

Tag-based Web Photo Retrieval Improved by Batch Mode Re-Tagging

Lin Chen Dong Xu Ivor W. Tsang
School of Computer Engineering
Nanyang Technological University
Singapore

{CHEN0631, DongXu, IvorTsang}@ntu.edu.sg

Jiebo Luo
Kodak Research Laboratories
Eastman Kodak Company
Rochester, USA

Jiebo.Luo@kodak.com

Abstract

Web photos in social media sharing websites such as Flickr are generally accompanied by rich but noisy textual descriptions (tags, captions, categories, etc.). In this paper, we proposed a tag-based photo retrieval framework to improve the retrieval performance for Flickr photos by employing a novel batch mode re-tagging method. The proposed batch mode re-tagging method can automatically refine noisy tags of a group of Flickr photos uploaded by the same user within a short period by leveraging millions of training web images and their associated rich textual descriptions. Specifically, for one group of Flickr photos, we construct a group-specific lexicon which contains only the tags of all photos within the group. For each query tag, we employ the inverted file method to automatically find loosely labeled training web images. We propose a SVM with Augmented Features, referred to as AFSVM, to learn adapted classifiers to refine the annotation tags of photos by leveraging the existing SVM classifiers of popular tags, which are associated with a large amount of positive training web images. Moreover, to further refine the annotation tags of photos in the same group, we additionally introduce an objective function that utilizes the visual similarities of photos within the group as well as the semantic proximities of their tags. Based on the refined tags, photos can be retrieved according to more reliable relevance scores. Extensive experiments demonstrate the effectiveness of our framework.

1. Introduction

With the rapid popularity of the Internet and digital cameras, a tremendous amount of richly labeled photos are being posted to photo sharing websites such as *Flickr.com* [19] and photo forums (e.g., *Photosig.com*) [20]. For example, *Flickr*, one of the most popular photo sharing websites, hosts more than two billion Flickr images. Everyday,

around three million photos are uploaded to *Flickr.com* and these photos are generally accompanied by rich descriptive keywords called tags. Such tags describe the contents of the photos and provide additional semantic information, which can be used to facilitate the retrieval of the shared Flickr photos.

However, the tags created by Flickr users are usually quite noisy. After comparing the performances of the classifiers trained based on Flickr images and their associated tags, Kennedy *et al.* revealed that only around 50% tags are relevant to the Flickr images [5]. Moreover, the tags may be also ambiguous, incomplete and overly personalized because the diverse Flickr users are with different knowledge and background [8]. Such inaccurate tags can significantly degrade the performance of tag-based retrieval of Flickr photos.

Recently, automatic tag ranking algorithms [8, 10] were proposed to rank the existing tags according to the relevance scores to the content of the given Flickr image. Li *et al.* [8] proposed to learn the tag relevance by counting the tag votes from visually similar photos. However, the performance of their work may significantly degrade when the visually similar images are not semantically relevant to the query image. In [10], Liu *et al.* adopted a kernel density estimation (KDE) algorithm to obtain the initial tag relevance estimation, and then employed a random walk-based method for tag refinement by exploiting the proximities between tags. However, both works [8, 10] did not utilize negative training samples, and they can not create new tags either.

We observe that Flickr users usually upload a group of photos captured within a short period for a memorable event, a particular activity, a trip, and so on, and the semantic concepts of photos in the same group are usually correlated (See Fig. 1). However, most Flickr users are reluctant to add tags to all photos in the same group. In this paper, the photos uploaded by the same user within a short period are considered to form a semantically related group. We propose a novel tag-based web photo retrieval framework to improve the retrieval performance by re-tagging a

group of Flickr photos (referred to as test Flickr photos). For each group of test Flickr photos, we respectively construct a group-specific lexicon which contains only the tags of all photos within the group. For each query tag (e.g., “dog”), we exploit the inverted file method [12] to automatically find the positive training web images that are related to the tag “dog” as well as the negative training web images that are irrelevant to the tag “dog”, as suggested in [12]. After that, one can train classifiers such as SVM based on these loosely labeled training web images.

