image	Assignment Retrieval	SemField + TagVote	$sim(t,t')$	$O(mTdp^2)$	$O(dp+dq)$	$O(nmd)$	$O(mdq)$
Transduction based:								
RobustPCA	tag + image	Refinement Retrieval	WordNet + KNN	L_i, L_t	$O(cm^2n+c'n^3)$		$O(cnm+c'\cdot(n^2+m^2))$	
TensorAnalysis	tag + image + user	Refinement	Postag sets	L_i, L_t, L_u	$O(\|P_1\|\cdot(r_T\cdot m^2+r_U\cdot r_I\cdot r_T))$		$O(n^2+m^2+u^2)$	

omitted. Having all tag-wise similarities in memory, applying Equation (4) requires l_x table lookups per tag. Hence, the computational complexity is $O(m \cdot l_x)$, and $O(m^2)$ for memory.

2. TagRanking [Liu et al. 2009]. The tag ranking algorithm consists of two steps. Given an image x and its tags, the first step produces an initial tag relevance score for each of the tags, obtained by (Gaussian) kernel density estimation on a set of $\bar{n} = 1{,}000$ images labeled with each tag, separately. Second, a random walk is performed on a tag graph where the edges are weighted by a tag-wise similarity. We use the same similarity as in SemanticField. Notice that when applied for tag retrieval, the algorithm uses the rank of t instead of its score, that is,

$$f_{TagRanking}(x, t) = -rank(t) + \frac{1}{l_x}, \tag{5}$$

where $rank(t)$ returns the rank of t produced by the tag ranking algorithm. The term $\frac{1}{l_x}$ is a tie-breaker when two images have the same tag rank. Hence, for a given tag t, TagRanking cannot distinguish relevant images from irrelevant images if t is the sole tag assigned to them. It explicitly exploits the coexistence of several tags per image. TagRanking has no learning stage. To derive tag ranks for Equation (5), the main computation is the kernel density estimation on $\bar{n}$ socially tagged examples for each tag, followed by an L iteration random walk on the tag graph of m nodes. All this results in a computation cost of $O(m \cdot d \cdot \bar{n} + L \cdot m^2)$ per test image. Because the two steps are executed sequentially, the corresponding memory cost is $O(\max(d\bar{n}, m^2))$.

3. KNN [Makadia et al. 2010]. This algorithm estimates the relevance of a given tag with respect to an image by first retrieving k-nearest neighbors from $\mathcal{S}$ based on a visual distance d, and then counting the tag occurrence in associated tags of the neighborhood. In particular, KNN builds $f_\Phi(x, t; \Theta)$ as

$$f_{KNN}(x, t) := k_t, \tag{6}$$

where k_t is the number of images with t in the visual neighborhood of x. The instance-based KNN requires no training. The main computation of f_{KNN} is to find k-nearest neighbors from $\mathcal{S}$, which has a complexity of $O(d \cdot |\mathcal{S}| + k \cdot \log|\mathcal{S}|)$ per test image, and a memory footprint of $O(d \cdot |\mathcal{S}|)$ to store all the d-dimensional feature vectors. It is worth noting that these complexities are drawn from a straightforward implementation of k-nn search and can be substantially reduced by employing more efficient search techniques (c.f. Jegou et al. [2011]). Accelerating KNN by the product quantization technique [Jegou et al. 2011] imposes an extra training step, where one has to construct multiple vector quantizers by K-means clustering and further use the quantizers to compress the original feature vector into a few codes.

4. TagVote [Li et al. 2009b]. The TagVote algorithm estimates the relevance of a tag t w.r.t. an image x by counting the occurrence frequency of t in social annotations of the visual neighbors of x. Different from KNN, TagVote exploits the user element, introducing a unique-user constraint on the neighbor set to make the voting result more objective. Each user has at most one image in the neighbor set. Moreover, TagVote takes into account tag prior frequency to suppress over frequent tags. In particular, the TagVote algorithm builds $f_\Phi(x, t; \Theta)$ as

$$f_{TagVote}(x, t) := k_t - k\frac{n_t}{|\mathcal{S}|}, \tag{7}$$

where n_t is the number of images labeled with t in $\mathcal{S}$. Following Li et al. [2009b], we set k to be 1,000 for both KNN and TagVote. TagVote has the same order of complexity as KNN.

5. TagProp [Guillaumin et al. 2009; Verbeek et al. 2010]. TagProp employs neighbor voting plus distance metric learning. A probabilistic framework is proposed where the probability of using images in the neighborhood is defined based on rank or distance-based weights. TagProp builds $f_\Phi(x, t; \Theta)$ as

$$f_{TagProp}(x, t) := \sum_{j}^{k} \pi_j \cdot \mathbf{I}(x_j, t), \tag{8}$$

where π_j is a nonnegative weight indicating the importance of the jth neighbor x_j, and $\mathbf{I}(x_j, t)$ returns 1 if x_j is labeled with t, and 0 otherwise. Following Verbeek et al. [2010], we use $k = 1{,}000$ and the rank-based weights, which showed similar performance to the distance-based weights. Different from TagVote that uses tag prior to penalize frequent tags, TagProp promotes rare tags and penalizes frequent ones by training a logistic model per tag upon $f_{TagProp}(x, t)$.

The use of the logistic model makes TagProp a model-based method. In contrast to KNN and TagVote, wherein visual neighbors are treated equally, TagProp employs distance metric learning to reweight the neighbors, yielding a learning complexity of $O(l \cdot m \cdot k)$, where l is the number of gradient descent iterations it needs (typically less than 10). TagProp maintains $2m$ extra parameters for the logistic models, though their storage cost is ignorable compared to the visual features. Therefore, running Equation (8) has the same order of complexity as KNN and TagVote.

6. TagCooccur [Sigurbjörnsson and Van Zwol 2008]. While both SemanticField and TagCooccur are tag based, the main difference lies in how they compute the contribution of a specific tag to the test tag's relevance score. Different from SemanticField, which uses tag similarities, TagCooccur uses the test tag's rank in the tag ranking list created by sorting all tags in terms of their co-occurrence frequency with the tag in $\mathcal{S}$. In addition, TagCooccur takes into account the stability of the tag, measured by its frequency. The method is implemented as

$$f_{tagcooccur}(x, t) = descriptive(t) \sum_{i=1}^{l_x} vote(t_i, t) \cdot rank\text{-}promotion(t_i, t) \cdot stability(t_i), \tag{9}$$

where $descriptive(t)$ is to damp the contribution of tags with a very high frequency, $rank\text{-}promotion(t_i, t)$ measures the rank-based contribution of t_i to t, $stability(t_i)$ is for promoting tags for which the statistics are more stable, and $vote(t_i, t)$ is 1 if t is among the top 25 ranked tags of t_i, and 0 otherwise. TagCooccur has the same order of complexity as SemanticField.

7. TagCooccur+ [Li et al. 2009b]. TagCooccur+ is proposed to improve TagCooccur by adding the visual content. This is achieved by multiplying $f_{tagcooccur}(x, t)$ with a content-based term, that is,

$$f_{tagcooccur+}(x, t) = f_{tagcooccur}(x, t) \cdot \frac{k_c}{k_c + r_c(t) - 1}, \tag{10}$$

where $r_c(t)$ is the rank of t when sorting the vocabulary by $f_{TagVote}(x, t)$ in descending order, and k_c is a positive weighting parameter, which is empirically set to 1. While TagCooccur+ is grounded on TagCooccur and TagVote, the complexity of the former is ignorable compared to the latter, so the complexity of TagCooccurs+ is the same as KNN.

8. TagFeature [Chen et al. 2012]. The basic idea is to enrich image features by adding an extra tag feature. A tag vocabulary that consists of d' most frequent tags in $\mathcal{S}$ is constructed first. Then, for each tag, a two-class linear SVM classifier is trained using LIBLINEAR [Fan et al. 2008]. The positive training set consists of p images labeled

with the tag in $\mathcal{S}$, and the same amount of negative training examples is randomly sampled from images not labeled with the tag. The probabilistic output of the classifier, obtained by Platt's scaling [Lin et al. 2007], corresponds to a specific dimension in the tag feature. By concatenating the tag and visual features, an augmented feature of the $d + d'$ dimension is obtained. For a test tag t, its tag relevance function $f_{TagFeature}(x, t)$ is obtained by retraining an SVM classifier using the augmented feature. The linear property of the classifier allows us to first sum up all the support vectors into a single vector and consequently to classify a test image by the inner product with this vector. That is,

$$f_{TagFeature}(x, t) := b + <x_t, x>, \tag{11}$$

where x_t is the weighted sum of all support vectors and b the intercept. To build meaningful classifiers, we use tags that have at least 100 positive examples. While d' is chosen to be 400 in Chen et al. [2012], the two smaller training sets, namely, Train10k and Train100k, have 76 and 396 tags satisfying the previous requirement. We empirically set p to 500 and do random down-sampling if the amount of images for a tag exceeds this number. For TagFeature, learning a linear classifier for each tag from p positive and p negative examples requires $O((d + d')p)$ in computation and $O((d + d')p)$ in memory [Fan et al. 2008]. Running Equation (11) for all the m tags and n images needs $O(nm(d + d'))$ in computation and $O(m(d + d'))$ in memory.

