Textual Query of Personal Photos Facilitated by Large-scale Web Data

Yiming Liu[1], Dong Xu[1], Ivor W. Tsang[1], Jiebo Luo[2]

[1]School of Computer Engineering, Nanyang Technological University, Singapore

[2]Intelligent Systems Research Center, Kodak Research Laboratories, Eastman Kodak Company, USA

Abstract

The rapid popularization of digital cameras and mobile phone cameras has lead to an explosive growth of personal photo collections by consumers. In this paper, we present a real-time textual query based personal photo retrieval system by leveraging millions of web images and their associated rich textual descriptions (captions, categories, etc.). After a user provides a textual query (*e.g.*, "water"), our system exploits the inverted file to automatically find the positive web images that are related to the textual query "water" as well as the negative web images that are irrelevant to the textual query. Based on these automatically retrieved relevant and irrelevant web images, we employ three simple but effective classification methods, k Nearest Neighbor (kNN), decision stumps and linear SVM, to rank personal photos. To further improve the photo retrieval performance, we propose two relevance feedback methods via cross-domain learning, which effectively utilize both the web images and personal images. In particular, our proposed cross-domain learning methods can learn robust classifiers with only a very limited amount of labeled personal photos from the user by leveraging the pre-learned linear SVM classifiers in real time. We further propose an incremental cross-domain learning method in order to significantly accelerate the relevance feedback process on large consumer photo databases. Extensive experiments on two consumer photo datasets demonstrate the effectiveness and efficiency of our system, which is also inherently not limited by any predefined lexicon.

Index Terms

Textual Query Based Consumer Photo Retrieval, Large-Scale Web Data, Cross-Domain Learning

I. INTRODUCTION

With the rapid popularization of digital cameras and mobile phone cameras, retrieving images from enormous collections of personal photos has become an important research topic and a practical problem at the same time. In the recent decades, many Content Based Image Retrieval (CBIR) systems [30], [33], [34], [47] have been proposed. These systems usually require users to provide example images as queries in order to retrieve personal photos, i.e., under the query by example framework. However, the paramount challenge in CBIR is the so-called semantic gap between the low-level visual features and the high-level semantic concepts. To bridge the semantic gap, relevance feedback methods were proposed to learn the user's intentions.

For consumer applications, it is more natural for the user to retrieve the desirable personal photos using textual queries. To this end, image annotation is commonly used to classify images

with respect to a set of high-level semantic concepts. This can be used as an intermediate stage for textual query based image retrieval because the semantic concepts are analogous to the textual terms that describe document contents. In general, image annotation methods can be classified into two categories, learning-based methods and web data-based methods [22]. Learning-based methods build robust classifiers based on a fixed corpus of labeled training data, and then use the learned classifiers to detect the presence of the predefined concepts in the test data. On the other hand, as an emerging paradigm, web data-based methods leverage millions of web images and the associated rich textual descriptions for image annotation.

Recently, Chang *et al.* presented the first systematic work for consumer video annotation. Their system can automatically detect 25 predefined semantic concepts, including occasions, scenes, objects, activities and sounds [6]. Observing that the personal photos are usually organized into collections by time, location and events, Cao *et al.* [3] proposed a label propagation method to propagate the concept labels from part of personal images to the other photos in the same album. In [22], Jia *et al.* proposed a web-based annotation method to obtain the conceptual labels for image clusters only, followed by a graph-based semi-supervised learning method to propagate the conceptual labels to the whole photo album. However, to obtain the initial annotations, the users are required to describe each photo album using textual terms, which are then submitted to an online image server (such as *Flickr.com*) to search for thousands of images related to the keywords. Therefore, the annotation performance of this method depends heavily on the textual terms provided by the users and the search quality of the web image server.

In this work, we propose a real-time textual query based retrieval system, which directly retrieves the desirable personal photos without undergoing any intermediate image annotation process. Our work is motivated by the advances in *Web 2.0* and the recent advances of web data-based image annotation techniques [22], [25], [35], [36], [38], [39], [41], [42]. Everyday, rich and massive social media data (texts, images, audios, videos, etc.) are posted to the web. Web images are generally accompanied by rich contextual information, such as tags, categories, titles, and comments. In particular, we have downloaded about 1.3 million images and the corresponding *high quality* surrounding textual descriptions (titles, categories, descriptions, etc.) from photo forum *Photosig.com*[1]. Note that in contrast to *Flickr.com*, the quality of the images

[1]http://www.photosig.com/

from this source can be considered higher and visually more characteristic of semantics of the corresponding textual descriptions. After the user provides a textual query (*e.g.*, "water"), our system exploits the inverted file to automatically retrieve the positive web images, which have the textual query "water" in the surrounding descriptions, as well as the negative web images, whose surrounding descriptions do not contain the query "water" and its descendants (such as "meltwater", "freshwater", etc.) according to *WordNet* [15]. The inverted file method has been successfully used in information retrieval to efficiently find all text documents where a given word occurs [44]. Based on these automatically retrieved positive and negative web images, we employ classifiers, including k Nearest Neighbor (kNN), decision stump ensemble, and linear SVM, to rank the photos in the personal collections. Observing that the total number of negative web images is much larger than the total number of positive web images, we randomly sample a fixed number of negative samples and combine these samples with the positive samples for training decision stump ensemble and SVM classifiers. Similar as in [33], the whole procedure is repeated multiple times by using different randomly sampled negative web images and the average output from multiple rounds is finally used for robust consumer photo retrieval.

To improve the retrieval performance in CBIR, relevance feedback has been frequently used to help acquire the search intention from the user. However, most users would prefer to label only a few images in a limited feedback, which frequently degrades the performance of the typical relevance feedback algorithms [17], [47]. A brute-force solution is to use a large number of web images and a limited amount of feedback images for relevance feedback. However, the classifiers trained from both the web images and labeled consumer images may perform poorly because the feature distributions from these two domains can be drastically different. To address this problem, we further propose two cross-domain learning methods to learn robust classifiers (referred to as target classifiers) using only a limited number of labeled feedback images by leveraging the pre-learned classifier (referred to as auxiliary classifier). Cross-domain methods have been used in real applications, such as sentiment classification, text categorization, and video concept detection [2], [11], [12], [13], [23], [46]. However, these methods are either variants of SVM or in tandem with non-linear SVM or other kernel methods, making it inefficient for large-scale applications. In addition, the recent cross-domain learning works on image annotation [12], [13], [23], [46] only cope with the cross-domain cases on news videos captured from different years or different channels. In contrast, this work tackles a more challenging cross-domain case from

the web image domain to the consumer photo domain.

Specifically, we first proposed a simple cross-domain learning method by directly combining the auxiliary classifier and SVM learned in the target domain. Then, we propose Cross-Domain Regularized Regression (CDRR) by introducing a new regularization term into regularized regression. This regularization term enforces a constraint such that the target classifier produces similar decision values as the auxiliary classifier on the unlabeled consumer photos. Our experiments demonstrate that the two cross-domain learning methods can significantly improve the photo retrieval performance. To significantly accelerate the relevance feedback process on large consumer photo databases, we further propose an incremental cross-domain learning method, referred to as Incremental CDRR, by incrementally updating the corresponding data matrices.

It is worth noting that the techniques used in *Google* image search cannot be directly used for textual query based consumer photo retrieval. *Google* image search[2] can only retrieve web images which are identifiable by rich semantic textual descriptions (such as filename, surrounding texts, and URL). However, raw consumer photos from digital cameras do not contain such semantic textual descriptions. In essence, we exploit a large-scale collection of web images and their rich surrounding textual descriptions as the training data to help retrieve the new input data in the form of raw, unlabeled consumer photos.

The main contributions of this paper include:

- We introduce a new framework for textual query based consumer photo retrieval by leveraging millions of web images and their associated rich textual descriptions. This framework is also inherently not limited by any predefined lexicon.
- Our proposed cross-domain learning approaches further improve the photo retrieve performance by using the pre-learned classifier (auxiliary classifier) from a large number of loosely labeled web images, and a small number of precisely labeled consumer photos from relevance feedback. To the best of our knowledge, this is the first time that the cross-domain learning methods are used for relevance feedback. Our cross-domain learning methods also outperform two conventional manifold ranking and SVM based relevance feedback methods[17], [47].
- Our proposed Incremental CDRR is a novel incremental cross-domain learning method,

[2]Fergus et al. proposed to use parts-based model to improve *Google* image search results in [16].

which is suitable for relevance feedback in large-scale consumer photo retrieval applications.

- Our system achieves real-time response thanks to the combined efficiency of decision stump ensemble classifier and linear SVM classifier, Incremental CDRR, and a number of speed-up techniques, including the utilization of the inverted file method to efficiently search relevant and irrelevant web images, PCA to reduce feature dimension, and computation on multiple threads.

A preliminary version of this work appeared in [27]. In this paper, we additionally use linear SVMs for initial photo retrieval and propose Incremental CDRR to achieve real-time retrieval performance on large photo datasets. This paper also provides additional experiments on the large *NUS-WIDE* dataset [8]. Moreover, we also systematically investigate the efficiency and effectiveness of linear SVM classifier and decision stump ensemble classifier for initial photo retrieval, as well as compare the retrieval performances of early fusion and late fusion schemes for fusing three types of global features (*i.e.*, Grid Color Moment, Edge Direction Histogram and Wavelet Texture).

The remainder of this paper is organized as follows. Sections II and III provide brief reviews of two related areas, content based image retrieval and image annotation. The proposed textual query based consumer photo retrieval system will be introduced in Section IV. Extensive experimental results will be presented in Section V, followed by concluding remarks in the final section.

