Lu B, Wang GR, Yuan Y. A novel approach towards large scale cross-media retrieval. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(6): 1140–1149 Nov. 2012. DOI 10.1007/s11390-012-1292-2

A Novel Approach Towards Large Scale Cross-Media Retrieval

Bo Lu (逯　波), Guo-Ren Wang (王国仁), *Member*, *CCF*, *ACM*, *IEEE*, and Ye Yuan (袁　野)

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

E-mail: mrcooler1982@gmail.com; {wanggr, yuanye}@ise.neu.edu.cn

Received September 5, 2012; revised October 5, 2012.

Abstract With the rapid development of Internet and multimedia technology, cross-media retrieval is concerned to retrieve all the related media objects with multi-modality by submitting a query media object. Unfortunately, the complexity and the heterogeneity of multi-modality have posed the following two major challenges for cross-media retrieval: 1) how to construct a unified and compact model for media objects with multi-modality, 2) how to improve the performance of retrieval for large scale cross-media database. In this paper, we propose a novel method which is dedicate to solving these issues to achieve effective and accurate cross-media retrieval. Firstly, a multi-modality semantic relationship graph (MSRG) is constructed using the semantic correlation amongst the media objects with multi-modality. Secondly, all the media objects in MSRG are mapped onto an isomorphic semantic space. Further, an efficient indexing MK-tree based on heterogeneous data distribution is proposed to manage the media objects within the semantic space and improve the performance of cross-media retrieval. Extensive experiments on real large scale cross-media datasets indicate that our proposal dramatically improves the accuracy and efficiency of cross-media retrieval, outperforming the existing methods significantly.

Keywords cross-media retrieval, multi-modality, semantic correlation, indexing structure

1 Introduction

Cross-media retrieval is coming as a new trend along with the rapid development of Internet and multimedia technology. Compared with the traditional content-based multimedia retrieval with single modality, cross-media retrieval is more in accordance with the user's experience. Because the modality of query example and returned results are often different, it is propitious to satisfy the various requirements of users[1-3]. For instance, as shown in Fig.1, if users submit a query example (an image of *eagle*), they may not only want to obtain some similar images about *eagle*, but also want to obtain the description of text or video clips about *eagle*.

Traditional content-based multimedia retrieval methods generally extract low-level features of media objects, which can be utilized to measure the similarity amongst media objects with single modality[4].

However, the similarity measure of media objects with multi-modality by only exploring low-level features is a very difficult problem. Because different kinds of low-level features of media objects cannot be computed in an isomorphic feature space. For example, it is difficult to measure the similarity between an image with visual features (e.g., color, shape and texture) and an audio with auditory features (e.g., timbre, pitch and amplitude) in an isomorphic feature space. In other words, the complexity and the heterogeneity of the multi-modality are the fundamental challenges of cross-media retrieval. In order to solve these issues, we use semantic concepts[5-6] as the high-level semantic features to measure the semantic correlation amongst media objects with multi-modality. The media objects

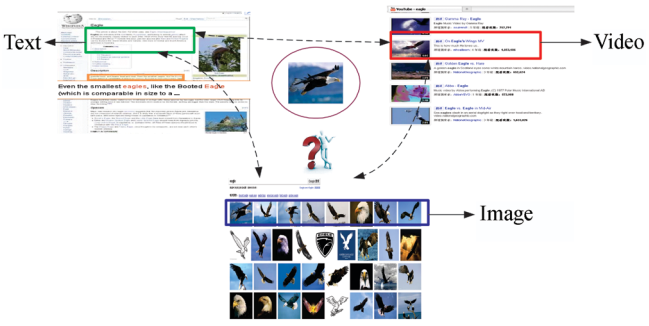

Fig.1. Example of user query of cross-media retrieval.

Regular Paper

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61025007, 60933001, 61100024, the National Basic Research 973 Program of China under Grant No. 2011CB302200-G, the National High Technology Research and Development 863 Program of China under Grant No. 2012AA011004, and the Fundamental Research Funds for the Central Universities of China under Grant No. N110404011.

*The preliminary version of the paper was published in the Proceedings of the 2012 Computational Visual media Conference.

of different modalities, such as text, image and video, generally possess some information of latent semantic correlation among each other. In addition, semantic concept is certainly closer to the natural representation of human which is in favor of unifying the features of media objects with multi-modality.

For the cross-media retrieval, the other major challenge is to manage and retrieve various types of media objects stored in large-scale multimedia database. When faced with the large-scale datasets, most of existing retrieval methods ignore the retrieval cost of cross-media retrieval, which usually leads to degrade the performance of cross-media retrieval. Therefore, it is important to effectively retrieve the results associated with the user request from large-scale multimedia database. Specifically, there is a urgent need of indexing techniques which can manage the cross-media database and support the execution of similarity queries.

Accordingly, an effective cross-media retrieval method should address following two problems: 1) how to construct a unified and compact model by exploring the semantic correlation of media objects with multi-modality, 2) how to improve the performance of cross-media retrieval for large-scale multimedia database. In this paper, we propose a new method which is dedicated to solving these difficulties to achieve effective and accurate cross-media retrieval. Firstly, a multi-modality semantic relationship graph (MSRG) is constructed by using the semantic correlation information of media objects with multi-modality. Specifically, semantic correlation among media objects with multi-modality is learned by canonical correlation analysis[7]. Further, all the media objects are mapped onto an isomorphic semantic space. To manage and retrieve all the media objects, an efficient indexing MK-tree based on heterogeneous data distribution is proposed to manage media objects within the semantic space and improve the performance of cross-media retrieval with the large scale cross-media database. Finally, we execute the *range* query and the *k-nearest neighbor* (kNN) query to examine the performance of cross-multimedia retrieval. An overview of our approach is shown in Fig.2.

In summary, the main contributions of the paper are:

1) In order to effectively address the problem of the heterogeneity of multiple media objects, we propose MSRG which covers multiple media objects by exploring the semantic concepts as high-level features of media objects.

2) We propose a unified indexing structure called MK-tree to efficiently improve the large scale cross-media retrieval. Specifically, we consider the characteristics of different media objects to implement the steps of data partitions of MK-tree based on the key dimension and data distribution of multiple media objects.