9. RelExample [Li and Snoek 2013]. Different from TagFeature [Chen et al. 2012], which directly learns from tagged images, RelExample exploits positive and negative training examples that are deemed to be more relevant with respect to the test tag t. In particular, relevant positive examples are selected from $\mathcal{S}$ by combining SemanticField and TagVote in a late fusion manner. For negative training example acquisition, they leverage Negative Bootstrap [Li et al. 2013], a negative sampling algorithm that iteratively selects negative examples deemed most relevant for improving classification. A T-iteration Negative Bootstrap will produce T metaclassifiers. The corresponding tag relevance function is written as

$$f_{RelExample}(x, t) := \frac{1}{T} \sum_{l=1}^{T} \left(b_l + \sum_{j=1}^{n_l} \alpha_{l,j} \cdot y_{l,j} \cdot \mathcal{K}(x, x_{l,j}) \right), \tag{12}$$

where $\alpha_{l,j}$ is a positive coefficient of support vector $x_{l,j}$, $y_{l,j} \in \{-1, 1\}$ is class label, and n_l is the number of support vectors in the lth classifier. For the sake of efficiency, the kernel function $\mathcal{K}$ is instantiated with the fast intersection kernel [Maji et al. 2008]. RelExample uses the same amount of positive training examples as TagFeature. The number of iterations T is empirically set to 10. For the SVM classifiers used in TagFeature and RelExample, Platt's scaling [Lin et al. 2007] is employed to convert prediction scores into probabilistic output. In RelExample, for each tag learning a histogram intersection kernel, SVM has a computation cost of $O(dp^2)$ per iteration, and $O(T\,dp^2)$ for T iterations. By jointly using the fast intersection kernel with a quantization factor of q [Maji et al. 2008] and model compression [Li et al. 2013], an order of $O(dq)$ is needed to keep all learned metaclassifiers in memory. Since learning a new classifier needs a memory of $O(dp)$, the overall memory cost for training RelExample is $O(dp + dq)$. For each tag, model compression is applied to its learned ensemble in advance to running Equation (12). As a consequence, the compressed classifier can be cached in an order of $O(dq)$ and executed in an order of $O(d)$.

10. RobustPCA [Zhu et al. 2010]. On the base of robust principal component analysis [Candès et al. 2011], RobustPCA factorizes the image-tag matrix D by a low-rank decomposition with error sparsity. That is,

$$D = \hat{D} + E, \tag{13}$$

where the reconstructed $\hat{D}$ has a low-rank constraint based on the nuclear norm, and E is an error matrix with an ℓ_1-norm sparsity constraint. Notice that the decomposition is not unique. So for a better solution, the decomposition process takes into account image affinities and tag affinities by adding two extra penalties with respect to a Laplacian matrix L_i from the image affinity graph and another Laplacian matrix L_t from the tag affinity graph. Consequently, two hyperparameters λ_1 and λ_2 are introduced to balance the error sparsity and the two Laplacian strengths. We follow the original paper and set the two parameters by performing a grid search on the very same proposed range. To address the tag sparseness, the authors employ a preprocessing step to refine D by a weighted KNN propagation based on the visual similarity. RobustPCA requires an iterative procedure based on the accelerated proximal gradient method with a quadratic convergence rate [Zhu et al. 2010]. Each iteration spends the majority of the required time performing Singular Value Decomposition that, according to Golub and Van Loan [2012], has a well-known complexity of $O(cm^2n + c'n^3)$, where c, c' are constants. Regarding memory, it has a requirement of $O(cn \cdot m + c' \cdot (n^2 + m^2))$ as it needs to maintain a full copy of D and Laplacians of images and labels.

11. TensorAnalysis [Sang et al. 2012a]. This method considers ternary relationships between images, tags, and the user by extending the image-tag association matrix to a binary user-image-tag tensor $F \in \{0, 1\}^{|\mathcal{X}|\times|\mathcal{V}|\times|\mathcal{U}|}$. The tensor is factorized by Tucker decomposition into a dense core C and three low-rank matrices U, I, T, corresponding to the user, image, and tag modalities, respectively:

$$F = C \times_u U \times_i I \times_t T. \tag{14}$$

Here $\times_j$ is the tensor product between a tensor and a matrix along dimension $j \in \{u, i, t\}$. The idea is that C contains the interactions between modalities, while each low-rank matrix represents the main components of each modality. Every modality has to be sized manually or by energy retention, adding three needed parameters $R = (r_I, r_T, r_U)$. The tag relevance scores are obtained by computing $\hat{D} = C \times_i I \times_t T \times_u \mathbf{1}_{r_u}$. Similar to RobustPCA, the decomposition in Equation (14) is not unique and a better solution may be found by regularizing the optimization process with a Laplacian built on a similarity graph for each modality (i.e., L_i, L_t, and L_u) and an ℓ_2 regularizer on each factor (i.e., C, U, I and T). For TensorAnalysis, the complexity is $O(|P_1| \cdot (r_T \cdot m^2 + r_U \cdot r_I \cdot r_T))$, proportional to the number of tags P_1 asserted in D and the dimension of low-rank r_U, r_I, r_T factors. The memory required is $O(n^2 + m^2 + u^2)$ for the Laplacians of images, tags, and users.

4.2. Considerations

An overview of the methods analyzed is given Table IV. Among them, SemanticField, counting solely on the tag modality, has the best scalability with respect to both computation and memory. Among the instance-based methods, TagRanking, which works on selected subsets of $\mathcal{S}$ rather than the entire collection, has the lowest memory request. When the number of tags to be modeled is substantially smaller than the size of $\mathcal{S}$, the model-based methods require less memory and run faster in the test stage, but at the expense of SVM model learning in the training stage. The two transduction-based methods have limited scalability and can operate only on small-sized $\mathcal{S}$.

5. EVALUATION

This section presents our evaluation of the 11 methods according to their applicability to the three tasks using the proposed experimental protocol, that is, KNN, TagVote, TagProp, TagFeature, and RelExample for tag assignment (Section 5.1); TagCooccur, TagCooccur+, RobustPCA, and TensorAnalysis for tag refinement (Section 5.2); and

Table V. Evaluating Methods for Tag Assignment (Given the Same Feature, Bold Values Indicate Top Performers on Individual Test Sets)

	MIRFlickr			NUS-WIDE		
Method	Train10k	Train100k	Train1m	Train10k	Train100k	Train1m
MiAP scores:						
RandomGuess	0.147	0.147	0.147	0.061	0.061	0.061
BovW + KNN	0.232	0.286	0.312	0.171	0.217	0.248
BovW + TagVote	0.276	0.310	**0.328**	0.183	0.231	0.259
BovW + TagProp	0.276	0.299	0.314	0.230	0.249	**0.268**
BovW + TagFeature	0.278	0.294	0.298	0.244	0.221	0.214
BovW + RelExample	0.284	0.309	0.303	0.257	0.233	0.245
CNN + KNN	0.326	0.366	0.379	0.315	0.343	0.376
CNN + TagVote	0.355	0.378	0.389	0.340	0.370	**0.396**
CNN + TagProp	0.373	0.384	**0.392**	0.366	0.376	0.380
CNN + TagFeature	0.359	0.378	0.383	0.367	0.338	0.373
CNN + RelExample	0.309	0.385	0.373	0.365	0.354	0.388
MAP scores:						
RandomGuess	0.072	0.072	0.072	0.023	0.023	0.023
BovW + KNN	0.231	0.282	0.336	0.094	0.139	0.185
BovW + TagVote	0.228	0.280	0.334	0.093	0.137	0.184
BovW + TagProp	0.245	0.293	**0.342**	0.102	0.149	**0.193**
BovW + TagFeature	0.200	0.199	0.201	0.090	0.096	0.098
BovW + RelExample	0.284	0.303	0.310	0.119	0.155	0.172
CNN + KNN	0.564	0.613	0.639	0.271	0.356	0.400
CNN + TagVote	0.561	0.613	0.638	0.257	0.358	**0.402**
CNN + TagProp	0.586	0.619	**0.641**	0.305	0.376	0.397
CNN + TagFeature	0.444	0.554	0.563	0.262	0.310	0.326
CNN + RelExample	0.538	0.603	0.584	0.300	0.346	0.373

all for tag retrieval (Section 5.3). For TensorAnalysis, we were able to evaluate only tag refinement with BovW features on MIRFlickr with Train10k and Train100k. The reason for this exception is that our implementation of TensorAnalysis performs worse than the baseline. Consequently, the results of TensorAnalysis were kindly provided by the authors in the form of tag ranks. Since the provided tag ranks cannot be converted to image ranks, we could not compute MAP scores. A comparison between our Flickr-based training data and ImageNet is given in Section 5.4.

5.1. Tag Assignment

Table V shows the tag assignment performance of KNN, TagVote, TagProp, TagFeature, and RelExample. Their superior performance against the RandomGuess baseline shows that learning purely from social media is meaningful. TagVote and TagProp are the two best-performing methods on both test sets. Substituting CNN for BovW consistently brings improvements for all methods.

In more detail, the following considerations hold. TagProp has higher MAP performance than KNN and TagVote in almost all the cases under analysis. As discussed in Section 4, TagProp is built upon KNN, but it weights the neighbor images by rank and applies a logistic model per tag. Since the logistic model does not affect the image ranking, the superior performance of TagProp should be ascribed to rank-based neighbor weighting. A per-tag comparison on MIRFlickr is given in Figure 2. TagProp is almost always ahead of KNN and TagVote. Concerning TagVote and KNN, recall that their main difference is that TagVote applies the unique-user constraint on the neighborhood

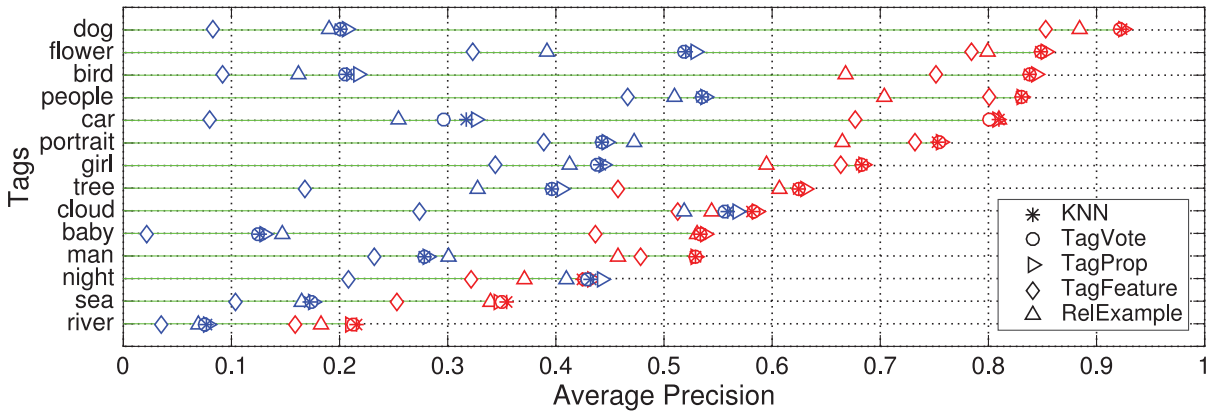


Fig. 2. **Per-tag comparison of methods for tag assignment on MIRFlickr**, trained on Train1m. The colors identify the features used: **blue** for BovW, **red** for CNN. The test tags have been sorted in descending order by the performance of CNN + TagProp.

