A Mixed Generative-Discriminative Based Hashing Method

Qi Zhang, Yang Wang, Jin Qian, and Xuanjing Huang

Abstract—Hashing methods have proven to be useful for a variety of tasks and have attracted extensive attention in recent years. Various hashing approaches have been proposed to capture similarities between textual, visual, and cross-media information. However, most of the existing works use a bag-of-words methods to represent textual information. Since words with different forms may have similar meaning, semantic level text similarities can not be well processed in these methods. To address these challenges, in this paper, we propose a novel method called semantic cross-media hashing (SCMH), which uses continuous word representations to capture the textual similarity at the semantic level and use a deep belief network (DBN) to construct the correlation between different modalities. To demonstrate the effectiveness of the proposed method, we evaluate the proposed method on three commonly used cross-media data sets are used in this work. Experimental results show that the proposed method achieves significantly better performance than state-of-the-art approaches. Moreover, the efficiency of the proposed method is comparable to or better than that of some other hashing methods.

Index Terms—Hashing method, word embedding, fisher vector

1 Introduction

With the rapid expansion of the World Wide Web, digital information has become much easier to access, modify, and duplicate. Hence, hashing based similarity calculation or approximate nearest neighbour searching methods have been proposed and received considerable attention in recent years. Various applications such as information retrieval, near duplicate detection, and data mining are performed by hashing based methods. Due to the rapid expansion of mobile networks and social media sites, information input through multiple channels has also attracted increasing attention. Images and videos are associated with tags and captions. According to research published on eMarketer, about 75 percent of the content posted by Facebook users contains photos.[1] The relevant data from different modalities usually have semantic correlations. Therefore, it is desirable to support the retrieval of information through different modalities. For example, images can be used to find semantically relevant textual information. On the other side, images without (or with little) textual descriptions are highly needed to be retrieved with textual query.

Along with the increasing requirements, in recent years, cross-media search tasks have received considerable attention [1], [2], [3], [4], [5], [6], [7]. Since each modality having different representation methods and correlational structures, a variety of methods studied the problem from the aspect of learning correlations between different modalities. Existing methods proposed to use Canonical Correlation Analysis (CCA) [8], manifolds learning [9], dual-wing harmoniums [10], deep autoencoder [11], and deep Boltzmann machine [12] to approach the task. Due to the efficiency of hashing-based methods, there also exists a rich line of work focusing the problem of mapping multi-modal high-dimensional data to low-dimensional hash codes, such as Latent semantic sparse hashing (LSSH) [13], discriminative coupled dictionary hashing (DCDH) [14], Cross-view Hashing (CVH) [15], and so on.

Most of the existing works use a bag-of-words to model textual information. The semantic level similarities between words or documents are rarely considered. Let us consider the following examples:

S1. The company *announces* new operating system.
S2. The company *releases* new operating system.
S3. The company *delays* new operating system.

From these examples, we can observe that although only one word differs between the three sentences, sentence *S3* should not be considered as the near duplicate sentence of *S1* and sentence *S2*. The meaning expressed by *S3* is much different with *S1* and *S2*'s. Since existing methods are usually based on lexical level similarities, this kind of issue cannot be well addressed by these methods.

In short text segments (e.g., microblogs, captions, and tags), the similarities between words are especially important for retrieval. For example: *journey* versus *travel*, *coast* versus *shore*. According to human-assigned similarity judgments [16], more than 90 percent of subjects thought that these pairs of words had similar meanings. Fig. 1 illustrates a set of images retrieved from Flickr using different queries. From these examples, we can see that images may express similar concepts, even though there is little overlap in terms of annotated tags. Since users rarely annotate a single image

1. http://www.socialmediaexaminer.com/photos-generate-engagement-research/

• *The authors are with the Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 201203, P.R. China.*
E-mail: {qz, 14210240023, 12110240030, xjhuang}@fudan.edu.cn.

Manuscript received 2 June 2015; revised 17 Oct. 2015; accepted 1 Dec. 2015. Date of publication 0 . 0000; date of current version 0 . 0000.
Recommended for acceptance by Y. Chang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2507127

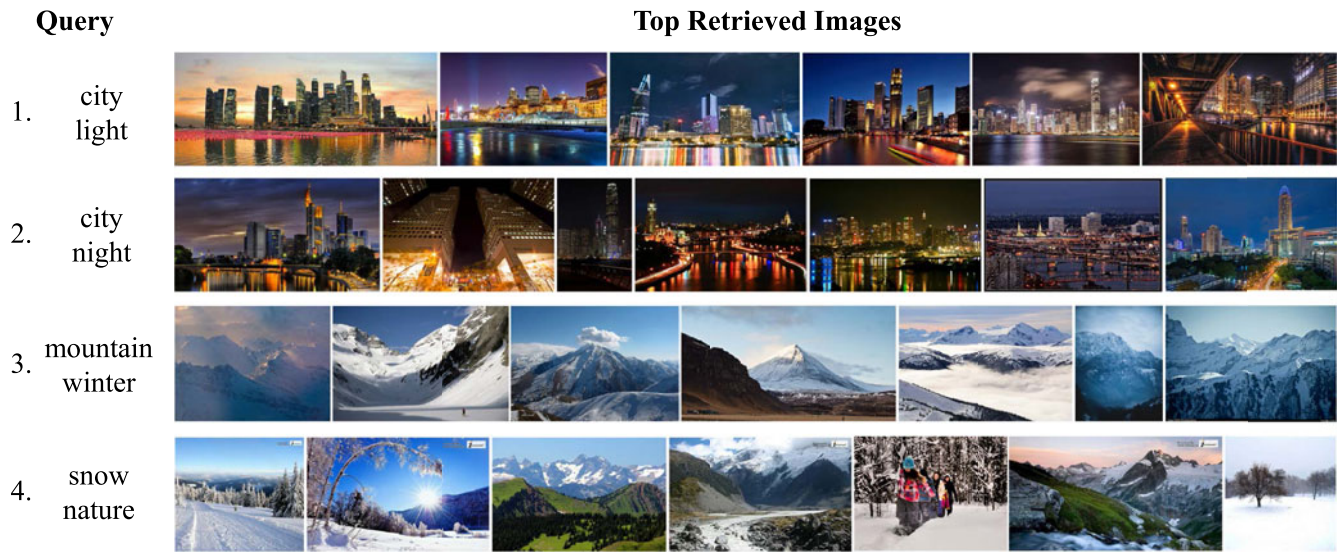


Fig. 1. An example of top retrieved images from Flickr with different tags.

using multiple words with similar meaning, semantic level textual similarities should be incorporated into the cross-media retrieval.

Motivated by the success of continuous space word representations (also called word embeddings) in a variety of tasks, in this work we propose to incorporate word embeddings to meet these challenges. Words in a text are embedded in a continuous space, which can be viewed as a Bag-of-Embedded-Words (BoEW). Since the number of words in a text is dynamic, in [17], we proposed a method to aggregate it into a fixed length Fisher Vector (FV), using a Fisher kernel framework [18]. However, the proposed method only focus on textual information. Another challenge in this task is how to determine the correlation between multi-modal representations. Since we propose the use of a Fisher kernel framework to represent the textual information, we also use it to aggregate the SIFT descriptors [19] of images. Through the Fisher kernel framework, both textual and visual information is mapped to points in the gradient space of a Riemannian manifold. However, the relationships that exist between FVs of different modalities are usually highly non-linear. Hence, to construct the correlation between textual and visual modalities, we introduce a DBN based method to model the mapping function, which is used to convert abstract representations of different modalities from one to another.

The main contributions of this work are summarized as follows.

- We propose to incorporate continuous word representations to handle semantic textual similarities and adopted for cross-media retrieval.
- Inspired by the advantages of DBN in handling highly non-linear relationships and noisy data, we introduce a novel DBN based method to construct the correlation between different modalities.
- A variety of experiments on three cross-media commonly used benchmarks demonstrate the effectiveness of the proposed method. The experimental results show that the proposed method can significantly outperform the state-of-the-art methods.

2 Related Work

Along with the increasing requirement, extensive Hashing-based methods have been proposed for cross-media retrieval. In this section, we briefly describe the related works, which can be categorized into the following four research areas: cross-media retrieval, text Reuse detection, and hashing methods.

