# Learning of Multimodal Representations With Random Walks on the Click Graph

Fei Wu, Xinyan Lu, Jun Song, Shuicheng Yan, *Senior Member, IEEE*, Zhongfei (Mark) Zhang, Yong Rui, *Fellow, IEEE*, and Yueting Zhuang

***Abstract*— In multimedia information retrieval, most classic approaches tend to represent different modalities of media in the same feature space. With the click data collected from the users' searching behavior, existing approaches take either one-to-one paired data (text–image pairs) or ranking examples (text–query–image and/or image–query–text ranking lists) as training examples, which do not make full use of the click data, particularly the implicit connections among the data objects. In this paper, we treat the click data as a large click graph, in which vertices are images/text queries and edges indicate the clicks between an image and a query. We consider learning a multimodal representation from the perspective of encoding the explicit/implicit relevance relationship between the vertices in the click graph. By minimizing both the truncated random walk loss as well as the distance between the learned representation of vertices and their corresponding deep neural network output, the proposed model which is named multimodal random walk neural network (MRW-NN) can be applied to not only learn robust representation of the existing multimodal data in the click graph, but also deal with the unseen queries and images to support cross-modal retrieval. We evaluate the latent representation learned by MRW-NN on a public large-scale click log data set Clickture and further show that MRW-NN achieves much better cross-modal retrieval performance on the unseen queries/images than the other state-of-the-art methods.**

***Index Terms*— Cross-media search, click log, latent representation, deep learning.**

## I. Introduction

WITH the rapid growth of multimedia data, cross-media retrieval is imperative to many applications of practical interest, such as finding a set of images that visually best illustrate a given text query. However, the *semantic-gap* between the low-level features and high-level semantics as well as the *heterogeneity-gap* between multimodal data have been widely understood as a fundamental barrier to successful cross-media retrieval. To reduce these gaps, a typical way is to map the multimodal data into a common semantic feature space, and then the retrieval procedure can be conducted in the newly mapped space. For example, automatic annotation translates the images from the image space into the text space to support the image retrieval from text queries.

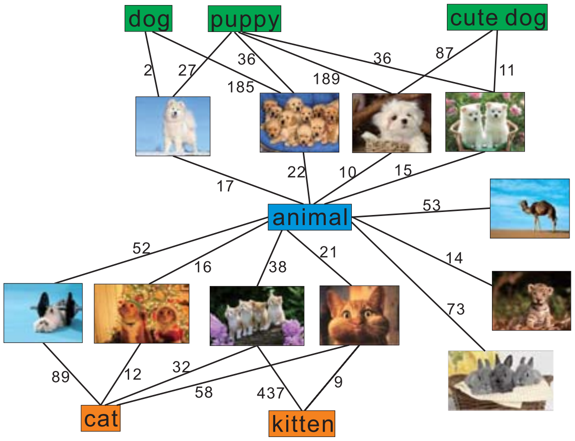


Fig. 1. An example of the subgraph of the click graph from a commercial image search engine [2]. Vertices are queries or images and the edges indicate the click count of the image given the text query.

In recent years commercial image search engines, such as Google[1] and Bing,[2] record the users' behavior from their queries and clicks, with little overhead [1]. The objective of recording click logs is to help improve the performance in image search engines by leveraging click data aggregated across users and sessions [2]. The click data is usually stored in a large table, with each line containing a triad $(D, Q, C)$. A triad $(D, Q, C)$ means that the image $D$ was clicked $C$ times in the search results of textual query $Q$. The click data can be also viewed as a bipartite graph, with two types of vertices (queries and images) and the edges weighted according to the total number of clicks from all the users. An example of a subgraph of the click graph is depicted in Figure 1. Intuitively, the more clicks, the more relevant the image w.r.t. the query. Therefore, the query-image pairs can be viewed as "soft" relevance judgments to help bridge the heterogeneity-gap. Compared with the other types of training data such as

Manuscript received May 7, 2015; revised September 17, 2015 and October 28, 2015; accepted November 30, 2015. Date of publication December 9, 2015; date of current version December 28, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guoliang Fan.

F. Wu, X. Lu, J. Song, and Y. Zhuang are with the School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: wufei@cs.zju.edu.cn; xinyanlu@zju.edu.cn; yzhuang@cs.zju.edu.cn).

S. Yan is with the Electrical Engineering Department, National University of Singapore, Singapore 119077 (e-mail: eleyans@nus.edu.sg).

Z. Zhang is with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: zhongfei@zju.edu.cn).

Y. Rui is with Microsoft Research Asia, Beijing 100080, China (e-mail: yongrui@microsoft.com).

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/TIP.2015.2507401

[1] https://images.google.com/

[2] http://www.bing.com/images/

manually labeled image datasets like ImageNet [3] and text-image symbiosis datasets like Wikipedia feature articles [4], the click data has several advantages:

1) users can quickly glance at the returned image thumbnails before they click, and thus the click logs have a relatively low level of noise: it was estimated that about 84% of the images were relevant among all the clicked images [5];
2) the high-quality click data is harvested by the collective intelligence of the users with no extra effort from the users;
3) the click data from commercial search engines increases all the time by the use of search engines, and can cover all the query intents from the users, including the new emerging topics and concepts.

Therefore, the click data has attracted a great deal of research devoted to the development of algorithms for learning an optimal common representation of different modalities. According to the difference on modeling the click data, the existing approaches can be categorized into two classes. One class of the approaches (e.g., [6], [7]) models the click data as a set of (weighted) query-image paired data, of which the training objective is that a more clicked query-image pair in the mapped latent space should be closer. The other class of the approaches is based on the techniques of *learning to rank* (e.g., [8], [9]), which model the click data as a set of cross-modal ranking examples (here one text-query-image ranking example consists of a text query and its corresponding ranked images). Given the training ranking examples, these approaches tend to learn a common latent space in which the distance of the mapped images to the query is in accordance with their relevance to the query.

However, all the aforementioned approaches only explore a limited part of the click data, i.e., the explicit link structure of the click graph, in which the implicit relationship among the vertices in the click graph is ignored. For example, two images both clicked by the same query may have both high visual similarity and high semantic similarity; two queries that click the same image may be highly related to each other (e.g., "labradoodle dog" and "puppy"). We argue that such approaches disregarding the implicit link structure of the click graph lead to an inferior performance on learning the common latent representation of the multimodal data.

We focus on learning the common latent representation of multimodal data in this paper. Moreover, we aim to learn the mapping to the latent space such that cross-modal retrieval can be performed, considering the utilization of the multimodal click data. The learned space is constrained to be a low-dimensional continuous space since the intrinsic dimensionality of a semantic space is usually much lower that that of original feature space. The latent representations should encode the relevance relation between the queries and the images accurately. That is, the indirect relationship between the vertices in the click graph should also be considered in learning the latent space such that the learned representation can capture more accurate semantics of both queries and images. More importantly, the learned model should be also generalized to new queries and new images such that the model can be applied to not only the training data but also the unclicked / emerging queries and images.

Our work extends DeepWalk proposed in [10]. In general, we seek to bridge the gap between multimodal click graph modeling and deep neural networks. By modeling the multimodal click graph by a stream of short random walks and adapting techniques of deep neural networks, we present an end-to-end solution, named *Multimodal Random Walk Neural Network* (MRW-NN), that takes a multimodal click graph as input to learn the common latent representation of text and imagery. The overview framework of MRW-NN is depicted in Figure 3. Specifically, each vertex in the click graph is associated with two representations: one (the *social* representation) encoding its context (neighborhood similarity) in the graph, and the other one (the *internal* representation) representing the semantics of the vertex by analyzing its content with deep neural networks. By optimizing both the truncated random walk loss as well as the distance between the social representation and the internal representation of the vertices, the social representations of the vertices and the parameters of the deep neural networks are learned. Thus the proposed model not only captures more accurate semantics of the training queries and images, but also generalizes to the unseen queries and images better. When used for cross-modal retrieval, MRW-NN simply outputs the latent representation of queries and images by the deep neural networks, with then the images are ranked by their distance from the query in the latent space.

