Multimedia Systems (2015) 21:525–539
DOI 10.1007/s00530-014-0397-6

REGULAR PAPER

Analyzing semantic correlation for cross-modal retrieval

Liang Xie · Peng Pan · Yansheng Lu

Received: 11 September 2013 / Accepted: 2 June 2014 / Published online: 25 June 2014

Abstract With the development of multimedia technology, effective cross-modal retrieval methods are increasingly demanded. The key point of cross-modal retrieval is analyzing the correlation of heterogeneous modalities. There are mainly two types of correlation: content correlation and semantic correlation. Semantic correlation is constructed at a high level of abstraction which is more close to the human understanding than content correlation. In this paper, we investigate a semantic model to construct the semantic correlation for cross-modal retrieval. We assume that the semantic correlation of multimedia data from different modalities can be conditionally generated by semantic concepts in a probabilistic generation framework. The cross-modal semantic generation model (CMSGM) is proposed based on this assumption. We consider three cases of the cross-modal retrieval task. The first is the ideal case that all manifest concepts exist in training data for constructing the correlation, and we propose manifest CMSGM (M-CMSGM) which directly uses CMSGM on the manifest semantic concepts for retrieval. The second is the case that there are no manifest concepts in training data, and latent CMSGM (L-CMSGM) based on latent semantic concepts is proposed for this case, where the latent semantic concepts are learned by asymmetric spectral clustering. The last is the most general case that some of the manifest concepts exist, and we combine M-CMSGM and L-CMSGM to get combinative CMSGM (C-CMSGM) to solve this case. Experimental results on Wikipedia featured articles and MIR Flickr show that our methods have better performance compared with previous state-of-the-art methods. And C-CMSGM can maintain good performance in the case that manifest concepts are lacking, which confirms the robustness and practicality of C-CMSGM.

Keywords Cross-modal retrieval · Semantic correlation · Generation model · Semantic concept

1 Introduction

In recent years, there has been a rapid development of multimedia content on the web. A variety of media data such as image, text, audio and video have become easy to be accessed and delivered. However, most of existing retrieval methods focus on unimodal data, such as content-based image retrieval and text-based retrieval, they cannot deal with multiple media modalities. With the development of multimedia technology, to facilitate the management of a variety of multimedia content, effective cross-modal retrieval methods are increasingly demanded.

In cross-modal retrieval, the modality of query data is different from the data to be retrieved. In this paper, we refer to the modality as the media type, thus image, text, and audio are different modalities. Cross-modal retrieval is different from multi-modal retrieval which has been studied by many works [1–3]. Although they both deal with multi-modal data, they are two different retrieval problems. In multi-modal retrieval, both query and retrieved data are multi-modal, and they always contain the same modalities.

Communicated by L. Zhang.

L. Xie · P. Pan (✉) · Y. Lu
School of Computer Science and Technology, Huazhong
University of Science and Technology, Wuhan 430074, China
e-mail: panpeng@mail.hust.edu.cn

L. Xie
e-mail: whutxl@hotmail.com

Y. Lu
e-mail: lys@mail.hust.edu.cn

The main purpose of multi-modal retrieval methods is to improve the performance of unimodal retrieval; texts are always utilized to enhance the traditional content-based image retrieval. Multi-modal retrieval methods usually fuse different modalities for retrieval rather than correlate them [39]. Unlike multi-modal retrieval, in cross-modal retrieval query data and retrieved data do not share the same modality, cross-modal retrieval methods have to analyze the correlation of different modalities. For example, if a user wants to use an image of tiger to retrieve the text description about tiger, then the cross-modal methods can work for him/her. General multi-modal methods cannot meet this requirement, because they cannot establish the correlation between the image and text documents.

The key point of cross-modal retrieval is analyzing the correlation of heterogeneous modalities. There are mainly two types of correlation to be analyzed. One is content correlation, which is the correlation of content features from different modalities. The content correlation is generally obtained according to the concurrence of different modalities. Different modal data occurring in a document, a web page, etc., are seen as correlated, and their correlation of content features can be analyzed by content-based cross-modal methods. Canonical correlation analysis (CCA) [4] is a typical method which is used to analyze the content correlation of multiple modalities. CCA learns two subspaces from two content features of two different modalities, and these subspaces maximize the correlation between the two modalities. The advantage of content-based cross-modal methods is that content correlation can be easily obtained from the multi-modal documents and multi-modal web pages, thus they are well suited to most of the cross-modal retrieval problems. The other is the semantic correlation. It is obvious that all the modalities are correlated by the semantic concepts. Different modalities can share the same semantic concepts, for example, the textual description of a bird, the image of the bird and the sound of the bird are all about the bird, and they are correlated according to the concept "bird".

The semantic-based cross-modal methods which analyze the semantic correlation can better correlate heterogeneous modalities, the reason is that they construct the correlation at a high level of abstraction which is more close to the human understanding. However, semantic correlation is learned from training data which is labeled by semantic concepts and hard to be collected. General semantic concepts include scenes (e.g., indoor, outdoor, landscape, etc.), objects (animal, people, car, etc.), or events (travel, work, etc.), we refer to them as manifest semantic concepts which can be easily understood by human. Due to the semantic gap between these concepts and the feature contents of multimedia data, training data with semantic information are required for learning true semantic concepts. Without the semantic knowledge in the training data, it is impossible to make the machine understand the manifest semantic concepts. Manually labeling manifest semantic concepts for training data is a traditional approach to construct the training data, but the cost is expensive [5]. On the other hand, latent semantic concepts can be automatically learned by machine, they are not similar to the manifest semantic concepts and may be more difficult to be understood by human. A good automatical learning method can learn latent concepts which are close to the manifest concepts, latent concepts can also be used to construct the correlation between the multi-modal data at a higher level of abstraction than content correlation.

In this paper, we propose the cross-modal semantic generation model (CMSGM) for cross-modal retrieval. CMSGM describes the semantic correlation of heterogeneous multimedia data in a probabilistic generative framework. The core idea of CMSGM is that multimedia data from different modalities share the same semantic concepts, they can be generated by these concepts, and their generation processes are conditional independent. In the estimation of CMSGM, unlike previous generative approaches which directly use an existing distribution such as Gaussian for the generation probabilities, we use an indirect discriminative approach to estimate the generation probabilities. At last, effective cross-modal similarities are inferred from the generation model to measure the semantic correlation. We consider three cases to use the CMSGM for cross-modal retrieval. For the ideal case where the training data contain complete manifest semantic information, we construct CMSGM on the manifest semantic concepts, and propose manifest CMSGM (M-CMSGM) for cross-modal retrieval. For the case that the training data do not contain any manifest semantic information, we propose asymmetric spectral clustering (ASC) which is an extension of spectral clustering to automatically learn latent semantic concepts from training data, and CMSGM is constructed on the latent semantic concepts, finally we get latent CMSGM (L-CMSGM) for cross-modal retrieval. For more general case where training data may contain incomplete semantic information, we combine M-CMSGM and L-CMSGM to get combinative CMSGM (C-CMSGM) for retrieval. The C-CMSGM takes the advantages of the two methods and avoids their disadvantages. Experimental results show the effectiveness of our CMSGM model, and C-CMSGM is shown to be practicable when we cannot know whether the manifest concepts is lacking.

This paper is organized as follows. We discuss related work in Sect. 2. In Sect. 3, we propose the ASC for learning the latent semantic concepts. In Sect. 4, we describe our CMSGM; and the cross-modal similarities for retrieval are derived from CMSGM. In Sect. 5, we propose three cross-modal retrieval methods based on CMSGM for three different cases. Section 6 shows experimental results of our methods based on CMSGM and compares it to other methods, as well as the performance of our three methods in different cases. Finally, we conclude in the last section.

2 Related work

There are many works on cross-modal problems. Automatic image annotation is a typical cross-modal problem; the purpose of it is solving the problem of keyword-based image retrieval which belongs to the problem of cross-modal retrieval. The main goal of image annotation is allocating appropriate textual words to an image by machine automatically. There are many works on the image annotation task. Most of them can be divided into three groups [6]. The first is the group of generative models; they usually analyze the correlation of images and text words which is based on latent variables, such as images in the training data, semantic topics, etc. Cross-media relevance model [7] and multiple Bernoulli relevance model [8] both choose the images in the training data as latent variables, and the correlation of images and words is analyzed on these variables. Semantic topics which can be seen as the latent semantic concepts in our paper are also widely used for image annotation. CORR-LDA [9] is the extension of the Latent Dirichlet Allocation (LDA) [10], it correlates words and images by semantic topics which are learned by LDA. Asymmetric probabilistic latent semantic analysis (PLSA) [11] first learns semantic topics from text words using PLSA, then folding-in method [12] is used to learn the relation between topics and images, at last images and words are correlated according to these topics. The idea of asymmetric learning is also adopted by our ASC. The second is the group of discriminative models, these methods learn a separate classifier for each word, and use them to predict whether the test image belongs to the class of images that are annotated with each particular word [13, 14]. The last is nearest neighbor-based methods; in fact they are special discriminative methods. The representative methods: JEC [15] and TagProp [16] use the nearest neighbor method to solve the classify problem which is similar to discriminative methods. The successes of generative and discriminative models on image annotations demonstrate their excellent ability in analyzing the correlation of heterogeneous media data.