II. Related Work in Content based Image Retrieval (CBIR)

Over the past two decades, a large number of CBIR systems have been developed to retrieve images from image databases in the hope for returns semantically relevant to the user's query image. Interested readers can refer to two comprehensive surveys in [32], [10] for more details. However, in consumer applications, it is more convenient and natural for a user to supply a textual query when performing image retrieval.

It is well-known that the major problem in CBIR is the semantic gap between the low-level features (color, texture, shape, etc.) and the high-level semantic concepts. Relevance feedback has proven to be an effective technique to improve the retrieval performance of CBIR systems. The early relevance feedback methods directly adjusted the weights of various features to adapt to the user's intention [30]. In [48], Zhou and Huang proposed Biased Discriminant Analysis (BDA) to select a small set of discriminant features from a large feature pool for relevance

feedback. Support Vector Machines (SVM) based relevance feedback techniques [33], [34], [47] were also proposed. The above methods have demonstrated promising performance for image retrieval, when a sufficient number of labeled images are marked by the users. However, users typically mark a very limited number of feedback images during the relevance feedback process, and this practical issue can significantly degrade the retrieval performance of these techniques [30], [33], [34], [47], [48]. Semi-supervised learning [19], [21] and active learning [21], [34] have also been proposed to improve the performance of image retrieval. He [19] used the information from relevance feedback to construct a local geometrical graph to learn a subspace for image retrieval. Hoi *et al.* [21] applied active learning strategy to improve the retrieval performance of Laplacian SVM. However, these methods usually require manifold assumption of unlabeled images, which may not hold with unconstrained consumer photos.

In this paper, we propose a real-time, textual query based retrieval system to directly retrieve the desired photos from personal image collections by leveraging millions of web images together with their accompanying textual descriptions. We further propose two efficient cross-domain relevance feedback methods to learn robust classifiers by effectively utilizing the rich but perhaps loosely annotated web images as well as the limited feedback images marked by the user. In addition, we also propose Incremental CDRR (ICDRR), an incremental cross-domain learning method, to significantly accelerate the relevance feedback process on large consumer photo dataset.

III. Related Work in Image Annotation

Image annotation is an important task and closely related to image retrieval. The methods can be classified into two categories, learning-based methods and web data-based methods [22]. In learning-based methods [3], [6], [24], robust classifiers (also called models or concept detectors) are first learned based on a large corpus of labeled training data, and then used to detect the presence of the concepts in any test data. However, the current learning-based methods can only annotate at most hundreds of semantic concepts [29], because the concept labels of the training samples need to be obtained through time consuming and expensive human annotation.

Recently, web data-based methods were developed and these methods can be used to annotate general images. Torralba *et al.* [35] collected about 80 million tiny images (color images with the size of 32 by 32 pixels), each of which is labeled with one noun from *WordNet*. They

demonstrated that with sufficient samples, a simple kNN classifier can achieve reasonable performance for several tasks such as image annotation, scene recognition, and person detection and localization. Subsequently, Torralba *et al.* [36] and Weiss *et al.* [43] also developed two indexing methods to speed up the image search process by representing each image with less than a few hundred bits. Zhang and his colleagues have also proposed a series of works [25], [38], [39], [41], [42] to utilize images and the associated high quality descriptions (such as surrounding title and category) in photo forums (*e.g.*, *Photosig.com* and *Photo.net*) to annotate general images. For a given query image, their system first searches for similar images among those downloaded images from the photo forums, and then "borrows" representative and common descriptions (concepts) from the surrounding descriptions of these similar images as the annotation for the query image. The initial system [41] requires the user to provide at least one accurate keyword to speed up the search efficiency. Subsequently, an approximate yet efficient indexing technique was proposed, such that the user no longer needs to provide keywords [25]. An annotation refinement algorithm [38] and a distance metric learning method [39] were also proposed to further improve the image annotation.

It is possible to perform textual query based image retrieval by using image annotation as an intermediate stage. Since the image annotation process needs to be performed before textual query based consumer photo retrieval, the user needs to perform image annotation again to assign these new textual terms to all the personal images, when the new text queries provided by the user are out of the current set of vocabularies. In addition, these image annotation methods do not provide a metric to rank the images.

IV. Textual Query Based Consumer Photo Retrieval

In this Section, we will present our proposed framework on how to utilize a large collection of web images to assist image retrieval using textual query for consumer photos from personal collections. It is noteworthy that myriads of web images are readily available on the *Internet*. These web images are usually associated with rich textual descriptions (referred to as surrounding texts hereon) related to the semantics of the web images. These surrounding texts can be used to extract high-level semantic labels for the web images without any cost of labor-intensive annotation efforts. In this framework, we propose to apply such valuable Internet assets to facilitate textual query based image retrieval. Recall that the consumer photos (from personal

collections) are usually organized in folders without any indexing structure to facilitate textual queries. To automatically retrieve consumer photos using textual queries, we choose to leverage millions of web images and their surrounding texts as the bridge between the domains of the web images and the consumer photos.

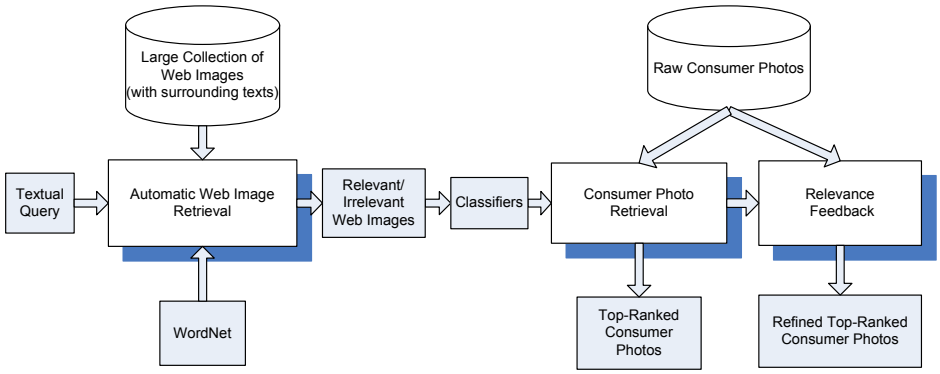

Fig. 1. Textual Query Based Consumer Photo Retrieval System.

A. Proposed Framework

The architecture of our proposed framework is depicted in Figure 1. It consists of several machine learning modules. The first module of this framework is automatic web image retrieval, which first interprets the semantic concept of textual queries by a user. Based on the semantic concept and *WordNet*, the sets of relevant and irrelevant web images are retrieved from the web image database using the inverted file method [44]. The second module then uses these relevant and irrelevant web images as a labeled training set to train classifiers (such as kNN, decision stumps, SVM, and boosting). These classifiers are then used to retrieve potentially relevant consumer photos from personal collections. To further improve the retrieval performance, relevance feedback and cross-domain learning techniques are employed in the last module to refine the image retrieval results.

B. Automatic Web Image Retrieval

In this framework, we first collect a large set of web images with surrounding texts related to a set of almost all the daily-life semantic concepts $\mathcal{C}_w$ from *Photosig.com*. Stop-word removal is also used to remove from $\mathcal{C}_w$ the high-frequency words that are not meaningful. Then, we

assume such a large-scale web image database contains sufficient images to cover almost all the daily-life semantic concepts in a personal collection. Then, we construct the inverted file, which has an entry for each word q in $\mathcal{C}_w$, followed by a list of all the images that contain the word q in the surrounding texts.

For any textual query q, we can efficiently retrieve all web images whose surrounding texts contain the word q by using the pre-constructed inverted file. These web images can be deemed as relevant images. For irrelevant web images, we use *WordNet* [15], [35], which models semantic relationships for commonly-used words, to define the set $\mathcal{C}_s$ as the descendant texts of q. Figure 2 shows the subtree representing the two-level descendants of the keyword "water" in *WordNet*. Based on this subtree, one can retrieve all irrelevant web images that do not contain any word in $\mathcal{C}_s$ in the surrounding texts. Thereafter, we can denote these automatically annotated (relevant and irrelevant) web images as $D^w = (\mathbf{x}_i^w, y_i^w)|_{i=1}^{n_w}$, where $\mathbf{x}_i^w$ is the ith web image and $y_i^w \in \{\pm 1\}$ is the label of $\mathbf{x}_i^w$.

C. Consumer Photo Retrieval

As discussed in Section IV-B, with the surrounding texts, we can automatically obtain annotated web images D^w based on the textual query. These annotated web images can be used as the training set for building classifiers. Any classifiers (such as SVM or Boosting) can be used in our framework. However, considering that the size of the web images in D^w can be up to millions, direct training of complex classifiers (e.g., nonlinear SVM and Boosting) may not be feasible for real-time consumer photo retrieval. We therefore choose three simple but effective classifiers, namely k Nearest Neighbor classifier, decision stump ensemble classifier, and linear SVM classifier. Note that boosting using decision stumps has shown the state-of-the-art performance in face detection [37], in which the training of boosting classifier is performed in an offline way. Boosting is not suitable for our real-time online photo retrieval application because of its high computational cost.

1) k Nearest Neighbors: For the given relevant web images in D^w (*i.e.*, web images with $y_i^w = 1$), the simplest method to retrieve the target consumer photos is to compute the average distance between each consumer photo and its k nearest neighbors (kNN) from the relevant web images (says, $k = 300$). Then, we rank all consumer photos with respect to the average distances to their k nearest neighbors.

