Using Large-Scale Web Data to Facilitate Textual Query Based Retrieval of Consumer Photos

Yiming Liu, Dong Xu, Ivor W. Tsang
School of Computer Engineering
Nanyang Technological University
Singapore
{YMLiu,DongXu,IvorTsang}@ntu.edu.sg

Jiebo Luo
Kodak Research Laboratories
Eastman Kodak Company
Rochester, USA
Jiebo.Luo@kodak.com

ABSTRACT

The rapid popularization of digital cameras and mobile phone cameras has lead to an explosive growth of consumer photo collections. In this paper, we present a (quasi) 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.*, "pool"), our system exploits the inverted file method to automatically find the positive web images that are related to the textual query "pool" as well as the negative web images which are irrelevant to the textual query. Based on these automatically retrieved relevant and irrelevant web images, we employ two simple but effective classification methods, k Nearest Neighbor (kNN) and decision stumps, to rank personal consumer photos. To further improve the photo retrieval performance, we propose three new relevance feedback methods via cross-domain learning. These methods effectively utilize both the web images and the consumer images. In particular, our proposed cross-domain learning methods can learn robust classifiers with only a very limited amount of labeled consumer photos from the user by leveraging the pre-learned decision stumps at interactive response time. Extensive experiments on both consumer and professional stock photo datasets demonstrated the effectiveness and efficiency of our system, which is also inherently not limited by any predefined lexicon.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Experimentation

Keywords

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

MM'09, October 19–24, 2009, Beijing, China.

1. 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 practical problem at the same time. In the recent decades, many Content Based Image Retrieval (CBIR) systems [21, 23, 24, 37] have been proposed. These systems usually require users to provide images as queries 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 high-level semantic concepts. This can be used as intermediate stage for textual query based image retrieval because the semantic concepts are analogous to the textual terms describing document contents. In general, the image annotation methods can be classified into two categories, learning-based methods and web-based methods [16]. 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-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 [3]. Observing that the personal photos are usually organized into collections by time, location and events, Cao *et al.* [2] proposed a label propagation method to propagate the concept labels from part of personal images to the other photos in the same album. In [16], 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 by 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-based image annotation techniques [16, 19, 25, 26, 28, 29, 31, 32]. 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 can be considered higher and visually more characteristic of semantics of the corresponding textual descriptions. After the user provides a textual query (*e.g.*, "pool"), our system exploits the inverted file to automatically retrieve the positive web images, which have the textual query "pool" in the surrounding descriptions, as well as the negative web images, whose surrounding descriptions do not contain the query "pool" and its descendants (such as "natatorium", "cistern", etc.) according to *WordNet* [11]. The inverted file method has been successfully used in information retrieval to efficiently find all text documents where a given word occurs [34]. Based on these automatically retrieved positive and negative web images, we employ classifiers, such as k Nearest Neighbor (kNN) and decision stumps, to rank the photos in the personal collections.

To improve the retrieval performance in CBIR, relevance feedback has been frequently used to acquire the search intention from the user. However, most of the users would prefer to label only a few images in a limited feedback, which frequently degrades the performance of the typical relevance feedback algorithms. 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 three new 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 decision stump ensemble classifier (referred to as auxiliary classifier). In particular, 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. Finally, we also propose a Hybrid scheme to take the advantages of the above two methods. It will be shown by the experimental results that our Hybrid method can significantly improve the final retrieval performance at interactive response time.

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 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 Hybrid approach further improves the photo retrieve performance by using the pre-learned classifier (auxiliary classifier) from loosely labeled web images, and precisely labeled consumer photos from relevance feedback.
- Our Hybrid approach significantly outperforms other cross-domain learning methods and two manifold ranking and SVM based relevance feedback methods [13, 37].
- Our system achieves interactive time (or quasi real-time) response thanks to the combined efficiency of decision stump classifiers, 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 the parallelization scheme OpenMP to take advantage of multiple threads.

The remainder of this paper is organized as follows. Sections 2 and 3 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 4. Extensive experimental results will be presented in Section 5, followed by concluding remarks in the final section.

2. RELATED WORK IN CONTENT BASED IMAGE RETRIEVAL (CBIR)

Over past 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 [22, 6] for more details. However, in consumer applications, it is more convenient 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 [21]. In [38], 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

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

[2]Fergus et al. proposed to use parts-based model to improve *Google* image search results in [12].

