CHAPTER 3

Topic Correlations for Cross-Modal Multimedia Information Retrieval

Jing Yu and Zengchang Qin

Advanced information retrieval systems face a great challenge arising from the emergence of massive and multi-modal data, including images, texts, video, audio and etc. One of the most important problems in this field is to accomplish effective and efficient search across various modalities of information. Given a query from one modality, it is desirable to retrieve semantically relevant answers in all the available modalities. This chapter first briefly reviews related works, including uni-modal information retrieval, multi-modal information retrieval, and cross-modal information retrieval. For cross-modal retrieval models, this chapter gives an introduction to manifold alignment-based

Jing Yu
Institute of Information Engineering, Chinese Academy of Sciences,
Beijing, 100029, China
Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University,
Beijing, 100191, China
e-mail: yujing02@iie.ac.cn

Zengchang Qin
Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University,
Beijing, 100191, China
e-mail: zcqin@buaa.edu.cn

Editors: S. Murala and S.K. Vipparthi, *Multimedia Retrieval - Perspectives and Challenges*

DOI: 10.15579/gcsr.vol4.ch3, GCSR Vol. 4, pp. 43-66, 2015

model (MAM), naive topic correlation model (NTC), and semantic topic correlation model (TCM) and their correspondence mapping techniques, particularly semantic topic correlation mapping. An extension of TCM applied to retrieve information in Chinese language is also introduced in this chapter.

3.1 Introduction

Various online multimedia information has been increasing explosively in disparate modalities, such as image, text, audio and video. Effective and efficient information search techniques are desirable to access such massive and multi-modal data. However, the predominant multimedia search engines today are still using keywords as queries, which have obvious shortcomings. For instance, a concept can be more accurately described by using more words. However, for a search engine, a query with a few keywords reflecting the main information of the concept you are looking for is more desired than detailed descriptions. Using long sentences, although it is semantically delicate yet redundant for a search engine. Another shortcoming is that successful search engines for large-scale information require that the latter be augmented with annotations provided by human annotators. But 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 the above problems. One popular research area is content-based image retrieval (CBIR) [31, 48], which aims to retrieve relevant images given an image query based on the similarity of visual features. However, the retrieval results exist serious "semantic gap" between the low level image features and human's high level semantic concepts. Another research area focuses on automatic image annotation [7, 10, 25]. The goal of automatic image annotation is to automatically assign relevant text to an out-of-sample image. But the annotations are usually in form of a few keywords which are limited to accurately describe the semantic content of an image. How to build a joint model to capture the relations between disparate modalities while supporting access to the content in each individual modality is essential for further progress in the multimedia research area.

In this chapter, we introduce a cross-modal information retrieval model, named as semantic topic correlation modal (TCM), by mapping different modalities, into a semantic space in order to find their semantic topic correlations in the space. The topic correlations between different modalities are measured by an effective statistical distance obtained from a hierarchical representation of the raw data. Although this model can be applied to any combination of modalities, we restrict our discussion to documents containing images and texts. Moreover, we also extend the TCM for cross-modal retrieval tasks on Chinese corpora since most of previous research only focused on English corpora. A text in English can be regarded as a collection of words, which are the basic semantic units in the majority of western languages. So techniques used in studying English can be easily extended to other alphabetic languages. However, in Chinese language, both characters and words play important roles in accurate expression. In this chapter, an extension of the TCM is studied to retrieve information in Chinese language by analyzing the influence of both characters and words serving as the basic semantic units.

The rest of this chapter is structured as follows: Section 3.2 briefly introduces the related work in three categories, including uni-modal information retrieval, multi-modal information retrieval, and cross-model information retrieval. In Section 3.3, we describe the definition of cross-modal information retrieval problem and the hierarchical image and textual representations for semantic content modeling. Cross-modal retrieval models based on manifold alignment is introduced in Section 3.4. The naive model and semantic model based on topic correlations are presented in Section 3.5 and Section 3.6, respectively. To verify the effectiveness of the newly proposed semantic topic correlation model, two applications are presented on three benchmark datasets, including both English and Chinese datasets. The retrieval performance are analyzed and compared to previous work in Section 3.7. The conclusions and future work are given in Section 3.8.

3.2 Related Work

3.2.1 Uni-Modal Information Retrieval

Multimedia information retrieval, mainly focusing on the text modality and image modality, has been well studied in both multimedia and information retrieval areas. The predominant research in these areas are focused on uni-modal approaches, which only consider single modality for both query and retrieved data. For instance, in [36], a query text is given to retrieve relevant text documents based on bag-of-words features. While in content-based image retrieval area [12, 31, 48], an image is used as a query to retrieve similar images by matching their visual features. However, uni-modal approaches are not always effective. They limit the applicability of multimedia retrieval models to retrieving the same modality with the query, without answers from other rich modalities. Moreover, the "semantic gap" between low level image features and human's high level semantic concepts severely influences the performance of the image retrieval models [38].

