Pattern Anal Applic
DOI 10.1007/s10044-015-0478-y

THEORETICAL ADVANCES

Topic correlation model for cross-modal multimedia information retrieval

Zengchang Qin[1] · Jing Yu[2] · Yonghui Cong[1] · Tao Wan[3]

Received: 22 August 2013 / Accepted: 17 April 2015

Abstract In this paper, we present a simple and effective topic correlation model (TCM) for cross-modal multimedia retrieval by jointly modeling the text and image components in multimedia documents. In this model, the image component is represented by the bag-of-features model based on local scale-invariant feature transform features, meanwhile the text component is described by a topic distribution learned from a latent topic model. Statistical correlations between these two mid-level features are investigated by mapping them into a semantic space. These cross-modality correlations are used to calculate the conditional probabilities of answers in one modality while given query in the other modality. The model is tested on three cross-modal retrieval benchmark problems including Wikipedia documents in both English and Chinese. Experimental results have demonstrated that the new TCM model achieves the best performance compared to recent state-of-the-art cross-modal retrieval models on the given benchmarks.

✉ Jing Yu
jing.emy.yu@gmail.com; yujing02@iie.ac.cn

Zengchang Qin
zcqin@buaa.edu.cn; zengchang.qin@gmail.com

Yonghui Cong
c_yonghui@163.com

Tao Wan
tao.wan.wan@gmail.com

[1] Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing 100191, China

[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

[3] School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China

Keywords Cross-modal multimedia retrieval · Topic correlation model · Topic models · Bag-of-features model

1 Introduction

In recent years, online multimedia information in various modalities, such as image, text, audio and video, has been increasing explosively. Effective and efficient information search techniques are required to access such massive and multi-modal data. However, the predominant multimedia search engines today are still text-based, which have obvious shortcomings. For example, when we convey a concept, it can be more accurately expressed by using more words. However, for a search engine, such as the ones from Google and Yahoo, a query with a few keywords reflecting the main properties of the concept you are looking for is more desired than detailed descriptions. Using long sentences, though semantic delicacy which can be detected by human, becomes redundant for a search engine. Another shortcoming is that manual information annotation is subjective that may cause retrieval deviations by users with different views of understanding.

In recent years, much effort has been made to solve these problems. Content-based image retrieval (CBIR) [1, 2] aims to retrieve relevant images given a query image based on their visual features. However, the retrieval results exhibit big "semantic gap" between the low-level image features and high-level semantic concepts. Another class of models focuses on automatic image annotation [3–5]. The goal of automatic image annotation is to assign relevant text to a given image. But the annotations are represented by a few words which are limited to accurately describe the semantic content of the image. How to build a joint model to capture the relations between different

Published online: 05 May 2015

modalities and support access to the content in multiple modalities is essential for further progress in the multimedia research area.

In this research, we aim to develop a cross-modal multimedia retrieval system by mapping different modalities, including texts and images, into a semantic space in order to find their semantic topic correlations in this space. The topic correlation can be obtained from a hierarchical representation for both texts and images. Most previous models for cross-modal information retrieval are focused on English corpuses. The techniques used in studying English can be easily extended to other alphabetic languages. A text in English can be regarded as a collection of words, which are the basic semantic units in the majority of Western languages. In this paper, we also consider to develop a new model to retrieve information in Chinese language. Topic modeling of Chinese language has been well studied [6, 7]. Though most previous studies choose Chinese words as the most basic units of the language [8, 9]. Zhao et al. [10] showed the computational evidence that character-based topic models outperform the word-based topic models in the text classification tasks. In this paper, we will investigate topic modeling based on both words and characters.

The rest of the paper is structured as follows: Section 2 briefly introduces the related work. In Sect. 3, we describe the hierarchical image and textual representations for semantic content modeling. The model of topic correlation is presented in details in Sect. 4. To verify the effectiveness of the newly proposed model, comprehensive experiments are conducted on three benchmark datasets, including English Wikipedia, TVGraz, and Chinese Wikipedia. Experimental results are analyzed and compared to previous work in Sect. 5. The conclusions and future work are given in Sect. 6.

2 Related work

Multimedia information retrieval, mainly focusing on the text modality and image modality, has been well studied in both multimedia and information retrieval areas. Roughly, there are two types of methods, one type of methods concentrates on automatic image annotation [3–5], where images are labeled by key words that will be used to match given text query later. The other type is content-based image retrieval [1, 2], where an image is given as a query to retrieve similar images by using visual feature matching techniques. However, these retrieval systems are all unimodals that only consider single modality for both the queries and the retrieval data [11]. Thus, the uni-modal search techniques are not enough to meet the requirements of multimedia search across multiple modalities. How to model semantic relationships between different modalities, such as documents with paired data of images and texts, is essential to many practical applications.

Moving beyond uni-modal retrieval, various models have been proposed to study cross-modal information retrieval. One class of models are based on combination strategies: (1) combining low-level features from different modalities into concise multi-modal features. In [12], a manifold learning algorithm based on Laplacian Eigenmaps is introduced to combine low-level descriptors of each separate modality and map them to a new low-dimensional multi-modal feature space. In this feature space, semantically similar multi-modal data are represented by multi-modal descriptor vectors close to each other. (2) Combining independent systems at different levels. For example, Kliegr et al. [13] utilized a combination of two independent systems at the output level, so that one system models the text data stream and the other models the image data stream. Similar studies with combinations at other levels were reported in [1] and [14]. Iyengar et al. [15] developed a joint retrieval framework, in which individual components are used to model different relationships between documents and queries, and then combined via a linear model. However, most of the combination-based models require queries having the same modalities with the retrieval data, e.g., both queries and retrieval data are constituted by image–text pairs, which are not always available for users. Therefore, these models are actually extensions of uni-modal retrieval methods, supporting retrieval of more than one modality simultaneously.

Regarding cross-modal retrieval, another class of models intends to build generative models for predictive tasks. The key technique involves building a joint model based on the correlations to bridge up the "semantic gap" between different modalities. Following the previous work of Blei et al. [3, 16] presented a correspondence latent Dirichlet allocation (Corr-LDA) to model the images and associated annotations within a shared mixture of latent factors. [3] also proposed a multi-modal LDA (mmLDA) to compute a mean topic distribution topic as the shared variable between different modalities. Other models such as [17–19] are focusing on either optimizing the likelihood of the topic model or the distance between different multimedia documents.

Recently, some attempts bring new perspective for solving the cross-modal retrieval problem. Yang et al. [20] constructed a multimedia correlation space (MMCS), where every multimedia document (including text, images and audio) is represented as a point, based on the multimedia content and the co-occurrence of the heterogeneous data. Then a novel ranking algorithm is used, which adopted a local linear regression model for each point and globally aligned all the regression models by minimizing a

global objective function. Though this method yields significant retrieval performance on the training dataset, the method achieves low retrieval accuracy when the query is out of the dataset, unless the relevance feedback is applied to the ranking procedure. Mahadevan et al. [21] computed the nearest neighbors of the query among the training samples in the original feature space and learned a mapping of the query as a weighted combination of these neighbors. The similarities between different modalities are computed in the mapping space. Mao et al. [22] proposed a parallel field alignment retrieval (PFAR) method, which considered the cross-modal information retrieval as a manifold alignment task employing parallel techniques. Rasiwasia et al. [23] demonstrated the benefits of jointly modeling text and image components by mapping these two modalities into a common space via the canonical correlation analysis (CCA). The joint model greatly improves the cross-modal retrieval accuracy and outperforms state-of-the-art unimodal retrieval approaches. Inspired by [23], we develop a new model for cross-modal retrieval, which is more simple and effective compared to [23]. Different from their work, we aim to design a cross-modal multimedia retrieval system based on the statistical correlations between these two components. As a general cross-modal retrieval system, our model will accomplish two tasks: (1) given a text query, retrieve relevant images, and (2) given an image query, retrieve relevant texts. Our work mainly concentrates on the joint modeling between different modalities by considering given category information.

3 Topic representations

The cross-modal multimedia retrieval task is to handle a large and heterogeneous collection of images accompanied by unstructured and noisy texts. Choosing appropriate content representations that are able to capture the semantic correlations between different modalities is a critical issue in the multimedia retrieval field. Low-level features, such as keywords and captions for texts or colors and textures for images, contain limited semantic information to describe the complex content in the modalities. Recently, the mid-level features, such as visual words in the bag-of-features model [2], and latent topics in topic models [10, 24], attract much attention for their effectiveness in semantic modeling. We have not been aware of any work to investigate the mid-level feature relations between different modalities. In the following section, we will first introduce these mid-level representations and then investigate the correlations between the topics of words and the topics of image features (i.e., visual words in the bag-of-features model) to alleviate the problem of the "semantic gap" between texts and images [2].