2.1 Cross-Media Retrieval

Cross-media retrieval, in which the modality of input query and the returned results can be of different, has received considerable attentions [1], [3], [6], [7], [8], [9], [10], [12], [20], [21]. Wu et al. [8] introduced a Canonical Correlation Analysis based method to construct isomorphic subspace and multi-modal correlations between media objects and polar coordinates to judge the general distance of media objects. Due to lack of sufficient training samples, relevance feedback of user was used to accurately refine cross-media similarities. Yang et al. [9] proposed manifold-based method, in which they used Laplacian media object space to represent media object for each modality and an multimedia document semantic graph to learn the multimedia document semantic correlations. In [22], a rich-media object retrieval method is proposed to represent data consisting of multiple modalities, such as 2-D images, 3-D objects and audio files. To tackle the large scale problem, a multimedia indexing scheme was also adopted.

Since the relationships across different modalities are typically highly non-linear and observations are usually noisy, Srivastava and Salakhutdinov [12] proposed a Deep Boltzmann Machine to learn joint representations of image and text inputs. The proposed model fuses multiple data modalities into a unified representation, which can be used for classification and retrieval. Xing et al. [10] introduced to use dual-wing harmoniums to build a joint model for images and text. The model incorporated Gaussian hidden units together with Gaussian and Poisson visible units into a linear RBM model. In [12], a multimodal deep Boltzmann machine was proposed for learning multimodal data representations. To reduce the training time complexity,

Zhang and Li [6] proposed to seamlessly integrate semantic labels into the hashing learning procedure for large-scale data modeling.

In past few years, deep neural networks (DNNs) have achieved tremendous success in various tasks. Cross-media retrieval is one of the tasks which DNNs and other neural network architectures obtained improvements. In [11], a deep autoencoder was proposed to learn features over multiple modalities. The method uses the hidden units to construct shallow representation for the data and builds deep bimodal representations by modeling the correlations across the learned shallow representations. Karpathy and Fei-Fei proposed a multimodal recurrent neural network for generating descriptions for images[23]. The generated descriptions can be used for cross-media retrieval.

Most of the existing works described above focused on constructing the correlations between multiple modalities from different aspects. They usually use bag-of-words model to represent text. However, we in this work propose to use Fisher kernel framework to represent both textual and visual information and use a deep network to construct the correlations between the two manifolds.

2.2 Near-Duplicate Detection

The task of detecting near duplicate textual information has received considerable attentions in recent years. Previous works studied the problem from different aspects such as fingerprint extraction methods with or without linguistic knowledge, hash codes learning methods, different granularities, and so on.

Broder [24] proposed Shingling method, which uses contiguous subsequences to represent documents. It does not rely on any linguistic knowledge. If sets of shingles extracted from different documents are appreciably overlap, these documents are considered exceedingly similar, which are usually measured by Jaccard similarity. In order to reduce the complexity of shingling, meta-sketches was proposed to handle the efficiency problem [25]. In order to improve the robustness of shingle-like signatures, Theobald et al. [26] introduced a method, SpotSigs. It provides more semantic pre-selection of shingles for extracting characteristic signatures from Web documents. SpotSigs combines stopword antecedents with short chains of adjacent content terms. The aim of it is to filter natural-language text passages out of noisy Web page components. They also proposed several pruning conditions based on the upper bounds of Jaccard similarity.

I-Match [27] is one of the methods using hash codes to represent input document. It filters the input document based on collection statistics and compute a single hash value for the remainder text. If the documents have same hash value, they are considered as duplicates. It hinges on the premise that removal of very infrequent terms and very common terms results good document representations for the near-duplicate detection task. Since I-Match signatures are respect to small modifications, Kołcz et al. [28] proposed the solution of several I-Match signatures, all derived from randomized versions in the original lexicon.

Local text reuse detection focused on identifing the reused and modified sentences, facts or passages, rather than whole documents. Seo and Croft [29] analyzed the task and defined six categories of text reuses. They proposed a general framework for text reuse detection. Several fingerprinting techniques under the framework were evaluated under the framework. Zhang et al. [30] also studied the partial-duplicate detection problem. They converted the task into two subtasks: sentence level near-duplicate detection and sequence matching. Except for the similarities between documents, the method can simultaneously output the positions where the duplicated parts occur. In order to handle the efficiency problem, they implement their method using three Map-Reduce jobs. Kim et al. [31] proposed to map sentences into points in a high dimensional space and leveraged range searches in this space. They used MD5 hash function to generate hash code for each word. File signature is then created by taking the bitwise-or of all signatures of words that appear in the file.

Different with these existing methods, in this paper, we propose to use aggregated word embeddings to capture the semantic level similarities to reduce the false matches.

2.3 Hashing-Based Methods

In recent years, hashing-based methods, which create compact hash codes that preserve similarity, for single-modal or cross-modal retrieval on large-scale databases have attracted considerable attention [4], [5], [12], [13], [14], [15], [32], [33], [34], [35], [36], [37], [38]. For single-modal, Hinton and Salakhutdinov [33] proposed a two-layer network, which is called a Restricted Boltzmann machine (RBM), with a small central layer to convert high-dimensional input vectors into low-dimensional codes. In [36], spectral hashing was defined to seek compact binary codes in order to preserve the semantic similarity between codewords. The criterion used in spectral hashing is related to graph partitioning. Norouzi and Fleet [39] introduced a method based on latent structural SVM framework for learning similarity-preserving hash functions. A specific loss function is designed to incorporating both Hamming distance and binary quantization into consideration. In [40], Self-Taught Hashing (STH) converted the hash codes learning problem into two stages. Unsupervised method, binarised Laplacian Eigenmap, is used to optimize l-bit binary codes. The classifiers were trained to predict the l-bit code for unseen documents.

A variety works studied the problem of mapping multimodal high-dimensional data to low-dimensional hash codes. Latent semantic sparse hashing [13] proposed the use of Matrix Factorization to represent text and sparse coding to capture the salient structures of images. Then, these representations are mapped to a joint abstraction space. However, LSSH requires the use of both visual and textual information to construct the data set. Although out-of-samples can be estimated, the performances may be heavily influenced. Yu et al. [14] introduced a discriminative coupled dictionary hashing approach, which generated a coupled dictionary for each modality based on category labels. Kumar and Udupa [15] formulated formulated the problem of learning hash functions as a constrained minimization problem. Since the optimization problem is NP hard, they transformed it into a tractable eigenvalue problem by means of a relaxation. Inter-media hashing (IMH) [4] uses a linear

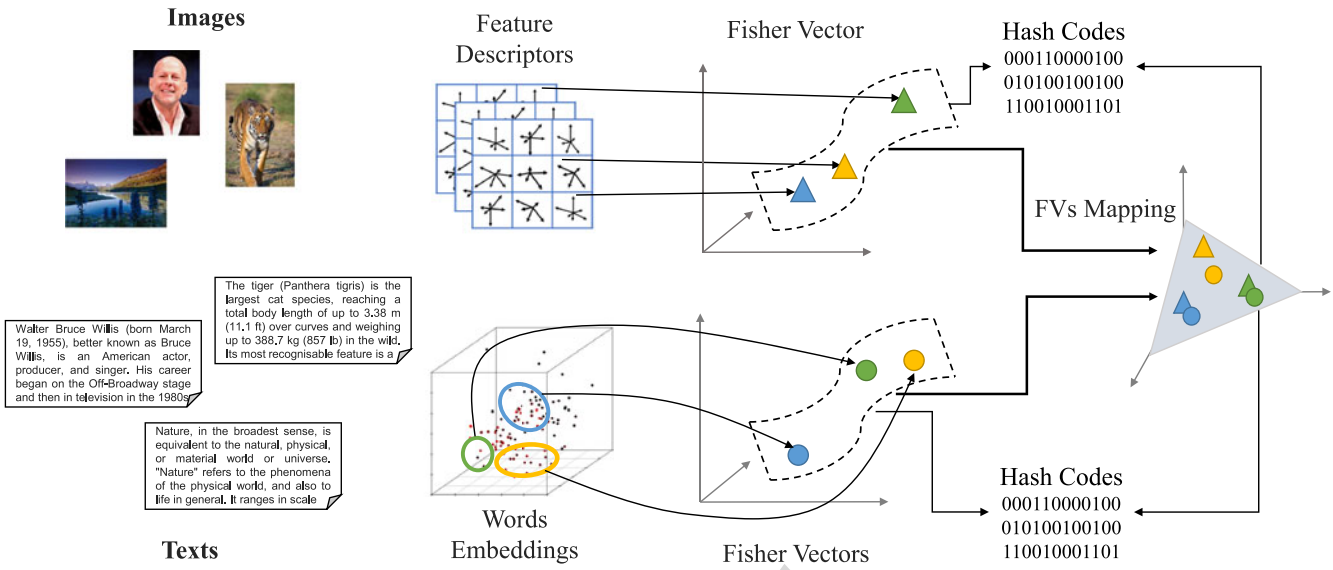


Fig. 2. An overview processing flow of the proposed SCMH for cross-media retrieval.

regression model to jointly learn a set of hashing functions for each individual media type.

