Unsupervised Semantic Feature Discovery for Image Object Retrieval and Tag Refinement

Yin-Hsi Kuo, Wen-Huang Cheng, Hsuan-Tien Lin, Winston H. Hsu

***Abstract*—We have witnessed the exponential growth of images and videos with the prevalence of capture devices and the ease of social services such as Flickr and Facebook. Meanwhile, enormous media collections are along with rich contextual cues such as tags, geo-locations, descriptions, and time. To obtain desired images, users usually issue a query to a search engine using either an image or keywords. Therefore, the existing solutions for image retrieval rely on either the image contents (e.g., low-level features) or the surrounding texts (e.g., descriptions, tags) only. Those solutions usually suffer from low recall rates because small changes in lighting conditions, viewpoints, occlusions or (missing) noisy tags can degrade the performance significantly. In this work, we tackle the problem by leveraging both the image contents and associated textual information in the social media to approximate the semantic representations for the two modalities. We propose a general framework to augment each image with relevant semantic (visual and textual) features by using graphs among images. The framework automatically discovers relevant semantic features by propagation and selection in textual and visual image graphs in an unsupervised manner. We investigate the effectiveness of the framework when using different optimization methods for maximizing efficiency. The proposed framework can be directly applied to various applications, such as keyword-based image search, image object retrieval, and tag refinement. Experimental results confirm that the proposed framework effectively improve the performance for these emerging image retrieval applications.**

***Index Terms*—semantic feature discovery, image graph, image object retrieval, tag refinement**

I. Introduction

MOST of us have been used to sharing personal photos on the social services (or media) such as Flickr and Facebook. More and more users are also willing to contribute related tags or comments on the photos for photo management and social communication [1]. Such user-contributed contextual information provides promising research opportunities for understanding the images in social media. Image retrieval (either content-based or keyword-based) over large-scale photo collections is one of the key techniques for managing the exponentially growing media collections. Lots of applications such as annotation by search [2][3] and geographical information estimation [4] are keen to the accuracy and efficiency of content-based image retrieval (CBIR) [5][6]. Nowadays, the existing image search engines employ not only the surrounding texts but also the image contents to retrieve images (e.g., Google and Bing).

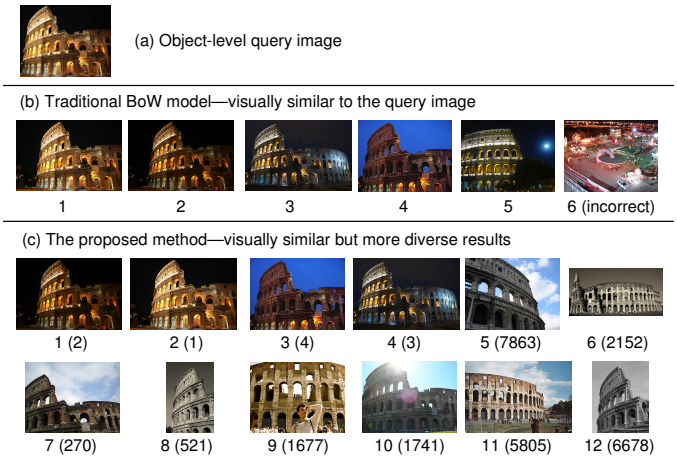


Fig. 1. Comparison in the image object retrieval performance of the traditional BoW model [5] and the proposed approach. (a) An example of object-level query image. (b) The retrieval results of a BoW model, which generally suffers from the low recall rate. (c) The results of the proposed system, which obtains more accurate and diverse images, with the help of automatically discovered visual features. 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.

For CBIR systems, bag-of-words (BoW) model is popular and shown effective [5]. 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 [6][7]. Thus, the methods suffer from low recall rate as shown in Figure 1(b). Due to varying capture conditions and large VW vocabulary (e.g., 1 million vocabulary), the features for the target images might have different VWs (cf. Figure 1(c)). Besides, it is also difficult to obtain these VWs through query expansion (e.g., [8]) or even varying quantization methods (e.g., [6]) because of the large differences in visual appearance between the query and the target objects.

For keyword-based image retrieval in social media, textual features such as tags are more semantically relevant than visual features. However, it is sill difficult to retrieve all the target images by keywords only because users might annotate non-specific keywords such as “Travel” [9]. Meanwhile, in most photo-sharing websites, tags and other forms of text are freely

Y.-H. Kuo is with the Graduate Institute of Networking and Multimedia, National Taiwan University (e-mail: kuonini@cmlab.csie.ntu.edu.tw). W.-H. Cheng is with Research Center for Information Technology Innovation, Academia Sinica, Taiwan (e-mail: whcheng@citi.sinica.edu.tw). H.-T. Lin is with the Department of Computer Science and Information Engineering, National Taiwan University (e-mail: htlin@csie.ntu.edu.tw). W. H. Hsu is with the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: winston@csie.ntu.edu.tw). Prof Hsu is the contact person.

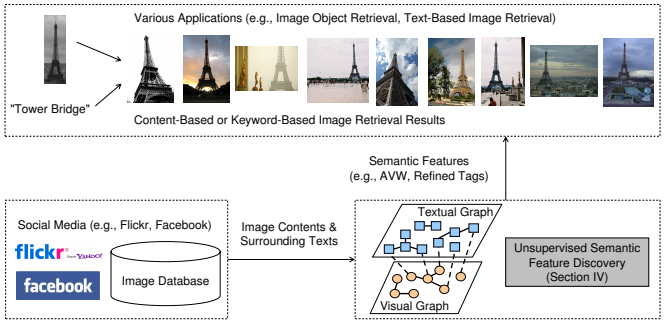


Fig. 2. A system diagram of the proposed method. Based on multiple modalities such as image contents and tags from social media, we propose an unsupervised semantic feature discovery which exploits both textual and visual information. The general framework can discover semantic features (e.g., semantically related visual words and tags) in large-scale community-contributed photos. Therefore, we can apply semantic features to various applications.

entered and are not associated with any type of ontology or categorization. Tags are therefore often inaccurate, wrong or ambiguous [10].