We observe that some unpopular tags are only associated with a limited number of positive training web images, which may degrade the performance of SVM. Therefore, we propose to learn new SVM classifiers with Augmented Features (AFSVM), which can be adapted from the linear combination of a set of pre-learned classifiers of popular tags that are associated with a large number of positive training web images. Intuitively, AFSVM can utilize the training images associated with concepts like “river” and “lake” when training a classifier for the concept “water”. The solution of AFSVM is to re-train SVM classifiers again based on augmented features, which combine the original features and the decision values obtained from the pre-learned SVM classifiers of popular tags.

Inspired by [11], we also introduce a new objective function to further refine the annotation tags of test Flickr photos by additionally utilizing the similarities of test photos within the group as well as the (semantic) proximities of tags in the same group-specific lexicon. The objective function can be readily solved with a linear equation. Based on the refined annotation tags, we finally conduct tag-based photo retrieval in terms of more reliable relevance scores [10].

The main contributions of this paper include: (1) a new tag-based retrieval framework for loosely tagged photos by using a batch-mode re-tagging method, which effectively utilizes both the visual similarities of photos within a semantically related group and the semantic proximities of their tags, and (2) a new classification method of AFSVM, which performs better than SVM for this specific problem.

2. Related Work

Our work is related to Content Based Image Retrieval (CBIR). The CBIR systems (See the recent survey [2]) usually require users to provide images as queries to retrieve photos, *i.e.*, under the query by example framework. However, it is more natural for users to retrieve the desirable photos using textual queries (*i.e.*, tags). The most related work is the textual query based consumer photo retrieval system proposed in [12], which also employed loosely labeled web images to learn SVM classifiers for photo retrieval. In contrast to the approach in [12], our batch mode framework can effectively utilize the similarities of photos



Figure 1. Sample Flickr images in one group.

within the group and the semantic proximities of tags in the lexicon to improve the retrieval performance. Moreover, we propose a new classification method of AFSVM, which outperforms SVM.

Our work is also related to automatic image tagging (also known as image annotation) because Flickr photos are re-tagged before conducting tag-based photo retrieval. The image tagging methods can be roughly categorized into learning-based methods and web data-based methods [12]. In learning-based methods [1, 4, 13, 7], robust classifiers are learned from the training data, and then used to detect the presence of the concepts in any test data. However, the current learning-based methods can only tag at most hundreds of semantic concepts because the concept labels of the training samples need to be obtained through time-consuming and expensive human annotation. Web data-based methods used data-driven approach for image annotation. Torralba *et al.* [15] used k NN classifier for image tagging by leveraging 80 million tiny images which are loosely labeled with one noun from *WordNet*. Two indexing methods [16, 21] were subsequently developed to speed up the image search process by representing each image with less than a few hundred bits. Wang *et al.* [20] also employed millions of web images and their associated high quality descriptions (such as surrounding title and category) in photo forums (*e.g.*, *Photosig.com*) to tag images. An annotation refinement algorithm [17] and a distance metric learning method [18] were also proposed to further improve the image tagging performances.

Our re-tagging method belongs to the web data-based approach, thus it is inherently not limited to any predefined lexicon. When compared with the prior data-driven approaches [15, 17, 18, 20], our batch mode re-tagging method can simultaneously tag a group of photos captured by the same user within a short period. Moreover, most web data-based image tagging methods [15, 20] only output binary decisions (presence or absence). They do not provide a metric to rank the photos while our proposed framework