In recent years, cross-modal retrieval has gained much attention. Some cross-modal methods analyze the content correlation of heterogeneous data. Hierarchical manifold learning is proposed in [17] for the content-based cross-media retrieval. CCA is used by [18] to localize visual events associated with sound sources. In [19], relative importance of object is leveraged to construct tag features, and then kernel CCA [20] is used for cross-modal retrieval. Besides, other methods analyze the semantic correlation. A cross-media correlation graph is constructed in [21] according to the media objects features and their co-existence information, and cross-modal retrieval is performed on this graph. [22] constructs Multimedia Correlation Space by exploring the semantic correlation of different multimedia modalities, then the ranking algorithm Local Regression and Global Alignment is proposed. However, [21] and [22] do not give the deep analyzing for the semantic correlation, the semantic concepts are not used in these methods, and they mainly correlate multiple modalities by contents of multimedia. Other works correlate heterogeneous modalities on manifest semantic concepts [23, 36], but they cannot be applied in the case that manifest concepts are missing in the training data. SCM [23] combines content correlation and semantic correlation for cross-modal retrieval. CCA is firstly used to learn a content correlation space for images and texts, then logistic regression is used to learn the semantic correlation based on manifest semantic concepts. However, CCA is the preliminary step of SCM, and only the semantic correlation is preserved for retrieval, thus SCM cannot work well in the case that the manifest semantic concepts are lacking.

The cross-modal methods also have other applications, many works use cross-modal analysis to improve the uni-modal and multi-modal retrieval. In [24], cross-media correlations are explored to improve the multimedia document retrieval. [25] uses Markov random field to model the correlation of heterogeneous modalities, then the cross-modality similarity is learned for the task of multi-modal retrieval. [26] uses the asymmetric non-negative matrix factorization to learn the latent factors which correlate the images and texts in the same representation space, and multi-modal retrieval is based on this representation. In [27], the visual–textual joint relevance is determined by a hypergraph learning approach, and then the relevance is used for social image retrieval. In [40], image similarity is learned from the cross-modal relation of images and associated textual documents, and query expansion is used for internet cross-media retrieval.

3 Learning latent semantic concepts

To learn the semantic correlation of heterogeneous modalities explicitly, semantic concepts are needed to label the training data. Traditional manually labeling uses existing manifest semantic concepts such as categories, objects, etc., it is time consuming and the cost is expensive. However, we can learn semantic concepts from training data automatically by unsupervised learning methods instead. These semantic concepts are different from manual labeled concepts; they may be more difficult to be understood by human than manifest concepts, so we refer to them as latent semantic concepts to make them different from traditional semantic concepts. A good automatic learning method can get latent semantic concepts which are nearly equal to the ground truth of the manifest semantic concepts, and basing on these latent concepts a good semantic correlation can be constructed for cross-modal retrieval.

In this section, we describe the method: ASC, an asymmetric type of spectral clustering (SC) [28], to learn the latent semantic concepts from training data. Unlike symmetric methods which use all modalities in training data for learning, ASC only uses one modality for learning. Despite asymmetric using of modalities, ASC is actually a standard SC algorithm. Text features tend to provide a more reliable information source to extract semantic information for retrieval than other modality such as visual features, and asymmetric methods have been proven to perform well on text features [11, 26]. Moreover, it is easy to collect textual information from web resources such as social tags and comments, wikipedia articles, etc. ASC also learns latent semantic concepts from text features of training data, and then these concepts are used to label all the training data including other modalities.

In this paper, our cross-modal analysis only concentrates on text and image modality which are the most prevalent on the web. However, our methods are also suitable for other modalities such as audio and video. Suppose the training data are composed of N multimedia documents, each document M_i contains text T_i and an image I_i. The purpose of ASC is to learn K latent semantic concepts $LS_k(k = 1, \ldots, K)$ and assign them to every document in the training data. The steps of ASC are:

1. Construct the similar graph W of text features;
2. Compute the laplacian matrix $L = D - W$ where D is a diagonal matrix and $d_i = \sum_{j=1}^{N} w_{ij}$;
3. Compute the first K generalized eigenvectors $u_1 \cdots u_K$ of the generalized eigenproblem $Lu = \lambda Du$;
4. Let U be the matrix containing the vectors $u_1 \cdots u_K$ as columns, for $i = 1, \ldots, n$, let v_i be the vector corresponding to the i-th row of U;
5. Use the k-means algorithm to cluster $(v_1 \cdots v_N)$ into K clusters $LS_1 \cdots LS_K$ which is also the latent semantic concepts. After clustering each text belongs to one of the K latent semantic concepts, the document which contains the text and the image in the document is also labeled by this concept.

The steps of ASC are similar to classical spectral clustering methods, but ASC only concentrates on text features. The computation of similar graph W on text features is:

$$W = \begin{cases} w_{ij} & w_{ij} > \delta \\ 0 & w_{ij} < \delta \qquad \text{or} \quad i = j \end{cases} \tag{1}$$

where w_{ij} is the similarity of text feature T_i and T_j in the training documents, δ is the threshold to control the sparsity of the similar matrix. Text features are histogram of term frequency and we use histogram intersection distance as the similarity, the details of text features will be discussed in the experimental section.

If the training data are labeled by manifest semantic concepts, ASC may be not necessary for our semantic-based cross-modal analysis. However, the role of ASC is very important to make our cross-modal model practicable. In reality, it is hard to get enough semantic labeled multi-modal data for training, and it is also not easy to know all semantic concepts about the training data. If the training data are labeled by semantic concepts, then the semantic correlation can be obtained based on the existing concepts. However, in most cases, the semantic information of training data is lacking, we need to use ASC to learn the latent semantic concepts instead, and then construct the semantic correlation based on the latent concepts. Besides, the latent semantic information is also supplementary for the semantic information.

4 CMSGM

Once we obtain the manifest semantic concepts or latent semantic concepts for the training data, we can learn the semantic correlation of heterogeneous modalities. CMSGM is proposed to construct the semantic correlation for cross-modal retrieval in the probabilistic framework. The core idea of CMSGM is that if multimedia data from different modalities share the same semantic concept, then they can be generated by this concept, and they should be generated independently. This means given a semantic concept, the generation processes of heterogeneous data are conditional independent. These generation processes are analogous to the generation of media data from different modalities in the real world. For example, given the concept "bird", using the camera to get an image of bird, is independent of using text words to describe the bird. Although the generation processes of the two media data are independent, they both describe the concept "bird", which makes them correlated. CMSGM has a good description for the real correlation of media data. In addition, introducing the conditional independent property into the generation processes makes the model concise and easy to be solved.

4.1 Description of CMSGM

Suppose there are K semantic concepts $S_k(k = 1, \ldots, K)$ which may be manifest concepts from the existing knowledge or be latent concepts learned by ASC. We let $\mathbf{S}$ be the semantic concept for multi-modal document $\mathbf{M} = (\mathbf{I}, \mathbf{T})$, where $\mathbf{I}$ represents the visual feature of the image in the document and $\mathbf{T}$ represents the text feature. $\mathbf{S}$ is a K dimensional vector which follows the 1-of-K scheme, one of the elements in the vector equals 1, and all remaining elements equal 0, and the k-th element equals 1 means the document has the semantic concept S_k. $\mathbf{S}$ follows the categorical distribution with parameter $\mu = [\mu_1, \ldots, \mu_K]$. Then, we denote two semantic conditional probability distributions: $P(\mathbf{I}|\mathbf{S}, \theta_I)$ and $P(\mathbf{T}|\mathbf{S}, \theta_T)$, they are the distributions on the semantic concepts to generate image $\mathbf{I}$ and text $\mathbf{T}$, θ_I and θ_T are the parameters of the distributions.