3.1 Image representation

The desired representation for images should be robust with small changes, such as scale, illumination, and transformation. Moreover, a good representation is required to map the original image to a lower-dimensional feature space where images within a category are ideally near to each other while keep large distances to the images belonging to other categories.

Among content-based image models, one of the most popular approaches is the bag-of-features (BoF) model [25]. Previous research has shown that the BoF model is robust in object and scene classification [26], image search [2], and video retrieval tasks [27]. The model is invariant to slight changes of features in the local regions by quantizing each feature to a representative visual word. The basic idea of BoF is to describe each image as an orderless collection of local features. The detailed methodology for generating an image representation is described as follows.

3.1.1 Local feature extraction

The first step of the BoF model is to extract discriminant local features, which capture the invariant properties of relevant image changes. The scale-invariant feature transform (SIFT) [28] feature has been proven to be powerful descriptors with respect to different geometrical variations, e.g., translation, scale, rotation, and small distortions. We first search the keypoints in the difference of Gaussian (DoG) space [29] and then compute gradients using Gaussian weighted derivatives in the local regions with region size of 16×16 pixels around the keypoints. Each region is divided into 4×4 spacial bins. In each bin, we linearly interpolate the gradients into 8 directions. Finally, we compute a normalized 128-dimensional vector as the SIFT descriptor for each region.

3.1.2 Codebook generation

Following the local feature extraction procedure, we utilize a k-means clustering to generate a codebook. The local descriptors $\mathbf{d} = [d_1, d_2, \ldots, d_n]$ are divided into k clusters $\mathbf{C} = [C_1, C_2, \ldots, C_k]$. The cluster centers J, also referred to as visual words, are defined as a set of vectors $\{v_i\}$ calculated by minimizing the following equation:

$$J = \arg\min_{\{v_i\}} \sum_{i=1}^{k} \sum_{d_j \in C_i} \left\| d_j - v_i \right\|, \tag{1}$$

where

$$v_i = \frac{1}{|C_i|} \sum_{d_j \in C_i} d_j. \tag{2}$$

3.1.3 Feature quantization

For each image, we use the "hard assignment" method [30] to assign each descriptor to one cluster center via the nearest-neighbor classifier and normalize the resulting histogram. So far, images are represented as distribution histograms over k visual words.

3.2 Text representation

In text modeling, statistical methods have become increasingly popular and attracted more attention compared to classical syntactic rule-based natural language processing (NLP) techniques. Based on the bag of words (BoW) assumption [31], natural language is considered as a set of orderless data and important semantic patterns can be detected and learned by using machine learning algorithms. For example, the topic model, a type of Bayesian generative model with a latent variable for modeling semantic topics, has attracted considerable attention in both machine learning and NLP communities. The main idea of topic models is that documents can be represented as a mixture of latent topics, and each topic is a probability distribution over the vocabulary. The topic models depict a probabilistic procedure to show how documents are described in a concise way. A most well used topic model is latent Dirichlet allocation (LDA) [16] in which a Dirichlet distribution is used to generate a k-dimensional random variable θ as the topic mixture weights. A k-dimensional Dirichlet variable α is conjugate to Multinomial distribution and this property is conductive for the inference and estimation. The LDA can be considered as the following generative process for each text document $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$ in a corpus:

1. Choose $\theta \sim \mathrm{Dir}(\alpha)$.
2. For each of the N words w_n:
   (i) Choose a topic $z_n \sim \mathrm{Multinomial}(\alpha)$.
   (ii) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In the above process, β is a $k \times V$ matrix, where k is the number of topics, and V is the size of vocabulary. Based on the LDA procedure, we can calculate the joint probability of θ, $\mathbf{z}$ and $\mathbf{w}$ given α and β as hyper-parameters, which is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3}$$

The marginal distribution $\mathbf{w}$ can be calculated by summing over z_n and integrating over θ as defined by:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) d\theta \tag{4}$$

The LDA has been widely used in different NLP tasks, such as information retrieval [32], text classification [33], and question answering system [34]. In this work, the text components in the documents with paired texts and images are described as distributions over pre-trained topics by using the standard LDA [43]. Figure 1 gives a schematic illustration of how a multimedia document is processed based on the image components and text components, respectively.

The representations of both images and texts here do not use the low-level features directly. We construct the mid-level representations for modeling the contents in a two-level hierarchical structure to make them more robust and abstract. In this paper, we use the term "topics of features" instead of the visual words in order to highlight the similarity between bag-of-features model and the topic model, because we are interested in studying the correlations between these topics of different modalities.

4 Topic correlation model

In multimedia information retrieval, documents generally contain multiple forms of contents, e.g., texts and images. The retrieval problem is to search semantically matched texts by given a query image and vice versa. Formally, given a set of documents denoted as $\mathbf{D} = [D_1, D_2, \ldots, D_K]$, we assume that each document D_k, $k = 1, \ldots, K$, contains at least an image and associated text. In fact, there can be multiple texts accompanied with more than one image or no image and vice versa. In this paper, we only consider a simplified case of an one-to-one mapping between image and text as shown in Fig. 2, which can be defined as:

$$D_k = [I_k, TX_k], \ k = 1, \ldots, K, \tag{5}$$

where $D_k \in \mathbf{D}$, and I_k and TX_k denote the image and corresponding text in D_k, respectively. The retrieval task is to find the most semantically related TX_k (or I_k) in $\mathbf{D}$ given a query I_q (or TX_q).

Given the above representations of two modalities, the key problem of the cross-modal retrieval is to model the correlations between the text modality and image modality. For simplicity, we introduce a score function to evaluate the correlation of an image I_k given a text query TX_q by

$$S(I_k) = P\left(I_k|TX_q\right) \tag{6}$$

Similarly, the score function for TX_k is:

$$S(TX_k) = P\left(TX_k|I_q\right). \tag{7}$$

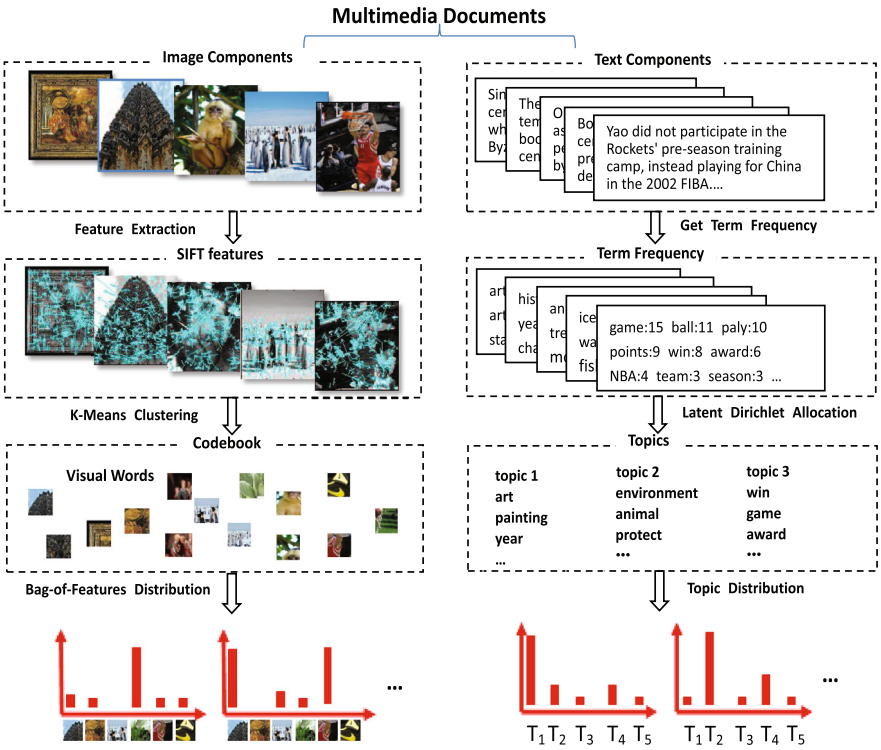

Fig. 1 Topic representations of texts and images in given multimedia documents. On the left-hand side, representations of the image components are based on the bag-of-features model. SIFT features are extracted on all the training images and a codebook is learned from these features. Each image is represented as a distribution over visual words in the codebook. On the right-hand side, representations of the text components are based on the latent Dirichlet allocation model. Texts are first pre-processed and represented in terms of word frequency. The latent Dirichlet allocation is used to learn the topics from the whole corpus and each text is represented as the distribution over these topics

Fig. 2 Definition of multimedia documents in our model. Each document contains a text and its corresponding image

Equations (6) and (7) are used to arrange the retrieval results in a descending order given a query image or text.