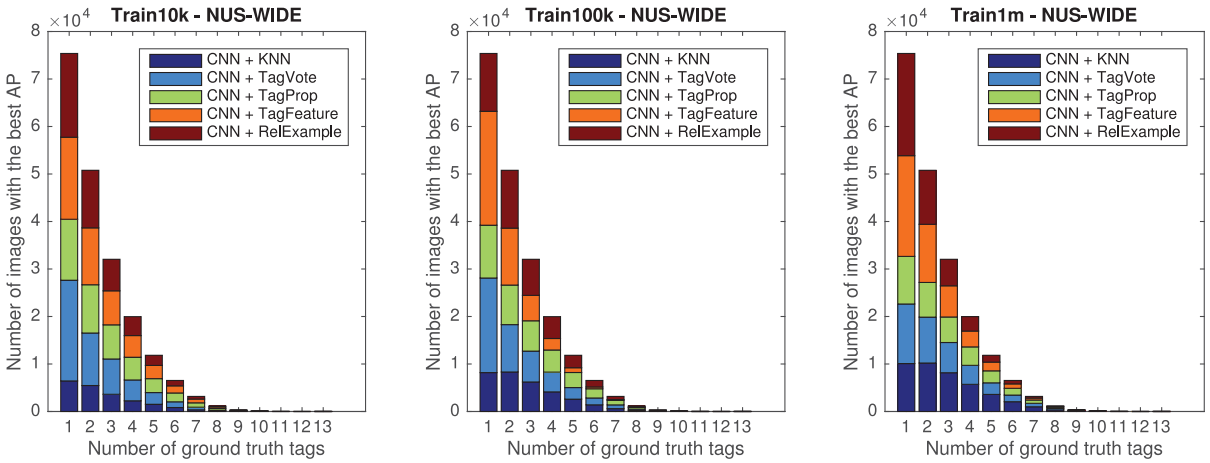


Fig. 3. **Per-image comparison of methods for tag assignment on NUS-WIDE**. Test images are grouped in terms of their number of ground-truth tags. The area of a colored bar is proportional to the number of images for which the corresponding method scores best.

and employs tag prior as a penalty term. The fact that the training data contains no batch-tagged images minimizes the influence of the unique-user constraint. While the penalty term does not affect image ranking for a given tag, it affects tag ranking for a given image. This explains why KNN and TagVote have mostly the same MAP. Also, the result suggests that the tag-prior-based penalty is helpful for doing tag assignment by neighbor voting.

We observe that RelExample has a better MAP than TagFeature in every case. The absence of a filtering component makes TagFeature more likely to overfit to training examples irrelevant to the test tags. For the other two model-based methods, the overfit issue is alleviated by different strategies: RelExample employs a filtering component to select more relevant training examples, while TagProp has fewer parameters to tune.

A per-image comparison on NUS-WIDE is given in Figure 3. The test images are put into disjoint groups so that images within the same group have the same number of ground-truth tags. For each group, the area of the colored bars is proportional to the number of images on which the corresponding methods score best. The first group (i.e., images containing only one ground-truth tag) has the most noticeable change as the training set grows. There are 75,378 images in this group, and for 39% of the images,

Table VI. Evaluating Methods for Tag Refinement

	MIRFlickr			NUS-WIDE		
Method	Train10k	Train100k	Train1m	Train10k	Train100k	Train1m
MiAP scores:						
UserTags	0.204	0.204	0.204	0.255	0.255	0.255
TagCooccur	0.213	0.242	0.253	0.269	0.305	0.317
BovW + TagCooccur+	0.217	0.262	0.286	0.245	0.297	0.324
BovW + RobustPCA	0.271	**0.310**	–	**0.332**	0.323	–
BovW + TensorAnalysis	*0.298	*0.297	–	–	–	–
CNN + TagCooccur+	0.234	0.277	0.310	0.305	0.359	0.387
CNN + RobustPCA	0.368	**0.376**	–	**0.424**	0.419	–
CNN + TensorAnalysis	–	–	–	–	–	–
MAP scores:						
UserTags	0.263	0.263	0.263	0.338	0.338	0.338
TagCooccur	0.266	0.298	0.313	0.223	0.321	0.308
BovW + TagCooccur+	0.294	0.343	**0.377**	0.231	0.345	**0.353**
BovW + RobustPCA	0.225	0.337	–	0.229	0.234	–
BovW + TensorAnalysis	–	–	–	–	–	–
CNN + TagCooccur+	0.330	0.381	0.420	0.264	0.391	0.406
CNN + RobustPCA	0.566	**0.627**	–	0.439	**0.440**	–
CNN + TensorAnalysis	–	–	–	–	–	–

The asterisk (*) indicates results provided by the authors of the corresponding methods, while the dash (–) means we were unable to produce results. Given the same feature, bold values indicate top performers on individual test sets per performance metric.

their single label is "person." When Train1m is used, RelExample beats KNN, TagVote, and TagProp for this frequent label. This explains the leading position of RelExample in the first group. The result also confirms our earlier discussion in Section 3.3 that MiAP is likely to be biased by frequent tags.

In summary, as long as enough training examples are provided, instance-based methods are on par with model-based methods for tag assignment. Model-based methods are more suited when the training data is of limited availability. However, they are less resilient to noise, and consequently a proper filtering strategy for refining the training data becomes essential.

5.2. Tag Refinement

Table VI shows the performance of different methods for tag refinement. We were unable to complete the table. In particular, RobustPCA could not go over 350,000 images due to its high demand in both CPU time and memory (see Table IV), while TensorAnalysis was provided by the authors only on MIRFlickr with Train10k, Train100k, and the BovW feature.

RobustPCA outperforms the competitors on both test sets when provided with the CNN feature. Figure 4 presents a per-tag comparison on MIRFlickr. RobustPCA has the best scores for nine out of the 14 tags with BovW and wins all the tags when CNN is used.

Concerning the influence of the media dimension, the tag + image-based methods (RobustPCA and TagCooccur+) are in general better than the tag-based method (TagCooccur). As shown in Figure 4, except for three out of 14 MIRFlickr test tags with BovW, using the image media is beneficial. As in the tag assignment task, the use of the CNN feature strongly improves the performance.

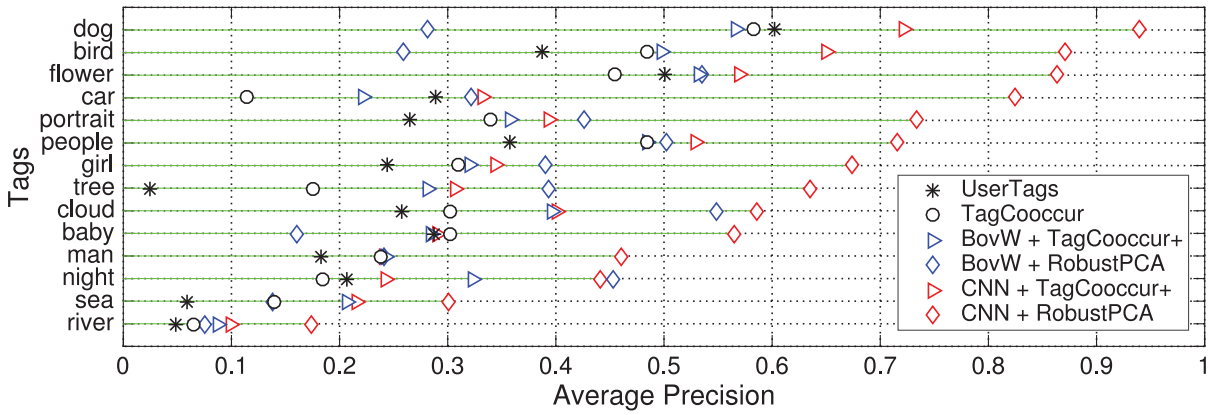


Fig. 4. **Per-tag comparison of methods for tag refinement on MIRFlickr**, trained on Train100k. The colors identify the features used: **blue** for BovW, **red** for CNN. The test tags have been sorted in descending order by the performance of CNN + RobustPCA.

Concerning the learning methods, TensorAnalysis has the potential to leverage tag, image, and user simultaneously. However, due to its relatively poor scalability, we were able to run this method only with Train10k and Train100k on MIRFlickr. For Train10k, TensorAnalysis yielded higher MiAP than RobustPCA, probably thanks to its capability of modeling user correlations. It is outperformed by RobustPCA when more training data is used.

As more training data is used, the performance of TagCooccur, TagCooccur+, and RobustPCA on MIRFlickr consistently improves. Since these three methods rely on data-driven tag affinity, image affinity, or tag and image affinity, a small set of 10,000 images is generally inadequate to compute these affinities. The effect of increasing the training set size is clearly visible if we compare scores corresponding to Train10k and Train100k. The results on NUS-WIDE show some inconsistency. For TagCooccur, MiAP improves from Train100k to Train1m, while MAP drops. This is presumably due to the fact that in the experiments, we used the parameters recommended in the original paper, appropriately selected to optimize tag ranking. Hence, they might be suboptimal for image ranking. BovW + RobustPCA scores a lower MAP than BovW + TagCooccur+. This is probably due to the fact that the low-rank matrix factorization technique, while being able to jointly exploit tag and image information, is more sensitive to the content-based representation.