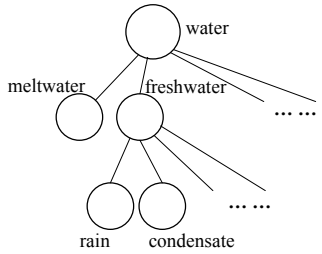

Fig. 2. The subtree representing the two-level descendants of "water" in *WordNet*.

2) Asymmetric Bagging with Decision Stumps: Note that the kNN approach cannot account for the irrelevant photos for consumer photo retrieval. To improve the retrieval performance, we also use the relevant and irrelevant web images in D^w to train a decision stump ensemble classifier. In particular, the size of the irrelevant images (up to millions) can be much larger than that of the relevant images, so the class distribution in D^w can be extremely unbalanced. To avoid such a highly skewed distribution in the annotated web images, following the method proposed in [33], we randomly sample a fixed number of irrelevant web images as the negative samples, and combine with the relevant web images as the positive samples to construct a smaller training set.

After sampling, a decision stump $f_d(\mathbf{x}) = h(s_d(x_d - \theta_d))$ is learned by finding the sign $s_d \in \{\pm 1\}$ and the threshold $\theta_d \in \Re$ of the dth feature x_d of the input $\mathbf{x}$ such that the threshold θ_d separates both classes with a minimum training error ϵ_d on the smaller training set. For discrete output, $h(x)$ is the sign function, that is, $h(x) = 1$ if $x > 0$; and $h(x) = -1$, otherwise. For continuous output, $h(x)$ can be defined as the symmetric sigmoid activation function, i.e., $h(x) = \frac{1-\exp(-x)}{1+\exp(-x)}$. We observe that it is difficult to rank the consumer photos by using discrete output because the responses of many consumer photos are the same in this case. In this work, we therefore use the continuous output of $h(x)$. The threshold θ_d can be determined by sorting all samples according to the feature x_d, and scanning the sorted feature values. In this way, the decision stump can be found efficiently. Next, the weighted ensembles of these decision stumps are computed for prediction, i.e.,

$$f^s(\mathbf{x}) = \sum_d \gamma_d h(s_d(x_d - \theta_d)), \tag{1}$$

where the weight γ_d for each stump is set to $0.5 - \epsilon_d$ and ϵ_d is the training error rate of the dth decision stump classifier. Note that γ_d is further normalized such that $\sum_d \gamma_d = 1$.

To remove the possible side effect of random sampling of the irrelevant images, the whole procedure is repeated n_s times by using different randomly sampled irrelevant web images. Finally, the average output is used for robust consumer photo retrieval. This sampling strategy is also known as Asymmetric Bagging[3] [33].

After asymmetric bagging with decision stumps, there are $n_s n_d$ decision stumps. We remove the 20% decision stumps with the largest training error rates. This removal process generally preserves the most discriminant decision stumps, and at the same time accelerates the initial photo retrieval process.

3) Asymmetric Bagging with Linear SVM: While decision stump ensemble classifier can effectively exploit both relevant and irrelevant web photos in D^w, it is inefficient to use this classifier on a large consumer photo dataset because all the decision stumps need to be applied on every test photo in the testing stage. Suppose we train $n_s n_d$ decision stump classifiers, where n_d is the feature dimension and n_s is the random sampling times for generating the negative samples in asymmetric bagging. Then, for each test image, all the decision stumps need to be applied in the test stage, which means the floating value comparison and the calculation of exponential function in symmetric sigmoid function will be performed for $0.8 n_s n_d$ times even after removal of 20% decision stumps with the largest training error rates. Moreover, one decision stump classifier only account for one single dimension of the whole feature space. Thus, each individual classifier may be still too weak.

To facilitate large scale consumer photo retrieval, we propose to use linear SVM classifier based on loosely labeled web images. Considering that the total number of irrelevant web images is much larger than that of relevant web images, we also construct a smaller training set by combining the positive web images and randomly sampled negative web images. As suggested in [20], feature vectors are normalized into unit hyper-spheres in the kernel space[4]. Assume that $f_{SVM}(\mathbf{x}) = \mathbf{w}_s' \mathbf{x} + b_s$ is the decision classifier, we then train the linear SVM classifier by

[3]In [33], the base classifier used in asymmetric bagging is non-linear SVM.

[4]For linear SVM, normalization in kernel space is equivalent to normalization in input space.

minimizing the following objective functional:

$$\frac{1}{2}\|\mathbf{w}_s\|^2 + C_{SVM}\sum_i \xi_i$$

$$s.t. \quad y_i^w(\mathbf{w}_s'\mathbf{x}_i^w + b_s) \geq 1 - \xi_i, \tag{2}$$

where ξ_i is the slack variable and C_{SVM} is the tradeoff parameter.

We also repeat the whole procedure for n_s times by using different random samples of irrelevant web images. Finally, the average output is used for robust consumer photo retrieval:

$$f^s(\mathbf{x}) = \sum_s \gamma_s g(\mathbf{w}_s'\mathbf{x} + b_s) \tag{3}$$

where $\gamma_s = 0.5 - \epsilon_s$, ϵ_s is the training error of the s-th linear SVM classifier, and $g(x)$ is the sigmoid activation function. Again, γ_s is normalized such that $\sum_s \gamma_s = 1$.

4) Decision Stumps vs. Linear SVM: With the same n_s, in general, it takes more time to train a linear SVM classifier than a decision stump ensemble classifier. However, the prediction of asymmetric bagging with linear SVM is much faster. For each test data, there are only n_s times of the calculation of exponential function in (3). Moreover, in the experiments, we observe that linear SVM usually achieves comparable or even better retrieval performances, possibly because it simultaneously considers multiple feature dimensions. Therefore, we generally prefer linear SVM for large-scale consumer photo retrieval.

D. Relevance Feedback via Cross-Domain Learning

With Relevance Feedback (RF), we can obtain a limited number of relevant and irrelevant consumer photos from the user to further refine the image retrieval results. However, the feature distributions of photos from different domains (web images and consumer photos) may differ considerably and thus have very different statistical properties (in terms of mean, intra-class and inter-class variance). To differentiate the images from these two domains, we define the labeled and unlabeled data from the consumer photos as $D_l^T = (\mathbf{x}_i^T, y_i^T)|_{i=1}^{n_l}$ and $D_u^T = \mathbf{x}_i^T|_{i=n_l+1}^{n_l+n_u}$, respectively, where $y_i^T \in \{\pm 1\}$ is the label of $\mathbf{x}_i^T$. We further denote D^w as the data set from the source domain, and $D^T = D_l^T \cup D_u^T$ as the data set from the target domain with the size $n_T = n_l + n_u$.

1) Cross-Domain Learning: To utilize all training data from both consumer photos (target domain) and web images (source domain) for image retrieval, one can apply cross-domain learning methods [45], [46], [11], [7], [23], [12], [13]. Yang *et al.* [46] proposed Adaptive Support Vector Machine (A-SVM), where a new SVM classifier $f^T(\mathbf{x})$ is adapted from an existing auxiliary SVM classifier $f^s(\mathbf{x})$ trained with the data from the source domain. Specifically, the new decision function is formulated as:

$$f^T(\mathbf{x}) = f^s(\mathbf{x}) + \Delta f(\mathbf{x}), \tag{4}$$

where the perturbation function $\Delta f(\mathbf{x})$ is learned using the labeled data D_l^T from the target domain. As shown in [46], the perturbation function can be learned by solving quadratic programming (QP) problem which is similar to that of SVM.

Besides A-SVM, many existing works on cross-domain learning attempted to learn a new representation that can bridge the source domain and the target domain. Jiang *et al.* [23] proposed cross-domain SVM (CD-SVM), which uses k-nearest neighbors from the target domain to define a weight for each auxiliary pattern, and then the SVM classifier is trained with re-weighted samples. Daumé III [11] proposed the Feature Augmentation method to augment features for domain adaptation. The augmented features are used to construct a kernel function for kernel methods. It is important to note that most cross-domain learning methods [45], [46], [11], [23] *do not* consider the use of unlabeled data in the target domain. Recently, Duan *et al.* proposed a cross-domain kernel-learning method, referred to as Domain Transfer SVM (DTSVM) [12], and a multiple-source domain adaptation method called Domain Adaptation Machine (DAM) [13]. These methods can be readily used to exploit the data from both source domain and target domain for relevance feedback component in our general photo retrieval framework. However, these methods are either variants of SVM or in tandem with non-linear SVM or other kernel methods. Therefore, these methods are not efficient enough for large-scale retrieval applications. Therefore, we propose two effective and efficient cross-domain methods for relevance feedback.

2) Cross-Domain Combination of Classifiers: To further improve photo retrieval performance, a brute-force solution is to combine the web images and the annotated consumer photos to re-train a new classifier. However, the feature distributions of photos from different domains are drastically different, causing such classifier to perform poorly. Moreover, it is also inefficient to re-train the classifier using the data from both domains for online relevance feedback. To

significantly reduce the training time, the decision stump ensemble classifier and the linear SVM classifier $f^s(\mathbf{x})$ discussed in Section IV-C can be reused as the auxiliary classifier for relevance feedback. Here, we propose a simple cross-domain learning method, referred to as Cross-Domain Combination of Classifiers (CDCC), by simply combining the source classifier learned from the labeled data in the source domain D^w, and the target classifier (non-linear SVM with RBF kernel, referred to as SVM_T) learned from limited labeled data in the target domain D_l^T. The output of SVM_T is also converted into the range $[-1, 1]$ by using the symmetric sigmoid activation function and then the outputs of source classifier and SVM_T are combined with equal weights.