3) We consider extensive experiments over real large scale cross-media datasets to evaluate our proposed method. Moreover we present theoretical analysis and comparison on the search and storage cost of the proposed indexing scheme.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the data representation and the method of extracting high-level semantic features. Section 4 introduces the construction of MSRG in details. In Section 5, we discuss cross-media retrieval based on MK-tree indexing. Section 6 reports the experimental evaluation of our method. We finally conclude the paper in Section 7.

2 Related Work

In recent years, the academic community has proposed concept-based multimedia retrieval by pooling a set of pre-trained semantic concept detectors which can be regarded as intermediate descriptors to bridge the semantic gap[5-6,8-9]. The semantic concepts generally cover a wide range of topics which include objects, scenes, people, events, etc. Some multimedia research communities have put tremendous efforts into manual annotation and released a large number of ground truth annotations, such as TRECVID[10], imageCLEF[11] and

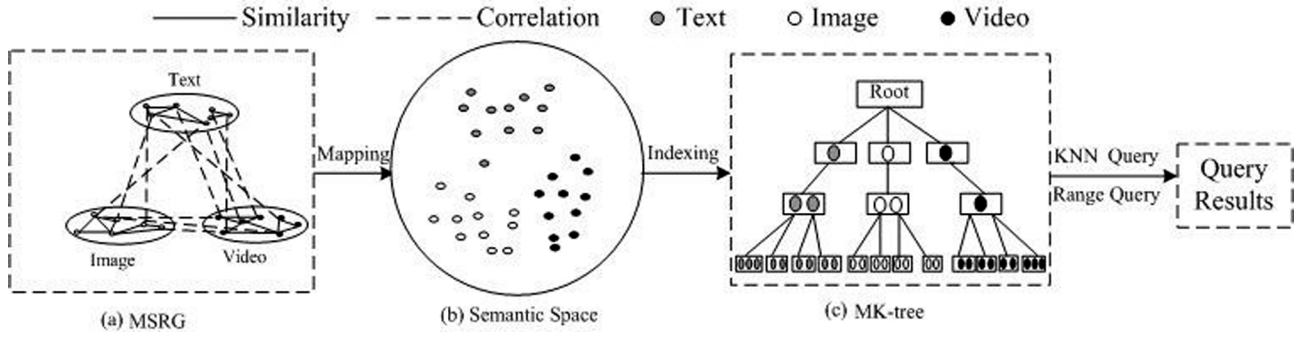

Fig.2. Framework of our approach. (a) Multi-modality semantic relationship graph (MSRG). (b) Semantic space. Note that heterogenous media objects have different data distributions. (c) MK-tree based on heterogeneous data distribution.

LSCOM[12]. These concept annotations involve image or video data complement with annotations, closed-caption information or speech recognition transcripts.

An intrinsic problem of cross-media retrieval is to investigate the semantic correlation amongst the heterogeneous multimedia data. There are many researches have focused on this issue[13-16].

Yang *et al.*[2] proposed a two-level manifold learning method for cross-media retrieval. Firstly three independent graphs are constructed for image objects, audio objects and text objects respectively. According to the graphs, media objects are projected onto three spaces which are then combined to obtain the final data representation in multimedia document semantic space. However, the semantic correlations among heterogeneous multimedia objects are not used when construct the independent spaces for image, audio and text objects. In addition, the two-level manifold learning method is so complex that it must simultaneously adjust more than 10 parameters, which making it less applicable in the real applications.

In [3], the researchers proposed a ranking algorithm for the cross-media retrieval, namely ranking with local regression and global alignment, which learns a robust Laplacian matrix for data ranking. For each data point, they employed a local linear regression model to predict the ranking value of its top-k nearest neighboring points. Furthermore, they proposed a global objective function to assign an optimal ranking value to each point.

The multimedia indexing is a kind of high-dimensional indexing problem. The academic community has made many efforts on solving the high-dimensional indexing problem. Existing techniques can be divided into three main categories. The first category is based on data space partition, hierarchical tree index structures such as the R-tree[17] and its variants. Their performance deteriorates rapidly as the dimensionality increases. The second category is to represent original feature vectors using smaller approximate representations, such as VA-file[18]. Although it can accelerate the sequential scan by data compression, it suffers from the higher computational cost for decoding the bit string. The last category uses a distance-based indexing method, such as iDistance[19]. iDistance is a distance-based scheme, in which high-dimensional points are mapped onto a single-dimensional distance values by computing their distance from the centroid respectively.

Ciaccia *et al.*[20] proposed a paged metric called M-tree, which is a paged and balanced dynamical index based on a bottom-up construction with node promotion and split mechanism. However, the overlap of subspaces in M-tree is usually considerable large, which leads to the decrease of the performance.

In this paper, we propose MK-tree to index heterogeneous media objects, which truly supports cross-media retrieval. Analogous to M-tree, MK-tree is a dynamical paged and balanced tree. Besides inheriting the merits of M-tree, MK-tree improves the data partitioning based on different data distributions and extends M-tree with key dimension to obtain higher performance of retrieval.

3 Preliminaries

3.1 Data Representation

We consider the problem of cross-media retrieval from a multimedia database which contains components of text, image and video. Some symbols to be used in the rest of the paper are shown in Table 1. Each media object is represented as an x-dimensional semantic feature vector, such as the text object t_i is denoted as $\boldsymbol{f}_{t_i} = (f_1^{t_i}, f_2^{t_i}, \ldots, f_x^{t_i})$. Image and video objects are represented in the same way.

Table 1. Notations

Symbol	Meaning
T	Set of text objects, $T = \{t_1, t_2, \ldots, t_w\}$
I	Set of image objects, $I = \{p_1, p_2, \ldots, p_m\}$
V	Set of video objects, $V = \{s_1, s_2, \ldots, s_n\}$
D	Multimedia dataset, $D = \{T, I, V\}$
f^{T}	High-level semantic features set of text, $f^{\mathrm{T}} = \{\boldsymbol{f}_{t_1}, \boldsymbol{f}_{t_2}, \ldots, \boldsymbol{f}_{t_w}\}$
f^I	High-level semantic features set of image, $f^I = \{\boldsymbol{f}_{p_1}, \boldsymbol{f}_{p_2}, \ldots, \boldsymbol{f}_{p_m}\}$
f^V	High-level semantic features set of video, $f^V = \{\boldsymbol{f}_{s_1}, \boldsymbol{f}_{s_2}, \ldots, \boldsymbol{f}_{s_n}\}$

3.2 Extraction of High-Level Semantic Features

In order to construct a unified and compact semantic correlation model, in this paper, we extract the high-level semantic concept features of heterogeneous media objects. The method of extraction is introduced as following.

