Multimedia Systems
DOI 10.1007/s00530-014-0430-9

SPECIAL ISSUE PAPER

Tag relevance fusion for social image retrieval

Xirong Li

Abstract Due to the subjective nature of social tagging, measuring the relevance of social tags with respect to the visual content is crucial for retrieving the increasing amounts of social-networked images. Witnessing the limit of a single measurement of tag relevance, we introduce in this paper tag relevance fusion as an extension to methods for tag relevance estimation. We present a systematic study, covering tag relevance fusion in early and late stages, and in supervised and unsupervised settings. Experiments on a large present-day benchmark set show that tag relevance fusion leads to better image retrieval. Moreover, unsupervised tag relevance fusion is found to be practically as effective as supervised tag relevance fusion, but without the need of *any* training efforts. This finding suggests the potential of tag relevance fusion for real-world deployment.

Keywords Social image retrieval · Tag relevance estimation · Tag relevance fusion

1 Introduction

Searching for the ever-growing amounts of varied and dynamically changing images on the social web is important for a number of applications. The applications include landmark visualization [14], visual query suggestion [48], training data acquisition [38], photo-based question answering [47], and photo-based advertisements [22], to name a few. As users often assign tags when posting their images on social media, one might expect tag-based retrieval to be a natural and good starting point for social image retrieval. Compared to content-based search [6], tag-based search potentially bypasses the semantic gap problem, and its scalability has been verified by decades of text retrieval research [2]. However, due to varied reasons, such as diversity in user knowledge, levels of expertise, and tagging intentions, social tagging is known to be ambiguous, subjective, and inaccurate [29]. Moreover, since individual tags are used only once per image in the social tagging paradigm, relevant and irrelevant tags for a specific image are not separable by tag statistics alone. Measuring social tag relevance with respect to the visual content they are describing is essential.

For tag relevance estimation, quite a few methods have been proposed. For example, Liu et al. [24] propose a non-parametric method to rank tags for a given image by kernel density estimation in a specific visual feature space. Chen et al. [4] train a Support Vector Machine classifier per tag. Given an image and its social tags, Zhu et al. [57] propose to measure the relevance of a specific tag in terms of its semantic similarity to the other tags. In our earlier work [19], a neighbor voting algorithm is introduced which exploits tagging redundancies among multiple users. Using learned tag relevance value as a new ranking criterion, better image search results are obtained, when compared to image search using original tags.

Positioned in a deluge of social data, however, tag relevance estimation is challenging. Visual concepts, for example, 'boat or 'garden', vary significantly in terms of their visual appearance and visual context. A single measurement of tag relevance as proposed in previous work is limited to tackle such large variations, resulting

X. Li (✉)
Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China
e-mail: xirong.li@gmail.com

X. Li
Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200443, China

Published online: 25 October 2014

in suboptimal image search. At the feature level, it is now evident that no single feature can represent the visual content completely [9, 26, 40, 49, 54]. Global features are suited for capturing the gist of scenes [31], while local features better depict properties of objects [33, 53]. As shown previously in content-based image search [41, 42], image annotation [10, 28], and video concept detection [43, 44], fusing multiple visual features is beneficial. So it is safe for us to envisage that tag relevance estimation will also benefit from the joint use of diverse features. The question is *what is the best strategy* to maximize such benefit?

Concerning fusion strategies, Snoek et al. [35] propose the taxonomy of early fusion and late fusion, which combine multiple sources of information at different stages. *Are early and late fusion schemes equally effective* for exploiting diverse features for measuring social tag relevance? Moreover, for both schemes, supervised learning techniques have been developed to optimize fusion weights, see for instance [25, 41]. In principle, the learned weights, obtained at the cost of learning from many manually labeled examples, should be better than uniform weights which simply treat individual features (in early fusion) and individual tag relevance estimators (in late fusion) equally. However, this "common sense" is not necessarily valid for social media, which is large scale, miscellaneous, and dynamically changes. Towards coping with the many tags and many images in social media, it is worthy to ask: *is supervised fusion a must?*

Towards answering the above questions, we make the following contributions:

1. We propose visual tag relevance fusion as an extension of tag relevance estimation for social image retrieval. Using the neighbor voting algorithm as a base tag relevance estimator [19], we present a systematic study on early and late tag relevance fusion. We extend the base estimator for both early and late fusion. Our previous work [20], which discusses late tag relevance fusion only, is a special case of this work.
2. Experiments on a large benchmark [5] show that tag relevance fusion leads to better image search. In particular, late fusion which combines both content-based [19, 24] and semantic-based [57] tag relevance estimators yields the best performance. Tag relevance fusion is also found to be helpful for acquiring better training examples from socially tagged data for visual concept learning.
3. This study offers a practical solution to exploit diverse visual features in estimating image tag relevance.

The problem we study lies at the crossroads of social tag relevance estimation and visual fusion. So next we present a short review of both areas.

2 Related work

2.1 Social tag relevance estimation

A number of methods have been proposed to attack the tag relevance estimation problem [4, 15, 19, 23, 24, 36, 45, 56, 57]. We structure them in terms of the main rationale they use, which is expressed in the following three forms, i.e., visual consistency [15, 19, 24, 36], semantic consistency [57], and visual–semantic consistency [23, 56]. Given two images labeled with the same tag, the visual consistency-based methods conjecture that if one image is visually closer to images labeled with the tag than the other image, then the former image is more relevant to the tag. Liu et al. [24] employ kernel density estimation in a visual feature space to find such visually close images, while Sun et al. [36] exploit visual consistency to quantify the representativeness of an image with respect to a given tag. We introduce a neighbor voting algorithm which infers the relevance of a tag with respect to an image by counting its visual neighbors labeled with that tag [19]. Lee et al. [15] first identify tags which are suited for describing the visual content by a dictionary lookup. Later, they apply the neighbor voting algorithm to the identified tags. To take into account negative examples of a tag which are ignored in the above works, Chen et al. [4] train SVM models for individual tags. Li and Snoek [18] take one step further by training SVM models with relevant positive and negative examples. Zhu et al. [57] investigate semantic consistency, measuring the relevance of a tag to an image in terms of its semantic similarity to the other tags assigned to the image, ignoring the visual content of the image itself. Sun et al. [37] propose to use the position information of the tags, and tags appearing top in the list are considered more relevant. To jointly exploit visual and semantic consistency, Liu et al. [23] perceive tag relevance estimation as a semi-supervised multi-label learning problem, while Zhu et al. [56] formulate the problem as decomposing an image tag co-occurrence matrix. Yang et al. [46] present a joint image tagging framework which simultaneously refines the noisy tags and learns image classifiers. Gao et al. [7, 8] propose to improve tag-based image search by visual-text joint hypergraph learning. Given initial image search results, the authors view the top ranked images as positive instances, and re-rank the search results by hypergraph label propagation. In all the above methods, only a single feature is considered. How to effectively exploit diverse features for tag relevance estimation remains open. It is also unclear whether fusing the individual and heterogeneous measurements of tag relevance is beneficial.