Schweikert et al. [31] also proposed to combine the source classifier and the target classifier for cross-domain learning. However, the source classifier used in their work is non-linear SVM with RBF kernel. It will be shown in our experiments such non-linear SVM cannot be used as the source classifier in this application because it cannot achieve real-time retrieval performance even on a small test dataset. Moreover, our system is the first work to apply Cross-Domain Combination of Classifiers for relevance feedback in photo retrieval applications.

3) Cross-Domain Regularized Regression: Besides CDCC, we also introduce a new learning method, namely Cross-Domain Regularized Regression (CDRR). In the following, we denote the transpose of vector or matrix by a superscript $'$. For the i-th sample $\mathbf{x}_i$, we denote $f_i^T = f^T(\mathbf{x}_i)$ and $f_i^s = f^s(\mathbf{x}_i)$, where $f^T(\mathbf{x})$ is the target classifier and $f^s(\mathbf{x})$ is the pre-learnt auxiliary classifier. Let us also denote $\boldsymbol{f}_l^T = [f_1^T, \ldots, f_{n_l}^T]'$ and $\mathbf{y}_l^T = [y_1^T, \ldots, y_{n_l}^T]'$. The empirical risk functional of $f^T(\mathbf{x})$ on the labeled data in the target domain is:

$$\sum_{i=1}^{n_l} (f_i^T - y_i^T)^2 = \|\boldsymbol{f}_l^T - \mathbf{y}_l^T\|^2. \tag{5}$$

For the unlabeled target patterns D_u^T in the target domain, let us define the decision values from the target classifier and the auxiliary classifier as $\boldsymbol{f}_u^T = [f_{n_l+1}^T, \ldots, f_{n_T}^T]'$ and $\boldsymbol{f}_u^s = [f_{n_l+1}^s, \ldots, f_{n_T}^s]'$, respectively. We assume that the target classifier $f^T(\mathbf{x})$ should have similar decision values as the pre-computed auxiliary classifier $f^s(\mathbf{x})$ [13]. We propose a regularization term to enforce the constraint that the label predictions of the target decision function $f^T(\mathbf{x})$ on the unlabeled data D_u^T in the target domain should be similar to the label predictions by the

auxiliary classifier $f^s(\mathbf{x})$ (see Figure 3), *i.e.*,

$$\frac{1}{2n_u}\sum_{i=n_l+1}^{n_T}(f_i^T - f_i^s)^2 = \frac{1}{2n_u}\|\boldsymbol{f}_u^T - \boldsymbol{f}_u^s\|^2. \tag{6}$$

We simultaneously minimize the empirical risk of labeled patterns in (5) and the penalty term in (6). The proposed method is then formulated as follows:

$$\min_{f^T}\Omega(f^T) + C\left(\lambda\|\boldsymbol{f}_l^T - \mathbf{y}_l^T\|^2 + \frac{1}{2n_u}\|\boldsymbol{f}_u^T - \boldsymbol{f}_u^s\|^2\right), \tag{7}$$

where $\Omega(f^T)$ is a regularizer to control the complexity of the target classifier $f^T(x)$, the second term is the prediction error of the target classifier $f^T(x)$ on the target labeled patterns D_l^T, and the last term controls the agreement between the target classifier and the auxiliary classifier on the unlabeled samples in D_u^T, and $C > 0$ and $\lambda > 0$ are the tradeoff parameters for the above three terms. Note that we use the factor $\frac{1}{2n_u}$ in the last term because we have very limited labeled data (less than 10 samples in our experiments) and much more unlabeled consumer photos.

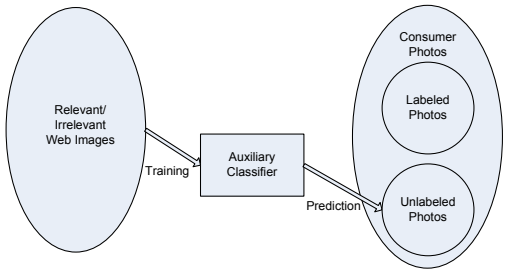

Fig. 3. Illustration of Cross-Domain Regularized Regression.

Assume that the target decision function is a linear regression function, *i.e.*, $f^T(\mathbf{x}) = \mathbf{w}'\mathbf{x}$ for image retrieval, and the regularizer as $\Omega(f^T) = \frac{1}{2}\|\mathbf{w}\|^2$, the optimal projection vector $\mathbf{w}$ in the structural risk functional (7) can be solved by a linear system:

$$\left(\mathbf{I} + \mathcal{C}\lambda\mathbf{X}_l\mathbf{X}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{X}_u'\right)\mathbf{w} = C\lambda\mathbf{X}_l\mathbf{y}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{f}_u^s, \tag{8}$$

where $\mathbf{X}_l = [\mathbf{x}_1^T, \ldots, \mathbf{x}_{n_l}^T]$ and $\mathbf{X}_u = [\mathbf{x}_{n_l+1}^T, \ldots, \mathbf{x}_{n_T}^T]$ are the data matrix of labeled and unlabeled consumer photos, and $\mathbf{I}$ is the identify matrix. Finally, we have the closed-form solution:

$$\mathbf{w} = \left(\mathbf{I} + C\lambda\mathbf{X}_l\mathbf{X}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{X}_u'\right)^{-1}\left(C\lambda\mathbf{X}_l\mathbf{y}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{f}_u^s\right). \tag{9}$$

4) Incremental Cross-Domain Regularized Regression: In the past several years, many incremental learning methods [1], [4] have been proposed for dimension reduction and classification. In this work, we propose an incremental cross-domain learning method, referred to as Incremental Cross-Domain Regularized Regression (ICDRR), to significantly accelerate the relevance feedback process in large-scale consumer photo retrieval.

In our ICDRR, we incrementally update two matrices $\mathbf{A}_1 = \boldsymbol{X}_l\boldsymbol{X}_l'$, $\mathbf{A}_2 = \boldsymbol{X}_u\boldsymbol{X}_u'$ and two vectors $\mathbf{b}_1 = \boldsymbol{X}_l\mathbf{y}_l'$, $\mathbf{b}_2 = \boldsymbol{X}_u\boldsymbol{f}_u^s$ in Eq. (9). Let us denote $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$ in the r-th round of relevance feedback as $\mathbf{A}_1^{(r)}$, $\mathbf{A}_2^{(r)}$, $\mathbf{b}_1^{(r)}$, $\mathbf{b}_2^{(r)}$, respectively. Before relevance feedback (*i.e.*, the 0-th round), we initialize $\mathbf{A}_1^{(0)} = \mathbf{0}$, $\mathbf{A}_2^{(0)} = \boldsymbol{X}\boldsymbol{X}'$, $\mathbf{b}_1^{(0)} = \mathbf{0}$, $\mathbf{b}_2^{(0)} = \boldsymbol{X}\boldsymbol{f}^s$, where $\boldsymbol{X}$ is the data matrix of all consumer photos, $\boldsymbol{f}^s$ is the output of source classifier on all consumer photos. In the r-th round of relevance feedback, we then incrementally update $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{b}_1$ and $\mathbf{b}_2$ by:

$$\mathbf{A}_1^{(r)} = \mathbf{A}_1^{(r-1)} + (\Delta\boldsymbol{X})(\Delta\boldsymbol{X})' \tag{10}$$

$$\mathbf{A}_2^{(r)} = \mathbf{A}_2^{(r-1)} - (\Delta\boldsymbol{X})(\Delta\boldsymbol{X})' \tag{11}$$

$$\mathbf{b}_1^{(r)} = \mathbf{b}_1^{(r-1)} + (\Delta\boldsymbol{X})(\Delta\mathbf{y}) \tag{12}$$

$$\mathbf{b}_2^{(r)} = \mathbf{b}_2^{(r-1)} - (\Delta\boldsymbol{X})(\Delta\boldsymbol{f}^s). \tag{13}$$

In the above equations, $\Delta\boldsymbol{X} \in \mathbb{R}^{n_d \times n_c}$, $\Delta\mathbf{y} \in \mathbb{R}^{n_c}$ and $\Delta\boldsymbol{f}^s \in \mathbb{R}^{n_c}$ are the data matrix, label vector, and the response vector from source classifier of the newly labeled consumer photos in the current round, where n_c is the number of user-labeled consumer photos in this round. The user only labels a very limited number of consumer photos in each round of relevance feedback, the computational cost for updating $\mathbf{A}_1^{(r)}$, $\mathbf{A}_2^{(r)}$, $\mathbf{b}_1^{(r)}$ and $\mathbf{b}_2^{(r)}$ becomes trivial in our ICDRR. Moreover, $\mathbf{A}_2^{(0)} = \boldsymbol{X}\boldsymbol{X}'$ can be computed offline because it does not depend on the source classifier, and $\mathbf{b}_2^{(0)} = \boldsymbol{X}\boldsymbol{f}^s$ can be computed when the user inspects the initial retrieval result (it costs less than 0.15 seconds with one single CPU thread even on the large *NUS-WIDE* dataset with about 270K images). Therefore in our experiments, we do not count the time for calculating $\mathbf{A}_2^{(0)}$ and $\mathbf{b}_2^{(0)}$. It will be shown in the experimental results that ICDRR significantly accelerates the relevance feedback process for large scale photo retrieval.

V. Experiments

We evaluate the performance of our proposed framework for textual query based consumer photo retrieval. First, we compare the initial retrieval performances of kNN classifier, decision

stump ensemble classifier, and linear SVM classifier *without* using relevance feedback. Second, we evaluate the performance of our proposed cross-domain relevance feedback methods CDCC and CDRR.