For the text objects, we use a state-of-the-art IE system developed for the automatic content extraction (ACE) program to process text and automatic speech recognition output[21-22]. The pipeline includes name tagging, nominal mention tagging, time expression extraction and normalization, relation extraction and event extraction. Entities include co-referred persons, geo-political entities, locations, organizations, facilities, vehicles and weapons; relations include 18 types (e.g., *a town some 50 miles south of Salzburg indicates* a located relation); events include the 33 distinct event types defined in ACE 2005 (e.g., "*Barry Diller on*

Wednesday quit as chief of Vivendi Universal Entertainment." indicates a personnel-start event). Names are identified and classified using an HMM-based name tagger. Nominals are identified using a maximum entropy based chunker and then semantically classified using statistics from ACE training corpora. Relation extraction and event extraction are also based on maximum entropy models, incorporating diverse lexical, syntactic, semantic and ontological knowledge.

For image and video objects, we employ a semantic concept extraction system, which is developed by IBM for the TREC retrieval evaluation[23]. This system can extract 2 617 semantic concepts defined by TRECVID. It uses Support Vector Machine (SVM) to learn the mapping between low level features extracted from visual modality as well as from transcripts and production related meta-features. It also exploits a correlative multi-label learner, a multi-instance kernel and label propagation through linear neighborhoods to extract all other high-level semantic features. For each classifier, different models are trained on a set of different modalities (e.g., the color moments, wavelet textures, and edge histograms), and the predictions made by these classifiers are combined together with a hierarchical linearly weighted fusion strategy across different modalities and classifiers.

4 Construction of Multi-Modality Semantic Relationship Graph

As we know from above, in order to effectively address the problem of the heterogeneity of media objects with multi-modality, we construct a unified and compact semantic correlation model. In this section, we describe the main steps of the construction of MSRG. The details of each step are then explained sequentially.

The formal definition of MSRG is given below.

Definition 1. *A multi-modality semantic relationship graph (MSRG) is denoted as MSRG* $= (M, E)$*, where* M *is a set of media objects,* E *is a set of edges. Note that, if* $\forall x_i,\ x_j \in D \wedge x_i,\ x_j \in T$ *(*I *or* V*),* E *refers to the similarity between* x_i *and* x_j*. Otherwise, i.e., if* $\forall x_i,\ x_j \in D \wedge \neg(x_i, x_j \in T$ *(*I *or* V*)),* E *refers to the semantic correlation between* x_i *and* x_j*.*

There are three kinds of media objects in MSRG, text, image and video. And the MSRG can be represented by an affinity matrix, which indicates the semantic correlation amongst different media objects.

4.1 Measuring Semantic Correlation

Let $\boldsymbol{R}$ be an n-by-n affinity matrix $(r_{ij})_{n\times n}$ to represent the MSRG, in which r_{ij} represents the semantic correlation among media objects with multi-modality and n represents the total number of media objects. Here, semantic correlation among media objects is learned by canonical correlation analysis[7]. The semantic correlation of the three kinds of media objects is measured with each other. We explore the correlation structure of two sets of variables, one represents a set of independent variables and the other one represents a set of dependent variables. The canonical correlation is optimized so that the linear correlation between two latent variables is maximized.

The mathematical formulation of semantic correlation metric is described as follows. Given two types media objects $\boldsymbol{X}$ and $\boldsymbol{Y}$, denoted as

$$\boldsymbol{X} = (f_1^X, f_2^X, \ldots, f_n^X)^{\mathrm{T}}, \boldsymbol{Y} = (f_1^Y, f_2^Y, \ldots, f_n^Y)^{\mathrm{T}}, \quad (1)$$

we extract the correlated modes between vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ by searching for a set of transformation pairs as $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. For maximizing the canonical correlation amongst latent variables, we give the canonical variants $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ as follows:

$$\boldsymbol{u}_i = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_i, \quad \boldsymbol{v}_i = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\beta}_i, \quad (2)$$

where symbol i denotes the i-th transformation pair. The transformation in (2) obtains the i-th pair of variates $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. Then, the maximum semantic correlation between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ is defined as

$$\rho_i = \max_{\alpha_i \neq 0, \beta_i \neq 0} \frac{\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{C}_{XY} \boldsymbol{\beta}_i}{\sqrt{\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{C}_{XX} \boldsymbol{\alpha}_i} \sqrt{\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{C}_{YY} \boldsymbol{\beta}_i}}, \quad (3)$$

where $\boldsymbol{C}_{XY}$ is the cross-covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{C}_{XX}$ and $\boldsymbol{C}_{YY}$ are auto-covariance matrices. To maximize (3), we obtain the partial derivative of ρ_i with respect to α_i and set the derivative to be zero. We have

$$\boldsymbol{C}_{XY}\boldsymbol{\beta}_i = \frac{\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{C}_{XY} \boldsymbol{\beta}_i}{\boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{C}_{XX} \boldsymbol{\alpha}_i} \boldsymbol{C}_{XX}\boldsymbol{\alpha}_i. \quad (4)$$

As the same way, setting the partial derivative of ρ_i with respect to $\boldsymbol{\beta}_i$ to be zero, we have

$$\boldsymbol{C}_{YX}\boldsymbol{\alpha}_i = \frac{\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{C}_{YX} \boldsymbol{\alpha}_i}{\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{C}_{YY} pmbbeta_i} \boldsymbol{C}_{YY}\boldsymbol{\beta}_i. \quad (5)$$

By combining (4) and (5), we have

$$\begin{aligned} \boldsymbol{C}_{XX}^{-1}\boldsymbol{C}_{XY}\boldsymbol{C}_{YY}^{-1}\boldsymbol{C}_{YX}\boldsymbol{\alpha}_i &= \rho_i^2 \boldsymbol{\alpha}_i, \\ \boldsymbol{C}_{YY}^{-1}\boldsymbol{C}_{YX}\boldsymbol{C}_{XX}^{-1}\boldsymbol{C}_{XY}\boldsymbol{\beta}_i &= \rho_i^2 \boldsymbol{\beta}_i. \end{aligned} \quad (6)$$

By solving the eigenvalue in (6), we obtain the correlation values in ascending order $\{\rho_1, \rho_2, \ldots, \rho_n\}$ and the corresponding transformation sets, $\overline{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n\}$ and $\overline{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n\}$. The corresponding sets of canonical variants can be expressed as $U = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ and $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$.