In response to the above challenges for content-based and keyword-based image retrieval in social media, we propose a general framework, which integrates both visual and textual information[1], for *unsupervised semantic feature discovery* as shown in Figure 2. In particular, we augment each image in the image collections with semantic features—additional features that are semantically relevant to the search targets (cf. Figure 1(c))—such as specific VWs for certain landmarks or refined tags for certain scenes and events. Aiming at large-scale image collections for serving different queries, we mine the semantic features in an unsupervised manner by incorporating both visual and (noisy) textual information. We construct graphs of images by visual and textual information (if available) respectively. We then automatically propagate and select the informative semantic features across the visual and textual graphs (cf. Figure 5). 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 frameworks (e.g., MapReduce).

We demonstrate the effectiveness of the proposed framework by applying it to two specific tasks, i.e., image object retrieval and tag refinement. The first task—*image object retrieval*—is a challenging problem because the target object may cover only a small region in the database images as shown in Figure 1. We apply the semantic feature discovery framework to augment each image with *auxiliary visual words (AVW)*. The second task is *tag refinement* which augments each image with semantically related texts. Similarly, we apply the framework on the textual domain by exchanging the role of visual and textual graphs so that we can propagate (in visual graph) and select (in textual graph) relative and representative tags for each image.

Experiments show that the proposed method greatly improves the recall rate for image object retrieval. In particular, the unsupervised auxiliary visual words discovery greatly outperforms BoW models (by 111% relatively) and is complementary to conventional pseudo-relevance feedback. Meanwhile, AVW discovery can also derive very compact (i.e., $\sim$1.4% of the original features) and informative feature representations which will benefit the indexing structure [5][14]. Besides, experimental results for tag refinement show that the proposed method can improve text-based image retrieval results (by 10.7% relatively).

The primary contributions of the paper[2] include,

- Observing the problems in image object retrieval by conventional BoW model (Section III).
- Proposing semantic feature discovery through visual and textual clusters in an unsupervised and scalable fashion, and deriving semantically related visual and textual features in large-scale social media (Section IV and Section VI).
- Investigating different optimization methods for efficiency and accuracy in semantic feature discovery (Section V).
- Conducting experiments on consumer photos and showing great improvement of retrieval accuracy for image object retrieval and tag refinement (Section VIII).

II. Related Work

In order to utilize different kinds of features from social websites, we propose a general framework for semantic feature discovery through image graphs in an unsupervised manner. The semantic visual features can be visual words or user-provided tags. To evaluate the effect of semantic feature discovery, we adopt the proposed framework to image object retrieval and tag refinement. Next, we introduce some related work for these issues in the following paragraphs.

Most image object retrieval systems adopt the scale-invariant feature transform (SIFT) descriptor [16] to capture local information and adopt bag-of-words (BoW) model [5] to conduct object matching [8][17]. 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 [18], textual information [19]. However, it usually needs extra (manual) information during the 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 [20]. It suggested

[1]We aim to integrate different contextual cues (e.g., visual and textual) to generate semantic (visual or visual) features for database images in the offline process. In dealing with the online query, i.e., when users issue either an image or keywords to the search engine, we can retrieve diverse search results as shown in Figure 1(c). Of course, if the query contains both image and keywords, we can utilize the two retrieval results or adopt advanced schemes like the re-ranking process for obtaining better retrieval accuracy [11][12][13].

[2]Note that the preliminary results were presented in [15]. We extend the original method to a general framework and further apply it in the text domain.

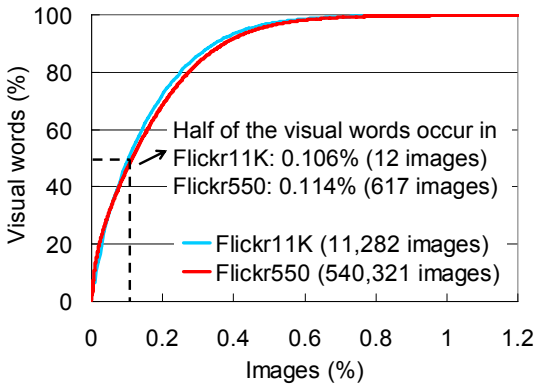


Fig. 3. Cumulative distribution of the frequency of VW occurrence in two different image databases, cf. Section III-A. It shows that half of the VWs occur in less than 0.11% of the database images (i.e., 12 and 617 images, respectively). The statistics represent that VWs are distributed over the database images sparsely. Note that the x-axis only shows partial values (0% 1.2% images) because the cumulative distribution almost saturates (∼99%) at 1.2% level, so we skip the remaining parts (1.2% 100%) in the figure.

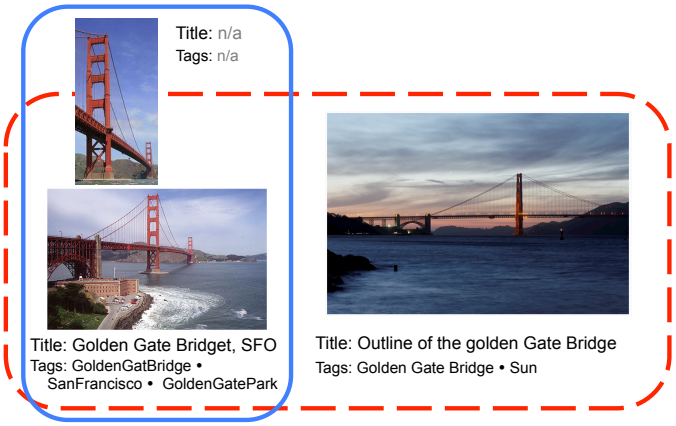


Fig. 4. Illustration of the roles of semantic related features in image object retrieval. Images in the blue rectangle are visually similar, whereas those images in the red dotted rectangle are textually similar. The semantic (textual) features are promising to establish the in-between connection (Section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

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 [21] and [22] 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.