4.1 Naive topic correlations

For a specified document, though its contents may be in different modalities, the underlying semantic contents are similar or even identical. In our framework, an image is represented by a distribution over visual words, and a text is described by a distribution over topics. Intuitively, the underlying relationships between some particular visual words and topics may imply a common latent semantic concept.

Let $\mathbf{V} = [V_1, V_2, \ldots, V_M]$ denote a set of visual words in the codebook (M is the codebook size), and $\mathbf{T} = [T_1, T_2, \ldots, T_N]$ is a set of topics (N is a predefined number of topics). For a visual word V_i and a topic T_j, the underlying probabilistic relation can be computed on the training document $\mathbf{D} = [D_1, D_2, \ldots, D_K]$:

$$P(V_i|T_j) = \sum_{k=1}^{K} P(V_i|I_k)P(I_k|TX_k)P(TX_k|T_j), \tag{8}$$

where $P(V_i|I_k)$ is the BoF distribution over V_i of the image I_k. Since the image I_k and the text TX_k appear in the same document D_k, then

$$P(I_k|TX_k) = P(TX_k|I_k) = 1.$$

For the third term $P(TX_k|T_j)$, according to the Bayes theorem, we can obtain

$$P(TX_k|T_j) = \frac{P(T_j|TX_k)P(TX_k)}{\sum_{k=1}^{K} P(T_j|TX_k)P(TX_k)}, \quad (9)$$

where $P(TX_k)$ is the prior probability of the text component in document D_k. $P(T_j|TX_k)$ is the probability of topic T_j given text TX_k that can be predicted by LDA. As no prior information is available on each document, we use the uniform distribution as the prior according to the principle of maximum entropy. Formally,

$$P(TX_k) = P(I_k) = \frac{1}{K}. \quad (10)$$

Similarly, the likelihood of topic T_j given a visual word V_i is computed by:

$$P(T_j|V_i) = \sum_k P(T_j|TX_k)P(TX_k|I_k)P(I_k|V_i), \quad (11)$$

where $P(T_j|TX_k)$ is the topic distribution over T_j given text TX_k. Using the Bayes theorem, $P(I_k|V_i)$ can be defined as:

$$P(I_k|V_i) = \frac{P(V_i|I_k)P(I_k)}{\sum_{k=1}^{K} P(V_i|I_k)P(I_k)}. \quad (12)$$

Based on the above correlation between the topics of words and the topics of features, we can calculate the relevance between any images (texts) and a text (image) query. The likelihood of being the image I_k given a query text TX_q and vice versa can be evaluated by:

$$P(I_k|TX_q) = \sum_i \sum_j P(I_k|V_i)P(V_i|T_j)P(T_j|TX_q) \quad (13)$$

$$P(TX_k|I_q) = \sum_i \sum_j P(TX_k|T_j)P(T_j|V_i)P(V_i|I_q). \quad (14)$$

However, such correlations do not take into account any complications regarding how images and texts are semantically related. The model only uses the naive probabilistic relations between the topics of words and the topics of features. In [35], the experimental results showed that this correlation is weak. If we are able to find the correlations between some specific topics of words and topics of features within documents belonging to one category, the correlations are relatively strong. However, this model intends to relate the images to texts based on the mid-level features and does not consider the category information. The entire dataset, including documents from different categories, is used to train the model, which weakens the desired correlations significantly. Moreover, the correlations between some specific topics of words and topics of features are weak when the correlations on the documents are from different categories.

4.2 Semantic topic correlations

Instead of directly mining the correlation between mid-level features of texts and images, we can represent both of these modalities at a semantic level and map them into a common semantic space, where correlations between texts and images can be built at this more abstracted level. A feasible way to correlate texts and images with semantic-level concepts is to assign a semantic concept to each multimedia document in the datasets, thus the text and image in the document will be labeled by the same concept. In the experiments, the semantic concept of each document is the same as its category predefined by the datasets. In our model, we consider local correlation based on the category information and map the topic representations of both images and texts to a semantic space, which has a meaningful concept for each dimension. We refer to this model as a local topic correlation model (TCM).

Given the category information, two semantic mappings are implemented by training two multi-class classifiers on the BoF descriptors of images and the topic descriptors of texts, respectively. Then each image I_k in the topics of features space can be mapped into a vector of posterior probabilities $P(C_i|I_k)$, where C_i is the ith category given the predefined categories of documents $\mathbf{C} = [C_1, C_2, \ldots, C_n]$. These posterior vectors exist in a new space call the semantic space. Similarly, each text T_k in the topics of words space can be mapped into a vector of posterior probabilities $P(C_i|T_k)$. These vectors are in the same semantic space as images' vectors because each dimension of these two kinds of vectors indicates the same document category. Figure 3 shows the schematic illustration of the semantic mapping procedure for this cross-modal retrieval model.

One of the possible ways to compute the vectors of posterior probabilities is to apply multi-class support vector machine (SVM) [44]. This builds multiple 'one-versus-one' binary classifiers and each binary classification is considered to be a voting casting for one category [36]. By normalizing the votes of all the categories, we obtain a vector of posterior probabilities over all the categories for each image ($P(C_i|I_k)$) or text ($P(C_i|T_k)$). A multi-class SVM classifier is trained for the images and texts, respectively, to map their mid-level features to the same semantic space. Since it is common in probability estimation that estimated probability can be inaccurate with small number of training data, it is not necessary to compute the posterior probabilities explicitly and other algorithms for multi-class classification, such as k-nearest neighbor, neural networks, or logistic regression can be used to obtain the posterior probabilities here.

After semantic mapping, correlations between texts and images can be established by computing the conditional

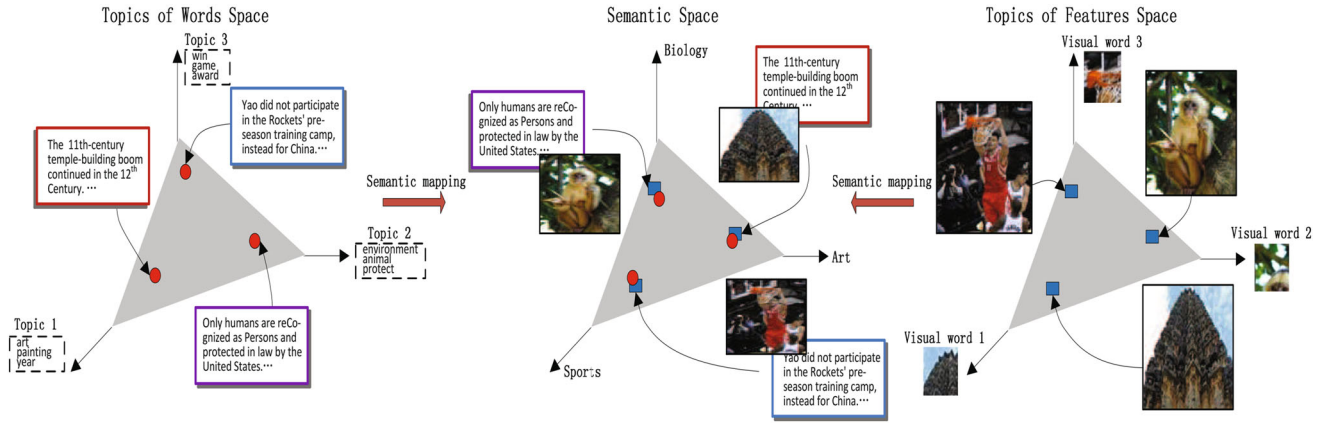

Fig. 3 Schematic illustration of the semantic mapping for the TCM model. (*Left*) Semantic mapping of the text components from corresponding topics of words space, learned by LDA, to the semantic space, learned by the multi-class classifier for texts. (*Right*) Semantic mapping of the image components from associated topics of features space, learned by BoF, to the semantic space, learned by the multi-class classifier for images

probability of a retrieved image given a query text or vice versa in the retrieval procedure. Given a query text (image), represented by a vector of probability $P(C_i|T_k)$ ($P(C_i|I_k)$), text classifier (image classifier) is utilized to predict its probability distribution over categories. The probability of the image I_k given a text query TX_q is then computed by summing up the conditional probabilities across all the categories as expressed by:

$$P(I_k|TX_q) = \sum_i P(I_k|C_i)P(C_i|TX_q) \tag{15}$$

Based on the Bayes theorem, we can obtain

$$P(I_k|C_i) = \frac{P(C_i|I_k)P(I_k)}{\sum_k P(C_i|I_k)P(I_k)}, \tag{16}$$

where $P(C_i|I_k)$ and $P(C_i|TX_q)$ can be obtained through the predictions from the learned multi-class SVM classifiers. These two probabilities are not necessarily the same since the classifiers are trained individually based on the contents of different modalities. The score function $S(TX_k)$ for text TX_k is assigned as the values of $P(I_k|TX_q)$, which is used to rank the retrieved images in a descending order.