Machines (SVM) based relevance feedback techniques [23, 24, 37] 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 [21, 23, 24, 37, 38]. Semi-supervised learning [14, 15] and active learning [15, 24] have also been proposed to improve the performance of image retrieval. He [14] used the information from relevance feedback to construct a local geometrical graph to learn a subspace for image retrieval. Hoi *et al.* [15] 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 (quasi) 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 the accompanying textual descriptions. In addition, we also propose three 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. Cross-domain methods have been used in real applications, such as sentiment classification, text categorization, and video concept detection [1, 7, 8, 9, 17, 36]. However, these methods are either variants of SVM or in tandem with SVM or other kernel methods, making it inefficient for large-scale applications. In addition, the recent cross-domain learning works on image annotation [8, 9, 17, 36] 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.

3. 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-based methods [16]. In learning-based methods [2, 3, 18], 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, because the concept labels of the training samples need to be obtained through time consuming and expensive human annotation.

Recently, web-based methods were developed and these methods can be used to annotate general images. Zhang and his colleagues have proposed a series of works [19, 28, 29, 31, 32] 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. On 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 [31] 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 [19]. An annotation refinement algorithm [28] and a distance metric learning method [29] were also proposed to further improve the image annotation. Torralba *et al.* [25] collected about 80 million tiny images (color images with the size of 32 by 32 pixels), each one 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.* [26] and Weiss *et al.* [33] also developed two indexing methods to speed up the image search process by representing each image with less than a few hundred bits.

It is possible to perform textual query based image retrieval by using image annotation as 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.

4. 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 to facilitate textual query based image retrieval. Recall that the consumer photos (from personal collections) are usually organized in folders without any indexing to facilitate textual queries. To automatically retrieve consumer photos using textual queries, we choose to leverage millions of web images [3] and their surrounding texts as the bridge between the domains of the web images and the consumer photos.

4.1 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 [34]. 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 boost-

[3]One can assume that such a large-scale web image database contains sufficient images to cover almost all daily-life semantic concepts.

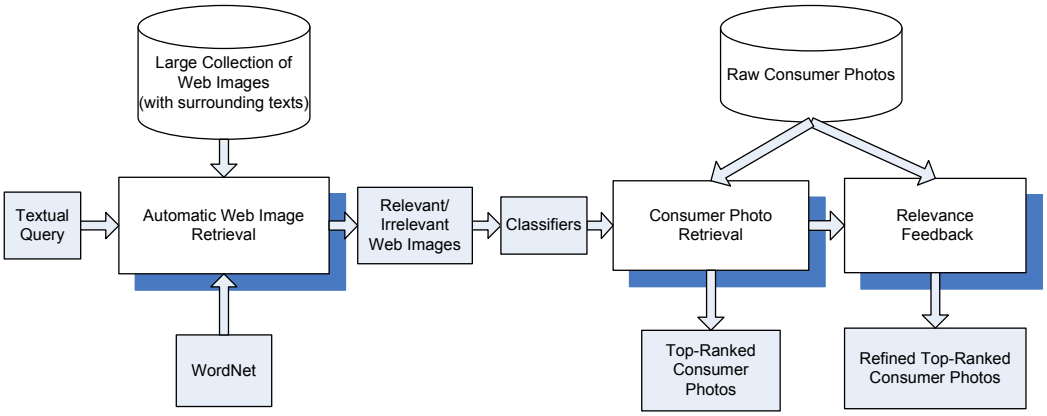


Figure 1: Textual Query Based Consumer Photo Retrieval System.

ing). 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.

4.2 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 daily-life semantic concepts $\mathcal{C}_w$ from *Photosig.com*. Stop-word removal is also used to remove the high-frequency words from $\mathcal{C}_w$ that are not meaningful. Then, we assume that the set of all concepts in a personal collection $\mathcal{C}_p$ is a subset of $\mathcal{C}_w$. In other words, almost all the possible concepts in a personal collection can be expected to be present in the web image database. 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* [11, 25], 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 "pool" 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 $\mathbf{x}_i^w$.

4.3 Consumer Photo Retrieval

As discussed in Section 4.2, 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., non-linear SVM and boosting) may not be feasible for real-time consumer photo retrieval. We therefore choose two simple but effective classifiers k Nearest Neighbors and Decision Stump Ensembles[4].

