# Facilitating Image Search With a Scalable and Compact Semantic Mapping

Meng Wang, *Member, IEEE*, Weisheng Li, Dong Liu, Bingbing Ni, Jialie Shen, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—This paper introduces a novel approach to facilitating image search based on a compact semantic embedding. A novel method is developed to explicitly map concepts and image contents into a unified latent semantic space for the representation of semantic concept prototypes. Then, a linear embedding matrix is learned that maps images into the semantic space, such that each image is closer to its relevant concept prototype than other prototypes. In our approach, the semantic concepts equated with query keywords and the images mapped into the vicinity of the prototype are retrieved by our scheme. In addition, a computationally efficient method is introduced to incorporate new semantic concept prototypes into the semantic space by updating the embedding matrix. This novelty improves the scalability of the method and allows it to be applied to dynamic image repositories. Therefore, the proposed approach not only narrows semantic gap but also supports an efficient image search process. We have carried out extensive experiments on various cross-modality image search tasks over three widely-used benchmark image datasets. Results demonstrate the superior effectiveness, efficiency, and scalability of our proposed approach.**

***Index Terms*—Compact semantic mapping (CSM), image search, semantic gap.**

## I. Introduction

At the beginning of the 21st century, the world wide web has brought about a fundamental change in the way how people access and consume media information [10], [30], [51]. The techniques to facilitate image search are receiving more and more attentions from different research communities, e.g., information retrieval and multimedia data management. In general, the existing approaches for image search can be roughly categorized into two widely recognized yet independent classes [10], [50], namely, text-based image retrieval (TBIR) and content-based image retrieval (CBIR).

Since none of these two paradigms can fully satisfy all user requirements for online image search, some recent research efforts have been made to combine the advantages from both TBIR and CBIR into a single search framework [5], [16], [21]. But a major obstacle is how to combine the features extracted from different modalities to support accurate and computationally efficient image search. Motivated by this fact, we develop a novel cross-modality search scheme that aims to minimize the semantic gap between high-level concepts and low-level visual features. Fig. 1 illustrates the flowchart of the online retrieval procedure facilitated by our proposed method. The scheme possesses good scalability in handling dynamic online image databases. To achieve this goal, we develop a method to explicitly map keyword terms and images into a compact latent semantic space with “semantically meaningful” distance metric. In particular, it places semantic concepts into the space in the form of concept prototypes. Then, a linear embedding matrix is learned to map low-level visual features into the semantic space. With the proposed approach, textual keywords can still be applied as query terms, and the mapped images around a related concept prototype can be viewed as the retrieved results of the query term. Consequently, the images are ranked according to their distances to the concept prototype. Since the images of a certain semantic category are enforced to compactly locate around the corresponding concept prototype in the semantic space, it essentially reduces the semantic gap between low-level features and high-level concepts. Moreover, our method possesses superior scalability in dealing with new semantic concepts via updating the embedding matrix without model reconstruction. The major technical contributions of this paper can be summarized as follows.

1) A novel cross-modality image search scheme is proposed that explicitly reduces the semantic gap between semantic concepts and low-level features extracted from

Manuscript received December 24, 2013; revised June 30, 2014, August 21, 2014, and August 27, 2014; accepted August 30, 2014. This work was supported in part by the National 973 Program of China under Grant 2014CB347600 and Grant 2013CB329604, in part by the National Nature Science Foundation of China under Grant 61272393 and Grant 61322201, in part by the Fundamental Research Funds for the Central Universities of China under Grant 2013HGCX0001, in part by the Doctoral Fund of Ministry of Education of China under Grant 20130111110010, in part by the Program for New Century Excellent Talents (NCET) in University under Grant NCET-12-0836, in part by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore’s Agency for Science, Technology, and Research, and in part by the Open Project Program of the National Laboratory of Pattern Recognition. This paper was recommended by Associate Editor X. He.

M. Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: eric.mengwang@gmail.com).

W. Li is with the Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: liws@cqupt.edu.cn).

D. Liu is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: dongliu@ee.columbia.edu).

B. Ni is with Advanced Digital Sciences Center, Singapore (e-mail: bingbing.ni@adsc.com.sg).

J. Shen is with Singapore Management University, Singapore (e-mail: jlshen@smu.edu.sg).

S. Yan is with the National University of Singapore, Singapore (e-mail: eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2356136

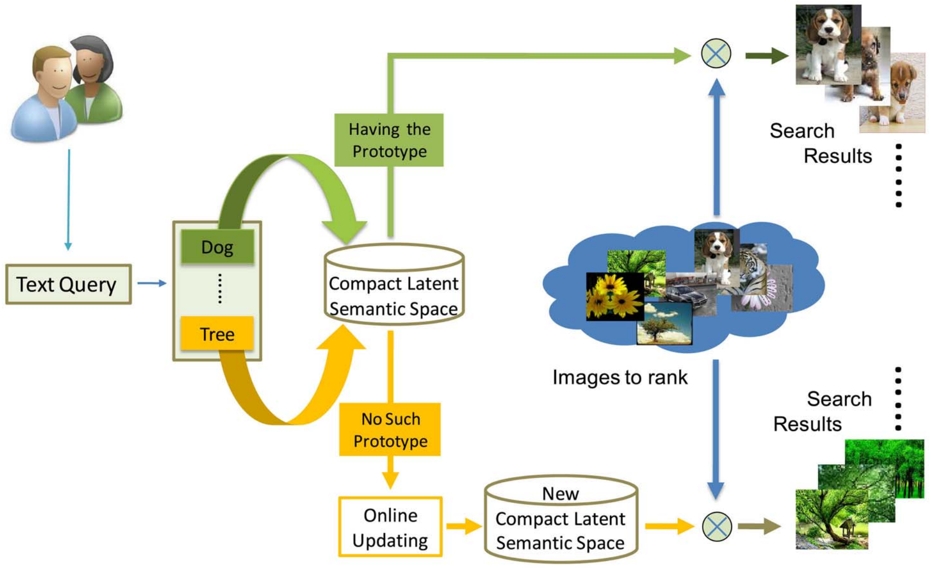


Fig. 1. Flowchart of online search procedure in the proposed image search scheme. For a given keyword, if the corresponding concept prototype has been established in the semantic space, the images indexed by the concept can be directly returned. Otherwise, an online updating procedure is triggered out to augment the semantic space with the newly arrived concept and images.

images. The approach also possesses superior scalability and can well handle dynamically increasing image repository.

2) A semantic embedding method called compact semantic embedding is proposed to project concept prototypes and visual features into a compact semantic space.
3) To deal with dynamically increasing semantic concepts, an online updating algorithm is proposed to refine the semantic space. The new semantic concepts and the corresponding images can be effectively retrieved in an incremental manner.

The remainder of this paper is organized as follows. Section II introduces some related works. Section III presents the proposed keyword query-based yet content aware image search paradigm. It also provides a detailed introduction about an online updating procedure for efficiently handling dynamic changes. Section IV presents the experiments over three benchmark image datasets. Finally, Section V concludes this paper.

## II. Related Work

The first research attempts to develop an image search system were made over 30 years ago. Since then, the related research problems have been a fundamental subject in areas such as computer vision, multimedia computing, and information retrieval. In each of these areas, plentiful literature has been published. In this section, we will focus only on the analysis of the technical developments that are most relevant to the study.

The most popular way to implement image search systems is based on unimodal approach, where query and retrieved documents are described using the same modality [1], [6], [26], [39], [42], [54]. One of the most typical examples is TBIR, which heavily relies on textual information to describe and annotate visual content of images [15], [33], [36], [40], [46]. But a major obstacle is the lack of the annotation of large-scale image collections. The most naive approach is to manually assign text labels to each image in database. However, since manual labeling requires a huge amount of time and domain expertise, it greatly restricts the feasibility of the approach in many real applications. Consequently, intelligent algorithm design for automatic image annotation has been extensively studied in last decades [3], [11], [13], [15], [20], [25], [41], [45], [48]. Many different algorithms and systems have been developed.

There exist a wide variety of features to describe an image (e.g., color, texture, shape, and spatial layout). Utilizing a single kind of features may not be able to represent an image comprehensively. Thus, the effective combination of various features to achieve optimal search performance becomes a very important research issue. Inspired by this observation, a multimodal paradigm for search has been developed and demonstrated promising performance in a series of domain applications [2], [21], [22], [38], [47], [55]. But for images in real world, different modalities do not appear in isolation. Instead, they demonstrate strong correlation and naturally interact with each other at different levels (from low-level visual features to high-level semantics). For example, there exists certain underlying association between textual and visual modalities. Through exploring and analyzing their association, certain patterns could be identified and applied for improving search accuracy. Based on this principle, the canonical correlation analysis (CCA) technique [14] has been applied to cross-modality image search. One typical example is the study presented in [35]. Rasiwasia *et al.* [35] investigated two important research questions: 1) whether modeling correlations effectively between textual components and visual content components can boost overall search performance and 2) whether the modeling can be more effective in feature spaces with semantic abstract. The empirical results show that the search systems accounting for cross-modality correlations and high-level abstraction demonstrate a promising performance. Although this method realizes the goal of querying image content through textual keywords, the learned statistical model only catches the latent relationship between text and visual components. The semantic gap between image content and semantic concepts remains unsolved.

## III. Our Approach

This section presents details of the proposed image search scheme. We first give a brief overview of the scheme in Section III-A. Then, we introduce the key technical component, namely the compact semantic embedding, in Section III-B. The algorithm for learning semantic embedding under dynamic environment is presented and discussed in Section III-C.

