Unsupervised Auxiliary Visual Words Discovery for Large-Scale Image Object Retrieval

Yin-Hsi Kuo[1,2], Hsuan-Tien Lin[1], Wen-Huang Cheng[2], Yi-Hsuan Yang[1], and Winston H. Hsu[1]
[1]National Taiwan University and [2]Academia Sinica, Taipei, Taiwan

Abstract

Image object retrieval – locating image occurrences of specific objects in large-scale image collections – is essential for manipulating the sheer amount of photos. Current solutions, mostly based on bags-of-words model, suffer from low recall rate and do not resist noises caused by the changes in lighting, viewpoints, and even occlusions. We propose to augment each image with auxiliary visual words (AVWs), semantically relevant to the search targets. The AVWs are automatically discovered by feature propagation and selection in textual and visual image graphs in an unsupervised manner. We investigate variant optimization methods for effectiveness and scalability in large-scale image collections. Experimenting in the large-scale consumer photos, we found that the the proposed method significantly improves the traditional bag-of-words (111% relatively). Meanwhile, the selection process can also notably reduce the number of features (to 1.4%) and can further facilitate indexing in large-scale image object retrieval.

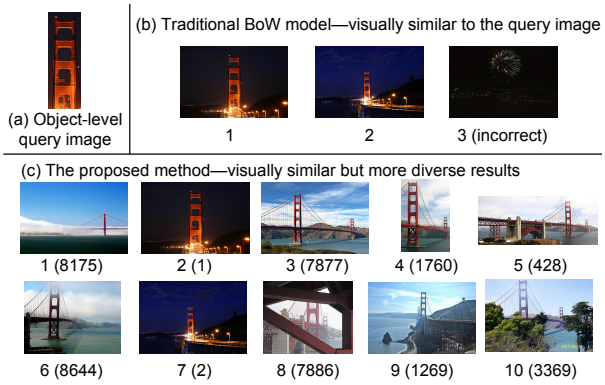


Figure 1. Comparison in the retrieval performance of the traditional BoW model [14] and the proposed approach. (a) An example of object-level query image. (b) The retrieval results of a BoW model, which generally suffers from the low recall rate. (c) The results of the proposed system, which obtains more accurate and diverse results. Note that the number below each image is its rank in the retrieval results and the number in a parenthesis represents the rank predicted by the BoW model.

1. Introduction

Image object retrieval – retrieving images (partially) containing the target image object – is one of the key techniques of managing the exponentially growing image/video collections. It is a challenging problem because the target object may cover only a small region in the database images as shown in Figure 1. Lots of promising applications such as annotation by search [17, 18], geographical information estimation [7], etc., are keen to the accuracy and efficiency of image object retrieval.

Bag-of-words (BoW) model is popular and shown effective for image object retrieval [14]. BoW representation quantizes high-dimensional local features into discrete visual words (VWs). However, traditional BoW-like methods fail to address issues related to noisily quantized visual features and vast variations in viewpoints, lighting conditions, occlusions, etc., commonly observed in large-scale image collections [12, 21]. Thus, it suffers from low recall rate as shown in Figure 1(b).

Due to variant capture conditions and large VW vocabulary (e.g., 1 million vocabulary), the features for the target image objects might have different VWs (cf. Figure 1(c)). Besides, it is also difficult to obtain these VWs through query expansion (e.g., [1]) or even varying quantization methods (e.g., [12]) because of the large differences in visual appearance between the query and the target objects.

We observe the sparseness for the visual words in BoW model to cover the whole search targets and the lack of semantic related features from these visual features only, as discussed in Section 3. In this work, we argue to augment each image in the image collections with *auxiliary visual words (AVW)*—additional visual features semantically relevant to the search targets (cf. Figure 1(c)). Targeting at large-scale image collections for serving different queries, we mine the auxiliary visual words in an unsupervised manner by incorporating bother visual and (noisy) textual information. We construct graphs of images by visual and textual information (if available) respectively. We then automatically propagate the semantic and select the informa-

tive AVWs across the visual and textual graphs since these two modalities can boost each other (cf. Figure 3). The two processes are formulated as optimization formulations iteratively through the subtopics in the image collections. Meanwhile, we also consider the scalability issues by leveraging distributed computation framework (e.g., MapReduce).

Experiments show that the proposed method greatly improves the recall rate for image object retrieval. Specifically, the unsupervised auxiliary visual words discovery greatly outperforms BoW models (by 111% relatively) and complementary to conventional pseudo-relevance feedback. Meanwhile, AVW discovery can also derive very compact (i.e., 1.4% of the original features) and informative feature representations which will benefit the indexing structure [14].