A. Dataset and Experimental Setup

We have downloaded about 1.3 million photos from the photo forum Photosig as the training dataset. Most of the images are accompanied by rich surrounding textual descriptions (*e.g.*, title, category and description). After removing the high-frequency words that are not meaningful (*e.g.*, "the", "photo", "picture"), our dictionary contains 21,377 words, and each image is associated with about five words on the average. Similarly to [42], we also observed that the images in Photosig generally are high quality with the sizes varying from 300×200 to 800×600. In addition, the surrounding descriptions reasonably describe the semantics of the corresponding images.

We test the performance of our retrieval framework on two datasets. The first test dataset is derived (under a usage agreement) from the Kodak Consumer Video Benchmark Dataset [28], which was collected by Eastman Kodak Company from about 100 real users over the period of one year. In this dataset, 5,166 key-frames (the image sizes vary from 320×240 to 640×480) were extracted from 1,358 consumer video clips. Key-frame based annotation was performed by the students at Columbia University to assign binary labels (presence or absence) for each visual concept. 25 semantic concepts were defined, including 22 visual concepts and three audio-related concepts (*i.e.*,"singing" , "music" and "cheer"). We also merge two concepts "group_of_two" and "group_of_three_or_more" into a single concept "people" for the convenience of searching the relevant and irrelevant images from the Photosig web image dataset. Observing that the key frames from the same video clip can be near duplicate images, we select only the first key frame from each video clip in order to perform a fair comparison of different algorithms. In total, we test our framework on 21 visual concepts and with 1,358 images.

The second dataset is *NUS-WIDE*[8], which was recently collected by the National University of Singapore (NUS). In total, this dataset has 269,648 images and their ground-truth annotations for 81 concepts. The images in *NUS-WIDE* dataset are downloaded from the online consumer photo sharing website Flickr.com. We choose *NUS-WIDE* dataset because it is the largest annotated consumer photo dataset available to researchers today, and is suitable for testing the

performances of our framework for large-scale photo retrieval. Moreover, it is also meaningful to use this dataset to test the retrieval precisions of our cross-domain relevance feedback methods CDCC and CDRR because the data distributions of photos downloaded from different websites (*i.e.*, Photosig.com and Flickr.com) are still different. It is also worth mentioning that the images in *NUS-WIDE* are used as raw photos, in other words, we do not consider the associated tag information in this work.

In our experiments, we use three types of global features. For Grid Color Moment (GCM), we extract the first three moments of three channels in the LAB color space from each of the 5×5 fixed grid partitions, and aggregate the features into a single 225-dimensional feature vector. The Edge Direction Histogram (EDH) feature includes 73 dimensions with 72 bins corresponding to edge directions quantized in five angular bins and one bin for non-edge pixels. Similar to [8], we also extract 128-D Wavelet Texture (WT) feature by performing Pyramid-structured Wavelet Transform (PWT) and Tree-structured Wavelet Transform (TWT). Finally, each image is represented as a single 426-D vector by concatenating the three types of global features. Please refer to [8] for more details about the features. While it is possible to use other local features, such as SIFT descriptors, we use the above global features because they can be efficiently extracted over the large image corpus and they have been shown to be effective for consumer photo annotation in [6], [8]. It is also convenient for fair assessment of other known systems that use the same types of visual features.

For the training dataset *photosig*, we calculate the original mean value μ_d and standard deviation σ_d for each dimension d, and normalize all dimensions to zero mean and unit variance. We also normalize the test datasets (*i.e.*, *Kodak* and *NUS-WIDE*) by using μ_d and σ_d. In our experiment, all algorithms are implemented with C++. Matrix and vector operations are performed using the Intel Math Kernel Library 10. Experiments are performed on a server machine with dual Intel Xeon 3.0GHz Quad-Core CPUs (eight threads) and 16GB Memory. In time cost analysis, we do not consider the time of loading the data from the hard disk because the data can be loaded for once and then used for subsequent queries.

B. Retrieval without Relevance Feedback

Considering that the queries by the CBIR methods and our framework are different in nature, we cannot compare our work directly with the existing CBIR methods before relevance feedback.

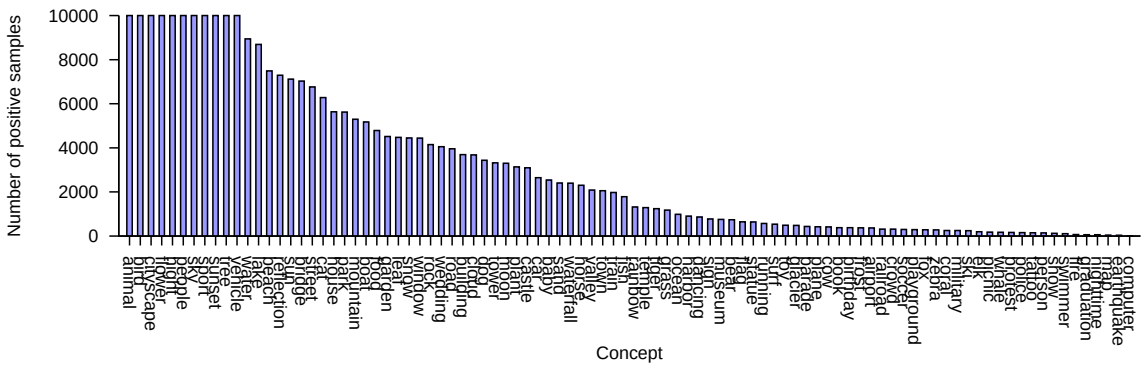

Fig. 4. Number of randomly selected positive samples for each concept in the training web image database.

We also cannot compare the retrieval performance of our framework directly with web data-based annotation methods, because of the following two aspects: 1) These prior works [25], [35], [36], [38], [41], [42] only output binary decisions (presence or absence) without providing a metric to rank the personal photos; 2) An initial textual term is required before image annotation in [22], [41], [42] and their annotation performances depend heavily on the correct textual term, making it difficult to compare their methods fairly with our automatic technique. However, we notice that the previous web data-based image annotation methods [25], [35], [36], [38], [41], [42] all used kNN classifier for image annotation, possibly owning to its simplicity and effectiveness. Therefore, we directly compare the retrieval performance of decision stump ensemble classifier, linear SVM classifier, and the baseline kNN classifier.

Suppose a user wants to use the textual query q to retrieve the relevant personal images. For both methods, we randomly select $n_p = \min(10000, n_q)$ positive web images from *photosig* dataset, where n_q is the total number of images that contain the word q in the surrounding textual descriptions. *Kodak* and *NUS-WIDE* contains 94 distinct concepts in total ("animal", "beach", "boat", "dancing", "person", "sports", "sunset" and "wedding" appear in both datasets). The average number of selected positive samples of all the 94 concepts is 3088.3, and Figure 4 plots the number of positive samples for each concept.

To improve the speed and reduce the memory cost, we perform Principal Component Analysis (PCA) using all the images in the photosig dataset. We also investigate the performances of two possible fusion methods to fuse three types of global features in this application.

- **Early Fusion**: We concatenate the three types of features before performing PCA. We observe that the first $n_d = 103$ principal components are sufficient to preserve 90% energy.

After dimension reduction, all the images in training and test datasets are projected into the 103-D space for further processing.

- **Late Fusion**: We perform PCA on three types of features independently. We observe that the first $n_{d1} = 91$, $n_{d2} = 24$, $n_{d3} = 5$ principal components are sufficient to preserve 90% energy for GCM, EDH and WT features, respectively. Then, these three types of features of all the images in the training and test datasets are projected to n_{d1}-D, n_{d2}-D, n_{d3}-D space after dimension reduction. We train independent classifiers based on each type of feature. Finally, the classifiers from different features are linearly combined with the combination weights determined based on the training error rates.

For each fusion method, we compare the following three methods:

- k**NN_S**: We only use the positive images from the web-image database as the training data. For each consumer photo from the testing dataset, we find the top-k nearest neighbors in the positive images, and use the average distance to measure the relevance between the textual query to the testing consumer photo. In the experiment, we set $k = 200$. We also perform exhaustive exact kNN search accelerated by SIMD CPU instructions and multiple threads. For kNN based method with late fusion, we combine the outputs of all kNN classifiers with equal weights because the training error rate of kNN classifier on each type of feature is unknown in this case. In the sequel, we denote kNN_S with early fusion and late fusion by kNN_SE and kNN_SL, respectively.
- **DS_S**: We randomly choose n_p negative samples for n_s times, and in total we train $n_s n_d$ decision stumps for early fusion (referred to as DS_SE) or $3n_s n_d$ (referred to as DS_SL) for late fusion. After removing the 20% decision stumps with the largest training error rates, we apply $0.8n_s n_d$ or $2.4n_s n_d$ decision stumps for the testing stage in DS_SE and DS_SL, respectively.
- **LinSVM_S**: We also randomly choose n_p negative samples for n_s times. In total, we train n_s linear SVM classifiers for early fusion (referred to as LinSVM_SE) or $3n_s$ classifiers for late fusion (referred to as LinSVM_SL). In this work, we use tools from LibLinear [14] in our implementations and use the default value 1 for the parameter C_{SVM}.

There are 21 and 81 concept names from the *Kodak* dataset and *NUS-WIDE* dataset, respectively. They are used as textual queries to perform image retrieval. Precision (defined as the

percentage of relevant images in the top I retrieved images) is used as the performance measure to evaluate the retrieval performance. Since online users are usually interested in the top ranked images only, we set I as 20, 30, 40, 50, 60 and 70 for this study, similarly to in [33].