The contribution in this work is summarized as follows:

1) Not only the proposed model learns high-level feature representation for data objects with different modalities (which reduces the semantic gap), but also the learned representation encodes the direct and indirect relevance relationship among the vertices in the click graph (which reduces the heterogeneity gap).
2) By constraining the distance between the latent representation of the vertices and their neural network output, robust high-level latent representation is learned (for example, the visually similar images are mapped closely). More importantly the loss is back-propagated to train the modality-specific neural networks with end-to-end training, by which the deep neural networks are optimized for encoding the relevance relationship between text and imagery, thus benefiting for performing cross-modal retrieval on the unseen queries and images.
3) We also conduct an empirical study on a commercial image search engine click logs with 11.7 million queries and 1 million images. It is observed that the proposed model has the new state-of-the-art cross-modal ranking performance.

The rest of this paper is organized as follows. In Section II, we introduce related work. In Section III, we describe the method in detail and show its feasibility. We analyze the learned latent representation and compare the proposed model with the existing cross-modal ranking approaches on the real-world multimodal click dataset in Section IV. Conclusions are given at the end.

## II. Related Work

In this section, we discuss the related work on the learning of multimodal click graph.

There has been a great deal of research denoted to the development of algorithms for learning an optimal common representation of different modalities. These popular approaches map the data of multiple modalities into a common space such that the distance between two similar objects is minimized, while the distance between two dissimilar objects is maximized. For example, Rasiwasia *et al.* [4] presents that modeling the correlations between modalities is more effective in feature spaces with higher levels of abstraction. The scalability of incorporating new semantic concepts into the semantic space is studied in [11], which allows the updated embedding function to be applied to dynamic image repositories. Furthermore, since there are a wide variety of visual features to describe an image (e.g., color, texture, shape and spatial layout), graph-based feature fusion techniques ([12], [13]) which explore the complementation of multiple features during the learning process are also investigated, showing better performance than that of using concatenated high-dimensional global feature vector (*early fusion*) or applying different features to learning algorithms and then fusing the results (*late fusion*) [14].

To learn the representation of the vertices in the multimodal click graph, a typical way is to first extract pairs of clicked query-image data from the graph, and then to take the paired data as the input to optimize the specific objective functions. As one of the most popular approaches, Canonical Correlation Analysis (CCA) [15] and its extensions (e.g., Deep CCA [6] and Generalize Multiview Analysis [16]) learn the (non)linear transformation that projects the query textual semantics and image content into the common subspace respectively, by maximizing the correlations between the two variables in the latent space. To utilize the information of the click count, the training of Click-through-based Cross-view Learning (CCL) [7] is performed simultaneously by minimizing the distance between query and image mappings in the latent subspace weighted by their clicks, and preserving the structure relationships between the paired training examples in the original feature space. The underlying assumption of CCL is that the higher the click number, the smaller the distance between the query and the image in the latent space. Furthermore, the similarity between examples in the original space can be preserved in the learned latent subspace.

Motivated by the fact that *learning to rank* methods have the intrinsic power for document retrieval, another category of methods views the click data as a set of cross-modal ranking examples. These methods are based on the techniques of learning to rank and take the ranking examples as the pairwise (or listwise) input to optimize a certain ranking loss. PAMIR [8] is the first attempt to address the problem of ranking images by text queries. PAMIR formulates the cross-modal retrieval problem in a way similar to that of RankSVM and derives an efficient training procedure by adapting the Passive-Aggresive algorithm. A model named DeViSE [17] has a similar ranking loss function to that of PAMIR, while the embedding of multimodal data is performed by deep neural networks. The goal of PAMIR and DeViSE is to minimize the average number of the inversions in ranking; that is, the more clicked images should be ranked higher than the less clicked ones. Unlike the above pairwise approaches, the listwise approaches notice that the ranking is a prediction task on a list of documents and take the ranking lists as the training instances. These methods explicitly minimize the ranking loss of a whole permutation listwise, not pairs of items. The authors of [18] propose a general cross-modal ranking algorithm to optimize the listwise ranking loss with a low rank embedding. The embedding space of queries and images is discriminatively learned by a structural large margin learning for certain ranking criteria (e.g., MAP) directly. Noticing that the click graph can be viewed as not only text-query-image ranking examples but also image-query-text ranking examples, Bi-CMSRM proposed in [19] takes bi-directional ranking examples into account, such that two directions of retrieval are optimized simultaneously, yielding a better representation for multimodal data.

Different from the aforementioned methods that consider only direct connections in the click graph, some approaches consider the sparsity problem of the click graph and try to model the implicit connections among the vertices. These approaches are usually based on the random walk process and obtain a probabilistic distribution over documents describing how likely they are relevant to a given query. Craswell et al. [20] propose a Markov random walk model to a large click log for finding relevant documents, including those that as-yet unclicked for a query, without analyzing the query content or image content. The drawback of the model is that it cannot be applied to the new emerging queries or images. Given a text query, the visual link structure among the related images is considered in the so-called *visual reranking* approaches (e.g., [13], [21], [22]) which help re-rank the initial returned list to make a better relevance score. Assuming that the majority of the related images are relevant to the query, the transition probabilities among the images are first computed by their visual similarities, and then the PageRank algorithm is applied to the image-image graph, assigning the relevance score to an image based on its connections to others. The images found to be "authorities" are chosen that answer the query well. We note that though these reranking approaches cannot deal with the heterogeneity-gap of the multimodal click graph, they can be incorporated with the proposed MRW-NN model to rerank the returned ranking result, which is out of the scope of the presented work.

On the other hand, deep neural networks (DNNs) that learn a transformation of a low-level representation to a high-level representation have shown their powerful ability to the tasks of learning multimodal representation. The authors of [23] propose a multimodal Deep Boltzmann Machine for learning a generative model of data that consists of multiple input modalities. Noting that the multimodal DBM is trained with paired multimodal training data, the model works by learning a joint representation over the space of multimodal inputs, which is useful for cross-modal retrieval. Methods like [24] based on autoencoders and [6] based on CCA have similar incentives. Some other DNNs are based on the cross-modal ranking examples which are trained to learn representation that minimizes a certain ranking loss. The ranking loss is back-propagated into the visual model and the textual model to fine-tune their representations. For example, DeViSE [17] is optimized for the pairwise ranking loss and CMRNN [25] for the listwise ranking loss. The visual model and the textual model in these methods are usually adapted from the promising deep neural networks, like CNN [26] for image representation and word2vec [27] for word representation. The proposed MRW-NN model can be integrated with the deep structure of these methods, while it remains an open question in our further work whether there is a specific (and better) deep structure for modeling the text queries and the images in the click graph.

Our work is closely related to DeepWalk [10] which first suggests the use of random walks to learn latent representation on the social community graph. In this work, we constrain ourselves to learn the latent representation of the multimodal click graph. The main goal of the proposed model is to perform cross-modal ranking, which differs from that of DeepWalk that aims to learning the latent representation for classifying the static members of a social network. Specifically, our work differs from [10], in which not only the relationship among the vertices in the graph is encoded in the latent space, but also the proposed model is required to generalize for the unseen queries and images (those not involved in the training click graph) by learning deep neural networks that transform the unseen data from their original feature space to the common space to support cross-modal retrieval.

## III. MRW-NN Model/Algorithm

We consider the problem of multimodal representation learning. The proposed method MRW-NN learns a general multimodal representation from the training click graph in the sense that it maps the two types of multimodal data into the same common space in which the cross-modal retrieval can be performed.

### A. Notation

Denote $m$ as the dimension of the image feature space (e.g., the number of pixel intensity levels × the number of pixels) and $n$ as the dimension of the text feature space (e.g., vocabulary size of bag-of-words). The dimension of the latent semantic space is denoted as $d$. In this work, the click data is viewed as a bipartite graph $G$ with vertices $D$ denoting the set of images and vertices $Q$ denoting the corpus of text queries. More formally, let $G = (V, E)$, where $V = \{D, Q\}$ are the vertices of the click graph, and $E$ denoting the undirected edges, $E \subseteq (D \times Q)$. The weight of an edge $e_{i,j} = (D_i, Q_j)$ is assigned with the click count representing how many times image $D_i$ is clicked for the text query $Q_j$ (maybe by different users at different times). For clarity, we list important notations and definitions throughout this paper in Table I.