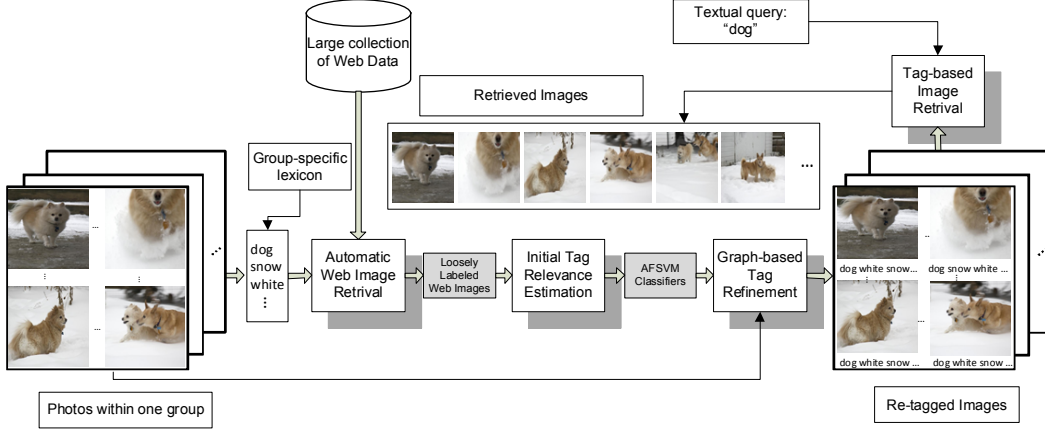


Figure 2. The framework of our tag-based photo retrieval framework.

does.

3. Our Framework

Our proposed framework is illustrated in Fig. 2. It consists of four modules: (1) automatic web image retrieval, (2) initial tag relevance estimation, (3) graph-based tag refinement, and (4) tag-based image retrieval.

3.1. Automatic Web Image Retrieval

In this module, for each group of Flickr photos, we first construct a group-specific lexicon which contains only the tags of all photos within the same group. Following [12], for each query tag (e.g., “dog”), we exploit the inverted file method to automatically find the positive training web images that are related to the tag “dog” as well as negative training web images which do not contain the tag “dog” in the surrounding texts. Considering the total number negative training web images (up to millions) is much larger than that of the positive training web images, we randomly sample a fixed number of negative web images and combine them with the positive web images to construct a smaller training set.

3.2. Initial Tag Relevance Estimation

Let us denote the tags from all the test Flickr images as $\mathcal{T} = \{t_1, \dots, t_H\}$, where H is the total number of tags. Based on the loosely labeled web images in the smaller training set, we can train a set of SVM classifiers $f_i(x)$'s ($i = 1, \dots, H$) for the tags t_i 's in \mathcal{T} , which can be directly employed to obtain the initial decision values for each test Flickr photo [12].

We observe that the unpopular tags are only associated with a limited number of positive training web images, which may significantly degrade the classification performances of SVM classifiers. We therefore propose SVM

with Augmented Features (AFSVM) to learn the adapted classifiers for all the tags in the set \mathcal{T} by leveraging the existing SVM classifiers of popular tags which are associated with a large amount of positive web images. As shown in our experiments (See Section 4.2), the average retrieval performances over all the tags in the set \mathcal{T} can be improved by using AFSVM classifiers, when compared with SVM classifiers.

Specifically, let us denote the set of popular tags as $\mathcal{T}_p = \{t_1^p, \dots, t_{K_p}^p\}$, where K_p is the total number of popular tags. For each tag $t \in \mathcal{T}$, we learn a one-versus-all classifier by using the pre-learned SVM classifiers $f_i^p(x)$'s of popular tags t_i^p 's as well as the corresponding loosely labeled training web data $(\mathbf{x}_i, y_i)_{i=1}^N$, where $y_i \in \{-1, 1\}$ is the label of sample \mathbf{x}_i and N is the total number of training samples. Motivated by [14], we assume the target classifier is in the following form:

$$\hat{f}(\mathbf{x}) = \mathbf{w}^\top \varphi(\mathbf{x}) + \beta^\top \phi(\mathbf{f}(\mathbf{x})) + b,$$

where $\mathbf{f}(\mathbf{x}) = [f_1^p(\mathbf{x}), \dots, f_{K_p}^p(\mathbf{x})]^\top$ is the vector of decision values of the pre-learned classifiers $f_i^p(\mathbf{x})$'s, and β is the weight vector, ϕ is the nonlinear feature mapping function for $\mathbf{f}(\mathbf{x})$, and $\mathbf{w}^\top \varphi(\mathbf{x}) + b$ is the decision function of standard SVM with φ being the nonlinear feature mapping function. The intuitive explanation for the above target classifier is based on the observation that some concepts are semantically correlated. For example, AFSVM can utilize the training images associated with concepts like “river” and “lake” when training a classifier for the concept “water”.