### A. Overview of the Approach

As illustrated in Fig. 1, the proposed scheme consists of two functional modules. In the first module, given a collection of training images with a fixed number of semantic concepts, we aim to create a unified semantic space $\mathcal{F}$, where each semantic concept $\alpha$ is represented as a topic prototype $\mathbf{p}_\alpha \in \mathcal{F}$ and each image $\mathbf{x}_i \in \mathcal{X}$ is represented by $\mathbf{z}_i \in \mathcal{F}$.

*Definition 1:* A topic prototype $\mathbf{p}_\alpha$ denotes the point which represents the semantic concept $\alpha$ in the latent semantic space $\mathcal{F}$.

As shown in Fig. 2(a), the purpose of learning a unified semantic space is to create a common representation of both images and semantic concepts, in which distance is able to indicate image concept membership. To facilitate this process, we derive semantic concept prototypes by exploring the semantic distance among the concepts. Based on the suitable prototype embedding of the semantic concepts, we further find an appropriate mapping $\mathbf{W}: \mathcal{X} \rightarrow \mathcal{F}$, which maps each image of one semantic concept as close as possible to its corresponding concept prototype. In this way, images of the same category are enforced to be located in a compact region around the concept prototype in the semantic space $\mathcal{F}$. This naturally leads to a cross-modality image search paradigm. Each of the concept keywords can still be used as query keyword, which inherits the simplicity of TBIR. Meanwhile, the embedded images can be represented by multidimensional vectors in the new feature space. For those images around a concept prototype, they can be regarded as indexed by the concept. In search process, these images can be ranked according to their distances to the concept prototype in the embedded semantic space.

As shown in Fig. 2(b), our proposed scheme can scale well when images and semantic concepts are dynamically changed. When a new semantic concept appears, instead of rebuilding the semantic space which is computational intensive, we only incorporate the appropriate concept prototype for the new concept and then update the mapping matrix $\mathbf{W}$ in an incremental way.

### B. Compact Semantic Mapping

In our image search scheme, input training images are assumed to be represented by a set of high-dimensional feature vectors: $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathcal{X}$ of dimension $d$. All the semantic concepts corresponding to the training images can be denoted by $y_1, y_2, \ldots, y_n \in \{1, \ldots, c\}$, where $c$ is the number of semantic concepts. In addition, the image indices are denoted by $i, j \in \{1, 2, \ldots, n\}$ and the concept indices are denoted by $\alpha, \beta \in \{1, 2, \ldots, c\}$. The semantic interactions among the semantic concepts can be represented by a semantic distance matrix $\mathbf{D} \in \mathcal{R}^{c \times c}$, where $\mathbf{D}_{\alpha,\beta} \geq 0$ estimates the semantic distance between concepts $\alpha$ and $\beta$. There are several existing approaches for estimating the distance between two textual keywords, such as Google distance [8] and WordNet distance [31]. Here, we adopt a Flickr distance method. It estimates $\mathbf{D}_{\alpha,\beta}$ as

$$\mathbf{D}_{\alpha,\beta} = \frac{\max(\log f(\alpha), \log f(\beta)) - \log f(\alpha, \beta)}{\log G - \min(\log f(\alpha), \log f(\beta))} \tag{1}$$

where $f(\alpha)$ and $f(\beta)$ are the numbers of images containing concepts $\alpha$ and $\beta$, respectively, $f(\alpha, \beta)$ is the number of images containing both $\alpha$ and $\beta$, and $G$ is the total number of images. All these numbers are obtained from Flickr website. It actually has the same formulation with Google distance [8] and the only difference is that the numbers $f(\alpha)$, $f(\beta)$, $f(\alpha, \beta)$, and $G$ are obtained from Flickr instead of Google. This is because Flickr is a dedicated image sharing website and thus it is expected to help to estimate a better distance between image concepts. This method has also been employed in [28], [29], and [49]. In Section IV-E, we will compare this method with Google distance and WordNet distance with an experiment.

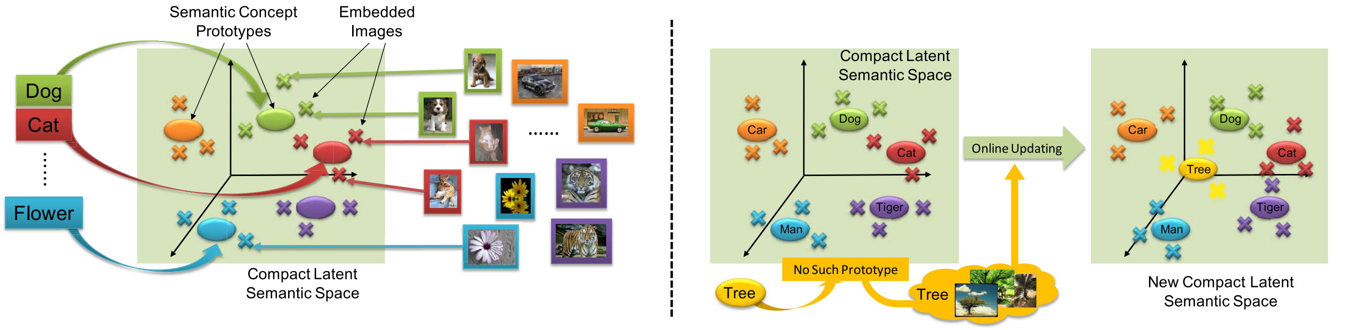
As summarized in Algorithm 1, the whole learning procedure consists of three major steps.

*1) Concept Prototype Embedding:* In the first step of our proposed compact semantic embedding process, the feature vectors belonging to individual semantic concepts are projected into an Euclidean vector space. In this paper, we derive the semantic concept prototypes $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_c \in \mathcal{R}^c$ based on the semantic distance matrix $\mathbf{D}$. It is worth noting that the dimension of the semantic space is set to $c$, i.e., the total number of semantic concepts residing in images. Such a straightforward setting prevents the MDS algorithm from complex parameter exploration and achieves good performance in our experimental results (see Section IV). To further simplify the notation, we group the columns together to generate a matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_c] \in \mathcal{R}^{c \times c}$ whose columns consist of all $c$ concept prototypes.

In this paper, we enforce the prototypes of similar semantic concepts to be close to each other than the other dissimilar concept prototypes. Here, the semantic distance matrix $\mathbf{D}$ is used as an estimation of semantic dissimilarity between the concepts. For any two given semantic concepts $\alpha$ and $\beta$, we aim to place the corresponding concept prototypes into the lower dimensional space such that the Euclidean distance $\|\mathbf{p}_\alpha - \mathbf{p}_\beta\|_2^2$ is as close to $\mathbf{D}_{\alpha\beta}$ as possible. Mathematically, we can formulate the given task as

$$\hat{\mathbf{P}} = \arg\min_{\mathbf{P}} \sum_{\alpha,\beta=1}^{c} \left( \|\mathbf{p}_\alpha - \mathbf{p}_\beta\|_2^2 - \left(\mathbf{D}_{\alpha\beta}\right)^2 \right)^2 \tag{2}$$

where $\hat{\mathbf{P}}$ denotes the optima of the objective function. By enforcing the two terms to be close to each other, the



(a) (b)

Fig. 2. Schematic layouts of the two functional modules in our approach. (a) At the first stage, we use the given collection of training images with a fixed number of semantic concepts to create a compact latent semantic space, where each concept is represented as a concept prototype and a mapping matrix is learned to map the images around the correct prototypes. (b) When there is a new semantic concept, it will be incorporated into the semantic space by inserting the corresponding concept prototype. Meanwhile, the mapping matrix is updated accordingly.

```
Algorithm 1 Image Search via Compact Semantic Embedding
1: Semantic Distance Matrix Learning: Given a set of semantic concepts with training images, learn the semantic distance matrix $\mathbf{D}$ by
       $\mathbf{D}_{\alpha,\beta} = \frac{\max\left(\log f(\alpha), \log f(\beta)\right) - \log f(\alpha,\beta)}{\log G - \min\left(\log f(\alpha), \log f(\beta)\right)}$
   where $\mathbf{D}_{\alpha,\beta}$ denotes the semantic distance between concepts $\alpha$ and $\beta$.
2: Concept Prototype Embedding: Derive the semantic concept prototypes $\mathbf{P}$ by
       $\hat{\mathbf{P}} = \arg\min_{\mathbf{P}} \sum_{\alpha,\beta=1}^{c} (\|\mathbf{p}_\alpha - \mathbf{p}_\beta\|_2^2 - (\mathbf{D}_{\alpha\beta})^2)^2.$
3: Image Content Mapping: Learn the mapping matrix $\mathbf{W}$ from $\hat{\mathbf{P}}$ and the labeled training images $\mathbf{X}$ by
       $\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_i \|\hat{\mathbf{p}}_{y_i} - \mathbf{W}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{W}\|_F^2.$
4: Indexing: Map all unlabeled images in the database with $\hat{\mathbf{W}}$. The mapped images and the concept prototypes form an inverted file structure, where the images indexed by a semantic concept are ranked with the compact semantic distances.
5: Online Search: At the stage of online query, for any given query keyword, the images around the corresponding concept prototype are directly returned.
```

obtained concept prototypes $\hat{\mathbf{P}}$ actually inherit the semantic relationships of the individual semantic concepts.

