Multimedia Systems (2014) 20:645–657
DOI 10.1007/s00530-012-0299-4

REGULAR PAPER

Relative image similarity learning with contextual information for Internet cross-media retrieval

Shuqiang Jiang · Xinhang Song · Qingming Huang

Published online: 1 February 2013

Abstract With the fast explosive rate of the amount of image data on the Internet, how to efficiently utilize them in the cross-media scenario becomes an urgent problem. Images are usually accompanied with contextual textual information. These two heterogeneous modalities are mutually reinforcing to make the Internet content more informative. In most cases, visual information can be regarded as an enhanced content of the textual document. To make image-to-image similarity being more consistent with document-to-document similarity, this paper proposes a method to learn image similarities according to the relations of the accompanied textual documents. More specifically, instead of using the static quantitative relations, rank-based learning procedure by employing structural SVM is adopted in this paper, and the ranking structure is established by comparing the relative relations of textual information. The learning results are in more accordance with the human's recognition. The proposed method in this paper can be used not only for the image-to-image retrieval, but also for cross-modality multimedia, where a query expansion framework is proposed to get more satisfactory results. Extensive experimental evaluations on large scale Internet dataset validate the performance of the proposed methods.

S. Jiang (✉) · X. Song
Key Lab of Intelligent Information Processing,
Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China
e-mail: sqjiang@ict.ac.cn; Shuqiang.jiang@vipl.ict.ac.cn

X. Song
e-mail: xinhang.song@vipl.ict.ac.cn

Q. Huang
University of Chinese Academy of Sciences,
Beijing 100049, China
e-mail: qmhuang@jdl.ac.cn

Keywords Image similarity learning · Structural SVM · Cross-media retrieval · Query expansion

1 Introduction

There exist big amounts of data in the Internet with multiple modalities. From the beginning of the Internet, textual data always acts as the main information communication type among all the modalities. Recently, with the fast development of capturing devices, storage abilities, web facilities, and displaying equipments, Internet image data are growing as an explosive rate, and are playing more and more important roles in the Internet. Besides the specific image/video/audio/flash sharing websites such as Youtube, Flickr, etc., the more widely existed web pages in the Internet are news, blogs, reports, twitters, comments, etc. Among these web pages, image data usually co-occurs with the textual documents. For example, in the news website of CNN, BBC, FoxNews, etc., it is difficult to find a webpage without contextual images. Figure 1 illustrates some examples of web pages containing image. Compared with the complex and lengthy text document, images are more vivid and intuitive; and users are more intent to include image data to make the webpage more informative and readability. On one hand, visual information can be regarded as an enhanced content of the textual document; on the other hand, images can bring some information that could not be delivered by the textual data. As the two main types of media modalities in the Internet, image and text data contained in the web pages bring abundant information to the users. In order to better utilize Internet multimedia data and satisfy users' diverse application requirements, we need to establish effective techniques to process images contained in the web pages and to analyze the coherence relations of image and text data.

Fig. 1 Web pages containing both image data and textual documents

Internet image analysis methods can be categorized into the following classes: (1) only relying on textual information without using any visual computing. This is a traditional method and is widely used in many real applications due to its simplicity. However, textual information does not equal to the visual information, so the analysis results may be biased. (2) Directly computing the visual content. This kind of method has been intensively investigated in the research community [1–7], which has two limitations in the Internet scenarios: the first is the semantic gap between digital signals and semantic understanding for images; the second is the lack of the cooperation with the other information in the web pages, especially the textual documents. (3) Using tag information to improve the image understanding abilities [8–12]. The large number of available image tags facilitates this research topic. The disadvantages are on two folds. The first is the widely existed noisy tags and the second is that several simple tags usually could not cover the messages delivered by the images. (4) Integrating other contextual information to visual computing [13–16]. This kind of method is relatively reasonable compared with the aforementioned methods. The challenge is how to take the full potential of the available information to fulfill the Internet multimedia application tasks. Furthermore, the functionality of image data in the Internet web pages has not been paid much research attentions. Based on the image functions, the relationship of images with other modalities should be clearly analyzed. In this procedure, the similarities of images play an important role for Internet multimedia computing [17–21].

Image similarity is a complex problem, and it has multiple aspects as shown in Fig. 2. Some images are visually similar; some images contain the same concept, and some images are related with the same topic although they are not visually and semantically similar at all. Moreover, different similar image pairs may have different similarity extent. It is an interesting problem of how to appropriately establish the computing methods on image similarities. Traditional image similarity modeling techniques are based on the mathematical computing or distance metric learning. However, it has been found that image similarities do not follow the exact mathematical distance metric conditions. A direct example is that image similarity does not have the property of similarity transitivity; image I_1 is similar with image I_2, and image I_2 is similar with image I_3, however, image I_1 and I_3 may not be similar at all. On the other hand, it has been well recognized in the research community that learning based similarity computing methods are superior to the non-learning based methods for the following reasons: (1) un-supervised learning based methods have a better sample organization; (2) supervised learning based methods are more semantically discriminative; (3) both supervised and un-supervised learning methods are more separable based on the training samples. Most of the image similarity learning methods is based on image labels or image classes. Due to polysemous property of images, label-based similarity learning may have the following limitations: (1) it is single sided, as the learning procedure only considers the image labels, and other semantic information contained in the images are overlooked; (2) it is usually used for general applications; however, different application scenarios may focus on different semantic explanation; (3) it is inclined to over fit to the training samples due to the limited number of available data. Other image similarity learning methods consider other contextual information, such as tag based [17, 22], hierarchical based [21, 23], etc. Image similarity learning without using the image labels has not been much investigated. More comprehensive and subjective learning

Fig. 2 Some examples of similar images. **a**, **b** are visually similar images where **a** is an illustration of partially object similar and **b** is globally similar on color, **c** is an example of concept similar and **d** is an example of two images with similar topic: "tennis"

methods are desired to fulfill various application requirements on the Internet multimedia data.

To process and understand Internet multimedia information, first we need to analyze the structure of the multimedia content, and the functionality of different component, as well as their inter-relationships. This paper works on the image similarities in the Internet web pages, as most of the images exist in the web pages and images are co-occurred in the Internet. In most cases, due to the variant characteristics of web images and texts, images usually function as the supplementary materials of textual documents. When users access a webpage, they usually have a rough idea of textual information, then watch the images to acquire the more detailed and vivid information. For example, in an web page P_1 containing textual document D_1 and image I_1, P_2 containing textual document D_2 and image I_2, if P_1 and P_2 are describing the same topic or similar information, D_1 and D_2 should contain similar keyword to show they describe the same content, and I_1 and I_2 can be visually similar, semantically similar or topic similar, as illustrated in Fig. 2.