Currently, most of the image retrieval systems in the famous search engines, such as Google, Yahoo, etc, provide image retrieval functions by textual queries. These systems, in fact, augment the image collections with related textual descriptions. The textual descriptions are typically a few keywords or short captions provided by manual annotation or automatic image annotation [7, 10, 25]. When these textual metadata is available, the query text is matched with the textual metadata describing images in the collections during the retrieval procedure. Thus the retrieval procedure is actually uni-modal regardless of the image data and the retrieval performance is highly depended on the accuracy of image annotation. However, both manual annotation and automatic annotation have obvious problems that a few annotated keywords cannot describe an image accurately especially for scene images. It is also possible that users and annotators use different keywords to describe the same image. These problems greatly limit the effectiveness of the systems. Thus, uni-modal search techniques are not enough to meet the requirements of information 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.

3.2.2 Multi-Modal Information Retrieval

Moving beyond uni-modal retrieval, various models have been proposed to study multi-modal information retrieval [18, 23, 24, 31, 42]. The multi-modal retrieval models are generally extended from uni-modal ones since they combine information from multiple modalities in different retrieval procedures of common retrieval models. There are typically two kinds of combination strategies: (1) Combination of low-level features from different modalities into concise multi-modal features. In [24], a manifold learning algorithm based on Laplacian Eigenmaps is introduced to combine low-level descriptors of each separate modality and map them to a common low-dimensional multi-modal feature space. In such feature space, semantically similar multi-modal data are represented by multi-modal vectors close to each other. (2) Combination of independent systems at different levels. For example, Kliegr et al. [23] utilized a combination of two independent systems at the output level, so that one system models the text data stream while the other models the image data stream. Iyengar et al. [18] 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. Similar research with combinations at other levels was presented in [31] and [42], which [13] gives a good overview of these models and also introduces combination of multi-modal and uni-modal retrieval models.

In general, multi-modal retrieval models are extensions of the uni-modal ones, which support retrieving more than one modality simultaneously but require queries having the same modalities with the retrieved data. Users cannot access each data modality individually, which limits the models' applicability.

3.2.3 Cross-Modal Information Retrieval

In recent years, progress in cross-modal information retrieval has overcome the limitation of both uni-modal and multi-modal retrieval models. In the literature, various modalities have been studied, including images and texts [7], texts and audio [37], images and audio [49], or texts, images, and audio [46, 45]. One kind of popular 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 previous work of Blei et al. [7, 8] introduces a correspondence latent Dirichlet allocation (Corr-LDA) to model the images and associated annotations within a shared mixture of latent factors. [7] also proposed a multi-modal LDA (mmLDA) to compute a mean topic distribution topic as the shared variable between different modalities. Other models like [19, 32, 40] pay attention to either distance between different multimedia documents or optimizing the likelihood of the topic model.

Recently, some attempts bring new perspective for solving the cross-modal retrieval problem. Yang et al. constructed a Multimedia Correlation Space (MMCS), where each multimedia document (including text, images and audio) is represented as a point, based on the the original content and the correspondence between the heterogeneous data. Then a novel ranking algorithm is used, which adopts a local linear regression model for each point and aligns 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-sample, unless the relevance feedback is applied to the ranking procedure. Mahadevan et al. [28] 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. [29] 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. [35] 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 uni-modal retrieval approaches.

Inspired by [35], we develop a new model for cross-modal retrieval, which is more simple and effective compared with [35]. Different from their work, our cross-modal multimedia retrieval system is 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 jointly modeling different modalities by considering category information.

3.3 Cross-Modal Problem Definition and Modality Representation

3.3.1 Cross-Modal Problem Definition

In multimedia information retrieval, documents generally contain multiple forms of contents. We consider the retrieval tasks from a dataset containing documents of image and text components. The retrieval problem is to search semantically matched texts by an image query and vice versa. Formally, given a set of documents denoted as $\mathbf{D} = [D_1, D_2, ..., D_K]$, we assume that each document D_k, $k = 1, ..., K$, contains at least an image and associated text. In fact, there can be multiple texts accompanied with more than one image or none image and vice versa. In this chapter, we only consider a simplified case of an one-to-one mapping between image and text as shown in Fig.(3.1), which can be defined as:

$$D_k = [I_k, TX_k],\ k = 1, ..., K \tag{3.1}$$

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

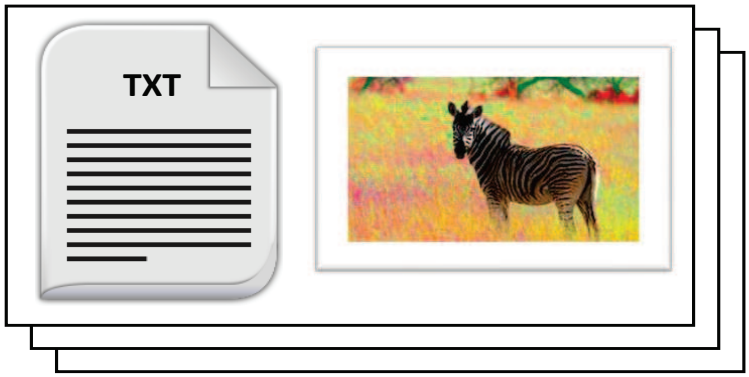

Figure 3.1: Definition of multimedia documents in our model. Each document contains a text and its corresponding image.

an image I_k given a text query TX_q by

$$S(I_k) = P\left(I_k|TX_q\right) \tag{3.2}$$

Similarly, the score function for TX_k is

$$S(TX_k) = P\left(TX_k|I_q\right) \tag{3.3}$$

Eq. (3.2) and (3.3) are used to arrange the retrieval results in a descending order given a query image or text.