[4]Boosting using decision stumps has shown state-of-the-art performance in face detection [27], in which the training of boosting classifier is performed in an offline way. Boosting is not suitable for our real-time online image retrieval application because of its high computational cost.

4.3.1 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.

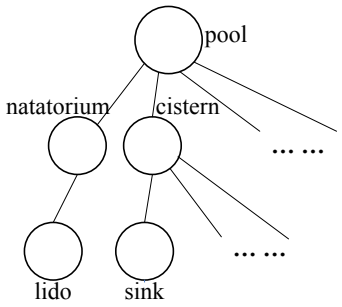


Figure 2: The subtree representing the two-level descendants of "pool" in *WordNet*.

4.3.2 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 can use both the relevant and irrelevant web images in D^w to train decision stump ensembles 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 very unbalanced. To avoid such a highly skewed distribution in the annotated web images, following the method proposed in [23], we randomly sample a fixed number of irrelevant web images as the negative samples, and combine 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., $h(x) = \frac{1-\exp(-x)}{1+\exp(-x)}$. 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.,

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

Here, we use the continuous output of $h(x)$, and the weight γ_d for each stump is set to $0.5 - \epsilon_d$, where ϵ_d is the training error rate of the dth decision stump classifier. 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 100 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[5] [23].

4.4 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 tremendously 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 $n_T = n_l + n_u$.

4.4.1 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 [35, 36, 7, 4, 17, 8, 9]. Yang *et al.* [36] 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:

$$f^T(\mathbf{x}) = f^s(\mathbf{x}) + \Delta f(\mathbf{x}), \quad (2)$$

where the perturbation function $\Delta f(\mathbf{x})$ is learned using the labeled data D_l^T from the target domain. As shown in [36], 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.* [17] 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 [7] proposed the Feature Augmentation method to augment features for domain adaptation. The augmented features are used to construct a kernel function for kernel methods. Note, most cross-domain learning methods [35, 36, 7, 17] 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) [8], and a multiple-source domain adaptation method, Domain Adaptation Machine (DAM) [9]. However, these methods are either variants of SVM or in tandem with SVM or other kernel methods. Therefore, these methods may not be efficient enough for large-scale retrieval applications.

4.4.2 Cross-Domain Combination of Classifiers

To further improve photo retrieval performance, the 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, making such classifier 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 classifier $f^s(\mathbf{x})$ discussed in Section 4.3 can be reused as the auxiliary classifier for relevance feedback. Here, we propose a simple cross-domain learning method, referred to as DS_S+SVM_T. This method simply combines the weighted ensembles of the decision stumps learned from the labeled data in the source domain D^w (referred to as DS_S), and the SVM classifier learned from limited labeled data in the target domain D_l^T (referred to as SVM_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 DS_S and SVM_T are combined with equal weights.

4.4.3 Cross-Domain Regularized Regression

Besides DS_S+SVM_T, we also introduce a new learning method, namely Cross-Domain Regularized Regression (CDRR). In the sequel, we denote the transpose of vector or matrix by the 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 the $f^T(\mathbf{x})$ on the labeled data in the target domain is:

$$\frac{1}{2n_l}\sum_{i=1}^{n_l}(f_i^T - y_i^T)^2 = \frac{1}{2n_l}\|\boldsymbol{f}_l^T - \mathbf{y}_l^T\|^2. \quad (3)$$

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 $f^s(\mathbf{x})$ [9]. We propose a regularization term to enforce 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.*,

$$\frac{1}{2n_u}\sum_{i=n_l+1}^{n_T}(f_i^T - f_i^s)^2 = \frac{1}{2n_u}\|\boldsymbol{f}_u^T - \boldsymbol{f}_u^s\|^2. \quad (4)$$

[5] In [23], the base classifier used in asymmetric bagging is SVM.

We simultaneously minimize the empirical risk of labeled patterns in (3) and the penalty term in (4). The proposed method is then formulated as follows:

$$\min_{f^T} \Omega(f^T) + C\left(\frac{\lambda}{2n_l}\|\boldsymbol{f}_l^T - \mathbf{y}_l^T\|^2 + \frac{1}{2n_u}\|\boldsymbol{f}_u^T - \boldsymbol{f}_u^s\|^2\right), \quad (5)$$

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.