2.2 Visual fusion

Snoek et al. [35] classify methods for visual fusion into two groups: early fusion and late fusion. We follow their

taxonomy to organize our literature review on visual fusion. In early fusion, a straightforward method is to concatenate individual features to form a new single feature [35]. As feature dimensionality increases, the method suffers from the curse of dimensionality [32]. Another disadvantage of the method is the difficulty to combine features into a common representation [35]. Instead of feature concatenation, another method is to combine visual similarities of the individual features [10, 28, 43]. In these works, multiple visual (dis)similarities are linearly combined, with the combination weights optimized by distance metric learning techniques. In the context of video concept detection, Wang et al. [43] also choose linear fusion to combine similarity graphs defined by different features. In a recent work for fine-grained image categorization [50], an image is divided into multi-level hierarchical cells, and spatially adjacent cells are employed to describe the discriminative object components in a coarse-to-fine manner. Graphlets are introduced in [51, 55] to describe multiple aspects of an image including spatial relationships between pixels and their color/texture distribution. In late fusion, models are obtained separately on the individual features and their output is later combined [40, 44]. In the work by Wu et al. [44], base classifiers are trained using distinct features, and the output of the base classifiers forms a new feature vector for obtaining a final classifier. Wang et al. [40] combine the base classifiers in a boosting framework. To the best of our knowledge, visual fusion in the tag relevance estimation context has not been well explored in the literature.

3 Base tag relevance estimators

For a valid comparison between early and late fusion, we shall choose the same base tag relevance estimators for both fusion schemes. Thus, before delving into the discussion about tag relevance fusion and its solutions, we first make our choice of base estimators. For the ease of consistent description, we use x to denote an image, and w for a social tag. Let $g(x, w)$ be a base tag relevance function whose output is a confidence score of a tag being relevant to an image. Further, let $\mathcal{S}$ be a source set of social-tagged images, and $\mathcal{S}_w$ the set of images labeled with w, $\mathcal{S}_w \subset \mathcal{S}$.

A base estimator should be data driven and favorably exploit the large amount of social data. Moreover, it should be generic enough to adapt to both early and late fusion. In that regard, we choose the neighbor voting algorithm proposed in our previous work [19]. Despite its simplicity, recent studies [37, 39] report that this algorithm remains the state of the art for tag relevance estimation. To find visual neighbors from S for a given image x, we use $z(x)$ to represent a specific visual feature vector. We also have to specify a distance function for the given feature. The optimal distance varies in terms of tasks [52]. As the visual features used in this work, e.g., color correlogram and bag of visual words, are histogram based, we choose the l_1 distance. We use $\mathcal{S}_{x,z,k}$ to represent the k nearest visual neighbors of x, retrieved by the l_1 distance on z. The neighbor voting version of $g(x, w)$ is computed as

$$g(x, w) = \frac{|\mathcal{S}_{x,z,k} \cap \mathcal{S}_w|}{k} - \frac{|\mathcal{S}_w|}{|\mathcal{S}|}, \tag{1}$$

where $|\cdot|$ is the cardinality of a set. The term $|\mathcal{S}_{x,z,k} \cap \mathcal{S}_w|$ is the number of neighbor images labeled with w. Equation (1) shows that more neighbor images labeled with the tag induce larger tag relevance scores, and in the meantime, common tags with high frequency and thus less descriptive are suppressed by the second term.

In what follows, we develop early and late fusion variants of the neighbor voting algorithm, with a conceptual diagram illustrated in Fig. 1.

4 Tag relevance fusion

4.1 Problem formalization

From an information fusion perspective [3], diversity in base tag relevance estimators is important for effective fusion. We generate multiple tag relevance estimators by varying the visual feature z, the number of neighbors k, or both. For a given feature, as a larger set of visual neighbors always include a smaller set of visual neighbors, the parameter k has a relatively limited impact on the diversity. Hence, we fix k and diversify the base estimators using diverse visual features. Let $\mathcal{Z} = \{z_1, \ldots, z_m\}$ be a set of such features, and $g_i(x, w)$ be a base estimator specified by feature z_i, $i = 1, \ldots, m$. We adapt the notion of early and late fusion, defining

Early tag relevance fusion Fusion schemes that integrate individual features before estimating social tag relevance scores.

Late tag relevance fusion Fusion schemes that first use individual features to estimate social tag relevance scores separately, and then integrate the scores.

We use $G^e(x, w)$ to denote a fused tag relevance estimator obtained by early fusion, and $G^l(x, w)$ to denote a late fused estimator. The goal of tag relevance fusion is to construct a $G(x, w)$, let it be $G^e(x, w)$ in early fusion and $G^l(x, w)$ in late fusion, so that when $G(x, w)$ is used as an image-ranking criterion, better image retrieval is obtained compared to image retrieval using a single-feature estimator.

Since linear fusion is a well-accepted choice for visual fusion as discussed in Sect. 2.2, we follow this convention for tag relevance fusion. For early fusion, we