Similarly, given an image query I_q, the probability of the text component is computed by:

$$P(TX_k|I_q) = \sum_i P(TX_k|C_i)P(C_i|I_q), \tag{17}$$

where $P(TX_k|C_i)$ is evaluated by:

$$P(TX_k|C_i) = \frac{P(C_i|TX_k)P(TX_k)}{\sum_k P(C_i|TX_k)P(TX_k)} \tag{18}$$

where $P(C_i|TX_k)$ and $P(C_i|I_q)$ can be obtained through the predictions from the learned multi-class SVM classifiers. The score function $S(I_k)$ for image I_k is assigned as the values of $P(TX_k|I_q)$ and that the values are used to rank the retrieved texts in a descending order.

5 Experimental studies

To evaluate the effectiveness of the TCM-based system, we conduct a number of experiments on three datasets, including English Wikipedia, TVGraz, and Chinese Wikipedia. The retrieval results are compared to the existing state-of-the-art approaches for the following tasks: (1) given an image query from the test set, the retrieval system returns a ranked set of all texts from the training dataset, and (2) query a text to obtain a ranked list of images. The mean average precision (MAP) [37] and precision-recall (PR) curves [38] are adopted to measure the retrieval performance.

5.1 Dataset description

Three benchmark datasets are tested to evaluate the retrieval performance of the presented TCM model.

5.1.1 English wikipedia

The English Wikipedia corpus [45] is a collection of "Wikipedia featured articles", which has been first used in [23]. We name it En-Wikipedia for short to make difference from the Ch-Wikipedia which will be mentioned later. It contains 2866 paired images and texts that are divided into 10 categories. The article in each document is split into sections according to the section headings. The first image associated with a particular section is chosen as its related image for this document. The sections within the document

Table 1 Summary of the En-Wikipedia dataset

Category	Training	Testing	Total
Art and architecture	138	34	172
Biology	272	88	360
Geograpgy and places	244	96	340
History	248	85	233
Literature and theatre	202	65	267
Media	178	58	236
Music	186	51	237
Royalty and nobility	144	41	185
Sports and recreation	214	71	285
Warfare	347	104	451

Table 2 Summary of the TVGraz dataset

Category	Training	Testing	Total
Brain	148	50	198
Butterfly	197	65	262
Cactus	144	48	192
Deer	224	74	298
Dice	192	64	256
Dolphin	201	56	257
Elephant	153	50	203
Frog	232	77	309
Harp	159	53	212
Pram	147	48	195

without images are ignored. In our experiments, the processed dataset is randomly divided into two parts with three-fourths the documents (2173) for training and the remaining one-fourth (693) for testing. The definition of each category and the numbers for training and testing documents are shown in Table 1. Three sample images for each category are shown in Fig. 4. In the Wikipedia dataset, texts are well expressed and can be representative to their semantic categories. However, images in each category are relatively ambiguous. For instance, a portrait of a historical figure can appear in multiple categories, such as "art", "history", "literature", and "warfare". This leads to ambiguity for correctly classifying these images, because categories in Wikipedia are abstract and have overlaid semantics.

5.1.2 TVGraz

The TVGraz dataset [46] is a collection of webpages including images and texts [39]. It contains the top 1000 results from Google image search for each of 10 categories from the Caltech-256 [40]. The database is pre-processed and contains 2594 image–text pairs. We choose the texts that have more than 10 words in our experiments and there are 2382 documents in total. The average length of the texts is 361 words.

The three-fourths of documents (1789) are randomly selected for training and the remaining one-fourth (593) of the documents are used for testing. The definition of each category and the numbers for training and testing documents are shown in Table 2. Figure 5 shows three sample images for each category.

For above two English datasets (En-Wikipedia, TVGraz), we first pre-process the raw text documents by parsing them into words and deleting punctuation as well as numbers. A stop-word list [47] is then applied to remove insignificant words, such as "if", "a", "with", and "I". Finally, a stemming process is used to represent words by their roots. For instance, "paint", "paints", "painted" , and "painting" are represented by the word "paint".

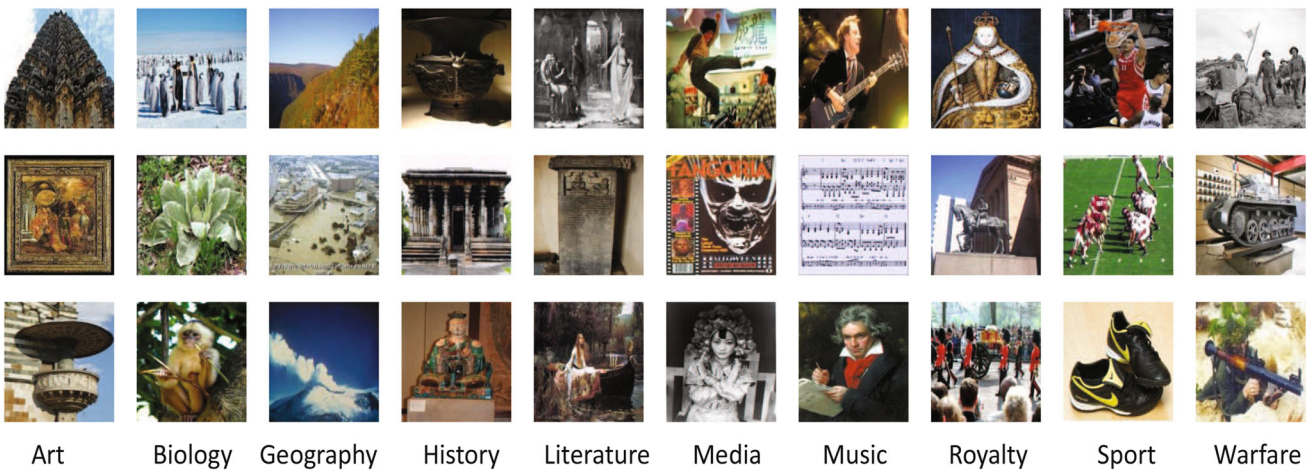

Fig. 4 Samples of the ten categories of the En-Wikipedia dataset

Fig. 5 Samples of the ten categories of the TVGraz dataset

Table 3 Summary of the Ch-Wikipedia dataset

Category	Training	Testing	Total
Culture	285	71	356
Biology and medicine	327	82	409
Natural science	279	70	349
Geography	374	93	467
History	424	106	530
Traffic	156	39	195
Warfare and militaty	206	52	258
Scholar and occupational figures	145	36	181
Political and military figures	286	72	358

5.1.3 Chinese wikipedia

Since there is no well established image–text paired Chinese corpus for cross-modal retrieval research, we create a dataset named Ch-Wikipedia [48]. It consists of 3103 documents of paired texts and images from 9 categories. The definition of each category and the numbers for training and testing documents are shown in Table 3. Three sample images for each category are shown in Fig. 6. The documents in this corpus are obtained from the contents of Chinese Wikipedia, which is one of the biggest online information websites in Chinese language. There are 20 classes in original corpus covering literature, media, sports, politics and other topics. Each article is split into multiple parts by section headings. The texts containing less than 100 Chinese characters are ignored. The first image associated with a text is chosen as its related image and the texts without images are removed. Topics of similar classes are integrated into one category. For example, "humanities" and "culture & society" are combined into "culture". Some independent classes with less than 150

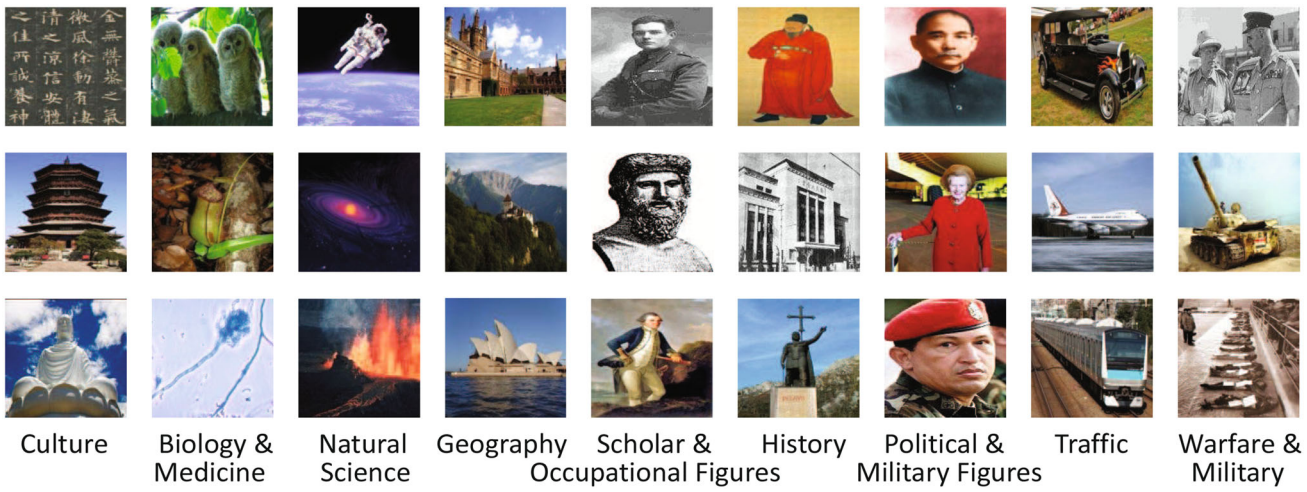

Fig. 6 Samples of the nine categories of the Ch-Wikipedia dataset

Table 4 Accuracy of SVM classifiers with different kernels on two English datasets