In this paper, we propose a method to learn image similarities by considering the relations of the accompanied textual documents, which take regard of the fact that images are usually accompanied with contextual textual information. These two heterogeneous modalities are mutually reinforcing to make the Internet content more informative, and in most cases, visual information can be regarded as an enhanced content of the textual document. Based on these observations, this work uses the relation of text-to-text similarity as the learning evidence to guide image-to-image similarity learning. However, the quantitative value of textual document similarity is not reliable enough to indicate the accompanied image similarities due to the heterogeneous characteristic properties of the two modalities. To make image-to-image similarity being more appropriately consistent with document-to-document similarity, rank-based learning procedure is adopted by employing the method of structural SVM [24, 25], which applies the relative ranking sequence of the samples to get the large margins of the training points. In this work, the ranking structure is established by comparing the relative relations of textual information based on the bag-of-word descriptor. Compared with the label-based and quantitative ground truth-based learning method, the learning results are in more accordance with the human's recognition on web pages and experimental evaluations validate this. The proposed method in this paper can be used not only for the image-to-image retrieval, but also for cross-modality multimedia retrieval. Here, we propose a method of cross-media query expansion to make the retrieval results more satisfactory, which not only considers the cross-modality similarities relations, but also take regards of the important factors of the top-ranked results. The contributions of this paper can be summarized as follows:

1. A new Internet image similarity learning method is proposed by considering the relative relations of the contextual textual documents;
2. The method of ranking-based large margin metric learning method is used to more appropriately model the Internet image similarities in the Internet web pages;
3. A cross-media query expansion method is proposed to more satisfactorily achieve the cross-media retrieval based on the similarity learning results.

The rest of the paper is structured as follows. In Sect. 2, we review the related works and present the overview of the proposed method. In Sect. 3, we describe the relative image similarity learning method. In Sect. 4, the cross-media retrieval based on query expansion is introduced. We report the evaluation results in Sect. 5 and conclude this paper with future work in Sect. 6.

2 Related work

As Internet image data are growing rapidly, many image annotation methods for Internet images were proposed in recent years. Wang et al. [26] first searched for similar images from the web by mining representative and common descriptions from the surrounding descriptions of similar images as the annotation for the query image. In [9], the authors propose a novel kNN-sparse graph-based

semi-supervised learning approach for simultaneously harnessing the labeled and unlabeled data, which exploits the problem of annotating a large-scale image corpus by label propagation over noisily tagged web images. Harchaoui et al. [27] introduce a new scalable learning algorithm for large-scale multi-class image classification, based on the multinomial logistic loss and the trace-norm regularization penalty. Using stochastic gradient descent optimization, the work in [28] can scale the learning to millions of images and thousands of classes.

Image tagging is widely used in Internet image sharing website as users are relatively easy to generate, manage and search web images through tags. In [17], the authors establish a large-scale automated photo tagging method by exploring the social images and present a retrieval-based approach for automated photo tagging. Wu et al. [22] propose a machine learning framework for mining social images and investigate its application to automated image tagging. We can use image tagging information to improve the image understanding abilities as in [18], in which they propose a novel multi-task multi-feature metric learning method which models the information sharing mechanism among different learning tasks. And their method is capable of simultaneous learning with semantic information and social tagging based on the multi-task learning framework, thus they both benefit from the information provided by each other.

Social images always contain tags and other textual descriptions, so cross-media retrieval methods are useful to improve image understandings. Zhang et al. [29] present the idea of semantic propagation based on relevance feedback and cross-modality query expansion that improves the retrieval performances when a query is initialized either with a keyword or an example image. The authors in [30] propose a cross-modality model, which does not require all of the training data to be pairs of corresponding image and text. As a result, their model can exploit the availability of only a few image-text pairs, together with image–image and text–text similarities, to learn the intrinsic relationships between images and text.

Image similarity learning methods become more and more comprehensive, which contain not only image visual similarity but also semantic similarity. Hwang et al. [21] learn a corresponding tree of metrics (ToM) with disjoint visual features using a hierarchical taxonomy that captures semantic similarity between the objects. They combine the idea of ToM and disjoint sparsity together as a new strategy for visual feature learning. Verma et al. [23] propose a novel framework to learn similarity metrics using the class taxonomy, and their metrics can also help determine the correct placement of a new category that was not part of the original taxonomy.

2.1 Our approach

Different with the above related works, this paper presents a method to learn a distance metric to compute image similarities under Internet environments, where images are usually contained in the web pages with textual documents. The similarity learning method is based on the ground truth of ranking sequence instead of category labels. Given an input image sample, its ranking structure is used as the characteristics for learning. In this paper, we adopt the comparisons of textual documents as the ranking sequences which are more reliable and can well reflect the similarity of web pages. Thus, the similarities of images in the web pages are adjusted to be more suitable for the Internet cross-media applications. Structural SVM is used in the similarity learning where the single slack formulation is adopted. Based on the learned metrics, a cross-media query expansion method is proposed to find more relevant web pages according to the input image with the top returned results. This method reduces the influence of the differences of input and output and establishes a solution to measure the similarity of the two modalities. Figure 3 illustrates the framework of this method.

3 Relative image similarity learning based on textual relations

In this section, we first describe the basic idea of similarity learning, then introduce the rank-based metric learning method [25] based on structural SVM [24]. After that, ranking structure of textual document is introduced. At the end, we describe the learning procedure in this work.

3.1 Image similarity learning

The solutions of image similarity learning (ISL) are usually based on the mahalanobis metric, which computes the dissimilarity extent of two samples in the input data space. The goal of ISL is to preserve the distance relations of training data. For two sample points $I_i \in \mathbb{R}^d$, $I_j \in \mathbb{R}^d$, the learned manhalanobis distance metric $d_{\mathrm{M}}(I_i, I_j)$ is usually to learn a positive semi-definite covariance matrix $M \in \mathbb{R}^{d\times d}$ in the form:

$$d_{\mathrm{M}}(I_i, I_j) = \left\|I_i - I_j\right\|_{\mathrm{M}} = \sqrt{(I_i - I_j)^T M (I_i - I_j)}. \quad (1)$$

This can be regarded as a linear projection of the samples, which transforms the sample space into another Euclidean space to make the similarity computing results more satisfying according to the provided training samples, and Eq. (1) can be rewritten as:

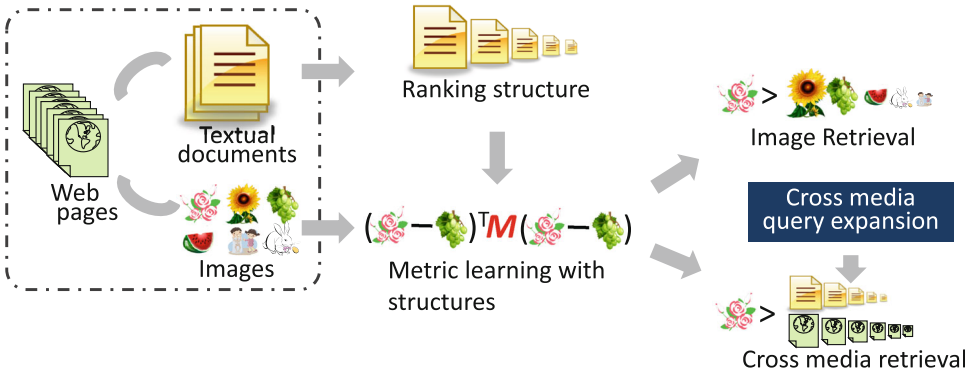


Fig. 3 Illustration of the proposed method

$$d_{\mathrm{M}}(I_i, I_j) = \|I_i - I_j\|_{\mathrm{M}} = \sqrt{(WI_i - WI_j)^T (WI_i - WI_j)},$$
$$M = W^T W. \tag{2}$$

In the literature, most of the learning methods are based on the category-based training sample such as LMNN [31], NCA [32], where samples in the same category should be near and those in different categories should be far. However, in the real conditions, image similarity is far more complex than the simple semantic categorizations. An image contains many semantic explanations, and different semantic concepts may have different importance in one image. In some cases, images in the same category may not as similar as images from different categories. Moreover, in the Internet images, the similarity of two images relies on the contextual constraints. Two images both containing the same context may belong to two different web pages which describe totally different topics. In image similarity learning, the idea of relative relations can be adopted, and the contextual information around the images can be of much helpful to obtain a more reliable similarity measure for Internet applications.

3.2 Rank-based metric learning

In the machine learning community, distance metric learning methods based on relative relations can be categorized into two types. The first type is based on pair-wise relations, such as [33–35], and the second is based on the sequential ranking-based structures, such as [25, 36]. In this work, we adopt the ranking-based learning strategy as it is more suitable for the retrieval tasks as well as for the Internet applications with multiple modalities.

To learn a mahalanobis distance metric, the samples are usually in the form $X = \{S_1, S_2, \ldots, S_n\} = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots (\alpha_n, \beta_n)\}$. For each sample S_i, α_i is the description of the sample, β_i is the response variables of α_i, usually β_i is a class name or the sample labels. Based on the given training samples with given response, various training solutions can be established to obtain the mahalanobis matrix. Ranking-based similarity learning assumes the output response β_i as the ranking sequence of α_i with descending similarities:

$$\beta_i = \{\alpha_{i\_1}, \alpha_{i\_2}, \ldots, \alpha_{i\_m}\}\ \alpha_{i\_1} \neq \alpha_i$$
$$\forall 0 < x < y \leq m\ \mathrm{sim}(\alpha_i, \alpha_{i\_x}) \leq \mathrm{sim}(\alpha_i, \alpha_{i\_y}). \tag{3}$$

Rank-based metric learning method [25] uses structural SVM [24, 37], whose output β can be structure such as ranking sequence, trees, or graphs. The goal of structural SVM is to solve the optimization problem in a similar way as classical SVM [37].

3.3 Classical SVM (β_i is class label of α_i)

$$\min_{w, \xi_i \geq 0} \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$
$$s.t. \forall i \in \{1, \ldots, n\} : \beta_i (w^T \alpha_i) \geq 1 - \xi_i \tag{4}$$

where w is the normal vector to the hyperplane, ξ_i is the slack variables for the soft margin, and $C > 0$ is a constant to balance the loss function and margin maximization.

3.4 Structural SVM (β_i is ranking sequence of α_i)

$$\min_{w, \xi_i \geq 0} \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$
$$s.t. \forall \beta_1^* \in \mathrm{B} : w^T [\psi(\alpha_1, \beta_1) - \psi(\alpha_1, \beta_1^*)] \geq \Delta(\beta_1, \beta_1^*) - \xi_1$$
$$\cdots$$
$$s.t. \forall \beta_n^* \in \mathrm{B} : w^T [\psi(\alpha_n, \beta_n) - \psi(\alpha_n, \beta_n^*)] \geq \Delta(\beta_n, \beta_n^*) - \xi_n. \tag{5}$$

Here we use the optimization problem with the margin-rescaling instead of slack rescaling [24, 37] because this can generate approximate results based on the subset selection procedure of the output samples. B is the output structural space, $\psi(\alpha, \beta)$ is a feature map vector to characterize the relationship of input α and output structure β. The loss function $\Delta(\beta_i, \beta_i^*) : \mathrm{B} \times \mathrm{B} \rightarrow \mathbb{R}$ quantifies the loss between the prediction β_i^* and the true output value β_i,

which satisfies: $\Delta(\beta_i, \beta_i^*) = 0$ if $\beta_i = \beta_i^*$, and $\Delta(\beta_i, \beta_i^*) > 0$ if $\beta_i \neq \beta_i^*$.

Learning large margins from structural output responses can be regarded as a generalization of multiclass SVM [38]. Instead of separating multiclass problem into several binary classification problems, multiclass SVM uses one optimization procedure. The difference of structural SVM and multiclass SVM is the output generalized from labels to sequential or tree structures, which is mechanically more suitable for the retrieval tasks. That is the reason why this work adopts structural-based learning method to characterize the distance relation of Internet images. As there are a very large number of output sets in structural SVM, the algorithm needs to deal with a lot of margin inequalities as shown in Eq. (5). This may lead to current optimization solutions computationally intractable in implementation. To solve this problem, two kinds of methods can be considered [37]. The first one is to reformulate the problem to a polynomial format. The second is to choose a polynomial-sized output subset from the original problem, which can be efficiently established based on a cutting plane model to optimize the discriminative function weight *w*within prescribed tolerance [24, 37].

In the rank-based metric learning method, the goal is to learn an optimal metric M-based on the training samples $\mathrm{X} = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots(\alpha_n, \beta_n)\}$, where α_i is the input query and β_i is the possible rankings for α_i. Based on the maximum margin method of structural SVM, the metric M can be obtained by implementing the optimization procedure.