A per-image comparison is given in Figure 5. As for tag assignment, the test images have been grouped according to the number of ground-truth tags associated. The size of the colored areas is proportional to the number of images where the corresponding method scores best. For the majority of test images, the three tag refinement methods have higher average precision than UserTags. This means more relevant tags are added, so the tags are refined. It should be noted that the success of tag refinement depends much on the quality of the original tags assigned to the test images. Examples are shown in Table VII: in row 6, although the tag “earthquake” is irrelevant to the image content, it is ranked at the top by RobustPCA. To what extent a tag refinement method shall count on the existing tags is tricky.

To summarize, the tag + image-based methods outperform the tag-based method for tag refinement. RobustPCA is the best and improves as more training data is employed. Nonetheless, implementing RobustPCA is challenging for both computation and memory footprint. In contrast, TagCooccur+ is more scalable and it can learn from large-scale data.

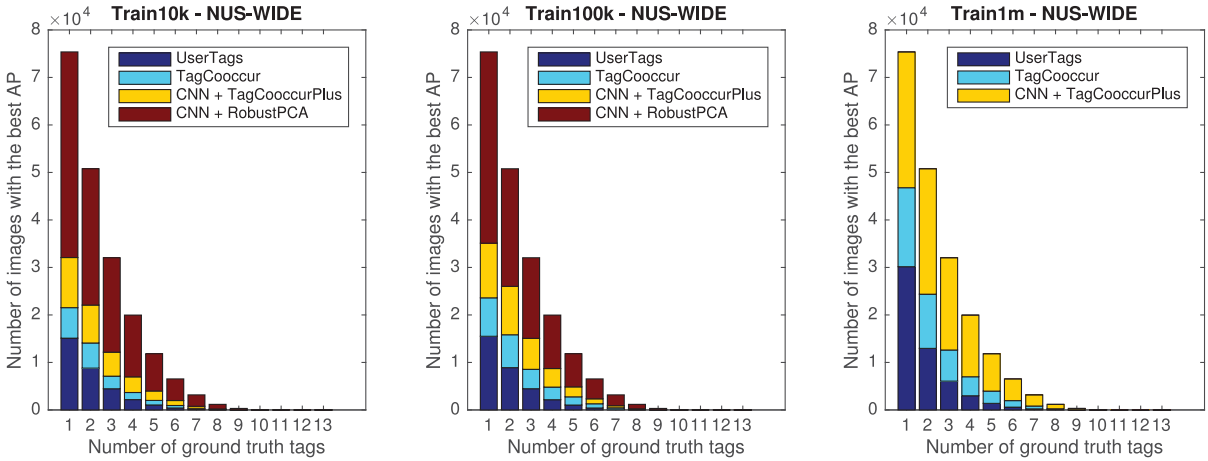


Fig. 5. **Per-image comparison of methods for tag refinement on NUS-WIDE**. Test images are grouped in terms of their number of ground-truth tags. The area of a colored bar is proportional to the number of images for which the corresponding method scores best.

5.3. Tag Retrieval

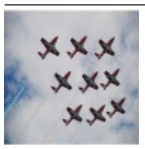
Table VIII shows the performance of different methods for tag retrieval. Recall that when retrieving images for a specific test tag, we consider only images that are labeled with this tag. Hence, MAP scores here are higher than their counterpart in Table VI.

We start our analysis by comparing the three baselines, namely, UserTags, TagNum, and TagPosition, which retrieve images simply by the original tags. As can be noticed, TagNum and TagPosition are more effective than UserTags, TagNum outperforms TagPosition on Flickr51, and the latter has better scores on NUS-WIDE. The effectiveness of such metadata-based features depends much on datasets and are unreliable for tag retrieval.

All the methods considered have higher MAP than the three baselines. All the methods have better performance than the baselines on Flickr51, and performance increases with the size of the training set. On NUS-WIDE, SemanticField, TagCooccur, and TagRanking are less effective than TagPosition. We attribute this result to the fact that, for these methods, the tag relevance functions favor images with fewer tags. So they closely follow similar performance and dataset dependency.

Concerning the influence of the media dimension, the tag + image-based methods (KNN, TagVote, TagProp, TagCooccur+, TagFeature, RobustPCA, RelExample) are in general better than the tag-based method (SemanticField and TagCooccur). Figure 6 shows the per-tag retrieval performance on Flickr51. For 33 out of the 51 test tags, RelExample exhibits average precision higher than 0.9. By examining the top retrieved images, we observe that the results produced by tag + image-based methods and tag-based methods are complementary to some extent. For example, consider “military,” one of the test tags of NUS-WIDE. RelExample retrieves images with strong visual patterns such as military vehicles, while SemanticField returns images of military personnel. Since the visual content is ignored, the results of SemanticField tend to be visually different, making it possible to handle tags with visual ambiguity. This fact can be observed in Figure 7, which shows the top 10 ranked images of “jaguar” by TagPosition, SemanticField, BovW + RelExample, and CNN + RelExample. Although their results are all correct, RelExample finds Jaguar-brand cars only, while SemanticField covers both cars and animals. However, for a complete evaluation of the capability of managing ambiguous tags, fine-grained ground truth beyond what we currently have is required.

Table VII. Selected Tag Assignment and Refinement Results on NUS-WIDE

Test Image	Ground Truth	User Tags	Tag Assignment				Tag Refinement		
			KNN	TagVote	TagProp	RelExample	TagCooccur	TagCooccur+	RobustPCA
	sign	***sign*** reptile zoo red white	animal flower car horse street	dog house bird ***sign*** bear	***sign*** street flower dog bird	soccer whale book toy moon	animal street ***sign*** water car	***sign*** bird dog animal toy	***sign*** bird flower animal street
	animal dog person	colour color ***dog*** hound	flower garden horse tree ***dog***	garden flower food cat ***dog***	flower ***dog*** garden car tree	garden ***dog*** fish fox ***animal***	***dog*** ***animal*** car beach flower	***dog*** flower ***animal*** cat food	***dog*** flower ***animal*** water garden
	cloud grass sky	***cloud*** ***grass***	***cloud*** ***sky*** beach water snow	***cloud*** ***sky*** water beach mountain	***cloud*** ***sky*** beach water lake	***cloud*** ocean surf ***sky*** beach	***grass*** ***sky*** tree flower water	***cloud*** ***sky*** water beach tree	***cloud*** ***grass*** ***sky*** water mountain
	animal bear water	brown ***bear*** salmon national park	snow beach ***animal*** ***water*** tree	snow ***animal*** waterfall tree ***water***	snow beach sand ***bear*** ***water***	***water*** sand rock surf ocean	waterfall ***water*** tree ***bear*** ***animal***	waterfall ***water*** ***animal*** snow tree	***water*** waterfall ***bear*** ***animal*** snow
	airplane cloud military sky	flag great	***sky*** ***cloud*** snow bird ***airplane***	snow ***cloud*** ***sky*** mountain bird	***airplane*** ***sky*** snow bird airport	snow frost bird ***airplane*** tattoo	car street snow water ***sky***	snow ***sky*** ***cloud*** mountain bird	flag ***sky*** snow ***cloud*** bird
	cloud garden sky water	china earthquake people hangzhou summer westlake	car beach ***water*** street tree	grass tree ***water*** road bridge	car road street ***sky*** bird	house road grass bird sand	***water*** flower street temple tree	tree ***water*** street ***garden*** car	earthquake ***water*** tree ***cloud*** ***sky***
	police road vehicle window	farmer dog motorcycle ***police*** train	car street ***police*** ***vehicle*** ***road***	car street ***police*** ***vehicle*** sport	***police*** car street ***road*** sport	***police*** ***vehicle*** street car sport	street car animal train bird	car street ***police*** food horse	***police*** train dog bird car
	airplane airport cloud military sky vehicle	vertical sunglass smoke pilot landing	car beach street water ***airplane***	car street sport ***airplane*** ***vehicle***	car sport ***airplane*** ***vehicle*** road	***airplane*** sport ***airport*** ***vehicle*** ***military***	***airplane*** car ***military*** ***airport*** street	car ***airplane*** street ***airport*** ***military***	***airplane*** car ***sky*** ***cloud*** water
	animal grass horse	***horse*** pony run field brown	waterfall tree garden water ***horse***	***animal*** tree ***horse*** garden waterfall	garden ***grass*** ***horse*** tree waterfall	cow elk ***animal*** ***grass*** ***horse***	***horse*** car ***animal*** street dog	***animal*** ***horse*** tree dog water	***animal*** ***horse*** tree water flower

Visual feature: BovW. The top five ranked tags are shown, with correct prediction marked by the ***bold italic*** font.