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 methods for post-processing in image retrieval. Similar limitations are observed in [23], where only the image similarity scores are propagated between textual and visual graphs. Different from the prior works, we use local features for image object retrieval and propagate features directly between the textual and visual graphs. The discovered semantic visual features are thus readily effective in retrieving diverse search results, eliminating the need to apply a random walk in the graphs again.

Similar to [9], we can also apply our general framework to augment keyword-based image retrieval by tag refinement to improve text (tag) quality for image collections. Through tag propagation and selection processes, we can annotate images and refine the original tags. Annotation by search [3] is a data-driven approach which relies on retrieving (near) duplicate images for better annotation results. The authors in [24] propose a voting-based approach to select proper tags via visually similar images. Different from annotation by search [3] and voting-based tag refinement [24], we propagate and select informative tags across images in the same image clusters. Meanwhile, the tag propagation step can also assign suitable tags for those images without any tags in database.

For both [25] and [26], they focus on modifying the weights of the tags originally existing in the photo and only retain those reliable tags based on the voting by visually similar images. Instead, our proposed method concentrates on obtaining more (new) semantically related tags from semantically related images. We further select those representative tags to suppress noisy or incorrect tags. [14] proposed to select the most descriptive visual words according to the TF-IDF weighting. Different from [14], our selection process further considers similar images to retain more representative tags.

III. Key Observations—Requiring Semantic Feature for Image Retrieval

Nowadays, bag-of-words (BoW) representation [5] is widely used in image retrieval and has been shown promising in several content-based image retrieval (CBIR) tasks (e.g., [17]). However, most existing systems simply apply the BoW model without carefully considering the sparse effect of the VW space, as detailed in Section III-A. Another observation (explained in Section III-B) 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 semantic feature discovery method is targeted to address these issues.

A. 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 [17]. As mentioned in [27], one observation is the uniqueness of VWs—visual words in images usually do not appear more than once. Moreover, 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 VII-A), and obtain similar curves as shown in Figure 3. 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

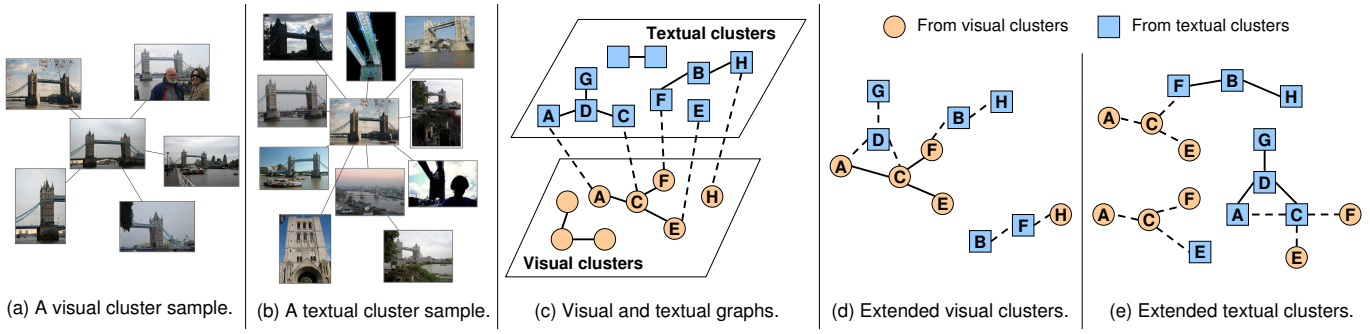


Fig. 5. The visual cluster (a) groups visually similar images in the same cluster, whereas the textual cluster (b) favors semantic similarities. The two clusters facilitate representative feature selection and semantic feature propagation, e.g., visual words, tags. Based on visual and textual graphs in (c), we can propagate auxiliary features among the associated images in the extended visual or textual clusters. (d) 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. Similarly, (e) shows three extended textual clusters include the semantically (by expanded tags) similar images and those co-occurrences in other visual clusters.

contain similar features. In other words, those similar images might only have few common VWs. This phenomenon is the sparseness of the VWs. It is partly due to some quantization errors or noisy features. Therefore, in Section IV, we propose to augment each image with auxiliary visual words.

B. Lacking Semantics Related Features

Since VWs are merely low-level visual features, it is very difficult to retrieve images with different viewing angles, lighting conditions, partial occlusions, etc. An example is shown in Figure 4. By using BoW models, the query image (e.g., 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 benefits of 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 4. If those semantically similar images can share (propagate) their VWs to each other, the query image can retrieve similar but more visually and semantically diverse results.

IV. Semantic Feature Discovery Framework

Based on the observations above, it is necessary to discover semantic features for each image. Unlike previous works that focus on constructing the features in one single domain, we propose a general framework for semantic feature discovery based on multiple modalities such as image contents and tags. Meanwhile, such framework can also discover semantically related visual words and tags in large-scale community-contributed photos. In this section, we first illustrate the framework from the view of the visual domain. Then we adapt the framework for applications in the textual domain in Section VI.

As mentioned in Section III, it is important to propagate VWs to those visually or semantically similar images. We follow the intuition to propose an offline stage for unsupervised semantic feature discovery. We augment each image with *auxiliary visual words (AVW)*—additional and important features relevant to the target image—by considering semantically related VWs in its textual cluster and representative VWs in its visual cluster. When facing large-scale datasets, we can deploy the processes in a parallel way (e.g., MapReduce [28]). Besides, AVW reduces the number of VWs to be indexed (i.e., better efficiency in time and memory [14]). Such AVW might potentially benefit the further image queries and can greatly improve the recall rate as demonstrated in Section VIII-A and in Figure 8. For mining AVWs, we first generate image graphs and image clusters in Section IV-A. Then based on the image clusters, we propagate auxiliary VWs in Section IV-B and select representative VWs in Section IV-C. Finally, we combine both selection and propagation methods in Section IV-D.