Since the distance matrix $\mathbf{D}$ quantifies the semantic divergence between the pairwise semantic concepts, we can solve (2) with the multidimensional scaling (MDS) algorithm [52]. To be detailed, denote $\overline{\mathbf{D}} = -1/2\mathbf{HDH}$, where the centering matrix is defined as $\mathbf{H} = \mathbf{I} - 1/c\mathbf{1}\mathbf{1}^\top$. Then, we perform eigenvector decomposition as $\overline{\mathbf{D}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$. Finally, the solution can be obtained by

$$\hat{\mathbf{P}} = \sqrt{\mathbf{\Lambda}}\mathbf{V} \tag{3}$$

where $\mathbf{\Lambda}$ is the eigenvalue diagonal matrix and $\mathbf{V}$ is the corresponding eignevector matrix. In the followings, we refer to the individual vectors residing in matrix $\hat{\mathbf{P}}$ as the semantic concept prototypes.

Note that the learned semantic concept prototypes matrix $\hat{\mathbf{P}}$ is actually independent of the image content $\{\mathbf{x}_i\}, i = 1, \ldots, n$. In the next section, we will further introduce an image embedding algorithm that maps the images into the obtained semantic space. In real applications, this procedure can be typically processed in an off-line manner which results in a set of prototypes corresponding to all the semantic concepts.

*2) Image Content Mapping:* Assume that a reliable embedding matrix $\hat{\mathbf{P}}$ of the semantic concept prototypes is ready, we now aim to find an appropriate mapping matrix $\mathbf{W} : \mathcal{X} \to \mathcal{F}$, which can map each input training image $\mathbf{x}_i$ with label $y_i$ as close as possible to its relevant concept prototype $\hat{\mathbf{p}}_{y_i}$. To achieve this target, a linear mapping $\mathbf{z}_i = \mathbf{W}\mathbf{x}_i$ can be found via the following mathematical formula:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_i \left\|\hat{\mathbf{p}}_{y_i} - \mathbf{W}\mathbf{x}_i\right\|_2^2 + \lambda \left\|\mathbf{W}\right\|_F^2 \tag{4}$$

where $\hat{\mathbf{W}}$ denotes the optima of the objective function, and $\lambda$ is the weight of the regularization of $\mathbf{W}$ which prevents potential overfitting. Although we only use a linear matrix $\mathbf{W}$ as the mapping matrix which is a common practice in cross-modality learning, we can also explore kernelization or other methods to seek for nonlinear mapping. Actually, the minimization in (4) is an instance of linear ridge regression, enforcing the images from a certain semantic category to locate compactly around the corresponding concept prototype. The above formulation has a closed-form solution as follows:

$$\hat{\mathbf{W}} = \left(\sum_{i=1}^{n} \hat{\mathbf{p}}_{y_i}\mathbf{x}_i^\top\right)\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top + \lambda\mathbf{I}\right)^{-1} \tag{5}$$

$$= \hat{\mathbf{P}}\mathbf{J}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1} \tag{6}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{J} \in \{0, 1\}^{c\times n}$, with $\mathbf{J}_{\alpha i} = 1$ if and only if $y_i = \alpha$.

*3) Indexing and Search Process:* After the training process, the learned embedding matrix $\hat{\mathbf{W}}$ can be applied to index unlabeled images in the repository. More specifically, given an unlabeled image $\mathbf{x}_t$, we first map it into the latent semantic space $\mathcal{F}$ and estimate its indexing keyword as the semantic concept with the closest prototype $\mathbf{p}_\alpha$ as

$$\hat{\alpha} = \arg\min_{\alpha} \left\| \mathbf{p}_\alpha - \hat{\mathbf{W}}\mathbf{x}_t \right\|_2^2. \tag{7}$$

This implies that the new image $\mathbf{x}_t$ is indexed by the semantic concept $\hat{\alpha}$.

After embedding all the unlabeled images in repository into the semantic space $\mathcal{F}$, we actually generate an inverted file structure whose indexing terms are the individual semantic concepts, and the images indexed by the semantic concept are ranked with their latent semantic distances to the concept prototype. At the online search stage, for a given query keyword, we can directly return the images around the corresponding concept prototype in the latent semantic space. In this way, we naturally obtain a keyword-based yet content aware cross-modality image search paradigm.

### *C. Online Updating*

In real-world applications, online image repositories can change dynamically and images belonging to new semantic concepts can emerge very frequently. In this case, the leant semantic mapping matrix $\hat{\mathbf{W}}$ may not be able to handle new semantic concepts effectively and thus degrade performance. The desirable solution for this problem is to incorporate these new concepts with the corresponding appropriate semantic concept prototypes. Then, the image mapping matrix $\hat{\mathbf{W}}$ also needs to be updated after the new semantic prototype terms are added into the latent semantic space. To achieve this target, we develop an online updating scheme that can: 1) generate the concept prototypes for the newly arrived semantic concepts and 2) adjust the image mapping matrix $\hat{\mathbf{W}}$ which incrementally handles the new concepts and the corresponding new training images appropriately. The whole procedure for the online updating can be illustrated as in Algorithm 2.

*1) Inserting New Semantic Concept Prototypes:* To achieve a more scalable image search scheme, we develop a set of novel methods to handle new semantic concept prototype insertion efficiently. When a new semantic concept $\alpha_{N+1}$ arrives, our system can incorporate the corresponding concept prototypes $\mathbf{p}_{N+1}$ into the latent semantic space $\mathcal{F}$ such that the newly generated concept prototype inherits the semantic distance between the pairwise semantic concepts. Intuitively, a natural solution to this task is to reimplement the concept prototype embedding procedure, in which the $N+1$ semantic concepts are considered together to learn the concept prototypes in a distance preserving manner. However, this approach is obviously infeasible since the learning procedure will lead to an extremely heavy computational cost, especially when the number of semantic concepts is huge under the Internet environment. Therefore, instead of reconstructing all semantic concept prototypes simultaneously, we fix the previously learned concept prototypes and only incorporate the new prototype $\mathbf{p}_{N+1}$. To achieve this target, the concept prototype $\mathbf{p}_{N+1}$ is learned in the semantic space $\mathcal{F}$ such that the embedding distances between $\mathbf{p}_{N+1}$ and the other $N$ prototypes maintain the semantic distances between the corresponding semantic concepts in the original semantic space. Mathematically, this task can be formulated as an optimization problem as follows:

**Algorithm 2** Online Updating

```
Given: An initial semantic space containing $N$ sematic concepts $\hat{\mathbf{P}}_N = \{\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_N\}$, an image mapping matrix $\hat{\mathbf{W}}_N$, and a newly arrived semantic concept $\alpha_{N+1}$ with a set of relevant images $\{\mathbf{x}'_i\}$, $i = 1, \ldots, q$.
Incorporating new semantic concept $\hat{\mathbf{p}}_{N+1}$ into $\hat{\mathbf{P}}_N$: Derive the $(N+1)$th concept prototype $\hat{\mathbf{p}}_{N+1}$ by
    $\hat{\mathbf{p}}_{N+1} = \arg\min_{\mathbf{p}} \sum_{i=1}^{N} (\|\mathbf{p} - \mathbf{p}_i\|_2^2 - (\mathbf{D}_{N+1,i})^2)^2$
    and directly incorporate it into the semantic space as $\hat{\mathbf{P}}_{N+1} = \hat{\mathbf{P}}_N \bigcup \hat{\mathbf{p}}_{N+1}$.
Updating the mapping matrix $\hat{\mathbf{W}}_N$: Incrementally update $\hat{\mathbf{W}}_N$ by
    $\mathbf{W}_{N+1} = (\mathbf{B} + \sum_{i=1}^{q} \mathbf{p}_c \mathbf{x}_i'^{\top})(\mathbf{Q} + \sum_{i=1}^{q} \mathbf{x}'_i \mathbf{x}_i'^{\top})^{-1}$.
Image mapping: Map all newly arrived images with $\hat{\mathbf{W}}_{N+1}$. Also, map the previous unlabeled images which have the same hashing binary codes with the newly arrived images with $\hat{\mathbf{W}}_{N+1}$.
```

$$\hat{\mathbf{p}}_{N+1} = \arg\min_{\mathbf{p}} \sum_{i=1}^{N} \left( \left\| \mathbf{p} - \hat{\mathbf{p}}_i \right\|_2^2 - \left( \mathbf{D}_{N+1,i} \right)^2 \right)^2 \tag{8}$$

where $\hat{\mathbf{p}}_i, i = 1, 2, \ldots, N$ denote the concept prototypes learned previously, and $\mathbf{D}_{N+1,i}$ denotes the semantic distance between the $(N+1)$th semantic concept and the $i$th semantic concept.

The objective function is a fourth-order function, and generally there does not exist a closed-form solution. In this paper, the conjugate gradient is applied to derive the minimization of the objective function in (8). Denote by $f$ the objective function in (8), i.e., $f = \sum_{i=1}^{N} (\|\mathbf{p} - \mathbf{p}_i\|_2^2 - (\mathbf{D}_{N+1,t})^2)^2$. Differentiating the objective function $f$ with respect to $\mathbf{p}$ yields a gradient rule

$$\frac{\partial f}{\partial \mathbf{p}} = 4 \left( \sum_{i=1}^{N} \left( (\mathbf{p} - \hat{\mathbf{p}}_i)^{\top} (\mathbf{p} - \hat{\mathbf{p}}_i) - \mathbf{D}_{N+1,i}^2 \right) \right) \left( \sum_{i=1}^{N} (\mathbf{p} - \hat{\mathbf{p}}_i) \right). \tag{9}$$

Based on the above gradient, the iterative gradient descent is applied to derive the optimal solution of (8). Denote by $l$ the index of iterations, the iterative gradient procedure updates the current solution $\mathbf{p}_{N+1}^{l}$ to $\mathbf{p}_{N+1}^{l+1}$ by the following rule:

$$\mathbf{p}_{N+1}^{l+1} = \mathbf{p}_{N+1}^{l} - \rho \frac{\partial f}{\partial \mathbf{p}} |_{\mathbf{p}=\mathbf{p}_{N+1}^{l}} \tag{10}$$

where $\rho$ is the learning rate and we set it to 0.1 throughout the experiments in this paper.