Concerning the learning methods, TagVote consistently performs well, as in the tag assignment experiment. KNN is comparable to TagVote, for the reason we have discussed in Section 5.1. Given the CNN feature, the two methods even outperform their model-based variant TagProp. Similar to the tag refinement experiment, the effectiveness of RobustPCA for tag retrieval is sensitive to the choice of visual features. While BovW + RobustPCA is worse than the majority on Flickrt51, the performance of CNN + RobustPCA is more stable and performs well. For TagFeature, its gain from

Table VIII. Evaluating Methods for Tag Retrieval

	Flickr51			NUS-WIDE		
Method	Train10k	Train100k	Train1m	Train10k	Train100k	Train1m
MAP scores:						
UserTags	0.595	0.595	0.595	0.489	0.489	0.489
TagNum	0.664	0.664	0.664	0.520	0.520	0.520
TagPosition	0.640	0.640	0.640	0.557	0.557	0.557
SemanticField	0.687	0.707	0.713	0.565	0.584	0.584
TagCooccur	0.625	0.679	0.704	0.534	0.576	0.588
BovW + TagCooccur+	0.640	0.732	0.764	0.560	0.622	0.643
BovW + TagRanking	0.685	0.686	0.708	0.557	0.574	0.578
BovW + KNN	0.678	0.742	0.770	0.587	0.632	0.658
BovW + TagVote	0.678	0.741	0.769	0.587	0.632	0.659
BovW + TagProp	0.671	0.748	0.772	0.585	0.636	0.657
BovW + TagFeature	0.689	0.726	0.737	0.589	0.602	0.606
BovW + RelExample	0.706	0.756	**0.783**	0.609	0.645	**0.663**
BovW + RobustPCA	0.697	0.701	–	0.650	0.650	–
BovW + TensorAnalysis	–	–	–	–	–	–
CNN + TagCooccur+	0.654	0.781	0.821	0.572	0.653	0.674
CNN + TagRanking	0.744	0.735	0.747	0.589	0.590	0.590
CNN + KNN	0.811	0.859	0.880	0.683	0.722	0.734
CNN + TagVote	0.808	0.859	**0.881**	0.675	0.724	**0.738**
CNN + TagProp	0.824	0.867	0.879	0.689	0.727	0.731
CNN + TagFeature	0.827	0.853	0.859	0.675	0.700	0.703
CNN + RelExample	0.838	0.863	0.878	0.689	0.717	0.734
CNN + RobustPCA	0.811	0.839	–	0.725	0.726	–
CNN + TensorAnalysis	–	–	–	–	–	–
$NDCG_{20}$ scores:						
UserTags	0.432	0.432	0.432	0.487	0.487	0.487
TagNum	0.522	0.522	0.522	0.541	0.541	0.541
TagPosition	0.511	0.511	0.511	0.623	0.623	0.623
SemanticField	0.591	0.623	0.645	0.596	0.622	0.624
TagCooccur	0.482	0.527	0.631	0.529	0.602	0.614
BovW + TagCooccur+	0.503	0.625	0.686	0.590	0.681	0.734
BovW + TagRanking	0.530	0.568	0.571	0.557	0.572	0.572
BovW + KNN	0.577	0.699	0.756	0.638	0.734	0.799
BovW + TagVote	0.573	0.701	0.754	0.629	0.734	0.804
BovW + TagProp	0.570	0.715	**0.759**	0.666	0.750	**0.809**
BovW + TagFeature	0.547	0.626	0.646	0.622	0.615	0.618
BovW + RelExample	0.614	0.722	0.748	0.692	0.736	0.776
BovW + RobustPCA	0.549	0.548	–	0.768	0.781	–
BovW + TensorAnalysis	–	–	–	–	–	–
CNN + TagCooccur+	0.504	0.615	0.724	0.571	0.705	0.738
CNN + TagRanking	0.577	0.607	0.597	0.578	0.594	0.583
CNN + KNN	0.709	0.830	0.897	0.773	0.832	0.863
CNN + TagVote	0.722	0.826	**0.899**	0.740	0.837	**0.879**
CNN + TagProp	0.768	0.857	0.865	0.764	0.839	0.845
CNN + TagFeature	0.755	0.813	0.818	0.704	0.807	0.787
CNN + RelExample	0.764	0.843	0.879	0.773	0.814	0.866
CNN + RobustPCA	0.733	0.821	–	0.865	0.862	–
CNN + TensorAnalysis	–	–	–	–	–	–

Given the same feature, bold values indicate top performers on individual test sets per performance metric.

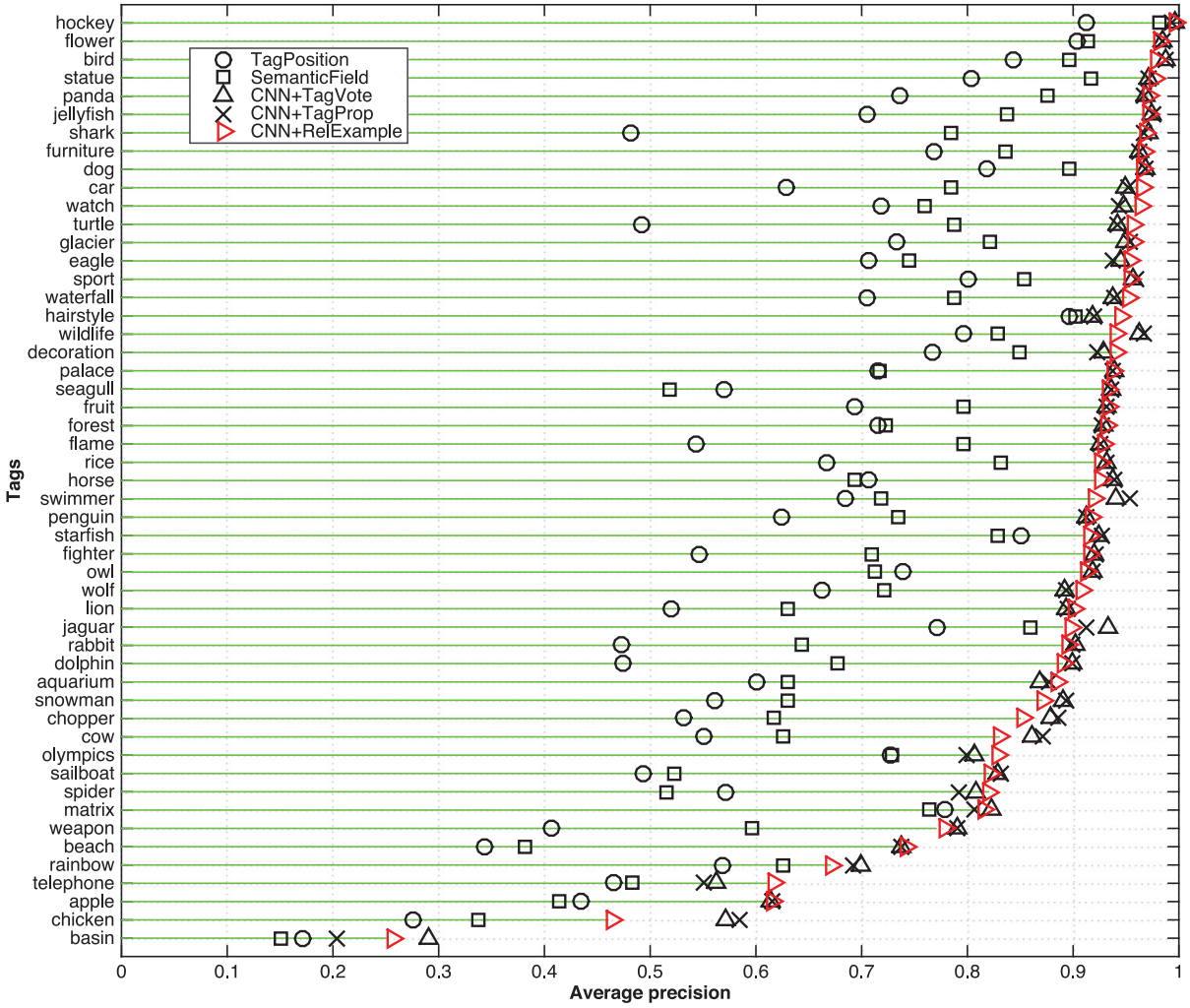


Fig. 6. **Per-tag comparison between TagPosition, SemanticField, TagVote, TagProp, and RelExample on Flickr51**, with Train1m as the training set. The 51 test tags have been sorted in descending order by the performance of RelExample.

using larger training data is relatively limited due to the absence of denoising. In contrast, RelExample, by jointly using SemanticField and TagVote in its denoising component, is consistently better than TagFeature.

The performance of individual methods consistently improves as more training data is used. As the size of the training set increases, the performance gap between the best model-based method (RelExample) and the best instance-based method (TagVote) reduces. This suggests that large-scale training data diminishes the advantage of model-based methods against the relatively simple instance-based methods.

In summary, even though the performance of the methods evaluated varies over datasets, common patterns have been observed. First, the more social data for training are used, the better performance is obtained. Since the tag relevance functions are learned purely from social data without any extra manual labeling, and social data are increasingly growing, this result promises that better tag relevance functions can be learned. Second, given small-scale training data, tag + image-based methods that conduct model-based learning with denoised training examples turn out to be the most effective solution. This, however, comes with a price of reducing the visual diversity in the retrieval results. Moreover, the advantage of model-based learning