Since we in this work learn the mapping functions between FVs of different modalities, all the hashing based methods for single modality can be incorporated into it.

2.4 Neural Networks for Representing Image and Text

The task of learning continuous space word representations have a long history[41], [42], [43], [44], [45], [46], [47]. It has demonstrated outstanding results across a variety of tasks. Hinton and Salakhutdinov [44] introduced a deep generative model to learn word-count vector and binary code for documents. In [45], the word representations are learned by a recurrent neural network language model. The proposed architecture consists of an input layer and a hidden layer with recurrent connections. Probabilistic neural network language model (NNLM) [48] simultaneously learns a distributed representation for each word and the probability function for word sequences. Bordes et al.[41] proposed to use multi-task training process to jointly learn representations of words, entities and meaning representations. The work described in [49] introduced a mix of unsupervised and supervised techniques to learn word vectors to capture both semantic and sentiment similarities among words.

On the image side, there are also a variety of studies tackling the problem of higher-level representations of visual information. Krizhevsky et al. [50] proposed to use a deep convolutional neural network to perform object detection. In [51], region proposals are combined with CNNs to generate features for object detection. Except for these supervised methods, unsupervised learning methods for training visual features have also been carefully studied. Lee et al. [52] introduced convolutional deep belief network , a hierarchical generative model, represent images. Taylor et al. [53] proposed a convolutional gated restricted Boltzmann machineto model the spatio-temporal features for videos.

Although, in this work, we use word embeddings and SIFT to represent texts and images respectively, the proposed method can also incorporate these representations.

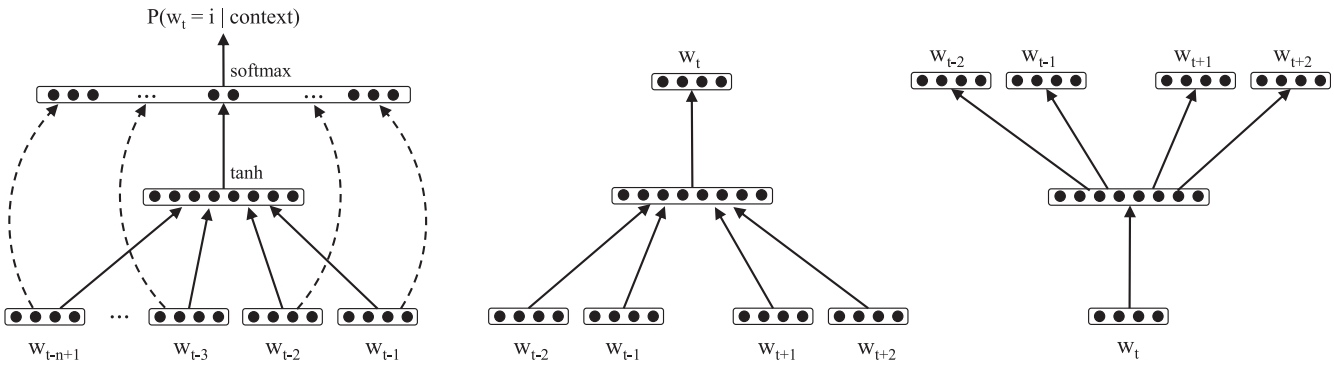
3 The Proposed Method

The processing flow of the proposed semantic cross-media hashing (SCMH) method is illustrated in Fig. 2. Given a collection of text-image bi-modality data, we firstly represent image and text respectively. Through table lookup, all the words in a text are transformed to distributed vectors generated by the word embeddings learning methods. For representing images, we use SIFT detector to extract image keypoints. SIFT descriptor is used to calculate descriptors of the extracted keypoints [19]. After these steps, a variable size set of points in the embeddings space represents the text, and a variable size set of points in SIFT descriptor space represents each image. Then, the Fisher kernel framework is utilized to aggregate these points in different spaces into fixed length vectors, which can also be considered as points in the gradient space of the Riemannian manifold. Henceforth, texts and images are represented by vectors with fixed length. Finally, the mapping functions between textual and visual Fisher vectors (FVs) are learned by a deep neural network. We use the learned mapping function to convert FVs of one modality to another. Hash code generation methods are used to transfer FVs of different modalities to short length binary vectors. In the following section, we provide detailed examples of practical applications of the proposed method.

3.1 Word Embeddings Learning

Representation of words as continuous vectors recently has been shown to benefit performance for a variety of NLP and IR tasks [44], [46], [47]. Similar words tend to be close to each other with the vector representation. Moreover, Mikolov et al. [54] also demonstrated the learned word representations could capture meaningful syntactic and semantic regularities. Hence, in this work, we propose to use word embeddings to capture the semantic level similarities between short text segments.

Fig. 3 shows three architectures used for learning word embeddings. w_i represents the ith words in the given words sequence $\{w_1, w_2, \ldots, w_T\}$. Fig. 3a shows the architecture of



(a) NNLM (b) CBOW (c) Skip-gram

Fig. 3. Methods used to learn word embeddings. The NNLM architecture predicates the probability of words based on the existing words [48]. CBOW predicts the current word based on the context [54]. Skip-gram predicts surrounding words given the current word [54].

the probabilistic neural network language model (NNLM) proposed by Bengio et al. in [48]. It can have either one hidden layer beyond the word features mapping or direct connections from the word features to the output layer. They also proposed to use $softmax$ function for the output layer to guarantee positive probabilities summing to 1. The word vectors and the parameters of that probability function can be learned simultaneously. In this work, we only use the learned word vectors.

Figs. 3b and 3c show the architectures of the methods proposed by Mikolov in [54]. The architecture of CBOW, which is similar to NNLM, is shown in Fig. 3b. The main differences are that (i) the non-linear hidden layer is removed; (ii) the words from the future are included; (iii) the training criterion is to correctly classify the current (w_t) word. The Skip-gram architecture, which is shown in Fig. 3c, is similar to CBOW. However, instead of predicting the current word based on the history and future words, it tries to maximize classification accuracy of words within a certain range before and after the current word based on only the current word as input.

Besides the methods mentioned above, there are also a large number of works addressing the task of learning distributed word representations [47], [49], [55]. Most of them can also be used in this work. The proposed framework has no limits in using which of the continuous word representation methods.

3.2 Fisher Kernel Framework

Fisher kernel framework [18] was proposed to directly obtain the kernel function from a generative probability model. A parametric class of probability models $P(X|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^l$ for some positive integer l. If the dependence on θ is sufficiently smooth, the collection of models with parameters from Θ can be viewed as a manifold M_Θ. Though applying a scalar product at each point $P(X|\theta) \in M_\Theta$, it can be turned into a Riemannian manifold [56].

We denote a text or an image $X = \{x_i, 1 \leq i \leq |X|\}$, where x_i is the embedding of ith word of a text or the SIFT descriptors of the ith keypoint of an image, $|X|$ is the number of words in a text or the number of the extracted keypoints in an image. x_i is D-dimension word embeddings or SIFT descriptors. We should note that there may be different parameters for different data sets. According to the Fisher kernel framework, X can be modeled by a probability density function. In this work, $P(X|\theta)$ is given by Gaussian mixture model (GMM), which a sum of N Gaussians $N(\mu_i, \Sigma_i)$ with weights ω_i. Let $\theta = \{\omega_i, \mu_i, \Sigma_i, \forall i = 1 \ldots N\}$ be the set of GMM parameters. The parameters θ are estimated through the optimization of Maximum Likelihood (ML) criterion using Expectation Maximization (EM) method [57].