Note that, the semantic correlation values $\{\rho_1, \rho_2, \ldots, \rho_n\}$ is the pairwise correlation among the high-level semantic features of media objects. As a result, we have $\rho_i = [r_{ij}]$, where r_{ij} means the correlation between two media objects. Then, we obtain the semantic correlation matrix $\boldsymbol{R}$.

4.2 Media Objects Mapping

In order to efficiently manage and retrieve all the media objects, we need to map the media objects onto an isomorphic semantic space. As mentioned above, we derive the semantic correlation of all the media objects from the MSRG. In this subsection, we decompose the semantic correlation matrix $\boldsymbol{R}$ and construct an isomorphic semantic space.

The semantic correlation matrix $\boldsymbol{R}$ is defined as:

$$\boldsymbol{R} = (r_{ij})_{n\times n} = \begin{pmatrix} r_{11} & r_{12} & \cdots \\ r_{21} & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}_{n\times n}. \quad (7)$$

Then, the eigenvalue decomposition of semantic correlation matrix $\boldsymbol{R}$ can be calculated by

$$\boldsymbol{R} = \boldsymbol{O}\boldsymbol{\Lambda}\boldsymbol{O}^{\mathrm{T}} = \boldsymbol{O}\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_v \end{pmatrix}\boldsymbol{O}^{\mathrm{T}}, \quad 0 \leqslant v \leqslant n, \quad (8)$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix. Its elements of diagonal correspond to the eigenvalues of correlation matrix $\boldsymbol{R}$. $\boldsymbol{O}$ is an orthogonal eigenvector matrix corresponding to all the eigenvalues, which is defined by $\boldsymbol{O} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v)^{\mathrm{T}}$. $\boldsymbol{O}^{\mathrm{T}}$ represents the transpose of $\boldsymbol{O}$. $\boldsymbol{q}_i$ is the normalized eigenvector of semantic correlation matrix $\boldsymbol{R}$ corresponding to the eigenvalue λ_i. Here, all the eigenvalues are real and all the eigenvectors are mutually orthogonal because the semantic correlation matrix $\boldsymbol{R}$ is symmetric.

We denote that $(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v)^{\mathrm{T}}$ is an orthogonal basis vector of semantic space. Thus, the isomorphic semantic space can be defined as:

$$\text{SemanticSpace} \rightarrow span(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v)^{\mathrm{T}},$$

which is an orthogonal space generated by linear combinations of $(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_v)^{\mathrm{T}}$.

5 Cross-Media Retrieval Based on MK-Tree

The cross-media datasets is usually large scale, and it is inefficient to retrieve over large scale cross-media datasets only using linear scan. In this section, we propose an efficient indexing MK-tree based on heterogeneous data distribution to index all the media objects which are mapped onto the isomorphic semantic space.

5.1 Data Partition Based on Data Distribution and Key Dimension

MK-tree is a dynamically index structure, which can be used to index large scale multimedia objects datasets. Specifically, we both consider heterogeneous data distribution and key dimension to improve the efficiency of data space partition and reduce the response time of similarity search for various media objects.

As we know, heterogeneous media types have different data distributions. For example, the set of video data may be normal distribution and the set of text data may be uniform distribution. In [24], it was confirmed that the optimal query processing depends on not only the number of objects stored in the database but also the underlying data distribution. Therefore, data distribution is an important factor for influencing query processing.

In an isomorphic semantic space, a key dimension is a dimension that mostly affects similarity computation. Meanwhile, it is crucial to select the key dimension for filtering irrelevant data. In addition, a key dimension can be used to minimize the overlap, and thus avoid a lot of unnecessary path traversals over the index.

The best strategy of the key dimension selection should keep the media objects nearer from each other in the same subspace so that the twin nodes are not overlapped. The most optimal partition method is to segment the data space along the axis with maximal variance, which has been proved to be efficient. This approach ensures the optimization of semantic space partition and reduces the number of paths traversed. Thus, a dimension with the maximal variance is selected to serve as the key dimension.

In this paper, data partition of semantic space is performed as follows: 1) according to heterogenous data distribution, we firstly segment the original semantic space based on the key dimension, 2) the partitioned subspace is split by m-RAD-2 way[20], 3) the subspace is further segmented into twin subspaces. An overview of steps of the data partition of semantic space is shown in Fig.3.

5.2 Filtering Principle

We consider the filtering principle based on the key dimension in the semantic space with respect to the *rang* query. The case is similar to the *k-nearest neighbor* (kNN) query. Let $Q = \{q_1, q_2, \ldots, q_n\}$ be a set of query objects, $X = \{x_1, x_2, \ldots, x_n\}$ be a set of media objects in the semantic space, which can be any kinds of text, image and video, and r be the search radius. Then the distance metric between Q and X can be calculated as

$$d(Q, X) = \sqrt{(q_1 - x_1)^2 + \cdots + (q_n - x_n)^2}. \quad (9)$$

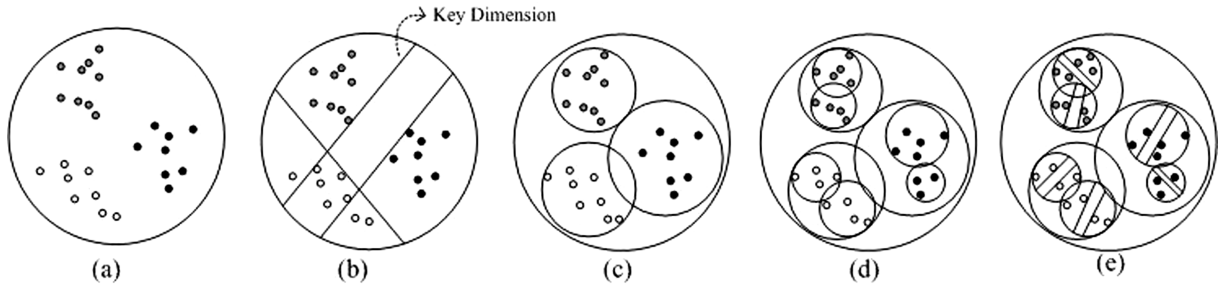

Fig.3. Overview of data partition with regard to different data distributions in semantic space. (a) Original isomorphic semantic space. (b) Partition the original semantic space based on the key dimension. (c) and (d) Partitioned subspaces are split by m-RAD-2 way. (e) Subspaces partitioned into twin subspaces based on the key dimension.