A. Graph Construction and Image Clustering

The proposed framework starts by constructing a graph which embed image similarities from the image collection. We adopt efficient algorithms to construct the large-scale image graph by MapReduce. We apply [29] to calculate the image similarity since we observe that most of the textual and visual features are sparse for each image and the correlation between images are sparse as well. 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. To cluster images on the image graph, we apply affinity propagation (AP) [30] 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.

In this work, the images are represented by 1M VWs and 90K text tokens expanded by Google snippets from their associated (noisy) tags. The image clustering results are sampled in Figure 5(a) and (b). 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.

B. Auxiliary Visual Word 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(c)). 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 5(b)) .

We actually conduct the propagation on each *extended visual cluster*, containing the images in a visual cluster and those additional ones co-occurring with these images in certain textual clusters. The intuition is to balance visual and semantic consistence for further VW propagation and selection (cf. Section IV-C). Figure 5(d) shows two extended visual clusters derived from Figure 5(c). More interestingly, image E has no tags and is thus singular in the textual cluster; however, E still belongs to a visual cluster and can 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. Let X_i be the VW histogram of image i and assume M among N images are from the same visual cluster. For example, $N = 8$ and $M = 4$ in the left extended visual cluster in Figure 5(d). 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. $P(i, j)$ weights the whole features propagated from image j to i. If we multiply the propagation matrix P and X (i.e., PX), we can obtain a new $M \times D$ VW histograms, as the AVWs. Each row of PX represents the new VW histogram for each image which 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). Here is an example of an initial propagation matrix P_0,

$$P_0 = \begin{array}{c} \\ A \\ C \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array}.$$

Each row represents the relationship between the image and its semantically similar images (i.e., in the same textual cluster). For example, image A (the first row) is related to image A, C, D and G as shown in Figure 5(c). Note that we can also modify the weights in P_0 based on the similarity score or AP score. We propose to formulate the propagation operation as

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

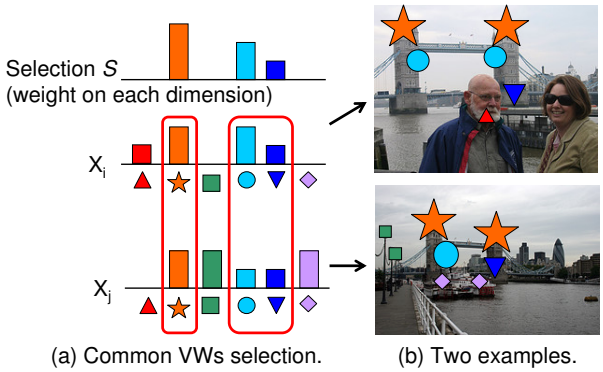


Fig. 6. Illustration of the selection operation for auxiliary visual words. The VWs should be similar in the same visual cluster; therefore, we select those representative visual features (red rectangle). (b) illustrates the importance (or representativeness) for different VWs. And we can further remove some noisy features (less representative) which appeared on the people or boat. The similar idea can be used to select informative tags from the noisy ones for each image.

The goal of the first term is to avoid from propagating too many VWs (i.e., propagating conservatively) 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). Here $||.||_F$ stands for the Frobenius norm. $N_{P1} = \|P_0X\|_F^2$ 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 VIII-C. Note that the propagation process updates the propagation matrix P on each extended visual cluster separately as shown in Figure 5 (d); therefore, this method is scalable for large-scale dataset and easy to adopt in a parallel way.

C. Common Visual Word 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 5(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 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 VIII-C.

D. 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 VIII-A.

V. 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

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

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

A. 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

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

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

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.

B. Gradient Descent Solver (GD)

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

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

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

For the selection formulation (Section IV-C), we can adopt similar steps with two changes. And let $\beta_1 = \frac{\beta}{N_{S1}}$ and $\beta_2 = \frac{1-\beta}{N_{S2}}$. First, Eq. (6) is replaced with

$$S^{new} = S^{old} - 2\eta\big(-\beta_1 X^T X(S_0 - S^{old}) + \beta_2 S^{old}\big). \tag{7}$$

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 occurred in the same visual cluster, because the other dimensions would contribute 0 to $X^T X$.

C. 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

$$S^* = \beta_1(\beta_1 X^T X + \beta_2 I_D)^{-1} X^T X S_0. \tag{8}$$

Nevertheless, as mentioned in the previous subsection, the X^TX 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^TX directly, we transform X^TX 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

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

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

$$S^* = \beta_1 X^T(\beta_1 XX^T + \beta_2 I_N)^{-1} XS_0. \tag{9}$$

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.

VI. Tag Refinement

Textual features are generally semantically richer than visual features. However, tags (or photo descriptions) are often missing, inaccurate, or ambiguous as annotated by the amateurs [10]; e.g., adding the tag "honeymoon" to all images of a newly married couple's trip. Traditional keyword-based image retrieval systems are thus limited in retrieving these photos with noisy or missing textual descriptions. Hence, there arise strong needs for effective image annotation and tag refinement. To tackle this problem, most recent researches focus on annotation by search [2][3] or discovering relevant tags from the votes by its visually similar images [24][25]. Those previous work solely rely on one feature modality to improve the tag quality. In this work, we further propose to annotate and refine tags by jointly leveraging the visual and textual information.

In Section IV, we propose a framework for semantic feature discovery, where we utilize the image graphs to propagate and select auxiliary visual words starting from the images' textual relations for introducing more diverse but semantically relevant visual features. In this section, we will show that the proposed framework is general and can be extended to tag refinement and photo annotation by exchanging the roles of visual and textual graphs. That is, starting from the visual graph, we propagate and then select representative tags in the textual graph. We will introduce tag propagation in Section VI-A and representative tag selection, where we further considered the sparsity of tags in Section VI-B. Note than we apply our proposed method on the same image graphs constructed in Section IV-A.