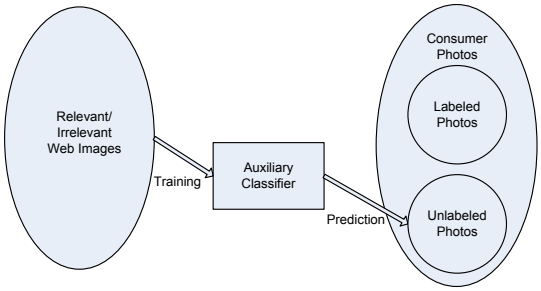


Figure 3: Illustration of Cross-Domain Regularized Regression.

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 $\Omega(f^T) = \frac{1}{2}\|\mathbf{w}\|^2$, then the weight vector $\mathbf{w}$ in the structural risk functional (5) can be solved efficiently by a linear system

$$\left(\mathbf{I} + \frac{C\lambda}{n_l}\mathbf{X}_l\mathbf{X}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{X}_u'\right)\mathbf{w} = \frac{C\lambda}{n_l}\mathbf{X}_l\mathbf{y}_l^T + \frac{C}{n_u}\mathbf{X}_u\mathbf{f}_u^s, \quad (6)$$

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. Thus, we have the closed-form solution $\mathbf{w} = \left(\mathbf{I} + \frac{C\lambda}{n_l}\mathbf{X}_l\mathbf{X}_l' + \frac{C}{n_u}\mathbf{X}_u\mathbf{X}_u'\right)^{-1}\left(\frac{C\lambda}{n_l}\mathbf{X}_l\mathbf{y}_l^T + \frac{C}{n_u}\mathbf{X}_u\mathbf{f}_u^s\right)$.

4.4.4 *Hybrid Method*

Finally, we propose a hybrid method to take the advantages of DS_S+SVM_T and CDRR. After the user marks the consumer photos in each feedback round, we measure the average distance $\bar{d}$ between the labeled positive images and their ρ nearest neighbor consumer photos (ρ is set as 30 in this work). We observe that: when $\bar{d}$ is larger than a threshold ϵ, DS_S+SVM_T is generally better than CDRR; otherwise, CDRR generally outperforms DS_S+SVM_T. We therefore propose a Hybrid approach, in which DS_S+SVM_T or CDRR is chosen as the relevance feedback method based on the value of ϵ.

5. EXPERIMENTS

We evaluate the performance of our proposed framework for textual query based consumer photo retrieval. First, we compare the retrieval performances of the kNN classifier based method and the decision stump classifier based method *without* using relevance feedback. Second, we evaluate the performance of our proposed relevance feedback methods DS_S+SVM_T and CDRR.

5.1 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 [32], we also observed that the images in Photosig generally are high resolution with the sizes varying from 300×200 to 800×600. In addition, the surrounding descriptions more or less 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 an agreement from the Kodak Consumer Video Benchmark Dataset [20], 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 were performed by the students at Columbia University to assign binary labels (presence or absence) for each visual concept. To the best of our knowledge, this dataset is the largest annotated dataset from personal collections. 25 semantic concepts were defined, including 22 visual concepts and three audio-related concepts (*i.e.*,"singing" , "music" and "cheer"). We also combine 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 keyframes from the same video clip are generally near duplicate images, we select only the first keyframe from each video clip in order to fairly compare different algorithms. In total, we test our framework on 21 visual concepts and with 1,358 images.

The second test dataset is the Corel stock photo dataset [30]. We recognized that Corel is not a consumer photo collection, but decided to include it nevertheless because it was used in other studies and also represents a cross-domain case. We use the same subset as in [10], in which 4,999 images (the image sizes are 192×128 or 128×192) are manually annotated in terms of over 370 concepts. Since many concepts have very few images, we only chose 43 concepts that contain at least 100 images.

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. Similarly to [5], 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 three types of global features. Please refer to [5] for more details about the features.

For the training dataset *photosig*, we calculate the original mean value μ_d and standard deviation σ_d for each dimension d, and also normalize all dimensions to zero mean and unit variance. We also normalize two test datasets (*i.e.*, Kodak and Corel) by using μ_d and σ_d.

To improve the speed and reduce the memory cost, we further perform Principal Component Analysis (PCA) using all the images in the photosig dataset. We observe that the first $n_d = 103$ principal components are sufficient to preserve 90% energy. Therefore, all the images in training and test datasets are projected into the 103-D space after dimension reduction.