Since $(q_k - x_k)^2 \leqslant (q_1 - x_1)^2 + (q_2 - x_2)^2 + \cdots + (q_n - x_n)^2$ is always true, that is to say, $|q_k - x_k| \leqslant d(Q, X)$ is always true. For any dimension k, if $q_k - x_k \geqslant r$, then $d(Q, X) \geqslant r$. In this case, the media object X can be filtered without the similarity computation. In general, for any similarity measure function d, the filtering principle of the key dimension is valid if and only if $|q_k - x_k| \leqslant d(Q, X)$.

5.3 Data Structure of MK-Tree

In indexing MK-tree, there are two types of mode objects, routing objects and leaf objects. The data structure for leaf entries denoted as $L(x_i, oid(x_i), d(x_i, P(x_i)))$, where x_i is a set of media objects in leaf node, which may be text, image and video. $oid(x_i)$ is an object identifier, and $d(x_i, P(x_i))$ is the distance of x_i from its parent.

The data structure of routing entries denoted as

$$R(o_r, r(o_r), d(o_r, P(o_r)), K_{NO}, leftTwinPtr(T_{lt}(o_r)), M_{l\max}, M_{r\min}, rightTwinPtr(T_{rt}(o_r))),$$

where o_r is a set of media objects in routing object, $r(o_r)$ is the covering radius of o_r, $d(o_r, P(o_r))$ is the distance of o_r from its parent, K_{NO} is the number of key dimension, $leftTwinPtr(T_{lt}(o_r))$ and $rightTwinPtr(T_{rt}(o_r))$ are two pointers to the left twin sub-tree and to the right twin sub-tree respectively, $M_{l\max}$ and $M_{r\min}$ are the maximal value of key dimension in the left twin sub-tree and the minimal value of key dimension in the right twin sub-tree. Fig.4 indicates the MK-tree index structure, corresponding to semantic space shown in Fig.3.

5.4 Query Algorithms

In this subsection, we introduce the details of the algorithms on *range* query and kNN query respectively.

5.4.1 Range Query

We first consider the *range* query. Given a set of query objects Q and a query radius $r(Q)$, the *range* query starts from the root node and recursively traverses all the paths in which the objects match the search condition. The algorithm is described in Fig.5.

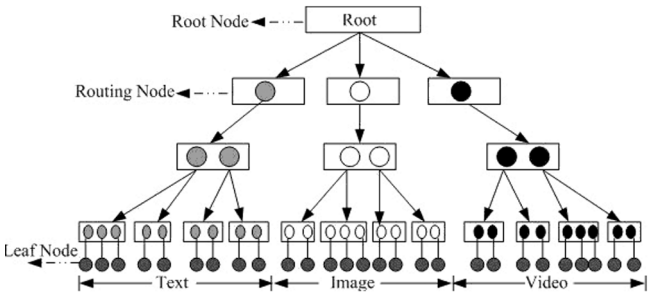

Fig.4. MK-tree index structure.

Input: N: node, Q: a set of query objects, $r(Q)$: query radius
Output: Top-k query results
1 **if** N is not leaf node **then**
2 o_r, o_p in N, **do**:
3 **if** $|d(o_p, Q) - d(o_r, o_p)| \leqslant r(Q) + r(o_r)$ **then**
4 Compute $d(o_r, Q)$
5 **if** $d(o_r, Q) \leqslant r(Q) + r(o_r)$ **then**
6 **if** *key* dim $Val(Q) \leqslant M_{l\max} + r(Q)$ **then**
7 *range-query*($*leftTwinPtr(Tlt(o_r))$, $Q, r(Q)$)
/* *key* dim $Val(Q)$ is the key dimension value of Q*/
8 if *key* dim $Val(Q) \geqslant M_{r\min} - r(Q)$ **then**
9 *range-query*($*rightTwinPtr(Trt(o_r))$, $Q, r(Q)$)
10 **else if** $|d(o_p - Q) - d(x_i - o_p)| \leqslant r(Q)$ **then**
11 Compute $d(x_i, Q)$
12 **else if** $d(x_i, Q) \leqslant r(Q)$ **then**
13 Add $oid(x_i)$ to the result

Fig.5. *Range* query algorithm.

As shown in Fig.5, *range* query begins from the root.

For all subspaces in the current space, those subspaces not containing any query result can be filtered according to the property of triangular inequality. If the subtree is active and cannot be filtered, the distance between the querying object and the routing object is calculated, and further filtering can be done according to the property of triangular inequality. Then, filtering based on the key dimension is performed on the twin nodes. The process is done recursively till the leaf node. In a leaf node, the results can be obtained by calculating directly.

5.4.2 *k*-Nearest Neighbor Query

Given a set of query objects Q and the number of objects to be searched k, the kNN query retrieves the k nearest neighbor of Q. Sharing the method proposed in M-tree, MK-tree uses PR, a priority queue that contains pointers to active sub-trees. We define NN as an array used to store the final search results. The kNN search algorithm is described in Fig.6.

Input: T: root node, Q: a set of query objects, k: integer
Output: Top-k query results
1 $PR \leftarrow [T, -]$
2 **for** $i = 1$ to k **do**
3 $\quad NN[i] \leftarrow [-,]$
4 **while** $PR \neq \varnothing$ **do**
5 $\quad$ *NextNode* = *ChooseNode*(PR)
$\quad$ /* *ChooseNode* is used to select the candidate results from PR */
6 $\quad$ **if** the flag of *NextNode* = True **then**
7 $\quad\quad$ *NextNode* = *TtwinNode*(*NextNode*)
$\quad\quad$ /* *TtwinNode* is to get the children node of *NextNode* */
8 $\quad\quad$ *k-nearest neighbor query* (*NextNode*, Q, k)

Fig.6. *k-nearest neighbor* query algorithm.

In the *k-nearest neighbor* query algorithm, the priority queues operation can be improved based on the key dimension filtering and described as follows. For each node N of PR, if its twin node is active, its flag of N is set to True; otherwise, its flag is set to False. When the two twin nodes of N are all active, only one PR access is needed to do. In this way, many PR accesses are saved. As a result, the cost of query is lowered. First, the root is kept into PR and the maximal distance is kept in NN. Then a priority node is chosen from PR and node search are performed. If the flag of this node is TRUE, the same search process is needed to do for its twin node. The improved *k-nearest neighbor* query algorithm is described in Fig.7.