3.3.2 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 [48], and latent topics in topic models [52, 51], attract much attention for their effectiveness in semantic modeling. In the following, we will introduce these mid-level representations since the cross-modal retrieval models introduced in Section 3.4, Section 3.5, and Section 3.6 are based on these representations for further investigation of the correlations between different modalities.

3.3.2.1 Image Representation

The desired representation for images should be robust with small changes, such as illumination, scale, 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 [15]. Previous research has shown that the BoF model is robust in object and scene classification [20], image search [48], and video retrieval tasks [14]. 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.

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) [26] feature has been proven to be powerful descriptors with respect to different geometrical variations, e.g. translation, scale, rotation, and small distortions. Keypoints are firstly searched in the difference of Gaussian (DoG) space [9] and then gradients are computed 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.

Codebook Generation Following the local feature extraction procedure, we utilize a k-means clustering to generate a codebook. The local descriptors $\mathbf{d} = [d_1, d_2, ...d_n]$ are divided into k clusters $\mathbf{C} = [C_1, C_2, ...C_k]$. 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:

$$J = \arg\min_{\{v_i\}} \sum_{i=1}^{k} \sum_{d_j \in C_i} \|d_j - v_i\| \tag{3.4}$$

where

$$v_i = \frac{1}{|C_i|} \sum_{d_j \in C_i} d_j \tag{3.5}$$

Feature Quantization For each image, we use the "hard assignment" method [21] 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.

3.3.2.2 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 [17], 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) [8] 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 $\mathbf{w}$ in a corpus:

1. Choose $\theta \sim Dir(\alpha)$.
2. For each of the N words w_n:

   i) Choose a topic $z_n \sim Multinomial(\alpha)$.

   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:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.6}$$

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

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) d\theta \tag{3.7}$$

The LDA has been widely used in different NLP tasks, such as information retrieval [41], text classification [39], and question answering system [33]. 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 [3]. Figure 3.2 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 chapter, 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.

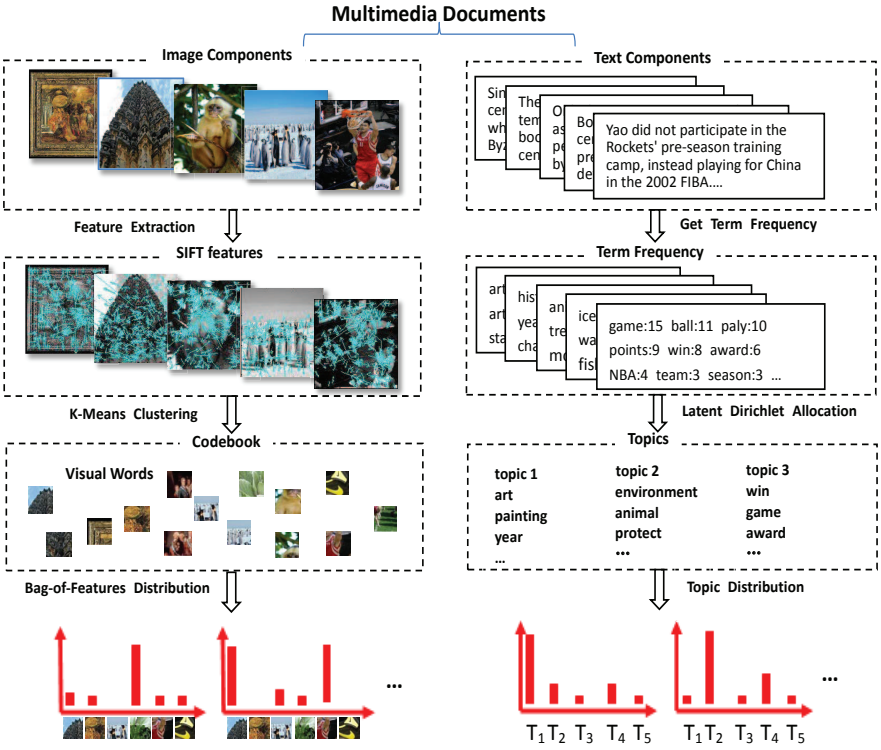

Figure 3.2: Topic representations of texts and images in given multimedia documents. On the left-hand side, representations of the image components are based on the bag-of-features model. SIFT features are extracted on all the training images and a codebook is learned from these features. Each image 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.