The primary contributions of the paper include,

- Observing the problems in large-scale image object retrieval by conventional BoW model (Section 3).
- Proposing auxiliary visual words discovery through visual and textual clusters in an unsupervised and scalable fashion (Section 4).
- Investigating variant optimization methods for efficiency and accuracy in AVW discovery (Section 5).
- Conducting experiments on consumer photos and showing great improvement of recall rate for image object retrieval (Section 7).

2. Related Work

Most image object retrieval systems adopt the scale-invariant feature transform (SIFT) descriptor [8] to capture local information and adopt bag-of-words (BoW) model [14] to conduct object matching [1, 11]. The SIFT descriptors are quantized to visual words (VWs), such that indexing techniques well developed in the text domain can be directly applied.

The learned VW vocabulary will directly affect the image object retrieval performance. The traditional BoW model adopts k-means clustering to generate the vocabulary. A few attempts try to impose extra information for visual word generation such as visual constraints [13], textual information [19]. However, it usually needs extra (manual) information during the supervised learning, which might be formidable in large-scale image collections.

Instead of generating new VW vocabulary, some researches work on the original VW vocabulary such as [15]. It suggested to select useful feature from the neighboring images to enrich the feature description. However, its performance is limited for large-scale problems because of the need to perform spatial verification, which is computationally expensive. Moreover, it only considers neighboring images in the visual graph, which provides very limited semantic information. Other selection methods for the useful features such as [6] and [10] are based on different criteria—the number of inliers after spatial verification, and pairwise constraints for each image, thus suffer from limited scalability and accuracy.

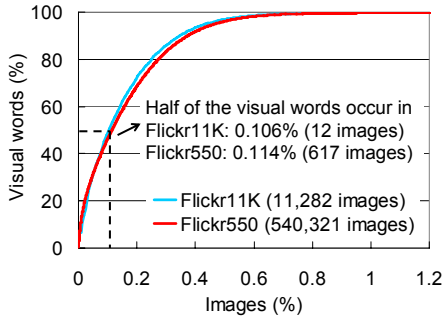


Figure 2. Cumulative distribution of the frequency of VW occurrence in two different image databases, cf. Section 3.1. It shows that half of the VWs occur in less than 0.11% of the database images (i.e. 12 and 617 images, respectively). The statistics represent that VWs are distributed over the database images very sparsely.

Authors in [9] consider both visual and textual information and adopt unsupervised learning methods. However, they only use global features and adopt random-walk-like process for post-processing in image retrieval. Similar limitations are observed in [16], where only the similarity image scores are propagated between textual and visual graphs. Different from the prior works, we use local features for image object retrieval and propagate the VWs directly between the textual and visual graphs. The discovered auxiliary VWs are thus readily effective in retrieving diverse search results, eliminating the need to apply a random walk in the graphs again.

To augment images with their informative features, we propose auxiliary visual words discovery, which can efficiently propagate semantically relevant VWs and select representative visual features by exploiting both textual and visual graphs. The discovered auxiliary visual words demonstrate significant improvement over the BoW baseline and are shown orthogonal and complementary to conventional pseudo-relevance feedback. Besides, when dataset size becomes larger, we can apply all the processes in a parallel way (e.g., MapReduce).

3. Key Observations—Requiring Auxiliary Visual Words

Nowadays, bag-of-words (BoW) representation [14] is widely used in image object retrieval and has been shown promising in several content-based image retrieval (CBIR) tasks (e.g., [11]). However, most existing systems simply apply the BoW model without carefully considering the sparse effect of the VW space [2], as detailed in Section 3.1. Another observation (explained in Section 3.2) is that

VWs are merely for describing visual appearances and lack the semantic descriptions for retrieving more diverse results (cf. Figure 1(b)). The proposed AVW discovery method is targeted to address these issues.

3.1. Sparseness of the Visual Words

For better retrieval accuracy, most systems will adopt 1 million VWs for their image object retrieval system as suggested in [11]. However, our statistics shows that the occurrence of VWs in different images is very sparse. We calculate it on two image databases of different sizes, i.e. Flickr550 and Flickr11K (cf. Section 6.1), and obtain similar curves as shown in Figure 2. We can find that half of the VWs only occur in less than 0.11% of the database images and most of the VWs (i.e. around 96%) occur in less than the 0.5% ones (i.e. 57 and 2702 images, respectively). That is to say, two images sharing one specific VW seldom contain similar features. In other words, those similar images might only have few common VWs. This phenomenon is known as the uniqueness of VWs [2]. It is partly due to some quantization errors or noisy features. Therefore, in Section 4, we propose to augment each image with auxiliary visual words.