Input: T: root node, Q: a set of query objects, k: integer
Output: Top-k query results
1 **if** N is not leaf node **then**
2 $\quad o_r$, o_p in N, **do**:
3 $\quad$ **if** $|d(o_p, Q) - d(o_r, o_p)| \leqslant d_k + r(o_r)$ **then**
4 $\quad\quad$ Compute $d(o_r, Q)$
5 $\quad$ **if** $d(o_r, T)_{\min} \leqslant d_k$ **then**
6 $\quad\quad$ Filter by the key dimension
$\quad\quad$ /* $d(o_r, T)_{\min}$ is the minimum distance from o_r to T */
7 $\quad\quad$ Set *Flag* = True or False
8 $\quad\quad$ Push result of NN into PR
9 $\quad$ **if** $d(o_r, T)_{\max} \prec d_k$ **then**
10 $\quad\quad$ $d_k = NN\_Update([-, d(o_r, T)_{\max}])$
$\quad\quad$ /* $d(o_r, T)_{\max}$ is the maximum distance from o_r to T */
11 $\quad$ **if** $d(o_r, T)_{\min} \prec d_k$ **then**
12 $\quad\quad$ Remove candidate nodes from PR
$\quad$ /* *NN_Update* is the updated result in NN */
13 $\quad$ **else**
14 $\quad\quad$ x_i in N, **do**:
15 $\quad$ **if** $|d(o_p, Q) - d(x_i, o_p)| \leqslant d_k$ **then**
16 $\quad\quad$ Compute $d(x_i, Q)$
17 $\quad$ **if** $d(x_i, Q) \leqslant d_k$ **then**
18 $\quad\quad$ $d_k = NN\_Update([oid(x_i), d(x_i, Q)])$
19 $\quad$ **if** $(d(x_i, T)_{\min} \prec d_k)$ **then**
$\quad$ /* $(d(x_i, T)_{\min}$ is the minimum distance from x_i to T */
20 $\quad$ Remove candidate nodes from PR

Fig.7. Improved *k-nearest neighbor* query algorithm.

6 Experimental Evaluation

In this section, we evaluate the performance of our proposed approach on real large-scale multimedia dataset through extensive experiments.

6.1 Experiments Setup

We introduce the setup of the experiments, including the data preparation, experimental environment and parameter setting.

In order to test the effectiveness and efficiency of our proposed method: Efficient Indexing-Based Cross-Media Retrieval (IBCR), we conduct the experiments on a real large-scale multimedia dataset. The experimental data includes 45 000 texts, 75 000 image objects and 15 000 video clips, which are downloaded from the Internet or collected from Microsoft Encarta. Specifically, the text objects are categorized into 29 categories by Wikipedia. These category labels were assigned to text components. Since some of the categories are

very scarce, we consider only the 15 most populated ones. For the video clips, we use the annotations of TRECVID05 dataset from Columbia374[25], which has a lexicon of 374 semantic concepts. The optimization is performed independently on each video for simultaneously labeling all concepts and video shots. In Table 2, we summarize the parameters and their varying ranges in our experiments. The default value of each parameter is highlighted in bold. All the experiments are conducted on Intel Core2 2.8 GHz CPU with 4 GB memory and a 500 GB hard disk.

Table 2. Parameters Setting

Parameter	Varying Range
Search range of text	0.1, 0.15, **0.2**, 0.25
Search range of image	0.3, 0.35, **0.4**, 0.45
Search range of video	0.3, 0.4, **0.5**, 0.6
k of k-NN query	4, 8, **16**, 32, 64
Ground distance	**Euclidean**

6.2 Effectiveness of Cross-Media Retrieval Method

We testify the effectiveness and efficiency of our cross-media retrieval method. As shown in Fig.8, we execute the kNN query, where $k = 20$. When the user submits a query image of *eagle*, 20 candidate video clips are retrieved by our approach.

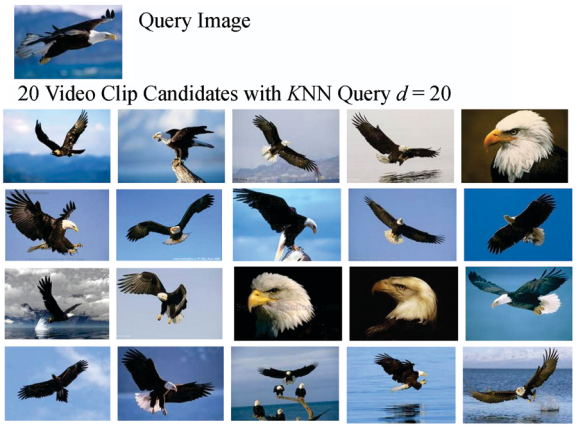

Fig.8. Example of *k-nearest neighbor* query with our approach, where $k = 20$.

Note that, it is difficult to browse large-scale multimedia datasets to generate the manually ground truth for a query. In this paper we obtain a query's ground truth (the top-k best results in kNN search) by comparing the query with database media objects using the similarity measure. Denote the set of ground truth as *rel*, and the set of results returned by a summarization method as *ret*, the recall and precision achieved by the retrieval method are defined as:

$$\text{recall} = \frac{|rel \cap ret|}{|ret|}, \quad \text{precision} = \frac{|rel \cap ret|}{|rel|}. \tag{10}$$

Fig.9 illustrates a recall-precision curve for the performance comparisons between our approach (IBCR) and CIndex[26]. In [26], the researchers employed index structure like B^+-tree by reducing the dimensions of the original space and used low level features of media object to measure the correlation of heterogenous media types. The drawback of CIndex is that it drops out some important correlation information when reduce the dimensions of the data space, and the low level features cannot well represent the semantic correlations of heterogenous media objects. In particular, we compare the average retrieval result (indicates the average precision rate under the average recall rate) of 20 media objects queries randomly chosen from the multimedia dataset. From Fig.9, we observe that the performance of our proposed retrieval method is better than that of CIndex by a large margin.