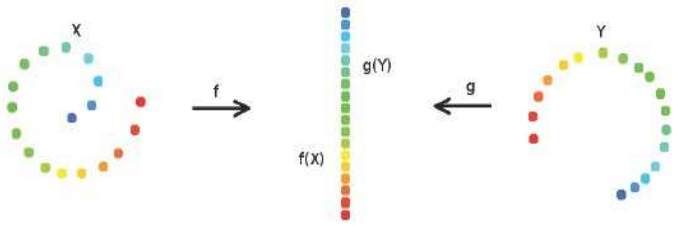

Figure 3.3: Schematic illustration of manifold alignment. Two datasets X and Y containing different modalities are mapped into a common space, where the corresponding instances from different datasets are equal, while the local structural similarities are consistent with the original datasets.

3.4 Manifold Alignment Model

One popular kind of models for cross-modal multimedia information retrieval is based on manifold alignment techniques [27, 29, 45, 46, 49, 55], which is also widely used in automatic machine translation, cross-lingual information retrieval, image interpretation, and social network analysis. Manifold alignment is a dimensionality reduction technique that maps datasets from two or more disparate modalities to a common lower-dimensional space by aligning their underlying manifold. After manifold alignment, data from disparate modalities will have a unifying representation in a common space that keeps the structural similarities of each original dataset and preserve the correspondence between disparate datasets (see Fig.(3.3)). Retrieval is then conducted by computing the nearest document on the manifold space to a multimedia query.

Figure 3.3 shows the illustration of the problem of manifold alignment. From the perspective of cross-modal retrieval, X and Y can respectively represent the feature space of the image dataset and the text dataset, whose instances lie on the same manifold Z. Manifold alignment aims to find two mapping functions f and g so that $f(x_i)$ and $f(y_j)$ are close to each other in terms of Euclidean distance if x_i and y_j are close to each other in manifold space. Image features in X can be represented by a $n \times p$ matrix containing n samples in p-dimensional feature space. Textual features in Y can be represented by a $m \times q$ matrix containing m samples in q-dimensional feature space. $f : \mathbb{R}^p \rightarrow \mathbb{R}^k$, and $g : \mathbb{R}^q \rightarrow \mathbb{R}^k$, where k is the dimension of the joint latent space.

Manifold alignment algorithms can be categorized into three types, supervised, semi-supervised, and unsupervised. If the correspondence information is completely available, the algorithm is supervised and corresponding samples from different datasets will be mapped to a unifying point in the joint latent space. If the correspondence information is incomplete, the algorithm is semi-supervised and the alignment depends on both the incomplete correspondence and the datasets' inner structure. Given no correspondence information, the algorithm is unsupervised and external corresponding information has to be learned from the data. [27] gives a good introduction on manifold alignment and alignment approached based on one-step alignment and two-step

alignment.

However, cross-modal retrieval models based on manifold alignment lacks out-of-sample generalization. As far as we know, there is no effective method to map the query into the manifold space and compute the similarity with documents in the datasets directly. All the queries have to be mapped to the nearest neighbors from the dataset of the same modality, which is depart from the intrinsic purpose of cross-modal retrieval.

3.5 Naive Topic Correlation Model

For a specified document, though its contents may be in different modalities, the underlying semantic contents are similar or even identical. In this model, 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 $\mathbf{V} = [V_1, V_2, ..., V_M]$ denote a set of visual words in the codebook (M is the codebook size), and $\mathbf{T} = [T_1, T_2, ..., T_N]$ 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 $\mathbf{D} = [D_1, D_2, ..., D_K]$:

$$P(V_i|T_j) = \sum_{k=1}^{K} P(V_i|I_k)P(I_k|TX_k)P(TX_k|T_j) \tag{3.8}$$

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

$$P(TX_k|T_j) = \frac{P(T_j|TX_k)P(TX_k)}{\sum_{k=1}^{K} P(T_j|TX_k)P(TX_k)} \tag{3.9}$$

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,

$$P(TX_k) = P(I_k) = \frac{1}{K} \tag{3.10}$$

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

$$P(T_j|V_i) = \sum_k P(T_j|TX_k)P(TX_k|I_k)P(I_k|V_i) \tag{3.11}$$

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:

$$P(I_k|V_i) = \frac{P(V_i|I_k)P(I_k)}{\sum_{k=1}^{K} P(V_i|I_k)P(I_k)} \tag{3.12}$$

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:

$$P(I_k|TX_q) = \sum_i \sum_j P(I_k|V_i)P(V_i|T_j)P(T_j|TX_q) \tag{3.13}$$

$$P(TX_k|I_q) = \sum_i \sum_j P(TX_k|T_j)P(T_j|V_i)P(V_i|I_q) \tag{3.14}$$

However, such correlations do not take into account of 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 [47], 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.