A. Tag Propagation

In order to obtain more semantically relevant tags for each image, we propose to propagate tags through its visually similar images. We will then remove noisy tags and preserve representative ones in Section VI-B. Following the auxiliary feature propagation in Section IV-B, we construct the *extended textual cluster* to propagate relevant tags. As shown in Figure 5(e), we conduct the propagation on each extended textual cluster which contains the images in a textual cluster and those additional ones co-occurring with any of these images in certain visual clusters (in the image graph).

To find a proper propagation matrix for each extended textual cluster, we can adopt the same formulation as mentioned in Section IV-B. That is, we can directly apply Eq. (1) to propagate related tags on the extended textual clusters. It brings some advantages as discussed in Section IV-B and is also applicable to the textual domain. For example, as shown in Figure 5(c), image E has no tags and thus is singular in the textual cluster. However, through the tag propagation, image E can obtain some related tags from the images of A, C, and F (cf. Figure 5(e)). Note that this process is similar to image annotation. In the same way, image H is singular in the visual cluster, we can still propagate related tags to image H through extended textual cluster. For example, an image might obtain different tags such as "Tower bridge," "London," or "Travel."

B. Tag Selection and Sparsity of Tags

After the previous tag propagation step, each image can obtain more different tags. However, it is possible to obtain some incorrect ones. Similar to visual feature selection in Section IV-C, we propose to retain important (representative) tags and suppress the incorrect ones. To select important tags for each image, we can directly adopt the same selection formulation (Eq. (2)) as mentioned in Section IV-C. Following Eq. (2), we select representative tags in each textual cluster since images in the same textual cluster are semantically similar to each other. For example, in Figure 5(b), the more specific tag, "Tower bridge," would have higher score than a general one, "London."

Through tag selection, we can highlight the representative tags and reject the noisy ones. However, as the system converges, we observed that each image tends to have many tags with very small confidence scores; an ad-hoc thresholding process is required to cut those low-confidence (probably noisy) tags. Meanwhile, users usually care about few important (representative) tags for each image rather than plenty of tags. Thus, we need to further consider the sparsity of selected tags. We do so by modifying the original regularization term (L2-norm) to L1-norm. That is, the objective function of Eq. (2) is adjusted as:

$$f_{SS} = \min_S \|XS_0 - XS\|_F^2 + \lambda \|S\|_1. \tag{10}$$

λ is a regularization parameter. Since the L1-norm regularization term is non-differentiable, we can not obtain the analytic solution directly. However, recent researches have provided certain solutions for this problem [31][32], we can derive the solution by way of [33] or SPAMS (SPArse Modeling Software).[3]

[3]http://www.di.ens.fr/willow/SPAMS/

TABLE I
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 [5] 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 (propagation first)			Selection → Propagation (selection first)		
	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

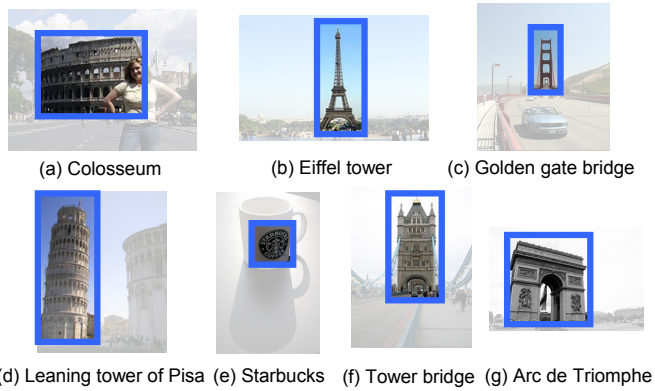


Fig. 7. Query examples in the Flickr11K dataset used for evaluating image object retrieval and text-based image retrieval. The query objects are enclosed by the blue rectangles and the corresponding query keywords are listed below each object image.

VII. EXPERIMENTAL SETUP

A. Dataset

We use Flickr550[4] [34] 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 (Eiffel), Golden Gate Bridge (Golden), Leaning tower of Pisa (Pisa), Starbucks logo (Starbucks), Tower Bridge (Tower), and Arc de Triomphe (Triomphe). Also, we randomly pick up 10,000 images from Flickr550 to form a smaller subset called Flickr11K.[5] Some query examples are shown in Figure 7.

B. Performance Metrics

In the experiments, we use the average precision, a performance metric commonly used in the previous work [17], [34], 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.

C. Evaluation Protocols

As suggested by the previous work [17], 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) [8], 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 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 8 and Section VIII-A. We take L1 distance as our baseline for BoW model [5]. The MAP for the baseline is 0.245 with 22M (million) feature points and the MAP after PRF is 0.297 (+21.2%).

For evaluating tag refinement, we seek text-based image retrieval to evaluate the overall tag quality. We also include semantic queries in text-based image retrieval tasks. We use the following keywords as the query for the 12 categories: Colosseum, Eiffel tower, Golden gate bridge, Leaning tower of Pisa, Starbucks, Tower bridge, Arc de Triomphe, Beach, Football, Horse, Louvre, and Park. Note that we use the same ground truth images as content-based image retrieval for evaluation.

VIII. RESULTS AND DISCUSSIONS

In this section, we conduct experiments on the proposed framework—unsupervised semantic feature discovery. Since we target a general framework for serving different applications, we will first adopt the proposed method to visual domain for image object retrieval in Section VIII-A and then the textual domain for tag refinement (by keyword-based retrieval and annotation) in Section VIII-B. Moreover, in Section VIII-C, we also investigate the impact of different parameters in the formulations.