Based on the learned parameters set θ, a text or an image X can be characterized by the gradient vector using the following function:

$$\begin{aligned} G_\theta^X &= \nabla_\theta \log P(X|\theta) \\ &= \left(\frac{\partial}{\partial\theta_1}\log\left(P(X|\theta),\ldots,\frac{\partial}{\partial\theta_l}\log\left(P(X|\theta)\right)\right)\right), \end{aligned} \tag{1}$$

where G_θ^X is a vector whose dimensionality is only dependent on the number of parameters in λ, not on the number of words or keypoints. The gradient describes the contribution of each individual parameters to the generative process. It can also be interpreted as how these parameter contribute to the process of generating an example. We follow the work described in [18] for normalizing these gradients by incorporating Fisher information matrix (FIM) F_θ. According to the theory of information geometry [58], $\mathscr{U} = \{P(X|\theta), \theta \in \Theta\}$, which is a parametric family of distributions, can be regarded as a Riemanninan manifold M_Θ with a local metric given by the FIM $F_\theta \in \mathbb{R}^{M\times M}$:

$$F_\theta = \mathbf{E}\Big(\nabla_\theta\, logP(X|\theta)\nabla_\theta logP(X|\theta)^T\Big). \tag{2}$$

The similarity between two samples X and Y can be measured by the Fisher kernel defined as:

$$K_{FK}(X,Y) = G_\theta^{X^T} F_\theta^{-1} G_\theta^Y. \tag{3}$$

Since F_θ is symmetric and positive definite, F_θ^{-1} can be transformed to $L_\theta^T L_\theta$ based on the Cholesky decomposition. Therefore, $K_{FK}(X,Y)$ can be rewritten as follows:

$$K_{FK}(X,Y) = \mathscr{G}_\theta^{X^T}\mathscr{G}_\theta^Y, \tag{4}$$

where

$$\mathscr{G}_\theta^X = L_\theta G_\theta^X = L_\theta \nabla_\theta \log P(X|\theta). \tag{5}$$

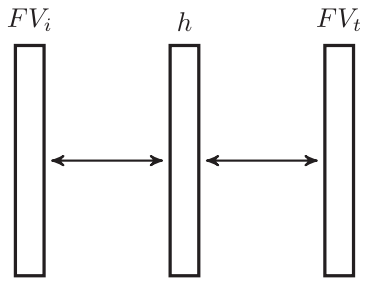


Fig. 4. A single hidden layer model for mapping FVs of different modalities. FV_i and FV_t denote the Fisher vector of image and text, respectively. h represents the hidden layer.

In this work, we assume that $x_i (1 \leq i \leq |X|)$ follows the naive independence model, $\mathscr{G}_\theta^X$ can be rewritten as follows:

$$\sum_{i=1}^{|X|} L_\theta \, \nabla_\theta \log P(x_i|\theta). \tag{6}$$

$\mathscr{G}_\theta^X$ is also referred to as the *Fisher Vector* of X.

Based on the specific probability density function GMM, which we used in this work, FV of X is respect to the mean μ and standard deviation σ of all the mixed Gaussian distributions. Let $\gamma_{x_i}(k)$ be the soft assignment of the x_i in X to Gaussian k:

$$\gamma_{x_i}(k) = \mathbb{P}(k|x_i, \theta) = \frac{\omega_i P_k(x_i|\theta)}{\sum_{j=1}^{N} \omega_j P_j(x_i|\theta)}. \tag{7}$$

Mathematical derivations lead to:

$$\begin{aligned} \mathscr{G}_{\mu,k}^X &= \frac{1}{|X|\sqrt{\omega_i}} \sum_{i=1}^{|X|} \gamma_{x_i}(k) \left(\frac{x_i - \mu_k}{\sigma_k} \right), \mathscr{G}_{\sigma,k}^X \\ &= \frac{1}{|X|\sqrt{2\omega_i}} \sum_{i=1}^{|X|} \gamma_{x_i}(k) \left[\frac{(x_i - \mu_k)^2}{\sigma_k^2} - 1 \right]. \end{aligned} \tag{8}$$

The division between vectors is as a term-by-term operation. The final gradient vector $\mathscr{G}_\lambda^X$ is is the concatenation of the $\mathscr{G}_{\mu,k}^X$ and $\mathscr{G}_{\sigma,k}^X$ vectors for $k = 1 \ldots N$. Let T be the dimensionality of the vector offsets. The final gradient vector $\mathscr{G}_\lambda^X$ is therefore 2NT-dimensional.

3.3 Mapping Function Learning

To transfer the FVs of one modality to another, we propose to use a deep belief network with one hidden layer to achieve the task. Fig. 4 shows the structure of the proposed method. The building block of the network used in this work is the Gaussian restricted Boltzmann machine. Because we have converted both textual and visual information into the gradient space of a Riemannian manifold, we in this work use a single hidden layer model to do it.

The restricted Boltzmann machine is a kind of an undirected graphical model with observed units and hidden units. The undirected graph of an RBM has an bipartite structure. It can be understood as a Markov random field with latent factors which explain the input observed data using binary hidden variables. Let $\mathbf{v}$ be the L dimensional observed data, which can take real values or binary values. The dimension of stochastic binary units h is K. Each visible unit is connected to each hidden unit. The graphical model

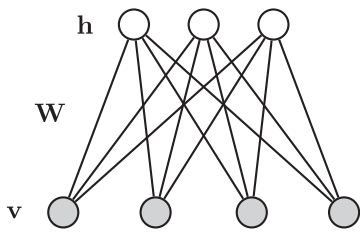


Fig. 5. A graphical model representation of restricted Boltzmann machine.

representation is illustrated in Fig. 5. The parameters of RBM consist of the weight matrix $\mathbf{W} \in \mathbb{R}^{L \times K}$, the biases $\mathbf{c} \in \mathbb{R}^L$ for observed units, and the biases $\mathbf{b} \in \mathbb{R}^K$ for hidden units. If the observed units are real-valued, the model is called the Gaussian RBM. Its joint probability distribution can be defined as follows:

$$\begin{aligned} P(\mathbf{v}, \mathbf{h}) &= \frac{1}{Z} exp(-E(\mathbf{v}, \mathbf{h})), E(\mathbf{v}, \mathbf{h}) \\ &= \frac{1}{2\sigma^2} \sum_i (v_i - c_i)^2 - \frac{1}{\sigma} \sum_{i,j} v_i W_{ij} h_j - \sum_j b_j h_j, \end{aligned} \tag{9}$$

where Z is a normalization constant. The conditional distribution of this model can be written as follows:

$$P(h_j = 1|\mathbf{v}) = sigm\left(\frac{1}{\sigma} \sum_i W_{ij} v_i + b_j \right), \tag{10}$$

$$P(v_i = 1|\mathbf{h}) = \mathcal{N}\left(v_i; \sigma \sum_j W_{ij} h_j + c_i, \sigma^2 \right), \tag{11}$$

where $sigm(s) = \frac{1}{1+exp(-s)}$ is the sigmoid function, and $\mathcal{N}(.;.,.)$ is a Gaussian distribution.

Although exact maximum likelihood learning in this model is intractable, sampling-based approximate maximum-likelihood methods can be used to estimated the parameters. Because the variables in a layer are conditionally independent, block Gibbs sampling can be performed in parallel. After training the RBM, Fisher vectors of different modalities can be transferred with the estimated parameters.

3.4 Hash Code Generation

Through the previous steps, a variable length of text segments or keypoints can be transferred to a fixed length vector. However, Fisher vectors are usually high dimensional and dense. It limits the usages of FVs for large-scale applications, where computational requirement should be studied. In this work, we propose to use hashing methods to address the efficiency problem.

The task of generating hash codes for samples can be formalized as learning a mapping $b(\mathbf{x})$, referred to as a hash function, which can project p-dimensional real-valued inputs $x \in \mathbb{R}^p$ onto q-dimensional binary codes $h \in \mathcal{H} \equiv \{-1, 1\}^q$, while preserving similarities between samples in original spaces and transformed spaces. The mapping $b(\mathbf{x})$ can be parameterized by a real-valued vector $\mathbf{w}$ as:

$$b(\mathbf{x}; \mathbf{w}) = \text{sign}(f(\mathbf{x}; \mathbf{w})), \tag{12}$$

where $\mathrm{sign}(\cdot)$ represents the element-wise sign function, and $f(\mathbf{x}; \mathbf{w})$ denotes a real-valued transformation from $\mathbb{R}^p$ to $\mathbb{R}^q$. In this work, Fisher vectors of text segments or keypoints are the $\mathbf{x}$ in mapping function $b(\mathbf{x}; \mathbf{w})$. A variety of existing methods have been proposed to achieve this task under this framework using different forms of f and different optimization objectives. Most of the learning to hash methods for dense vectors can be used in this framework. In this work, we evaluated several state-of-the-arts hashing methods, whose performances are shown in the experiment section.