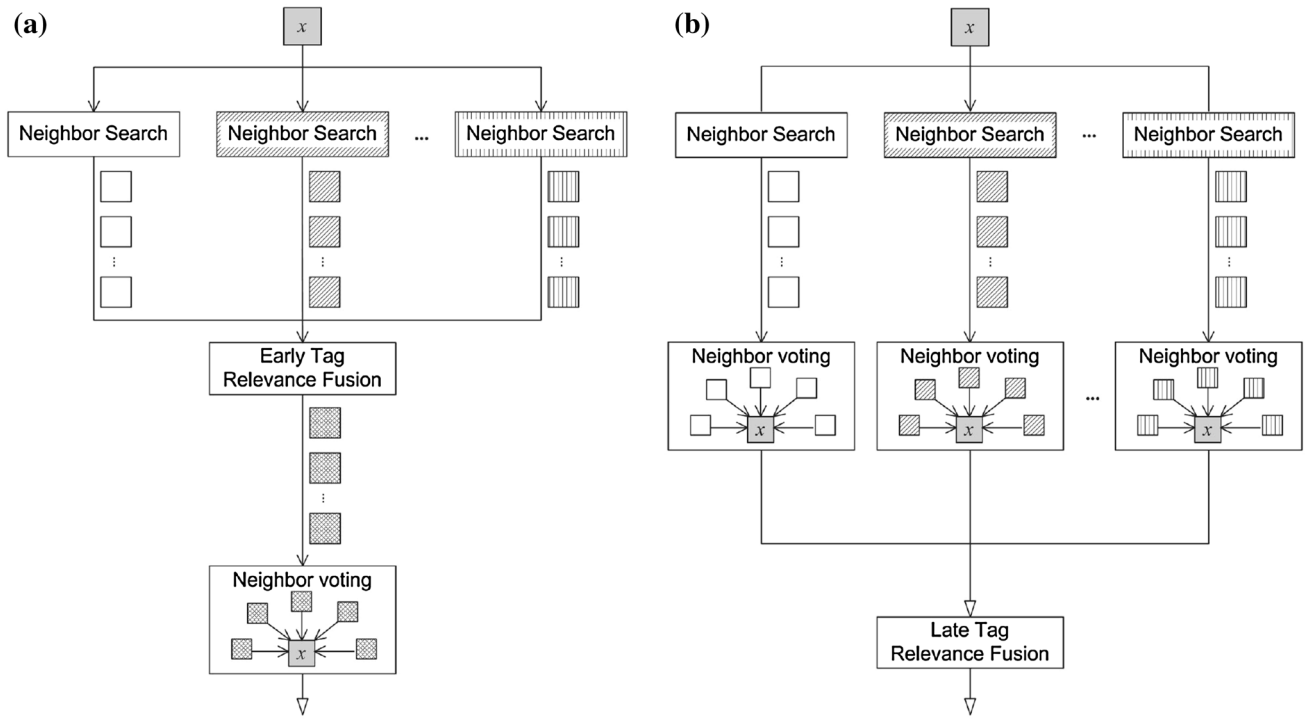


Fig. 1 Extending the neighbor voting algorithm to (**a**) early tag relevance fusion and (**b**) late tag relevance fusion. Given an image x, *different textured backgrounds* indicate its visual neighbors obtained by distinct visual features. In ealry tag relevance fusion, multiple visual neighbor sets are combined to obtain a better neighbor set for tag relevance estimation, while in late tag relevance fusion, we fuse multiple tag relevance estimates

aim for a better neighbor set by combining visual similarities defined by the m features. Concretely, given two images x and x', let $d_i(x, x')$ be their visual distance computed in terms of feature z_i. We define the combined distance as

$$d_{\Lambda}(x, x') = \sum_{i=1}^{m} \lambda_i \cdot d_i(x, x'), \tag{2}$$

where λ_i is a weight indicating the importance of z_i. The subscript Λ is to make the dependence of the fused distance on $\{\lambda_i\}$ explicit. We choose features which are intellectually devised, so we assume that they are better than random guess, meaning adding them is helpful for measuring the visual similarity. Hence, we constrain our solution with $\lambda_i \geq 0$. Since normalizing weights by dividing by their sum does not affect image ranking, any linear fusion with nonnegative weights can be transformed to a convex combination. So we enforce $\sum_{i=1}^{m} \lambda_i = 1$.

Let $\mathcal{S}_{x,\Lambda,k}$ be the k nearest neighbors retrieved by $d_{\Lambda}(x, x')$. Substituting it for $\mathcal{S}_{x,z,k}$ in (1) leads to the early fused tag relevance function:

$$G^{e}_{\Lambda}(x, w) = \frac{|\mathcal{S}_{x,\Lambda,k} \cap \mathcal{S}_w|}{k} - \frac{|\mathcal{S}_w|}{|\mathcal{S}|}. \tag{3}$$

In a similar fashion, we define the linear late fused tag relevance function:

$$G^{l}_{\Lambda}(x, w) = \sum_{i=1}^{m} \lambda_i \cdot g_i(x, w). \tag{4}$$

4.2 Solutions for tag relevance fusion

As distinct features are of varied dimensions and scales, the resultant visual distance scores (and tag relevance scores) often reside at varied scales. Score normalization is, thus, necessary before fusion.

4.2.1 Score normalization

We employ two popular strategies, i.e., MinMax and RankMax. Using a specific tag relevance estimator $g_i(x, w)$ as an example, its MinMax normalized version is defined as:

$$\tilde{g}_i(x, w) = \frac{g_i(x, w) - \min(g_i(x, w))}{\max(g_i(x, w)) - \min(g_i(x, w))}, \tag{5}$$

where the min (max) function returns the minimum (maximum) possible score. The RankMax normalized $g_i(x, w)$ is defined as:

$$\hat{g}_i(x,w) = 1 - \frac{rank(g_i(x,w))}{n_w}, \quad (6)$$

where $rank(g_i(x,w))$ returns the rank of image x when sorting images by $g_i(x,w)$ in descending order. Compared to MinMax, RankMax quantizes scores into discrete ranks, making it more robust to outliers.

Intuitively, for early (late) tag relevance fusion, better features (estimators) should have larger weights. Compared to the simplest solution that treats individual features and base estimators equally, it is not surprising that when we have access to many well-labeled examples, a better solution can be learned. However, for many tags, well-labeled examples are often of limited availability, making the study of unsupervised fusion necessary. Therefore, we study tag relevance fusion in both unsupervised and supervised settings.

4.2.2 Unsupervised tag relevance fusion

In an unsupervised setting, we have no prior knowledge of which feature or its resultant estimator is most appropriate for a given tag. According to the principle of maximum entropy [13], one shall make the least assumption about things we do not know. Hence, when no prior information concerning $\{\lambda_i\}$ is available, we shall use uniform weights. Following this thought, we consider fusion by averaging.

4.2.2.1 Unsupervised early fusion The fused distance $d_\Lambda(x,x')$ is the averaged value of $\{d_i(x,x')\}$, i.e.,

$$d_{\text{avg}}(x,x') = \frac{1}{m}\sum_{i=1}^{m} d_i(x,x'). \quad (7)$$

4.2.2.2 Unsupervised Late Fusion The corresponding $G^l_\Lambda(x,w)$ is simply the average of $\{g_i(x,w)\}$:

$$G^l_{\text{avg}}(x,w) = \frac{1}{m}\sum_{i=1}^{m} g_i(x,w). \quad (8)$$

Notice that fusing the RankMax normalized functions with the uniform weights is equal to Borda Count, a common algorithm for combining rankings generated by multiple sources of evidence [1].

4.2.3 Supervised tag relevance fusion

In an supervised setting, we aim to learn optimal fusion weights from many labeled examples. For early tag relevance fusion, this is to optimize the combined distance so that the percentage of relevant neighbors will increase, and consequently better tag relevance estimation is achieved. For late tag relevance fusion, this is to optimize the combined tag relevance estimator. In the following, we describe two learning algorithms for the two fusion schemes, respectively.

