Multimed Tools Appl (2019) 78:13169–13188
https://doi.org/10.1007/s11042-018-5767-1

Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval

Yuhua Jia[1] · Liang Bai[1] · Shuang Liu[1] · Peng Wang[2] · Jinlin Guo[1] · Yuxiang Xie[1]

Received: 30 July 2017 / Revised: 26 January 2018 / Accepted: 9 February 2018 /
Published online: 27 February 2018

Abstract Aiming at measuring the inter-media semantic similarities, cross-modal retrieval tries to align heterogenous features to an intermediate common subspace in which they can be reasonably compared. This is based on the same understanding of the semantics which are represented by different modalities. However, the semantics can usually be reflected by multiple concepts since concepts co-occur in real-world rather than occur in isolation. This leads to a more challenging task of multi-label cross-modal retrieval in which multiple concepts are annotated as labels for images as an example. More importantly, the co-occurrence patterns of concepts result in correlated pairs of labels whose relationships need to be considered in an accurate cross-modal retrieval. In this paper, we propose multi-label kernel canonical correlation analysis (ml-KCCA), a novel approach for cross-modal retrieval

Yuhua Jia and Liang Bai are both first authors.

✉ Peng Wang
pwang@tsinghua.edu.cn

Yuhua Jia
jiayuhua11@outlook.com

Liang Bai
xabpz@163.com

Shuang Liu
liushuangkd@163.com

Jinlin Guo
gjlin99@gmail.com

Yuxiang Xie
yxxie@nudt.edu.cn

[1] Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China

[2] National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

which enhances kernel CCA with high-level semantic information reflected in multi-label annotations. By kernelizing correlation extraction from multi-label information, more complex non-linear correlations between different modalities can be measured in order to learn a discriminative subspace which is more suitable for cross-modal retrieval tasks. Extensive evaluations on public datasets have validated the improvements of our approach over the state-of-the-art cross-modal retrieval approaches including other CCA extensions.

Keywords Cross-modal retrieval · Kernel CCA · Multi-label information · Concept correlations

1 Introduction

Cross-modal multimedia retrieval is especially needed in the era of Web 2.0, due to the explosive multimedia contributions in social network and media sharing websites. It is imperative to many real-world applications, e.g., to find a set of images that visually best illustrate a given text description or to find a set of sentences that textually best illustrate a given image. A large volume of multimedia is generated by users with informal content structures and various media types. To retrieve among heterogenous instances, the key problem is how to measure distances or similarities between them in a cross-modal manner. Many previous works [1, 2, 7, 8, 13, 25, 27, 32, 43] aim to align two feature spaces by learning a common latent space so that they can be reasonably comparable. Of these proposed approaches, Canonical Correlation Analysis (CCA) [9] shows its simplicity and efficiency in learning the common subspace by maximizing the correlation between the linear projections of two modalities.

While CCA has been popular for its advantages, it also suffers from several drawbacks. CCA relies on explicit pairings between two modalities to establish correspondences and in this procedure, the multi-label information still remains un-utilized. However, semantic concepts usually co-occur in real-world instead of occurring in isolation. For example, Fig. 1 demonstrates the automatic tagging of an image of horse with multiple labels through service provided by www.imagga.com. As we can see from the highlighted tags in boxes which are more representative in the exemplar image, more labels can help to interpret the content of an image, such as a horse as a foreground and grass or farm in the background. Besides the concepts which can be used to label the content of an image, the correlations between different concept pairs form another part of visual semantics. Similar to the hierarchy of concepts organized in WordNet [22] lexicon, images are also pre-organized into class hierarchies in Imagenet [4] where an image labeled with the child node class can also be categorized into its parent class. This "*is-a*" relationship is also illustrated in Fig. 1 where both "horse" and "mammal" are children of concept "animal". Relationship of "*is-part-of*" is also shown in Fig. 1 to reflect the inherent correlations of concepts of "grass" and "farm".