3.6 Semantic Topic Correlation Model

Instead of directly mining the correlation between mid-level features of texts and images, semantic topic correlation model represents 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 one document will be labeled by the same concept. In this 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 $\mathbf{C} = [C_1, C_2, ..., C_n]$. These posterior vectors are existed in a new space called 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 3.4 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) [4]. This builds multiple 'one-versus-one'

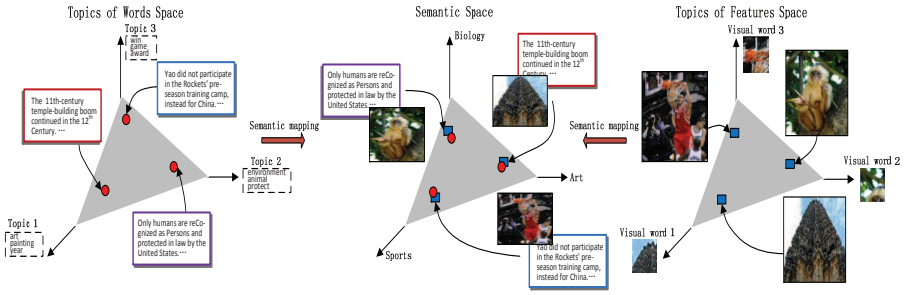

Figure 3.4: Schematic illustration of the semantic mapping for the TCM model. (Left) Semantic mapping of the text components from corresponding topics of words space, learned by LDA, to the semantic space, learned by the multi-class classifier for texts. (Right) Semantic mapping of the image components from associated topics of features space, learned by BoF, to the semantic space, learned by the multi-class classifier for images.

binary classifiers and each binary classification is considered to be a voting casting for one category [11]. 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 probability of a retrieved image given a text query or vice versa in the retrieval procedure. Given a text (image) query, 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:

$$P(I_k|TX_q) = \sum_i P(I_k|C_i)P(C_i|TX_q) \tag{3.15}$$

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. Given a query text TX_q, the value of score function $S(I_k)$ for a retrieved image I_k in the dataset is assigned by the value 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:

$$P(TX_k|I_q) = \sum_i P(TX_k|C_i)P(C_i|I_q) \tag{3.16}$$

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

$$P(TX_k|C_i) = \frac{P(C_i|TX_k)P(TX_k)}{\sum_k P(C_i|TX_k)P(TX_k)} \tag{3.17}$$

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.

Given a query imageI_q, the value of score function $S(TX_k)$ for a retrieved text TX_k in the dataset is assigned by the value of $P(TX_k|I_q)$, which is used to rank the retrieved texts in a descending order.

3.7 Application Examples

To evaluate the effectiveness of some popular cross-modal retrieval models, we applied some representative models on three datasets, including English Wikipedia, TVGraz, and Chinese Wikipedia. The retrieval results are compared 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) [54] is adopted to measure the retrieval performance.

3.7.1 Dataset Description

Three benchmark datasets are tested to evaluate the retrieval performance of some representative cross-modal retrieval models.

English Wikipedia

The English Wikipedia corpus [1] is a collection of "Wikipedia featured articles", which has been first used in [35]. 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 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. Three sample images for each category are shown in Fig.(3.5). 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 be appeared 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.

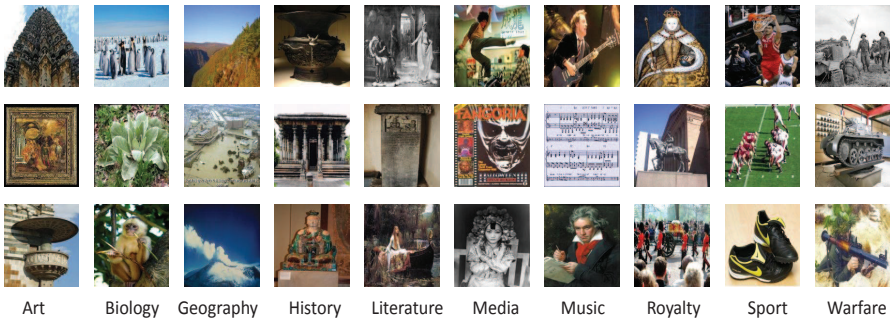

Figure 3.5: Samples of the ten categories of the En-Wikipedia dataset.

TVGraz

The TVGraz dataset [5] is a collection of webpages including images and texts [22]. It contains the top 1000 results from Google image search for each of 10 categories from the Caltech-256 [16]. 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. Figure 3.6 shows three sample images for each category.

For the 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 [2] 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".

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 [6]. It consists of 3103 documents of paired texts and images from 9 categories. Three sample images for each category are shown in Fig.(3.7). 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 documents are discarded. The raw tests are pre-processed by removing punctuation as well as numbers.

Figure 3.6: Samples of the ten categories of the TVGraz dataset.