Like in SVM, we also minimize the structural risk functional. After incorporating the target classifier $\hat{f}(\mathbf{x})$, we arrive at the objective function:

$$\begin{aligned} \min_{\mathbf{w}, \beta, \xi_i, b} \quad & \frac{1}{2} (\|\mathbf{w}\|^2 + \|\beta\|^2) + C \sum_{i=1}^N \xi_i, \quad (1) \\ \text{s.t.} \quad & y_i \hat{f}(\mathbf{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, N. \end{aligned}$$

Note that, unlike the semi-parametric SVM in [14], here we also penalize the complexity of the weight vector β to control the complexity of the prelearned classifiers.

By introducing the Lagrange multipliers α_i 's for inequality constraints of (1), and the nonlinear feature mapping $\vartheta([\mathbf{x}_i^\top, \mathbf{f}(\mathbf{x}_i)^\top]^\top) = [\varphi(\mathbf{x}_i)^\top, \phi(\mathbf{f}(\mathbf{x}_i))^\top]^\top$, we can arrive at the following dual form:

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j k(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \end{aligned} \quad (2)$$

where $k(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \vartheta([\mathbf{x}_i^\top, \mathbf{f}(\mathbf{x}_i)^\top]^\top)^\top \vartheta([\mathbf{x}_j^\top, \mathbf{f}(\mathbf{x}_j)^\top]^\top)$ is the ϑ induced kernel function, and $\hat{\mathbf{x}}_i = [\mathbf{x}_i^\top, \mathbf{f}(\mathbf{x}_i)^\top]^\top$.

It is interesting to observe that the resultant optimization problem in (2) shares a similar form with the dual of SVM. The only difference is that in our method the training features are augmented as $[\mathbf{x}_i^\top, \mathbf{f}(\mathbf{x}_i)^\top]^\top$, which can be easily solved with SVM solvers such as LIBSVM¹. The implementation details are given in Procedure-1. For computational efficiency, the decision vector $\mathbf{f}(\mathbf{x}_i)$ of each training image \mathbf{x}_i is obtained by using linear SVM classifiers of popular tags (See Step-1 of Procedure-1) because the training and testing processes of linear SVM using LIBLINEAR [3] are much faster.

-
1. Train a linear SVM classifier for each popular tag $t_i^p \in \mathcal{T}_p$, and then apply all the K_p learned linear classifiers on each training sample \mathbf{x}_i of the tags in \mathcal{T} to obtain the decision value vector $\mathbf{f}(\mathbf{x}_i)$.
 2. For each tag $t \in \mathcal{T}$, train a new SVM classifier $\hat{f}(x)$ using RBF kernel with the default bandwidth parameter based on the augmented features of the corresponding loosely labeled web training images.
 3. Output H non-linear AFSVM classifiers $\hat{f}(x)$'s.
-

Procedure-1: The procedure of AFSVM

To compare the results from different classifiers for the subsequent operations, the decision values from AFSVM classifiers are converted into probabilities by using the sigmoid function:

$$\hat{g}(x) = \frac{1}{1 + \exp(a\hat{f}(x) + b)},$$

where the optimal a and b can be learnt by using the method in [9].

¹<http://www.csie.ntu.edu.tw/~cjlin/libsvm>

3.3. Graph-Based Tag Refinement

To further improve the tag refinement performance, we propose a graph-based tag refinement method by using the visual similarities of test photos within the same group as well as the semantic proximities of their tags.