Kernel	Training images	Test images	Training text	Test text	Average	Dataset
Linear	**.347**	**.285**	**.804**	.812	**.562**	En-Wikipedia
RBF	.297	.273	.794	**.821**	.546	
Polynomial	.201	.166	.638	.541	.341	
Sigmoid	.290	.263	.763	.812	.532	
Linear	**.602**	**.594**	**.922**	**.985**	**.776**	TVGraz
RBF	.515	.487	.895	.902	.700	
Polynomial	.258	.218	.686	.593	.439	
Sigmoid	.484	.426	.871	.917	.676	

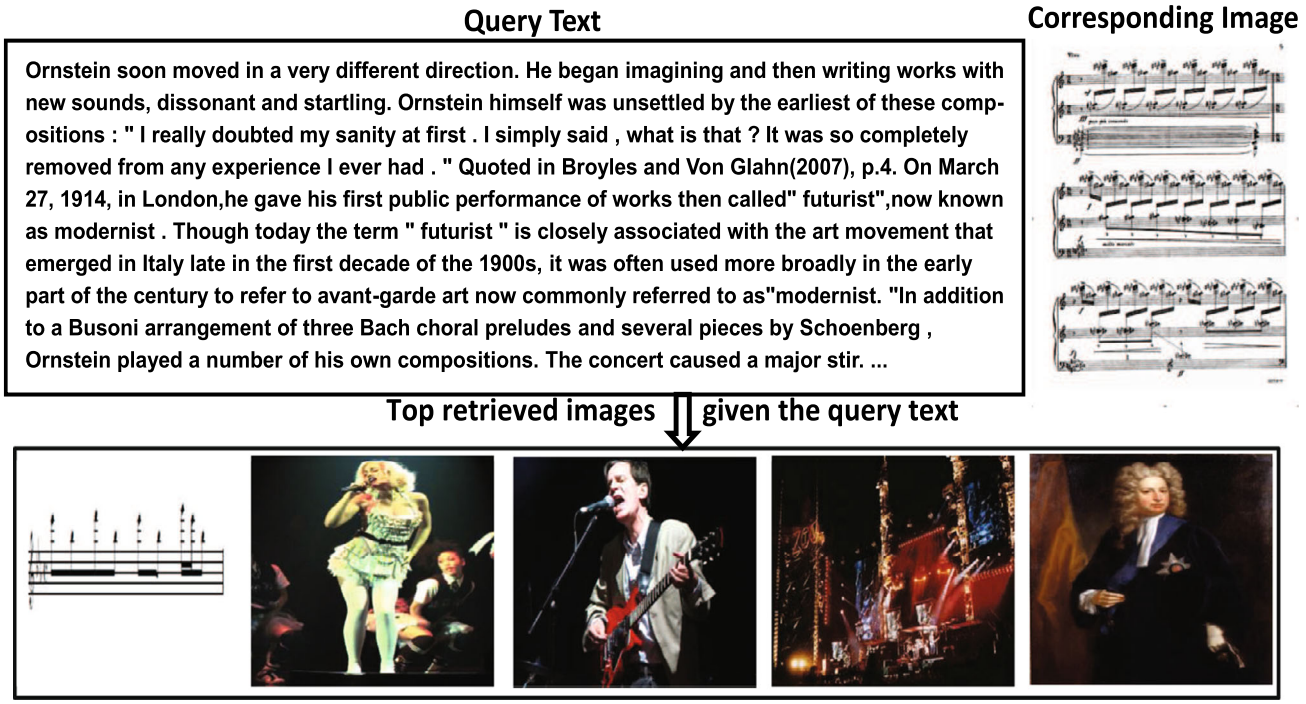

Fig. 7 Retrieval samples of TCM for text query task on the En-Wikipedia dataset. Given the query text on the top left, the top five retrieved images are shown on the *bottom*. We also show the image corresponding to the query image to have a clear semantic comparation with the retrieval results

documents are discarded. The raw tests are pre-processed by removing punctuation as well as numbers.

5.2 Experiments on English corpus

We first conduct the retrieval experiments on the En-Wikipedia dataset. We compare our TCM with the semantic correlation matching (SCM) model [23], Fast version of Maximum Covariance Unfolding (Fast-MCU) [21], and parallel field alignment retrieval (PFAR) [22]. Since Fat-MUC and PFAR model only published MAP on the Eng-Wikipedia dataset, we compare our TCM with these models on this dataset only. In our experiments, we set the topic number and the codebook size as 100 on the En-Wikipedia, which are the same parameters as used in [23].

The SVM classifiers are trained on the text features and image features in both the topics of word space and feature space, respectively. The experimental results using the SVM classifiers trained with different kernels are shown in Table 4. Each of the bold values is the highest accuracy among all the classification results predicted by the four classifiers with different kernels in the same dataset. We found that the linear kernel obtains the best classification accuracy on average. Therefore, the linear kernel SVM classifier is chosen to perform the experiments on all the test datasets.

For visual examination, two examples for retrieval task given a query test or image are displayed in Figs. 7 and 8, respectively. In Fig. 7, the text query is a paragraph related to “music”. The corresponding image (shown on the top

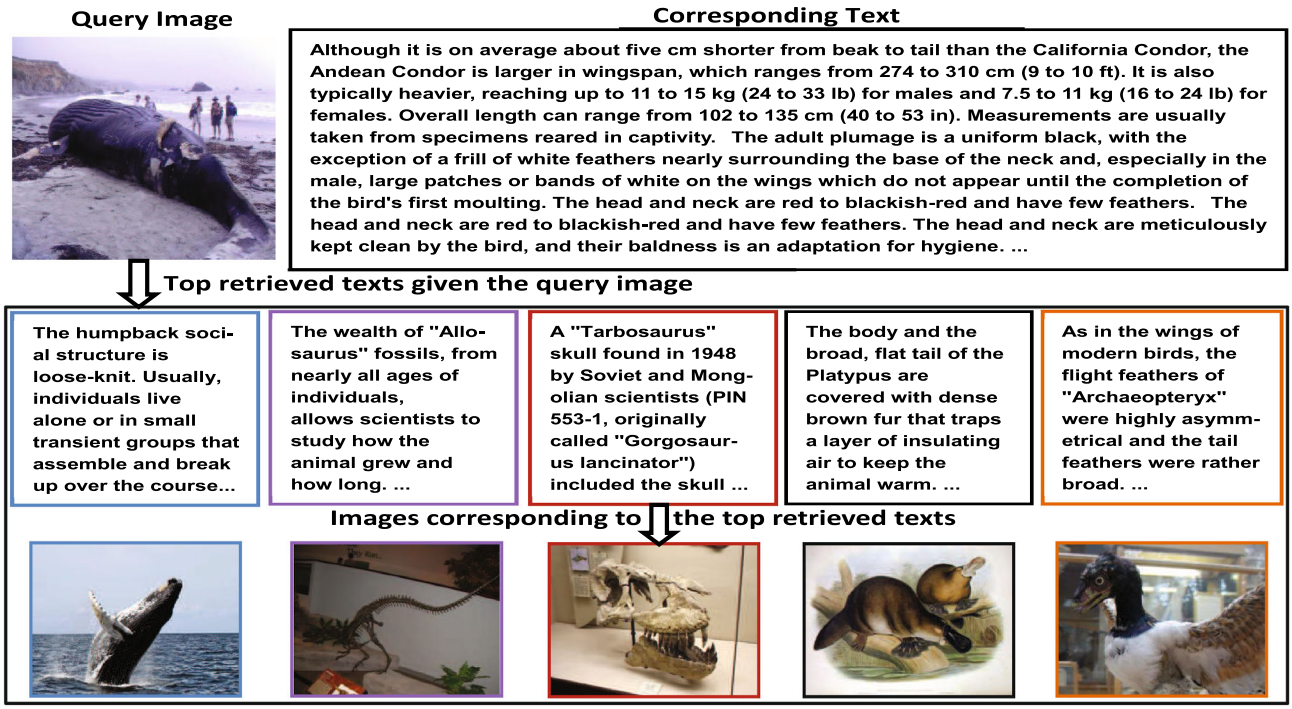

Fig. 8 Retrieval samples of TCM for image query task on the En-Wikipedia dataset. Given the query image on the top left, the top five retrieved texts are shown on the *bottom*. For clear semantic comparisons, we also show the text corresponding to the query image and images corresponding to the retrieved texts