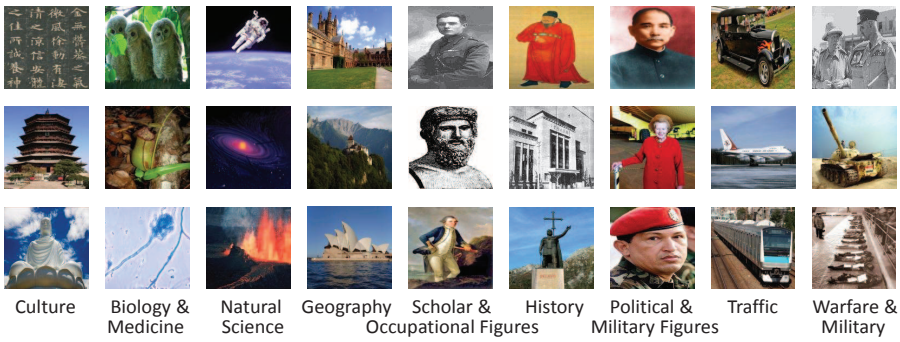

Figure 3.7: Samples of the nine categories of the Ch-Wikipedia dataset.

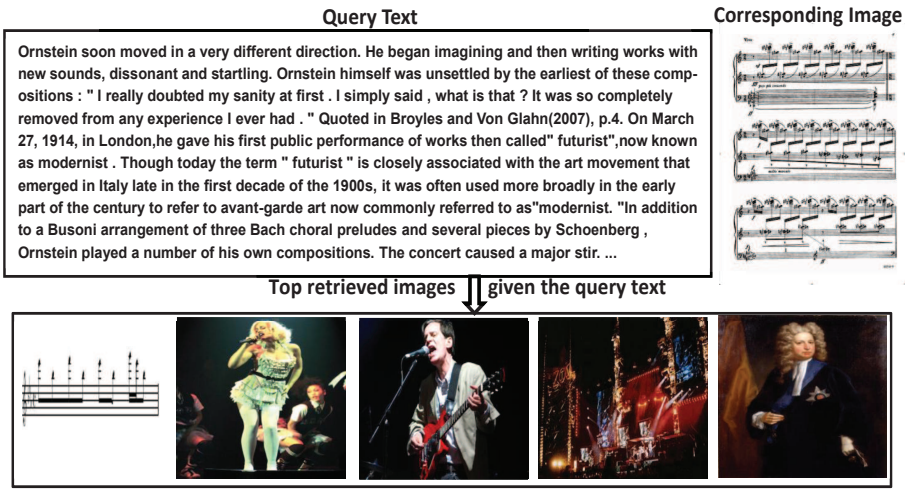

Figure 3.8: Retrial samples of TCM for text query task on the En-Wikipedia dataset. Given the text query on the top left, the top five retrieved images are shown on the bottom. We also show the image corresponding to the text query to have a clear semantic comparison with the retrieval results.

3.7.2 Retrieval on English Text and Image Datasets

We first conduct the retrieval experiments on the En-Wikipedia dataset. We compare our TCM with the semantic correlation matching (SCM) model [35], Fast version of Maximum Covariance Unfolding (Fast-MCU) [28], and Parallel Field Alignment Retrieval (PFAR) [29]. Since Fast-MCU and PFAR models published only the 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 [35].

For visual examination, two examples for retrieval task given a query test or image are displayed in Fig.(3.8) and Fig.(3.9), respectively. In Fig.(3.8), the text query is a paragraph related to "music". The corresponding image (shown on the top right of Fig.(3.8)) is 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 3.9 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 3.1. The baseline is computed on the random retrieval results [35]. 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 Fast-MCU model in both image and text queries and is comparable with PFAR model in image queries. Since only SCM published the MAP

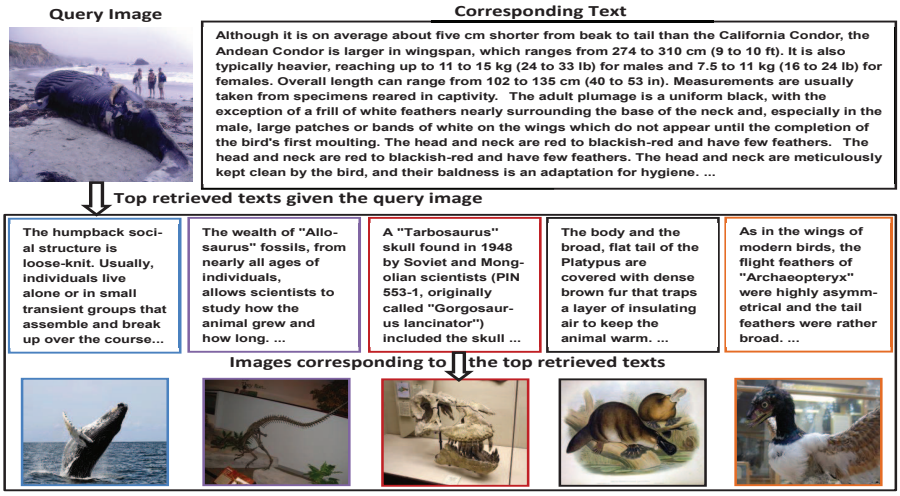

Figure 3.9: Retrieval samples of TCM for image query task on the En-Wikipedia dataset. Given the image query on the top left, the top five retrieved texts are shown on the bottom. For clear semantic comparisons, we also show the text corresponding to the image query and images corresponding to the retrieved texts.

scores of each category, we compare the histograms of our TCM, SCM and the random case.