Let us denote $Y \in \mathcal{R}^{M \times K}$ as the probability matrix obtained from AFSVM classifiers, and $F \in \mathcal{R}^{M \times K}$ as the matrix of the final prediction labels after tag refinement, where M is the number of test images in one group and K is the number of tags in the group-specific lexicon. Let us denote the i -th row and the i -th column of F as F_i and $F_{\cdot i}$, respectively. We also define a similarity matrix $W \in \mathcal{R}^{M \times M}$ to represent the similarities of photos in the group. The corresponding Laplacian matrix is denoted as $L = D - W$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_{j=1}^M W_{ij}$. We assume that if two photos in the same group are visually similar, their associated tags should be semantically related to each other. To this end, we can minimize $E_1(F) = \frac{1}{2} \sum_{i,j=1}^M W_{ij} \|F_i - F_j\|^2 = Tr(F^\top L F)$. We similarly define $W' \in \mathcal{R}^{K \times K}$ to characterize the proximities of tags, and its corresponding Laplacian matrix as G . To utilize the (semantic) proximities of the tags within the group specific lexicon, we similarly minimize $E_2(F) = \frac{1}{2} \sum_{i,j=1}^K W'_{ij} \|F_{\cdot i} - F_{\cdot j}\|^2 = Tr(F G F^\top)$. We propose the following objective function for tag refinement:

$$E(F) = \|F - Y\|^2 + \mu Tr(F^\top L F) + \lambda Tr(F G F^\top). \quad (3)$$

Setting the derivative of (3) with respect to F to zeros, we have:

$$\frac{\partial E(F)}{2\partial F} = (F - Y) + \mu L F + \lambda F G = 0.$$

Then we have:

$$A F + F B = C,$$

where $A = \mu L + I_{M \times M}$, $B = \lambda G$ and $C = Y$.

As shown in the prior work on multi-label learning [22], this is a well known Sylvester Equation in control theory. It can be rewritten as:

$$(I_{K \times K} \otimes A + B^\top \otimes I_{M \times M}) \cdot \text{vec}(F) = \text{vec}(C), \quad (4)$$

where \otimes is the Kronecker product defined as $U \otimes V = [U_{ij} V]$ for any two matrices U and V , and $\text{vec}(C)$ is the unfolded vector of matrix C . Then, (4) can be solved by linear equation, namely:

$$\text{vec}(F) = (I_{K \times K} \otimes A + B^\top \otimes I_{M \times M})^{-1} \cdot \text{vec}(C).$$

For the i -th test image, the associated tags can be ranked based on $F_{i \cdot}$. Based on the refined tags, tag-based web photo retrieval can be carried out using more reliable relevance scores (See Section 3.4).

3.4. Tag-based Flickr Photo Retrieval

Following [10], for each query tag q , we can conduct tag-based image retrieval to rank images based on their relevance scores in descending order. We do not use the decision values of the classifiers to retrieve images directly.

Instead for any test image x_i , we define the relevance score as:

$$r(x_i) = -\tau_i + 1/n_i, \quad (5)$$

where n_i is the total number of tags in the image x_i , and τ_i is rank position of the query tag q in the tag rank list of the image x_i which is decided according to the matrix F . If $\tau_i < \tau_j$, we have $r(x_i) > r(x_j)$. This indicates the image that contains the query tag q at the top positions in its tag rank list is assigned higher relevance scores. When the positions of the query tag q are the same for two images (*i.e.* $\tau_i = \tau_j$), the relevance score is decided by n_i and n_j , namely, the image that has fewer tags is preferred in this case.

4. Experiments

We compare our proposed framework with the methods proposed in [8, 10, 12]. We refer to the methods used in [8] and [10] as k NN and KDE.RW, because k NN classifier and Kernel Density Estimation (KDE) + Random Walk (RW) based methods are employed in [8] and [10], respectively. It is worth mentioning that the method proposed in [8] was also similarly used for web data-based image annotation in [20]. We also refer to the results after using the second and the third modules of our framework as AFSVM and AFSVM.Refine. In k NN, KDE.RW, the baseline SVM, AFSVM, AFSVM.Refine, we rank the tags for each test photo and then perform tag-based photo retrieval by using the relevance scores defined in Eq. (5). For comparison, we also report the retrieval performances of [12] (referred to as SVM.DV), for which we directly perform tag-based photo retrieval using SVM Decision Values. In this work, we do not compare our method with the existing learning based image annotation methods [1, 4, 13, 7] because these conventional methods currently can only tag at most hundreds of semantic concepts. In contrast, our image tagging method is intrinsically not limited by any predefined lexicon.