4.2.3.1 Supervised early fusion Optimizing fusion weights at the distance level is essentially distance metric learning. We opt to use the distance learning algorithm introduced by Wang et al. [41], for its effectiveness for multi-feature neighbor search. The basic idea is to find a combined distance to force images from the same class to be close, whilst images from different classes to be distant. This is achieved by solving the following objective function:

$$\underset{\Lambda}{\text{argmin}} \sum_{x,x'} \left(\exp\left(-\sum_{i=1}^{m} \lambda_i \cdot d_i(x,x') \right) - y(x,x') \right)^2, \quad (9)$$

where (x,x') is a pair of images randomly sampled from the training data, $y(x,x') = 1$ if the two images have labels in common, and $y(x,x') = 0$ otherwise.

4.2.3.2 Supervised late fusion Viewing the based estimators $\{g_i(x,w)\}$ as individual ranking criteria for image retrieval, we tackle supervised late tag relevance fusion as a learning-to-rank problem. Let $E_{metric}(G^l_\Lambda(x,w))$ be a performance metric function which measures the effectiveness of $G^l_\Lambda(x,w)$ on a training set. We seek Λ that maximizes E_{metric}:

$$\underset{\Lambda}{\text{argmax}} E_{metric}(G^l_\Lambda(x,w)). \quad (10)$$

Among many learning-to-rank algorithms, the coordinate ascent algorithm, developed by Metzler and Croft in the domain of document retrieval [30], can directly optimize (non-differentiable) rank-based performance metrics, e.g., Average Precision and NDCG. In the context of image auto-annotation [17], we observe that weights learned by coordinate ascent consistently outperform uniform weights for combining multiple meta classifiers. We, therefore, employ coordinate ascent for supervised late tag relevance fusion.

As a variant of hill climbing, coordinate ascent attempts to find Λ that maximizes E_{metric} in an iterative manner. In each iteration, a better solution is found by changing a single element of the solution, i.e., the weight corresponding to a specific base estimator. In particular, let λ_i be the parameter being optimized. We conduct a bi-direction line search with increasing steps to find the optimal value λ_i^*. If the search succeeds, i.e., λ_i^* yields a larger E_{metric}, we update λ_i with λ_i^*. Then, the next parameter λ_{i+1} is activated, and the same procedure applies. The optimization process continues until the objective function no longer increases.

The two fusion schemes, combined with specific normalization and weighting methods, result in the following 12 solutions:

1. Early-minmax-average: early fusion with MinMax normalization and uniform weights;
2. Early-rankmax-average: early fusion with RankMax normalization and uniform weights;
3. Early-minmax-learning: early fusion with MinMax normalization and fusion weights optimized by distance metric learning;
4. Early-rankmax-learning: early fusion with RankMax normalization and fusion weights optimized by distance metric learning;
5. Early-minmax-learning$^+$: early fusion with MinMax normalization and fusion weights optimized per concept by distance metric learning;
6. Early-rankmax-learning$^+$: early fusion with RankMax normalization and fusion weights optimized per concept by distance metric learning;
7. Late-minmax-average: late fusion with MinMax normalization and uniform weights;
8. Late-rankmax-average: late fusion with RankMax normalization and uniform weights;
9. Late-minmax-learning: late fusion with MinMax normalization and fusion weights optimized by coordinate ascent;
10. Late-rankmax-learning: late fusion with RankMax normalization and fusion weights optimized by coordinate ascent;
11. Late-minmax-learning$^+$: late fusion with MinMax normalization and fusion weights optimized per concept by coordinate ascent;
12. Late-rankmax-learning$^+$: late fusion with RankMax normalization and fusion weights optimized per concept by coordinate ascent.

4.3 Constructing base tag relevance estimators

As discussed in Sect. 4.1, the parameter k does not contribute significantly for diversifying the base estimators. We empirically fix k to be 500. Concerning the features $\{z_i\}$, we choose the following four visual features which describe image content in different aspects: COLOR, CSLBP, GIST, and DSIFT. COLOR is a 64-dimensional global feature [16], combining a 44-d color correlogram, a 14-d texture moments, and a 6-d RGB color moments. CSLBP is a 80-d center-symmetric local binary pattern histogram [11], capturing local texture distributions. GIST is a 960-d feature describing dominant spatial structures of a scene by a set of perceptual measures such as naturalness, openness, and roughness [31]. DSIFT is a 1,024-d bag of visual words depicting local information of the visual content, obtained by quantizing densely sampled SIFT descriptors using a precomputed codebook of size 1,024 [33]. We will refer to the four base estimators using the corresponding feature names.

Table 1 Datasets used in our experiments

	Source set	NUS-WIDE	
		Training	Test
No. images	815,320	155,545	103,688
No. users	177,871	40,202	32,415
No. tags	34,429	28,367	25,278
No. ground-truthed tags	N.A.	81	81

5 Experimental setup

5.1 Datasets

5.1.1 Source set for constructing base estimators

To instantiate $\mathcal{S}$, we use a public set of 3.5 million images[1] collected from Flickr in our previous work [19]. Since batch-tagged images tend to be visually redundant, we remove such images. Also, we remove images having no tags corresponding to WordNet. After this preprocessing step, we obtain a compact set of 815K images.

5.1.2 Benchmark data

We choose NUS-WIDE [5], a widely used benchmark set for social image retrieval. This set contains over 250K Flickr images,[2] with manually verified annotations for 81 tags which correspond to an array of objects, scenes, and event. As given in Table 1, the NUS-WIDE set consists of two predefined subsets, one training set with 155,545 images and one testing set of 103,688 images.

5.2 Experiments

5.2.1 Tag-based image retrieval

We evaluate the effectiveness of tag relevance fusion in the context of tag-based image retrieval, that is, for each of the 81 test tags, we sort images labeled with that tag in descending order by (fused) tag relevance scores.

[1] http://pan.baidu.com/s/1gdd3dBH.

[2] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm. As some images are no longer available on Flickr, the dataset used in this paper are a bit smaller than the original release.

Baselines As our goal is to study whether tag relevance fusion helps, the single-feature neighbor voting [19] is a natural baseline. For a more comprehensive comparison, we implement the following three present-day methods: tag position [37], tag ranking [24], and semantic field [57]. As tag ranking requires a specific visual feature for kernel density estimation in the feature space, we try tag ranking with each of the four features.

Evaluation criteria We use average precision (AP), which is in wide use for evaluating visual search engines. We also report normalized discounted cumulative gain (NDCG), commonly used to assess the top few ranked results of web search engines [12]. We compute NDCG for the top 100 ranked results. For overall comparisons, we average AP and NDCG scores over concepts, reporting mAP and mNDCG.