### B. Task: Learning Representations and Mapping Functions

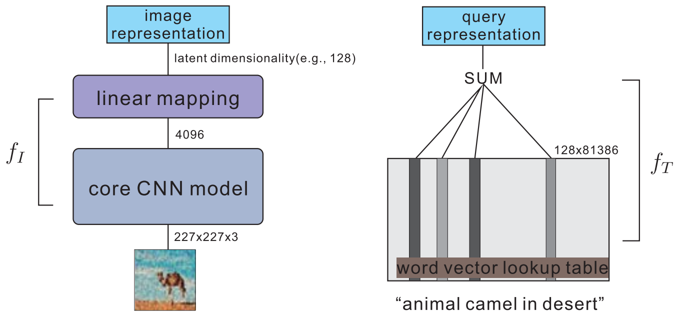
The click graph consists of objects from multiple modalities that come from different feature spaces. We seek to perform cross-modal ranking over the new images and queries that are not involved in the training click graph. Rather than knowing what the representations $\Phi$ of the vertices in the click graph are, we are more interested in *how* the vertices are transformed to the latent representation from their original feature representation. Thus, one of the main goals is to learn the mapping functions $f_I : \mathbb{R}^m \mapsto \mathbb{R}^d$ and $f_T : \mathbb{R}^n \mapsto \mathbb{R}^d$ that map the images and the queries to the common latent space, respectively.

TABLE I
Notations and Definitions

| Notation | Definition |
|---|---|
| $m$ | The dimension of the image feature space. |
| $n$ | The dimension of the query (text) feature space. |
| $d$ | The dimension of the latent semantic space. |
| $D$ | The set of images. |
| $Q$ | The set of queries. |
| $V$ | $V = \{D, Q\}$, the vertices of the click graph. |
| $v$ | $v \in V$, a vertex of the click graph. |
| $E$ | $E \subseteq (D \times Q)$, the undirected edges of the click graph. |
| $G$ | $G = (V, E)$, the bipartite click graph. |
| $e_{i,j}$ | $e_{i,j} = (D_i, Q_j)$, the click count representing how many times image $D_i$ is clicked for the text query $Q_j$. |
| $\Phi(v)$ | The *social* representation of $v$ encoding its context in the graph (that to be learned). |
| $f_I(p)$ | The *internal* representation of image $p$, where $f_I : \mathbb{R}^m \mapsto \mathbb{R}^d$ is the mapping function for images (that to be learned). |
| $f_T(t)$ | The *internal* representation of text query $t$, where $f_T : \mathbb{R}^n \mapsto \mathbb{R}^d$ is the mapping function for text queries (that to be learned). |
| $f(v)$ | When $v$ is an image, $f(v) = f_I(v)$; otherwise when $v$ is a text query, $f(v) = f_T(v)$. |
| $W_v$ | The random walk rooted at vertex $v \in V$. |
| $P_{i,i+1}$ | The probability of moving from vertex $v_i$ to vertex $v_{i+1}$ in random walks. |
| $w$ | The pre-defined window size which controls the neighborhood of a vertex. |
| $C$ | The parameter which controls the trade-off between the random walk error and the regularization penalty. |

We argue that an appropriate multimodal modeling approach over the click data should have the following characteristics:

- *Click Aware*: The more the explicit clicks between a query and an image, the closer their latent representation should be. The mapped representation should also encode the implicit connections between the vertices in the click graph. In other words, the distance of the vertices in the latent space should represent the semantic similarity between them.
- *Intra-modal Consistency*: While there's no explicit connection between the intra-modal nodes in the bipartite click graph, the visual similar images should mapped closely, as well as the queries that have similar semantics.
- *Generalization Aptitude:* It is insufficient to learn latent representation for the present members of the click graph only and the proposed modal should be able to perform cross-modal ranking in the future. Most importantly, the mapping functions should generalize for the unseen images and emerging queries well. To support cross-modal retrieval, given an image $p \in \mathbb{R}^m$ and a text query $t \in \mathbb{R}^n$, we consider their relevance measured by the cosine similarity of the two mapped vectors in



(A) visual mapping function (B) texual mapping function

Fig. 2. (A) the image mapping function $f_I$ is composed by a deep CNN model with the softmax layer replaced with a full-connected layer; (B) a query's representation $f_T$ is the summation of the representations of the contained keywords. Both the parameters of $f_I$ and $f_T$ are learned from the training click data.

the $d$-dimensional space:

$$\mathrm{Rel}(p, t) = \frac{f_I(p)^\top f_T(t)}{\|f_I(p)\| \, \|f_T(t)\|} \tag{1}$$

which is commonly used to measure the matching between textual vectors [28] ($f_I$ and $f_T$ are the mapping functions for images and textual queries respectively, which we will discuss shortly).

where we also constrain the learned representations to be continuous and low dimensional for robust statistical learning.

The hypothesis of the mapping functions in this work is depicted in Figure 2. Inspired by the recent advances of deep neural networks for learning representation, we employ the architecture of the deep convolutional neural network (DCNN) proposed in [26] for image modeling (however our work could be also integrated with any powerful DCNNs, e.g., a 16-layer VGGNet model [29]). The DCNN model consists of several convolutional filtering, local contrast normalization and max-pooling layers, followed by several fully connected neural network layers. The softmax prediction layer is replaced with a linear transformation that maps the 4,096-dimensional representation into the latent $d$-dimensional representation at the top. Thus, the images are fed through the visual model to be represented as real-valued feature vectors. For query modeling, we adapt the idea of learning distributed representations of words: each word in the vocabulary is embedded into a vector lookup table in such way a $d$-dimensional distributed representation is associated with each word, which is to be learned. Then a query with multiple words is represented as the summation of the corresponding word vectors. Therefore, the images and the queries can be both represented in the $d$-dimensional latent space.

The rest is to learn the parameters of both the visual model and the word vectors from the training click data to satisfy the above requirements. For simplicity, we denote the mapping function as $f$ (omitting the subscript of $f_I$ and $f_T$) when it is capable of being applied to the queries or the images.

### C. The Proposed Model

To meet the above characteristics, one straightforward consideration is a two-stage mechanism: firstly get the vertices' social representation with methods like Lapacian Eigenmap [30] or DeepWalk [10] without considering the content of the queries and images; and then train individual neural networks separately for different modality with the loss function as Euclidean distance between the neural network output and the previous learned latent representation. However, this cannot guarantee to satisfy all the above requirements, since two visual similar images or two semantic queries may be mapped far to each other in the previous step, resulting in the slow convergence of the training of deep neural networks as well as the limited generalization performance in cross-modal retrieval.

Instead, we propose an end-to-end learning solution and formulate our objective in the view of well-known "*empirical risk + regularization*" framework.

First consider capturing the structure of the click graph only, i.e., learning the latent representation of the vertices in the click graph without analysis of their content. Inspired by the DeepWalk model [10], the representation of the vertices of the click graph is learned from a stream of truncated random walks, using optimization techniques originally designed for language modeling. In this way, vertices which have similar neighborhoods will acquire similar representations.

Denote a random walk rooted at vertex $v \in V$ as $W_v$. It is a stochastic process with random variables $v_1, v_2, \ldots, v_k$ such that $v_{k+1}$ is a vertex chosen at random from the neighbors of vertex $v_k$. The random walk is performed on the weighted bipartite graph. The click count (i.e., the weight of an edge) serves as an important role in the click graph which measures the relevance between a query and an image. The transition probability $P_{i,i+1}$ (the probability of moving from vertex $v_i$ to vertex $v_{i+1}$ in the Markov random walk) is defined as

$$P_{i,i+1} = \frac{e_{i,i+1}}{\sum_j e_{i,j}} \tag{2}$$

and thus, a more clicked vertex is more likely to be chosen.