We consider two settings. In Setting A, for each image, we only consider its original tags created by Flickr users. This setting is used to evaluate the results of tag re-ranking. In Setting B, we assume that each image can be assigned with any of the tags in the group-specific lexicon, and new tags can therefore be created for a photo that initially has few or even no tags. This setting is used to evaluate the results of tag re-ranking and the creation of new tags. In both

settings, for any query tag q , only the photos that are associated with q are considered as test photos for performance evaluation.

4.1. Dataset and Experimental Setup

Our training dataset consists of about 1.3 million photos downloaded from the photo forum *Photosig.com* [12]. Most of the Photosig images are accompanied by rich surrounding textual descriptions including titles, categories and descriptions. After removing the high-frequency words (*e.g.*, “the”, “photo”, “picture”) that are not meaningful, our dictionary contains 21,377 words that almost cover all the daily-life concepts in a personal collection. Each image is associated with about five words on the average. While it is possible to use millions of images from *Flickr.com* as the training data, we choose *Photosig* dataset for training and use *Flickr* dataset for testing in order to avoid the overlap of training and test datasets. We note that the surrounding textual descriptions of *Photosig* are also less noisy than those of Flickr.

We have collected about 3500 test images from *Flickr.com* by using keywords to perform tag-based search. We choose 36 most popular tags² including scenes/landscapes, objects and colors. In total, we download 291 groups of images from 280 Flickr users. The images within each group were captured by the same user and within one day, and each group contains 12 images on the average. The tags of *Flickr* test images are manually annotated by two independent annotators who are not involved in the algorithmic design. Similar to [10], we first remove some over personalized, misspelling or meaningless tags that are not defined in WordNet before manual annotation, and convert the remaining words into their prototypes. After that, we have $H = 1033$ remaining tags in the test Flickr dataset. For each image, all the tags in the group-specific lexicon are shown to the annotators in order to compare different methods in both settings. The annotators can remove the inaccurate tags that are irrelevant to the images during the annotation process. Only the relevant tags retained by both annotators are finally considered as the ground-truth labels for performance evaluation.

Three types of global features (*i.e.*, Grid Color Moment (225 dim.), Wavelet Texture (128 dim.), and Edge Direction Histogram (73 dim.)) are used as the default features because of their efficiency and effectiveness. Each image is further represented as a single 426-D vector by concatenating the three types of global features. Using Principal Component Analysis (PCA), all the images in training and test datasets are finally projected into the 103-D space after

²These tags are *beach, bee, bird, blue, bridge, building, butterfly, candle, city, cloud, eye, firework, flower, garden, glass, green, home, island, lake, leaf, moon, mountain, peacock, pink, rain, red, river, rock, rose, sky, snow, sunset, tree, water, window, yellow.*

		<i>k</i> NN [8]	KDE_RW [10]	SVM_DV [12]	SVM	AFSVM	AFSVM_Refine
Setting A	all tags	52.5%	50.9%	49.2%	50.8%	54.2%	54.5%
	unpopular tags	58.0%	58.8%	54.4%	55.0%	59.1%	59.5%
Setting B	all tags	31.7%	36.0%	36.1%	38.0%	40.3%	41.9%
	unpopular tags	33.6%	39.5%	35.8%	37.3%	40.7%	42.4%

Table 1. MAPs of different algorithms in two settings. For each setting, we report MAPs over all the tags and the unpopular tags.

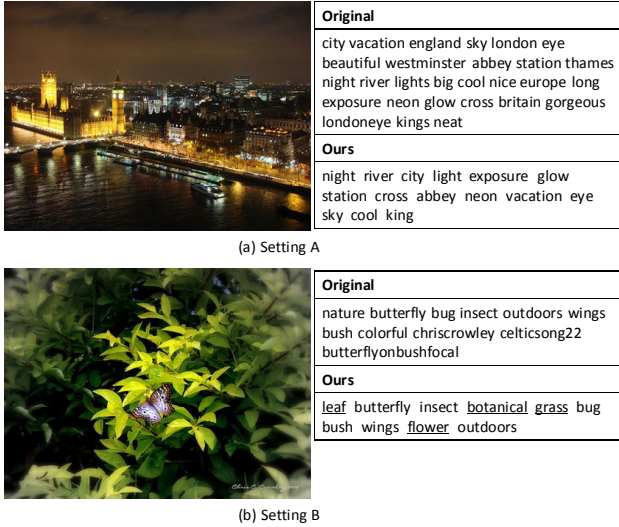


Figure 3. Re-tagging results by AFSVM_Refine. The tag rank lists of AFSVM_Refine are obtained according to the matrix F . In setting B, the newly created words are underlined.

dimension reduction. Please refer to [12] for more details.