Test of statistical significance We conduct significance tests, with the null hypothesis that there is no difference in mAP (or mNDCG) of two image retrieval systems. In particular, we use the randomization test as recommended by Smucker et al. [34].

5.2.2 Visual concept learning with weak labeling

In this experiment, we apply tag relevance fusion to select better training examples for visual concept learning. The resultant concept classifiers will enable us to search images that are totally unlabeled. Concretely, for each test tag, we select its positive training examples from the NUS-WIDE training set, by sorting images in descending order by Late-minmax-average, and preserve the top 100 ranked images. We consider SemanticField and $\text{TagRel}_{\text{COLOR}}$ as two baselines, applying them separately to acquire another two sets of 100 positive training examples. As the focus is to compare which positive set is better, the same negative training data shall be used. We take a random subset of 1,000 images from the NUS-WIDE training set as the common negative set, albeit more advanced methods for negative sampling exist [21]. Fast intersection kernel SVMs [27] are trained with the DSIFT feature, and later applied to classify the NUS-WIDE test set.

6 Results

6.1 Tag-based image retrieval

6.1.1 Tag relevance fusion versus single tag relevance

As Table 2 shows, the best base estimator is $\text{TagRel}_{\text{DSIFT}}$, with mAP of 0.636 and mNDCG of 0.719. Except for Early-minmax-average, all the other fusion solutions are significantly better than $\text{TagRel}_{\text{DSIFT}}$, at the significance level of 0.01. For a better understanding of the results, we make a per-concept comparison, see Fig. 2. Compared to the best base estimator, tag relevance fusion improves AP scores for the majority of the concepts. This can be observed from Fig. 2 that the blue markers, representing early fusion, and the red markers, representing late fusion, are mostly on the right side. Further, for each concept, we check the best performer among the four base estimators. We find that for 21 concepts $\text{TagRel}_{\text{COLOR}}$ is the best, 2 concepts for $\text{TagRel}_{\text{CSLBP}}$, 25 concepts for $\text{TagRel}_{\text{GIST}}$, and 34 concepts for $\text{TagRel}_{\text{DSIFT}}$. Then, for every concept, we compare Early-rankmax-average and Late-minmax-average with the concept's best performer, which are concept dependent. For 30 concepts, Early-rankmax-average outperforms the best performers, while Late-minmax-average beats the best performers for 46 concepts. These results justify the effectiveness of visual fusion for improving tag relevance estimation.

Table 2 Performance of social image retrieval with and without tag relevance fusion

Method	mAP	mNDCG
Baselines		
TagPosition	0.560	0.605
SemanticField	0.577	0.607
$\text{TagRanking}_{\text{COLOR}}$	0.578	0.596
$\text{TagRanking}_{\text{CSLBP}}$	0.577	0.591
$\text{TagRanking}_{\text{GIST}}$	0.575	0.589
$\text{TagRanking}_{\text{DSIFT}}$	0.577	0.596
$\text{TagRel}_{\text{COLOR}}$	0.625	0.712
$\text{TagRel}_{\text{CSLBP}}$	0.588	0.657
$\text{TagRel}_{\text{GIST}}$	0.621	0.710
$\text{TagRel}_{\text{DSIFT}}$	0.636	0.719
Early tag relevance fusion		
Early-minmax-average	0.646	0.734
Early-rankmax-average	0.662*	0.756*
Early-minmax-learning	0.657*,#	0.749*,#
Early-rankmax-learning	0.664*	0.755*
Early-minmax-learning$^+$	0.658*,#	0.749*,#
Early-rankmax-learning$^+$	**0.665***	**0.756***
Late tag relevance fusion		
Late-minmax-average	0.660*	0.749*
Late-rankmax-average	0.652*	0.739
Late-minmax-learning	0.665*,#	0.753*
Late-rankmax-learning	0.659*,#	0.745*
Late-minmax-learning$^+$	**0.677***,#	0.773*,#
Late-rankmax-learning$^+$	0.673*,#	0.767*,#

At the significance level of 0.01, the symbol * indicates that a fused tag relevance is better than the best single-feature tag relevance ($\text{TagRel}_{\text{DSIFT}}$), while the symbol # indicates that a supervised fusion is better than its unsupervised counterpart