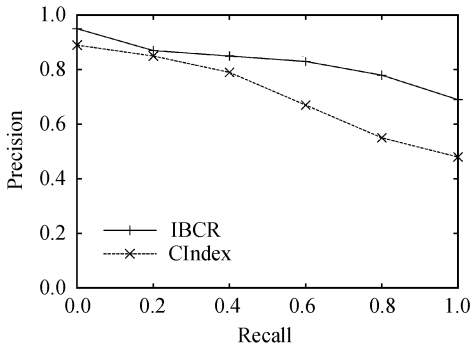

Fig.9. Recall vs precision.

6.3 Sensitivity of Dataset Size

We measure the performance behavior with varied number of media objects, as shown in Fig.10. The comparison of IBCR, CIndex and sequential scan method (Seq. scan) is conducted on range search with the number of media objects varying from 2 000 to 35 000. Fig.10 shows the performance of query processing for all the three media types in terms of CPU cost. It is evident that IBCR outperforms CIndex and the sequential scan method significantly. The CPU cost of IBCR increases slowly as the data size grows. We also notice that the gap between IBCR and the sequential scan is considerable, since the sequential scan is a CPU-intensive operation for large-scale datasets.

6.4 Experiments on Range Query

In this subsection, we discuss the influence of query radius of *range* query on the average precision of retrieve, CPU and I/O cost respectively. As shown in

Fig.11, the performance of our proposed method is superior to CIndex and sequential scan. For processing of *range* query, the CPU and I/O cost of our approach is much lower than the two other methods. Specifically, when the query radius is smaller, more twin nodes in indexing MK-tree can be filtered by the key dimension. The runtime of our method is averagely four times faster than the sequential scan. Simultaneously, we consider the average precision of *range* query, and the accuracy of retrieval by exploring our method does not change more along with the larger query radius.

6.5 Experiments on *k*NN Query

We also conduct experiments to compare the performance of the kNN query of our approach with CIndex and sequential scan. As shown in Fig.12, the performances of CIndex and sequential scan gradually decrease with larger k, but our method is obvious advantageous over them. Note that the average precision of our retrieve approach changes smoothly with different k. It is obvious that our method is appropriate for the cross-media retrieval with large-scale multimedia datasets.

7 Conclusions

In this paper, we presented a novel and efficient method for cross-media retrieval. We firstly constructed a multi-modality semantic relationship graph (MSRG) by exploring the semantic correlation of media objects with multi-modality. Further, all the media objects within MSRG were mapped onto an isomorphic semantic space, which is used to encapsulate the heterogenous media objects. Finally, an efficient indexing MK-tree was proposed to manage media objects and

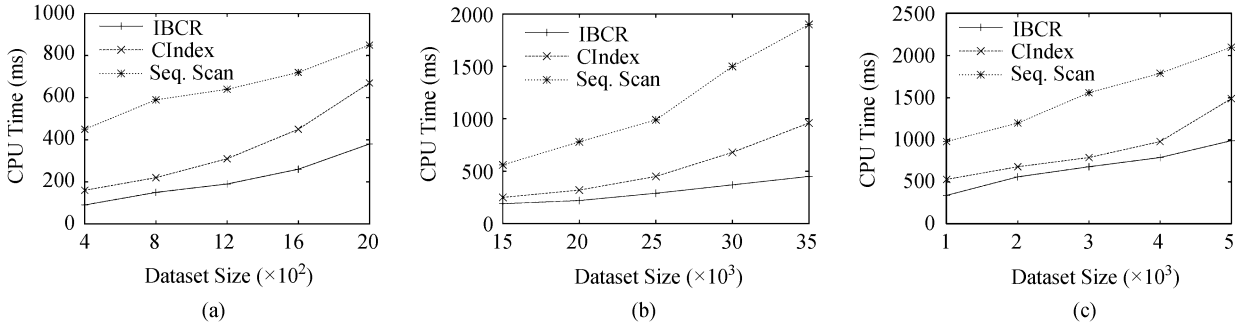

Fig.10. Sensitivity of datasets size. (a) Text. (b) Image. (c) Video.

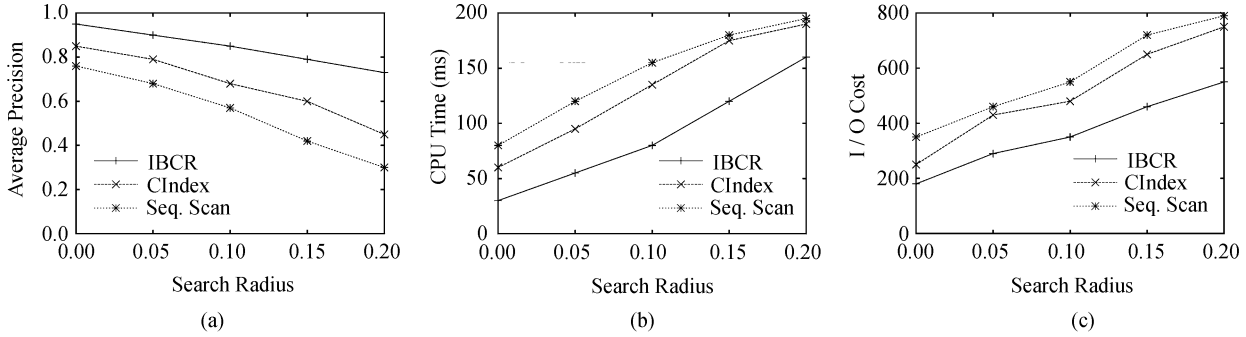

Fig.11. Analysis of the performance of *Range* query. (a) Average precision. (b) CPU cost. (c) I/O cost.

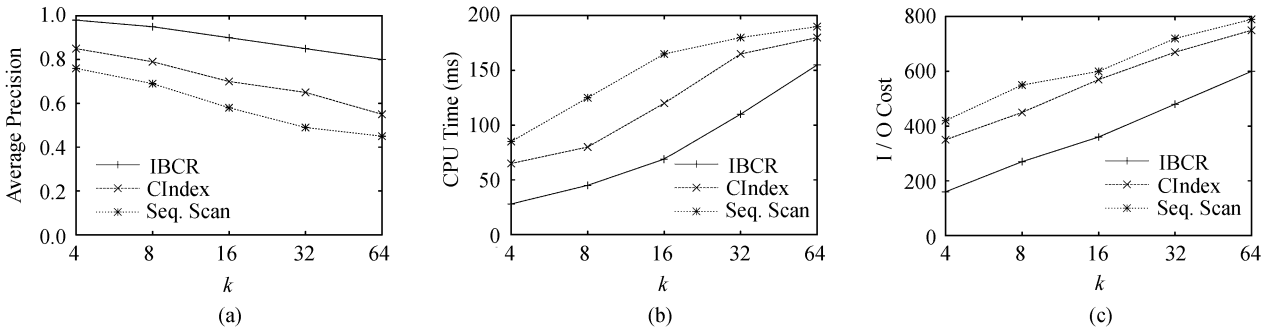

Fig.12. Analysis of the performance of kNN query. (a) Average precision. (b) CPU cost. (c) I/O cost.

effectively speedup the cross-media retrieval performance for handling the large-scale multimedia datasets. In order to effectively index the heterogenous media objects, MK-tree partitioned the data space based on the different media data distribution and key dimensions. Extensive experiments on real large-scale multimedia datasets indicate that our proposal dramatically improves the accuracy and efficiency of cross-media retrieval, outperforming existing methods significantly.