A. The Performance of Auxiliary Visual Words

The overall retrieval accuracy is listed in Table I. As mentioned in Section IV-D, we can iteratively update the features according to Eq. (1) and (2). It shows that the iteration with propagation first (propagation → selection) lead to 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 results because those informative

[4]http://mpac.ee.ntu.edu.tw/%7Eyihsuan/reranking/contextseer
[5]http://www.cmlab.csie.ntu.edu.tw/%7Ekuonini/Flickr11K

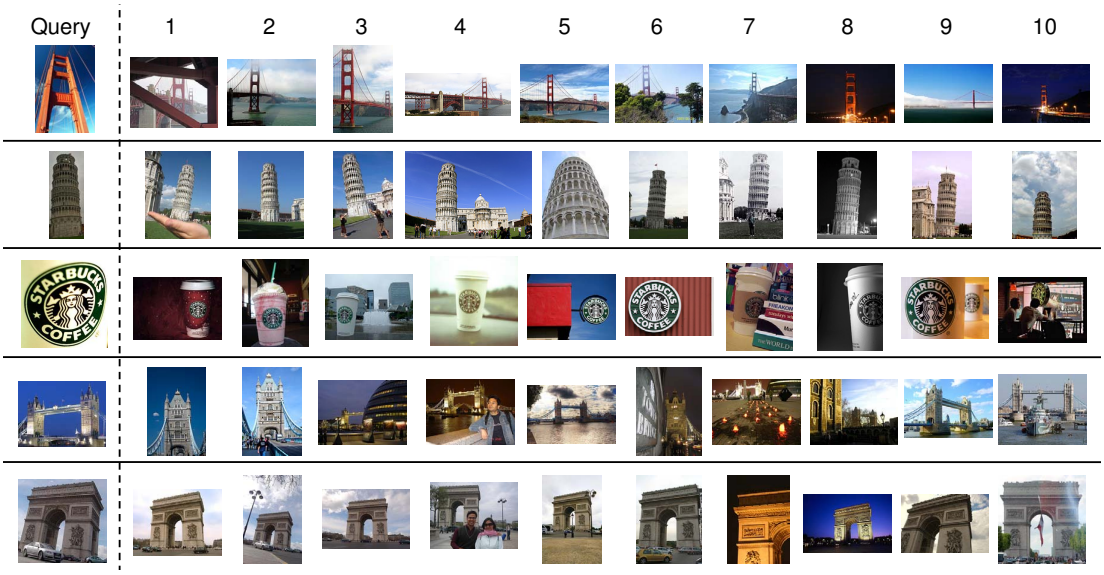


Fig. 8. More search results by auxiliary VWs. The number represents its retrieval ranking. The results show that the proposed AVW method, though conducted in an unsupervised manner in the image collections, can retrieve more diverse and semantic related results.

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 [5][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 I 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.[6] 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 iteratively. 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 8, 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.

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

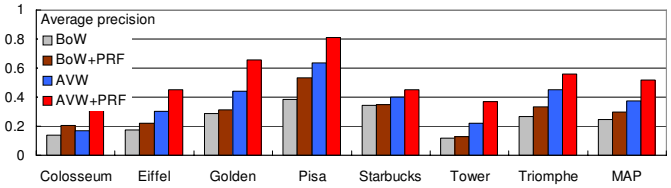


Fig. 9. Performance breakdown with auxiliary VWs (AVW) and PRF for image object retrieval. Consistent improvements across queries are observed. The right most is the average performance across seven queries (by MAP).

Figure 9 shows the performance breakdown for the seven queries. It can be found that the combination of AVW and PRF consistently improves the performance across all query categories. Especially, the proposed method works well for small objects such as "Starbucks logo," whereas the combination of BoW and PRF just marginally improves the retrieval accuracy. Besides, it is worthy to notice that the proposed method can achieve large improvements in "Tower bridge" query although the ground-truth images of "Tower bridge" usually have various lighting conditions and viewpoint changes as shown in Figure 5(b) and the fourth row of Figure 8.

B. The Performance of Tag Refinement

For the tag refinement task introduced in Section VI, we employed text-based image retrieval to evaluate the MAP by using predefined queries as mentioned in Section VII. The goal is to evaluate the overall tag quality before and after the tag refinement in the image collection. The overall retrieval accuracy is shown in Table II. It shows that our proposed method (Propagation + Selection) in general achieves better retrieval accuracy (+10.7%) because the tag propagation process obtains more semantically related tags and the tag selection process further preserves representative ones. However, the proposed method might slightly degrade after the tag refinement. For example, the "Starbucks" query does not gain

TABLE II
THE MAP OF TAG REFINEMENT RESULTS EVALUATED BY TEXT-BASED IMAGE RETRIEVAL. NOTE THAT WE USE THE FIRST COLUMN OF THE TABLE AS QUERY KEYWORDS TO RETRIEVE IMAGES. IT SHOWS THAT THE COMBINATION OF TAG PROPAGATION AND TAG SELECTION CAN ENHANCE TAG QUALITY AND THEN IMPROVE RETRIEVAL RESULTS. BESIDES, WE FURTHER IMPROVE THE PERFORMANCE BY COMBINING THE VOTING-BASED METHOD AND OUR APPROACH. '%' INDICATES THE RELATIVE MAP GAIN OVER THE ORIGINAL TAGS (0.498 MAP).

Query	Original tags	Voting-based method [24]	Propagation + Selection	Voting + Ours
Colosseum	0.694	0.716	0.790	0.831
Eiffel	0.467	0.468	0.676	0.699
Golden	0.463	0.463	0.671	0.699
Pisa	0.136	0.137	0.303	0.353
Starbucks	0.855	0.855	0.640	0.866
Tower	0.515	0.522	0.576	0.651
Triomphe	0.460	0.448	0.701	0.668
Beach	0.349	0.389	0.227	0.398
Football	0.543	0.655	0.628	0.686
Horse	0.784	0.783	0.601	0.774
Louvre	0.430	0.521	0.628	0.647
Park	0.281	0.285	0.178	0.301
MAP	0.498	0.520 (+4.4%)	0.551 (+10.7%)	0.631 (+26.7%)

from the proposed method because "Starbucks" images in the visual cluster tend to have more semantically diverse tags as the small objects do not necessarily correlate with semantically and visually similar images. In addition, incorrect (noisy) tags might be kept through the tag propagation process. Although the tag selection mechanism can help to alleviate this problem, it sometimes degrades the retrieval accuracy due to the loss of some important tags. For example, the "Triomphe" query obtains higher retrieval accuracy right after the tag propagation (0.729) but decreases slightly after the selection (0.701).