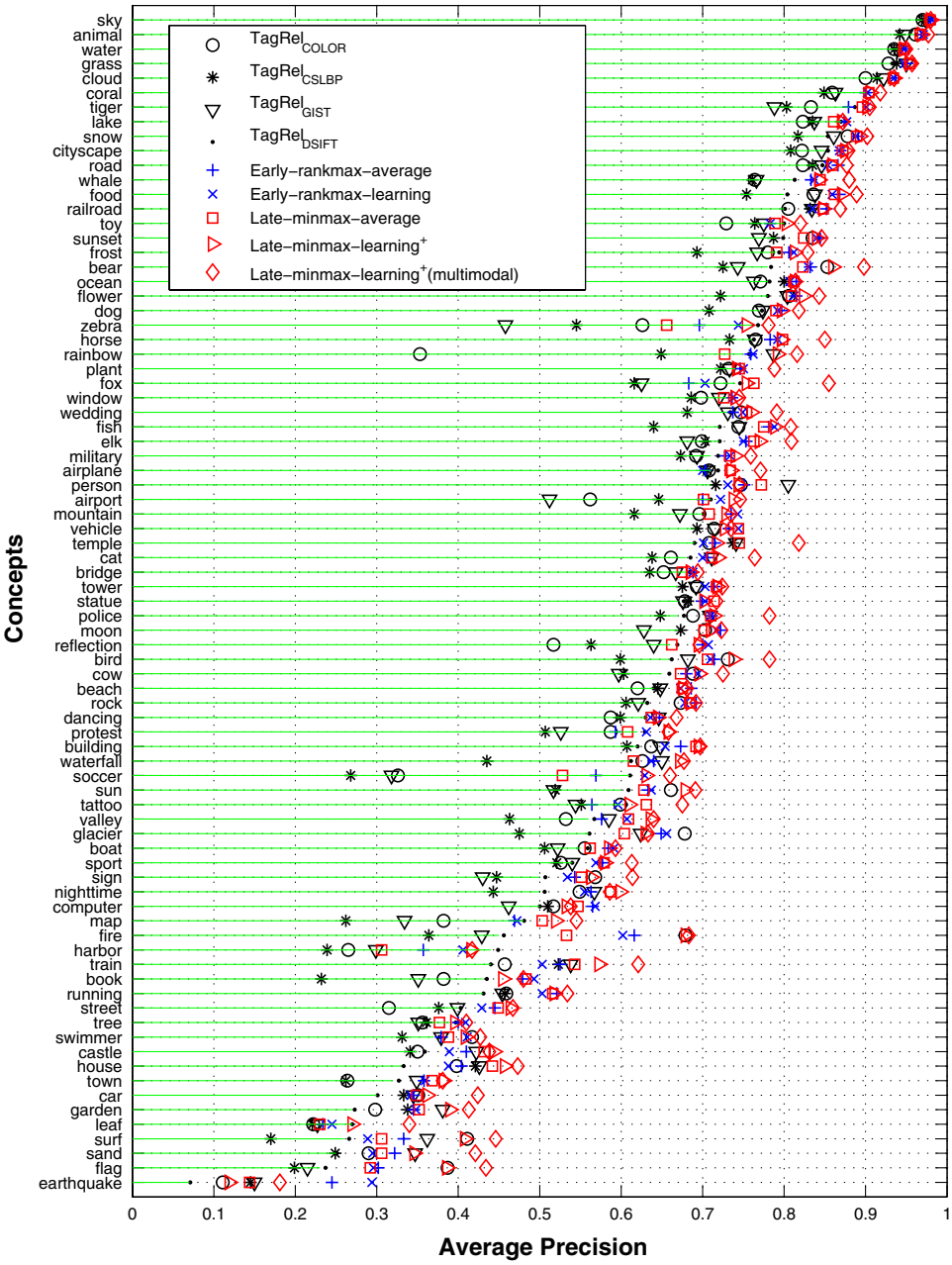


Fig. 2 Tag relevance fusion versus single tag relevance: a per-concept comparison. The concepts are sorted in descending order by $\text{TagRel}_{\text{DSIFT}}$. Best viewed *in color*

6.1.2 *Early tag relevance fusion versus late tag relevance fusion*

There is no significant difference between early and late fusion in unsupervised settings. Nevertheless, we observe the power of early fusion for addressing concepts that are rarely tagged. Consider 'earthquake' for instance. There are only 113 images labeled with the concept in $\mathcal{S}$. The rare occurrence makes the base estimators mostly produce zero score for the concept. Late fusion, with learning or not, does not add much in this case. In contrast, by directly manipulating the neighbor sets, Early-rankmax-learning yields the best result for 'earthquake'. Notice that early fusion needs to combine tens of thousands of visual neighbors, making it computationally more expensive than late fusion. Taking into account both effectiveness and efficiency, we recommend late fusion for tag relevance fusion.

For late fusion, Late-minmax-average, with mAP of 0.660 and mNDCG of 0.749, is slightly better than Late-rankmax-average, with mAP of 0.652 and 0.739. For 54 concepts, Late-minmax-average outperforms Late-rankmax-average. This result is mainly due to the fact that the base estimators already include an effect of smoothing by quantizing the visual neighborhood via neighbor voting. Extra quantization by RankMax makes tag relevance

Table 3 Performance of tag-based image retrieval by fusing heterogeneous tag relevance estimators, including the previous four base estimators, semantic field [57], and four variants of tag ranking [24]

Method	mAP	mNDCG
Late-minmax-average (multimodal)	0.673	0.759
Late-minmax-learning (multimodal)	0.679#	0.763
Late-minmax-learning$^+$ (multimodal)	0.700#	0.796#

At the significance level of 0.01, the symbol # indicates that a supervised fusion is better than its unsupervised counterpart

Table 4 Searching unlabeled images by visual concept classifiers learned from weakly labeled data

Positive example selection	mAP	mNDCG
SemanticField	0.119	0.271
$\text{TagRel}_{\text{COLOR}}$	0.119	0.298
Late-minmax-average	**0.127**	**0.339**

Classifierstrained on examples selected by Late-minmax-average-beats classifiers trained on examples selected by the two baselines

Bold values indicate the top performer

estimates less discriminative. Only when some base estimators yield large yet inaccurate values such as $\text{TagRel}_{\text{COLOR}}$ for 'rainbow', Late-rankmax-average is preferred.

6.1.3 Supervised fusion versus unsupervised fusion

The supervised methods achieve the best performance for both early and late fusion, see Table 2. Supervised methods work particularly well for those concepts where there is large variance in the performance of the base estimators. For early fusion, however, the difference between Early-rankmax-learning and Early-rankmax-average is not statistically significant. For late fusion, the difference in mNDCG of Late-minmax-learning and Late-minmax-average is not statistically significant. We also look into individual concepts. Although for 49 concepts Late-minmax-learning improves over Late-minmax-average, there are only eight concepts having a relative improvement of more than 5 %.

Learning weights per concept is beneficial. For 65 concepts, Late-minmax-learning$^+$ is better than Late-minmax-average, and the number of concepts that have more than 5 % relative improvement increases from 8 to 17. Nevertheless, because the weights are concept dependent, they are inapplicable to unseen concepts.

Overall, the performance of unsupervised fusion is close to supervised fusion. The result seems counter-intuitive as one would expect a larger improvement from supervised learning. We attribute this to the following two reasons. First, due to vagaries of social data, for a number of concepts, the models learned from the training data do not generalize well to unseen test data. Second, different from traditional learning-to-rank scenarios where features or rankers might be just better than random guess [25], the features employed in this study were intellectually designed and shown to be effective. As shown in Table 2, the base estimators already provide a strong starting point. Moreover, distinct features result in complementary neighbor sets for early fusion and complementary tag relevance estimates for late fusion. All this makes fusion with uniform weights a decent choice.

6.1.4 Fusing heterogeneous tag relevance estimators

To study the effect of fusing heterogeneous tag relevance estimators, we include semantic field and the four variants of tag ranking. Comparing Tables 2 and 3, we find that fusing the varied estimators is helpful. Again, Late-minmax-average is comparable to Late-minmax-learning in terms of NDCG. With mAP of 0.700 and mNDCG of 0.796, Late-minmax-learning$^+$ performs best. Note that the performance difference between Late-minmax-learning$^+$ and Late-minmax-average becomes larger. The result shows that concept-dependent weights are more needed for fusing tag relevance estimators driven by varied modalities.