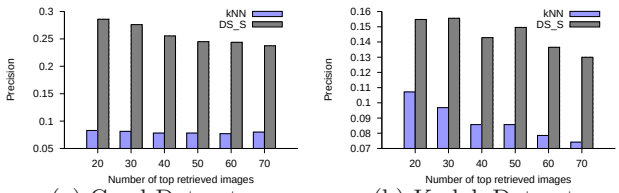
5.2 Retrieval without Relevance Feedback

Considering that the queries by the CBIR methods and our framework are different in nature, we cannot compare our work with the existing CBIR methods before relevance feedback. We also cannot compare the retrieval performance of our framework with web-based annotation methods, because of the following two aspects: 1) These prior works [19, 25, 26, 28, 31, 32] 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 [16, 31, 32] and their annotation performances depend heavily on the correct textual term, making it difficult to fairly compare their methods with our automatic technique. However, we notice that the previous web-based image annotation methods [19, 25, 26, 28, 31, 32] all used kNN classifier for image annotation, possibly owning to its simplicity and effectiveness. Therefore, we directly compare the retrieval performance of decision stumps 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 *Corel* datasets contain 61 distinct concepts in total (concept "beach", "boat" and "people" appear in both datasets). The average number of selected positive samples of all the 61 concepts is 3703.5, and Figure 4 plots the number of positive samples for each concept. For decision stumps, we also randomly choose n_p negative samples with n_s repetitions (n_s is set to 100 in this work), and in total we train $n_s \times n_d = 10300$ decision stumps. The 20% decision stumps with the largest training error rates are removed before computing the weighted ensemble output.

There are 21 concept names from *Kodak* dataset and 43 concept names of *Corel* 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 as in [23]. The average precisions on *Kodak* and *Corel* datasets are shown in Figure 5. We observe that decision stumps based on the training data from the source domain (referred to as DS_S) generally outperform kNN. This is possibly because DS_S employs both positive and negative samples to train the robust classifier while kNN only utilizes the positive samples.

A visual example is shown in Figure 6. We use the keyword "pool" to retrieve images from the Kodak dataset. Note that this query is *undefined* in the concept lexicon of the Kodak dataset. Our retrieval system produces 8 relevant images out of the top 10 retrieved images. One more example for the concept "ruins" on the Corel dataset is also shown in Figure 7(a), in which four correct images are initially retrieved in the top 10 images. In the subsequent subsection, we will show that our proposed Hybrid relevance feedback method can significantly improve the retrieval performance (See Figure 7(b)).



(a) Corel Dataset (b) Kodak Dataset

Figure 5: Retrieval precision using kNN and decision stumps on Corel dataset (4,999 images, 43 concepts) and Kodak dataset (1,358 videos, 21 concepts).

5.3 Retrieval with Relevance Feedback (RF)

In this subsection, we evaluate the performance of relevance feedback methods. For fair comparison, we use DS_S to obtain the initial retrieval results for all the methods except for the baseline kNN based RF method kNN_RF and A-SVM [36], which use kNN and SVM for initial retrieval respectively. For CDRR, we empirically fix $C = 20.0$, and set $\lambda = 0.02$ for the first feedback round and $\lambda = 0.04$ for the remaining rounds. In addition, we also observe 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 top ranked images, namely, these images are not the *extremely* negative images. In our Hybrid method, we empirically fix ρ as 30 and set ϵ as 14.0 and 10.8 for Kodak and Corel datasets, respectively.

We compare our DS_S+SVM_T, CDRR and the Hybrid method with the following methods:
1) **kNN_RF**: The initial retrieval results are obtained by using kNN. In each feedback round, kNN is performed again on the enlarged training set, which includes the labeled positive feedback images marked by the user in the current and all previous rounds, as well as the original n_p positive samples from *photosig* dataset obtained before relevance feedback. The rank of each test image is determined based on the average distance to the top-300 nearest neighbors from the enlarged training set.
2) **SVM_T**: SVM have been used for RF in several existing CBIR methods [23, 24, 37]. We train SVM based on the labeled images in the target domain, which are marked by the user in the current and all previous rounds. We set $C = 1$ and γ in RBF kernel as $\frac{1}{103}$.
3) **A-SVM**: Adaptive SVM (A-SVM) is a recently proposed method [36] for cross-domain learning as described in Section 4.4.1. SVM based on RBF kernel is used to obtain the initial retrieval results. The parameter setting is the same as that in SVM_T.
4) **MR**: Manifold Ranking (MR) is a semi-supervised RF method proposed in [13]. The parameters α and γ for this method are set according to [13].