Besides, the voting-based method [24] reaches better accuracy in few queries (e.g., "Beach") since it merely reweighs the tags originally existing in the photo. Different from [24], the proposed method aims to obtain more semantically related tags through the propagation process. Therefore, the proposed method might slightly degrade in few queries (e.g., "Park") due to the limitation of the BoW feature for describing the visual graph among scene-related images.[7] Although the propagation process highly relies on the visual similarity, the selection process can alleviate this effect by retaining more representative tags (e.g., "Football:" from 0.366 (propagation) to 0.628 (+ selection)) so that the overall retrieval accuracy is still better. Moreover, we notice that there is an advantage for the voting-based method as mentioned above so that we further combine it and our method (Voting + Ours) to achieve the best results (+26.7%).

We also show tag refinement examples in Figure 10. As mentioned above, each image can obtain more related (new) tags after tag propagation as shown in Figure 10(b) (e.g., "Colosseum" or "Eiffel tower"). And each image can further retain those representative tags and reject incorrect (or less

Fig. 10. Examples for tag refinement by tag propagation and selection (Section VI). (a) shows the original tags from the Flickr website. After tag propagation, each image can have more related (new) tags (b). To reduce incorrect (noisy) tags, we adopt tag selection to select the informative and representative tags (c). We further consider sparsity for tag selection to retain few (salient) tags (d). Note that the correct tags are indicated in bold style.

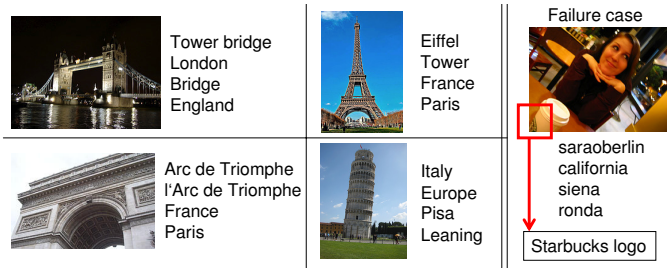


Fig. 11. Example results for image annotation on those images originally not associated with any tags. Though initially each image is singular in the textual cluster, through extended textual cluster and the following semantic feature (tag) discovery, each image can obtain semantically related tags. However, if the image object is too small to derive visually similar images (e.g., Starbucks logos), it might incur poor annotations.

frequent) ones (e.g., "Visiteiffel") after the tag selection in Figure 10(c). Interestingly, through the processes, we could also correct typos or seldom used tags such as "Colloseum" (widely used: "Colosseum," "Coliseum" or "Colosseo"). To further consider the tag sparsity, we can retain few representative tags (e.g., "Eiffel tower") as shown in Figure 10(d). However, it is possible to retain only some common tags such as "Paris" or "London."

Moreover, the tag refinement process can also annotate those images which initially do not have any tags. Figure 11 shows some image annotation results after the tag refinement process. During the tag propagation step, a single image (node) in the textual graph will obtain tags via its related visual clusters. This approach is similar to annotation by search; however, we base on the extended textual clusters to propagate tags rather than the search results.[8] As shown in Figure 11, we can correctly annotate some images on the left-hand side; nevertheless, it is still possible to propagate some incorrect tags such as the rightmost case because the visual (textual) clusters might be noisy. This can be improved if more effective clustering methods and contextual cues are employed.

To provide another view for evaluating annotation quality, we first remove the original (user-provided) tags before con-

[7]We believe the fusion of further visual features (e.g., texture, color) will enhance this part. In this work, we emphasize the proposed general framework for deriving semantic (visual or textual) features by leveraging more contextual cues along with the large-scale photo collections.

[8]Note that image annotation is a by-product of tag refinement and it only annotated database images rather than a new query image. Therefore, we do not compare with the other methods such as annotation by search [3].

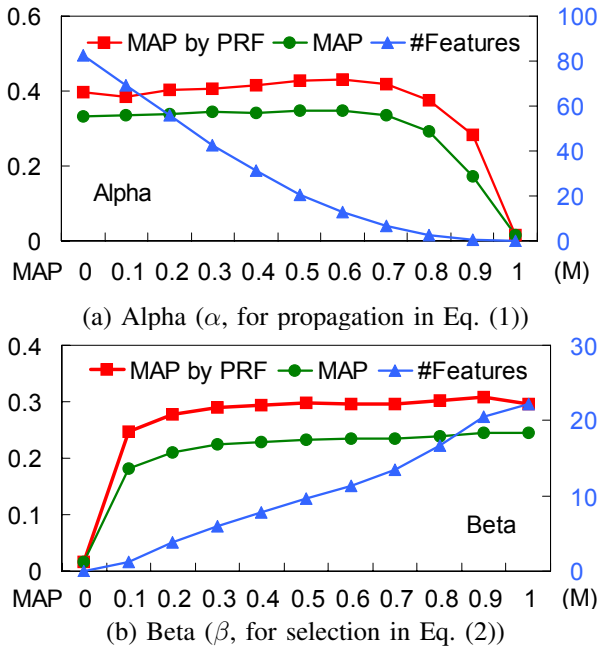
Fig. 12. The illustration for another evaluation for image annotation by removing the original (user-provided) tags before conducting the proposed method. In the left image, we can provide more specific tags than the original ones. However, we may annotate incorrect tags due to the limitation of the BoW feature for describing the visual graph among scene-related images.

ducting our proposed method. As shown in Figure 12, the proposed method can annotate semantically related tags if the image has more supporting photos from its visual cluster. It is interesting that we may annotate more specific tags (e.g., "London" or "Tower bridge") than the original ones (e.g., "Europe") as the left example given in Figure 12. This is because other photographers may accurately name the exact spot [25] and our approach can effectively leverage such cues to provide more specific tags. Note that it is still possible that we annotate incorrect tags since the BoW feature is limited in describing scene-related images (e.g., park). As we observe and many other literatures [26] have shown, it is effective to include more visual features for building the visual similarities. In this work, rather than optimizing the visual features, we emphasize the proposed general framework for deriving more semantic (visual and textual) features through the propagation and selection processes over the supplemental contextual cues (e.g., (noisy) tags, photos, geo-locations) commonly observed in social media.