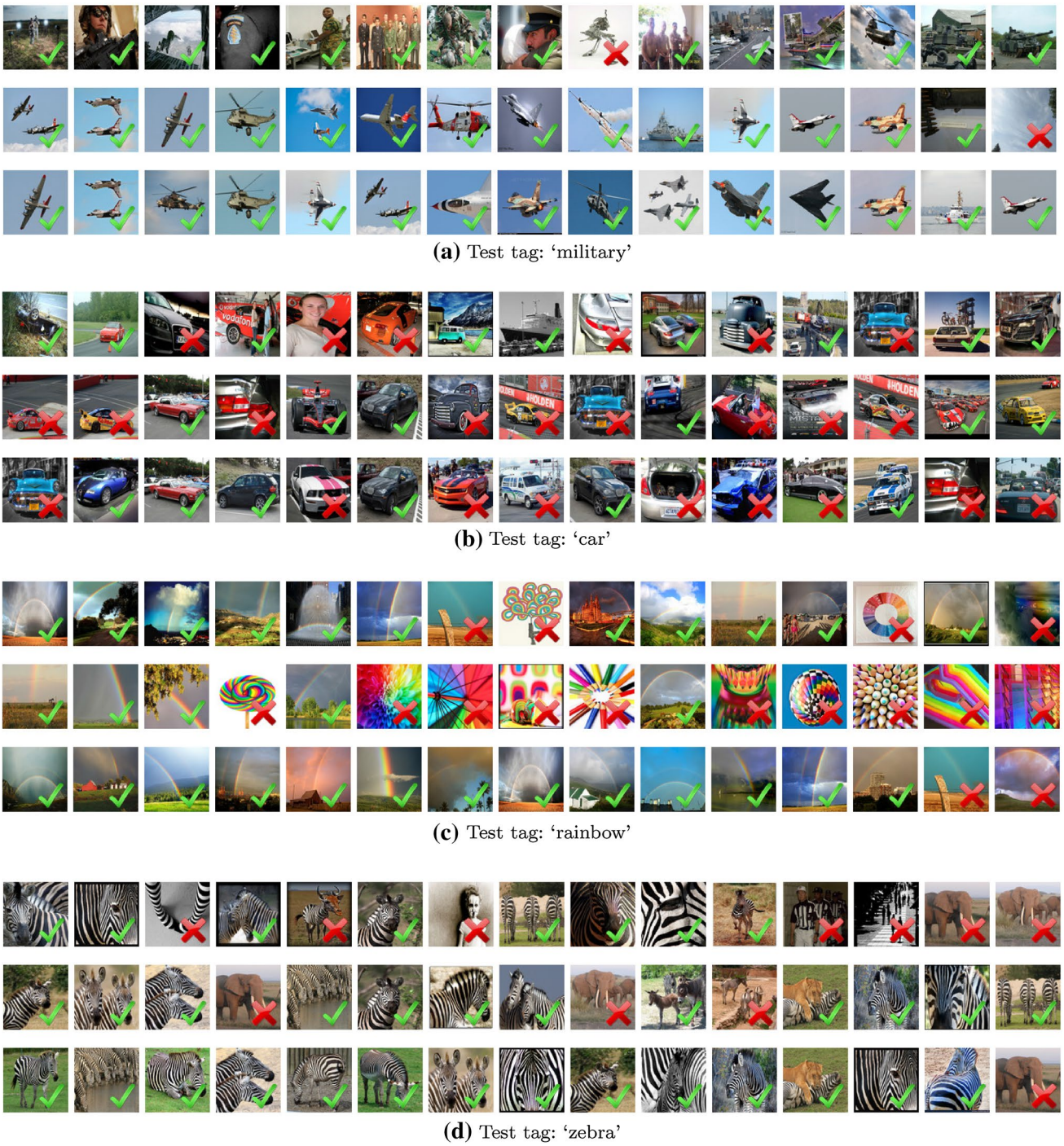
We present some image search results in Fig. 3. By exploiting diverse features, tag relevance fusion is helpful for concepts having larger inter-concept visual ambiguity such as rainbow versus colorful things like balloons. We observe from Fig. 3b that the annotation of NUS-WIDE is incomplete: a number of car images are not labeled as positive examples of 'car'. This is probably because the dataset developers used a kind of active learning strategy to ease the workload, without exhaustively labeling the dataset.

6.2 Visual concept learning with weak labeling

Table 4 shows the result of searching for the 81 test tags by the learned classifiers. Notice that because the test set is treated as totally unlabeled in this experiment, the scores are much lower than their counterparts in Table 2. We see from Table 4 that classifiers trained on positive examples selected by Late-minmax-average outperform classifiers trained on positive examples selected by the other methods. Hence, tag relevance fusion is also helpful for acquiring better training examples for visual concept learning.

7 Discussion and conclusions

Tag relevance estimation is important for social image retrieval. On recognizing the limitations of a single measurement of tag relevance, we promote in this paper tag



(a) Test tag: 'military'

(b) Test tag: 'car'

(c) Test tag: 'rainbow'

(d) Test tag: 'zebra'

Fig. 3 Image retrieval results for test tags (**a**) 'military', **b** 'car', **c** 'rainbow', and **d** 'zebra'. From the *top row* to the *bottom row*, each subfigure shows the top 15 results returned by SemanticField [57], $\text{TagRel}_{\text{COLOR}}$ [19], and the proposed Late-minmax-Learning^{+}, respectively. *Cross marks* indicate false positives according to the NUS-WIDE annotation.

relevance fusion as an extension to tag relevance estimation. We develop early and late fusion schemes for a neighbor voting based tag relevance estimator, and systematically study their characteristics and performance. Image retrieval experiments on a popular benchmark set of 250K images justify our findings as follows.

1. Tag relevance fusion improves tag relevance estimation. Comparing to the four base estimators whose mAP scores range from 0.588 to 0.636, fused tag relevance results in higher mAP ranging from 0.646 to 0.677. Adding extra heterogeneous estimators lifts mAP to 0.700.

2. The two fusion schemes each have their merit. By directly manipulating the visual neighbors, early tag relevance fusion is more effective for addressing concepts that are rarely tagged. Late fusion allows us to directly optimize image retrieval, and it is more flexible to handle varied tag relevance estimators.
3. Supervised fusion is meaningful only when one can afford per-concept optimization. Concept-independent weighting is marginally better than averaging the base estimators. For tag relevance fusion, we recommend the use of Late-minmax-average as a practical strategy.

Acknowledgments The author is grateful to Dr. Cees Snoek and Dr. Marcel Worring for their very useful comments on this work. The 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), and Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2014-002).