After the optimization of the above procedure, we actually obtain a new semantic concept prototype $\hat{\mathbf{p}}_{N+1}$ without updating the semantic concept prototypes learned previously. Once a new concept appears, a lightweight learning procedure is triggered out, which incrementally updates the semantic concept prototypes efficiently.

We can also consider changing the dimension of latent semantic subspace, but it will need to reimplement the whole MDS and image mapping process, even for those concepts that are already in database. One approach only adds some new dimensions for concept prototypes, such as increasing the dimension of concept prototype from $c$ to $c+r$ and keeping the original $c$ dimensions unchanged. But how to efficiently learn the $r$ new dimensions is a problem and it is beyond the scope of this paper. Therefore, in this paper we keep the dimension of latent semantic subspace unchanged.

*2) Updating Image Mapping Matrix* $\mathbf{W}$*:* Even after the insertion of new semantic concept prototypes, a mechanism is needed for updating the image embedding matrix $\hat{\mathbf{W}}$ such that the newly inserted semantic concepts and the corresponding images can have the compact property in the latent semantic space $\mathcal{F}$.

In contrast to the offline embedding matrix learning procedure introduced above, the online updating procedure assumes that the semantic concepts and the corresponding training images are received sequentially. That is, we assume that we have constructed a semantic space containing $N$ semantic concepts, and there already exists an embedding projection matrix $\hat{\mathbf{W}}_N$. Meanwhile, we also assume that a new concept $c_{N+1}$ and its corresponding training images, denoted as $\mathcal{X}_{N+1} = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_q\}$, are received, where $q$ denotes the number of images in $\mathcal{X}_{N+1}$. In this setting, we aim to update the image embedding matrix from $\hat{\mathbf{W}}_N$ to $\hat{\mathbf{W}}_{N+1}$.

To derive the updated embedding matrix $\mathbf{W}_{N+1}$ after new semantic concept $c_{N+1}$ and the corresponding training images $\mathcal{X}_{N+1}$ arrive, we proceed by exactly solving for the updated embedding projection matrix $\mathbf{W}_{N+1}$ based on the property in the closed-form solution of the embedding matrix $\mathbf{W}$. More specifically, from (5), we can observe that the closed-form solution of the embedding matrix $\mathbf{W}$ can be expressed as $(\sum_{i=1}^{n} \mathbf{p}_{y_i}\mathbf{x}_i^\top)(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top + \lambda\mathbf{I})^{-1}$ where the two summation terms involved are performed on the training images $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ witnessed until now. Actually, it potentially provides an efficient solution for us to derive the incremental solution. More specifically, if we keep the two matrices $\mathbf{B} = \sum_{i=1}^{n} \hat{\mathbf{p}}_{y_i}\mathbf{x}_i^\top$ and $\mathbf{Q} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$, each of which corresponds to the summation of all the training samples witnessed until now, then the updating procedure can be calculated as follows:

$$\hat{\mathbf{W}}_{N+1} = \left(\mathbf{B} + \sum_{i=1}^{q} \hat{\mathbf{p}}_{N+1}\mathbf{x}_i'^\top\right)\left(\mathbf{Q} + \sum_{i=1}^{q} \mathbf{x}_i'\mathbf{x}_i'^\top\right)^{-1} \quad (11)$$

which indicates that, after each updating procedure, we only need to store the matrices $\mathbf{B}$ and $\mathbf{Q}$. When a new concept and its corresponding training images are available, we only need to add the additional training samples into $\mathbf{B}$ and $\mathbf{Q}$ and perform a matrix inversion calculation, of which the computational cost only relies on the dimension of image features. Typically, the dimension of image feature will be in a considerate scale, and thus the computation can be performed efficiently.

Based on $\hat{\mathbf{W}}_{N+1}$, we need to remap the previous unlabeled images along with the newly arrived images of concept $c_{N+1}$ into $\mathcal{F}$ such that the compact semantic embedding property can be well maintained. However, directly mapping all the images with $\hat{\mathbf{W}}_{N+1}$ is infeasible since it takes linear time with respect to the number of images. Instead, we resort to locality semantic hashing (LSH) to perform a fast remapping. Specifically, suppose we have produced a hashing table for all the previous unlabeled images $\{\mathbf{x}_1, \ldots, \mathbf{x}_u\}$ in the semantic space $\mathcal{F}$. The length of binary codes is fixed to be 64 for LSH. When we get the new $\hat{\mathbf{W}}_{N+1}$, for each new image $\mathbf{x}_i^{'}$, we map it into 64 binary codes with the LSH hashing vectors. Finally, only the previous unlabeled images that have the same binary codes are considered as the most similar images with respect to $\mathbf{x}_i^{'}$. Comparing with other images, these images have closer relationships with the newly arrived semantic concept. Therefore, we simply only remap these images into $\mathcal{F}$, which dynamically updates the image distributions in $\mathcal{F}$. This strategy thus significantly reduces computational cost.

### D. Scalability Analysis

As shown above, our online updating scheme is able to scale well for dynamic online image repositories. The whole system does not need to be rebuilt after data objects from new semantic classes are inserted. It can save overall maintenance cost, especially for large-scale and dynamic data. The key idea can be summarized as follows.

1) When a new semantic concept arrives, we only need to learn its corresponding concept prototype while simply keeping the previous concept prototypes unchanged. The learning procedure is based on an iterative gradient descent, which is a quite fast process. For example, in our experiment on the NUS-WIDE dataset (see Section IV-C), the learning of the 41st concept prototype based on the existing 40 concept prototypes takes only 0.35 s on the MATLAB platform of an Intel XeonX5450 workstation with 3.0 GHz CPU and 16 GB memory. On the other hand, simultaneously learning all the 41 concept prototypes with the optimization strategy in (2) takes 2.58 s. Therefore, our method is very efficient in terms of computational cost.
2) For newly arrived images, we only need to map them with the updated $\mathbf{W}$ matrix, which takes linear time with respect to the number of new images. For previous unlabeled images, the LSH-based method only selects those that have the same binary codes with the newly arrived images. In this paper, the updating takes only 0.0032 s in average for each new image, which achieves thousands times speedup.

### E. Discussion

Up to now, we have introduced the semantic embedding approach and how to search images with the embedded concepts. We can also use the semantic space to facilitate image search with free text queries. Actually, query-to-concept mapping is already a widely investigated topic in multimedia research field [23], [32], [37], [43]. There are many different methods available. For example, for each existing concept, we can find its textual explanation from a dictionary. Then, we can calculate the similarity of a free text query and the textual explanation. In this way, we can map the query to the weighted combination of several existing concepts. Then, based on the semantic space, we can easily rank images for the given query. In Section IV-D, we will provide several experimental results with queries that contain multiple concepts.

## IV. Experiments

In this section, we systematically evaluate the effectiveness of our proposed solution, referred to as compact semantic mapping (CSM), by conducting experiments over three widely used datasets, namely, Caltech-101 [24], NUS-WIDE [7], and MIRFlickr dataset [19]. More specifically, the following three sub-tasks are exploited for performance evaluation.

1) Cross-modality image search via CSM. In this sub-task, we map semantic concepts and images into a unified latent semantic space, and then we use textual queries to perform cross-modality image search.
2) Online updating via incremental learning. The main purpose of this sub-task is to validate the effectiveness of the proposed online learning procedure in updating latent semantic space in order to handle new semantic concepts.
3) Image search with multiple query terms. Notably, multiple-query-term image search paradigm tries to find images that simultaneously contain multiple semantic concepts, which is a nontrivial but rarely investigated task. Since our proposed approach attacks image search simply based on the distances between semantic concepts and image contents, it provides a natural way to accomplish the task as the images residing closely to the multiple query terms can be deemed as search results. Given the fact that multiple-query-term image search is challenging for image search engines, the performance on this sub-task can well validate the generalization capability of the cross-modality image search approach.

For the purpose of performance comparison, the following four baseline image retrieval methods are implemented and their details can be found as follows.

1) *Content-Based Image Retrieval (CBIR):* This inputs a query image and searches over the entire image database based on low-level visual features.
2) *Semantic-Based Image Retrieval (SBIR):* For each semantic category, a set of positive and negative images are selected to train a support vector machine (SVM) [9] classifier with radius basis function kernel. Then, all the left images in the database can be ranked according to the prediction scores from the SVM classifier on the individual images. The libSVM toolkit [4] is utilized to implement the classification, and the parameters are tuned by a 10-fold cross-validation process.
3) *CCA-Based Cross-Modality Image Search [14]:* This is actually the most popular method for cross-modality image search. Given the image space $\mathcal{R}^I$ and the semantic concept space $\mathcal{R}^T$, CCA learns two mapping matrices $\mathbf{W}_{I\times r}$ and $\mathbf{W}_{T\times r}$ along which the two kinds of modality are maximally correlated. This results in a unified $r$-dimensional representation of both modalities in a common subspace for cross-modality search, and the images are ranked according to their distances to the semantic concepts in the common space. Regarding the parameter settings for this method, we tune the transformation dimensionality $r$ to its best value in our experiments.
4) *Compact Embedding With Random Matrix (CERM):* To verify the importance of the semantic correlation, we generate the initial semantic matrix $\mathbf{D}$ with random similarities and then perform cross-modality search based on our proposed approach.