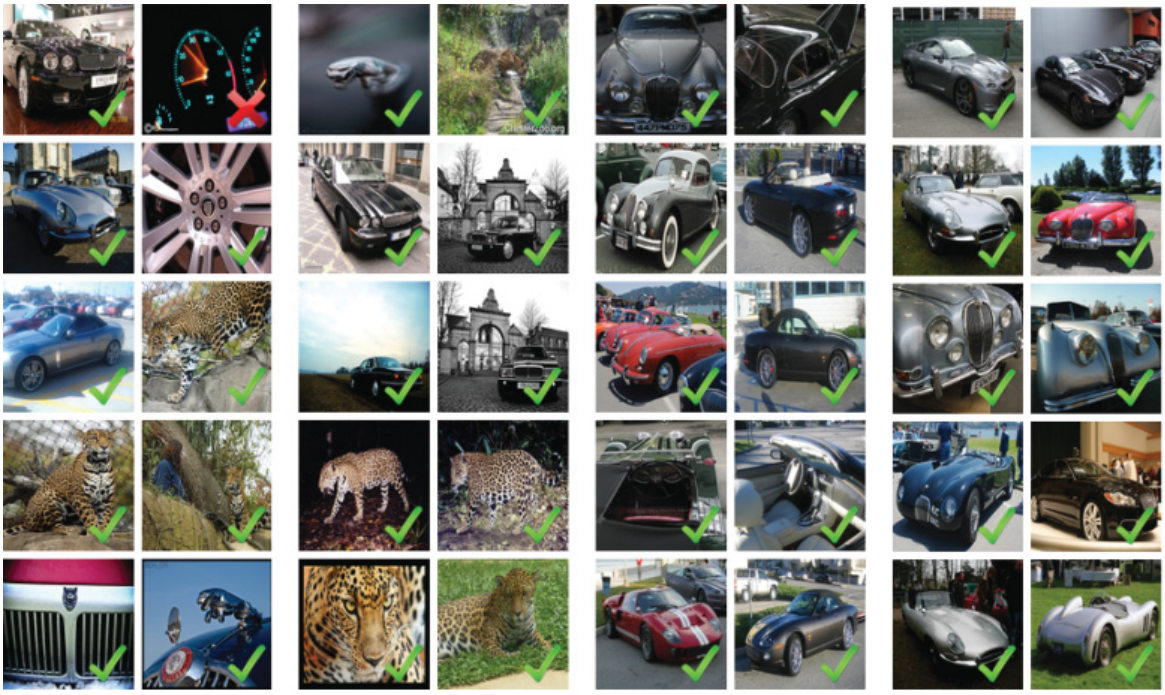


Fig. 7. **Top 10 ranked images of "jaguar," by (a) TagPosition, (b) SemanticField, (c) BovW + RelExample, and (d) CNN + RelExample**. Checkmarks (✓) indicate relevant results. While both RelExample and SemanticField outperform the TagPosition baseline, the results of SemanticField show more diversity for this ambiguous tag. The difference between (c) and (d) suggests that the results of RelExample can be diversified by varying the visual feature in use.

vanishes as more training data and the CNN feature are used, and TagVote performs the best.

5.4. Flickr Versus ImageNet

To address the question of whether one shall resort to an existing resource such as ImageNet for tag relevance learning, this section presents an empirical comparison between our Flickr-based training data and ImageNet. A number of methods do not work with ImageNet or require modifications. For instance, tag + image + user information-based methods must be able to remove their dependency on user information, as such information is unavailable in ImageNet. Tag co-occurrence statistics are also strongly limited, because an ImageNet example is annotated with a single label. Because of these limitations, we evaluate only the two best-performing methods, TagVote and TagProp. TagProp can be directly used since it comes from classic image annotation, while TagVote is slightly modified by removing the unique user constraint. The CNN feature is used for its superior performance against the BovW feature.

To construct a customized subset of ImageNet that fits the three test sets, we take ImageNet examples whose labels precisely match with the test tags. Notice that some test tags (e.g., "portrait" and "night") have no match, while some other tags (e.g, "car" and "dog") have more than one match. In particular, MIRFlickr has two missing tags, while the number of missing tags on Flickr51 and NUS-WIDE is nine and 15, respectively. For a fair comparison these missing tags are excluded from the evaluation. Putting the remaining test tags together, we obtain a subset of ImageNet, containing 166 labels and over 200k images, termed ImageNet200k.

The left half of Table IX shows the performance of tag assignment. TagVote/TagProp trained on the ImageNet data are less effective than their counterparts trained on the

Table IX. Flickr Versus ImageNet

Tag Assignment					Tag Retrieval				
	MIRFlickr		NUS-WIDE			Flickr51		NUS-WIDE	
Training Set	TagVote	TagProp	TagVote	TagProp	**Training Set**	TagVote	TagProp	TagVote	TagProp
MiAP scores:					***MAP scores:***				
Train100k	0.377	0.383	0.392	0.389	Train100k	0.854	0.860	0.742	0.745
Train1M	0.389	**0.392**	**0.414**	0.393	Train1M	**0.874**	0.871	0.753	0.745
ImageNet200k	0.345	0.304	0.325	0.368	ImageNet200k	0.873	0.873	**0.762**	**0.762**
MAP scores:					***$NDCG_{20}$ scores:***				
Train100k	0.641	0.647	0.386	0.405	Train100k	0.838	0.863	0.849	0.856
Train1M	0.664	**0.668**	**0.429**	0.420	Train1M	0.894	0.851	**0.891**	0.853
ImageNet200k	0.532	0.532	0.363	0.362	ImageNet200k	**0.920**	0.898	0.843	0.847

NoticeThat the numbers on Train100k and Train1M are different from Tables V and VIII due to the use of a reduced set of test tags. Bold values indicate top performers on a specific test set per performance metric.

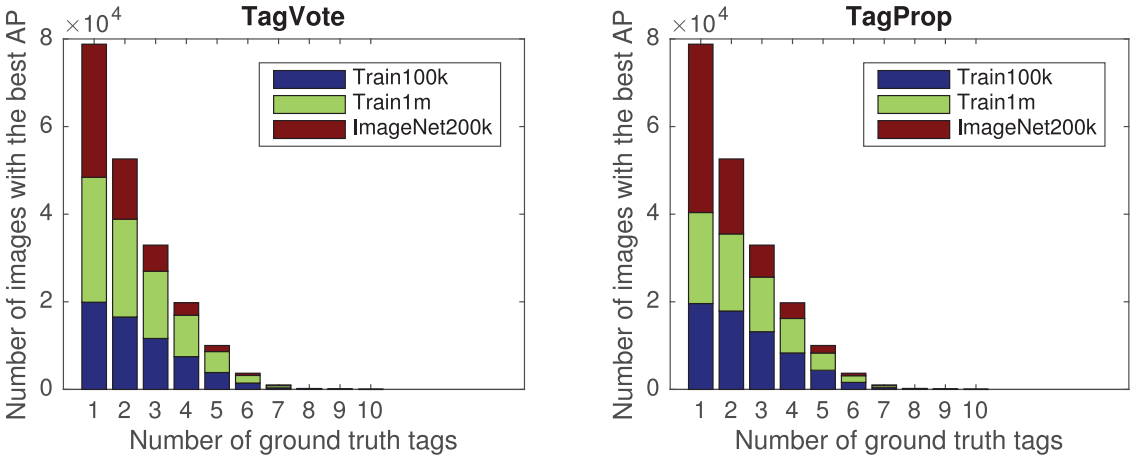


Fig. 8. **Per-image comparison of TagVote/TagProp learned from different training datasets**, tested on NUS-WIDE. Test images are grouped in terms of the number of ground-truth tags. Within each group, the area of a colored bar is proportional to the number of images for which (the method derived from) the corresponding training dataset scores the best. ImageNet200k is less effective for assigning multiple labels to an image.

Flickr data. For a better understanding of the result, we employ the same visualization technique as used in Section 5.1, that is, grouping the test images in terms of the number of their ground-truth tags, and subsequently checking the performance per group. As shown in Figure 8, while ImageNet200k performs better on the first group (i.e., images with a single relevant tag), it is outperformed by Train100k and Train1M on the other groups. For its single-label nature, ImageNet is less effective for assigning multiple labels to an image.

For tag retrieval, as shown in the right half of Table IX, TagVote and TagProp learned from ImageNet200k in general have higher MAP and NDCG scores than their counterparts learned from the Flickr data. By comparing the performance difference per concept, we find that the gain is largely contributed by a relatively small amount of concepts. Consider, for instance, TagVote + ImageNet200k and TagVote + Train1M on NUS-WIDE. The former outperforms the latter for 25 out of the 66 tested concepts. By sorting the concepts according to their absolute performance gain, the top three winning concepts of TagVote + ImageNet200k are “sand,” “garden,” and “rainbow,” with AP gain of 0.391, 0.284, and 0.176, respectively. Here, the lower performance of TagVote + Train1M is largely due to the subjectiveness of social tagging. For instance, Flickr images labeled with “sand” tend be much more diverse, showing a wide range of things

visually irrelevant to sand. Interestingly, the top three losing concepts of TagVote + ImageNet200k are "running," "valley," and "building," with AP loss of 0.150, 0.107, and 0.090, respectively. For these concepts, we observe that their ImageNet examples lack diversity. For example, "running" in ImageNet200k mostly shows a person running on a track. In contrast, the subjectiveness of social tagging now has a positive effect on generating diverse training examples.

In summary, for tag assignment, social media examples are a preferred resource of training data. For tag retrieval, ImageNet yields better performance, yet the performance gain is largely due to a few tags where social tagging is very noisy. In such a case, controlled manual labeling seems indispensable. In contrast, with clever tag relevance learning algorithms, social training data demonstrate competitive or even better performance for many of the tested tags. Nevertheless, where the boundary between the two cases is precisely located remains unexplored.

6. CONCLUSIONS AND PERSPECTIVES

6.1. Concluding Remarks

This article presents a survey on image tag assignment, refinement, and retrieval, with the hope of illustrating connections and differences between the many methods and their applicabilities, and consequently helping the interested audience to either pick up an existing method or devise a method of their own given the data at hand. As the topics are being actively studied, inevitably this survey will miss some papers. Nevertheless, it provides a unified view of many existing works, and consequently eases the effort of placing future works in a proper context, both theoretically and experimentally.

Based on the key observation that all works rely on tag relevance learning as the common ingredient, existing works, which vary in terms of their methodologies and target tasks, have been interpreted in a unified framework. Consequently, a two-dimensional taxonomy has been developed, allowing us to structure the growing literature in light of what information a specific method exploits and how the information is leveraged in order to produce their tag relevance scores. Having established the common ground between methods, a new experimental protocol has been introduced for a head-to-head comparison with the state of the art. A selected set of 11 representative works was implemented and evaluated for tag assignment, refinement, and/or retrieval. The evaluation justifies the state of the art on the three tasks.

Concerning what media is essential for tag relevance learning, tag + image is consistently found to be better than tag alone. While the joint use of tag, image, and user information (via TensorAnalysis) demonstrates its potential on small-scale datasets, it becomes computationally prohibitive as the dataset size increases to 100,000 and beyond. Comparing the three learning strategies, instance-based and model-based methods are found to be more reliable and scalable than their transduction-based counterparts. As model-based methods are more sensitive to the quality of social image tagging, a proper filtering strategy for refining the training media is crucial for their success. Despite their leading performance on the small training dataset, we find that the performance gain over the instance-based alternatives diminishes as more training data is used. Finally, the CNN feature used as a substitute for the BovW feature brings considerable improvements for all the tasks.