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 [34]. The MAP values shown in Table 3.1 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.

3.7.3 Retrieval on the Chinese Text and Image Dataset

In this section, we evaluate the new model on Chinese multimedia datasets. Topic modeling of Chinese language has been well studied [43, 50]. Most of the previous studies choose Chinese words as the most basic units of the language [30, 53]. However, 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 [44] and verified by computational evidence [52, 51] concluded that character-based topic models outperform the word-based topic models in the text classification tasks.

In practice, both the characters and words serve as indispensable parts for Chinese language. For example, character $b\bar{a}o$ means bag. By combining with the character *qián* (money), it becomes the word *qián* $b\bar{a}o$, which means wallet. By combining with

Table 3.1: Result comparisons using MAP measure on the English datasets.

Model	Image query	Text query	Average	Dataset
Random [35]	.118	.118	.118	En-Wikipedia
SCM [35]	.277	.226	.252	
Fast - MCU [28]	.287	.224	.256	
PFAR [29]	.298	.273	.286	
TCM	**.293**	**.232**	**.266**	
Random [34]	.119	.119	.119	TVGraz
SCM [34]	.693	.696	.694	
TCM	**.694**	**.706**	**.700**	

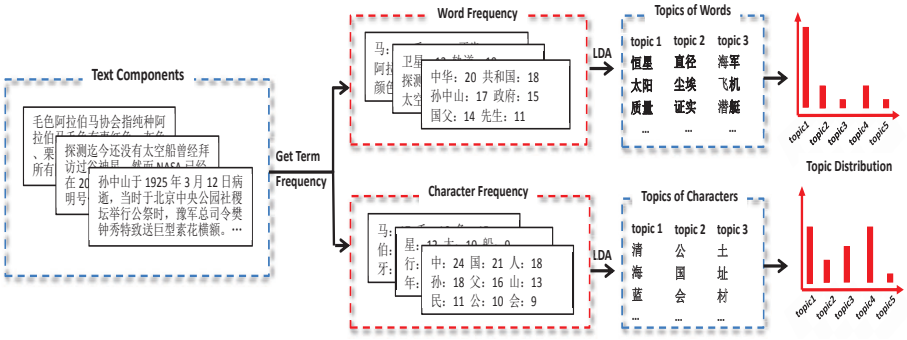

Figure 3.10: Chinese representation based on latent Dirichlet allocation [8].

the character $sh\bar{u}$ (book), it becomes the word $sh\bar{u}\ b\bar{a}o$ that means schoolbag. By combining with the character *pí* (leather), it becomes the word *pí* $b\bar{a}o$ (briefcase). Each Chinese character carries ambiguous semantic meaning. By forming a word, the semantic meaning is refined. In this chapter, we conduct comprehensive experiments to validate the new model on a Chinese dataset by using topic models based on both words and characters. We refer these two kinds of topic models as the word-based topic model and character-based topic model, respectively.Figure 3.10 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 appear in over 50% of the documents [52]. After the preprocessing procedure, we obtain 21240 unique Chinese words and 3419 unique Chinese characters.

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

Table 3.2: 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**	

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 Table3.2. 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 3.2, 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 [52].

Detailed experiments have also been carried out to compare the histogram 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. It is consistent with the results that were reported in [52], suggesting better retrieval performance can be achieved by using the character-based model for Chinese corpus.

3.8 Conclusions and Future Work

Cross-modal multimedia information retrieval aims to jointly model data from disparate modalities and support access to each individual modality when given a query from one arbitrary modality. In this chapter, a topic correlation model (TCM) for cross-modal multimedia retrieval shows relatively good performance compared with some representative state-of-the-art approaches. The statistical relations between mid-level features from different modalities were investigated. 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 how to filter irrelevant texts which have no corresponding images, or vice versa. This noise control process may significantly improve the quality of retrieval.

Acknowledgements

This work is supported by the National Science Foundation of China No. 61305047.

References

[1] http://www.svcl.ucsd.edu/projects/crossmodal/.
[2] http://www.textfixer.com/resources/common-english-words.txt.
[3] http://www.cs.princeton.edu/~blei/lda-c/.
[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[5] http://www.icg.tugraz.at/Members/kahn/TVGraz_dataset.tar.gz/view.
[6] http://icmll.buaa.edu.cn/zh_wikipedia.
[7] D.M. Blei and M.I. Jordan. Modeling annotated data. In *26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, 2003.
[8] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[9] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.
[10] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.
[11] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):Article No. 27, 2011.
[12] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):Article No. 5, 2008.
[13] H.J. Escalante, C.A. Hernadez, L.E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *1st ACM International Conference on Multimedia Information Retrieval (MIR)*, pages 172–179, 2008.
[14] J. Fehr, A. Streicher, and H. Burkhardt. A bag of features approach for 3D shape retrieval. In *5th International Symposium on Visual Computing (ISVC)*, pages 34–43, 2009.