Table 5 Result comparisons using MAP measure on the English datasets

Model	Image query	Text query	Average	Dataset
Random [23]	.118	.118	.118	
SCM [23]	.277	.226	.252	En-Wikipedia
Fast-MCU [21]	.287	.224	.256	
PFAR [22]	.298	.273	.286	
TCM	**.293**	**.232**	**.266**	
Random [41]	.119	.119	.119	TVGraz
SCM [41]	.693	.696	.694	
TCM	**.694**	**.706**	**.700**	

right of Fig. 7) served as ground truth. The top five retrieved images obtained from the TCM model include images of music sheet, singers, and concerts which are semantically related to "music". Figure 8 shows that the corresponding images of top five retrieved texts are semantically related to the query image. Both examples demonstrated that the TCM model is an effective cross-modal retrieval model by jointly estimating the correlations between images and texts.

The MAP performance of TCM, SCM, Fast-MCU, and PFAR models on the En-Wikipedia are shown in Table 5. Each of the bold values is the highest accuracy among all the retrieval results by different models in the same dataset and its corresponding retrieval task. The baseline is computed on the random retrieval results [23]. It is noted that the TCM model significantly improves the retrieval results compared to the baseline, particularly for the average MAP. Further, the TCM model outperforms the SCM and Fat-MUC model in both image and text queries and is comparable with PFAR model in image queries. Since only SCM published the MAP scores of each category, we compare the histograms of our TCM, SCM and the random case. Figure 9 shows the MAP histograms of these two models and baseline for each category of the En-Wikipedia dataset. TCM yields the highest values of MAP over all categories for the text query. Moreover, TCM is more efficient in the training and retrieving procedures compared to the SCM model, which requires to map texts and images

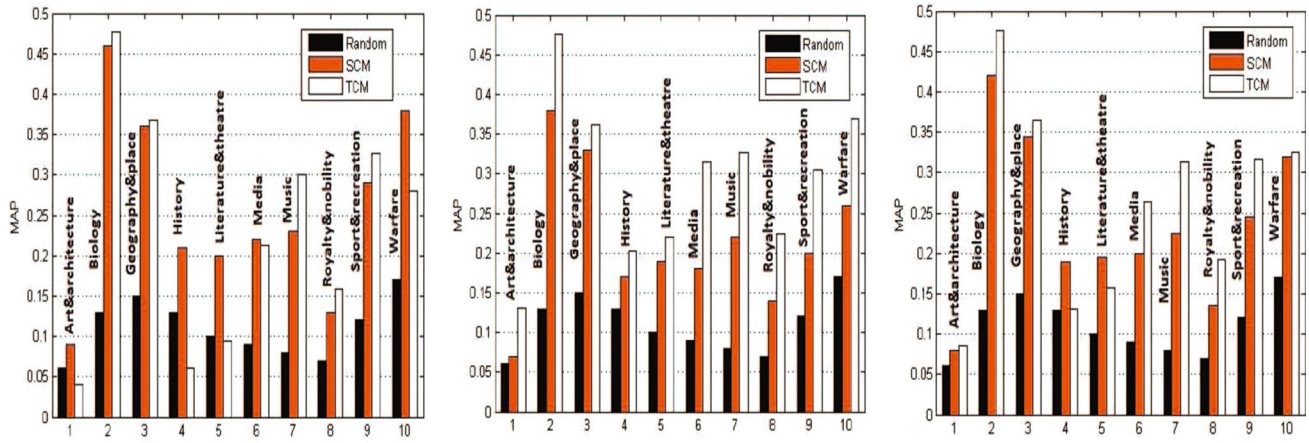

Fig. 9 Comparison of retrieval MAP on all the categories on the En-Wikipedia dataset. (*Left*) MAP of the query images; (*middle*) MAP of the query texts; (*right*) average performance for both the query images an the query texts

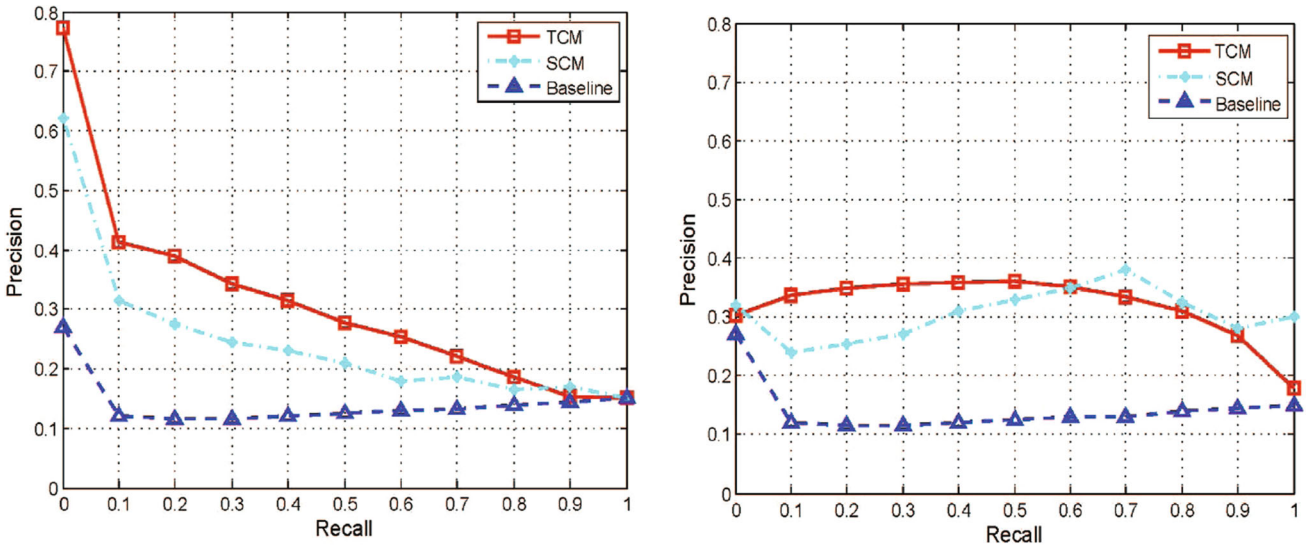

Fig. 10 Comparison of PR curves for (*left*) text query and (*right*) image query on En-Wikipedia

to a correlation space via CCA method to enhance the joint information between two modalities. Figure 10 shows PR curves of cross-modal retrieval results using TCM, SCM, and the baseline on the En-Wikipedia corpus. It is clear to see that the TCM model greatly improves the precision for both text and image retrieval tasks compared to the baseline method. For text query, the TCM model outperforms the SCM model at all levels except the level 1. For image query, the TCM model yields higher precision than SCM at five levels of recall and comparable results at three levels of recall. The curves of first six levels of recall in Fig. 10 indicated that more related retrieval texts obtained from the TCM model were highly ranked than the results from the SCM model. This is more applicable in practice since users are more concerned about the top retrieval results. The retrieval performance suggests that the cross-modal topic correlations based on mid-level topic representations are benefited by modeling the joint relations between different modalities, which is consistent for both image and text retrieval tasks.