To quantitatively measure the algorithmic performance, we utilize the following three criteria.

1) *Mean Precision @ n*: The proportion of relevant images in the top $n$ ranked images. We calculate precision@n for different queries and further calculate their mean value as the final evaluation metric.
2) *Mean Recall @ n:* The proportion of successfully retrieved relevant images when $n$ results are returned. We calculate the average value over all queries and report the obtained average value.
3) *Mean Average Precision (MAP) @ n:* AP measures the ranking quality of a whole list. Since it is an approximation of the area under a precision-recall curve, AP is commonly considered as a good combination of precision and recall [17]. Given a ranked list containing $n$ images, the AP value is calculated as

$$\frac{1}{R}\sum_{i=1}^{n}\frac{R_i}{i}\delta_i \tag{12}$$

where $R$ is the number of relevant images in the list, $R_i$ the number of relevant images in the top $i$ ranked ones, and $\delta_i = 1$ if the $i$th image is relevant and 0 otherwise. To evaluate the overall performance, we report MAP, the mean value of the AP values over all queries.

### A. Image Datasets

In our experiments, three publicly available benchmark image datasets, Caltech-101, NUS-WIDE, and MIRFlickr, are used to evaluate the performance of different methods. It is worth noting that all these datasets are collected from Internet, e.g., from Google and Flickr, and thus they precisely reflect the properties of the real-world web images.

*Caltech-101 Dataset [12]:* This dataset has been frequently used in image search tasks [18], [44]. It contains around 10 000 images from 101 object categories and 1 background category. Here, we directly use the 101 semantic concepts (categories) as the textual query keywords to perform image search.

*NUS-WIDE Dataset [7]:* This dataset is a more challenging collection of real-world social images from Flickr. It contains 269 648 images from 81 semantic categories, with a total number of 5018 unique tags, and the average number of tags for each image is 2. It is worth noting that there are some subordinate tags (such as "water" and "lake") and abstract tags (such as "military" and "earthquake") contained in this dataset. Therefore, search on this large-scale web image dataset is quite challenging for all the state-of-the-art methods. We use the 81 semantic category as query keywords to perform retrieval.[1]

*MIRFlickr Dataset [19]:* The MIRFlickr-25k collection consists of 25 000 images collected from Flickr. It contains 1386 tags which occur in at least 20 images. The average number of tags per image is 8.94. Here, we use the 25 most common tags whose ground truth annotations with respect to all the images have been provided by the MIRFlickr dataset: "sky," "water," "portrait," "night," "nature," "sunset," "clouds," "flower," "beach," "landscape," "street," "dog," "architecture," "graffiti," "tree," "people," "city," "sea," "sun," "girl," "snow," "food," "bird," "sign," and "car."

For low-level visual representation, we combine the following two different low-level features.

1) *Color Features:* We use the 64-dimensional color histogram and 144-dimensional color correlogram as image color features. More specifically, the color histogram serves as an effective representation for the color content of an image, whereas the color correlogram can be used to characterize color distributions and spatial correlations.
2) *Local Features:* Here, we use 128-dimensional SIFT descriptor for describing the local information of each image and then quantize them into a vocabulary of 1000 visual words generated by the *k*-means method. Then, each SIFT feature point is mapped to an integer (visual word index) between 1 and 1000, leading to the BoW image representation. Finally, we obtain a $64 + 144 + 1000 = 1208$ dimensional feature vector to describe each image.

### B. Exp-I: Cross-Modality Image Search

*1) Experimental Configuration:* In this experiment, all the three datasets are used for the evaluation. For Caltech-101, we randomly select 15 samples from each category to form a training set. So, there are $15 \times 101 = 1515$ training images in total. For the proposed CSM method, these training images are used to learn the semantic embedding matrix and all the left samples in this database are projected into the latent semantic space, which are further used for the consequent image search experiments. For the CBIR method, the images in the training set are used as the queries, resulting in 1515 query images in total. To realize SBIR, for each semantic category in Caltech-101, we use the selected 15 training images from this category as positive samples and all the left samples in the training set as negative samples to train an SVM model. The predicted scores obtained by the SVM model are then used to measure the relevance levels of the database images with respect to the query. Finally, for CCA, we also use the selected training set to learn two mapping matrices that maximize the correlation between the visual contents and semantic labels. Then, the visual contents and semantic labels are mapped into a common CCA space by the corresponding mapping matrices. For the NUS-WIDE and MIRFlickr datasets, the numbers of training samples are both set to 50 for each category.

TABLE I
CROSS-MODALITY SEARCH RESULTS ON THE NUS-WIDE DATASET WITH DIFFERENT CONCEPT DISTANCES

| Algorithm # | Google Distance | WordNet Distance | Employed Method |
|---|---|---|---|
| MAP@100 | 0.312 | 0.301 | **0.318** |
| MAP@500 | 0.227 | 0.208 | **0.229** |
| MAP@1000 | 0.175 | 0.16 | **0.183** |

*2) Results and Analysis:* The comparisons between the baseline methods and the proposed CSM method on Caltech-101, NUS-WIDE, and MIRFlickr are shown in Fig. 3(a)–(c), (d)–(f), and (g)–(i), respectively. From the results, we have the following observations.

1) The CSM method achieves much better performance than CBIR. The precision, recall, and MAP curves all clearly validate the effectiveness of CSM. The main reason is that CSM method well captures the semantic correlation between textual queries and image contents.
2) The CSM method outperforms the SBIR method, which is substantially owing to the exploration of the intermediate representation of both textual and visual modalities in the CSM method.
3) The CSM method clearly outperforms the CCA method. One explanation is that CCA method does not explicitly enforce that the images related to the same keyword should be close in the projected common CCA space.
4) The CSM method produces much better results than the CERM method. This clearly indicates that, besides the dimensionality reduction process, the semantic similarity also contributes significantly to the overall performance, which confirms the advantage of exploring the correlation of semantic concepts.

Fig. 4 illustrates the exemplary cross-modality image search results for queries "waterfall" and "horse" produced by the CSM method.

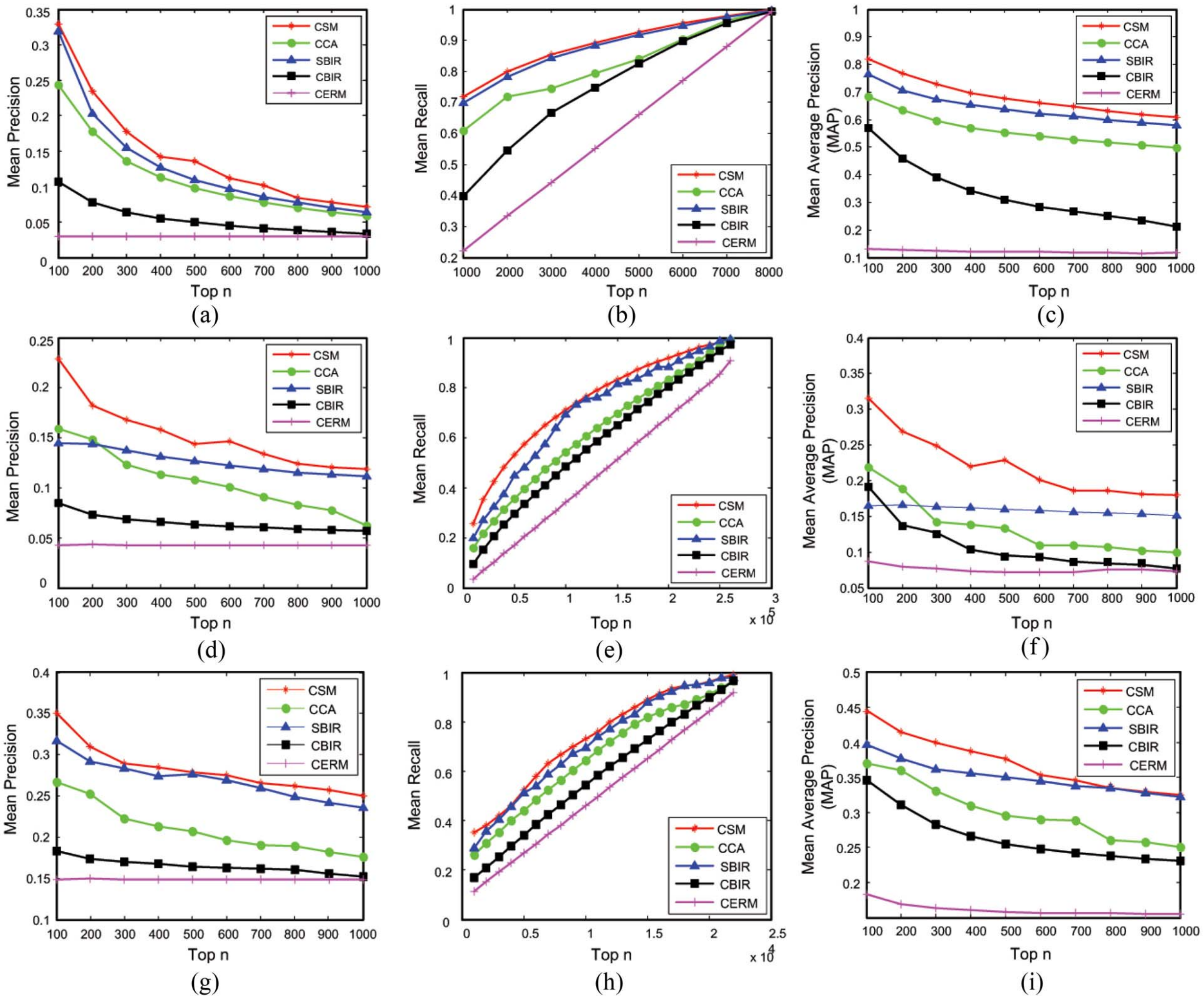
*3) On the Impact of Concept Distance Estimation:* Here, we also compare different methods for the estimation of concept distances $\mathbf{D}$ [see (1)]. We compare our approach against Google distance and WordNet distance. That means, we employ different methods to estimate $\mathbf{D}_{\alpha,\beta}$ and then observe the cross-modality search results. Here, we tabulate the MAP@100, MAP@500, and MAP@1000 results on the NUS-WIDE dataset with different methods in Table I.