3.2. Lacking Semantics Related Features

Since VWs are merely low-level visual features, it is very difficult to retrieve object images with different viewing angles, lighting conditions, partial occlusions, etc. An example is shown in Figure 3. By using BoW models, the query image (the top-left one) can easily obtain visually similar results (e.g., the bottom-left one) but often fails to retrieve the ones in a different viewing angle (e.g. the right-hand side image). This problem can be alleviated by taking benefit from the textual semantics. That is, by using the textual information associated with images, we are able to obtain semantically similar images as shown in the red dotted rectangle in Figure 3. If those semantically similar images can share (propagate) their VWs to each other, the query image can still obtain similar but more visually and semantically diverse results.

4. Auxiliary Visual Words Discovery

Based on the previous observations, it is necessary to propagate VWs to those visually or semantically similar images. Consequently, we propose an offline stage for unsupervised auxiliary visual words discovery. We augment each image with auxiliary visual words (features) by considering semantically related VWs in its textual cluster and representative VWs in its visual cluster. When facing large-scale dataset, we can deploy all the processes in a parallel way (e.g., MapReduce [3]). Besides, AVW reduces the number of VWs to be indexed (i.e., better efficiency in time and memory). Such AVW might potentially benefit the further image queries and can greatly improve the recall rate as demonstrated in Section 7.1 and in Figure 7. For mining AVWs, we first generate image graphs and image clusters in Section 4.1. Then based on the image clusters, we propagate auxiliary VWs in Section 4.2 and select representative VWs in Section 4.3. Finally, we combine both selection and propagation methods in Section 4.4.

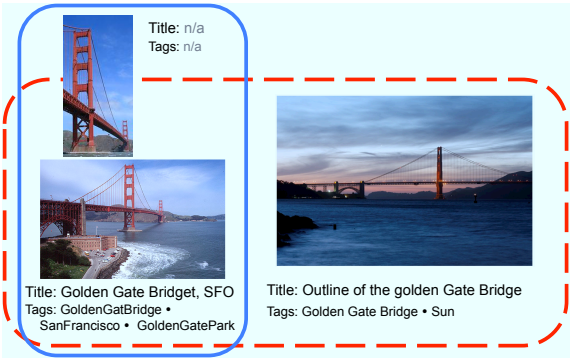
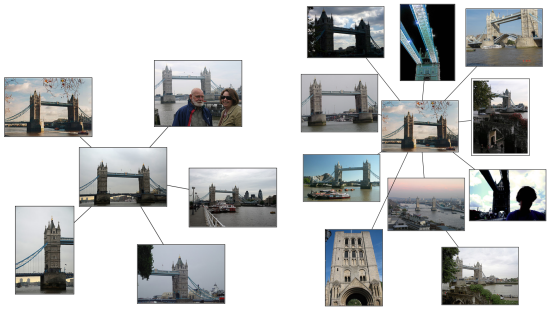


Figure 3. Illustration of the roles of semantic related features in the image object retrieval. Images in the blue rectangle are visually similar, whereas those images in the red dotted rectangle are textually similar. The semantic (textual) features are promising to establish the in-between connection (Section 4) to help the query image (the top-left one) retrieve the right-hand-side image.

4.1. Graph Construction and Image Clustering

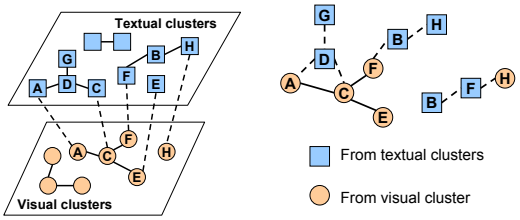
The image clustering is first based on a graph construction. The images are represented by 1M VWs and 90K text tokens expanded by Google snippets from their associated (noisy) tags. We take the advantage of the sparseness and use cosine measure as the similarity measure. The measure is essentially an inner product of two feature vectors and only the non-zero dimensions will affect the similarity value—i.e., skipping the dimensions that either feature has a zero value. We observe that the textual and visual features are sparse for each image and the correlation between images are sparse as well. We adopt efficient algorithms (e.g., [4]) to construct the large-scale image graph by MapReduce. To cluster images on the image graph, we apply affinity propagation (AP) [5] for graph-based clustering. AP passes and updates messages among nodes on graph iteratively and locally—associating with the sparse neighbors only. AP's advantages include automatic determining the number of clusters, automatic exemplar (canonical image) detection within each cluster.

The image clustering results are sampled in Figure 4. Note that if an image is close to the canonical image (center image), it has a higher AP score, indicating that it is more strongly associated with the cluster.



(a) A visual cluster sample. (b) A textual cluster sample.