References

[1] Zhang H, Zhuang Y, Wu F. Cross-modal correlation learning for clustering on image-audio dataset. In *Proc. the 15th ACM Int. Conf. Multimeida*, September 2007, pp.273-276.

[2] Yang Y, Zhuang Y, Wu F, Pan Y. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 2008, 10(3): 437-446.

[3] Yang Y, Xu D, Nie F P *et al.* Ranking with local regression and global alignment for cross-meida retrieval. In *Proc. the 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp.175-184.

[4] Lew M, Sebe N, Djeraba C, Jain R. Content-based multimedia information retrieval: State of the art and challenges. *ACM TOMCCAP*, 2006, 2(1): 1-19.

[5] Adams W, Iyengar G, Lin C Y *et al.* Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP J. Adv. Signal. Process*, 2003, 10(2): 170-185.

[6] Kennedy L, Chang S F. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proc. the 6th ACM CIVR*, July 2007, pp.333-340.

[7] Hotelling H. Relations between two sets of variates. *Biometrike*, 1936, 28(3/4): 321-377.

[8] Chang S F, Ma W Y, Smeulders A. Recent advances and challenges of semantic image/video search. In *Proc. ICASSP*, April 2007, pp.12-16.

[9] Snoek C, Worring M. Multimodal video indexing: A review of the state-of-the-art. *MTA*, 2005, 25(1): 5-35.

[10] Smeaton A F, Over P, Kraaij W. Evaluation campaigns and TRECVid. In *Proc. the 8th ACM MIR*, Oct. 2006, pp.321-330.

[11] Paramita M, Sanderson M, Clough P. Diversity in photo retrieval: Overview of the ImageCLEF photo task 2009. In *Lecture Notes in Computer Science 6242*, Peters C, Caputo B, Gonzalo J *et al.* (eds.), Springer-Verlag, 2009, pp.45-59.

[12] Naphade M, Smith J R, Tesic J, Chang S F, Hsu W, Kennedy L, Hauptmann A, Curtis J. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 2006, 13(3): 86-91.

[13] Chen T, Cheng M M, Tan P *et al.* Sketch2Photo: Internet image montage. *ACM TOG*, 2009, 28(5), Article No. 124.

[14] Ajorloo H, Lakdashti A. HBIR: Hypercube-Based Image Retrieval. *J. Comput. Sci. Technol.*, 2012, 27(1): 147-162

[15] Feng B L, Cao J, Bao X G *et al.* Graph-based multi-space semantic correlation propagation for video retrieval. *The Visual Computer*, 2011, 27(1): 21-34.

[16] Csurka G, Skaff S, Marchesotti L, Saunders C. Building look&feel concept models from color combinations with applications in image classification, retrieval, and color transfer. *The Visual Computer*, 2011, 27(12): 1039-1053.

[17] Guttman A. R-trees: A dynamic index structure for spatial searching. In *Proc. SIGMOD*, June 1984, pp.47-57.

[18] Weber G, Schek R, Blott H. A quantiative analysis and performance study for similarity search methods in high-dimensional spaces. In *Proc. the 24th VLDB*, August 1998, pp.194-205.

[19] Jagadish H V, Ooi B C, Tan K L *et al.* iDistance: An adaptive B^+-tree based indexing method for nearest neighbor search. *ACM TODS*, 2005, 30(2): 364-397.

[20] Ciaccia P, Patella M, Zezula P. M-tree: An efficient access method for similarity search in metric spaces. In *Proc. the 23rd VLDB*, August 1997, pp.426-435.

[21] Ji H, Grishman R. Refining event extraction through cross-document inference. In *Proc. the 46th ACL*, Jun. 2008, pp.254-262.

[22] Ji H, Grishman R, Freitag D *et al.* Name extraction and translation for distillation. In *Handbook of Natural Language Processing and Machine Translation*, Olive J, Christianson C, John M (eds.), Springer, 2009, pp.21-29.

[23] Naphade M R, Kennedy L, Kender J *et al.* A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical report RC23612, IBM, May 2005.

[24] Böhm C. A cost model for query processing in high dimensional data spaces. *ACM TODS*, 2000, 25(2): 129-178.

[25] Yanagawa A, Chang S F, Kennedy L, Hsu W. Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. ADVENT Technical Report #222-2006-8, Columbia University, March 2007.

[26] Zhuang Y, Li Q, Chen L. A unified indexing structure for efficient cross-media retrieval. In *Proc. the 14th DASFAA*, April 2009, pp.677-692.

[27] Lu B, Wang G R, Yuan Y. Towards large scale cross-media retrieval via modeling heterogeneous information and exploring an efficient indexing scheme. In *Proc. Computational Visual Media*, Nov. 2012, pp.202-209.

Bo Lu received the B.S. and M.S. degrees in computer science from Northeastern University, China in 2006 and 2008 respectively. Currently, he is a Ph.D. candidate of Northeastern University. His main research interest includes concept-based video retrieval, semantic concept detection and cross-media retrieval.

Guo-Ren Wang is a lecturer at Northeastern University, China. He received his B.E., M.E. and Ph.D. degrees from Northeastern University in 1988, 1991, 1996, respectively. His research interests include machine learning, data mining, data management, bioinformatics and multimedia technology.

Ye Yuan received the B.S., M.S. and Ph.D. degrees in computer science from Northeastern University, China, in 2004, 2007 and 2011, respectively. He is now an associate professor with the College of Information Science and Engineering in Northeastern University. His research interests include graph databases, probabilistic databases, data privacy-preserving and cloud computing.