3.5 Ranking-based metric learning

$$\min_{w,\xi \geq 0} tr(M) + C\xi$$
$$s.t.\ \frac{1}{n}\sum_{i=1}^{n} \left\langle M, \psi(\alpha_i, \beta_i^*) - \psi(\alpha_i, \beta_i) \right\rangle_F \geq \frac{1}{n}\sum_{i=1}^{n} \Delta(\beta_i^*, \beta_i) - \xi \tag{6}$$

In this formulation the normal vector to the hyperplane is changed to the metric M, and this transforms the classification problem to the distance metric learning problem. On the other hand, the use of $tr(M)$ instead of $\frac{1}{2}tr(M^T M)$ is to generate sparse results [25]. The metric learning algorithm uses the "1-slack formulation" with a single ξ analogous to the "multi-slack" to computing efficiency with cutting plane methods [37].

3.6 Ranking structure of textual document

Suppose we have a set of web pages $\mathrm{Z} = \{P_1, P_2, \ldots, P_n\}$. For each web page P_i, it contains both textual document D_i and image I_i. For simplicity, we assume W_i only has one document and one image. In the following description, it could be observed that it can be processed in a similar way for multiple cases. We denote the document set in Z as $\mathrm{T} = \{D_1, D_2, \ldots, D_n\}$ and image set as $\mathrm{I} = \{I_1, I_2, \ldots, I_n\}$. For each document D_i, its feature can be represented as $Fea(D_i)$, then the similarity of two textual documents D_i and D_j can be computed as $\mathrm{sim}(D_i, D_j)$. Thus, given a document query D_i, the similarity ranking order of D_i is $ro(D_i) = \{D_{i\_1}, D_{i\_2}, \ldots, D_{i\_m}\}$, $m \leq n-1$, which satisfies:

$$\forall 0 < x < y \leq m\ \mathrm{sim}(D_i, D_{i\_x}) \leq \mathrm{sim}(D_i, D_{i\_y}). \tag{7}$$

Based on these settings, the notations in rank-based learning are changed as follows and we can derive the image similarities by the structural-based distance metric learning method [25].

$$\begin{aligned} X &= \{S_1, S_2, \ldots, S_n\} = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots(\alpha_n, \beta_n)\} \\ &\rightarrow \{(I_1, \beta_1), (I_2, \beta_2), \ldots(I_n, \beta_n)\} \\ &= \{(I_1, ro(D_1)), (I_2, ro(D_2)), \ldots(I_n, ro(D_n))\} \end{aligned} \tag{8}$$

3.7 The learning procedure

From Eqs. (6) and (8), the optimization problem can be written as:

$$\min_{w,\xi \geq 0} tr(M) + C\xi$$
$$\begin{aligned} s.t.\ \frac{1}{n}\sum_{i=1}^{n} &\left\langle M, \psi(I_i, ro(D_i^*)) - \psi(I_i, ro(D_i)) \right\rangle_F \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \Delta(ro(D_i^*), ro(D_i)) - \xi \end{aligned} \tag{9}$$

To solve this problem, we need to decide the settings of the feature map ψ, the loss function Δ and the cutting plane solutions [25]. For the feature map ψ, the widely used partial order feature is adopted as in [25, 39–41], which is defined in Eq. (10) and can be explained as the summation of the feature differences of relevant and non-relevance image pairs.

$$\begin{aligned} \psi_{po}(I, ro(D)) = \frac{1}{\left|\Gamma_I^+\right| \cdot \left|\Gamma_I^-\right|} \sum_{I' \in \Gamma_I^+} \sum_{I'' \in \Gamma_I^-} rel(I', I'')(\phi(I, I') \\ - \phi(I, I'')) \end{aligned} \tag{10}$$

where $\left|\Gamma_I^+\right|$ and $\left|\Gamma_I^+\right|$ represent the relevant and non-relevant images for the query I, respectively, $\phi(I_1, I_2)$ is the relations of image I_1 and I_2, which is used as the inverse Euclidian distance: $\phi(I_1, I_2) = -\|I_1 - I_2\|^2$ to follow the ordering relations of $rel(I_1, I_2)$, which is defined as:

$$rel(I_1, I_2) = \begin{cases} +1, & \text{the partial order places } I_1 \text{ before } I_2 \text{ in } ro(D) \\ -1, & \text{the partial order places } I_1 \text{ after } I_2 \text{ in } ro(D) \end{cases}. \tag{11}$$

The separation oracle approach [25, 37] is used for the cutting plane solutions to find the structure $ro(D)$ with most violated margin constraint (Eq. (12)), which changes from multi-slack to the 1-slack formulation as discussed earlier. The goal of this kind of operation is to find the most suitable structures from numerous candidates in the distance metric learning procedure.

$$\begin{aligned} ro(D) \leftarrow & \arg\max\{\Delta(ro(D_i), ro(D)) \\ & - \langle W, \psi_{po}(I_i, ro(D_i)) - \psi_{po}(I_i, ro(D))\rangle_F\} \\ = & \arg\max\{\Delta(ro(D_i), ro(D)) \\ & - \langle W, \psi_{po}(I_i, ro(D)\rangle_F\} \end{aligned} \tag{12}$$

Another important issue is the determination of the ranking loss function, which allows the learning algorithms to incorporate specific performance measures including mean average precision (MAP), precision-at-k, AUC area, mean reciprocal rank (MRR), normalized discounted cumulative gain (NDCG), etc. [25, 39–41]. In Internet search applications, users are most interested in the first few returned results, and they more care about the relatedness of these results. The percentage of the relevant items among the first returned results reflects the quality of the search engine. The localized evaluation measure of precision-at-k can directly satisfy this, where k is the top returned results. In this work, we use precision-at-k as the ranking measures to compute the loss function of $\Delta(ro(D_i), ro(D))$ in learning the distance metric.In the optimization procedure, the learning algorithm alternatively optimizes the metric M and the ranking constraint set of the ranking structure. The algorithm will stop when there is no constraint violating a prescribed tolerance.

4 Cross-media retrieval based on query expansion

Query expansion has been widely used in textual information retrieval. The idea is to expand the search query besides the original one to find more relevant documents. If the input query is clear and specific enough to describe the user's required information, the retrieval system would be more likely to return satisfactory results. However, in real applications, the input queries are always ambiguous, which may lead to error and omission results. In the scenario of Internet cross-media retrieval, whose task is to input the query with one modality, returning results may be obtained with other modalities. The most common input is textual information, which can more clearly show the intention of the users. However, if the query is other modalities such as image, or audio, the ambiguity information may lead to unsatisfactory results using simple retrieve model. The strategy of query expansion can enhance the underrepresented aspects in the original query, so the query ambiguity can be alleviated to some extent in this way. Thus, the method of cross-media query expansion could be an effective solution to bridge the modality gaps and intention gaps in the Internet multimedia retrieval applications.