Much progress has been made. Given the current test tag set, the best-performing methods already outperform user-provided tags for tag assignment (MiAP of 0.392 vs. 0.204 on MIRFlickr and 0.396 vs. 0.255 on NUS-WIDE). Image retrieval using learned tag relevance also yields more accurate results compared to image retrieval using original tags (MAP of 0.881 vs. 0.595 on Flickr55 and 0.738 vs. 0.489 on NUS-WIDE). For tag assignment and tag retrieval, methods that exploit tag + image media

by instance-based learning take the leading position. In particular, for tag assignment, TagProp and TagVote perform best. For tag retrieval, TagVote achieves the best overall performance. Methods that exploit tag + image by transduction-based learning are more suited for tag refinement. RobustPCA is the choice for this task. These baselines need to be compared against when one advocates a new method.

6.2. Reflections on Future Work

Much remains to be done. Several exciting recent developments open up new opportunities for the future. First, employing novel deep-learning-based visual features is likely to boost the performance of the tag + image-based methods. What is scientifically more interesting is to devise a learning strategy that is capable of jointly exploiting tag, image, and user information in a much more scalable manner than currently feasible. The importance of the filter component, which refines socially tagged training examples in advance to learning, is underestimated. As denoising often comes with the price of reducing visual diversity, more research attention is required to understand what an acceptable level of noise shall be for learning tag relevance. Having a number of collaboratively labeled resources publicly available, research on joint exploration of social data and these resources is important. This connects to the most fundamental aspect of content-based image retrieval in the context of sharing and tagging within social media platforms: to what extent a social tag can be trusted remains open. Image retrieval by multitag query is another important yet largely unexplored problem. For a query of two tags, it is suggested to view the two tags as a single bigram tag [Li et al. 2012; Nie et al. 2012; Borth et al. 2013], which is found to be superior to late fusion of individual tag scores. Nonetheless, due to the increasing sparseness of n-grams, how to effectively answer generic queries of more than two tags is challenging. Test tags in the current benchmark sets were picked based on availability. It would be relevant to study what motivates people to search images on social media platforms and how the search is conducted. We have not seen any quantitative study in this direction. Last but not least, fine-grained ground truth that enables us to evaluate various tag relevance learning methods for answering ambiguous tags is currently missing.

"One way to resolve the semantic gap comes from sources outside the image" Smeulders et al. [2000] wrote at the end of their seminal paper. While what such sources would be was mostly unknown at that time, it is now becoming evident that the many images shared and tagged in social media platforms are promising to resolve the semantic gap. By adding new relevant tags, refining the existing ones, or directly addressing retrieval, the access to the semantics of the visual content has been much improved. This is achieved only when appropriate care is taken to attack the unreliability of social tagging.

ACKNOWLEDGMENTS

The authors thank Dr. Jitao Sang for providing the TensorAnalysis results, and Dr. Meng Wang and Dr. Yue Gao for making the Flickr51 dataset available for this survey.

REFERENCES

Morgan Ames and Mor Naaman. 2007. Why we tag: Motivations for annotation in mobile and online media. In *Proc. of ACM CHI*. 971–980.

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Support vector machines for multiple-instance learning. In *Proc. of NIPS*. 561–568.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16, 6 (2010), 345–379.

Lamberto Ballan, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2015. Data-driven approaches for social image and video tagging. *Multimedia Tools and Applications* 74, 4 (2015), 1443–1468.

Lamberto Ballan, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo. 2014. A cross-media model for automatic image annotation. In *Proc. of ACM ICMR*. 73–80.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. of ACM MM*. 223–232.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM* 58, 3 (2011), 11.

Lin Chen, Dong Xu, Ivor W. Tsang, and Jiebo Luo. 2012. Tag-based image retrieval improved by augmented features and group-based refinement. *IEEE Transactions on Multimedia* 14, 4 (2012), 1057–1067.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *Proc. of ACM CIVR*. 48:1–48:9.

Rudi L. Cilibrasi and Paul M. B. Vitanyi. 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19, 3 (2007), 370–383.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *Computing Surveys* 40, 2 (2008), 5:1–5:60.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR*. 248–255.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé, III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *Proc. of NAACL*. 762–772.

Kun Duan, David J. Crandall, and Dhruv Batra. 2014. Multimodal learning in loosely-organized web images. In *Proc. of CVPR*. 2465–2472.

Lixin Duan, Wen Li, Ivor Wai-Hung Tsang, and Dong Xu. 2011. Improving web image search by bag-based reranking. *IEEE Transactions on Image Processing* 20, 11 (2011), 3280–3290.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision* 111, 1 (2015), 98–136.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

Songhe Feng, Congyan Lang, and Bing Li. 2012. Towards relevance and saliency ranking of image tags. In *Proc. of ACM MM*. 917–920.

Zheyun Feng, Songhe Feng, Rong Jin, and Anil K. Jain. 2014. Image tag completion by noisy matrix recovery. In *Proc. of ECCV*. 424–438.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* 4 (2003), 933–969.

Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2013. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* 22, 1 (2013), 363–376.

Alexandru Lucian Ginsca, Adrian Popescu, Bogdan Ionescu, Anil Armagan, and Ioannis Kanellos. 2014. Toward an estimation of user tagging credibility for social image retrieval. In *Proc. of ACM MM*. 1021–1024.

Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32, 2 (2006), 198–208.

Gene H. Golub and Charles F. Van Loan. 2012. *Matrix Computations*. Johns Hopkins University Press.

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. 2009. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. of ICCV*. 309–316.

Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. *SIGKDD Explorations Newsletter* 12, 1 (2010), 58–72.

Xian-Sheng Hua, Linjun Yang, Jingdong Wang, Jing Wang, Ming Ye, Kuansan Wang, Yong Rui, and Jin Li. 2013. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proc. of ACM MM*. 243–252.

Mark J. Huiskes, Bart Thomee, and Michael S. Lew. 2010. New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative. In *Proc. of ACM MIR*. 527–536.

Fouzia Jabeen, Shah Khusro, Amna Majid, and Azhar Rauf. 2016. Semantics discovery in social tagging systems: A review. *Multimedia Tools and Applications* 75, 1 (2016), 573–605.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Intelligent Systems and Technology* 20, 4 (2002), 422–446.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2011. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 117–128.

Yu-Gang Jiang, Chong-Wah Ngo, and Shih-Fu Chang. 2009. Semantic context transfer across heterogeneous sources for domain adaptive video search. In *Proc. of ACM MM*. 155–164.

Yohan Jin, Latifur Khan, Lei Wang, and Mamoun Awad. 2005. Image annotations by combining multiple evidence & wordNet. In *Proc. of ACM MM*. 706–715.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proc. of ICML*. 200–209.

Justin Johnson, Lamberto Ballan, and Li Fei-Fei. 2015. Love thy neighbors: Image annotation by exploiting image metadata. In *Proc. of ICCV*.

Mahdi M. Kalayeh, Haroon Idrees, and Mubarak Shah. 2014. NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization. In *Proc. of CVPR*. 184–191.

Lyndon S. Kennedy, Shih-Fu Chang, and Igor V. Kozintsev. 2006. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *Proc. of ACM MIR*. 249–258.

Lyndon S. Kennedy, Malcolm Slaney, and Kilian Weinberger. 2009. Reliable tags using image similarity: Mining specificity and expertise from large-scale multimedia databases. In *Proc. of ACM MM Workshop on Web-Scale Multimedia Corpus*. 17–24.

Gunhee Kim and Eric P. Xing. 2013. Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In *Proc. of ACM WSDM*. 163–172.

Yin-Hsi Kuo, Wen-Huang Cheng, Hsuan-Tien Lin, and Winston H. Hsu. 2012. Unsupervised semantic feature discovery for image object retrieval and tag refinement. *IEEE Transactions on Multimedia* 14, 4 (2012), 1079–1090.

Tian Lan and Greg Mori. 2013. A max-margin riffled independence model for image tag ranking. In *Proc. of CVPR*. 3103–3110.

Sihyoung Lee, Wesley De Neve, and Yong Man Ro. 2013. Visually weighted neighbor voting for image tag relevance learning. *Multimedia Tools and Applications* 72, 2 (2013), 1363–1386.

Mingling Li. 2007. Texture moment for content-based image retrieval. In *Proc. of ICME*. 508–511.

Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. 2011a. Text-based image retrieval using progressive multi-instance learning. In *Proc. of ICCV*. 2049–2055.

Xirong Li. 2016. Tag relevance fusion for social image retrieval. *Multimedia Systems*. In press (2016). DOI:http://dx.doi.org/10.1007/s00530-014-0430-9

Xirong Li, Efstratios Gavves, Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2011b. Personalizing automated image annotation using cross-entropy. In *Proc. of ACM MM*. 233–242.

Xirong Li and Cees G. M. Snoek. 2013. Classifying tag relevance with relevant positive and negative examples. In *Proc. of ACM MM*. 485–488.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009a. Annotating images by harnessing worldwide user-tagged photos. In *Proc. of ICASSP*. 3717–3720.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009b. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia* 11, 7 (2009), 1310–1322.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proc. of ACM CIVR*. 10–17.

Xirong Li, Cees G. M. Snoek, Marcel Worring, Dennis Koelma, and Arnold W. M. Smeulders. 2013. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia* 15, 4 (2013), 933–945.

Xirong Li, Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2012. Harvesting social images for bi-concept search. *IEEE Transactions on Multimedia* 14, 4 (2012), 1091–1104.