4 Experiments

To demonstrate the effectiveness of the proposed method, we compare and contrast the experimental results of SCMH and state-of-the-art hashing methods on three commonly used data sets for cross-media retrieval.

4.1 Data Sets

The three data sets used in this example contain both texts and images. They have been chosen for the purpose of evaluating various cross-media retrieval methods.

Flickr. The MIR Flickr data set[2] [59], which consists of one million images along with their user assigned tags, was collected from Flickr. Out of all the images, 25,000 images are annotated for 24 concepts, including object categories (e.g., bird, people) and scene categories (e.g., sky, night). A stricter annotation was made on 14 concepts where a subset of the positive images was selected only if the concept is salient in the image. Therefore, this leads to a total of 38 concepts for this data set. Following previous works, each image may belong to one or more concepts. Image-text pairs are considered to be similar if they share the same concept.

LabelMe. The LabelMe data set[3] [60] contains 2,688 images, which belong to eight outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. All the objects in these images have been fully labeled and used as tags of the images. Following the work described in [13], tags occurring in fewer than three images are discarded. Therefore, there are a total of 245 unique tags remaining. To construct the golden standards, we also follow previous works and assume that image-text pairs are regarded as similar if they share the same scene label.

NUS-WIDE. The NUS-WIDE data set[4] [61] contains images and their associated tags from Flickr. The total number of images and unique tags are 269,648 and 5,018 respectively. The dataset includes six kinds of low-level features extracted from these images and 81 manually constructed ground-truth concepts. For comparison with previous methods, we also used the 10 most common concepts, and randomly selected 20,000 images from them for evaluation. We treat as similar image-text pairs labeled with the same concepts

4.2 Experiment Settings

For multimodal documents, we use the SIFT framework to represent images and use word embeddings to represent text. We use the SIFT keypoint detector to extract a variable number of keypoints for each image and calculate the descriptors of the keypoints using 128-dimensional SIFT descriptors. The toolkit we used in this work is VLFeat 0.9.19.[5] The word embeddings we used in this work are pre-trained vectors trained on part of a Google News dataset (about 100 billion words). A Skip-gram model [62] was used to generate these 300-dimensional vectors for three million words and phrases. For generating Fisher vectors, we use the implementation of INRIA [63].

To demonstrate the effectiveness of the propose method, we evaluated the following state-of-the-art methods on the three data sets:

- *Cross-view Hashing* [15] maps similar objects to similar codes across the views to enable cross-view similarity search.
- *Discriminative coupled dictionary hashing* [14] generates a coupled dictionary for each modality based on category labels.
- *Multi view discriminative coupled dictionary hashing (MV-DCDH)* [14] is extended from DCDH with multi-view representation to enhance the representing capability of the relatively "weak" modalities.
- *Latent semantic sparse hashing* [13] uses Matrix Factorization to represent text and sparse coding to capture the salient structures of images.
- *Collective matrix factorization hashing (CMFH)* [1] generates unified hash codes for different modalities of one instance through collective matrix factorization with latent factor model.
- *Semantic correlation maximization (SCM)* [6] integrates semantic labels into the hashing learning procedure for preserving the semantic similarity cross modalities.

The toolkits of LSSH, DCDH, and MV-DCDH are kindly provided by the authors. As we mentioned in the previous section, the proposed method SCMH can incorporate any hashing methods for single modality. In this work, we use Semantic Hashing to generate hash codes for both textual and visual information. Semantic Hashing [33] is a multilayer neural network with a small central layer to convert high-dimensional input vectors into low-dimensional codes.[6] For the length of hash codes, all the methods generate 32, 64, and 128 bits hash codes.

Following previous literatures on this task, we also adopt the widely used *Mean Average Precision* (MAP) as the evaluation metric. For a single query and top-K retrieved instances, *Average Precision* (AP) is defined as follows:

$$AP = \frac{1}{R}\sum_{k=1}^{K} P(k)\delta(k),$$

where R denotes the number of ground-truth instances in the retrieved set, $P(k)$ denotes the precision of top-k retrieved instances, and $\delta(k)$ is an indicator function which equals to 1 if the kth instance is relevant to query or 0 otherwise. In the experiments, we set $K = 50$. Besides MAP, we

2. http://press.liacs.nl/mirflickr/
3. http://people.csail.mit.edu/torralba/code/spatialenvelope/
4. http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
5. http://www.vlfeat.org/
6. http://www.cs.toronto.edu/~hinton/

TABLE 1
MAP Comparison on Flickr

Tasks	Methods	Code Length		
		32	64	128
Text → Image	CVH	0.615	0.613	0.610
	DCDH	0.577	0.598	0.611
	MV-DCDH	0.600	0.603	0.614
	LSSH	0.623	0.634	0.626
	CMFH	0.625	0.630	0.632
	SCM	0.624	0.606	0.600
	SCMH	**0.640**	**0.644**	**0.645**
Image → Text	CVH	0.609	0.601	0.602
	DCDH	0.610	0.621	0.622
	MV-DCDH	0.604	0.614	0.619
	LSSH	0.618	0.630	0.617
	CMFH	0.619	0.626	0.621
	SCM	0.614	0.620	0.623
	SCMH	**0.643**	**0.650**	**0.649**

also report precision-recall curve to represent the precision at different recall level.

We report the results of Text → Image and Image → Text tasks on all three databases. For Text → Image task, a text query, which contains the annotated tags of an image, is input to search images. The text query is firstly represented by a Fisher vector based on word embeddings. Then, the FV of text is mapped into a FV in image space. Finally, hamming distance is used to measure the similarities between the hash code of the converted FV and other hash codes of images. The top-K images are selected as the results. The procedure of Image → Text task is similar as the Text → Image task. Since the Fisher vector mapping function needs training data, for each data set, we select 40 percent of the data to train the mapping function between text and image. 35 percent of the data are chosen as the retrieval database and the others are formed the query set. All the methods use the same data splits.

4.3 Results and Discussions

4.3.1 Results on Flickr

Table 1 shows the comparisons of the proposed method with the state-of-the-art methods on the Flickr MIR data set. From the results, we observe that the proposed method SCMH achieves the best performance among all the methods on both Text → Image and Image → Text tasks. LSSH achieves the second best results in most of the cases. It approaches the best result when the hash code length is 64. However, if we increase the hash code length to 128, performances of LSSH and SCM are decrease. On the contrary, the performances of SCMH with different length of hash codes are more robust. The main possible reason is that the performances of SCMH are highly impacted by the mapping functions between FVs of different modalities. If we use the cosine similarity between Fisher vectors to rank candidates, the MAP results can reach 0.682 and 0.678 in Text → Image and Image → Text task respectively.

The precision and recall curves (PR-curves) are plotted in Fig. 6, where the x-axis denotes the recall and the y-axis indicates the corresponding precision. From these figures, we observe that SCMH outperforms the other methods on all tasks, especially with long hash codes. The performance of CVH, DCDH, MV-DCDH, LSSH, CMFH, and SCM decrease much more quickly than SCMH. This also confirms that the proposed SCMH better suits the tasks for cross-media retrieval.