In this paper, we propose a method of cross-media retrieval based on the image similarity learning results. As the textual input is common in the current search services, in this task, we consider the input as image data and the outputs can be web pages or textual documents. Given an input image query I, we want to retrieve a sequence of relevant web pages $\{P_1, P_2, \ldots, P_k\}$which are ranked with similarity descending order. Each web page at least contains one textual document and may possibly contain images, which is the usual appearance in most of the website. One important issue in cross-media retrieval is to derive the similarities among different modalities, for example the similarity of an image I and web page P: $\text{sim}(I, P)$ or the similarity of an image I and a textual document D: $\text{sim}(I, D)$. There are two kinds of cross-media similarity computing method. One is the transitive way, and the other is the direct way. The transitive way assumes all the web pages contain both text and image, so the similarity of two images can be transferred to the similarity of image and webpage (text document). The direct way needs to obtain a direct computing method to measure the similarity of two modalities. This way does not need the strict requirement of web pages both containing text and image. However, this kind of solution involves with sophisticated issues of how to map the two modalities of information into one metric space, so its reliability needs further being validated. In this work, the cross-media retrieval is realized through the strategy of query expansion. In this way, the cross-media similarity is performed in the transitive way, and this method does not need the strict requirement of the web page as described before.

In the cross-media retrieval based on the query expansion, the input is image I_q, and the web page dataset $\text{Z} = \{P_1, P_2, \ldots, P_n\}$. The goal is to find the web pages $\{P_{o-1}, P_{o-2}, \ldots, P_{o-k}\}$ that are similar with the I_q in the descending order. For the web page dataset Z, there exists an image set I_Z, which satisfies: $\forall I \in \text{I}_\text{Z}$, $\exists P \in \text{Z}$, image Ibelongs to the web page P. All the web pages containing images in I_Z is represented as Z^{img}, $Z^{\text{img}} \in \text{Z}$. We denote $\text{sim}(I_q, I)$ as the similarity of two images. By following the similarity transitive way, we denote $\text{sim}(I_q, P) = \text{sim}(I_q, D) = \text{sim}(I_q, I)$, where D is the textual document that co-occurs with the image I in P. First, we compute the

similarities of I_q with the images in $\mathrm{I_Z}$ based on the similarity learning method described above.

$$\mathrm{sim}(I_q, I) = 1 - \log(d_{\mathrm{M}}(I_q, I)) \tag{13}$$

where M is the learned metric, and the log function is used to normalize the similarity computing. By following the transitive cross-media computing way, the image-to-image similarity can be transferred to the image to web page similarity of Z^{img}. The algorithm selects top k_1 images in $\mathrm{I_Z}$ that are most similar with I_q, $k_1 < n$. Then there are k_1 web pages in Z^{img} that are most similar with I_q, represented as Z^{img}_{k1}. The textual documents in the k_1 web pages are represented as $\mathrm{T}^{\mathrm{img}}_{k1}$, and we use these textual documents in $\mathrm{T}^{\mathrm{img}}_{k1}$ as the input queries to extract more relevant web pages. Given a document $\tilde{D} \in \mathrm{T}^{\mathrm{img}}_{k1}$, compute the similarities of $\tilde{D}$ with the document in T, which satisfies $\forall D' \in \mathrm{T}, \exists P' \in Z, D'$ is a document in web page P', thus $\mathrm{T}^{\mathrm{img}}_{k1} \in \mathrm{T}$. The similarity of $\tilde{D}$ and D' $(\mathrm{sim}(\tilde{D}, D'))$ can be computed in the way of the textual document similarity computing. In this work, we adopt the bag of word representation model and L_2 distance norm to compute similarities. Suppose $\mathrm{sim}(I_q, \tilde{D}) = \mu_1$, $\mathrm{sim}(\tilde{D}, D') = \mu_2$, the similarity of query image with all other textual documents and web pages can be obtained:

$$\mathrm{sim}(I_q, D') = \mathrm{sim}(I_q, P') = \mu_1 \times \mu_2 \tag{14}$$

$\forall P'' \in Z^{\mathrm{img}}$, which contains an image I'', suppose the direct similarity of the query with this web page is $\mathrm{sim}(I_q, I'') = \mathrm{sim}(I_q, P'') = \mu_3$. If $P' = P''$, the similarity of the query with this web page is defined as:

$$\mathrm{sim}(I_q, P') = \mathrm{sim}(I_q, P'') = \max\{\mu_1 \times \mu_2, \mu_3\}. \tag{15}$$

For $\vec{D}' \in \mathrm{T}$, $\overrightarrow{P}' \in Z$, $\overrightarrow{D}'$ is a document in web page $\overrightarrow{P}'$, and $\vec{D}' \in \mathrm{T}^{\mathrm{img}}_{k1}, \tilde{D}' \neq \tilde{D}$, suppose $\mathrm{sim}(I_q, \tilde{D}') = \vec{\mu}_1, \mathrm{sim}(\tilde{D}', \vec{D}') = \overrightarrow{\mu}_2$, then $\mathrm{sim}(I_q, \overrightarrow{D}') = \mathrm{sim}(I_q, \overrightarrow{P}') = \overrightarrow{\mu}_1 \times \overrightarrow{\mu}_2$. If $P' = \overrightarrow{P}'$, the similarity of the query with this web page is defined as:

$$\mathrm{sim}(I_q, P') = \mathrm{sim}(I_q, \overrightarrow{P}') = \max\{\mu_1 \times \mu_2, \overrightarrow{\mu}_1 \times \overrightarrow{\mu}_2\} \tag{16}$$

After the above procedures, $\forall P \in \mathrm{Z}$, $\mathrm{sim}(I_q, P)$ can be obtained. The retrieval algorithm returns the top k similar web pages $\{P_{o-1}, P_{o-2}, \ldots, P_{o-k}\}$ with the query I_q as the retrieval results. Through this algorithm, all the web pages can be extracted without establishing the direct similarity computing method of image and web page, and the influence of input ambiguities can be reduced based on the procedure of top similar textual document extension, which transfers the image similarity to document similarity and leads to more reliable results. So the effectiveness of cross-media interactions can be realized.