CMSGM assumes the following generative process from which a multi-modal document $\mathbf{M}$ consists of image and text is generated:

1. Choose a semantic concept $\mathbf{S} \sim \text{Categorical}(\mu)$;
2. Generate image $\mathbf{I}$ from concept $\mathbf{S}$ by $\mathbf{I} \sim P(\mathbf{I}|\mathbf{S}, \theta_I)$;
3. Generate text $\mathbf{T}$ from concept $\mathbf{S}$ by $\mathbf{T} \sim P(\mathbf{T}|\mathbf{S}, \theta_T)$.

From the generative process above, we obtain the joint distribution of the document and a semantic concept:

$$P(\mathbf{I}, \mathbf{T}, \mathbf{S}) = P(\mathbf{I}|\mathbf{S}, \theta_I)P(\mathbf{T}|\mathbf{S}, \theta_T)P(\mathbf{S}|\mu) \tag{2}$$

Marginalizing over $\mathbf{S}$, we can get the joint distribution of image and text features, which also describes the probability of the multimedia document:

$$P(\mathbf{M}) = \mathbf{P}(\mathbf{I}, \mathbf{T}) = \sum_{\mathbf{S}} \mathbf{P}(\mathbf{I}|\mathbf{S}, \theta_{\mathbf{I}})\mathbf{P}(\mathbf{T}|\mathbf{S}, \theta_{\mathbf{T}})\mathbf{P}(\mathbf{S}|\mu) \tag{3}$$

The graphical illustration of the generation process is shown in Fig. 1. Image and text are correlated according to their generation from the same semantic concept, and the semantic correlation between them can be described by the joint probability. If images and texts are semantically correlated closely, then they are likely to be generated by the same semantic concept, and the joint probability of them is higher. The model has two advantages. One is that images and texts are correlated in the high-level semantics which makes the cross-modal and multi-modal retrieval to become more effective. The other is the conditional independence which enables images and texts to be modeled separately, which makes the model easy to be learned and extended to the semantic model with more than two modalities. While using CMSGM for retrieval, individual image and text are assumed to be in the same document, then their semantic correlation can be described by the joint probability. Our CMSGM is somewhat similar to the traditional generative models, the difference is that our CMSGM purely focuses on the semantic correlation which is based on the existing semantic concepts, they may be manifest concepts provided by people, or latent concepts automatically learned by ASC.

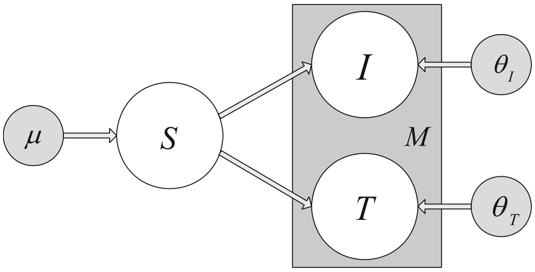


Fig. 1 The graphical illustration of the CMSGM

4.2 Estimation for CMSGM

To estimate the joint probability distribution of our model, we need to estimate the prior parameter μ, as well as the semantic conditional probability distribution parameters θ_I and θ_T. Let the number of the multi-modal document $\mathbf{M}_n = (\mathbf{I}_n, \mathbf{T}_n)$ in the training set be N. Using maximum likelihood estimation, the log-likelihood function of the training data can be expressed as:

$$\max_{\mu,\theta_I,\theta_T} \sum_{n=1}^{N} (\log P(\mathbf{I}_n|\mathbf{S}_n, \theta_I) + \log P(\mathbf{T}_n|\mathbf{S}_n, \theta_T) + \log P(\mathbf{S}_n|\mu)) \tag{4}$$

Then, we can write (4) in the summation of three parts:

$$\max_{\mu,\theta_I,\theta_T} \sum_{n=1}^{N} \log P(\mathbf{I}_n|\mathbf{S}_n, \theta_I) + \sum_{n=1}^{N} \log P(\mathbf{T}_n|\mathbf{S}_n, \theta_T) + \sum_{n=1}^{N} \log P(\mathbf{S}_n|\mu) \tag{5}$$

From (5), it can be seen that the semantic conditional probability of image feature $\mathbf{I}_n$, text feature $\mathbf{T}_n$, and the prior parameters of semantic concepts, are independent of each other. Thus, we can maximize these three terms separately, which will maximize the entire Eq. (4).

It is easy to estimate the prior parameters $\mu = [\mu_1, \ldots, \mu_K]$, we can get the optimal estimation function of the parameter from the third term of Eq. (5):

$$\arg\max_{\mu} \sum_{n=1}^{N} \log p(\mathbf{S}_n|\mu) \tag{6}$$

According to the properties of categorical distribution, we can get the constraint $\sum_{k=1}^{K} \mu_k = 1$. Then, using the method of Lagrange multipliers, we obtain the estimation:

$$\mu_k = \frac{N_k}{N} \quad k = 1, \ldots, K \tag{7}$$

where N_k is the number of training data with the k-th semantic concept S_k. The prior probability of a semantic vector is:

$$P(\mathbf{S}|\mu) = \prod_{k=1}^{K} \mu_k^{s_k} \tag{8}$$

where s_k is the k-th element of vector $\mathbf{S}$, it is 1 or 0 which denotes the presence/absence of the concept S_k.

To estimate the semantic distribution, we need to maximize the first and second terms of the Eq. (5), respectively. Without loss of generality, we use $\mathbf{X}$ to represent image feature as well as text feature. Then, the semantic conditional distribution of a media data $\mathbf{X}$ can be expressed as $P(\mathbf{X}|\mathbf{S}, \theta_X)$, the optimal function is:

$$\arg\max_{\theta_X} \sum_{n=1}^{N} \log P(\mathbf{X}_n|\mathbf{S}_n, \theta_X) \tag{9}$$

The general methods are directly estimating the semantic conditional distribution $P(\mathbf{X}|\mathbf{S}, \theta_X)$, it can be defined as parametric distribution, such as Gaussian. Assume that each semantic concept S_k corresponds to a Gaussian distribution with mean μ_k, and all the Gaussians share the same covariance matrix Σ, which makes the model easy to be estimated and more effective. Using the maximum likelihood estimation, the estimations of means and covariance matrix are:

$$\mu_k = \frac{\sum_{n=1}^{N} M_{nk} X_n}{\sum_{n=1}^{N} M_{nk}} \quad k = 1, \ldots, K \tag{10}$$

$$\Sigma = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} M_{nk}(X_n - \mu_k)(X_n - \mu_k)^{\mathrm{T}}}{N} \tag{11}$$

where M_{nk} indicates whether $\mathbf{X}_n$ corresponds to the k-th semantic concept S_k, if $\mathbf{X}_n$ belongs to S_k, then $M_{nk} = 1$, otherwise $M_{nk} = 0$. We denote this method, which directly estimates Gaussian distribution, as SGM-Gaussian.

However, it is difficult to directly model the distribution $P(\mathbf{X}|\mathbf{S}, \theta_X)$ for multiple high-dimensional multimedia features; using an existing probability distribution may not be well to depict $P(\mathbf{X}|\mathbf{S}, \theta_X)$. We consider the discriminative approach to estimate semantic conditional distributions indirectly. Unlike general direct approach, discriminative approach estimates the posterior probability distributions of semantic concepts, and then the posterior probability can be used to infer the semantic conditional probability by Bayes rule. Discriminative approach has the superiority that it avoids directly estimating the complex distribution of multimedia features, while obtaining a relative accurate estimation of the posterior distribution, then it will lead to more accurate estimation of semantic conditional distribution. When the posterior has been estimated, we can use the equation below to obtain the semantic conditional distribution:

$$P(\mathbf{X}|\mathbf{S}, \theta_X) = \frac{P(\mathbf{S}|\mathbf{X}, \theta_X) P(\mathbf{X})}{P(\mathbf{S}|\mu)} \tag{12}$$

where $P(\mathbf{X})$ is the prior of feature $\mathbf{X}$. It is difficult to estimate the priors of features in the discriminative approach, fortunately they can be ignored in the retrieval, which will be shown in Sect. 4.3. $P(S|\mu)$ can be obtained using Eq. (8). At this time, θ_X is now the discriminative model parameter which is used to predict the posteriors on $\mathbf{X}$. And the estimation of semantic conditional distribution is converted to estimation of the posterior of the semantic concept.

From Eq. (12), we can know if the posterior is accurate, then the semantic conditional probability calculated by posterior is also accurate. Many discriminative methods can obtain the posteriors of semantic concepts, such as logistic regression, support vector machine (SVM), and ensemble learning methods, etc. We adopt the SVM [29] for both image and text features. SVM is a good discriminative method for semantic learning, classification and object detection and has been successfully applied to various modalities, it can obtain precise discriminative results but do not cost much computing resource, thus it is suitable for large-scale data. LibSVM [30] is used as the implementation of SVM and it has the ability to predict the posterior of each semantic concept. We also adopt the one-versus-all scheme [31] to learn multiple semantic concepts.