While it is widely accepted that incorporating the above discussed multi-label relationships can help computer to understand the semantics of multimedia, it is still challenging to quantify these correlations in fulfilling tasks for multimedia retrieval. To alleviate this, concept correlations are exploited both statistically [15–17, 20, 28, 34] from annotation sets or semantically [18, 36, 41] from knowledge bases aiming at improving multi-label tagging performances for image retrieval. The high-level semantics is also utilized in [29] for location visualization. Instead of highly relying on the annotation sets or pre-constructed knowledge bases, [37, 39] proposed a training-free method which can utilize concept correlations reflected by the underlying co-occurrence and re-occurrence patterns in

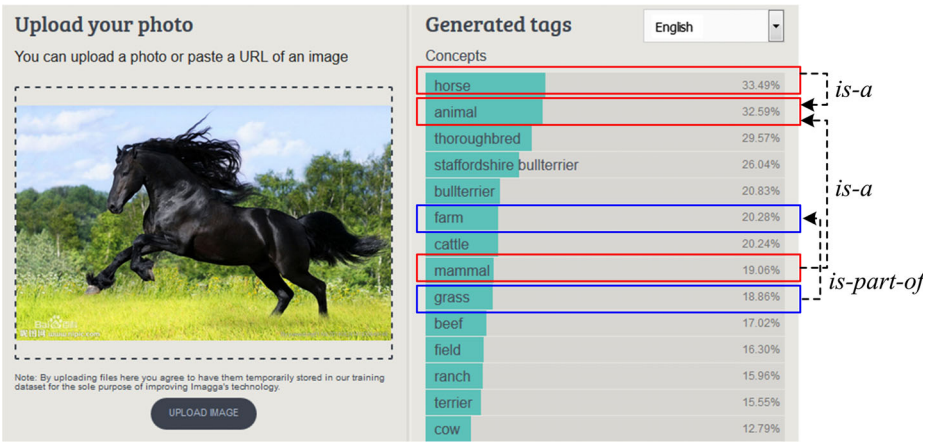


Fig. 1 Automatic tagging of an image of horse with multiply labels through service provided by www.imagga.com. Representative labels are highlighted in red and blue boxes, which are linked with concept relationships of "*is-a*" and "*is-part-of*", respectively

improving image multi-labeling performances. This method avoids the difficulty in explicitly quantifying the concept correlations through concept graph which are usually highly non-linear but is tackled through global and local pattern analysis in [37, 39]. Moreover, to retrieve images accurately, [47–49] proposed novel ranking models in which visual features and click features are simultaneously utilized to obtain the ranking model. For example, Yu et al. [49] proposed Deep-MDML method which adopts a structured ranking model to utilize both visual and click features in distance metric learning. Similar problem is also faced with when applying CCA to cross-modal retrieval as the incapacity of CCA in measuring complex non-linear correlations between different modalities can limit its performance. Since many semantic correlations cannot be simply represented in linear form, less discriminative subspaces might be constructed which cannot better suit for cross-modal retrieval tasks involving multiple labels. Though some extensions of CCA [7, 27] have been proposed to utilize label information, most of them tackle the problem of single label, i.e., assuming data sample is annotated with only one label. It is common that an image has multiple concept in presence hence should have multi-label annotations. Thus, it is imperative to take multi-label into account in order to precisely model the correlations between different modalities. Based on this rationale, Viresh Ranjan et al. proposed ml-CCA [25], an extension of CCA which outperforms most of other CCA extensions by combining high level semantics in the form of multi-label annotations. Because ml-CCA still relies on linear establishment of modality correspondences, ml-CCA will not perform consistently well when involved in more complex correlations which are difficult to be modeled linearly.

To tackle the challenge of learning complex correlations embedded in multi-label semantics, we introduce a novel multi-label Kernel Canonical Correlation Analysis (ml-KCCA) method for effective cross-modal retrieval with multi-label properties. By introducing semantic similarity matrix and embedding it into KCCA, semantic information can be utilized by the proposed method to learn a more discriminative common subspace for different modalities. Moreover, the algorithm structure is compatible with different multi-label semantics as long as they can be quantified and represented as matrices. The contributions of this paper can be summarized as follows:

- A novel cross-modal retrieval method named ml-KCCA, which accounts for multi-label information as well as utilizes kernel function to mine non-linear correlations between data from different modalities.
- A kernelized CCA with multi-label embedding is formalized in order to provide a non-linear solution for multi-label semantic correspondence estimation.
- Extensive empirical evaluation on public datasets validates our approach and shows improvement over other extensions of CCA and other state-of-the-art cross-modal retrieval approaches.

The paper is organized as follows: Section 2 discusses the most related work in areas of cross-modal information retrieval and multi-label approaches. Section 3 presents the overview of the proposed method for multi-label settings in cross-modal retrieval tasks. To deal the proposed problem, Section 4 presents a mathematical formulation and solution for the proposed ml-KCCA framework. Section 5 reports an extensive experimental evaluation on benchmark multi-label datasets to verify the efficacy of the proposed approach for cross-modal retrieval tasks. Finally, the paper is closed by conclusions and future work in Section 6.