Then each walk sequence is treated as a sentence (thus a vertex as a word) and the likelihood of observing the neighbors of vertex $v_i$ in the walk is maximized (as an analogy to the context of a word in a sentence). The proposed optimization problem is given as

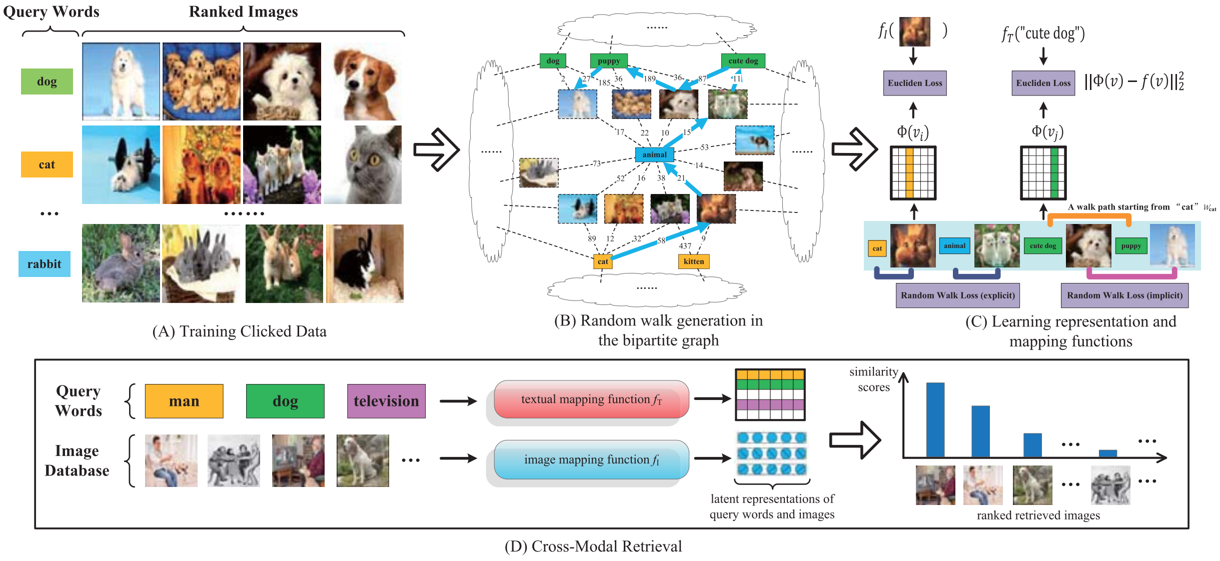
$$\max_{\Phi} \log \Pr(\{v_{i+1}, \ldots, v_{i+w}\} | v_i) \tag{3}$$

where $w$ is the predefined window size which restricts the size of the random walk context. Like the SkipGram model, the ordering constraint is removed and the objective is transformed to:

$$\max_{\Phi} \sum_{1 \le j \le w} \log \Pr(v_{i+j} | v_i)$$

with $\Pr(v_{i+j}|v_i)$ defined using the softmax function:

$$\Pr(v_{i+j}|v_i) = \frac{\exp(\Phi(v_{i+j})^\top \Phi(v_i))}{\sum\limits_{v} \exp(\Phi(v)^\top \Phi(v_i))}, \quad 1 \le j \le w \tag{4}$$



(A) Training Clicked Data

(B) Random walk generation in the bipartite graph

(C) Learning representation and mapping functions

(D) Cross-Modal Retrieval

Fig. 3. In this work, each vertex $v$ in the click graph is associated with two representations: one (the *social* representation $\Phi(v)$) encoding its context (neighborhood similarity in its vicinity) in the graph, and the other one (the *internal* representation $f(v)$) representing the semantics of $v$ by analyzing its content. The proposed method MRW-NN consists of two stages in the training process: (1) first generate random walk paths with transition probability $P_{i,i+1}$ defined in Eq. (2) on the click graph; (2) given walk paths, the proposed method attempts to minimize the random walk error as well as the difference between the learned representation $\Phi(v)$ and the modality-specific neural network output $f(v)$. As is depicted above (with the window size $w = 2$), the model considers both the explicit and implicit connections among the vertices. Vertices with the same color have the same semantics. The random walk loss and the Euclidean loss are back-propagated to train the parameters of the visual mapping function $f_I$, the textual mapping function $f_T$, and the latent representation $\Phi$. In the prediction process, the learned $f_I$ and $f_T$ are utilized to perform cross-modal retrieval. In this Figure, (a),(b) and (c) are training process while (d) is the prediction process. It should be noted that $f_I$, $f_T$ and $\Phi$ are trained over the training set and utilized for the test data.

where $\Phi(v)$ denotes the *social* representation of the vertex $v$ that captures neighborhood similarity and encodes social relations of $v$ in a continuous space, which is represented by a row in a $|V| \times d$ matrix of free parameters. Solving the Optimization Problem (3) will assign those vertices which have similar neighborhoods with similar social representation $\Phi$ (refer to [10] and [27]). The full softmax can be efficiently approximated with Negative Sampling or Hierarchical Softmax [31].

Unlike DeepWalk that computes the vector representation of both a vertex and its "context", we define $\Pr(v_{i+j}|v_i)$ as follows which does not distinguish the representation of a vertex and its "context". Here we regard a vertex's "context" as the vertex itself, since we find that this constraint not only is particularly desirable for cross-modal retrieval (recalling that the relevance scoring function in Equation (1)), but also reduces the complexity of the model.

The proposed model is required to maximize the probability of observing only vertices appearing to the right side of the given vertex in the random walk, rather than that appearing to both sides in the DeepWalk model (refer to Equation (3)). We note that this is equivalent to the original strategy of DeepWalk by not distinguishing the representation of a vertex and that of its "context" and approximating Equation (4) with Negative Sampling. This is a natural way to model the random walk process and reduces the computation cost by half.

More formally, the procedure is as follows: given a click graph with $|V|$ vertices, for each vertex $v$, a Markov random walk $W_v$ is performed on the click graph started at $v$ with fixed length $L$; the transition probability is defined in Equation (2). The objective is to maximize the likelihood of observing the walks:

$$\max_{\Phi} \sum_{v \in V} \sum_{v_i \in W_v} \sum_{1 \leq j \leq w} \log \Pr(v_{i+j}|v_i) \tag{5}$$

which serves as the *empirical risk* term in the loss function.

Second, we consider analyzing the content of the vertices. Recall that the regularization technique is used to control the over-fitting phenomenon, which involves adding a penalty term to the error function. From a Bayesian point of view, many regularization techniques correspond to imposing certain prior distributions to the model parameters, e.g., L2 regularization assumes a Gaussian prior with zero mean [32]. The parameter $\Phi$ to learn in Equation (5) is the latent representation of the queries and the images in the click graph. As is discussed above, the visually similar images or the semantic similar queries should be mapped closely. Thus, we would like to add a prior on $\Phi$ which depends on the content of the vertices and the mapping function, resulting in the following regularized optimization problem:

$$\begin{aligned} &\min_{\Phi, f} \sum_{v \in V} \sum_{v_i \in W_v} \sum_{1 \leq j \leq w} - \log \Pr(v_{i+j}|v_i) \\ &\text{s.t. } \|\Phi(v) - f(v)\|_2 \leq C, \quad v \in V \end{aligned} \tag{6}$$

where $C \geq 0$ controls the trade-off between the random walk error and the regularization penalty and $f$ is a (non)linear mapping function that maps a vertex from its original feature space to the latent semantic space by analyzing its content. Ideally for any vertex $v$, the social representation $\Phi(v)$ and the internal representation $f(v)$ would be the same in the semantic space, since it is the content of the vertex that determines the context of the vertex in the click graph (e.g., to be clicked or not, by a specific text query). While in the case that the click graph is very noisy, $C > 0$ allows a tolerance of click noise (e.g., very strange images clicked by mistake) and helps to generalize. By solving the Optimization Problem (6), the parameters of the visual mapping function $f_I$, the textual mapping function $f_T$, and the latent representation $\Phi$ can be obtained. We also require $f$ to be "smooth" such that two similar intra-modal vertices are mapped together closely (which is not presented in the regularized optimization problem for simplicity).

We reiterate here that Equation (6) is attractive to deal with the clicked data rather than other data, since a random walk is generated on the click graph started at any vertex $v$, the social representation of $v$ (e.g., $\Phi(v)$) and the internal representation of $v$ (e.g., $f(v)$) would be the same in the semantic space due to the enforced regularized term.

We discuss two special cases of the regularized Optimization Problem (6) here. In the case of $C = +\infty$, the optimization problem is (almost) equivalent to that of the DeepWalk model which only learns the latent representation of the vertices and ignores the content of the vertices. In the other case of $C = 0$, we get the following interesting optimization problem:

$$\min_f \sum_{v \in V} \sum_{v_i \in W_v} \sum_{1 \le j \le w} -\log \bar{\Pr}(v_{i+j}|v_i) \tag{7}$$

where $\bar{\Pr}(v_{i+j}|v_i) = \frac{\exp(f(v_{i+j})^\top f(v_i))}{\sum\limits_v \exp(f(v)^\top f(v_i))}$, which means that the latent representation should be exactly the output of the mapping function. The Optimization Problem (6) can be also viewed as the *slack* version of the Optimization Problem (7).