4.3 Cross-modal similarity

After constructing and estimating the CMSGM, the cross-modal similarity can be inferred from it. CMSGM describes the semantic correlation of heterogeneous modalities; it can measure the cross-modal similarity between heterogeneous modalities. In the previous section, we have estimated the joint probability of the image feature $\mathbf{I}$ and text feature $\mathbf{T}$, the joint probability shows the possibility that the image and the text are generated by the same semantic concept, thus it can be used for computing the similarity between the image and the text. We consider two types of cross-modal retrieval: image query, where an image query example is used to retrieve texts; and text query, where a text query example is used to retrieve images. For both two types of cross-modal retrieval, the corresponding CMSGM similarities are inferred by the joint probability in CMSGM.

In image query, suppose the query image is $\mathbf{I}_q$, for each text $\mathbf{T}_n$ ($n = 1, \ldots, N_T$, N_T is the number of texts in the database) to be retrieved, the similarity of the query image $\mathbf{I}_q$ and text $\mathbf{T}_n$ is expressed by the conditional probabilities of texts on the image:

$$\mathrm{Sim}(\mathbf{I}_q, \mathbf{T}_n) = P(\mathbf{I}_q|\mathbf{T}_n) = \frac{P(\mathbf{I}_q, \mathbf{T}_n)}{P(\mathbf{T}_n)} \tag{13}$$

Using the joint probability of CMSGM and substituting it into (13), the similarity can be expressed as:

$$\mathrm{Sim}(\mathbf{I}_q, \mathbf{T}_n) = \sum_{k=1}^{K} \frac{P(\mathbf{I}_q|S_k, \theta_I) P(\mathbf{T}_n|S_k, \theta_T) P(S_k)}{P(\mathbf{T}_n)} \tag{14}$$

The semantic conditional probabilities $P(\mathbf{I}_q|S_k,\theta_I)$ and $P(\mathbf{T}_q|S_k,\theta_T)$ in (14) can be substituted by Eq. (12), then (14) is transformed to the equation expressed by posteriors:

$$\mathrm{Sim}(\mathbf{I}_q,\mathbf{T}_n)=\sum_{k=1}^{K}\frac{P(S_k|\mathbf{I}_q,\theta_I)P(S_k|\mathbf{T}_n,\theta_T)P(\mathbf{I}_q)}{P(S_k)} \quad (15)$$

In text query, the similarity of query text $\mathbf{T}_q$ and each image $\mathbf{I}_n$ ($n=1,\ldots,N_I$, N_I is the number of images in retrieval set) is also a conditional probability which is slightly diferent to the similarity of image query:

$$\mathrm{Sim}(\mathbf{T}_q,\mathbf{I}_n)=P(\mathbf{T}_q|\mathbf{I}_n) \quad (16)$$

Then, we can compute the similarity which is similar to Eq. (15):

$$\mathrm{Sim}(\mathbf{T}_q,\mathbf{I}_n)=\sum_{k=1}^{K}\frac{P(S_k|\mathbf{T}_q,\theta_T)P(S_k|\mathbf{I}_n,\theta_I)P(\mathbf{T}_q)}{P(S_k)} \quad (17)$$

We can find in the two similarities (15) and (17), the prior $P(\mathbf{I}_q)$ and $P(\mathbf{T}_q)$ can be ignored for the same query. If we get the semantic posteriors of image and text, then we can easily compute the similarity of CMSGM. Unlike traditional approaches which use joint probability $P(\mathbf{I},\mathbf{T})$ for similarity, we use conditional probability $P(\mathbf{I}|\mathbf{T})$ for image query and $P(\mathbf{T}|\mathbf{I})$ for text query. The main difference between two types of probability is the prior $P(\mathbf{X})$, it describes the probability that each feature occurs, and it may decrease the performance of our method in cross-modal retrieval.

5 CMSGM for cross-modal retrieval

If the manifest semantic information exists in the training data, and we know all the manifest semantic concepts for the training data, then we can use CMSGM to compute the cross-modal similarity based on the manifest semantic concepts. However, this is only the ideal case. In general, we do not know the manifest semantic information of the training data, or the manifest semantic information in the training data is incomplete. To cope with these two more general cases, we need to design different retrieval procedures. In the next three subsections, we propose three different cross-modal retrieval methods based on CMSGM for different cases, respectively.

5.1 CMSGM on manifest semantic concepts

In the ideal case that we know all the manifest semantic concepts in the training data, CMSGM based on the manifest semantic concepts can be directly constructed for the cross-modal retrieval. Suppose there are N_I images and N_T texts in the retrieval set, they are individual document and are not correlated. When there are manifest semantic concepts MS for training data, the preliminary procedure for cross-modal retrieval is:

1. Train the SVM model θ_I for the images in the training data;
2. Train the SVM model θ_T for the texts in the training data;
3. Estimate the prior of semantic concepts from training data using Eq. (7);
4. Estimate the semantic posteriors $P(MS_k|\mathbf{I}_n,\theta_I)(k=1,\ldots,K,\ n=1,\ldots,N_I)$ for each image in the retrieval set using SVM θ_I;
5. Estimate the semantic posteriors $P(MS_k|\mathbf{T}_n,\theta_T)$ for each text in the retrieval set using SVM θ_T.

The preliminary procedure estimates the parameters of conditional semantic distribution for CMSGM, and maps the images and texts into the semantic space which is represented by semantic posteriors. Then, the images and texts in the retrieval set can be retrieved by the CMSGM retrieval procedure.

While the new query comes, for image query, the retrieval procedure of CMSGM is:

1. Estimate the semantic posteriors $P(MS_k|\mathbf{I}_n,\theta_I)$ of the query image using SVM θ_I;
2. Compute the CMSGM similarity $\mathrm{Sim}(\mathbf{I}_q,\mathbf{T}_n)$ of query image $\mathbf{I}_q$ and each text $\mathbf{T}_n(n=1,\ldots,K)$ from the retrieval set using Eq. (15);
3. Rank the similarities in descending order, and return the texts which are most similar to the query image in semantic.

For text query, the retrieval procedure is similar to the image query, and only the roles of text and image are reversed in retrieval. The method which uses CMSGM on manifest semantic concepts for retrieval is called as M-CMSGM. Since the manifest semantic concepts are close to the human understanding, the semantic correlation constructed on manifest concepts can grasp almost all the semantic aspects of human understanding. Thus, M-CMSGM can perform well on complete semantic concepts of training data.

5.2 CMSGM on latent semantic concepts

There is usually no manifest semantic information in the multi-modal document for training, and it is hard to obtain the semantic concepts for the training data manually. The ASC described in the previous section of this paper can automatically learn latent semantic concepts, which can be seen as substitution for the manifest semantic concepts for the training data. Thus, we first use ASC for learning the latent semantic concepts, and then we can use CMSGM based on the latent semantic concepts for the cross-modal retrieval. For the case that there is no manifest semantic information in the training data, the preliminary procedure for cross-modal retrieval is:

1. Learn the latent semantic concepts from training data by ASC;
2. Train the SVM model θ'_I and θ'_T for the images and texts in the training data, respectively, the two SVM models are different to the SVM models in Sect. 5.1, they are about the latent semantic concepts;
3. Estimate the semantic posteriors $P(LS_k|\mathbf{I}_n, \theta'_I)(k = 1, \ldots, K, n = 1, \ldots, N_I)$ for each image in the retrieval set using SVM θ'_I;
4. Estimate the semantic posteriors $P(LS_k|\mathbf{T}_n, \theta'_I)$ for each text in the retrieval set using SVM θ'_T.

Besides the step (1) which uses ASC to learn latent semantic concepts, the rest steps are similar to the steps of preliminary procedure in Sect. 5.1, the only difference is that the manifest semantic concepts in Sect. 5.1 are replaced by latent semantic concepts. To make the two types of retrieval distinguishable, we denote the latent semantic concepts as $LS_k(k = 1, \ldots, K)$ to distinguish it from the manifest semantic concepts MS_k.

While the new query comes, for image query, the retrieval procedure of CMSGM on latent semantic concepts is:

1. Estimate the latent semantic posteriors $P(LS_k|\mathbf{I}_n, \theta'_I)$ of the query image $\mathbf{I}_q$ using SVM θ'_I;
2. Compute the CMSGM similarity $\mathrm{Sim}(\mathbf{I}_q, \mathbf{T}_n)$ of query image $\mathbf{I}_q$ and each text $\mathbf{T}_n(n = 1, \ldots, K)$ from the retrieval set using Eq. (15), it should be noted here that in this equation only the posteriors of latent semantic concepts are used;
3. Rank the similarities in descending order, and return the texts which are most similar to the query image in semantic.