Figure 4. Sample image clusters (cf. Section 4.1). The visual cluster groups visually similar images in the same cluster, whereas the textual cluster favors semantic similarities. The two clusters facilitate representative VWs selection and semantic (auxiliary) VWs propagation, respectively.



(a) Visual and textual graphs. (b) Two extended visual clusters from the (left) visual and textual clusters.

Figure 5. Illustration of the propagation operation. Based on visual and textual graphs in (a), we can propagate auxiliary VWs among the associated images in the extended visual clusters. (b) shows the two extended visual clusters as the units for propagation respectively; each extended visual cluster include the visually similar images and those co-occurrences in other textual clusters.

4.2. Semantic Visual Features Propagation

Seeing the limitations in BoW model, we propose to augment each image with additional VWs propagated from the visual and textual clusters (Figure 5(a)). Propagating the VWs from both visual and textual domains can enrich the visual descriptions of the images and be beneficial for further image object queries. For example, it is promising to derive more semantic VWs by simply exchanging the VWs among (visually diverse but semantically consistent) images of the same textual cluster (cf. Figure 4(b)) .

We actually conduct the propagation on each *extend visual cluster*, containing the images in a visual cluster and those additional ones co-occuring with these images in certain textual clusters. The intuition is to balance visual and semantic consistence for further VW propagation and selection (cf. Section 4.3). Figure 5(b) shows two extended visual clusters derived from Figure 5(a). More interestingly, image E is singular in textual cluster due to having no tags; however, E still belongs to a visual cluster and can still receive AVWs in its associated extended visual cluster. Similarly, if there is a single image in a visual cluster such as image H, it can also obtain auxiliary VWs (i.e., from image B and F) in the extended visual cluster.

Assuming matrix $X \in \mathbb{R}^{N\times D}$ represents the N image histograms in the extended visual cluster and each image has D (i.e., 1 million) dimensions. And X_i stands for the VW histogram of image i. Assume M among N are from the same visual cluster; for example, $N = 8$ and $M = 4$ in the left extended visual cluster in Figure 5(b). The visual propagation is conducted by the propagation matrix $P \in \mathbb{R}^{M\times N}$, which controls the contributions from different images in the extended visual cluster[1]. $P(i, j) \in [0, 1]$ weights the whole features propagated from image j to i. If we multiply the propagation matrix P and X (PX), we can obtain a new $M \times D$ VW histograms, as the AVWs, for the M images augmented by the N images.

For each extended visual cluster, we desire to find a better propagation matrix P, given the initial propagation matrix P_0 (i.e., $P_0(i, j) = 1$, if both i and j are semantically related and within the same textual cluster). We propose to formulate the propagation operation as

$$f_P = \min_P \alpha \frac{\|PX\|_F^2}{N_{P1}} + (1-\alpha)\frac{\|P - P_0\|_F^2}{N_{P2}}, \qquad (1)$$

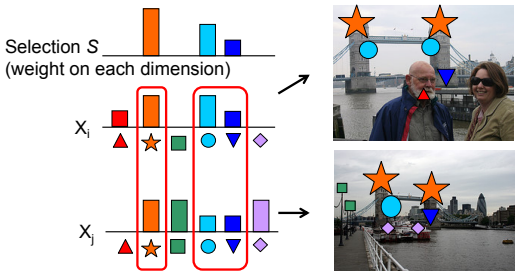
The goal of the first term is to avoid from propagating too many VWs since PX becomes new VW histogram matrix after the propagation. And the second term is to keep the similarity to the original propagation matrix (i.e., similar in textual cluster). $N_{P1} = \|P_0X\|_F^2$ and $N_{P2} = \|P_0\|_F^2$ are two normalization terms and α modulates the importance between the first and the second terms. We will investigate the effects of α in Section 7.2. Note that the propagation process updates the propagation matrix P on each extended visual cluster separately as shown in Figure 5(b); therefore, this method is scalable for large-scale dataset and easy to adopt in a parallel way.

4.3. Common Visual Words Selection

Though the propagation operation is important to obtain different VWs, it may include too many VWs and thus decrease the precision. To mitigate this effect and remove those irrelevant or noisy VWs, we propose to select those representative VWs in each visual cluster. We observe that images in the same visual cluster are visually similar to each other (cf. Figure 4(a)); therefore, the selection operation is to retain those representative VWs in each visual cluster.

As shown in Figure 6(a), X_i (X_j) represents VW histogram of image i (j) and selection S indicates the weight

[1]Note that here we first measure the images from the same visual cluster only. However, by propagating through each extended visual clusters, we can derive the AVWs for each image.



(a) Common VWs selection. (b) Two examples.