C. Parameter Sensitivity

Finally, we report the impact of sensitive tests on two important parameters—propagation formulation (α) and selection formulation (β). Here we evaluate the effect on image object retrieval only and we find the same parameters are applicable to other applications. The results are shown in Figure 13. In the propagation formulation, α decides the number of features needed to be propagated. Figure 13(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, 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 13(b) shows that if we just keep a few dimensions of VWs, the MAP is still



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

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

Fig. 13. Parameter sensitivity on alpha and beta in AVW discovery for image object retrieval. (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 VIII-C. Note that we can further improve retrieval accuracy by iteratively updating semantic features by the proposed propagation and selection processes.

similar to BoW baseline though with some retrieval accuracy decrease. Because of the spareness of large VW vocabulary as mentioned in Section III-A, we only need to keep those important VWs.

IX. Conclusions and Future Work

In this work, we present a general framework for semantic feature discovery which utilizes both the visual and textual graphs to propagate and select important (visual or textual) features. First, 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 (graphs). 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) for image object retrieval. Besides, we extend the proposed method to textual domain. It can not only help to retain representative tags for each image but also automatically derive meaningful tags to annotate unlabeled images. Experiments in text-based image retrieval show that tag refinement can improve the retrieval accuracy effectively (+10.7% relatively). We are to investigate more advanced contextual features, such as geo-tags, time, user attributes, along with the proposed framework to leverage the rich contexts from the emerging social media [35][36].

REFERENCES

[1] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 971–980.

[2] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "Annosearch: Image auto-annotation by search," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006, pp. 1483–1490.

[3] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma, "ARISTA - image search to annotation on billions of web photos," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 2987–2994.

[4] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[5] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[7] X. Zhang, Z. Li, L. Zhang, W. Ma, and H. Shum, "Efficient indexing for large scale visual search," in *IEEE International Conference on Computer Vision*, 2009, pp. 1103–1110.

[8] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[9] H. Ma, J. Zhu, M. R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Transactions on Multimedia*, vol. 12 (5), pp. 462–473, 2010.

[10] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: context and content in community-contributed media collections," in *ACM Multimedia*, 2007, pp. 631–640.

[11] V. Mezaris, H. Doulaverakis, S. Herrmann, B. Lehane, I. Kompatsiaris, and M. G. Strintzis, "Combining textual and visual information processing for interactive video retrieval: SCHEMA's participation in TRECVID 2004," in *TRECVID Workshop*, 2004.

[12] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Multimedia information retrieval*, 2006.

[13] X. Wang, K. Liu, and X. Tang, "Query-specific visual semantic spaces for web image re-ranking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 857–864.

[14] B. Thomee, E. M. Bakker, and M. S. Lew, "TOP-SURF: a visual words toolkit," in *Proceedings of the international conference on Multimedia*, 2010.

[15] Y.-H. Kuo, H.-T. Lin, W.-H. Cheng, Y.-H. Yang, and W. H. Hsu, "Unsupervised auxiliary visualwords discovery for large-scale image object retrieval," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60 (2), pp. 91–110, 2004.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[18] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *European Conference on Computer Vision*, 2010.

[19] L. Wu, S. C. Hoi, and N. Yu, "Semantics-preserving bag-of-words models for efficient image annotation," in *ACM workshop on Large-scale multimedia retrieval and mining*, 2009, pp. 19–26.

[20] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV Workshop on Emergent Issues in Large Amounts of Visual Data*, 2009, pp. 2109–2116.

[21] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "I know what you did last summer: Object-level auto-annotation of holiday snaps," in *IEEE International Conference on Computer Vision*, 2009, pp. 614–621.

[22] P. K. Mallapragada, R. Jin, and A. K. Jain, "Online visual vocabulary pruning using pairwise constraints," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

[23] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li, "Multi-model similarity propagation and its application for web image retrieval," in *ACM Multimedia*, 2004, pp. 944–951.

[24] X. Li, C. G. M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Multimedia Information Retrieval*, 2008, pp. 180–187.

[25] L. Kennedy, M. Slaney, and K. Weinberger, "Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases," in *Proceedings of the 1st workshop on Web-scale multimedia corpus*, 2009.

[26] X. Li, C. G. M. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010, pp. 10–17.

[27] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 17–24.

[28] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *Symposium on Operating Systems Design and Implementation*, 2004, pp. 137–150.

[29] T. Elsayed, J. Lin, and D. Oard, "Pairwise document similarity in large collections with mapreduce," in *the Association for Computational Linguistics*, 2008, pp. 265–268.

[30] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *International Conference on Machine Learning*, 2009.

[32] J. Mairal, F. R. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, pp. 19–60, 2010.

[33] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Neural Information Processing Systems*, 2006, pp. 801–808.

[34] Y.-H. Yang, P.-T. Wu, C.-W. Lee, K.-H. Lin, and W. H. Hsu, "Contextseer: Context search and recommendation at query time for shared consumer photos," in *ACM Multimedia*, 2008.

[35] D. Chen, G. Baatz, Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[36] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and Fisher vectors for efficient image retrieval," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.