From the table, we can see that the WordNet distance has a relatively large performance gap in comparison with the other two methods. The performances of the employed method and Google distance are close, but the employed method performs slightly better.

### C. Exp-II: Online Updating

As aforementioned, an online learning procedure is provided in the CSM method for dynamically updating the model

[1] Actually, we can use all the 5018 tags as query keywords to perform image search. But the lack of ground truth annotations brings difficulty for evaluation. Therefore, we only report the search results on the 81 semantic categories (tags) for which ground truth annotations have been provided in the NUS-WIDE dataset.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3. Mean precision, mean recall, and MAP curves at different returned depths (Top n) on the (a)–(c) Caltech-101 dataset, (d)–(f) NUS-WIDE dataset, and (g)–(i) MIRFlickr dataset, respectively, for Exp-I: cross-modality image search.

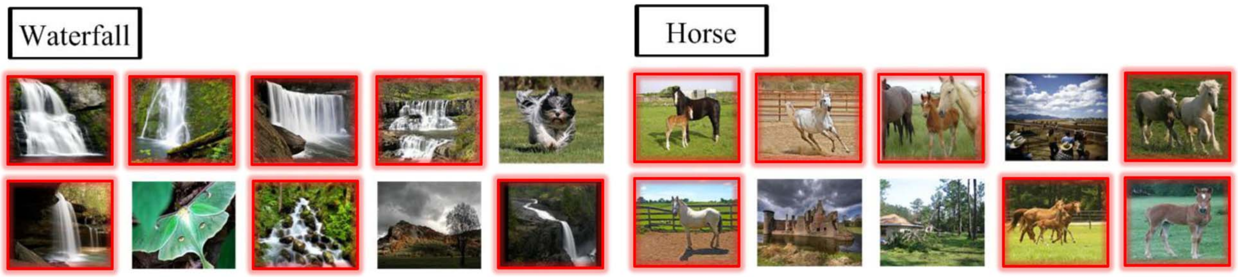
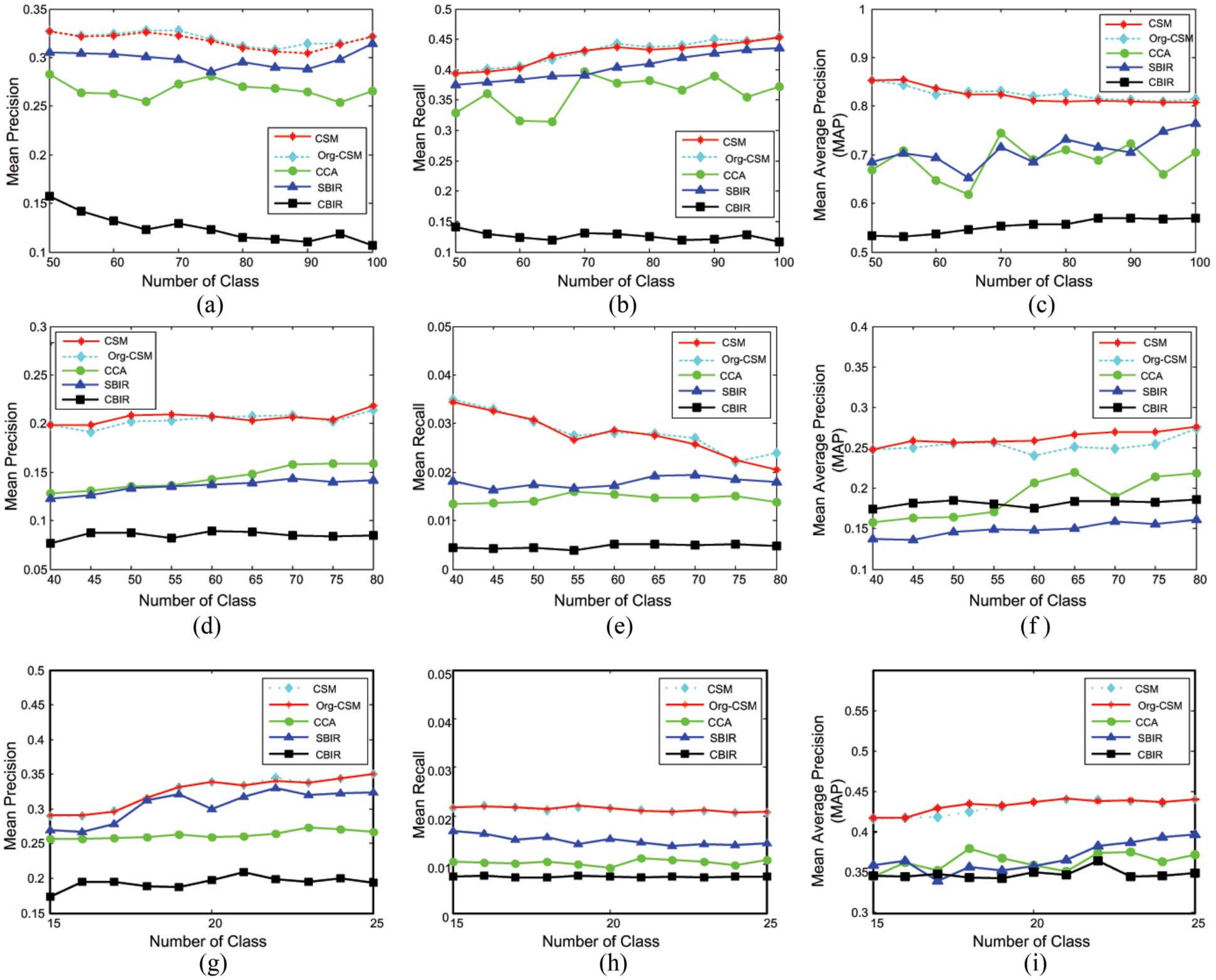


Fig. 4. Top ten results of the CSM method for queries "waterfall" and "horse." The relevant images are highlighted with red boxes.

when new semantic concepts are involved. In this subsection, we evaluate the performance of the proposed online updating procedure.

*1) Experimental Configuration:* For the Caltech-101 dataset, we randomly choose 50 semantic categories to build the semantic prototype set **P** and mapping matrix **W**, in which the implementation details are the same as in the first experiment in Section IV-B. So, the number of training samples is $15 * 50 = 750$. In this case, we can treat the generated 50 semantic concept prototypes and the corresponding semantic mapping matrix as the initial component of the latent semantic space, based on which the proposed online updating procedure can be further implemented. Here, we assume that the semantic concepts arrive one-by-one. Each time when a new semantic concept is involved, 15 images randomly selected from this semantic category are employed as the training images of this category and added to the training set. Based on the new semantic concept and the corresponding training images, we employ our online updating procedure to incorporate the new concept prototype and update the semantic



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 5. Mean precision, mean recall, and MAP curves (Top 100) for the online updating task (Exp-II) on (a)–(c) Caltech-101 dataset, (d)–(f) NUS-WIDE dataset, and (g)–(i) MIRFlickr dataset, respectively.

mapping matrix. For CBIR, SBIR, and CCA, the implementation details are also the same as in the first experiment in Section IV-B. Similarly, on the NUS-WIDE dataset, we randomly select 50 training samples from each of the 40 semantic categories and utilize all the training images to construct the latent semantic space which consists of 40 semantic concepts. Furthermore, when a new semantic concept is involved, additional 50 training samples of this semantic category are randomly selected from the dataset, which are then used for the consequent online updating procedure. For MIRFlickr dataset, the beginning number of semantic categories is 15, and the left process is the same with that for the NUS-WIDE dataset. In addition, we also employ the Original CSM method (termed as Org-CSM) which, instead of performing incremental learning, directly works on all concerned semantic categories and corresponding training images.

*2) Results and Analysis:* The experimental results on different datasets are shown in Fig. 5(a)–(c) (Caltech-101), (d)–(f) (NUS-WIDE), and (g)–(i) (MIRFlickr), respectively. From these results, we have the following observations.

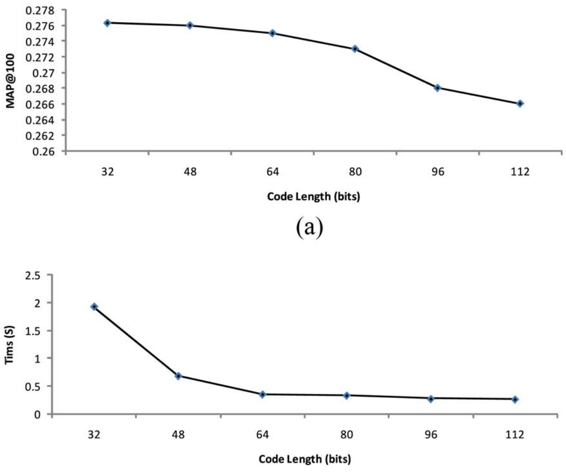
1) The proposed CSM with online updating procedure clearly outperforms CBIR, SBIR, and CCA, which well validates the effectiveness of the online updating strategy.
2) The performance of CSM with online updating procedure is quite close to Org-CSM, which clearly shows that CSM with online updating procedure is well generalizable.