4.3.2 Results on LabelMe

Table 2 compares the relative performances of the different methods on the LabelMe dataset. Fig. 7 gives the PR-curves of different methods on the dataset. From the results, we observe that SCMH achieves better performance than state-of-the-art methods on all tasks. From analyzing the data, we find that different tags belonging to the same category may express similar or related meaning. Since semantic relations can be readily captured by the proposed method, SCMH outperforms the other methods. As the length of hash code increases, the MAP performance of SCMH improves. However, when the hash code length approaches 128, the performances of most of the methods except SCMH decrease. Comparing with Flickr dataset, the total number of images and unique tags are much smaller than it. Hence, the main possible reason is that longer hash codes encode more explicit information and thus the inability to capture the

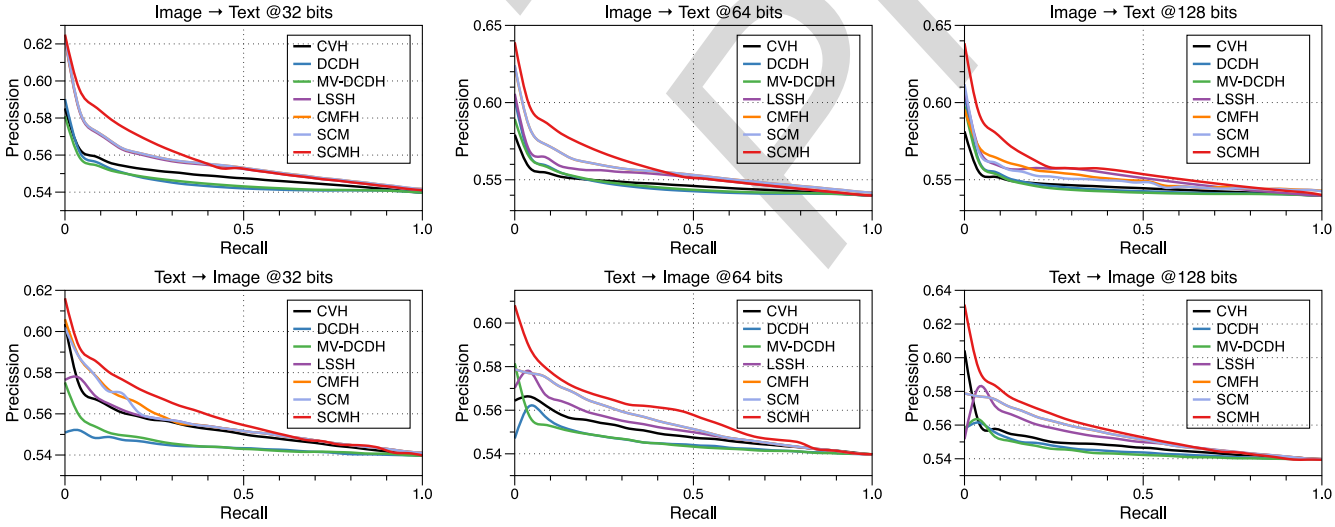


Fig. 6. The precision-recall curves of different hash code generation methods on the Flickr data set.

TABLE 2
MAP Comparison on LabelMe

Tasks	Methods	Code Length		
		32	64	128
Text → Image	CVH	0.400	0.370	0.349
	DCDH	0.410	0.449	0.424
	MV-DCDH	0.437	0.476	0.448
	LSSH	0.665	0.695	0.671
	CMFH	0.589	0.601	0.610
	SCM	0.613	0.621	0.615
	SCMH	**0.694**	**0.701**	**0.714**
Image → Text	CVH	0.343	0.342	0.338
	DCDH	0.416	0.466	0.428
	MV-DCDH	0.448	0.480	0.455
	LSSH	**0.670**	0.673	0.687
	CMFH	0.636	0.644	0.652
	SCM	0.622	0.630	0.636
	SCMH	0.662	**0.676**	**0.688**

TABLE 3
MAP Comparison on NUS-WIDE

Tasks	Methods	Code Length		
		32	64	128
Text → Image	CVH	0.435	0.426	0.418
	DCDH	0.468	0.486	0.484
	MV-DCDH	0.479	0.487	0.484
	LSSH	0.504	0.509	0.504
	CMFH	0.504	0.510	0.508
	SCM	0.526	0.528	0.530
	SCMH	**0.552**	**0.560**	**0.556**
Image → Text	CVH	0.437	0.426	0.421
	DCDH	0.460	0.476	0.481
	MV-DCDH	0.462	0.474	0.478
	LSSH	0.504	0.501	0.498
	CMFH	0.512	0.514	0.511
	SCM	0.531	0.539	0.541
	SCMH	**0.590**	**0.597**	**0.593**

semantic level similarities between tags decreases the performance. We also observe that SCMH achieves better performance on the Text → Image task than the Image → Text task. The DCDH, MV-DCDH, LSSH, CMFH, and SCM methods all behave differently from SCMH, achieving better performance on the Image → Text task. The main reason is possibly that word level semantic similarities can be better captured than keypoints represented by SIFT descriptors though SCMH. In Image → Text task, the performance of SCMH is slightly worse than LSSH when the hash code length is 32 bits. We think that the main reason is that size of LabelMe dataset and number of tags occurred in this dataset are both too small.

From the PR-curves shown in Fig. 7, we also observe that although SCMH has similar performance as LSSH in Image → Text task, the precision of SCMH decreases much more slowly. This means that SCMH can achieve better results when the user needs more candidates. We also observe from the figure that the improvements of SCMH on the Image → Text task are relatively marginal compared to those on the Text → Image tasks at all recall levels. This also confirms the phenomenon described above.

4.3.3 *Results on NUS-WIDE*

The results of different methods on the NUS-WIDE dataset are shown in Table 3. The corresponding PR-curves of them are given in the Fig. 8. From the results, we observe that SCMH achieves significantly better performance than state-of-the-art methods on all tasks. The relative improvements of SCMH over the second best results are 10.0 and 18.5 percent on the Text → Image and Image → Text task respectively. Comparing with the results of SCMH on LabelMe and Flickr dataset, the improvement of SCMH on NUS-WIDE is more significant. The main possible reason is that the number of tags based on their frequency we used in this dataset is bigger than LabelMe and Flickr. There are only a total of 245 unique tags which occur more than three times in the whole LabelMe dataset. For comparing with other methods, we selected top 500 most frequent tags in Flickr data set. Since NUS-WIDE is a more practical dataset, which contains more unique tags, we propose to use the top 1,000 most frequent tags. Hence, the weakness of the other methods in capturing the semantic level similarities between tags decreases the performance.

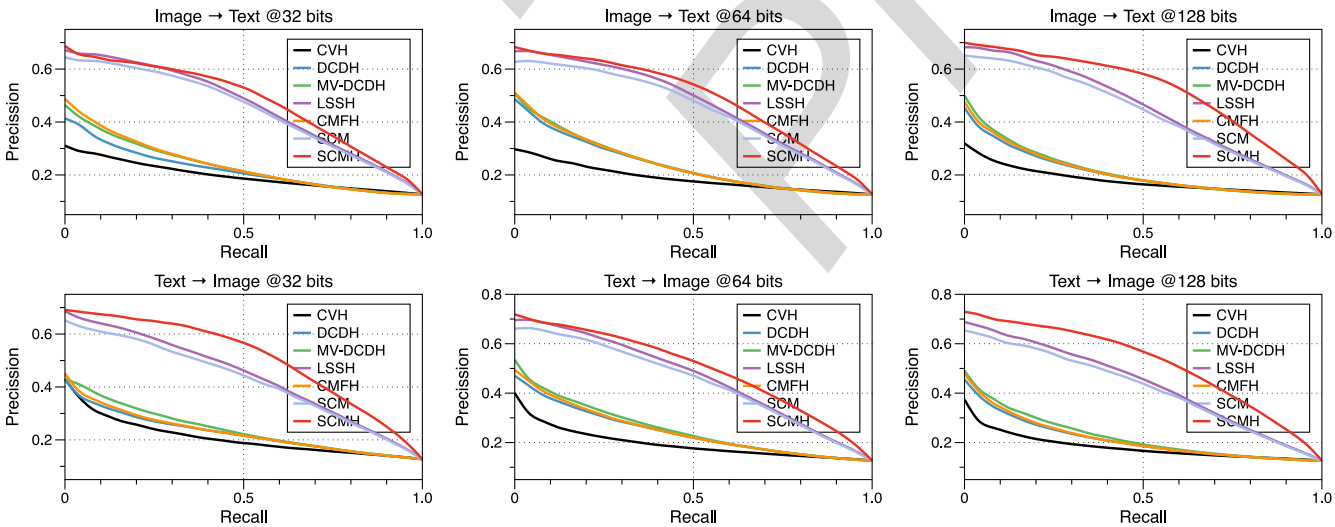


Fig. 7. The precision-recall curves of different hash code generation methods on the LabelMe data set.