To calculate the similarity matrix W of the test Flickr images within the same group (See Section 3.3), we adopt the Spatial Pyramid Matching (SPM) method [6]. As suggested in [6], we extract dense SIFT features from 16×16 pixel patches over a grid with spacing of 8 pixels, and we use 4 pyramid levels. We calculate the proximity matrix of tags W' using Google distance, namely, $W'(t_i, t_j) = \exp(-\psi_g(t_i, t_j))$, where $\psi_g(t_i, t_j)$ is the Google distance between two tags t_i and t_j (See [10] for more details).

4.2. Results

For performance evaluation, we use non-interpolated Average Precision (AP). Mean Average Precision (MAP) is the mean of APs over all the tags or the unpopular tags. Since the test dataset is unknown before training, we choose 400 popular tags based on the frequency of images in *Photosig* dataset to form the set \mathcal{T}_p . Among the 1033 tags in the *Flickr* dataset (*i.e.*, the set \mathcal{T}), there are 310 popular tags and 723 unpopular tags. In SVM related methods, we train one-versus-all classifiers by using the default setting in LIBSVM. We empirically fix $\mu = 3$ and $\lambda = 0.5$ in the proposed AFSVM_Refine.

We test the retrieval performances in two settings (*i.e.*,

Setting A and Setting B) using all the 1033 tags and the 723 unpopular tags as queries. The results are shown in Table 1. We have the following observations: 1) SVM is better than SVM_DV, which demonstrates the effectiveness of using the relevance scores defined in Eq. (5) for tag-based web photo retrieval. 2) AFSVM outperforms both SVM and SVM_DV, especially for the cases that the unpopular tags are used as queries. This demonstrates the effectiveness of using the prelearned classifiers of popular tags in AFSVM to improve the photo retrieval performance. 3) AFSVM_Refine achieves the best results, thanks to the additional utilization of the graph-based tag refinement method. We also observe that the MAP improvements of AFSVM_Refine over other existing algorithms [8, 10, 12] for the unpopular tags in Setting A are less pronounced. We believe it is reasonable given only a limited number of training samples in this case. 4) All the algorithms achieve worse results in Setting B, because the total number of test photos in Setting B is generally larger than that in Setting A.

Typical re-tagging results in the two settings are illustrated in Fig. 3. In our tag-based photo retrieval framework, the photos are more likely to be retrieved by using their associated top-ranked tags as queries (See Eq. (5)). Therefore, the top-ranked tags are more important. For the example in Setting A, our method can place the most relevant words such as “night”, “river”, “city”, and “light” near the top of the tag rank list. For the example in Setting B, three new tags “leaf”, “botanical” and “grass” are correctly added to the top of the tag rank list. Fig. 4 shows the top-10 retrieved images in Setting A by using query tag “peacock” and Fig. 5 shows the top-10 retrieved images in Setting B by using the query tag “dog”. Clearly, the results of AFSVM_Refine are much better than those of *k*NN, KDE_RW and SVM_DV.

5. Conclusions

We have proposed a novel tag-based photo retrieval framework by re-tagging a group of semantically related Flickr photos. In our framework, we first construct a group-specific lexicon consisting of only the tags of all the photos within the group. For any query tag, we obtain loosely labeled positive and negative training web images by using inverted file based method. Based on these loosely labeled training web images, we train SVMs with Augmented Fea-