TABLE II
AVERAGE RUNNING TIMES (SECOND) OF DIFFERENT METHODS FOR UPDATING A NEW CONCEPT IN ONLINE LEARNING TASK ON NUS-WIDE DATASET

| Algorithm # | CSM | CBIR | SBIR | CCA |
|---|---|---|---|---|
| Time (s) | **0.35** | 10.49 | 29.96 | 8.27 |

Besides performance superiority, our proposed CSM with online updating method also possesses the advantages in time complexity. Table II lists the running times of different methods in the task of online learning on NUS-WIDE dataset. As can be seen, our proposed method can update the semantic space and the mapping matrix for a new semantic concept and the corresponding images within 0.35 s in average, which is much faster than the other methods. This further validates the scalability and efficiency of our proposed algorithm.

*3) On the Impact of LSH Code Length:* In online updating, we have used LSH to select images for mapping with the newly obtained projection matrix. Here, we also carry an



(a)

(b)

Fig. 6. Variation of performance and time cost with different number of hashing bits in the online updating task. (a) MAP@100 scores on the NUS-WIDE dataset after 41 new concepts are all processed. (b) Time costs for processing a new concept in average.

experiment to test the influence of the code length in our LSH. We vary the code length from 32 to 112 bits and observe the final MAP@100 results on the NUS-WIDE dataset after 41 concepts are all processed by our approach. Fig. 6 illustrates the results. We can see that, generally the increasing of code length reduces time cost, but the performance will also degrade. This is because, with a large number of bits, less images will be remapped, but it may miss some images that are relevant to the new concept and should be remapped. We choose 64 bits as this length well compromises performance and computational cost.

### *D. Exp-III: Multiple-Query-Term Image Search*

In this subsection, we evaluate the effectiveness of the proposed CSM method under the setting where a query contains multiple keyword terms. Notably, multiple-query-term image search usually requires complex structural learning models to facilitate such function [27], [34]. But our proposed scheme can be directly applied to such scenario as the image subset close enough to the prototypes of the query terms in the unified space can be returned as search results. This provides a simple paradigm for multiple-query-term image search and further verifies the advantages of the proposed algorithm.

*1) Experimental Setup:* We conduct experiments on the NUS-WIDE dataset.[2] To be detailed, we select ten multiple-terms queries, each of which is composed of two different keywords.[3] To properly choose the two-term queries, we count the TF-IDF [53] of each tag pair in the image set and select the top ten tag pairs as the query terms. In this way, the combination of each tag pair corresponds to a reasonable portion of the images in the dataset. The detailed list of the two-term queries is shown in Table III. In the online search stage, when a two-term query is input, we can obtain two image ranking lists from the latent semantic space. Each of the ranking lists is indexed by one of the concepts and the corresponding images are ranked according to their semantic closeness to the concept. To obtain the desired search results, the images are ranked based on the sum of their square distances to both of these two concept prototypes.[4]

TABLE III
MULTIPLE-TERM QUERIES USED IN EXP-III

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Coral | Cow | Tree | Horse | Sand | Grass | Bear | Railroad | Castle | Water |
| Fish | Grass | Waterfall | Water | Tiger | House | Tree | Tower | Sand | Grass |

TABLE IV
MEAN PRECISION AND MAP RESULTS (TOP 100) FOR THE MULTIPLE-QUERY-TERM IMAGE SEARCH SUB-TASK

| Algorithm # | CSM | CBIR | SBIR | CCA |
|---|---|---|---|---|
| Precision | **0.2142** | 0.0431 | 0.1625 | 0.0881 |
| MAP | **0.1925** | 0.0513 | 0.1456 | 0.1032 |

For the other three baseline methods, in order to realize multiple-query-term image search, we implement each of them in the following manner.

1) *CBIR:* We select 50 query images, and the images in database are ranked based on the sum of their square distances to each of these two categories.
2) *SBIR:* For each semantic concept, we train an SVM model. We rank the images in database according to the sum of the predicted scores of the two models.
3) *CCA:* After mapping tags and images into a CCA space, we rank the images based on the sum of their square distances to each of these two tags. For each algorithm, the implementation settings are the same with those used in the experiment in Section IV-B.

*2) Results and Analysis:* Table IV shows the results generated by different methods in the sub-task of multiple-query-term image search. We can find that the proposed CSM method can respond complex queries with better effectiveness than the other methods. For example, in comparison with CCA, CSM achieves 59.9% and 46.3% improvements in terms of Precision and MAP, respectively. It can be attributed to the fact that the proposed CSM method well organizes the semantic concepts and the images in a unified compact semantic space, whereby the complex queries can be simply analyzed according to the spatial layout between the individual query terms and the images.

Fig. 7 further shows the top search results of some example multiple-term queries, which well illustrates the effectiveness of the CSM method in multiple-query-term image search.

[2] For the Caltech-101 dataset, the images solely contain a single semantic concept. For the MIRFlickr dataset, we observed the fact that, for many multiple-query terms, there are very few relevant images in the returned results. Therefore, we only conduct experiments on the largest one among the three datasets, namely the NUS-WIDE dataset.

[3] It is worth noting that our proposed method can support complex queries that contain more than two terms. However, to simplify the implementation, here we only use the query with two terms.

[4] Semantic concepts that simultaneously appear in one image typically have strong semantic correlation, and thus their concept prototypes should be close to each other. In this way, an image may locate near to multiple concept prototypes.

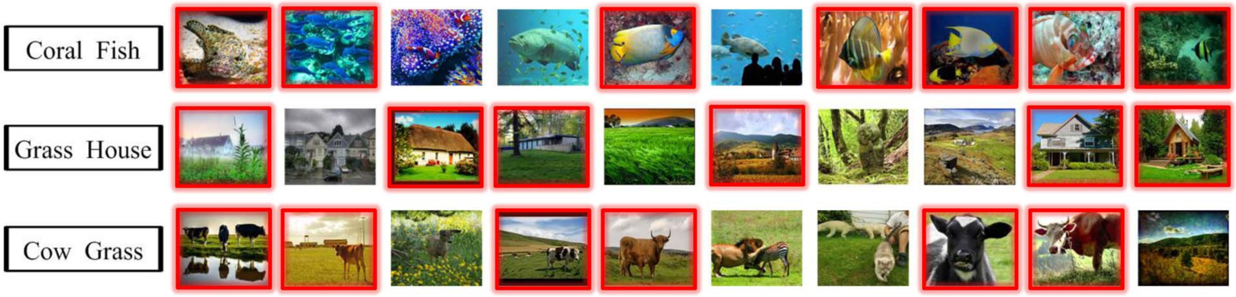


Fig. 7. Top ten results generated by the CSM method for example multiple-term queries "Coral Fish," "Grass House," and "Cow Grass." The relevant images are highlighted using red boxes.

### E. Exp-IV: Example-Based Image Search

Finally, we also conduct a simple experiment to test the effectiveness of the learned semantic space in conventional CBIR, i.e., example-based image search. Although our approach is designed for cross-modality image search, the learned semantic space can also be employed in example-based image search, as the distance of two images can be estimated in the learned space.

*1) Experimental Setup:* We conduct experiments on the NUS-WIDE dataset. Here, 100 images are randomly chosen as queries from the whole dataset and the other images are ranked according to their distances to the query images. The first ten returned images are labeled as relevant or irrelevant to each query by human (there are three labelers involved in the experiments and they label relevance according to the semantic relatedness of the returned image and the query). We estimate mean precision and MAP scores use them to evaluate the performance of different methods.

We compare the following three methods.

1) Image search with Euclidean distance (denoted as "Euclidean"). Here, we directly rank the images according to their Euclidean distances to the query image by comparing visual feature vectors.
2) Image search with distance metric learning (denoted as "metric learning"). We first perform a distance metric learning method and then rank images based on the learned distance metric. Here, we employ the relevance component analysis algorithm. More specifically, we select 100 images for each of the 81 semantic categories, and so there are 8100 samples in 81 classes, whereby we then perform RCA to learn a distance metric.
3) Image search with our semantic space (denoted as "proposed"). We rank the images according to their Euclidean distances to the query image in our semantic space.

*2) Results and Analysis:* Table V demonstrates the results. From the results we can see that, although the semantic space learning approach is designed for cross-modality image search, it can also facilitate example-based image search to some extent. By employing the learned semantic space, image search can yield much better performance than simply comparing the visual features of images with Euclidean distance. Its performance is also better than firstly learning a distance metric with RCA and then performing image search with the learned distance metric.