In real circumstances, the users typically would be reluctant to perform many rounds of relevance feedback or annotate many images for each round. Therefore, we only report

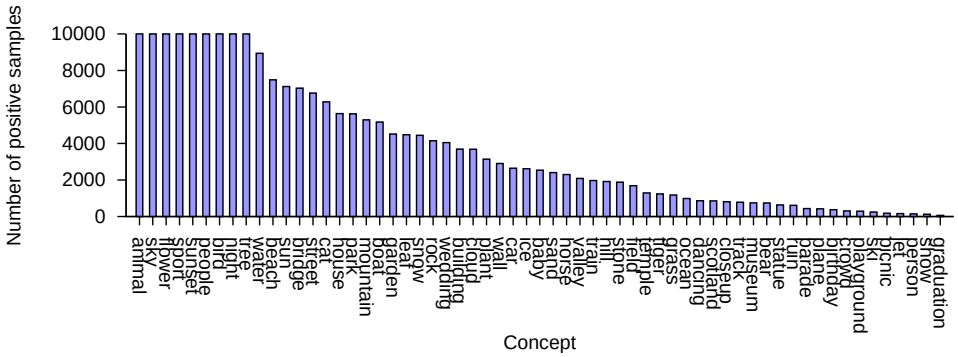


Figure 4: Number of randomly selected positive samples for each concept in the training web image database.

Figure 6: Top-10 initial retrieval results for query 'pool' on Kodak dataset. Incorrect results are highlighted by green boxes.

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 7(b), we show top-10 retrieved images after 4 rounds of relevance feedback for the query "ruins" on the Corel dataset. We observe that the results are improved considerably after using our proposed Hybrid relevance feedback algorithm. Figures 8 and 9 compare different relevance feedback methods on the *Corel* and *Kodak* datasets, respectively.

From these results, we have the following observations:

1) Our CDRR and DS_S+SVM_T outperform the RF methods kNN_RF, SVM_T and MR as well as the existing cross-domain learning method A-SVM in most cases, because they successfully utilize the images from both domains. By taking the advantages of DS_S+SVM_T and CDRR, our Hybrid method generally achieves the best results. When comparing our Hybrid approach with SVM_T after the first round of relevance feedback , the relative improvements are no less than 18.2% and 19.2% on Corel and Kodak datasets, respectively. Moreover, the retrieval performance of our CDRR, DS_S+SVM_T and the Hybrid method 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 RBF kernel function, which will be investigated in the future.

2) The retrieval performances of kNN_RF are almost the same, even after 4 rounds of feedback, possibly because the limited number of user-labeled images in the target domain cannot influence the average distance from the nearest neighbors and kNN's inability to utilize negative feedbacks;

3) For SVM_T, the retrieval performances sometimes drop after the first round of RF, but increase from the second iteration. The explanation is that SVM_T trained based on a limited number of 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.

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 DS_S on the *Kodak* dataset because of the utilization of SVM for initialization. However, as shown in Section 5.4, it takes more than 10 minutes to train an SVM classifier, making it infeasible for the practical image retrieval application.

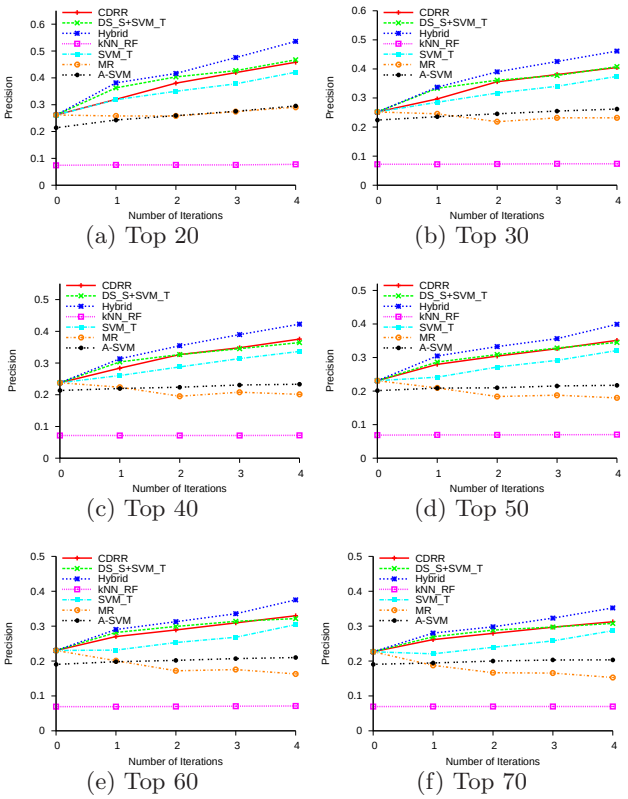
5) Semi-supervised learning method MR can improve the retrieval performance only in some cases on *Kodak* dataset, possibly because the manifold assumption does not hold well for unconstrained consumer images.