Figure 6. Illustration of the selection operation. The VWs should be similar in the same visual cluster; therefore, we select those representative visual features (red rectangle). (b) illustrates the importance (or representativeness) for different VWs. And we can further remove some noisy features (less representative) which appeared on the people or boat.

on each dimension. So XS indicates the total number of features retained after the selection. The goal of selection is to keep those common VWs in the same visual cluster (cf. Figure 6(b)). That is to say, if S emphasizes more on those common (representative) VWs, the XS will be relatively large. Then the selection operation can be formulated as

$$f_S = \min_S \beta \frac{\|XS_0 - XS\|_F^2}{N_{S1}} + (1-\beta)\frac{\|S\|_F^2}{N_{S2}}. \quad (2)$$

The second term is to reduce the number of selected features in the visual clusters. The selection is expected to be compact but should not incur too many distortions from the original features in the visual clusters and thus regularized in the first term, showing the difference of feature numbers before (S_0) and after (S) the selection process. Note that S_0 will be assigned by one which means we select all the dimensions. $N_{S1} = \|XS_0\|_F^2$ and $N_{S2} = \|S_0\|_F^2$ are the normalization terms and β stands for the influence between the first and the second terms and will be investigated in Section 7.2.

4.4. Iteration of Propagation and Selection

The propagation and selection operations described above can be performed iteratively. The propagation operation obtains semantically relevant VWs to improve the recall rate, whereas the selection operation removes visually irrelevant VWs and improves memory usage and efficiency. An empirical combination of propagation and selection methods is reported in Section 7.1.

5. Optimization

In this section, we study the solvers for the two formulations above (Eq. (1) and (2)). Before we start, note that the two formulations are very similar. In particular, let $\tilde{S} = S - S_0$, the selection formulation (2) is equivalent to

$$\min_{\tilde{S}} \quad \beta \frac{\|X\tilde{S}\|_F^2}{N_{S1}} + (1-\beta)\frac{\|\tilde{S} + S_0\|_F^2}{N_{S2}}. \quad (3)$$

Given the similarity between Eq. (1) and (3), we can focus on solving the former and then applying the same technique on the latter.

5.1. Convexity of the Formulations

We shall start by computing the gradient and the Hessian of Eq. (1) with respect to the propagation matrix P. Consider the M by N matrices P and P_0. We can first stack the columns of the matrices to form two vectors $p = \text{vec}(P)$ and $p_0 = \text{vec}(P_0)$, each of length MN. Then, we replace $\text{vec}(PX)$ with $(X^T \otimes I_M)p$, where I_M is an identity matrix of size M and $\otimes$ is the Kronecker product. Let $\alpha_1 = \frac{\alpha}{N_{P1}} > 0$ and $\alpha_2 = \frac{1-\alpha}{N_{P2}} > 0$, the objective function of Eq. (1) becomes

$$\begin{aligned} & f(p) \\ =& \alpha_1 \|(X^T \otimes I_M)p\|_2^2 + \alpha_2 \|p - p_0\|_2^2 \\ =& \alpha_1 p^T (X \otimes I_M)(X^T \otimes I_M)p + \alpha_2 (p-p_0)^T (p-p_0) \end{aligned}$$

Thus, the gradient and the Hessian are

$$\nabla_p f(p) = 2\big(\alpha_1 (X \otimes I_M)(X^T \otimes I_M)p + \alpha_2 (p - p_0)\big) \quad (4)$$

$$\nabla_p^2 f(p) = 2\big(\alpha_1 (X \otimes I_M)(X^T \otimes I_M) + \alpha_2 I_{MN}\big). \quad (5)$$

Note that the Hessian (Eq. (5)) is a constant matrix. The first term of the Hessian is positive semi-definite, and the second term is positive definite because $\alpha_2 > 0$. Thus, Eq. (1) is strictly convex and enjoys an unique optimal solution.

From the analysis above, we see that Eq. (1) and (2) are strictly convex, unconstrained quadratic programming problems. Thus, any quadratic programming solver can be used to find their optimal solutions. Next, we study two specific solvers: the gradient descent solver which iteratively updates p and can easily scale up to large problems; the analytic one which obtains the optimal p by solving a linear equation and reveals a connection with the Tikhonov regularization technique in statistics and machine learning.