### *D. Algorithm and Implementation*

To solve the regularized optimization problem (6), the general optimization procedure of MRW-NN alternates between two steps, one generating random walks and the other updating the parameters of $f_I$, $f_T$ and $\Phi$, as listed in Algorithm 1. The procedure runs at most $\gamma$ epochs over the training data. At the start of each epoch, a random ordering of all the vertices is performed so as to speed up the convergence of stochastic gradient descent. The random walk $W_{v_i}$ starting at vertex $v_i$ samples a vertex from the neighbors of the last visited vertex with the transition probability defined in Equation (2), until the maximum length $L$ is reached.

Given a walk sequence, the representation $\Phi(v)$ is first projected onto the set

$$B(f(v), C) = \{\mathbf{x} : \|\mathbf{x} - f(v)\|_2 \le C\}$$

where $B$ is the candidate set that satisfies the regularization constraint in Optimization Problem (6). Specifically, the projection is obtained by

$$\Phi(v) \leftarrow f(v) + (\Phi(v) - f(v)) \cdot \min\{1, \frac{C}{\|\Phi(v) - f(v)\|}\} \tag{8}$$

Then the SkipGram algorithm is used to update these representations $\Phi(v)$ in accordance with the objective function in Equation (5), as listed in Algorithm 2. We use Negative Sampling to approximate the softmax likelihood function (4), with $J(\Phi)$ replaced with

$$-(\log \sigma(\Phi(u_j)^\top \Phi(v_i)) + \sum_{r=1}^{R} E_{u_r \sim P(v)}[\log \sigma(-\Phi(u_r)^\top \Phi(v_i))])$$

where $R$ is the number of randomly selected negative samples and $P(v)$ is the probability of vertex $v$ appearing in a random walk.

**Algorithm 1** Multimodal Random Walk Neural Network Model (MRW-NN)

```
Input: a bipartite click graph $G = (V, E)$ with multimodal data vertices, number of epochs $\gamma$, window size $w$, length of random walk $L$, trade-off control parameter $C \ge 0$, the latent space dimension $d$
Output: mapping function $f_I$ and $f_T$, the social representation $\Phi$ of $V$
Initialization: initialize the parameters of $f_I \in \mathbb{R}^{m \times d}$ and $f_T \in \mathbb{R}^{n \times d}$, and sample $\Phi$ from $\mathbb{R}^{|V| \times d}$ (see the experiment part in Section IV-B for more details)
for $t = 1$ to $\gamma$ do
    $\mathcal{O} = Shuffle(V)$ // order the vertices randomly
    for each $v_i \in \mathcal{O}$ do
        $W_{v_i} = RandomWalk(G, v_i, L)$ // generate a random walk $W_{v_i}$ starting at vertex $v_i$, where the transition probability is given by Equation (2)
        Obtain the internal representation $f_t(v)$ of each vertex in $W_{v_i}$ via $f_I$ or $f_T$
        Compute $\Phi_{t+\frac{1}{2}}$ to be the projection of $\Phi_t$ onto the set $B(f(v), C)$, via Equation (8)
        Given $\Phi_{t+\frac{1}{2}}$, back propagate the empirical random walk error (given by Equation (5)) to calculate the gradient $\Delta\Phi$
        Update $\Phi_{t+1} = \Phi_{t+\frac{1}{2}} + \alpha * \Delta\Phi$, where $\alpha$ is an adjustable learning rate
        Similarly, back propagate the Euclidean loss $\frac{1}{2}\|\Phi_{t+1} - f_t\|_2^2$ to update $f_{t+1}$
    end for
end for
```

**Algorithm 2** SkipGram($\Phi$,$W_v$,$w$)

```
Input: current latent representation $\Phi$, walk sequence $W_v$, window size $w$
Output: updated latent representation $\Phi$
for each $v_i \in W_v$ do
    for each $u_j \in W_v[i+1, i+w]$ do
        $J(\Phi) = -\log \Pr(u_j|v_i)$
        $\Phi = \Phi - \eta \cdot \frac{\partial J}{\partial \Phi}$
    end for
end for
```

TABLE II
THE STATISTICS OF CLICKTURE DATASET

| | Train | Dev |
|---|---|---|
| # of images | 1M | 79665 |
| # of queries | 11.7M | 1000 |
| # of image/query pairs | 23.1M | 79926 |
| # of unique terms | 915K | / |
| # of unique terms (after preprocessing[a]) | 81K | / |
| # of clicks | 82.3M | / |

[a] We do word stemming and filter the stop words (like "the", "photo" and "pic") and the rare words with occurrence less than 10.

Then the parameters of the mapping function are updated by the back-propagation algorithm with the top derivative of the regularized term as $\Delta = f(v) - \Phi(v)$.

The code is implemented upon Caffe [33] and its MATLAB interface. To employ the computation capabilities efficiently, we use the data parallelism training. Each mini-batch consists of 200 random walks of length $L$. In the training period of a mini-batch, the negative samples of a vertex in a given walk are obtained from the vertices appearing in the other walks.

## IV. EXPERIMENTS

The main goal of the experiments is to evaluate the effectiveness of the learned representation and the mapping functions by the proposed MRW-NN model. To show its competitive performance, MRW-NN is compared with the other state-of-the-art approaches for cross-modal retrieval.

### A. Dataset

To the best of our knowledge, the Clickture dataset [2] is the only public, large-scale multimodal click log dataset, which is collected from one year click-through data of one commercial image search engine. To make our experiments reproducible and comparable, we conduct the experiments on the Clickture dataset. The statistics of the Clickture dataset is shown in Table II. The dataset comprises two parts, i.e., the training and development (dev) sets. The training set consists of 1 million images and 11.7 million unique queries. The click count between an image and a query is summed from different users at different times. Among all the image-query pairs, there are 23.1 million of them with click count equal or more than 1. Figure 1 shows a few exemplar images with their clicked queries and click counts in the Clickture dataset. It is worth noting that there is no any other information (e.g., user information, surrounding text and time stamp of click) provided in the Clickture dataset.

In the dev dataset, there are 79,926 query-image pairs generated from 1,000 queries, where each image to the corresponding query is manually annotated on three relevance measurement: Excellent, Good, and Bad. In the experiments, the training set is used for learning the mapping functions, while the dev set is used for performance evaluation.

### B. Experimental Settings

*1) Parameter Tuning:* For the proposed MRW-NN, the parameters to adjust are respectively the learning rate, the dimensionality of the latent space $d$, the input window size $w$, the length of walk path $L$, and the trade-off $C$ between the empirical risk and the prior of Optimization Problem (6). Empirically, $d$ is set to be a multiple of 32 (the size of GPU warp). Some parameters should be fixed during the training procedure like the dimensionality of the latent space $d$, while others can be adjusted (or better to be adjusted) like the learning rate. Many choices are almost equally as good; we end up with a particular choice of parameters with $d = 128$, $w = 2$, $L = 10$, $C = 0.0001$.

*2) Neural Network Settings:* We use the AlexNet model previously trained on ILSVRC 2012 [26] to initialize the image-specific neural network with the parameter of the last linear transformation layer sampled from Gaussian distribution $N(0, 0.01^2)$. The query-specific word vector lookup table of size 128 × 81386 is initialized by randomly sampling from Gaussian distribution $N(0, 0.01^2)$. Similar to [26], the core visual model is denoted as Image - C96 - P - N - C256 - P - N - C384 - C384 - C256 - P - F4096 - F4096 - F128. To train the neural networks, we first fixed the parameters of the core visual model and updated the other parameters only. After the loss came to be stable, we set a small learning rate to update the core visual model as well while we observed very limited performance improvement by updating the core visual model.

### C. Qualitative Analysis on the Learned Representations

The learned representations encode both the explicit connections and the implicit connections among the vertices in the click graph. In this subsection, we give some qualitative examples to show the semantic similarities of the learned word representations and how the learned representations help to improve the search experience by correcting misspelling words and re-ranking images.