5.4 Running Time

We report the average running time of our system for the initial retrieval and RF in Table 1 and Table 2, respectively. The experiments are performed on a server machine with dual Intel Xeon 3.0GHz Quad-Core CPUs (eight threads) and 16GB Memory. Our system is implemented in C++. Matrix and vector operations are performed using the Intel Math Kernel Library 10.0. In this work, each decision stump classifier can be trained and used independently. Therefore, we also use the simple but effective parallelization scheme, OpenMP, to take advantages of multiple threads. In Table 1 and 2, 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.

As shown in Table 1, for our method, the average running time of the initial retrieval for all the concepts is about 8.5 seconds with single thread and 2 seconds with 8 threads. As can be seen from Table 2, the RF process of DS_S+SVM_T and CDRR is very responsive, because we only need to train SVM with less than 10 training samples for DS_S+SVM_T or solve a linear system for CDRR (See Eq. (6)). In practice, DS_S+SVM_T, CDRR and the Hybrid method all take less than 0.1 seconds per round. Therefore, our system is able to achieve real-time retrieval. All the other methods, except for A-SVM, can also achieve real-time retrieval. Similarly as in [36], we train an SVM classifier based on RBF kernel to obtain the initial retrieval result for A-SVM. While the initial retrieval performance of A-SVM is better than DS_S on the

Figure 7: Top-10 retrieval results for query "ruins" on Corel dataset. (a) Initial results; (b) Results after 4 rounds of relevance feedback (one positive and one negative images are labeled in each round). Incorrect results are highlighted by green boxes.

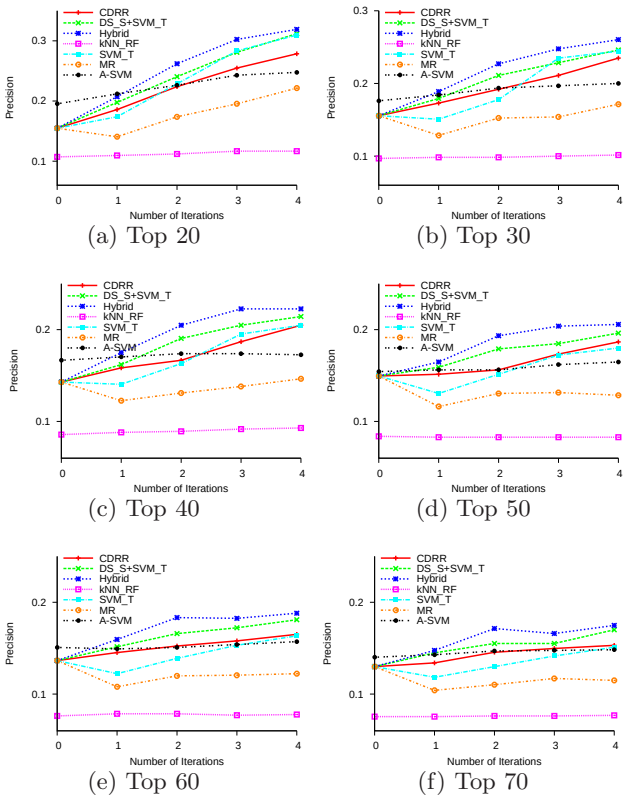


(a) Top 20 (b) Top 30

(c) Top 40 (d) Top 50

(e) Top 60 (f) Top 70

Figure 8: Retrieval results after relevance feedback (one positive and one negative feedbacks per round) on the Corel dataset (4999 images, 43 concepts).



(a) Top 20 (b) Top 30

(c) Top 40 (d) Top 50

(e) Top 60 (f) Top 70

Figure 9: Retrieval results after relevance feedback (one positive and one negative feedbacks per round) on the Kodak dataset (1,358 images, 21 concepts).