5.2. Gradient Descent Solver (GDS)

The gradient descent solver optimizes Eq. (1) by starting from an arbitrary vector p^{start} and iteratively updates the vector by

$$p^{new} \leftarrow p^{old} - \eta \nabla_p f(p^{old}),$$

where a small $\eta > 0$ is called the learning rate. We can then use Eq. (4) to compute the gradient for the updates. Nevertheless, computing $(X \otimes I_M)(X^T \otimes I_M)$ may be unnecessarily time- and memory-consuming. We can re-arrange the matrices and get

$$(X \otimes I_M)(X^T \otimes I_M)p = (X \otimes I_M)\text{vec}(PX) = \text{vec}(PXX^T)$$

Then,

$$\begin{aligned}\nabla_p f(p) &= 2\alpha_1 \text{vec}(PXX^T) + 2\alpha_2 \text{vec}(P - P_0) \\ &= \text{vec}(2\alpha_1 PXX^T + 2\alpha_2(P - P_0)).\end{aligned}$$

That is, we can update p^{old} as a matrix P^{old} with the gradient also represented in its matrix form. Coupling the update scheme with an adaptive learning rate η, we get update propagation matrix by

$$P^{new} = P^{old} - 2\eta\big(\alpha_1 P^{old} XX^T + \alpha_2(P^{old} - P_0)\big). \quad (6)$$

Note that we simply initialize p^{start} to $\text{vec}(P_0)$.

For the selection formulation (Section 4.3), we can adopt similar steps with two changes. First, Eq. (6) is replaced with

$$S = S - 2\eta\left(\beta\frac{-X^T X(S_0 - S)}{N_{S1}} + (1-\beta)\frac{S}{N_{S2}}\right). \quad (7)$$

Second, the initial point S^{start} is set to a zero matrix since the goal of selection formulation is to select representative visual words (i.e., retain a few dimensions).

There is one potential caveat of directly using Eq. (7) for updating. The matrix $X^T X$ can be huge (e.g., $1M \times 1M$). To speed up the computation, we could keep only the dimensions that occured in the same visual cluster, because the other dimensions would contribute 0 to $X^T X$.

5.3. Analytic Solver (AS)

Next, we compute the unique optimal solution p^* of Eq. (1) analytically. The optimal solution must satisfy $\nabla_p f(p^*) = 0$. Note that From Eq. (4),

$$\nabla_p f(p^*) = Hp^* - 2\alpha_2 p_0,$$

where H is the constant and positive definite Hessian matrix. Thus,

$$p^* = 2\alpha_2 H^{-1} p_0.$$

Similar to the derivation in the gradient descent solver, we can write down the matrix form of the solution, which is

$$P^* = \alpha_2 P_0(\alpha_1 XX^T + \alpha_2 I_M)^{-1}.$$

For the selection formulation, a direct solution from the steps above would lead to

$$S^* = \beta(\beta\frac{X^T X}{N_{S1}} + (1-\beta)\frac{I_D}{N_{S2}})^{-1}\frac{X^T X S_0}{N_{S1}}. \quad (8)$$

Nevertheless, as mentioned in the previous subsection, the $X^T X$ matrix in Eq. (8) can be huge (e.g., $1M \times 1M$). It is a time-consuming task to compute the inverse of an $1M \times 1M$ matrix. Thus, instead of calculating $X^T X$ directly, we transform $X^T X$ to XX^T which is N by N and is much smaller (e.g., 100×100). The transformation is based on the identity of the inverse function

$$(A + BB^T)^{-1}B = A^{-1}B(I + B^T A^{-1}B)^{-1}.$$

Then, we can re-write Eq. (8) as

$$S^* = \beta X^T(\beta\frac{XX^T}{N_{S1}} + (1-\beta)\frac{I_M}{N_{S2}})^{-1}\frac{XS_0}{N_{S1}}. \quad (9)$$

Note that the analytic solutions of Eq. (1) and (2) are of a similar form to the solutions of ridge regression (Tikhonov regularization) in statistics and machine learning. The fact is of no coincidence. Generally speaking, we are seeking to obtain some parameters (P and S) from some data (X, P_0 and S_0) while regularizing by the norm of the parameters. The use of the regularization not only ensures the strict convexity of the optimization problem, but also eases the hazard of overfitting with a suitable choice of α and β.

6. Experimental Setup

6.1. Dataset

We use Flickr550 [20] as our main dataset in the experiments. To evaluate the proposed approach, we select 56 query images (1282 ground truth images) which belong to the following 7 query categories: Colosseum, Eiffel Tower, Golden Gate Bridge, Tower de Pisa, Starbucks logo, Tower Bridge, and Arc de Triomphe. Also, we randomly pick up 10,000 images from Flickr550 to form a smaller subset called Flickr11K. Some query examples are shown in Figure 7.

6.2. Performance Metrics

In the experiments, we use the average precision, a performance metric commonly used in the previous work [11, 20], to evaluate the retrieval accuracy. It approximates the area under a non-interpolated precision-recall curve for a query. A higher average precision indicates better retrieval accuracy. Since average precision only shows the performance for a single image query, we also compute the mean average precision (MAP) over all the queries to evaluate the overall system performance.