1) Comparison of precision: We tested all the methods above for initial retrieval without using relevance feedback. For *Kodak* dataset, we set the random sampling times n_s for generating negative samples as 50 for DS_SE and DS_SL, and 10 for LinSVM_SE and LinSVM_SL in order to make the running time of initial retrieval process under 1 second. The precisions of all methods are shown in Figure 5. We observe that DS_SE, DS_SL, LinSVM_SE and LinSVM_SL are much better than kNN_SE and kNN_SL. This is possibly because kNN_SE and kNN_SL only utilize the positive web images while other methods take advantage of both the positive and negative web images to train the more robust classifiers. Moreover, the average values of the top-20,30,40,50,60 and 70 precisions from LinSVM_SL, DS_SL, LinSVM_SE and DS_SE, are 14.50%, 14.47%, 14.39% and 14.21%, respectively. We conclude that the linear SVM classifier and decision stump ensemble classifier achieve comparable retrieval performances on the *Kodak* dataset.

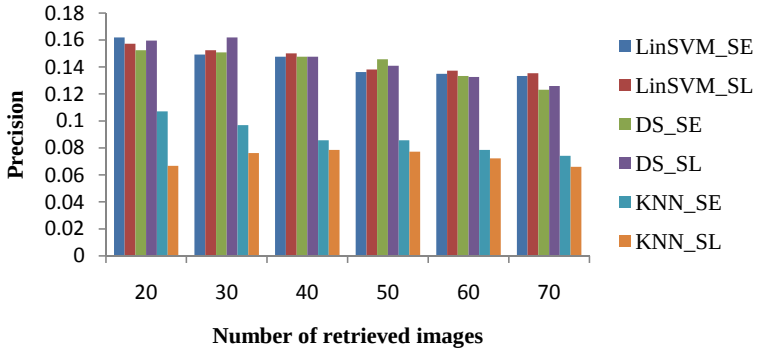

Fig. 5. Retrieval precisions using kNN classifier, decision stump ensemble classifier, and linear SVM classifier on the Kodak dataset (1358 images, 21 concepts).

To better compare the performances of different algorithms, we also test them on the large *NUS-WIDE* dataset. In Figure 6, we plot the precision variations of different algorithms with respect to different values of n_s, in which n_s is set to 1,3,5,7 and 10. We have the following observations:

1) Again, kNN_SE and kNN_SL achieve much worse performances, when compared with the

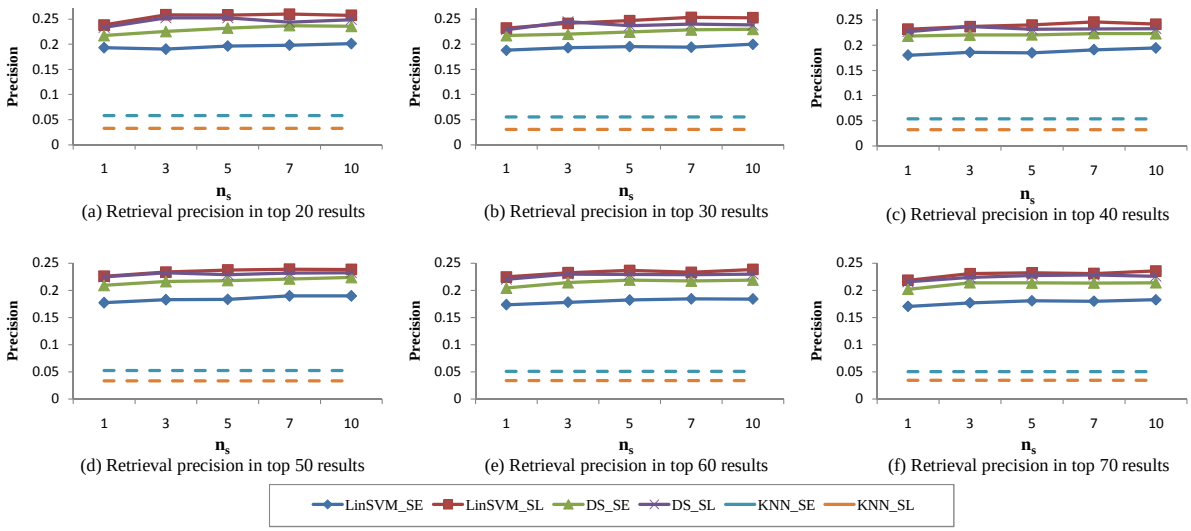

Fig. 6. Retrieval precisions using kNN classifier, decision stump ensemble classifier, and linear SVM classifier on NUS-WIDE dataset (269,648 images, 81 concepts). Since the precisions of kNN_SE and kNN_SL are irrelevant with respect to n_s , their precisions are presented with dashed curves.

Fig. 7. Top-10 retrieval results for query "water" on the Kodak dataset. Incorrect retrieval results are highlighted with green boxes.

other four algorithms. LinSVM_SL generally achieves the best results and it is slightly better than DS_SL in most cases.

2) When n_s increases, DS_SE, DS_SL, LinSVM_SE, and LinSVM_SL improve in most cases, which is consistent with the recent work [33].

3) It is interesting to observe that LinSVM_SE is the worst among four algorithms related to linear SVM and decision stump ensemble classifiers. We employ three types of features (color, edge and texture) in this work and it is well known that none of them can work well for all concepts. LinSVM_SL, DS_SL and DS_SE achieve better performance, possibly because they can fuse and select different type of features or even feature dimensions based on the training error rates.

4) Except for kNN classifier based algorithms, we also observe that the late fusion based methods are generally better than the corresponding early fusion based methods for photo retrieval on the *NUS-WIDE* dataset. kNN_SL is worse than kNN_SE. However, in kNN_SL, all types of features are combined with equal weights, namely, feature selection is not performed in kNN_SL.

Fig. 8. Top-10 retrieval results for query "animal" on the NUS-WIDE dataset. (a) Initial results; (b) Results after 1 round of relevance feedback (one positive and one negative images are labeled in each round). Incorrect results are highlighted by green boxes.

A visual example is shown in Figure 7. We use the keyword "water" to retrieve images from the *Kodak* dataset using LinSVM_SL with 10 SVM classifiers. Note that this query is *undefined* in the concept lexicon of the *Kodak* dataset. Our retrieval system produces eight diverse yet relevant images out of the top 10 retrieved images. One more visual example of our system using LinSVM_SL with 10 SVM classifiers is shown in Figure 8(a). We use the keyword "animal" to retrieve images from the *NUS-WIDE* dataset ("animal" is defined in the concept lexicon of *NUS-WIDE*). Our retrieval system produces six relevant images out of the top 10 retrieved images. In the subsequent subsection, we will show that our proposed CDRR relevance feedback method can significantly improve the retrieval performance (See Figure 8(b)).

2) Comparison of running time: We also compare the running time of all algorithms on the two datasets. In this work, each decision stump classifier and SVM classifier can be trained and used independently, and exhaustive kNN search is also easy to parallelize. We therefore use a simple but effective parallelization scheme, OpenMP, to take advantages of eight threads of our server for each method.

On the *Kodak* dataset, kNN_SE and kNN_SL spend 0.872 and 1.033 seconds, respectively, for the initial retrieval process. DS_SE and DS_SL with $n_s = 50$, LinSVM_SE and LinSVM_SL with $n_s = 10$ spend 0.912, 0.969, 0.830, and 0.852 seconds, respectively. All methods can achieve real-time retrieval performance on this small dataset.

The comparison of the running time on the *NUS-WIDE* dataset is plotted in Figure 9. On this dataset, kNN_SE and kNN_SL spend 213.35 and 225.73 seconds, respectively. We implement kNN based on exhaustive search, thus it takes much more time when compared with decision stump ensemble classifier and linear SVM classifier. When n_s is 10, the total running time of LinSVM_SE, LinSVM_SL, DS_SE and DS_SL are 0.782, 0.878, 1.373 and 1.575 seconds,

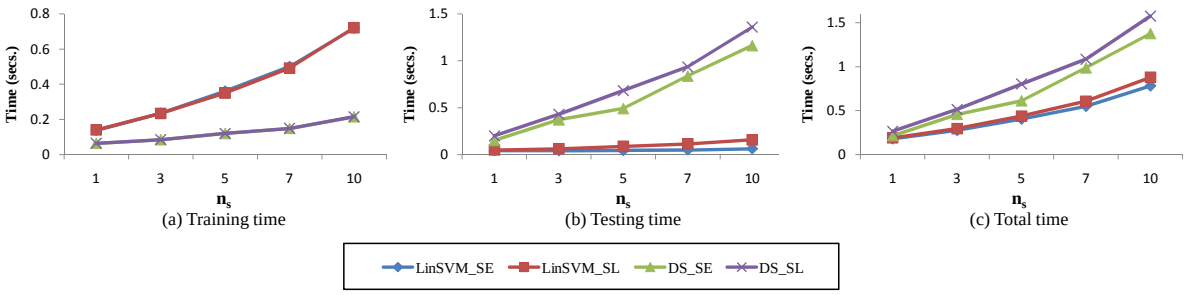

Fig. 9. Time cost of retrieval using decision stumps and SVMs with linear kernel on NUS-WIDE dataset (269,648 images, 81 concepts). Note that "total time" stands for the sum of training time and testing time.

respectively. We also observe that LinSVM_SE and LinSVM_SL generally cost more time than DS_SE and DS_SL in the training stage. However, the testing stage of LinSVM_SE and LinSVM_SL is much faster, making the total running time of initial retrieval process much shorter than DS_SE and DS_SL.