Kodak dataset, it takes 610.9 s. In the relevance feedback stage, the target classifier is adapted from the initial SVM classifier. Its speed is also very slow (about 26 seconds per round), making it infeasible for interactive photo retrieval.

6. CONCLUSIONS

By leveraging a large collection of web data (images accompanied by rich textual descriptions) and *WordNet*, we propose a quasi 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 apply two efficient and effective classification methods, kNN and asymmetric bagging with decision stumps for fast consumer photo retrieval.

We also propose three novel relevance feedback methods, namely DS_S+SVM_T, CDRR, and the Hybrid approach by utilizing the pre-learned auxiliary decision stump ensemble classifier and the feedback images to effectively improve the retrieval performance at interactive response time.

Extensive experimental results on *Corel* and *Kodak* photo datasets clearly demonstrate that our Hybrid approach requires an extremely limited amount of feedback from the user and it outperforms other popular relevance feedback methods. Our proposed system can also retrieve consumer photos with a textual query that is not included in the predefined lexicons.

Our current system used the simple decision stump classifier as the source classifier in order to achieve (quasi) real-time response. Some efficient linear SVM implementations (e.g., LIBLINEAR) may be also used in our system. In addition, non-linear functions may be also employed in CDRR to

Method	DS_S		kNN	
# Threads	1	8	1	8
Time (in Sec.)	8.528	2.042	3.265	0.913

Table 1: Average CPU time of initial retrieval.

Method	DS_S+SVM_T	CDRR	Hybrid
Time	0.056	0.052	0.097
Method	MR	SVM_T	A-SVM
Time	0.051	0.054	26.179

Table 2: Average CPU time (in Sec.) of relevance feedback (per round) with one single thread.

further improve the performance of our system. The above issues will be investigated in the future.

Acknowledgment

This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF2008IDM-IDM004-018. The authors also thank Yi Yang for his helpful discussions and suggestions.

7. REFERENCES

[1] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

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

[3] S.-F. Chang et al. Large-scale multimodal semantic concept detection for consumer video. In *ACM SIGMM Workshop on MIR*, 2007.

[4] S.-F. Chang et al. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *NIST TRECVID Workshop*, 2008.

[5] T.-S. Chua et al. NUS-WIDE: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[6] R. Datta et al. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*,1–60, 2008.

[7] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[8] L. Duan et al. Domain Transfer SVM for Video Concept Detection. In *CVPR*, 2009.

[9] L. Duan et al. Domain Adaptation from Multiple Sources via Auxiliary Classifiers. In *ICML*, 2009.

[10] P. Duygulu et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

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

[12] R. Fergus, P. Perona,and A. Zisserman. A Visual Category Filter for Google Images. In *ECCV*, 2004.

[13] J. He et al. Manifold-ranking based image retrieval. In *ACM MM*, 2004.

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

[15] S. Hoi et al. Semi-supervised svm batch mode active learning for image retrieval. In *CVPR*, 2008.

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

[17] W. Jiang et al. Cross-domain learning methods for high-level visual concept classification. In *ICIP*, 2008.

[18] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *T-PAMI*, 985–1002, 2008.

[19] X. Li et al. Image annotation by large-scale content-based image retrieval. In *ACM MM*, 2006.

[20] A. Loui et al. Kodak's consumer video benchmark data set: concept definition and annotation. In *ACM Workshop on MIR*, 2007.

[21] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *ICIP*, 1997.

[22] A. Smeulders et al. Content-based image retrieval at the end of the early years. *T-PAMI*,1349–1380, 2000.

[23] D. Tao et al. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *T-PAMI*, 1088–1099, 2006.

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

[25] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *T-PAMI*, 1958–1970, 2008.

[26] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *CVPR*, 2008.

[27] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 137–154, 2004.

[28] C. Wang et al. Content-based image annotation refinement. In *CVPR*, 2007.

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

[30] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *T-PAMI*,947–963, 2001.

[31] X. Wang et al. AnnoSearch: Image auto-annotation by search. In *CVPR*, 2006.

[32] X. Wang et al. Annotating images by mining image search results. *T-PAMI*, 1919–1932, 2008.

[33] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

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

[35] P. Wu and T. G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *ICML*, 2004.

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

[37] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *ICIP*, 2001.

[38] X. Zhou and T. Huang. Small sample learning during multimedia retrieval using biasmap. In *CVPR*, 2001.