The retrieval procedure of text query is also similar to the image query and the roles of image and text are reversed. We can find that the retrieval procedure of CMSGM based on latent semantic concepts is similar to the retrieval procedure in Sect. 5.1 which is based on manifest semantic concepts. And we denote the CMSGM retrieval methods based on latent semantic concepts as L-CMSGM.

5.3 Combining latent and manifest semantic concepts for retrieval

In the previous two sections, we consider the two cases that the training data either contain complete manifest semantic information or do not contain any manifest semantic information. Sometimes the multi-modal data for training may also contain incomplete manifest semantic information. Documents in the training data may not be labeled by all manifest semantic concepts completely, part of the manifest semantic concepts corresponding to the documents are labeled, and the other concepts are missed. So using the incomplete manifest semantic information of the training data can only obtain the incomplete semantic correlation. Combing the latent semantic correlation can supplement the incomplete manifest semantic correlation for cross-modal retrieval. While combining the latent and manifest semantic concepts for cross-modal retrieval, the preliminary procedure is:

1. Execute the preliminary procedure based on manifest semantic concepts in Sect. 5.1 to obtain SVM θ_I and θ_T, semantic posteriors $P(MS_k|\mathbf{I}_n, \theta_I)$ and $P(MS_k|\mathbf{T}_n, \theta_T)$ for each image and text, respectively, in the retrieval set;
2. Execute the preliminary procedure based on latent semantic concepts in Sect. 5.2 to obtain SVM θ'_I and θ'_T, semantic posteriors $P(LS_k|\mathbf{I}_n, \theta'_I)$ and $P(LS_k|\mathbf{T}_n, \theta'_T)$;

In the retrieval procedure, we need to combine the retrieval results of the two retrieval methods. We linearly combine the similarities of the two retrieval methods with equal weights, which is simple but effective. For both image query and text query, suppose the similarity of M-CMSGM is $\mathrm{MSim}(\mathbf{I}, \mathbf{T})$, and the similarity of L-CMSGM is $\mathrm{LSim}(\mathbf{I}, \mathbf{T})$, then the final combination of similarity is:

$$\mathrm{Sim}(\mathbf{I}, \mathbf{T}) = \alpha \cdot \mathrm{MSim}(\mathbf{I}, \mathbf{T}) + \mathrm{LSim}(\mathbf{I}, \mathbf{T}) \tag{18}$$

where α is used to control the importance of the two similarities. However, in real world it is difficult to determine which part is more important. If the manifest concepts in the training data are extremely incomplete, α should be very small. If the manifest concepts are relatively complete, then α should be relatively large. In our work, we set $\alpha = 1$, it may be optimized by the relevance feedback from users.

The retrieval procedure of the combination of M-CMSGM and L-CMSGM is:

1. Estimate the manifest semantic posteriors $P(MS_k|\mathbf{I}_n, \theta_I)$ of the query image $\mathbf{I}_q$ using SVM θ_I, and estimate the latent semantic posteriors $P(LS_k|\mathbf{I}_q, \theta'_I)$ of the query image $\mathbf{I}_q$ using SVM θ'_I;
2. Compute the M-CMSGM similarity $\mathrm{MSim}(\mathbf{I}_q, \mathbf{T}_n)$ of query image $\mathbf{I}_q$ and each text $\mathbf{T}_n(n = 1, \ldots, K)$ from the retrieval set as well as the L-CMSGM similarity $\mathrm{LSim}(\mathbf{I}_q, \mathbf{T}_n)$;
3. Compute the combination of the two similarities to get the final similarity $\mathrm{Sim}(\mathbf{I}_q, \mathbf{T}_n)$ using Eq. (18);

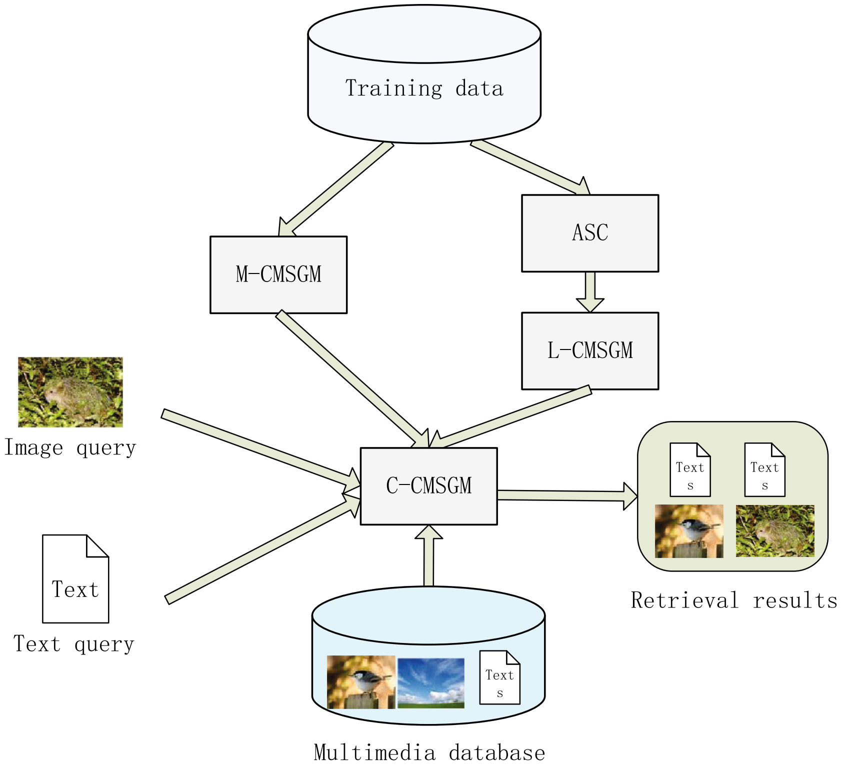


Fig. 2 The framework of our C-CMSGM for cross-modal retrieval, it also contains the procedures of M-CMSGM and L-CMSGM

4. Rank the final similarities in descending order, and return the texts which are most similar to the query image in semantic.

For text query, the retrieval procedure is the same and the roles of image and text are reversed. The method which combines M-CMSGM and L-CMSGM for cross-modal retrieval, is denoted as C-CMSGM. C-CMSGM gains the advantages of both M-CMSGM and L-CMSGM. When most manifest semantic concepts exist in the training data, the manifest part of C-CMSGM plays a more important role. And when the manifest semantic concepts are lacking, the latent part of C-CMSGM has a more important effect. C-CMSGM has a wide range of application; it can also obtain a good performance while manifest concepts are complete; thus if we cannot know whether the manifest semantic concepts are complete, we can consider the C-CMSGM. Figure 2 shows the framework of cross-modal retrieval by C-CMSGM, M-CMSGM and L-CMSGM are also shown in it.

6 Experiments

In this section, we will describe details of two datasets used and also discuss the visual features and textual feature extracted from these two datasets. Then, we will test our three methods on these two datasets. We also compare the experimental results of our methods with the varying amount of preserved concepts. Finally, we show the illustrative examples of our retrieval methods.

6.1 Datasets

We use two datasets: Wikipedia featured articles and MIR Flickr for our experiments.

The dataset of Wikipedia featured articles was firstly used in [23]. It is a continually updated collection in which articles have been selected and reviewed by Wikipedia's editors. The articles are accompanied by one or more pictures from the Wikimedia Commons. Each featured article is categorized by Wikipedia into 29 original concepts. These concept labels are assigned to both the text and image components of each article. Since some of the concepts are very scarce, only ten most popular ones are considered. The dataset is finally pruned by removing the sections without images. The final corpus contains a total of 2,866 documents. These documents are text–image pairs, annotated with a label from the vocabulary of ten concepts, which are treated as the manifest semantic concepts in this paper. And the dataset is divided into training set of 2,173 documents and test set of 693 documents.

The MIR Flickr dataset [32] contains 25,000 images downloaded from the popular online photo-sharing service Flickr. These images were collected directly from the web, to provide a realistic dataset for multimedia retrieval research, with high-resolution images and associated metadata. Images of this dataset were annotated for 24 semantic concepts, including not only object categories but also more general scene concepts such as sky, water and indoor. For 14 of the 24 concepts a second, stricter, annotation was made: for each concept, a subset of the positive images was selected where the concept is salient in the image. In total, there are 38 manifest semantic concepts for this dataset. Images in this dataset are also associated with the Flickr tags given by users, which can be considered as the text information. Thus, we also get the image–text pairs with semantic concepts from this dataset. We kept the tags that appear at least 50 times, resulting in a vocabulary of 457 tags. Finally, we remove the images without tags, and obtain the training set of 9,359 images, test set of 9,335 images.