3) Discussions: From the experiments on the *Kodak* dataset, we observe that linear SVM and decision stump ensemble classifiers based methods are generally comparable in terms of initial retrieval precision and speed. Since all the algorithms can achieve real-time speed, any of them can be used for initial retrieval on a small dataset. However, for large-scale photo retrieval, LinSVM_SL is preferred for the initial retrieval process because of its effectiveness and real-time response.

C. Retrieval with Relevance Feedback (RF)

In this subsection, we evaluate the performance of a few relevance feedback methods. For fair comparison, we choose LinSVM_SL with 10 SVM classifiers, the best algorithm in terms of overall performances (See Section V-B), for initial retrieval before relevance feedback. LinSVM_SL is also accordingly chosen as the source classifier in our methods CDCC and CDRR. From here on, we also refer to CDCC as LinSVM_SL+SVM_T, in which the responses from LinSVM_SL and SVM_T are equally combined. In our LinSVM_SL+SVM_T, CDRR and two conventional manifold ranking and SVM based relevance feedback algorithms [17], [47], we also adopt the late fusion scheme used in LinSVM_SL to integrate the three types of global features, namely, the three types of features are used independently at first and the decisions or responses are finally fused. The early fusion approach is used for the prior cross-domain learning method A-SVM [46] because it is faster.

We compare our LinSVM_SL+SVM_T method and CDRR with the following methods:

1) **SVM_T**: SVM has been used for RF in several existing CBIR methods [33], [34], [47]. We train non-linear SVM with an RBF kernel based on the labeled images in the target domain, which are marked by the user in the current and all previous rounds. We use LibSVM package [5] in our implementation and use its default setting for RBF kernel (*i.e.* C is set as 1 and γ in RBF kernel is set as $\frac{1}{91}$, $\frac{1}{24}$ and $\frac{1}{5}$ for GCM, EDH and WT features, respectively).

2) **MR**: Manifold Ranking (MR) is a semi-supervised RF method proposed in [17]. The two parameters α and γ for this method are set according to [17].

3) **A-SVM**: Adaptive SVM (A-SVM) is a recently proposed method [46] for cross-domain learning as described in Section IV-D.1, in which SVM based on an RBF kernel is used as the source classifier to obtain the initial retrieval results. The parameter setting is the same as that in SVM_T. Considering the running time of A-SVM is much higher than other methods even on the small *Kodak* dataset, we do not test it on the large *NUS-WIDE* dataset because it cannot achieve real time response.

As in other methods [17], [46], [47], several parameters needed to be decided beforehand. In LinSVM_SL+SVM_T, we need to determine the parameters in SVM_T and we use the same parameters setting as that in SVM_T. For CDRR, we empirically fix $C = 70.0$ and set $\lambda = 0.05$ on the *Kodak* dataset and $\lambda = 0.02$ on the *NUS-WIDE* dataset. In addition, we also observe that CDRR generally achieves better performance, if we respectively set $y_i^T = 1$ and $y_i^T = -0.1$ for positive and negative consumer photos, when compared with the setting $y_i^T = 1$ and $y_i^T = -1$. We set $y_i^T = -0.1$ for negative images because the negative images marked by the user in relevance feedback are still top ranked images, namely, these images are not the *extremely* negative images. Note that similar observations are also reported in [17]. It is still an open problem to automatically determine the optimal parameters in CDRR, which will be investigated in the future.

1) Comparison of precision: In real circumstances, users typically would be reluctant to perform many rounds of relevance feedback or annotate many images for each round. Therefore, we only report the results from the first four rounds of feedback. In each feedback round, the top one relevant image (*i.e.*, the highest ranked image with the same semantic concept as the textual query) is marked as a positive feedback sample from among the top 40 images. Similarly, one negative sample is marked out of the top 40 images. In Figure 8(b), we show top-10 retrieved images after 1 round of relevance feedback for the query "animal" on the *NUS-WIDE* dataset.

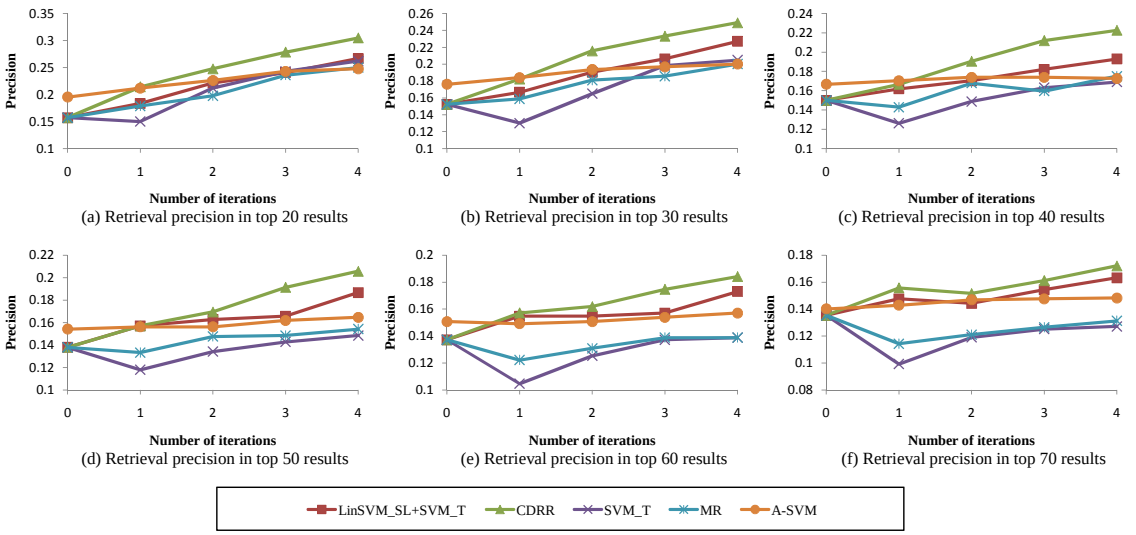

Fig. 10. Retrieval results after relevance feedback(one positive and one negative feedbacks per round) on the Kodak dataset (1358 images, 21 concepts).

We observe that the results are improved considerably after using our proposed CDRR relevance feedback algorithm. Figures 10 and Figure 11 compare different relevance feedback methods on the *Kodak* dataset and the *NUS-WIDE* dataset, respectively.

From these results, we have the following observations:

1) Our CDRR and LinSVM_SL+SVM_T outperform the conventional RF methods SVM_T and MR, because of the successful utilization of the images from both domains. When comparing CDRR with SVM_T and MR, the relative precision improvements after RF are more than 14.7% and 13.5% on the *Kodak* and *NUS-WIDE* datasets, respectively. CDRR is generally better than or comparable with LinSVM_SL+SVM_T, and the retrieval performances of our CDRR and LinSVM_SL+SVM_T increase monotonically with more labeled images provided by the user in most cases. For CDRR, we believe that the retrieval performance can be further improved by using non-linear function in CDRR. However, it is a non-trivial task to achieve the real-time retrieval performance with an RBF kernel function. This will be investigated in the future.

2) For SVM_T, the retrieval performance drops after the first round of RF, but increase from the second iteration. The explanation is that SVM_T trained based on two labeled training images is not reliable, but its performance can improve when more labeled images are marked by the user in the subsequent feedback iterations.

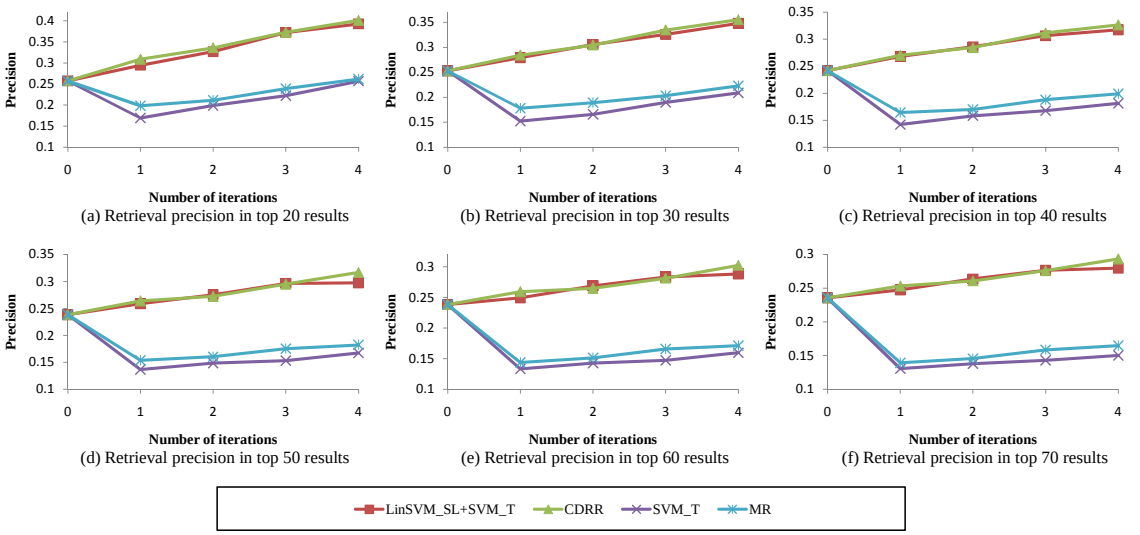

Fig. 11. Retrieval results after relevance feedback(one positive and one negative feedbacks per round) on the NUS dataset (269,648 images, 81 concepts).

Method	**ICDRR**	**CDRR**	**LinSVM_SL+SVM_T**	**SVM_T**	**MR**	**A-SVM**
Time	0.015	0.032	0.015	0.015	0.037	9.92

TABLE I

AVERAGE CPU TIME (IN SEC.) OF RELEVANCE FEEDBACK (PER ROUND) ON THE KODAK DATASET.