5 Performance evaluation

To evaluate the performance of the proposed methods, we use the Image CLEF public dataset [42], where the web pages are collected from the Wikipedia website. This website is a free information sharing site and it contains plenty of textual documents and images, which can well indicates the properties of most of the multimedia websites. Image CLEF dataset include more than 200,000 web pages, and each web page contains both document and image. Part of the web pages with document and text pairs are selected as the similarity learning set, whose page number is more than 20,000. The other part of the dataset is selected as the test set to evaluate the performance of both similarity learning method and cross-media query expansion method. The number of testing samples around 180,000. Two kinds of image features are used; one is the LBP [43], which is powerful texture feature by comparing the differences of nearby pixels and using local binary code to represent the patterns, the feature dimension is 256. The other feature is color moment, which is a traditional color feature and has been used for a long time. The feature is computed based on the image blocks and its dimension is 384. To validate the performance of the proposed method, we use precision-at-k as the evaluation measure for two reasons. The first is that in the metric learning procedure, the parameter of precision-at-k is adopted. The second reason is that for large scale Internet multimedia retrieval, this evaluation measure could well reflect the effectiveness of the search system. The ROC and MAP evaluation need the information of recalling from the true positives. However, in the web scale dataset, it would not be possible to determine all the relevant true samples with the query. In the following part of this section, we will provide the experimental evaluations on similarity learning method and cross-media query expansion method, respectively.

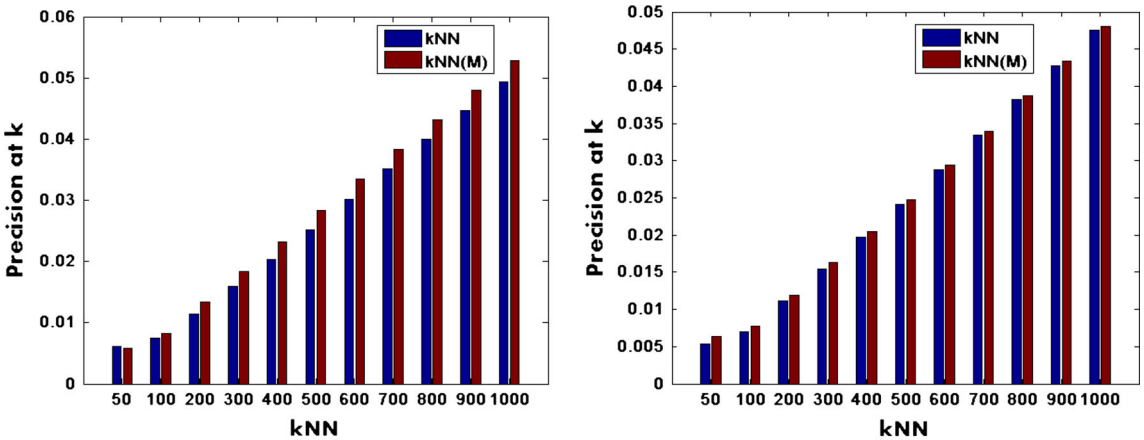
5.1 Performance of image similarity learning

This procedure is realized based on the distance metric learning according to the ranking structures. We use the traditional k nearest neighbor (KNN) method as the evaluation baseline. We compute the precision score of both KNN method using distance metric learning *(KNN(M))* and KNN without using distance metric learning (*KNN*). Experimental results on both features of LBP and color moment are illustrated in Fig. 4. It can be observed that *KNN(M)* outperforms *KNN* on both features. The effectiveness of the ranking-based metric learning can be

validated, because this learning technique considers the contextual information in the Internet by including the ranking relations of textual document. Through this way, the polysemia characteristics of visual content can be adjusted based on the contextual information and the relations on image similarities are more consistent with Internet applications. As the ground truth of ranking-based distance metric learning is different with the category-based learning, the category-based information is not available on the large scale Internet applications. So we did not compare with the other distance metric learning method. In this experiment, the parameter k is selected from 50 to 1,000. The larger k is, the better performance is achieved. This can be explained that the system will retrieve more relevant samples when k grows on both of the *KNN(M)* and *KNN*. The performance of *KNN(M)* remains steady compared with that of *KNN* when k changes, and this reveals the robustness of the metric learning method. From Fig. 4, it could be observed that color feature performs better than texture feature, not only on the quantitative precision, but also on the improvement of *KNN(M)* compared with *KNN*. The reason for this phenomenon is that the global color descriptors are usually more related with the content of the web pages. However, the LBP feature may represent the structural or textural patterns of the image, which may lead to the fact that this feature is less consistent with the contextual semantic information on the Internet.

5.2 Cross-media query expansion

In this part, the retrieval performance using cross-media query expansion will be evaluated. In this experimentation, k represents the number of final retrieved results and k_1 represents the number of selected expansion queries. Tables 1 and 2 provide the detailed results using color moment feature and Tables 3 and 4 provide the results using LBP feature. The difference of Table 1 (Table 3) with Table 2 (Table 4) is the expansion selection strategies on k_1: top n % (top n). Top n % means that the selected expansion query number is the top returned k_1 samples: $k_1 = k \times n\%$ and top n means $k_1 = n$. In Tables 1, 2, 3, and 4, the second row is the results without expansions, and this is the experimental baseline. The other rows are the results with query expansion. Experimental results using both features and various parameters in Tables 1, 2, 3, and 4 clearly reveal that query expansion improve the performance of cross-media retrieval. It can also be observed that top n % strategy performs better than the top n strategy from $k = 50$ to $k = 1{,}000$, because top n % strategy includes more expansion queries that enlarge the candidate relevant samples. From these experiments, we could also find that the larger k_1 is, the better performance of cross-media retrieval is. This phenomenon also coincides with the experimental results of expansion query selection strategy. Figure 5 provides the summarization performance of query expansion, where *KNN(M)* represents no query expansion with the learned metric, *KNN(M)-QE-Top50* represents query expansion with top 50 expansion selection and *KNN(M)-QE-Top50 %* represents query expansion with top 50 % expansion selection. Only top 50 and top 50 % are chosen because they perform best compared with other parameters. Figure 5a shows the results using color feature and Fig. 5b shows the results using texture features. It could be observed that color feature performs better without using query expansion. Texture feature is better with the cross-media query expansion method, where more improvements are achieved with top 50



(a) color moment feature **(b)** LBP feature

Fig. 4 Illustration of the performance of the image similarity learning method

Table 1 Query expansion using color moment feature with top n % expansion selection

k	k_1					
	0	10 %	20 %	30 %	40 %	50 %
50	0.0058611	**0.005932**	**0.005932**	**0.005932**	**0.005932**	0.0059288
100	0.0082584	0.0082998	0.0082998	0.0082987	0.0083003	**0.0083019**
200	0.013341	0.013374	0.01339	0.013392	0.01342	**0.013485**
300	0.018341	0.018358	0.018374	0.018498	0.01851	**0.018513**
400	0.023139	0.023201	0.02325	0.023361	0.023365	**0.023484**
500	0.028272	0.028325	0.028487	0.028504	0.028626	**0.028772**
600	0.033454	0.033505	0.033693	0.033799	0.033899	**0.034058**
700	0.038361	0.038414	0.038606	0.038746	0.038964	**0.039312**
800	0.043211	0.043324	0.043467	0.043676	0.043878	**0.04462**
900	0.048023	0.048251	0.048385	0.048654	0.04921	**0.049387**
1000	0.052802	0.053029	0.05317	0.05345	0.054189	**0.054195**