References

1. Aslam, J., Montague, M.: Models for metasearch. In: SIGIR (2001)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Boston (1999)
3. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Int. J. Inf. Fusion **6**(1), 5–20 (2005)
4. Chen, L., Xu, D., Tsang, I., Luo, J.: Tag-based image retrieval improved by augmented features and group-based refinement. IEEE Trans. Multimed. **14**(4), 1057–1067 (2012)
5. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: NUS-WIDE: a real-world web image database from National University of Singapore. In: CIVR (2009)
6. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 1–60 (2008)
7. Gao, Y., Wang, M., Luan, H., Shen, J., Yan, S., Tao, D.: Tag-based social image search with visual-text joint hypergraph learning. In: ACM multimedia (2011)
8. Gao, Y., Wang, M., Zha, Z.J., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. IEEE Trans. Image Process. **22**(1), 363–376 (2013)
9. Gehler, P., Nowozin, S.: Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In: CVPR (2009)
10. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
11. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogn. **42**, 425–436 (2009)
12. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**, 422–446 (2002)
13. Jaynes, E.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
14. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How Flickr helps us make sense of the world: context and content in community-contributed media collections. In: ACM multimedia (2007)
15. Lee, S., De Neve, W., Ro, Y.: Image tag refinement along the 'what' dimension using tag categorization and neighbor voting. In: ICME (2010)
16. Li, M.: Texture moment for content-based image retrieval. In: ICME (2007)
17. Li, X., Liao, S., Liu, B., Yang, G., Jin, Q., Xu, J., Du, X.: Renmin University of China at ImageCLEF 2013 scalable concept image annotation. In: CLEF working notes (2013)
18. Li, X., Snoek, C.: Classifying tag relevance with relevant positive and negative examples. In: ACM multimedia (2013)
19. Li, X., Snoek, C., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Trans. Multimed. **11**(7), 1310–1322 (2009)
20. Li, X., Snoek, C., Worring, M.: Unsupervised multi-feature tag relevance learning for social image retrieval. In: CIVR (2010)
21. Li, X., Snoek, C., Worring, M., Koelma, D., Smeulders, A.: Bootstrapping visual categorization with relevant negatives. IEEE Trans. Multimed. **15**(4), 933–945 (2013)
22. Li, Z., Zhang, L., Ma, W.Y.: Delivering online advertisements inside images. In: ACM Multimedia (2008)
23. Liu, D., Hua, X.S., Wang, M., Zhang, H.J.: Image retagging. In: ACM Multimedia (2010)
24. Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J.: Tag ranking. In: WWW (2009)
25. Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retr. **3**(3), 225–331 (2009)
26. Lu, Y., Zhang, L., Liu, J., Tian, Q.: Constructing concept lexica with small semantic gaps. IEEE Trans. Multimed. **12**(4), 288–299 (2010)
27. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR, pp. 1–8 (2008)
28. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. Int. J. Comput. Vis. **90**(1), 88–105 (2010)
29. Matusiak, K.: Towards user-centered indexing in digital image collections. OCLC Syst. Serv. **22**(4), 283–298 (2006)
30. Metzler, D., Croft, B.: Linear feature-based models for information retrieval. Inf. Retr. **10**(3), 257–274 (2007)
31. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
32. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. J. Mach. Learn. Res. **11**, 2487–2531 (2010)
33. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 1582–1596 (2010)
34. Smucker, M., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM (2007)
35. Snoek, C., Worring, M., Smeulders, A.: Early versus late fusion in semantic video analysis. In: ACM Multimedia (2005)
36. Sun, A., Bhowmick, S.: Quantifying tag representativeness of visual content of social images. In: ACM multimedia (2010)
37. Sun, A., Bhowmick, S., Nguyen, K., Bai, G.: Tag-based social image retrieval: an empirical evaluation. J. Am. Soc. Inf. Sci. Technol. **62**(12), 2364–2381 (2011)
38. Tang, J., Hong, R., Yan, S., Chua, T.S., Qi, G.J., Jain, R.: Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. ACM Trans. Intell. Syst. Technol. **2**, 14:1–14:15 (2011)
39. Uricchio, T., Ballan, L., Bertini, M., Del Bimbo, A.: An evaluation of nearest-neighbor methods for tag refinement. In: ICME (2013)
40. Wang, D., Liu, X., Luo, L., Li, J., Zhang, B.: Video diver: generic video indexing with diverse features. In: ACM MIR (2007)

41. Wang, G., Hoiem, D., Forsyth, D.: Building text features for object image classification. In: CVPR (2009)
42. Wang, J., Li, J., Wiederhold, G.: SIMPLIcity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 947–963 (2001)
43. Wang, M., Hua, X.S., Hong, R., Tang, J., Qi, G.J., Song, Y.: Unified video annotation via multigraph learning. IEEE Trans. Circuit Syst. Video Technol. **19**, 733–746 (2009)
44. Wu, Y., Chang, E., Chang, K., Smith, J.: Optimal multimodal fusion for multimedia data analysis. In: ACM multimedia (2004)
45. Xu, H., Wang, J., Hua, X.S., Li, S.: Tag refinement by regularized LDA. In: ACM multimedia (2009)
46. Yang, Y., Gao, Y., Zhang, H., Shao, J., Chua, T.S.: Image tagging with social assistance. In: ICMR (2014)
47. Yeh, T., Lee, J., Darrell, T.: Photo-based question answering. In: ACM multimedia (2008)
48. Zha, Z.J., Yang, L., Mei, T., Wang, M., Wang, Z., Chua, T.S., Hua, X.S.: Visual query suggestion: Towards capturing user intent in internet image search. ACM Trans. Multimed. Comput. Commun. Appl. **6**(3), 13:1–13:19 (2010)
49. Zhang, L., Gao, Y., Hong, C., Feng, Y., Zhu, J., Cai, D.: Feature correlation hypergraph: exploiting high-order potentials for multimodal recognition. IEEE Trans. Cybernet. **44**(8), 1408–1419 (2014)
50. Zhang, L., Gao, Y., Xia, Y., Dai, Q., Li, X.: A fine-grained image categorization system by cellet-encoded spatial pyramid modeling. IEEE Trans. Ind. Electron. (2014). doi:10.1109/TIE.2014.2327558
51. Zhang, L., Han, Y., Yang, Y., Song, M., Yan, S., Tian, Q.: Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans. Image Process.**22**(2), 5071–5084 (2013)
52. Zhang, L., Rui, Y.: Image search-from thousands to billions in 20 years. ACM Trans. Multimed. Comput. Commun. Appl. **9**(1), 36:1–36:20 (2013)
53. Zhang, L., Song, M., Liu, X., Bu, J., Chen, C.: Fast multi-view segment graph kernel for object classification. Signal Process. **93**(6), 1597–1607 (2013)
54. Zhang, L., Song, M., Liu, X., Sun, L., Chen, C., Bu, J.: Recognizing architecture styles by hierarchical sparse coding of blocklets. Inf. Sci. **254**, 141–154 (2014)
55. Zhang, L., Yang, Y., Gao, Y., Yu, Y., Wang, C., Li, X.: A probabilistic associative model for segmenting weakly supervised images. IEEE Trans. Image Process. **23**(9), 4150–4159 (2014)
56. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: ACM multimedia (2010)
57. Zhu, S., Jiang, Y.G., Ngo, C.W.: Sampling and ontologically pooling web images for visual concept learning. IEEE Trans. Multimed. **14**(4), 1068–1078 (2012)