6.3. Evaluation Protocols

As suggested by the previous work [11], our image object retrieval system adopts 1 million visual words as the basic vocabulary. The retrieval is then conducted by comparing (indexing) the AVW features for each database image. To further improve the recall rate of retrieval results, we apply the query expansion technique of pseudo-relevance feedback (PRF) [1], which expands the image query set by taking the top-ranked results as the new query images. This step also helps us understand the impacts of the discovered AVWs because in our system the ranking of retrieved

Table 1. The MAP of AVW results with the best iteration number and PRF in Flickr11K with totally 22M (SIFT) feature points. Note that the MAP of the baseline BoW model [14] is 0.245 and after PRF is 0.297 (+21.2%). #F represents the total number of features retained; M is short for million. '%' indicates the relative MAP gain over the BoW baseline.

Solver	Propagation → Selection			Selection → Propagation		
	MAP	MAP by PRF (%)	#F	MAP	MAP by PRF (%)	#F
Gradient descent solver (GD)	0.375	0.516 (+110.6%)	0.3M	0.342	0.497 (+102.9%)	0.2M
Analytic solver (AS)	0.384	0.483 (+97.1%)	5.2M	0.377	0.460 (+87.8%)	13.0M

Figure 7. More search results by auxiliary VWs. The number represents its retrieval ranking. The results show that the proposed AVW method, thought conducted in an unsupervised manner in the image collections, can retrieve more diverse and semantic related results.

images is related to the associated auxiliary visual words. They are the key for our system to retrieve more diverse and accurate images as shown in Figure 7 and Section 7.1. We take L1 distance as our baseline for BoW model [14]. The MAP for the baseline is 0.245 with 22M (millions) feature points and the MAP after PRF is 0.297 (+21.2%).

7. Results and Discussions

7.1. The Performance of Auxiliary Visual Words

The overall retrieval accuracy is listed in Table 1. As mentioned in Section 4.4, we can iteratively update the features according to Eq. (1) and (2). It shows that the iteration with propagation first (propagation → selection) can have the best results. Since the first propagation will share all the VWs with related images and then the selection will choose those common VWs as representative VWs. However, if we do the iteration with selection first (i.e., selection → propagation), we might lose some possible VWs after the first selection. Experimental results show that we only need one or two iterations to achieve better result because those informative and representative VWs have been propagated or selected in the early iteration steps. Besides, the number of features are significantly reduced from 22.2M to 0.3M (only 1.4% retained), essential for indexing those features by inverted file structure [14]. The required memory size for indexing is proportional to the number of features.

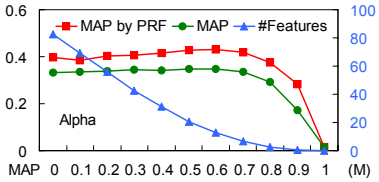
In order to have the timely solution by gradient descent solver, we set a loose convergence criteria for both propagation and selection operations. Therefore, the solution of the two solvers might be different. Nevertheless, Table 1 still shows that the retrieval accuracy of the two solvers are very similar. The learning time for the first propagation is 2720s (GD) and 123s (AS), whereas the first selection needs 1468s and 895s for GD and AS respectively. Here we fixed $\alpha = 0.5$ and $\beta = 0.5$ to evaluate the learning time.[2] By using analytic solver, we can get a direct solution and much faster than the gradient descent method. Note that the number of features will affect the running time directly; therefore, in the remaining iteration steps, the time required will decrease further since the number of features is greatly reduced iteratviely. Meanwhile, only a very small portion of visual features retained.

Besides, we find that the proposed AVW method is complementary to PRF since we yield another significant improvement after conducting PRF on the AVW retrieval results. For example, the MAP of AVW is 0.375 and we can have 0.516 (+37.6%) after applying PRF. The relative improvement is even much higher than PRF over the traditional BoW model (i.e., 0.245 to 0.297, +21.2%). More retrieval results by AVW + PRF are illustrated in Figure 7, which shows that the proposed AVW method can even retrieve semantically consistent but visually diverse images. Note that the AVW is conducted in an unsupervised manner in the image collections and requires no manual labels.

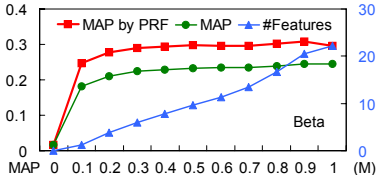
7.2. Parameter Sensitivity

Finally, we report the impact of sensitive tests on two important parameters—propagation formulation (α) and selection formulation (β). The results are shown in Figure 8. In the propagation formulation, α decides the number of features needed to be propagated. Figure 8(a) shows that if we propagate all the possible features to each image (i.e., $\alpha = 0$), we will obtain too many irrelevant and noisy features which is helpless for the retrieval accuracy. Besides,

[2]The learning time is evaluated in MATLAB at a regular Linux server with Intel CPU and 16G RAM.