To further verify the effectiveness of the TCM model, we also tested the model on the TVGraz dataset. In the experiments, we set the topic number to 100 and the size of codebook to 200, which have been reported to yield the best average retrieval performance in [41]. The MAP values shown in Table 5 demonstrated that the TCM model achieves the best retrieval results with up to 600 % improvement compared to the baseline method. Again, TCM outperforms the SCM model in both image and text queries on the TVGraz dataset. Figure 11 shows the histogram comparison between TCM, SCM and baseline method for each category. For most categories, TCM obtains higher

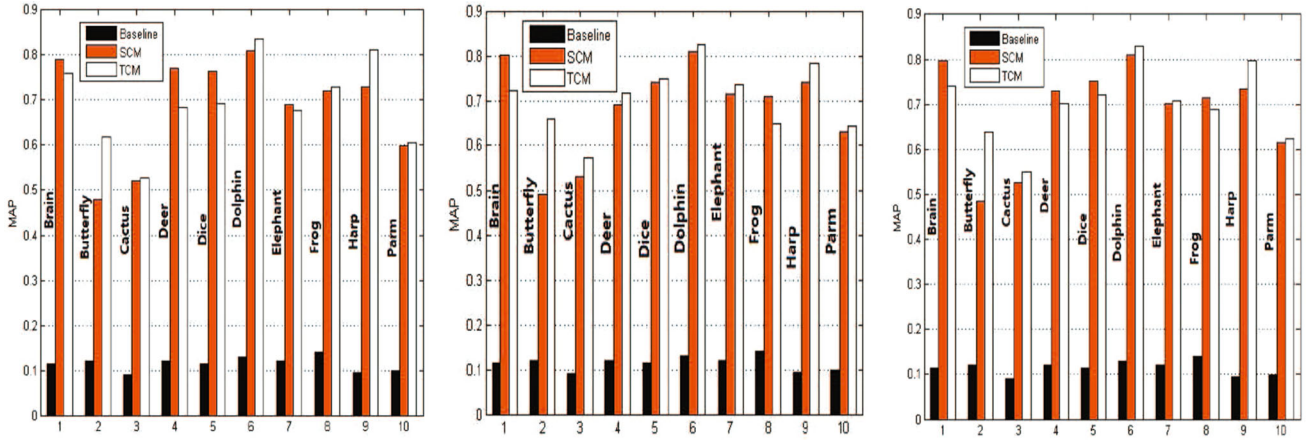

Fig. 11 Comparison of retrieval MAP on all the categories on the TVGraz dataset. (*Left*) MAP of the query images; (*middle*) MAP of the query texts; (*right*) average performance for the query images and the query texts

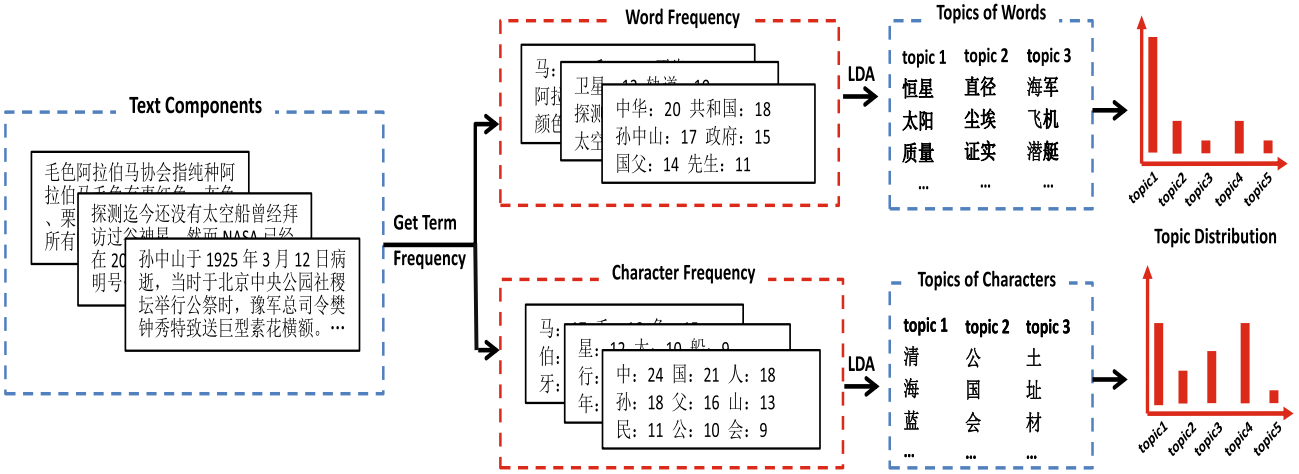

Fig. 12 Chinese representation based on latent Dirichlet allocation [16]

MAP values than the SCM model, thus leading to overall superior retrieval performance.

5.3 Experiments on Chinese corpus

In this section, we evaluate the new model on Chinese multimedia datasets. The morphology of Chinese language is different from Western languages, such as English, since characters, instead of words, are the basic structure units for Chinese language. This has been discussed in Chinese linguistics [42] and verified by using computational models [10, 24]. In practice, both the characters and words serve as indispensable parts for Chinese language. For example, character *bāo* means bag. By combining with the character *qián* (money), it becomes the word *qián bāo*, which means wallet. By combining with the character *shū* (book), it becomes the word *shū bāo* that means schoolbag. By combining with the character *pí* (leather), it becomes the word *pí bāo* (briefcase). Each Chinese character carries ambiguous semantic meaning. By forming a word, the semantic meaning is refined. In this research, we consider both words and characters as the basic units to model the text components in Chinese dataset. We refer these two kinds of topic models as the word-based topic model and character-based topic model, respectively. Figure 12 shows these two topic models that are applied to Chinese texts.

For the Chinese corpus, the image representations are as the same as used in English corpus. Since the morphology of Chinese language is different from Western languages, both characters and words are used as basic terms to model the Chinese text components by LDA.

These two topic models are named as word-based and character-based topic models. To build the vocabulary for the Chinese character-based topic model, the characters that appear less than 3 times in the whole corpus are removed. The characters are considered as stop words if they

Table 6 Accuracy of SVM classifiers with different kernels on the Ch-Wikipedia

Data	Linear	Polynomial	RBF	Sigmoid
Training images	**.356**	.203	.316	0.303
Test images	**.309**	.198	.308	**.309**
Word-based training texts	.642	.548	**.663**	0.627
Word-based test texts	.654	.385	.665	**.668**
Character-based training texts	**.820**	.639	.811	.776
Character-based test texts	.712	.533	**.721**	.704
Average	**.582**	.418	.576	.564

Table 7 Comparison results using MAP measure on the Ch-Wikipedia dataset

Model	Image query	Text query	Average	Dataset
Word-based TCM	.241	.298	.269	Ch-Wikipedia
Character-based TCM	**.310**	**.317**	**.313**	

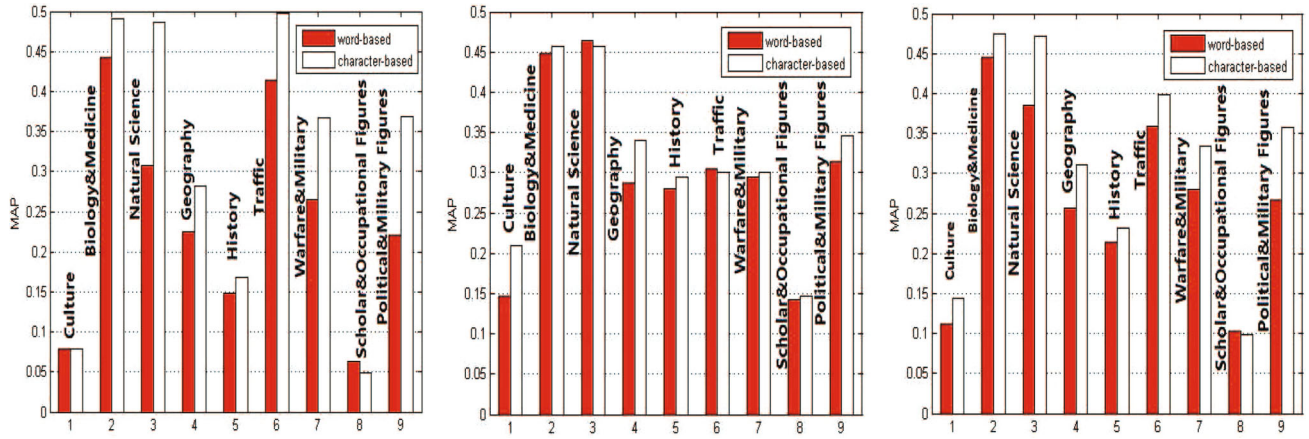

Fig. 13 Comparison of retrieval results in MAP on all the categories on the Ch-Wikipedia dataset. (*Left*) MAP of the query images; (*Middle*) MAP of the query texts; (*right*) average performance for both the query images and the query texts

appear in over 50 % of the documents [10]. After the preprocessing procedure, we obtain 21240 unique Chinese words and 3419 unique Chinese characters.

The SVM classifiers using different kernels are trained on the obtained data and the values of classification accuracy are listed in Table 6. In each row, the bold values is the highest accuracy among all the classification results predicted by the four classifiers with different kernels in the same data. For both word-based and character-based topic models, the topic number and codebook size are assigned as 100. We can see that the SVM classifier with linear kernel yielded the highest classification accuracy compared to other kernels, similar to the English corpus. It is utilized for evaluating the word-based and character-based TCM models.

The comparison results for image and text query using the word-based and character-based models on Chinese corpus are tabulated in Table 7. For each retrieval task listed in each column, the bold values is the highest accuracy among the two retrieval results respectively obtained by word-based TCM and character-based TCM. By testing different parameter settings, we found that the topic number of 500 and the codebook size of 100 gave the best retrieval performance for both models. By inspecting the MAP values shown in Table 7, we noted that both word-based and character-based models are effective for modeling multiple modalities. Moreover, the character-based model yielded better retrieval results than the word-based model with 6.9 % improvement for image query and 1.9 % for text query. The main reason lies in that the size of word vocabulary is larger than the size of character vocabulary. Less words in the vocabulary than the characters appear in the text corpus, which leads to a lower log likelihood of a perplexity measure. The perplexity is used to evaluate the ability of a language model to generalize to unseen data [10].