Best results for each k value are highlighted in bold

Table 2 Query expansion using color moment feature with top n expansion selection

k	k_1					
	0	10	20	30	40	50
50	0.0058611	0.005932	0.005932	0.0059342	0.005932	**0.0059363**
100	0.0082584	0.0082998	0.0082998	0.0082987	0.0083003	**0.0083019**
200	0.013341	0.013374	0.013374	0.013374	0.01339	**0.013392**
300	0.018341	0.018359	0.018359	0.018358	0.018371	**0.018374**
400	0.023139	0.023191	0.023191	0.023188	0.023201	**0.023204**
500	0.028272	0.028316	0.028317	0.02831	0.028323	**0.028325**
600	0.033454	0.033496	0.033497	0.033491	0.033505	**0.033507**
700	0.038361	0.038397	0.038398	0.038391	0.038406	**0.038408**
800	0.043211	0.043246	0.043247	0.043239	0.043253	**0.043256**
900	0.048023	0.048054	0.048055	0.048041	0.048056	**0.048059**
1000	0.052802	0.052824	0.052825	0.05281	0.052826	**0.052829**

Best results for each k value are highlighted in bold

Table 3 Query expansion using LBP feature with top n % expansion selection

k	k_1					
	0	10 %	20 %	30 %	40 %	50 %
50	0.0058611	0.005932	0.005932	0.0059342	0.005932	**0.0059363**
100	0.0082584	0.0082998	0.0082998	0.0082987	0.0083003	**0.0083019**
200	0.013341	0.013374	0.013374	0.013374	0.01339	**0.013392**
300	0.018341	0.018359	0.018359	0.018358	0.018371	**0.018374**
400	0.023139	0.023191	0.023191	0.023188	0.023201	**0.023204**
500	0.028272	0.028316	0.028317	0.02831	0.028323	**0.028325**
600	0.033454	0.033496	0.033497	0.033491	0.033505	**0.033507**
700	0.038361	0.038397	0.038398	0.038391	0.038406	**0.038408**
800	0.043211	0.043246	0.043247	0.043239	0.043253	**0.043256**
900	0.048023	0.048054	0.048055	0.048041	0.048056	**0.048059**
1000	0.052802	0.052824	0.052825	0.05281	0.052826	**0.052829**

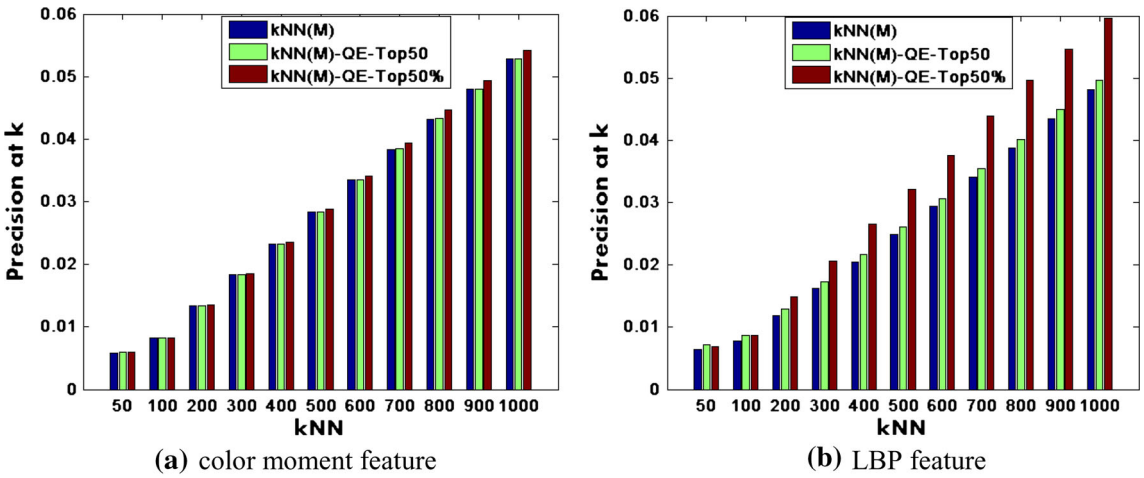
Best results for each k value are highlighted in bold

compared with no expansion. The reason for this may due to that LBP feature with local structure patterns is more discriminative when selecting more relevant samples using the expanded queries, which can be regarded as complementary visual information with color feature in Internet multimedia retrieval. So in designing the retrieval system, the first round retrieval can select color feature and the second expansion step could use LBP features.

Table 4 Query expansion using LBP feature with top n expansion selection

k	k_1					
	0	10	20	30	40	50
50	0.0064304	0.0067225	0.006831	0.005932	0.0069416	**0.0071167**
100	0.0078342	0.0081891	0.008253	0.0083529	0.0084952	**0.0087068**
200	0.011913	0.012312	0.012408	0.012531	0.012698	**0.012931**
300	0.016258	0.016678	0.016791	0.016939	0.017118	**0.017367**
400	0.020485	0.020945	0.021072	0.021228	0.021421	**0.021696**
500	0.024818	0.025277	0.025423	0.025598	0.025805	**0.0261**
600	0.029362	0.029813	0.029966	0.030141	0.030363	**0.030677**
700	0.034021	0.03446	0.034622	0.034808	0.035053	**0.035388**
800	0.038742	0.039172	0.039342	0.03954	0.039803	**0.04016**
900	0.043433	0.043856	0.044033	0.044232	0.044524	**0.044897**
1000	0.048137	0.048538	0.048719	0.048922	0.049254	**0.049637**

Best results for each k value are highlighted in bold



(a) color moment feature **(b)** LBP feature

Fig. 5 Illustration of the performance with the query expansion method

From the above analysis, we could draw the conclusion that the proposed query expansion method is an effective solution to improve the query expansion performance in cross-media retrieval.