6.2 Features and kernels

We extract three visual features from images for both two datasets, including SIFT histogram, HOG histogram and GIST. For SIFT histogram, local SIFT descriptors are first computed on 16 × 16 overlapping patches with a spacing of 1 pixels. Then, we perform *k*-means clustering of a random subset of computed SIFT descriptors to form a visual vocabulary of 200 visual words. Each SIFT descriptor is quantized into a visual word using the nearest cluster center. At last, SIFT descriptors of each image map to a spatial pyramid histogram with two spatial scales [33], resulting in the 1,000-dimensional SIFT histogram. The extraction of HOG histogram is similar to SIFT histogram; the difference is that local HOG descriptors [34] are computed in the extraction procedure. The dimension of the HOG histogram computed is also 1,000. Furthermore, we use the GIST descriptor [35], which roughly encodes the image layout. For text modality, the vector of words frequency is used to represent each text. In Wikipedia featured articles, we kept the words which appear more than 20 times in the whole dataset, resulting in a vocabulary of 6,603 words, thus we will get the 6,603-dimension text features. In MIR Flickr, the text features are binary vectors; this is attributed to the fact that each Flickr tag appears no more than two times in each image-text pair. Finally, to compare with previous works on Wikipedia articles, we also use 128-D SIFT histograms and 10-D LDA features in [23] for this dataset.

For SIFT and HOG histograms, we use histogram intersection kernel, the kernel is calculated by the following equation:

$$K_{\text{intersection}}(x_i, x_j) = \sum_{d} \min(x_{id}, x_{jd}) \tag{19}$$

where x_i and x_j are the features, and x_{id} is the d-th element of x_i. For Gist descriptors, RBF kernel is used, it is calculated by:

$$K_{\text{RBF}}(x_i, x_j) = \exp(-d(x_i, x_j)/\lambda) \tag{20}$$

where $d(x_i, x_j)$ is the L2 distance, and λ is the mean of all distance values. After calculating kernels of the three visual features, we linearly combine three kernels with equal weights to obtain the final kernel for SVM. For text features, we also choose the histogram intersection kernel.

6.3 Retrieval results

The performance of all retrieval methods is evaluated on the test sets of the two datasets. The training sets are used for learning our models. We evaluate our methods on two types of cross-modal retrieval: text retrieval using image query, and image retrieval using text query. In the first case, each image in the test set is used as a query, producing a ranking of all texts in the test set. In the second case, the roles of images and texts are reversed. In all cases, performance is measured by mean average precision (MAP) and precision–recall (PR) curves. Average

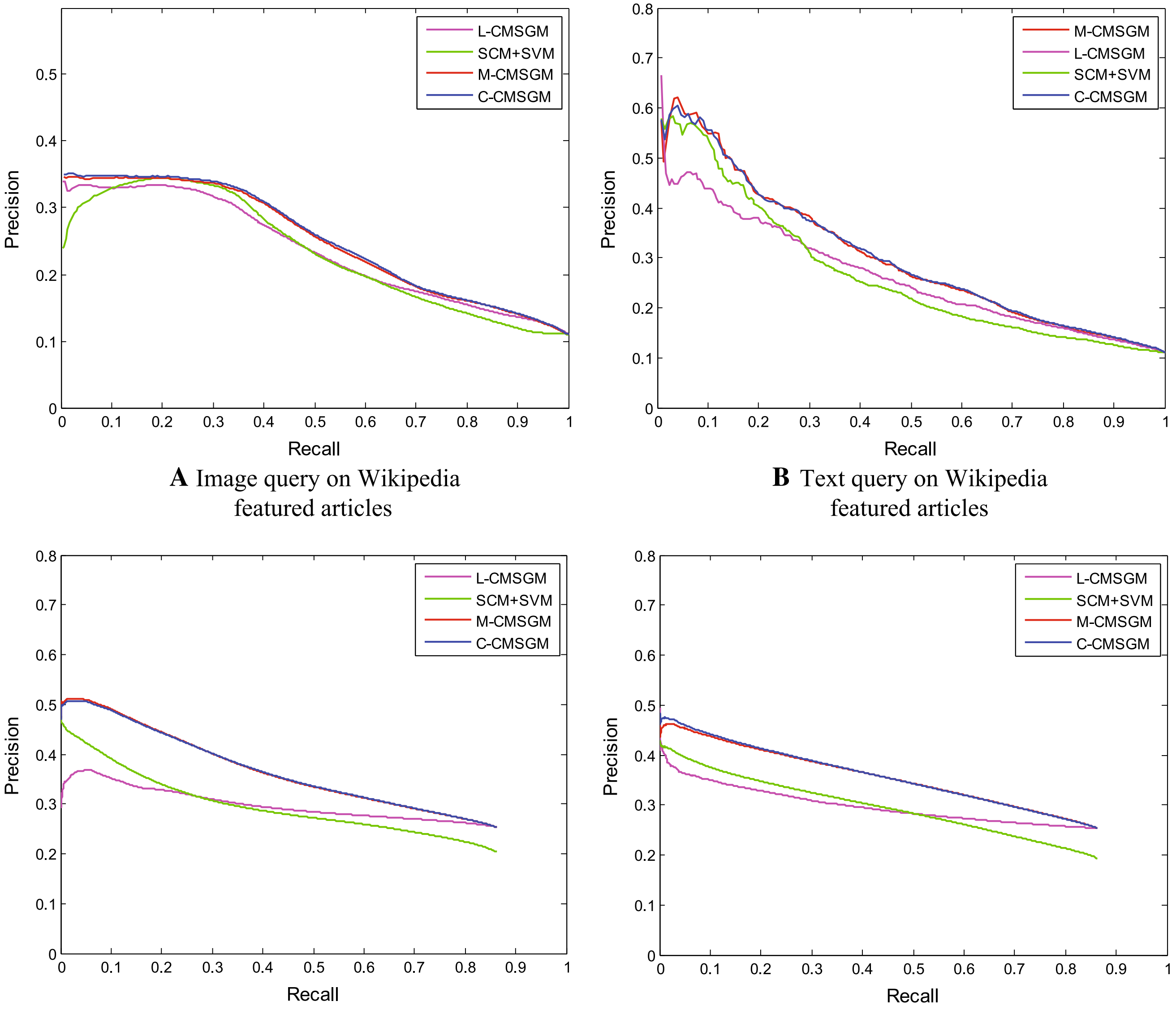
Table 1 The comparison of Map scores on two datasets for the case that all manifest semantic concepts are available

	Image query	Text query	Average
Wikipedia featured articles			
SCM [23]	0.277	0.226	0.252
MSCP [37]	0.329	0.256	0.293
AHSM [38]	0.347	0.259	0.303
SM+SVM(SIFT+LDA)	0.299	0.272	0.286
SGM-Gaussian	0.321	0.253	0.287
SGM-Gusssian(joint)	0.219	0.127	0.173
L-CMSGM(SIFT+LDA)	0.347	0.266	0.306
M-CMSGM(SIFT+LDA)	0.352	0.276	0.314
C-CMSGM(SIFT+LDA)	0.355	0.277	0.316
KCCA	0.302	0.254	0.278
ASC+GM+SVR	0.265	0.243	0.254
SCM+SVM	0.364	0.309	0.337
L-CMSGM	0.385	0.305	0.345
M-CMSGM	0.425	0.339	0.382
C-CMSGM	**0.426**	**0.340**	**0.383**
MIR FLickr			
KCCA	0.288	0.278	0.283
ASC+GM+SVR	0.289	0.294	0.292
SCM+SVM	0.332	0.322	0.327
L-CMSGM	0.321	0.319	0.320
M-CMSGM	**0.403**	0.389	0.396
C-CMSGM	**0.403**	**0.391**	**0.397**

The best results are marked as bold

precision is the average of precision values at the ranks where relevant items occur, which is further averaged over all queries to give MAP. MAP is widely used in the information retrieval. For the retrieval of Wikipedia featured articles, the documents which share the same semantic concept with the query are the relevant documents to the query. For the retrieval of MIR Flickr, because each document has multiple semantic concepts, we choose the documents which have at least two same semantic concepts to the query as relevant documents.