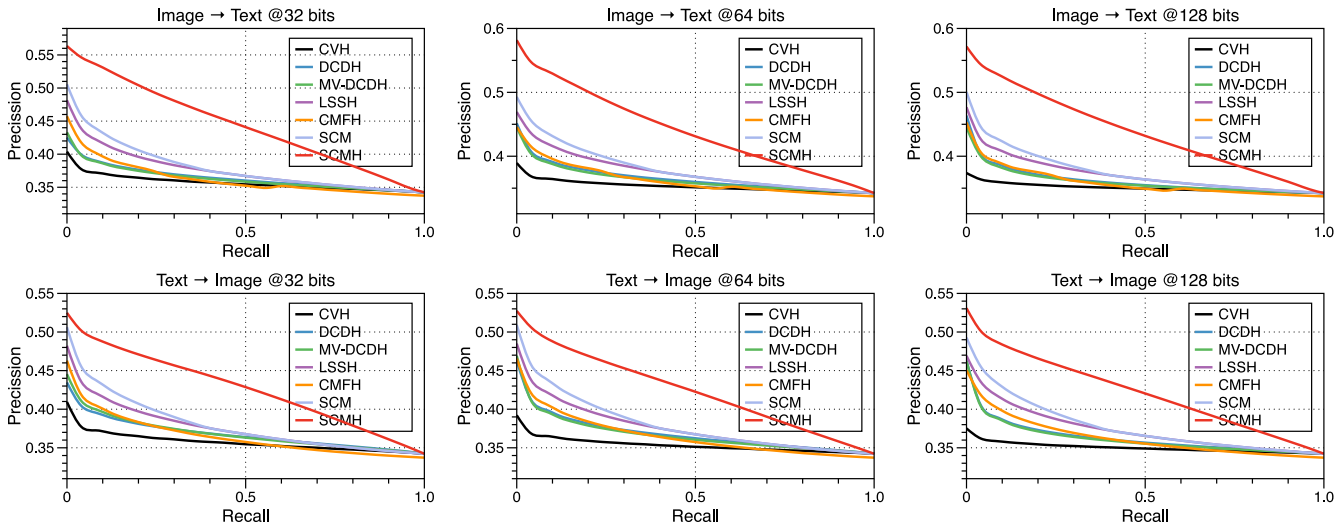


Fig. 8. The precision-recall curves of different hash code generation methods on the NUS-WIDE data set.

It can, in some degree, demonstrate that the proposed method SCMH is more appropriate for practical environment. From the PR-curves illustrated in Fig. 8, we observe the similar phenomenon as LabelMe and Flickr that the precision of SCMH decreases much more slowly. When recall achieves 20 percent, the relative improvement of SCMH over LSSH is more than 28.2 percent on Image → Text task.

To further analyze the results given by different methods, we calculate the cosine similarities between textual descriptions of queries and correct results in the top 50 lists on the Image → Text. Fig. 9 shows the distribution of cosine similarities. In the figure, x-axis denotes the ranges of cosine similarity the y-axis the number of correct results in the range. From the results, we can see that SCMH can find more correct results whose cosine similarity with corresponding query are less 10 percent. It can in some degree demonstrates the effectiveness of SCMH in capturing the semantic textual similarities.

In summary, the evaluation results on three data sets demonstrate conclusively that the proposed SCMH method is superior to the state-of-the-art methods when measured using commonly accepted performance metrics on data sets that are commonly used for evaluating cross-media retrieval.

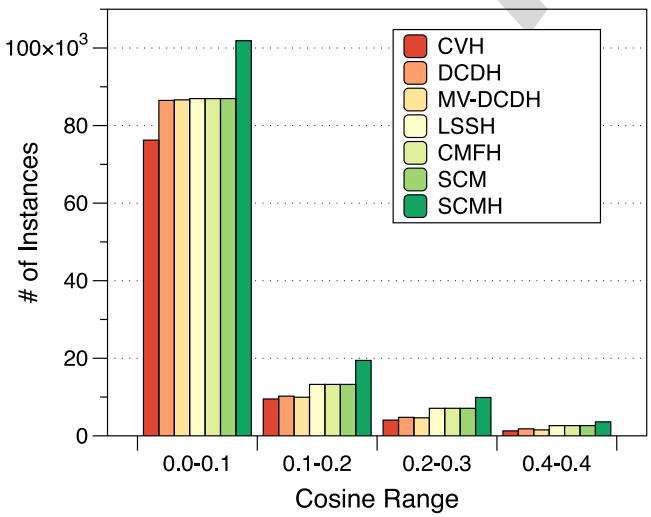


Fig. 9. Distribution of cosine similarities between queries and results on the Image → Text on NUS-WIDE dataset.

4.3.4 *Parameter Sensitivity*

To analysis the sensitivity of the hyper-parameters of SCMH, we conduct several empirical experiments on all the datasets. For easy comparison with previous methods, we set the hash code length to be 64 bits. Fig. 10 shows the performances of SCHM with different percentages of training data. In the two figures, the x-axis denotes the percentages of data used for training and the y-axis denotes the MAP performance. The data used for constructing retrieval set and query set are same as we used in previous section. From the figures, we observe that as the number of training data increases, the MAP performances of SCMH consequently improve on all data sets. When the percentages of training data are over 30 percent of the whole dataset, the MAP performances increase slowly. The main reason may possibly be that the number of categories or concepts included in these data sets are small. However, on the other side, we can say that the proposed method SCMH can achieve acceptable results with a few of ground truths. Hence, it can be easily adopted for achieving other data sets.

Since the training process of mapping function is solved by an iterative procedure, we also evaluate its convergency property. Fig. 11 shows the MAP performances of SCMH on Image → Text and Text → Image tasks. In the two figures, the x-axis denotes the number of iterations for optimizing the mapping function and the y-axis denotes the

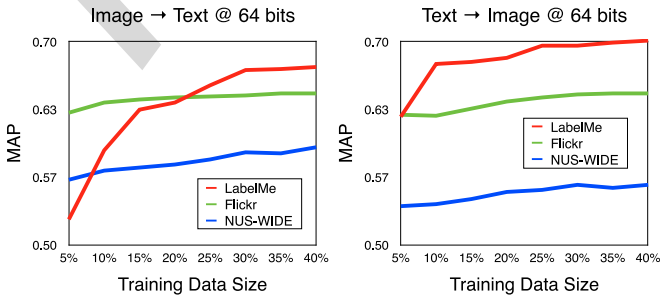


Fig. 10. Effects of training size on MAP performance on the Image → Text and Text → Image tasks.

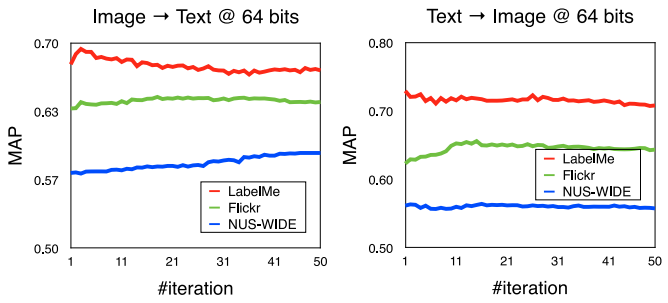


Fig. 11. Effects of the number of iterations on MAP performance on the Image → Text and Text → Image tasks.

MAP performance. From these figures, we observe that SCMH can coverage with less than 10 iterations on all three data sets. It means that SCMH can achieve stable and superior performance under a wide range of parameter values. A strange point occurs on the LableMe dataset on Image → Text task. The result achieve the best with three iterations. The main possible reason is that the size of LabelMe is relatively small comparing to Flickr and NUS-WIDE. Hence, the results may more sensitive on the LabelMe dataset.

4.3.5 Efficiency Evaluation

Due to the requirement of processing huge amounts of data, efficiency is also an important issue. In this work, we compare the running time of the proposed approach with other hashing learning methods. Although the offline stage of the proposed framework requires massive computation cost, the computational complexity of online stage is small or comparable to other hashing methods.

Fig. 12 shows the efficiency comparison of different hashing methods. We implement the all methods to run on single thread in the same machine, which contains Xeon quad core CPUs (2.53 GHz) and 32 GB RAM. All the methods take the text query as inputs. The processing time is calculated from receiving the inputs to generating hash codes. Since in practical usages queries are usually out-of-sample ones, we compare the proposed method with Spectral Hash and Semantic hash. For processing out-of-sample extension of spectral hashing, we propose to use the Nystrom method [64] to do it. From the results, we can observe that the computational complexity of the proposed method is comparable with and state-of-the hashing methods. Comparing to the methods based on the matrix factorization, the proposed method is much more efficient. In this work, we use semantic hash to generate hash codes of FVs. Hence, additional processing time is required to perform the calucuation. However, if we use less complex hashing method, the efficiency can be further improved. It demonstrates that the proposed method is applicable for large scale applications.