*1) Word Representations:* We first investigate the learned word representations. Traditionally, a word is represented by a high-dimensional one-hot-spot vector. In this work, each word is learned to be associated with a low-dimensional continuous vector in the word look-up table. One simple way to investigate the learned representation is to find the closest words for a specific word. As shown in Table III and Table IV, we compare the learned word representation of the proposed MRW-NN and the ranking-based comparative method CMRNN [25] by finding the closest words of *france* (name of country), *shanghai* (name of city), *huawei* (name of brand), *ps4* (PlayStation4, a game device) and *nba* (National Basketball Association). Reminding that MRW-NN considers the implicit connections among the queries while CMRNN only considers the explicit query-to-image connections, MRW-NN has more potential to capture the relationship among words appearing across different queries. For example, the closest words to *shanghai* learned by MRW-NN are the other cities like *Seattle* and *New York City*, while those that learned by CMRNN are related concepts to the city like *fleet* and *arena*. The similar phenomenon is also observed for the word *ps4*. To give more insights, the closest words that learned by the word2vec model [27] is shown in Table V, which is trained on part of

TABLE III

MRW-NN: THE CLOSEST WORDS AND THEIR SIMILARITIES FOR GIVEN WORDS (MEASURED IN COSINE DISTANCE)

| france | | shanghai | | huawei | | ps4 | | nba | |
|---|---|---|---|---|---|---|---|---|---|
| germany | 0.6391 | seattle | 0.5422 | smartphone | 0.6451 | playstation | 0.7656 | basketball | 0.6393 |
| spain | 0.5658 | singapore | 0.5331 | cellphone | 0.6184 | ps | 0.6564 | drose | 0.5988 |
| ukrain | 0.5575 | seatle | 0.5215 | phone | 0.6065 | playstation4 | 0.6468 | derrick | 0.5963 |
| city | 0.5571 | skyline | 0.5204 | iphone | 0.5989 | plastation | 0.6443 | lebronjames | 0.5800 |
| espana | 0.5545 | hongkong | 0.5197 | samsung | 0.5893 | xbox720 | 0.6412 | baketball | 0.5589 |
| parisfrance | 0.5523 | nyc | 0.5112 | lg | 0.5834 | ps3 | 0.6039 | dunk | 0.5120 |
| prague | 0.5498 | chicago | 0.5019 | nokia | 0.5802 | xbox | 0.5960 | miamiheat | 0.5051 |

TABLE IV

CMRNN: THE CLOSEST WORDS AND THEIR SIMILARITIES FOR GIVEN WORDS (MEASURED IN COSINE DISTANCE)

| france | | shanghai | | huawei | | ps4 | | nba | |
|---|---|---|---|---|---|---|---|---|---|
| brazil | 0.6993 | fleet | 0.5731 | htc | 0.7617 | sticker | 0.6380 | nfl | 0.8014 |
| germany | 0.6822 | entrance | 0.5615 | ipod | 0.7169 | 4s | 0.6329 | character | 0.7208 |
| peru | 0.6798 | elsavador | 0.5288 | lg | 0.6999 | import | 0.6067 | ncaa | 0.7167 |
| mexico | 0.6748 | obstacle | 0.5277 | ipad | 0.6954 | gel | 0.5994 | math | 0.7117 |
| downtown | 0.6747 | arena | 0.5272 | nook | 0.6795 | amd | 0.5984 | kindergarten | 0.7094 |
| va | 0.6736 | haiti | 0.5235 | app | 0.6795 | suitcase | 0.5941 | pain | 0.7085 |
| nevada | 0.6707 | vary | 0.5164 | kindle | 0.6794 | external | 0.5939 | crip | 0.7065 |

TABLE V

WORD2VEC: THE CLOSEST WORDS AND THEIR SIMILARITIES FOR GIVEN WORDS (MEASURED IN COSINE DISTANCE)

| france | | shanghai | | huawei | | ps4 | | nba | |
|---|---|---|---|---|---|---|---|---|---|
| french | 0.7001 | shenzhen | 0.7923 | zte | 0.7028 | ps3 | 0.6718 | knick | 0.6480 |
| belgium | 0.6933 | guangzhou | 0.7868 | ericsson | 0.5870 | 3ds | 0.6218 | wnba | 0.6428 |
| paris | 0.6335 | beijing | 0.7192 | motorola | 0.5852 | ssbb | 0.6205 | piston | 0.6159 |
| germany | 0.6271 | chongqing | 0.7064 | lte | 0.5612 | bf3 | 0.6116 | celtics | 0.6093 |
| italy | 0.6135 | hangzhou | 0.6977 | cisco | 0.5536 | ps2 | 0.6024 | nfl | 0.6019 |
| spain | 0.6064 | chengdu | 0.6941 | 3g | 0.5370 | mw3 | 0.5906 | lebron | 0.5977 |
| nant | 0.6042 | guangdong | 0.6779 | telecom | 0.5332 | halo3 | 0.5823 | artest | 0.5976 |

TABLE VI

MISSPELLING WORDS FOR GIVEN WORDS

| Misspelling words | Correct word |
|---|---|
| wrld worl wourld wrold wolrd | world |
| aple appple aplle appele applwe aaple | apple |
| anmal animla amimal aniaml anime anmiale | animal |
| gerl gril gile gilr grls | girl |
| goole googele gogle gooogle googal goolge | google |
| gutar gitar guiter gutair giutar gituar quitar | guitar |

TABLE VII

QUERY-TO-QUERY SUGGESTIONS

<table>
<tr><th>Input query</th><th>Query suggestions</th></tr>
<tr><td>barack obama</td><td>2012bracket barack obama | barack chilhood obama</td></tr>
<tr><td>bathroom shelf</td><td>bathroom shelf small | bathroom shelf wall</td></tr>
<tr><td>hairstyle short</td><td>hairstyle olderwomen short | hairstyle kinkey short</td></tr>
<tr><td>cuban flag</td><td>cuban flag large | cuban flag small</td></tr>
<tr><td>cupcake pail</td><td>collect cupcake | cost cupcake | cupcake price</td></tr>
</table>

Google News dataset with about 100 billion words.[3] We can see that the closest words learned by MRW-NN and word2vec are very different. To take *ps4* as an example, in an image retrieval system, users that search *ps4* may be more likely to search images for *playstation* and *xbox* (closest words learned by MRW-NN, both are game devices), but much less likely for *bf3* and *halo3* (closest words learned by word2vec, both are game names).

Though we do not observe the additive compositionality for performing analogical reasoning (e.g., vector(*king*) − vector(*man*)+vector(*woman*) = vector(*queen*)) on the learned word representation, the word representation learned from the click graph allows other interesting applications. Contributed to mining the implicit connections among the queries, the learned word representation can be another cue to correct the input misspelling words, besides the traditional cues like Edit Distance. Table VI shows some misspelling words and their corresponding correct words, where the correct words are found from the nearest neighbors of the misspelling words in the latent space (by filtering out those words in dictionary). Another application is query-to-query suggestion: given a query, find the other queries that the user might like to submit. Table VII shows some query suggestions by finding the closest queries to the input query in the latent space.

*2) Image Representations:* Though the clicked images are in general relevant to the corresponding query, the noises exist in the Clickture (in a relatively low level). Some images may be clicked by mistake as they attracted users' attention for whatever reason (e.g., very unique or strange images, even though they are not relevant to the current query). One of the interesting property of the learned representation is that the learned representation is more robust to noise than the mere click count based ranking. We present some query examples (*animal best*, *beer* and *dell computer*) in Figure 4. For example, as search engines typically show images indexed by the surrounding text in the same page, the top clicked images to the query *dell computer* involve other Dell related images like

[3] https://code.google.com/p/word2vec/

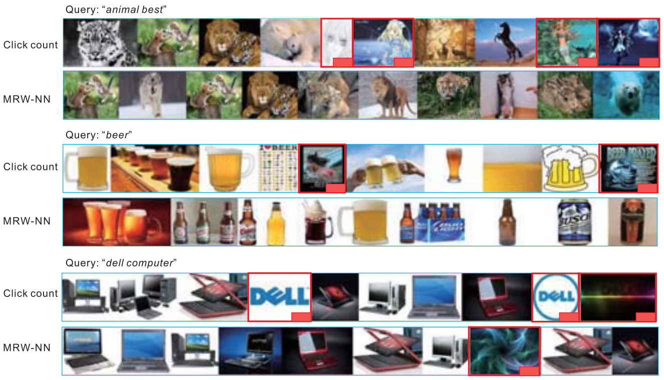


Fig. 4. Ranking comparison on the training data. For each query, top 10 images are returned. The upper ranking list is simply sorted by the click count of the query-image pairs (which may be noisy) and the lower list is sorted by the cosine similarity of the learned representation between the query and the images. The images in red indicate that the images may be irrelevant to the corresponding query.

the logo of Dell as well. In the proposed method MRW-NN, the latent representation is learned in a way considering the content of the image. Thus in the latent space, an image that consists of computers is closer to the query *dell computer* than those that consists of logos only, accordingly having a higher rank position.