Table 1 shows the MAP scores of our three methods: L-CMSGM, M-CMSGM and C-CMSGM on two datasets. State-of-the-art methods on Wikipedia featured articles are also shown in Table 1. Some of them are absence in MIR Flickr, because there are few cross-modal experiments on this dataset. To make a fair comparison, we also show the performance of our three methods using only 128-D SIFT histograms and 10-D LDA features, which are used by previous works on Wikipedia articles. SGM-Gaussian uses Eqs. (15) and (18) as similarity, and SGM-Gaussian(joint) uses the joint probability $P(\mathbf{I}, \mathbf{T})$ as similarity. SCM+SVM is our implementation of SCM [23], we replace the logistic regression in SCM by probabilistic kernel SVM and use three visual features, which causes a significant improvement in MAP score. SM+SVM(LDA+SIFT) is similar to SM [23], the only difference is that logistic regression is replaced by probabilistic kernel SVM, and we can observe that using kernel SVM can improve the retrieval performance. In the implementation of KCCA all three visual features are used, and normalized correlation distance is



A Image query on Wikipedia featured articles

B Text query on Wikipedia featured articles

C Image query on MIR Flickr

D Text query on MIR Flickr

Fig. 3 Precision–recall curves

used for KCCA. For most methods, all manifest semantic concepts in the training set are available for cross-modal analysis, only L-CMSGM and KCCA do not use any manifest semantic concepts in training set, L-CMSGM only uses latent semantic concepts learned from content features (image features and text features) of training set. We use ASC to learn 10 latent semantic concepts in Wikipedia featured articles, and 30 latent semantic concepts in MIR Flickr, and the threshold of text similarity is set to 0.15 in Wikipedia featured articles, and 0 in MIR Flickr. In addition, we also evaluate a baseline latent method: ASC+GM+SVR to show the performance of soft assignment for detection of latent concepts. It is similar to L-CMSGM, the differences are that Gaussian mixture (GM) is used to replace *K*-means in ASC, and support vector regression (SVR) is used to train on the continuous labels. NC distance which is the best choice we have found, is used for ASC+GM+SVR.

From Table 1, we can observe on both two datasets our three CMSGM methods perform best, except on MIR Flickr SCM+SVM obtains a slightly higher MAP score than L-CMSGM. This result confirms the effectiveness of our CMSGM model for cross-modal retrieval. SGM-Gaussian performs better than SGM-Gaussian(joint), which confirms that $P(\mathbf{I}|\mathbf{T})$ for image query and $P(\mathbf{T}|\mathbf{I})$ for text query are more effective than $P(\mathbf{I}, \mathbf{T})$ in our framework. M-CMSGM outperforms SGM-Gaussian, which shows that indirect discriminative estimation for posteriors is better than direct generative estimation for Gaussian. We can also find that ASC+GM+SVR performs worse than L-CMSGM on both two datasets, which demonstrates that hard assignment for latent concepts is reasonable. Although soft assignment seems better than hard assignment, the advantages of CMSGM can make our methods obtain well retrieval results.

In addition, using three features makes our methods perform better than only using SIFT. Since HOG and GIST can be easily extracted from images, it is reasonable to use all three visual features for retrieval. L-CMSGM outperforms other methods except our two methods on Wikipedia featured articles, and on MIR Flickr L-CMSGM obtains only a slightly lower MAP score than SCM+SVM. We think the reason of the high performance of L-CMSGM is that the semantic correlation is based on latent semantic concepts which are an approximation of the manifest semantic concepts, and ASC can make the approximation quite close to the manifest correlation.

Our CMSGM can model a good correlation of heterogeneous modalities, and thus L-CMSGM has a competitive performance in comparison to the previous methods which use manifest concepts. In the case that no manifest concepts exist in the training data, we can use L-CMSGM for retrieval and the performance is even better than previous methods. C-CMSGM and M-CMSGM obtain similar MAP scores, and C-CMSGM has a slightly better performance. The PR curves of our three methods and SCM+SVM are

Table 2 The MAP scores of cross-modal retrieval for different methods and various amounts of preserved manifest concepts

Preserved concepts	Method	Image query	Text query	Average
Wikipedia featured articles				
2	L-CMSGM	**0.385**	0.305	**0.345**
	SCM+SVM	0.217	0.161	0.189
	M-CMSGM	0.220	0.172	0.196
	C-CMSGM	0.374	**0.313**	0.344
4	L-CMSGM	**0.385**	0.305	0.345
	SCM+SVM	0.272	0.192	0.232
	M-CMSGM	0.293	0.214	0.254
	C-CMSGM	0.379	**0.321**	**0.350**
6	L-CMSGM	0.385	0.305	0.345
	SCM+SVM	0.299	0.220	0.278
	M-CMSGM	0.342	0.245	0.250
	C-CMSGM	**0.391**	**0.325**	**0.358**
8	L-CMSGM	0.385	0.305	0.345
	SCM+SVM	0.304	0.233	0.269
	M-CMSGM	0.356	0.259	0.308
	C-CMSGM	**0.391**	**0.329**	**0.360**
10	L-CMSGM	0.385	0.305	0.345
	SCM+SVM	0.364	0.309	0.337
	M-CMSGM	0.425	0.339	0.382
	C-CMSGM	**0.426**	**0.340**	**0.383**
MIR Flickr				
8	L-CMSGM	0.321	0.319	0.320
	SCM+SVM	0.258	0.268	0.263
	M-CMSGM	0.270	0.286	0.278
	C-CMSGM	**0.322**	**0.323**	**0.322**
16	L-CMSGM	0.321	0.319	0.320
	SCM+SVM	0.291	0.282	0.287
	M-CMSGM	0.320	0.310	0.315
	C-CMSGM	**0.336**	**0.332**	**0.334**
24	L-CMSGM	0.321	0.319	0.320
	SCM+SVM	0.307	0.302	0.305
	M-CMSGM	0.355	0.323	0.339
	C-CMSGM	**0.364**	**0.344**	**0.354**
32	L-CMSGM	0.321	0.319	0.320
	SCM+SVM	0.320	0.316	0.318
	M-CMSGM	0.390	0.374	0.382
	C-CMSGM	**0.393**	**0.378**	**0.386**
38	L-CMSGM	0.321	0.319	0.320
	SCM+SVM	0.332	0.322	0.327
	M-CMSGM	**0.403**	0.389	0.396
	C-CMSGM	**0.403**	**0.391**	**0.397**

The best results are marked as bold

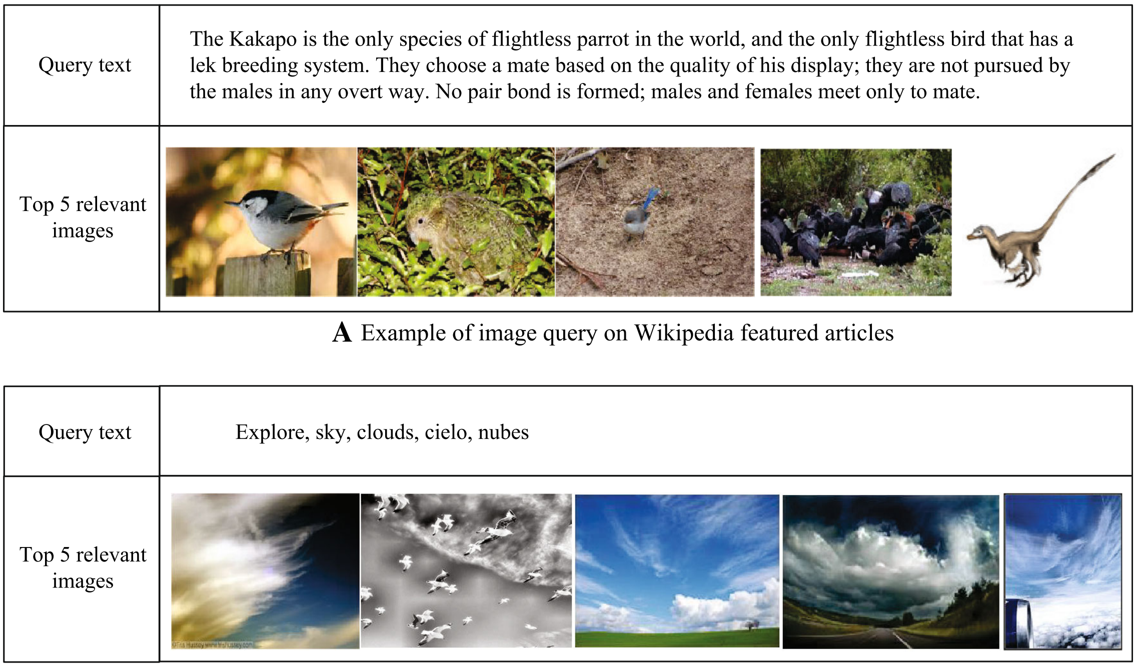
shown in Fig. 3. We can also see that the PR curves are consistent with the MAP scores, and C-CMSGM has almost the same PR curves with M-CMSGM. The advantage of C-CMSGM is not significant in the case that all manifest semantic concepts are available. And the latter experiment will show the advantage of C-CMSGM.