6 Conclusion

There exist numerous images in the Internet. How to better utilize image information by fully considering the contextual environment and fulfilling the Internet applications is an important problem. In this paper, an image similarity learning method is proposed by the ranking-based structures. The method establishes the training ground truth of ranking relations based on the co-occurred textual document. The image relations can be more consistent with the web pages which usually contain both textual and visual information. The learning technique uses structural SVM and adopts the cutting plane strategy to make the learning process being realized with a large number of output samples. Based on the similarity learning method, the technique of cross-media query expansion is proposed to alleviate influence of query ambiguities and improve the retrieval performance. In the future, we will investigate more techniques on image relation structures by considering more social factors to obtain more suitable image similarity computing method under the Internet environment.

Acknowledgments This work was supported in part by National Basic Research Program of China (973 Program):2012CB316400, in part by National Natural Science Foundation of China: 61070108, 61025011, and 61035001, in part by the Key Technologies R&D Program of China under Grant no. 2012BAH18B02.

References

1. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp. 524–531 (2005)

2. Li, L., Jiang, S., Huang, Q.: Learning hierarchical semantic description via mixed-norm regularization for image understanding. IEEE Trans. Multimedia **14**(5), 1401–1413 (2012)
3. Tang, J., Zha, Z.-J., Tao, D., Chua, T.-S.: Semantic-gap oriented active learning for multi-label image annotation. IEEE Trans. Image Process. **21**(4), 2354–2360 (2012)
4. Wang, S., Huang, Q., Jiang, S., Tian, Q.: S3MKL: scalable semi-supervised multiple kernel learning for real world image applications. IEEE Trans. Multimedia **14**(4), 1259–1274 (2012)
5. Wang, M., Hua, X., Hong, R., Tang, J., Qi, G., Song, Y.: Unified video annotation via multi-graph learning. IEEE Trans. Circ. Syst. Video Technol. **19**(5), 733–746 (2009)
6. Jiang, S., Huang, Q., Ye, Q., Gao, W.: An effective method to detect and categorize digitized traditional Chinese paintings. Pattern Recogn. Lett. **27**(7), 734–746 (2006)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
8. Tang, J., Yan, S., Hong, R., Qi, G.-J., Chua, T.-S.: Inferring semantic concepts from community-contributed images and noisy tags. In: Proceedings of ACM Multimedia, pp. 223–232 (2009)
9. Li, X., Snoek, C.G.M., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Trans. Multimedia **11**(7), 1310–1322 (2009)
10. Tang, J., Hong, R., Yan, S., Chua, T.-S., Qi, G.-J., Jain, R.: Image annotation by knn-sparse graph-based label propagation over noisily-tagged web images. ACM Trans. Intell. Syst. Technol. **2**, 2 (2011)
11. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: Proceeding of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, pp. 317–326 (2009)
12. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of ACM Multimedia, pp. 461–470 (2010)
13. Cai, D., He, X., Li, Z., Ma, W.-Y., Wen, J.-R.: Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proceedings of ACM Multimedia, pp. 952–959 (2004)
14. Gao, B., Liu, T.-Y., Qin, T., Zheng, X., Cheng, Q.-S., Ma, W.-Y.: Web image clustering by consistent utilization of visual features and surrounding texts. In: Proceedings of ACM Multimedia, pp. 112–121 (2005)
15. Rege, M., Dong, M., Hua, J.: Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In: Proceeding of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, pp. 317–326 (2008)
16. Jin, Y., Khan, L., Wang, L., Awad M.: Image annotations by combining multiple evidence and Wordnet. In: Proceedings of ACM Multimedia, pp. 706–715 (2008)
17. Wu, L., Hoi, S.C., Zhu, J., Jin, R., Yu, N.: Distance metric learning from uncertain side information with application to automated photo tagging. In: Proceedings of ACM Multimedia, pp. 135–144 (2009)
18. Wang, S., Jiang, S., Huang, Q., Tian, Q.: Multi-feature metric learning with knowledge transfer among semantics and social tagging. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2626–2633 (2012)
19. Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., Li, S.: Flickr distance. In: Proceedings of ACM Multimedia, pp. 31–40 (2008)
20. Hoi, S.C.H., Liu, W., Lyu, M.R., Ma, W.-Y.: Learning distance metrics with contextual constraints for image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2072–2078 (2006)
21. Hwang, S.J., Grauman, K., Sha, F.: Learning a tree of metrics with disjoint visual features. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS (2011)
22. Wu, P., Hoi, S.C.H., Zhao, P., He, Y.: Mining social images with distance metric learning for automated image tagging. In: WSDM, pp. 197–206 (2011)
23. Verma, N., Mahajan, D., Sellamanickam, S., Nair, V.: Learning hierarchical similarity metrics. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2280–2287 (2012)
24. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**, 1453–1484 (2005)
25. McFee, B., Lanckriet, G.: Metric learning to rank. In: International Conference on Machine Learning, Haifa, Israel (2010)
26. Wang, X.-J., Zhang, L., Li, X., Ma, W.-Y.: Annotating images by mining image search results. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1919–1932 (2008)
27. Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., Malick, J.: Large-scale image classification with trace-norm regularization. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp. 3386–3393 (2012)
28. Perronnin, F., Akata, Z., Harchaoui, Z., Schmid, C.: Towards good practice in large-scale learning for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3482–3489 (2012)
29. Zhang, H.J., Su, Z.: Improving CBIR by semantic propagation and cross modality query expansion. In: Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MM-CBIR'01), September, pp. 83–86 (2001)
30. Jia, Y., Salmann, M., Darrell, T.: Learning cross-modality similarity for multinomial data. In: Proceedings of IEEE International Conference on Computer Vision, pp. 2407–2414 (2011)
31. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
32. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (2005)
33. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (2009)
34. Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., Belongi, S.: Generalized non-metric multi-dimensional scaling. In: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (2007)
35. McFee, B., Lanckriet, G.R.G.: Learning multi-modal similarity. J. Mach. Learn. Res. (JMLR), February, pp. 491–523 (2011)
36. Lee, J.-E., Jin, R., Jain, A.K.: Rank-based distance metric learning: an application to image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2009)
37. Thorsten, J., Finley, T., John Yu C.-N.: Cutting-plane training of structural SVMs. Mach. Learn. 77(1):27–59 (2009). ISSN 0885-6125
38. Crammer, K., Singer, Y.: On the algorithmic implementation of multi-class kernel-based vector machines. Mach. Learn. Res. **2**, 265–292 (2001)
39. Joachims, T.: A support vector method for multivariate performance measures. In: International Conference on Machine Learning, pp. 377–384 (2005)
40. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of acm

special interest group on information retrieval conference, pp. 271–278 (2007)
41. Chakrabarti, S., Khanna, R., Sawant, U., Bhattacharyya, C.: Structured learning for non smooth ranking losses. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, pp. 88–96 (2008)
42. http://www.imageclef.org
43. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)