3) Semi-supervised learning method MR can improve the retrieval performance only in some cases on the *Kodak* dataset, possibly because the manifold assumption does not hold well for unconstrained consumer images.

4) The performance of A-SVM is slightly improved after using RF in most cases. It seems that the limited number of labeled target images from the user are not sufficient to facilitate robust adaptation for A-SVM. We also observe that initial results of A-SVM is better than other algorithms on the *Kodak* dataset because of the utilization of non-linear SVM for initialization. However, it takes 324.3 seconds with one thread for the initial retrieval process even on the small-scale *Kodak* dataset, making it infeasible for practical image retrieval applications even with eight threads.

2) Comparison of running time: In this Section, we compare the running time of all relevance feedback algorithms used in our experiment. Considering that all the algorithms except A-SVM and MR on the *NUS-WIDE* dataset are very responsive, we test all the algorithms by using only

Method	ICDRR	CDRR	LinSVM_SL+SVM_T	SVM_T	MR
Time	0.110	1.534	1.277	1.277	60.533

TABLE II

AVERAGE CPU TIME (IN SEC.) OF RELEVANCE FEEDBACK (PER ROUND) ON THE NUS-WIDE DATASET.

one single thread for relevance feedback.

The comparison of time cost on the *Kodak* dataset is shown in Table I. All methods except A-SVM are able to achieve the interactive speed on this small dataset. In addition, the incremental cross-domain learning method ICDRR is faster than CDRR.

In Table II, we report the running time of different algorithms on the *NUS-WIDE* dataset. MR is no longer responsive in this case because the label propagation process based on the graph with much more vertices becomes much slower. The RF process of CDRR and LinSVM_SL+SVM_T (or SVM_T) is still responsive (1.534 seconds and 1.277 seconds only), because we only need to train SVM with less than 10 training samples for LinSVM_SL+SVM_T and SVM_T or solve a linear system for CDRR.

Moreover, ICDRR only takes about 0.1 seconds per round after incrementally updating the corresponding matrices, which is much faster than CDRR. We also observe that the running time of LinSVM_SL+SVM_T (or SVM_T) increases when the number of user-labeled consumer photos increases in the subsequent iterations. Specifically, When the user labels 1, 2, 3, 4 positive consumer photos and the same number of negative photos, LinSVM_SL+SVM_T (or SVM_T) costs about 0.7, 1.1, 1.5 and 1.9 seconds, respectively. However, ICDRR takes about 0.1 seconds in all the iterations.

In short, ICDRR can learn the same projection vector $\mathbf{w}$ and achieve the same retrieval precisions as CDRR, but it is much more efficient than CDRR and LinSVM_SL+SVM_T for relevance feedback in large scale photo retrieval.

VI. CONCLUSIONS

By leveraging a large collection of web data (images accompanied by rich textual descriptions), we have proposed a real-time textual query based personal photo retrieval system, which can retrieve consumer photos without using any intermediate image annotation process. For a given textual query, our system can automatically and efficiently retrieve relevant and irrelevant web

images using the inverted file method and *WordNet*. With these retrieved web images as the training data, we employ three efficient classification methods, kNN classifier, decision stump ensemble classifier and linear SVM classifier, for consumer photo retrieval. We also propose two novel relevance feedback methods, namely CDCC and CDRR by utilizing the pre-learned auxiliary classifier and the feedback images to effectively improve the retrieval performance at interactive response time. Moreover, an incremental cross-domain learning method, referred to as ICDRR, is also developed for large scale consumer photo retrieval.

Extensive experimental results on the *Kodak* and *NUS-WIDE* consumer photo datasets clearly demonstrate that decision stump ensemble and linear SVM classifiers based methods are much better than kNN based methods for initial photo retrieval. Linear SVM classifier based method is preferred on a large photo dataset like *NUS-WIDE*, thanks to its effectiveness and faster and real-time response. Our experiments also demonstrate that the proposed relevance feedback approaches CDRR and LinSVM_SL+SVM_T require an extremely limited amount of feedback from the user and it outperforms two conventional manifold ranking and SVM based relevance feedback methods, and Incremental CDRR is much faster than CDRR and LinSVM_SL+SVM_T on the large *NUS-WIDE* dataset. Moreover, our proposed system can also retrieve consumer photos with a textual query that is not included in the predefined lexicons.

In summary, we have proposed a general photo retrieval framework by using textual query. Our work falls into the recent research trend of "*Internet Vision*" where the massive and valuable web data including texts and images are used for various computer vision and computer graphics tasks (e.g., [9], [18], [40]). Other efficient and effective learning techniques can be readily developed and incorporated into our framework to further improve the initial photo retrieval and relevance feedback. For example, the fast Stochastic Intersection Kernel MAchine (SIKMA) training algorithm may be used in our framework for initial photo retrieval [40] and non-linear functions may be employed in CDRR to replace the current linear regression function. In addition, this framework also lends itself to personal video retrieval because key frames in videos can be used to retrieve videos readily for non-motion related textual queries. In the long run, such a framework can also be extended to process action related concepts [26] by explicitly incorporating motion related features.

REFERENCES

[1] M. Artae, M. Jogan, and A. Leonardis. Incremental PCA for on-line visual learning and recognition. In *International Conference on Pattern Recognition*, 2002.

[2] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of Association for Computational Linguistics*, 2007.

[3] L. Cao, J. Luo, and T. S. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *ACM Multimedia*, 2008.

[4] G. Cauwenberghs and T. Poggio. Incremental and Decremental Support Vector Machine Learning. In *Neural Information Processing Systems*, 2000.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/˜cjlin/libsvm, 2001.

[6] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *ACM SIGMM Workshop on Multimedia Information Retrieval*, 2007.

[7] S.-F. Chang, J. He, Y. Jiang, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *NIST TRECVID Workshop*, 2008.

[8] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, 2009.

[9] N.I. Cinbis, R.G. Cinbis, and S. Sclaroff. Learning Actions From The Web. In *International Conference on Computer Vision*, 2009.

[10] R. Datta, D. Joshi, J. Li, and J.-Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 1–60, 2008.

[11] H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of Association for Computational Linguistics*, 2007.

[12] L. Duan, I. W. Tsang, D. Xu, and S. Maybank. Domain Transfer SVM for Video Concept Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[13] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain Adaptation from Multiple Sources via Auxiliary Classifiers. In *International Conference on Machine Learning*, 2009.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. In *Journal of Machine Learning Research* , 2008.

[15] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[16] R. Fergus, P. Perona, and A. Zisserman. A Visual Category Filter for Google Images. In *European Conference on Computer Vision*, 2004.

[17] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *ACM Multimedia*, 2004.

[18] J. Hays and A. Efros. Scene Completion Using Millions of Photographs. ACM Transactions on Graphics (SIGGRAPH 2007), 2007.

[19] X. He. Incremental semi-supervised subspace learning for image retrieval. In *ACM Multimedia*, 2004.

[20] R. Herbrich and T. Graepel. A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In *Neural Information Processing Systems*, 2001.

[21] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Semi-supervised svm batch mode active learning for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[22] J. Jia, N. Yu, and X.-S. Hua. Annotating personal albums via web mining. In *ACM Multimedia*, 2008.

[23] W. Jiang, E. Zavesky, S.-F. Chang. Cross-domain learning methods for high-level visual concept classification. In *IEEE International Conference on Image Processing*, 2008.

[24] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 985–1002, 2008.

[25] X. Li, L. Chen, L. Zhang, F. Lin, and W. Ma. Image annotation by large-scale content-based image retrieval. In *ACM Multimedia*, 2006.

[26] J. Liu, J. Luo and M. Shah. Recognizing Realistic Actions from Videos in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[27] Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Using Large-Scale Web Data to Facilitate Textual Query Based Retrieval of Consumer Photos. In *ACM Multimedia*, 2009.

[28] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, A. Yanagawa. Kodak's consumer video benchmark data set: concept definition and annotation. In *ACM Workshop on Multimedia Information Retrieval*, 2007.

[29] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE Multimedia Magazine*, 86–91, 2006.

[30] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *IEEE International Conference on Image Processing*, 1997.

[31] G. Schweikert, C. Widmer, B. Schölkopf, G. Rätsch An Empirical Analysis of Domain Adaptation Algorithms for Genomic Sequence Analysis. In Neural Information Processing Systems, 1433–1440, 2008.

[32] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1349–1380, 2000.

[33] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1088–1099, 2006.

[34] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM Multimedia*, 2001.

[35] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1958–1970, 2008.

[36] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[37] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision* , 137–154, 2004.

[38] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[39] C. Wang, L. Zhang, and H. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *ACM SIGIR*, 2008.

[40] G. Wang, D. Hoiem, and D. Forsyth. Learning Image Similarity from Flickr Groups Using Stochastic Intersection Kernel Machines. In *International Conference on Computer Vision*, 2009.

[41] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. AnnoSearch: Image auto-annotation by search. In *IEEE Conference on Computer Vision and Pattern Recognition* , 2006.

[42] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1919–1932, 2008.

[43] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Neural Information Processing Systems*, 2008.

[44] I. H. Witten, A. Moffat, and T. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Kaufmann Publishers, 1999.

[45] P. Wu and T. G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *International Conference on Machine Learning* , 2004.

[46] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM Multimedia*, 2007.

[47] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *IEEE International Conference on Image Processing* , 2001.

[48] X. Zhou and T. Huang. Small sample learning during multimedia retrieval using bias map. In *IEEE Conference on Computer Vision and Pattern Recognition* , 2001.