Zechao Li, Jing Liu, and Hanqing Lu. 2013. Nonlinear matrix factorization with unified embedding for social tag relevance learning. *Neurocomputing* 105 (2013), 38–44.

Zechao Li, Jing Liu, Xiaobin Zhu, Tinglin Liu, and Hanqing Lu. 2010. Image annotation using multi-correlation probabilistic matrix factorization. In *Proc. of ACM MM*. 1187–119.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. 2007. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning* 68, 3 (2007), 267–276.

Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Xiaojun Ye. 2013. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Proc. of CVPR*. 1618–1625.

Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. 2010. Image retagging. In *Proc. of ACM MM*. 491–500.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proc. of WWW*. 351–360.

Dong Liu, Xian-Sheng Hua, and Hong-Jiang Zhang. 2011. Content-based tag processing for internet social images. *Multimedia Tools and Applications* 51, 2 (2011), 723–738.

Dong Liu, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang. 2011b. Image retagging using collaborative tag propagation. *IEEE Transactions on Multimedia* 13, 4 (2011), 702–712.

Jing Liu, Zechao Li, Jinhui Tang, Yu Jiang, and Hanqing Lu. 2014. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Transactions on Multimedia* 16, 3 (2014), 588–600.

Jing Liu, Yifan Zhang, Zechao Li, and Hanqing Lu. 2013. Correlation consistency constrained probabilistic matrix factorization for social tag refinement. *Neurocomputing* 119, 7 (2013), 3–9.

Yang Liu, Fei Wu, Yin Zhang, Jian Shao, and Yueting Zhuang. 2011a. Tag clustering and refinement on semantic unity graph. In *Proc. of ICDM*. 417–426.

Hao Ma, Jianke Zhu, Michael Rung-Tsong Lyu, and Irwin King. 2010. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia* 12, 5 (2010), 462–473.

Subhransu Maji, Alexander C. Berg, and Jitendra Malik. 2008. Classification using intersection kernel support vector machines is efficient. In *Proc. of CVPR*. 1–8.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2010. Baselines for image annotation. *International Journal of Computer Vision* 90, 1 (2010), 88–105.

Julian McAuley and Jure Leskovec. 2012. Image labeling on a network: Using social-network metadata for image classification. In *Proc. of ECCV*. 828–841.

Philip McParlane, Stewart Whiting, and Joemon Jose. 2013b. Improving automatic image tagging using temporal tag co-occurrence. In *Proc. of MMM*. 251–262.

Philip J. McParlane, Yashar Moshfeghi, and Joemon M. Jose. 2013a. On contextual photo tag recommendation. In *Proc. of ACM SIGIR*. 965–968.

Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. 2014. Multimedia search reranking: A literature survey. *Computing Surveys* 46, 3 (2014), 38.

Ryszard S. Michalski. 1993. A theory and methodology of inductive learning. In *Readings in Knowledge Acquisition and Learning*. Morgan Kaufmann Publishers, 323–348.

Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *Proc. of ACM MM*. 59–68.

Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. 2014. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *Proc. of CVPR*. 4233–4240.

Oded Nov and Chen Ye. 2010. Why do people tag?: Motivations for photo tagging. *Communications of the ACM* 53, 7 (2010), 128–131.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 521–535.

Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji, and Thomas Huang. 2012. Exploring context and content links in social media: A latent space method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 5 (2012), 850–862.

Xueming Qian, Xian-Sheng Hua, Yuan Yan Tang, and Tao Mei. 2014. Social image tagging with diverse semantics. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2493–2508.

Zhiming Qian, Ping Zhong, and Runsheng Wang. 2015. Tag refinement for user-contributed images via graph learning and nonnegative tensor factorization. *IEEE Signal Processing Letters* 22, 9 (2015), 1302–1305.

Fabian Richter, Stefan Romberg, Eva Hörster, and Rainer Lienhart. 2012. Leveraging community metadata for multimodal image ranking. *Multimedia Tools and Applications* 56, 1 (2012), 35–62.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

Jitao Sang, Changsheng Xu, and Jing Liu. 2012a. User-aware image tag refinement via ternary semantic analysis. *IEEE Transactions on Multimedia* 14, 3 (2012), 883–895.

Jitao Sang, Changsheng Xu, and Dongyuan Lu. 2012b. Learn to personalized image search from the photo sharing websites. *IEEE Transactions on Multimedia* 14, 4 (2012), 963–974.

Neela Sawant, Ritendra Datta, Jia Li, and James Z. Wang. 2010. Quest for relevant tags using local interaction networks and visual content. In *Proc. of ACM MIR*. 231–240.

Neela Sawant, Jia Li, and James Z. Wang. 2011. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools and Applications* 51, 1 (2011), 213–246.

Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. 2006. Tagging, communities, vocabulary, evolution. In *Proc. of CSCW*. 181–190.

Börkur Sigurbjörnsson and Roelof Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proc. of WWW*. 327–336.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*.

Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380.

Nitish Srivastava and Ruslan R. Salakhutdinov. 2014. Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research* 15, 1 (2014), 2949–2980.

Aixin Sun, Sourav S. Bhowmick, Nam Nguyen, Khanh Tran, and Ge Bai. 2011. Tag-based social image retrieval: An empirical evaluation. *Journal of the American Society for Information Science and Technology* 62, 12 (2011), 2364–2381.

Jinhui Tang, Richang Hong, Shuicheng Yan, Tat-Seng Chua, Guo-Jun Qi, and Ramesh Jain. 2011. Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology* 2, 2 (2011), 14:1–14:15.

Jinhui Tang, Shuicheng Yan, Richang Hong, Guo-Jun Qi, and Tat-Seng Chua. 2009. Inferring semantic concepts from community-contributed images and noisy tags. In *Proc. of ACM MM*. 223–232.

Ba Quan Truong, Aixin Sun, and Sourav S. Bhowmick. 2012. Content is still king: The effect of neighbor voting schemes on tag relevance for social image retrieval. In *Proc. of ACM ICMR*. 9:1–9:8.

Ledyard R. Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 3 (1966), 279–311.

Tiberio Uricchio, Lamberto Ballan, Marco Bertini, and Alberto Del Bimbo. 2013. An evaluation of nearest-neighbor methods for tag refinement. In *Proc. of ICME*. 1–6.

Koen E. A. Van De Sande, Theo Gevers, and Cees G. M. Snoek. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2010), 1582–1596.

Jakob Verbeek, Matthieu Guillaumin, Thomas Mensink, and Cordelia Schmid. 2010. Image annotation with TagProp on the MIRFLICKR set. In *Proc. of ACM MIR*. 537–546.

Daan T. J. Vreeswijk, Cees G. M. Snoek, Koen E. A. van de Sande, and Arnold W. M. Smeulders. 2012. All vehicles are cars: Subclass preferences in container concepts. In *Proc. of ACM ICMR*. 8:1–8:7.

Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. 2006. Image annotation refinement using random walk with restarts. In *Proc. of ACM MM*. 647–650.

Gang Wang, Derek Hoiem, and David Forsyth. 2009. Building text features for object image classification. In *Proc. of CVPR*. 1367–1374.

Jingdong Wang, Jiazhen Zhou, Hao Xu, Tao Mei, Xian-Sheng Hua, and Shipeng Li. 2014. Image tag refinement by regularized latent Dirichlet allocation. *Computer Vision and Image Understanding* 124 (2014), 61–70.

Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *Computing Surveys* 44, 4 (2012), 25:1–25:24.

Meng Wang, Kuiyuan Yang, Xian-Sheng Hua, and Hong-Jiang Zhang. 2010. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia* 12, 8 (2010), 829–842.

Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. 2008. Flickr distance. In *Proc. of ACM MM*. 31–40.

Lei Wu, Rong Jin, and Anubhav K. Jain. 2013. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 3 (2013), 716–727.

Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. 2009. Learning to tag. In *Proc. of WWW*. 361–370.

Pengcheng Wu, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. 2011. Mining social images with distance metric learning for automated image tagging. In *Proc. of ACM WSDM*. 97–206.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proc. of ACL*. 133–138.

Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. 2009. Tag refinement by regularized LDA. In *Proc. of ACM MM*. 573–576.

Xing Xu, Akira Shimada, and Rin-ichiro Taniguchi. 2014. Tag completion with defective tag assignments via image-tag re-weighting. In *Proc. of ICME*. 1–6.

Kuiyuan Yang, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. 2011. Tag tagging: Towards more descriptive keywords of image content. *IEEE Transactions on Multimedia* 13, 4 (2011), 662–673.

Yang Yang, Yue Gao, Hanwang Zhang, Jie Shao, and Tat-Seng Chua. 2014. Image tagging with social assistance. In *Proc. of ACM ICMR*. 81–88.

Bolei Zhou, Vignesh Jagadeesh, and Robinson Piramuthu. 2015. ConceptLearner: Discovering visual concepts from weakly labeled image collections. In *Proc. of CVPR*.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2006. Learning with hypergraphs: Clustering, classification, and embedding. In *Proc. of NIPS*. 1601–1608.

Guangyu Zhu, Shuicheng Yan, and Yi Ma. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proc. of ACM MM*. 461–470.

Shiai Zhu, Chong-Wah Ngo, and Yu-Gang Jiang. 2012. Sampling and ontologically pooling web images for visual concept learning. *IEEE Transactions on Multimedia* 14, 4 (2012), 1068–1078.

Xiaofei Zhu, Wolfgang Nejdl, and Mihai Georgescu. 2014. An adaptive teleportation random walk model for learning social tag relevance. In *Proc. of ACM SIGIR*. 223–232.

Jinfeng Zhuang and Steven C. H. Hoi. 2011. A two-view learning approach for image tag ranking. In *Proc. of ACM WSDM*. 625–634.

Amel Znaidia, Hervé Le Borgne, and Céline Hudelot. 2013. Tag completion based on belief theory and neighbor voting. In *Proc. of ACM ICMR*. 49–56.

Received March 2015; revised December 2015; accepted March 2016