TABLE VIII

COMPARATIVE METHODS: THE COMPLEXITY OF CALCULATING RANKING SCORE

| Method | Complexity |
|---|---|
| BoWDNN-R | $O(n)$, $n$ is the dimension of the *query* space |
| CCL | $O(d)$, $d$ is the dimension of the latent space |
| PSI | $O(d)$, $d$ is the dimension of the latent space |
| PAMIR | $O(m)$, $m$ is the dimension of the *image* space |
| CMRNN | $O(d)$, $d$ is the dimension of the latent space |
| MRW-NN | $O(d)$, $d$ is the dimension of the latent space |

TABLE IX

THE DCG@25 (%) OF THE COMPARATIVE METHODS

| BoWDNN-R | CCL | PAMIR | PSI | CMRNN | MRW-NN |
|---|---|---|---|---|---|
| 50.89 | 50.59 | 50.17 | 49.91 | 50.71 | **51.04** |

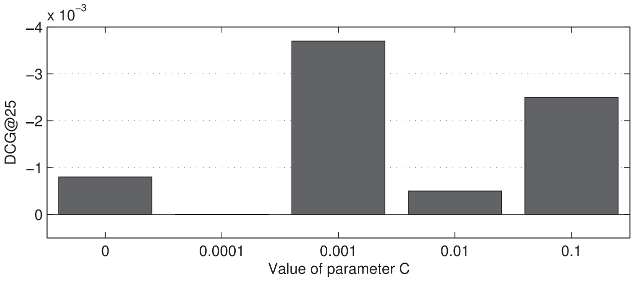


Fig. 5. Sensitivity of the parameter $C$ with best value $C = 0.0001$.

## D. Cross-Modal Retrieval Performance Comparison

One important characteristics of MRW-NN is that it can deal with those queries and images that do not appear in the training click graph, which allows the proposed method to perform cross-media retrieval on the unseen data by mapping the unseen data into the latent space. We compare MRW-NN with the other state-of-the-art methods for cross-modal retrieval performance evaluation on the Clickture dev set. More specifically, the text-query-image in cross-media retrial is compared [34], [35].

*1) Comparative Methods:* The comparative methods are elaborately chosen for the fair comparison, which includes paired-based methods (BoWDNN-R and CCL) and ranking-based methods (PAMIR, PSI and CMRNN). The comparative methods are listed as following:

- BoWDNN-R (Bag-of-Words similarity based ranking method [36]). The idea of this method is to measure the image-query relevance based on the cosine similarity between the query BoW representation and the image BoW representation learned by a proposed BoWDNN which maps the image to the high-dimensional (50,000D in their settings) text space.
- CCL (Click-through-based Cross-view Learning [7]). The training of CCL is performed simultaneously by minimizing the distance between query and image mappings in the latent subspace weighted by theirs clicks, and preserving the structure relationships among the training examples in the original feature space.
- PAMIR (Passive-Aggressive Model for Image Retrieval [8]). Passive-Aggressive model measures the match between a query and an image by first projecting the query into the image space with optimization for text-query-image pairwise ranking loss. The click count is used to measure the ranking priority of an image given a query.
- PSI (Polynomial Semantic Indexing [37]). A little different from PAMIR, a polynomial ranking model with 2-degree projection of the image and the query into a latent subspace with the aim of optimizing for the margin pairwise ranking loss.
- CMRNN (Cross-Model Ranking Neural Network [25]). Similar to the proposed MRW-NN, CMRNN takes two modality-specific neural networks for mapping the queries and the images into a common subspace, while it optimizes the listwise ranking loss of the cross-modal ranking examples.

All the comparative methods first compute the similarities between the query and the images, and then rank the images by their similarities in descending order. For BoWDNN-R, CCL, PAMIR and PSI, we compare the reported results in [36] and [38] for the MSR-Bing Image Retrieval Challenge. For CMRNN and the proposed MRW-NN, we train the models respectively on an Nvidia GTX Titan GPU. It takes about 10 hours to train the MRW-NN model. In the production environment for cross-modal retrieval, the test time for each query is more crucial. Given the comparative methods, the runtime complexity of computing the relevance similarity is given in Table VIII (the time of sorting are all the same).

*2) Evaluation:* For the evaluation of cross-modal ranking, we adapt Discounted Cumulative Gain (DCG) [39] which takes into account the measure of multi-level relevance degree as the performance metrics. Given an image ranked list,

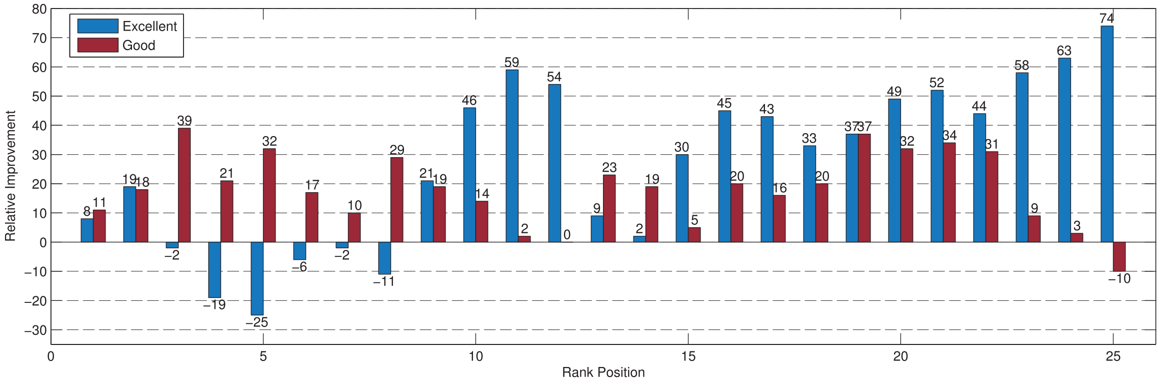


Fig. 6. The relative improvement of the proposed MRW-NN to CMRNN over the dev set. For example, *Relative Improvement@10* means that among all the top 10 returned results (given different queries), MRW-NN ranks 46 excellent and 14 good images more than CMRNN (and thus 60 bad images less than CMRNN).

the DCG score at the position of $p$ is defined by:

$$DCG@p = Z_p \sum_{j=1}^{p} \frac{2^{r_j} - 1}{\log(1+j)}$$

where $p = 25$, $Z_p = 0.01757$ and $r_j = \{Excellent = 3; Good = 2; Bad = 0\}$ is the manually judged relevance for each image with respect to the query. At last, the average of the DCGs on all the queries is the final evaluation result. The higher the DCG score, the better retrieval performance.

The DCG performance of cross-modal retrieval over 1,000 queries in the Clickture dev set is reported in Table IX, showing that MRW-NN outperforms all the comparative methods including paired-based methods and ranking-based methods. The sensitivity of parameter $C$ is shown in Figure 5. It is also worth noting that the second best performing method BoWDNN-R has a different structure of neural network and a much higher test time complexity (see Table VIII). The relative improvement of the proposed MRW-NN to CMRNN on the top 25 ranking positions is also reported in Figure 6. Each $(x, y)$ in the figure means that MRW-NN ranks $y$ excellent or good images than CMRNN in the top $x$ returned results ($y > 0$ means that MRW-NN has better performance). It is observed that MRW-NN ranks more relevant (excellent and good) images than CMRNN in every position of ranking list, showing that the improvement is consistent.

We also note that the performance of all the comparative methods are far to be perfect. One type of failure cases comes from the unseen words out of the training vocabulary, where 41 queries out of all the 1,000 queries cannot be recognized. We demonstrate some examples of the cross-modal retrieval results on the dev set in Figure 7, which contains both good and bad cases that MRW-NN performs. Take as an example the query *sunflower live show 2010*. MRW-NN ranks the plant sunflowers at the top while *sunflower live show* here is a proper noun, which suggests MRW-NN should find out a way to better understand the users' input queries.

The proposed model could be also effective in the application of image-to-text annotation. We give three examples of annotation in Figure 8. Given an image, the words are ranked according to their cosine similarities to the image in the latent space. The top 10 ranked words are listed for each image from the dev set, while including some misspelling words as well. It is also observed that the annotation results are diverse (for example, the third *sphinx* image).