Figure 13 illustrates the histogram comparison of MAP values on each category by using the word-based and character-based models to perform image and text query. For most categories, the character-based model outperforms the word-based models, especially for the image query task. There is significant improvement that can be observed by using the character-based model as shown in the left figure. It is consistent with the results that were reported in [10], suggesting better retrieval performance can be achieved by using the character-based model for Chinese corpus.

6 Conclusions and future work

In this paper, we presented a topic correlation model (TCM) for cross-modal multimedia retrieval. The statistical relations between mid-level features from different modalities were investigated. In order to verify the effectiveness of the new model, extensive experiments were conducted on three benchmark multimedia datasets containing paired images and texts. The results showed that the new model achieved the best performance compared to SCM and Fast-MCU. The main contributions of this research can be summarized as follows: (1) a simple and effective topic correlation model is presented for cross-modal information retrieval by modeling statistical correlation between mid-level features of different modalities; (2) the new model outperforms most of the state-of-the-art cross-modal retrieval models on given benchmark problems; and (3) the model can be applied to retrieval tasks in another languages. By considering the morphology of Chinese, word-based and character-based models were studied and evaluated on a Chinese Wikipedia dataset. The experimental results demonstrated that the character-based TCM works better than the word-based model.

The future work will be focused on the following issues: (1) a better understanding of deep semantic correlation between different modalities is necessary. A generative process can be considered that images and texts in one document are generated by the same hidden semantic concepts; (2) given a multimedia document, the information presented in both modalities (image and text) are actually redundant, due to the fact that there is unrelated information when considering the cross-modality semantic correlations. It remains unknown that how to filter irrelevant texts that has no corresponding images, or vice versa. This noise control process may significantly improve the quality of retrieval.

Acknowledgments This research is funded by the National Science Foundation of China No. 61305047 and No. 61401012.

References

1. Pham T, Maillot N, Lim J, Chevallet J (2007) Latent semantic fusion model for image retrieval and annotation. In: Proceedings of ACM conference on information and knowledge management, pp 439–444
2. Yuan X, Yu J, Qin Z, Wan T (2011) A SIFT-LBP retrieval model based on bag-of features. In: Proceedings of international conference on image processing. IEEE, New Jersey, pp 1061–1064
3. Blei D, Jordan M (2003) Modeling annotated data. In: Proceedings of ACM SIGIR conference on research and development in information retrieval, pp 127–134
4. Carneiro G, Chan A, Moreno P, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29(3):394–410
5. Li J, Wang J (2008) Real-time computerized annotation of pictures. IEEE Trans Pattern Anal Mach Intell 30(6):985–1002
6. Wu Y, Ding Y, Wang X, Xu J (2010) A comparative study of topic models for topic clustering of Chinese web news. In: Proceedings of international conference on computer science and information technology, pp 236–240
7. Zhang Y, Qin Z (2010) A topic model of observing Chinese characters. In: Proceedings of International conference on intelligent human-machine systems and cybernetics, pp 7–10
8. Gong Z, Zhou G (2011) Employing topic modeling for statistical machine translation. In: Proceedings of international conference on computer science and automation engineering, vol 3, pp 24–28
9. Ni X, Sun J, Hu J, Chen Z (2009) Mining multilingual topics from wikipedia. In: Proceedings of international conference on World Wide Web, pp 1155–1156
10. Zhao Q, Qin Z, Wan T (2011) What is the basic semantic unit of Chinese language? A computational approach based on topic models. In: Proceedings of Meeting on Math of Lang, vol 6878, pp 143–157
11. Datta R, Joshi D, Li J, Wang J (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
12. Lazaridis M, Axenopoulos A, Rafailidis D, Daras P (2013) Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. Signal Process 28(4):351–367
13. Kliegr T, Chandramouli K, Nemrava J, Svatek V, Izquierdo E (2008) Combining image captions and visual analysis for image concept classification. In: Proceedings of international workshop on multimedia data mining, pp 8–17
14. Westerveld T (2002) Probabilistic multimedia retrieval. In: Proceedings of ACM SIGIR conference, pp 438–439
15. Iyengar G, Duygulu P, Feng S et al. (2005) Joint visual-text modeling for automatic retrieval of multimedia documents. In: Proceedings of ACM conference on multimedia, pp 21–30
16. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
17. Putthividhy D, Attias H, Nagarajan S (2010) Topic regression multi-modal latent dirichlet allocation for image annotation. In: Proceedings of international conference on computer vision and pattern recognition, pp 3408–3415
18. Jia Y, Salzmann M, Darrell T (2011) Learning cross-modality similarity for multinomial data. In: Proceedings of international conference on computer vision, pp 2407–2414
19. Virtanen S, Jia Y, Klami A, Darrell T (2012) Factorized multimodal topic model. In: Proceedings of international conference on uncertainty in artificial intelligence, pp 843–851
20. Yi Y, Dong X, Fei N, et al. (2009) Ranking with local regression and global alignment for cross media retrieval. In: Proceedings of international conference on multimedia, pp 175–184

21. Mahadevan V, Wong W, Pereira C et al. (2011) Maximum covariance unfolding: manifold learning for bimodal data. Adv Neural Inf Process Syst, pp 918–926
22. Mao X, Lin B, Cai D et al. (2013) Parallel field alignment for cross media retrieval. In: Proceedings of the 21st ACM international conference on multimedia, pp 897–906
23. Rasiwasia N, Pereira J, Coviello E, Doyle G, Lanckriet G, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM conference on multimedia, pp 251–260
24. Zhao Q, Qin Z, Wan T (2011) Topic modeling of Chinese language using character-word relations. In: Proceedings of international conference on neural information processing, pp 139–147
25. Li F, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE international conference on computer vision, pp 524–531
26. Jiang Y, Ngo C, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of ACM conference on image and video retrieval, pp 494–501
27. Streicher A, Burkhardt H, Fehr J (2009) A bag of features approach for 3D shape retrieval. In: Proceedings of international symposium on advances in visual computing, pp 34–43
28. Lowe D (1999) Object recognition from local scale-invariant features. In: Proceedings of international conference on computer vision, pp 1150–1157
29. Burt P, Adelson A (1983) The laplacian pyramid as a compact image code. IEEE Trans Commun 31(4):532–540
30. Kearns M, Mansour Y, Ng A (1997) An information-theoretic analysis of hard and soft assignment methods for clustering. In: Proceedings of conference on uncertainty in artificial intelligence, pp 282–293
31. Harris Z (1981) Distributional structure. Papers on syntax 14:3–22
32. Wei X, Croft B (2006) LDA-based document models for ad-hoc retrieval. In: Proceedings of ACM SIGIR conference on research and development in information retrieval, pp 178–185
33. Steyvers M, Griffiths T (2007) Probabilistic topic models. Signal Processing Magazine 27(6):55–65
34. Qin Z, Thint M, Huan Z (2009) Ranking answers by hierarchical topic models. In: Proceedings of international conference on industrial, engineering and other applications of applied intelligent systems, vol 5579, pp 103–112
35. Yu J, Cong Y, Qin Z, Wan T (2012) Cross-modal topic correlations for multimedia retrieval. In: Proceedings of international conference on pattern recognition, pp 246–249
36. Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27
37. Zhu M (2004) Recall, precision and average precision. Technical Report, University of Waterloo
38. Gordon M, Kochen M (1989) Recall-precision trade-off: a derivation. J Am Soc Inf Sci 40(3):145–151
39. Khan I, Saffari A, Bischof H (2009) TVGraz: multi-modal learning of object categories by combining textual and visual features. In: Proceedings of workshop of the Austrian association for pattern recognition, pp 213–224
40. Griffin G, Holub A, Perona P (2007) The caltech-256. Technical report, Caltech
41. Rasiwasia N (2011) Semantic image representation for visual recognition. Dissertation, University of Caliifornia
42. Xu T (2001) Fundamental structural principles of Chinese semantic syntax in terms of Chinese characters. Appl Linguist 1:3–13 (In Chinese)
43. http://www.cs.princeton.edu/∼blei/lda-c/
44. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
45. http://www.svcl.ucsd.edu/projects/crossmodal/
46. http://www.icg.tugraz.at/Members/kahn/TVGraz_dataset.tar.gz/view
47. http://www.textfixer.com/resources/common-english-words.txt
48. http://icmll.buaa.edu.cn/zh_wikipedia