We have shown that in the case all manifest semantic concepts are available in the training data, M-CMSGM which learns semantic correlation based on manifest semantic concepts performs better than L-CMSGM. However, in the case that not all the manifest semantic concepts are available, M-CMSGM may be worse than L-CMSGM. Moreover, if there is no semantic information in training data, M-CMSGM cannot work but the performance L-CMSGM is unchanged.

We also evaluate the performance our methods in the case that training data lack the manifest semantic concepts. In the experiment, we preserve part of the manifest semantic concepts and the other concepts in the training data cannot be used by any methods, this can simulate the case that manifest semantic information is lacking. Table 2 shows the MAP scores of our three methods and SCM+SVM with varying amounts of preserved concepts. We randomly preserve 2, 4, 6, 8 manifest semantic concepts on Wikipedia featured articles, and 8, 16, 24, 32 on MIR Flickr. The results of all semantic concepts preserved in two datasets are also shown in Table 2 for comparison. The MAP score of M-CMSGM decreases significantly when the number of preserved concepts becomes less. And the M-CMSGM performs even worse than L-CMSGM when small amounts of manifest semantic concepts are preserved. SCM+SVM which is also based on the manifest semantic concepts, has the same decreasing of MAP score with M-CMSGM. In most cases, C-CMSGM obtains higher MAP scores than other two methods, which confirms the robustness of this method. C-CMSGM always has a good performance and adapts to most cases, if we do not know whether all manifest concepts are preserved in the training data (even if in fact the manifest concepts are complete in training data), C-CMSGM is the best choice. In the case that there are no manifest semantic concepts in the training data, L-CMSGM is the only choice. M-CMSGM performs well in the ideal case that all manifest semantic concepts are preserved in the training data, but it may be not practicable.

6.4 Illustrative examples

In this section, we show the illustrative examples of the C-CMSGM on the two datasets. We only show the text query in the case that all manifest concepts are preserved, which is enough to demonstrate the effectiveness of our CMSGM model. The examples of text query on two dataset are shown in Fig. 4, we can see from Fig. 4 that our cross-modal retrieval method can find the heterogeneous media data semantically correlated to the query. The text query of the Wikipedia featured articles is about the biology, and all the five relevant images are also about biology. The text query of MIR Flickr is about the sky and clouds, and all



Query text

The Kakapo is the only species of flightless parrot in the world, and the only flightless bird that has a lek breeding system. They choose a mate based on the quality of his display; they are not pursued by the males in any overt way. No pair bond is formed; males and females meet only to mate.

Top 5 relevant images

A Example of image query on Wikipedia featured articles

Query text

Explore, sky, clouds, cielo, nubes

Top 5 relevant images

B Example of image query on MIR Flickr

Fig. 4 Illustrative examples of text query

five relevant images contain sky and cloud. The text queries in the two datasets have different structures; text in Wikipedia featured articles consists of sentences and text in MIR Flickr consists of individual words; this shows that our methods can be applied to various types of text. Although our C-CMSGM constructs the latent semantic correlation, its performance is still affected by the completeness of manifest semantic concepts. In the retrieval of Wikipedia featured articles, the queries are also about bird, and one of the relevant images is about dinosaur, the reason is the two concepts are lacking in the dataset.

7 Conclusion

In this paper, we propose the CMSGM which has a good description for the semantic correlation of multiple modalities. And we design three methods M-CMSGM, L-CMSGM and C-CMSGM based on CMSGM for three different cases of the cross-modal retrieval. Experimental results show that CMSGM-based methods outperform the other cross-modal retrieval methods. L-CMSGM which constructs semantic correlation on latent semantic concepts learned by ASC, performs better than previously proposed methods which model the manifest concepts. This shows that our ASC can learn the latent concepts which are quite close to manifest concepts. Thus, even in the case that no manifest concepts exist, we can use L-CMSGM to obtain a good performance. Moreover, we also find when the manifest latent concepts are lacking, C-CMSGM always performs best, thus it is practicable in the more general case and M-CMSGM is only suited to the ideal case that all manifest concepts exist.

In the future work, some aspects of our methods still can be improved. The generation process of multiple modalities can be more sophisticated to better model the cross-modal correlation. In the estimation of posteriors SVM is used, and recent state-of-the-art multi-task learning [41] in discriminative task can be exploited to enhance the accuracy of estimation. The combination of L-CMSGM and M-CMSGM is proven to be effective, and weights learning method is needed to be designed for a better combination. Besides, the ability of our methods on dealing with large-scale multimedia data is needed to be tested.

References

1. Chandrika, P., Jawahar, C.V.: Multi modal semantic indexing for image retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, pp 342–349 (2010)
2. Wang, X.-J., et al.: Multi-model similarity propagation and its application for web image retrieval”. In: Proceedings of the 12th annual ACM international conference on Multimedia. ACM (2004)
3. Hoi, S.C.H., Lyu, M.R.: A multimodal and multilevel ranking scheme for large-scale video retrieval. IEEE Trans. Multimed. **10**(4), 607–619 (2008)
4. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
5. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. Advances in neural Information Processing Systems (2003)
6. Zhang, S., et al.: Automatic image annotation using group sparsity. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
7. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and development in information retrieval. ACM (2003)
8. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2. IEEE (2004)
9. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM (2003)
10. Blei, David M., Ng, Andrew Y., Jordan, Michael I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
11. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image indexing. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1802–1817 (2007)
12. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1–2), 177–196 (2001)
13. Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. IEEE Trans. Pattern Anal. Mach. Intell. **30**(8), 1371–1384 (2008)
14. Hertz, T., Bar-Hillel, A., Weinshall, D.: Learning distance functions for image retrieval. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (. CVPR), vol. 2. IEEE (2004)
15. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Conference on Computer Vision, ECCV 2008. Springer, Berlin, pp. 316–329
16. Guillaumin, M., et al.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. IEEE 12th International Conference on Computer Vision. IEEE (2009)
17. Yang, Y., et al.: Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Trans. Multimed. **10**(3), 437–446 (2008)
18. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1. IEEE (2005)
19. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int. J. Comput. Vis. **100**(2), 134–153 (2012)
20. Lai, P.L., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. Int. J. Neural Syst. **10**(05), 365–377 (2000)
21. Zhuang, Y.-T., Yang, Y., Wu, F.: Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Trans. Multimed. **10**(2), 221–229 (2008)
22. Yang, Y., et al.: Ranking with local regression and global alignment for cross media retrieval. In: Proceedings of the 17th ACM international conference on Multimedia. ACM (2009)
23. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the international conference on Multimedia. ACM (2010)

24. Chu, W.-T., Chen, H.-Y.: Toward better retrieval and presentation by exploring cross-media correlations. Multimed. Syst. **10**(3), 183–198 (2005)
25. Jia, Y., Salzmann, M., Darrell, T.: Learning cross-modality similarity for multinomial data. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE (2011)
26. Caicedo, J.C., et al.: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing **76**(1), 50–60 (2012)
27. Gao, Y., et al.: Visual–textual joint relevance learning for tag-based social image search. IEEE Trans. Image Process. **22**(1), 363–376 (2013)
28. Von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
29. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
30. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011)
31. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. **13**(2), 415–425 (2002)
32. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. ACM (2008)
33. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE (2006)
34. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1. IEEE (2005)
35. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
36. Xie, Liang, Pan, Peng, Lu, Yansheng.: A semantic model for cross-modal and multi-modal retrieval. Proceedings of the 3rd ACM conference on International conference on multimedia retrieval. ACM (2013)
37. Lu, Z., Ip, H.H.S., Peng, Y..: Exhaustive and efficient constraint propagation: a semi-supervised learning perspective and its applications. arXiv preprint arXiv:1109.4684 (2011)
38. Zhai, X., Peng, Y., Xiao, J.: Cross-media retrieval by intra-media and inter-media correlation mining. Multimedia Systems: 1–12
39. Atrey, P.K., Anwar Hossain, M., Saddik, AEl, Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. **16**(6), 345–379 (2010)
40. Jiang, S., Song, X., Huang, Q.: Relative image similarity learning with contextual information for Internet cross-media retrieval. Multimed. Syst. 1–13 (2013)
41. Yang, Y., Ma, Z., Hauptmann, A., Sebe, N.: Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans. Multimed. **15**(3), 661–669 (2013)