TABLE V
MEAN PRECISION AND MAP RESULTS (TOP 100) FOR DIFFERENT EXAMPLE-BASED IMAGE SEARCH METHODS

| Algorithm # | Euclidean | metric learning | proposed |
|---|---|---|---|
| Mean Precision | 0.48 | 0.57 | **0.62** |
| MAP | 0.56 | 0.61 | **0.68** |

## V. CONCLUSION

In this paper, we present a scalable cross-modality image search scheme based on a CSM method. The proposed method narrows semantic gap and scales well to dynamically increasing semantic concepts and images. To achieve this, we derive a method to construct a linear mapping matrix which maps images into a semantic space such that the images and the corresponding semantic concepts are as close as possible. In this way, mapped images around a concept prototype can be viewed as the retrieved results of the corresponding query. To deal with dynamically increasing semantic concepts, we also introduce an online learning procedure to incorporate new semantic concepts and update the mapping matrix. Extensive experiments on three widely-used benchmark datasets have well demonstrated the effectiveness of the proposed scheme.

## REFERENCES

[1] K. Barnard *et al.*, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Jan. 2003.

[2] J. C. Caicedo, J. G. Moreno, E. A. Niño, and F. A. González, "Combining visual features and text data for medical image retrieval using latent semantic kernels," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 359–366.

[3] G. Carneiro, A. B. C. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27.1–27.27, Apr. 2011.

[5] E. Y. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.

[6] L. Chen, D. Xu, I. W. Tsang, and X. Li, "Spectral embedded hashing for scalable image retrieval," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 2168–2216, Jul. 2013.

[7] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 8–10.

[8] R. Cilibrasi and P. M. B. Vitányi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, 1st ed. New York, NY, USA: Cambridge Univ. Press, 2000,

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *J. ACM Comput. Surv.*, vol. 40, no. 2, p. 2007, Apr. 2008.

[11] F. Dornaika and A. Bosaghzadeh, "Exponential local discriminant embedding and its application to face recognition," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 921–934, Jun. 2013.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPR)*, 2004, p. 178.

[13] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2004, pp. 1002–1009.

[14] D. R. Hardoon, S. R. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[15] R. Hong *et al.*, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[16] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[17] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 35–44.

[18] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3376–3383.

[19] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.

[20] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 119–126.

[21] Y. Jing and S. Baluja, "Pagerank for product image search," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 307–316.

[22] L. Kennedy, S. F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proc. IEEE*, vol. 96, no. 4, pp. 567–588, Apr. 2008.

[23] A. Ksibi, A. B. Ammar, and C. B. Amar, "Enhanced context-based query-to-concept mapping in social image retrieval," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Veszprem, Hungary, 2003, pp. 85–89.

[24] F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Und.*, vol. 106, no. 1, pp. 59–70, 2007.

[25] H. Li, Y. Wei, L. Li, and C. L. P. Chen, "Hierarchical feature extraction with local neural response for image recognition," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 412–424, Apr. 2013.

[26] P. Li, M. Wang, J. Cheng, C. Xu, and H. Lu, "Spectral hashing with semantically consistent graph for image indexing," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 141–152, Jan. 2013.

[27] X. Li, C. G. Snoek, M. Worring, and A. W. Smeulders, "Harvesting social images for bi-concept search," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1091–1104, Aug. 2012.

[28] D. Liu, X. S. Hua, L. Yang, M. Wang, and H. J. Zhang, "Tag ranking," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 351–360.

[29] D. Liu, S. Yan, X. S. Hua, and H. J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug. 2011.

[30] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[31] G. A. Miller, "WordNet: A lexical database for English," *ACM Commun.*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[32] S. Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Proc. Int. Conf. Content-Based Image Video Retrieval*, Tempe, AZ, USA, 2006, pp. 143–152.

[33] L. Nie, M. Wang, Y. Gao, Z. J. Zha, and T. S. Chua, "Beyond text QA: Multimedia answer generation by harvesting web information," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 426–441, Feb. 2013.

[34] L. Nie, S. Yan, M. Wang, R. Hong, and T. S. Chua, "Harvesting visual concepts for image search with complex queries," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 59–68.

[35] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia*, 2010, pp. 251–260.

[36] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE MultiMedia*, vol. 4, no. 3, pp. 12–20, Jul./Sep. 1997.

[37] C. G. Snoek *et al.*, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, Aug. 2007.

[38] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402. [Online]. Available: http://doi.acm.org/10.1145/1101149.1101236

[39] J. Song, Y. Yang, X. Li, Z. Huang, and Y. Yang, "Robust hashing with local models for approximate similarity search," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1225–1236, Jul. 2014.

[40] R. K. Srihari, "Automatic indexing and content-based retrieval of captioned images," *IEEE Comput.*, vol. 28, no. 9, pp. 49–56, Sep. 1995.

[41] D. Tao, L. Jin, Z. Yang, and X. Li, "Rank preserving sparse learning for Kinect based scene classification," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1406–1417, Oct. 2013.

[42] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2322–2336, Aug. 2004.

[43] D. Wang, X. Li, J. Li, and B. Zhang, "The importance of query-concept-mapping for automatic video retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 285–288.

[44] J. Wang *et al.*, "Brain state decoding for rapid image retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 945–954.

[45] M. Wang and X. S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 10–31, 2011.

[46] M. Wang, G. Li, Z. Lu, Y. Gao, and T. S. Chua, "When Amazon meets Google: Product visualization by exploring multiple information sources," *ACM Trans. Internet Technol.*, vol. 12, no. 4, Jul. 2013, Art. ID 12.

[47] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[48] M. Wang, B. Ni, X. S. Hua, and T. S. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," *ACM Comput. Surv.*, vol. 4, no. 4, Aug. 2012, Art. ID 25.

[49] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, Dec. 2010.

[50] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[51] S. Wei, D. Xu, X. Li, and Y. Zhao, "Joint optimization toward effective and efficient image search," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2216–2227, Dec. 2013.

[52] K. Q. Weinberger and O. Chapelle, "Large margin taxonomy embedding with an application to document categorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 1737–1744.

[53] H. C. Wu, R. W. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TFIDF weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–37, 2008.

[54] B. Xie, Y. Mu, D. Tao, and K. Huang, "m-SNE: Multiview stochastic neighbor embedding," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 4, pp. 1088–1096, Aug. 2011.

[55] J. Yu, D. Liu, D. Tao, and H. S. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 5, pp. 1413–1427, Oct. 2012.

**Meng Wang** (M'09) received the B.E. and the Ph.D. degrees in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China, Hefei, China.

He is a Professor with the Hefei University of Technology, Hefei. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing.

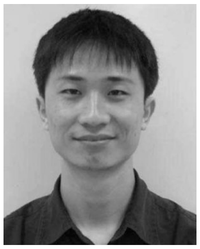
Dr. Wang was the recipient of several best paper awards, including from the 17th and 18th ACM International Conference on Multimedia, the 16th International Multimedia Modeling Conference, the 4th International Conference on Internet Multimedia Computing and Service, and the best demo award from the 20th ACM International Conference on Multimedia.

**Weisheng Li** received the B.S. and M.S. degrees from the School of Electronics and Mechanical Engineering, Xidian University, Xi'an, China, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, in 1997, 2000, and 2004, respectively.

He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include intelligent information processing and pattern recognition.

**Dong Liu** received the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2012.

He is currently an Associate Research Scientist with Digital Video and Multimedia Laboratory, Columbia University, New York City, NY, USA, where he is leading several research projects on video content analysis and retrieval. From 2011 to 2013, he was a Post-Doctoral Research Scientist with Columbia University, where he led projects on video event detection and object detection. From 2009 to 2010, he was a Research Engineer with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Prior to this, he was a Research Intern with Internet Media Group, Microsoft Research Asia, Beijing, China. His current research interests include multimedia event detection in videos, social image tag analysis and processing, novel machine-learning algorithms and their applications in computer vision, and multimedia retrieval.



**Bingbing Ni** received the B.Eng. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. degree from the National University of Singapore, Singapore, in 2005 and 2011, respectively.

He is currently a Research Scientist with Advanced Digital Sciences Center, Singapore. He was a Research Intern with Microsoft Research Asia, Beijing, China, in 2009, and also a Software Engineer Intern at Google Inc., Mountain View, CA, USA, in 2010. His current research interests include the areas of computer vision, machine learning, and multimedia.

Dr. Ni was the recipient of the Best Paper Award from PCM'11 and the Best Student Paper Award from PREMIA'08. He won the first prize in International Contest on Human Activity Recognition and Localization in conjunction with International Conference on Pattern Recognition, in 2012, and also won the second prize in ChaLearn 2014 Human Action Recognition Challenge.

**Jialie Shen** received the Ph.D. degree in computer science in the area of large-scale media retrieval and database access methods from the University of New South Wales (UNSW), Sydney, NSW, Australia.

He is an Assistant Professor with the School of Information Systems, Singapore Management University (SMU), Singapore. Before joining SMU, he was a Faculty Member with UNSW, and a Researcher with the Information Retrieval Research Group, the University of Glasgow, Glasgow, U.K. His current research interests include information retrieval in the textual and multimedia domain, economic-aware media analysis, and artificial intelligence, particularly machine perception and its applications on infrared and business intelligence.

**Shuicheng Yan** (SM'09) received the Ph.D. degree from Peiking University, Beijing, China, in 2004.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and a Founding Lead of the Learning and Vision Research Group (http://www.lv-nus.org). His current research interests include machine learning, computer vision, and multimedia. He has authored/co-authored hundreds of technical papers over a wide range of research topics, with over 13 000 Google Scholar citations and an H-index of 50.

Dr. Yan was the recipient of the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, International Conference on Multimedia & Expo'10, and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prize of ILSVRC'14 detection task, the winner prizes of the classification task in PASCAL VOC 2010–2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC'10, the 2010 TCSVT Best Associate Editor Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 NUS Young Researcher Award. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *ACM Transactions on Intelligent Systems and Technology*.