[15] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:524–531, 2005.

[16] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology (Caltech), 2007.

[17] Z. Harris. Distributional structure. *Word*, 10(2):146–162, 1954.

[18] G. Iyengar, P. Duygulu, and S. Feng et al. Joint visual-text modeling for automatic retrieval of multimedia documents. In *13th Annual ACM International Conference on Multimedia (MULTIMEDIA)*, pages 21–30, 2005.

[19] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2407–2414, 2011.

[20] Y.G. Jiang, C.W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *6th ACM International Conference on Image and Video Retrieval (CIVR)*, pages 494–501, 2007.

[21] M. Kearns, Y. Mansour, and A.Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 282–293, 1997.

[22] I. Khan, A. Saffari, and H. Bischof. TVGraz: multi-modal learning of object categories by combining textual and visual features. In *33rd Workshop of the Austrian Association for Pattern Recognition*, pages 213–224, 2009.

[23] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. Combining image captions and visual analysis for image concept classification. In *9th International Workshop on Multimedia Data Mining (MDM)*, pages 8–17, 2008.

[24] M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras. Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing: Image Communication*, 28(4):351–367, 2013.

[25] J. Li and J.Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[26] D.G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, volume 2, page 1150, 1999.

[27] Y. Ma and Y. Fu. *Manifold learning theory and applications*. CRC Press, 2011.

[28] V. Mahadevan, C.W. Wong, J.C. Pereira, T.T. Liu, N. Vasconcelos, and L.K. Saul. Maximum covariance unfolding: manifold learning for bimodal data. In *Advances in Neural Information Processing Systems*, pages 918–926, 2011.

[29] X. Mao, B. Lin, D. Cai, X. He, and J. Pei. Parallel field alignment for cross media retrieval. In *21st ACM International Conference on Multimedia (MM)*, pages 897–906, 2013.

[30] X. Ni, J. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *International Conference on World Wide Web*, pages 1155–1156, 2009.

[31] T.T. Pham, N.E. Maillot, J.H. Lim, and J.P. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *6th ACM Conference on Information and Knowledge Management (CIKM)*, pages 439–444, 2007.

[32] D. Putthividhy, H.T. Attias, and S.S. Nagarajan. Topic regression multi-modal latent Dirichlet allocation for image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3415, 2010.

[33] Z. Qin, M. Thint, and Z. Huang. Ranking answers by hierarchical topic models. In *22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)*, pages 103–112, 2009.

[34] N. Rasiwasia. *Semantic image representation for visual recognition*. PhD thesis, University of California, 2011.

[35] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *18th ACM International Conference on Multimedia (MM)*, pages 251–260, 2010.

[36] G. Salton. *The SMART retrieval system - experiments in automatic document processing*. Prentice-Hall, NJ, USA, 1971.

[37] M. Slaney. Semantic-audio retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages IV–4108–IV–4111, 2002.

[38] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[39] M. Steyvers and T. Griffiths. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models, pages 55–65. 2007.

[40] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In *28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 843–851, 2012.

[41] X. Wei and W.B. Croft. LDA-based document models for ad-hoc retrieval. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, 2006.

[42] T. Westerveld. Probabilistic multimedia retrieval. In *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 437–438, 2002.

[43] Y. Wu, Y. Ding, X. Wang, and J. Xu. A comparative study of topic models for topic clustering of Chinese web news. In *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, volume 5, pages 236–240, 2010.

[44] T. Xu. Fundamental structural principles of Chinese semantic syntax in terms of Chinese characters. *Applied Linguistics*, 1:3–13, 2001.

[45] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *17th ACM International Conference on Multimedia (MM)*, pages 175–184, 2009.

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

[47] J. Yu, Y. Cong, Z. Qin, and T. Wan. Cross-modal topic correlations for multimedia retrieval. In *21st International Conference on Pattern Recognition (ICPR)*, pages 246–249, 2012.

[48] X. Yuan, J. Yu, Z. Qin, and T. Wan. A SIFT-LBP image retrieval model based on bag-of-features. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 1061–1064, 2011.

[49] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clus-

tering on image-audio dataset. In *15th International Conference on Multimedia (MULTIMEDIA)*, pages 273–276, 2007.

[50] Y. Zhang and Z. Qin. A topic model of observing Chinese characters. In *2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 7–10, 2010.

[51] Q. Zhao, Z. Qin, and T. Wan. Topic modeling of Chinese language using character-word relations. In *18th International Conference on Neural Information Processing (ICONIP)*, pages 139–147, 2011.

[52] Q. Zhao, Z. Qin, and T. Wan. What is the basic semantic unit of Chinese language? In *12th Biennial Conference on The Mathematics of Language*, pages 143–157, 2011.

[53] G. Zhengxian and Z. Guodong. Employing topic modeling for statistical machine translation. In *IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, volume 4, pages 24–28, 2011.

[54] M. Zhu. Recall, precision and average precision. Technical report, University of Waterloo, 2004.

[55] Y.T. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.