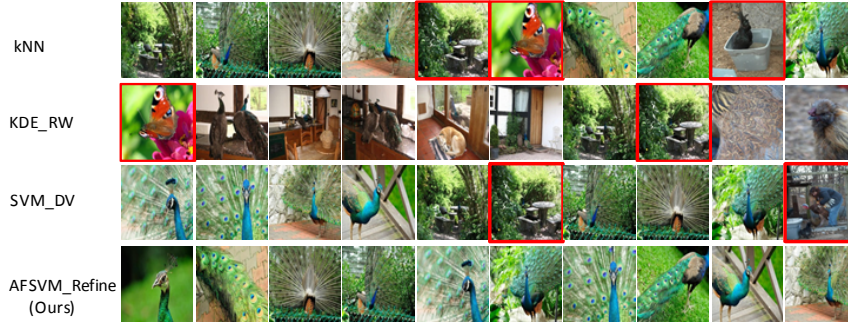


Figure 4. Top-10 retrieval results for query “peacock” in Setting A. Incorrect results are highlighted by red boxes.

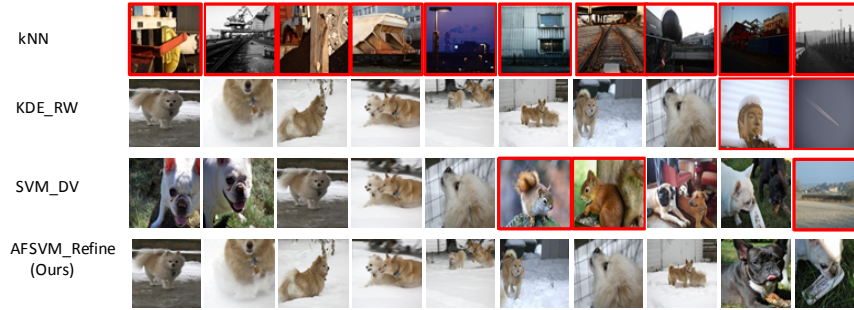


Figure 5. Top-10 retrieval results for query “dog” in Setting B. Incorrect results are highlighted by red boxes.

tures (AFSVM) classifiers for all the tags in the test dataset by leveraging the prelearned SVM classifiers of popular tags. Next, we use a graph-based method to further refine the annotation tags. Finally, we conduct tag-based photo retrieval by using the relevance scores suggested in [10]. Extensive experiments demonstrate the effectiveness of our framework.

Acknowledgements This work is supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF2008IDM-IDM004-018.

References

- [1] S.-F. Chang et al. Large-scale multimodal semantic concept detection for consumer video. In *MIR*, 2007.
- [2] R. Datta et al. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.
- [3] R.-E. Fan et al. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 2008.
- [4] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *T-PAMI*, 30(8):1371–1384, 2008.
- [5] L. S. Kennedy, S. F. Chang, and I. V. Kozintsev. To search or to label? predicting the performance of search-based automatic image classifiers. In *ACM MM Workshop on MIR*, 2006.
- [6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [7] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *T-PAMI*, 30(6):985–1002, 2008.
- [8] X. Li, C. G. M. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *ACM Conference on MIR*, 2008.
- [9] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt’s probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.
- [10] D. Liu et al. Tag ranking. In *WWW*, 2009.
- [11] L. Cao et al. Annotating Photo Collections by Label Propagation According to Multiple Similarity Cues. In *ACM MM*, 2008.
- [12] Y. Liu, D. Xu, I. W. Tsang and J. Luo. Textual Query of Consumer Photos Facilitated by Large-scale Web Data. In *T-PAMI*, 2010.
- [13] J. V. M. Guillaumin, T. Mensink and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.
- [14] A. J. Smola, T. T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines. In *NIPS*, 1999.
- [15] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *T-PAMI*, 30(11):1958–1970, 2008.
- [16] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *CVPR*, 2008.
- [17] C. Wang et al. Content-based image annotation refinement. In *CVPR*, 2007.
- [18] C. Wang, L. Zhang, and H. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR*, 2008.
- [19] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In *ICCV*, 2009.
- [20] X. Wang et al. Annotating images by mining image search results. *T-PAMI*, 30(11):1919–1932, 2008.
- [21] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.
- [22] Z.-J. Zha et al. Graph-based semi-supervised learning with multi-label. In *ICME*, 2008.