(a) Alpha (α, for propagation in Eq. (1))



(b) Beta (β, for selection in Eq. (2))

Figure 8. Parameter sensitivity on alpha and beta. (a) shows that propagating too many features is not helpful for the retrieval accuracy. (b) shows that only partial features are important (representative) to each image. More details are discussed in Section 7.2. Note that we can further improve retrieval accuracy by iteratively updated AVW by propagation and selection processes.

the curve drops fast after $\alpha \geq 0.8$ because it preserved few VWs which might not appear in the query images. The figure also shows that if we set α around 0.6 we can have better result but with fewer features which are essential for large-scale indexing problem.

And for selection formulation, similar to α, β also influences the number of dimensions needed to be retained. For example, if $\beta = 0$, we will not select any dimensions for each image. And $\beta = 1$ means we will retain all the features, and the result is equal to the BoW baseline. Figure 8(b) shows that if we just keep a few dimensions of VWs, the MAP is still similar to BoW baseline though with some retrieval accuracy decrease. Because of the spareness of large VW vocabulary as mentioned in Section 3.1, we only need to keep those important VWs.

8. Conclusions and Future Work

In this work, we show the problems of current BoW model and the needs for semantic visual words to improve the recall rate for image object retrieval. We propose to augment each database image with semantically related auxiliary visual words by propagating and selecting those informative and representative VWs in visual and textual clusters. Note that we formulate the processes as unsupervised optimization problems. Experimental results show that we can greatly improve the retrieval accuracy compared to the BoW model (111% relatively). In the future, we will further look into the problem by L2-loss L1-norm optimization which might preserve the sparseness for visual words. We will also investigate different solvers to maximize the retrieval accuracy and efficiency.

References

[1] O. Chum et al. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE ICCV*, 2007. 1, 2, 6

[2] O. Chum et al. Geometric min-hashing: Finding a (thick) needle in a haystack. In *IEEE CVPR*, 2009. 2, 3

[3] J. Dean et al. Mapreduce: Simplified data processing on large clusters. In *OSDI*, 2004. 3

[4] T. Elsayed et al. Pairwise document similarity in large collections with mapreduce. In *Proceedings of ACL-08: HLT, Short Papers*, pages 265–268, 2008. 3

[5] B. J. Frey et al. Clustering by passing messages between data points. *Science*, 2007. 3

[6] S. Gammeter et al. I know what you did last summer: Object-level auto-annotation of holiday snaps. In *IEEE ICCV*, 2009. 2

[7] J. Hays et al. im2gps: estimating geographic information from a single image. In *IEEE CVPR*, 2008. 1

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[9] H. Ma et al. Bridging the semantic gap between image contents and tags. *IEEE TMM*, 2010. 2

[10] P. K. Mallapragada et al. Online visual vocabulary pruning using pairwise constraints. In *IEEE CVPR*, 2010. 2

[11] J. Philbin et al. Object retrieval with large vocabularies and fast spatial matching. In *IEEE CVPR*, 2007. 2, 3, 6

[12] J. Philbin et al. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE CVPR*, 2008. 1

[13] J. Philbin et al. Descriptor learning for efficient retrieval. In *ECCV*, 2010. 2

[14] J. Sivic et al. Video Google: A text retrieval approach to object matching in videos. In *IEEE CVPR*, 2003. 1, 2, 7

[15] P. Turcot et al. Better matching with fewer features: The selection of useful features in large database recognition problems. In *IEEE ICCV Workshop on WS-LAVD*, 2009. 2

[16] X.-J. Wang et al. Multi-model similarity propagation and its application for web image retrieval. In *ACM MM*, 2004. 2

[17] X.-J. Wang et al. Annosearch: Image auto-annotation by search. In *IEEE CVPR*, 2006. 1

[18] X.-J. Wang et al. Arista - image search to annotation on billions of web photos. In *CVPR*, 2010. 1

[19] L. Wu et al. Semantics-preserving bag-of-words models for efficient image annotation. In *ACM workshop on LSMMRM*, 2009. 2

[20] Y.-H. Yang et al. Contextseer: Context search and recommendation at query time for shared consumer photos. In *ACM MM*, 2008. 6

[21] X. Zhang et al. Efficient indexing for large scale visual search. In *IEEE ICCV*, 2009. 1