5 Conclusions

In this work, we propose a novel hashing method, SCMH, to perform the near-duplicate detection and cross media retrieval task. We propose to use a set of word embeddings to represent textual information. Fisher kernel framework is incorporated to represent both textual and visual information with fixed length vectors. For mapping the Fisher vectors of different modalities, a deep belief network is

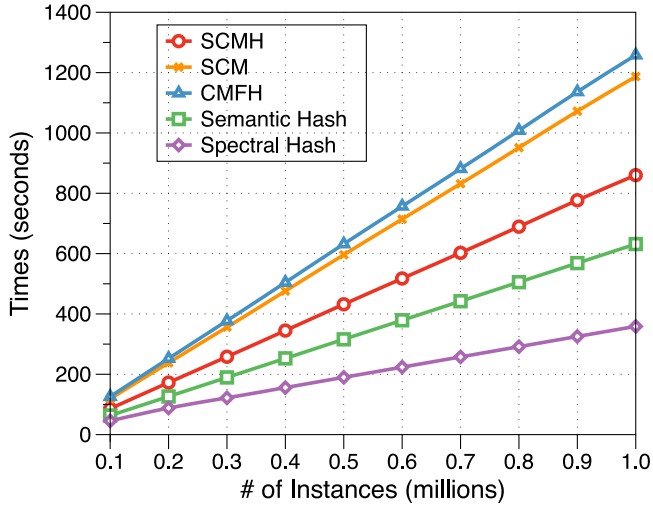


Fig. 12. The efficiency comparison of different hashing methods.

proposed to perform the task. We evaluate the proposed method SCMH on three commonly used data sets. SCMH achieves better results than state-of-the-art methods with different the lengths of hash codes. In NUS-WIDE data set, the relative improvements of SCMH over LSSH, which achieves the best results in these datasets, are 10.0 and 18.5 percent on the Text → Image and Image → Text tasks respectively. Experimental results demonstrate the effectiveness of the proposed method on the cross-media retrieval task.

Acknowledgments

This work was partially funded by National Natural Science Foundation of China (No. 61532011, 61473092, and 61472088), the National High Technology Research and Development Program of China (No. 2015AA015408), and Shanghai Science and Technology Development Funds (13dz226020013511504300). J. Qian is the corresponding author.

References

[1] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2083–2090.
[2] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.
[3] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, "Click-through-based cross-view learning for image search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 717–726.
[4] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. Int. Conf. Manage. Data*, 2013, pp. 785–796.
[5] D. Zhai, H. Chang, Y. Zhen, X. Liu, X. Chen, and W. Gao, "Parametric local multimodal hashing for cross-view similarity search," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2754–2760.
[6] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2177–2183.
[7] Y. Zhuang, Y. Yang, F. Wu, and Y. Pan, "Manifold learning based cross-media retrieval: A solution to media object complementary nature," *J. VLSI Signal Process. Syst. Signal, Image Video Technol.*, vol. 46, pp. 153–164, 2007.
[8] F. Wu, H. Zhang, and Y. Zhuang, "Learning semantic correlations for cross-media retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 1465–1468.

[9] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[10] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, 2005, pp. 633–641.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[12] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[13] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 415–424.

[14] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 395–404.

[15] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1360–1365.

[16] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 406–414.

[17] Q. Zhang, J. Kang, J. Qian, and X. Huang, "Continuous word embeddings for detecting local text reuses at the semantic level," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 797–806.

[18] T. Jaakkola, D. Haussleret al., "Exploiting generative models in discriminative classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 487–493.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis.*, 1999, p. 1150.

[20] X. Wang, Y. Liu, D. Wang, and F. Wu, "Cross-media topic mining on wikipedia," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 689–692.

[21] H. Zhang, J. Yuan, X. Gao, and Z. Chen, "Boosting cross-media retrieval via visual-auditory feature analysis and relevance feedback," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 953–956.

[22] P. Daras, S. Manolopoulou, and A. Axenopoulos, "Search and retrieval of rich media objects supporting multiple multimodal queries," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 734–746, Jun. 2012.

[23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Boston, MA, USA, Jun. 2015, pp. 3128–3137.

[24] A. Z. Broder, "On the resemblance and containment of documents," in *Proc. SEQUENCES*, 1997, p. 21.

[25] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Proc. Combinatorial Pattern Matching*, 2000, pp. 1–10.

[26] M. Theobald, J. Siddharth, and A. Paepcke, "Spotsigs: Robust and efficient near duplicate detection in large web collections," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 563–570.

[27] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe, "Collection statistics for fast duplicate document detection," *ACM Trans. Inf. Syst.*, vol. 20, no. 2, pp. 171–191, 2002.

[28] A. Kołcz, A. Chowdhury, and J. Alspector, "Improved robustness of signature-based near-replica detection via lexicon randomization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 605–610.

[29] J. Seo and W. B. Croft, "Local text reuse detection," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 571–578.

[30] Q. Zhang, Y. Zhang, H. Yu, and X. Huang, "Efficient partial-duplicate detection based on sequence matching," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 571–578.

[31] J. W. Kim, K. S. Candan, and J. Tatemura, "Efficient overlap and content reuse detection in blogs and online news articles," in *Proc. Int. Conf. World Wide Web*, 2009, pp. 571–578.

[32] Y. Gong, S. Kumar, V. Verma, and S. Lazebnik, "Angular quantization-based binary codes for fast similarity search," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1196–1204.

[33] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[34] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[35] K. Grauman and R. Fergus, "Learning binary hash codes for large-scale image search," in *Proc. Mach. Learn. Comput. Vis.*, 2013, pp. 49–87.

[36] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008.

[37] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 940–948.

[38] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: The state-of-the-art," *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, 2015.

[39] M. Norouzi and D. Fleet, "Minimal loss hashing for compact binary codes," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 353–360.

[40] D. Zhang, J. Wang, D. Cai, and J. Lu, "Self-taught hashing for fast similarity search," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 18–25.

[41] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 127–135.

[42] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Mach. Learn.*, vol. 7, pp. 195–225, 1991.

[43] G. E. Hinton, "Learning distributed representations of concepts," in *Proc. 8th Annu. Conf. Cognitive Sci. Soc.*, 1986, pp. 1–12.

[44] G. Hinton and R. Salakhutdinov, "Discovering binary codes for documents by learning deep generative models," *Topics Cognitive Sci.*, vol. 3, pp. 74–91, 2010.

[45] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, 2010, pp. 1045–1048.

[46] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 801–809.

[47] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proc. 48th Annu. Meeting Assoc. Comput. Ling.*, 2010, pp. 384–394.

[48] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.

[49] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Ling.: Human Lang. Technol.-Vol. 1*, 2011, pp. 142–150.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.

[52] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.

[53] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. Comput. Vis.*, 2010, pp. 140–153.

[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Workshop ICLR*, 2013, pp. 1–2.

[55] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proc. 50th Annu. Meeting Assoc. Comput. Ling.*, 2012, pp. 873–882.

[56] J. Jost and J. Jost, *Riemannian Geometry and Geometric Analysis*. New York, NY, USA: Springer, 2008.

[57] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, 1984.

[58] S. Amari and H. Nagaoka, *Methods of Information Geometry*, American Mathematical Soc., 2000.

[59] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 902–909.
[60] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, pp. 145–175, 2001.
[61] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-wide: A real-world web image database from national university of singapore," in *Proc. ACM Conf. Image Video Retrieval*, pp. 48:1–48:9.
[62] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[63] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2011.
[64] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet, "Learning eigenfunctions links spectral embedding and kernel PCA," *Neural Comput.*, vol. 16, pp. 2197–2219, 2004.

Qi Zhang received the PhD degree in computer science from Fudan University. He is an associate professor of computer science at Fudan University, Shanghai, China. His research interests include natural language processing and information retrieval.

Yang Wang received the bachelor's degree in computer science from Xidian University. He is currently working toward the master's degree at Fudan University. His research interests include information retrieval.

Jin Qian received the master's degree in computer science from Shandong University. He is currently working toward the PhD degree at Fudan University. His research interests include data mining.

Xuanjing Huang received the PhD degree in computer science from Fudan University. She is a professor of computer science at Fudan University, Shanghai, China. Her research interests include natural language processing and information retrieval.