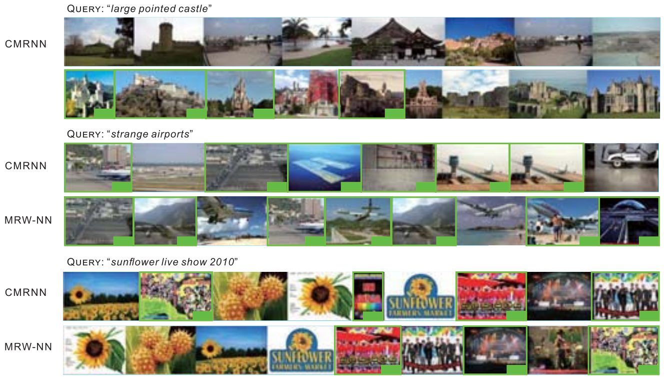

QUERY: "large pointed castle"
CMRNN
QUERY: "strange airports"
CMRNN
MRW-NN
QUERY: "sunflower live show 2010"
CMRNN
MRW-NN
SUNFLOWER FARMERS MARKET

Fig. 7. Cross-modal retrieval performance comparison between CMRNN and MRW-NN. The images with green indicate the image is relevant (Excellent or Good) to the corresponding query. Both good and bad cases are reported.

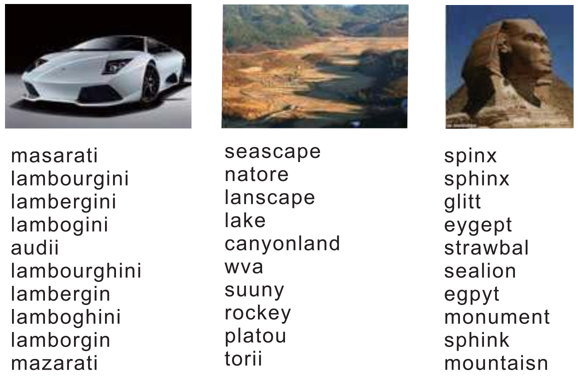


Fig. 8. Annotation of exemplar images in the Clickture dev set. Top 10 nearest words to the images in the latent space are listed including misspelling words as well.

## V. Conclusions

In this work, we have presented a new approach to learning latent representation of the multimodal data from a click graph. By the minimization of the random walk error and the regularization penalty from the output of the modal-specific neural networks, the learned model has the ability not only to represent the explicit connections and the implicit connections of the vertices in the click graph with low-dimensional continuous vectors, but also to map the unseen queries and images to the latent subspace to support cross-modal retrieval. We have demonstrated the effectiveness of the learned representation by the proposed method MRW-NN and shown its superior to the comparative methods on cross-modal retrieval on a large-scale click log dataset.

## ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research Program of China under Grant 2015CB352300, in part by the China Knowledge Centre for Engineering Sciences and Technology, and in part by the Fundamental Research Funds for the Central Universities and the Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis.

## REFERENCES

[1] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.

[2] X.-S. Hua *et al.*, "Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 243–252.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

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

[5] G. Smith, C. Brien, H. Ashman, "Evaluating implicit judgments from image searchclickthrough data," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2451–2462, 2012.

[6] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[7] 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. Retr.*, 2014, pp. 717–726.

[8] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.

[9] F. Wu *et al.*, "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1497–1509, May 2015.

[10] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.

[11] M. Wang, W. Li, D. Liu, B. Ni, J. Shen, and S. Yan, "Facilitating image search with a scalable and compact semantic mapping," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1561–1574, Aug. 2015.

[12] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

[13] 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.

[14] 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.

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

[16] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.

[17] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[18] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang, "A low rank structural large margin method for cross-modal ranking," in *Proc. 36th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2013, pp. 433–442.

[19] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2013, pp. 877–886.

[20] N. Craswell and M. Szummer, "Random walks on the click graph," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 239–246.

[21] Y. Jing and S. Baluja, "VisualRank: Applying PageRank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[22] C.-C. Wu, K.-Y. Chu, Y.-H. Kuo, Y.-Y. Chen, W.-Y. Lee, and W. H. Hsu, "Search-based relevance association with auxiliary contextual cues," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 393–396.

[23] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 2231–2239.

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

[25] X. Lu *et al.*, "Learning multimodal neural network with ranking examples," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 985–988.

[26] 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.

[27] 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.

[28] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA, USA: Addison-Wesley, 1999.

[29] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: http://arxiv.org/abs/1409.1556

[30] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[31] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1081–1088.

[32] C. M. Bishop *et al.*, *Pattern Recognition and Machine Learning*, vol. 4. New York, NY, USA: Springer, 2006, p. 4.

[33] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[34] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multimodal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[35] 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.

[36] Y. Bai *et al.*, "Bag-of-words based deep neural network for image retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 229–232.

[37] B. Bai *et al.*, "Polynomial semantic indexing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 64–72.

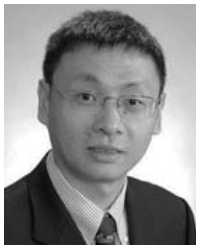
[38] Y. Pan, T. Yao, X. Tian, H. Li, and C.-W. Ngo, "Click-through-based subspace learning for image search," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 233–236.

[39] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

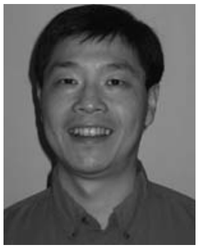
**Fei Wu** received the B.Sc. degree from Lanzhou University, Lanzhou, China, in 1996, the M.Sc. degree from the University of Macau, Macau, China, in 1999, and the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2002, all in computer science. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include multimedia retrieval, sparse representation, and machine learning.

**Xinyan Lu** received the B.S. degree from the College of Computer Science, Zhejiang University, and the Ph.D. degree from the School of Mathematics and Computational Science, Sun Yat-sen University. He is currently with Tencent Inc. His research interests are cross-media retrieval and learning to rank.

**Jun Song** received the B.E. degree from Tianjin University, Tianjin, China, in 2013. He is currently pursuing the Ph.D. degree in computer science with the Digital Media Computing and Design Laboratory, Zhejiang University. His research interests include machine learning, cross-media information retrieval, and understanding.



**Shuicheng Yan** is currently the Chief Scientist of 360, the Director of 360 AI Institute, and the Dean's Chair Associate Professor with the National University of Singapore (NUS). His research areas include machine learning, computer vision, and multimedia, and he has authored or co-authored hundreds of technical papers over a wide range of research topics, with Google Scholar citation over 20 000 times and h-index of 61. He is ISI Highly Cited Researcher in 2014 and 2015, and an IAPR Fellow in 2014. He received the best paper awards from ACM MM'13 (both best paper and best student paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, ICME'10, and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prize of ILSVRC'14 detection task without extra data, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honourable mention prize of the detection task in PASCAL VOC'10, the 2010 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Associate Editor (BAE) Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 NUS Young Researcher Award.



**Zhongfei (Mark) Zhang** received the B.S. (*cum laude*) degree in electronics engineering and the M.S. degree in information science from Zhejiang University, and the Ph.D. degree in computer science from the University of Massachusetts at Amherst. He is currently a Full Professor of Computer Science with the State University of New York at Binghamton. He directs the Multimedia Research Laboratory at Binghamton.

**Yong Rui** (F'10) received the B.S. degree from Southeast University, the M.S. degree from Tsinghua University, and the Ph.D. degree from the University of Illinois at Urbana–Champaign. He is currently the Deputy Managing Director with Microsoft Research Asia (MSRA), leading research groups in multimedia search and mining, and big data analysis, and engineering groups in multimedia processing, data mining, and software/hardware systems. He has authored 16 books and book chapters, and more than 100+ refereed journal and conference papers. His publications are among the most cited 15 000+ citations and his h-index of 52. He holds 60 issued U.S. and international patents. He is a fellow of IAPR and SPIE, a Distinguished Scientist of ACM, and a Distinguished Lecturer of both ACM and IEEE. He is an Executive Member of ACM SIGMM, and the Founding Chair of its China Chapter. He is recognized as a leading expert in his research areas. He is the Editor-in-Chief of the *IEEE Multimedia Magazine*, an Associate Editor of the *ACM Transactions on Multimedia Computing, Communication and Applications*, a Founding Editor of the *International Journal of Multimedia Information Retrieval*, and a Founding Associate Editor of the IEEE ACCESS. He was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2004-2008), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGIES (2006-2010), the ACM/Springer *Multimedia Systems Journal* (2004-2006), and the *International Journal of Multimedia Tools and Applications* (2004-2006). He also serves on the Advisory Board of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.

**Yueting Zhuang** received the B.S., M.S., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1986, 1989, and 1998, respectively. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University. His research interests include artificial intelligence, multimedia retrieval, digital library, and video-based animation.