2 Related work

Cross-modal information retrieval is a challenging research topic due to the socalled semantic gap which means that queries and their corresponding results might involve different media modalities and in such cases, the two counterparts cannot be directly compared. To tackle this challenge, a great number of approaches have been proposed to in cross-modal retrieval in the past few years, among which an effective method is to learn an optimal common representation of different modalities. This kind of method projects different modalities into a common space, in which the distance of similar semantics is minimized and the distance of dissimilar semantics is maximized. In building semantic correlation among multimodal instances, Canonical Correlation Analysis (CCA) is one of workhorses for cross-modal retrieval tasks due to its simplicity and efficiency. CCA is a method of correlating linear relationships between two multidimensional variables. It makes use of two views of the same semantic object to extract the representation of the semantics and has become one of the most popular unsupervised cross-modal subspace learning methods due to its generalization capability.

Various extensions of CCA have been proposed to emphasize different challenge aspects for the task of cross-modal retrieval in recent years [1, 2, 7, 8, 13, 25, 27]. First proposed by Hotelling [9], CCA is a method of data analysis used to discover a subspace of multiple data spaces. It is a useful method on how to seek optimal basic vectors for two sets of variables to model the multi-modal correlation. More than one canonical correlations can be found and each corresponds to a different set of basis vectors. PLS [10] aims to find a linear regression model by projecting the predicted variables and the observable variables to a new space, which is equivalent with CCA in many situations [8]. CCA can also be used as a complimentary preprocessing for other learning tasks. For example, based on the subspaces learned by CCA, Rasiwasia et al. [26] proposed to learn cross-modal topic classifiers to measure the semantic divergence of Web data. Wu et al. [42] constructed a semantic distance measurement model and Gong et al. [6] developed a binary codes learning approach which leverages the label information with CCA. More recently, Yao et al. [45] explored relative relationship by firstly finding a latent space by CCA and then re-adjusting the space to

incorporate ranking preferences from click-through data. The heterogeneous discriminative analysis of canonical correlation (HDCC) [40] utilizes discriminative information from the source domain as well as topology information from the target domain to learn two different projection matrices to discover a common feature subspace in which heterogeneous features can be compared.

However, the classic CCA ignores additional high-level semantic information which significantly limits its performance in real-world multimodal retrieval tasks. To alleviate this, Rasiwasia et al. proposed cluster-CCA [27] to incorporate high level features represented by single labels. Though demonstrated to be effective in single-label datasets in which instances have to be separated into distinct clusters, the disadvantage of cluster-CCA is obvious in multi-label scenarios because there is no natural separation of multi-label datasets into distinct clusters. To adapt CCA to multi-label settings, 3-view CCA was introduced in [7] and in this CCA variant, multi-label vectors are used as representations of high-level semantics. However, 3-view CCA highly depends on a priori correspondence information across modalities, hence cannot be directly applied to those datasets where such correspondence is not available for its requirements. Another typical extension of CCA to multi-label information is ml-CCA proposed by Viresh Ranjan et al. in [25]. Ml-CCA utilizes multi-label information while learning a common semantic space for the two modalities, and can learn a discriminative semantic space which is more suitable for cross-modal tasks. Unlike CCA, ml-CCA does not rely on explicit pairings between modalities, instead it uses the multi-label information to establish correspondences, which results in a more discriminative subspace that is better suited for cross-modal retrieval tasks. Benefited from taking multi-label information into account, ml-CCA has shown its merit and outperforms most of other extensions of CCA. However, ml-CCA fails to exploit non-linear inter-modal relationships which also limits its performance in multi-label cross-modal tasks in which the modality correspondence is usually complex and cannot precisely modeled by linear projections.

One stream of research related to multi-label semantics as this paper investigates is multi-label multimedia indexing, for which multi-label training and indexing refinement are two main approaches to utilizing multi-label information. A typical multi-label training method is presented in [24], in which concept correlations are modeled in the classification model using Gibbs random fields. Similar multi-label training methods can be found in [44]. Since all concepts are learned from one integrated model, one shortcoming is the lack of flexibility, which means that the learning stage needs to be repeated when the concept lexicon is changed. As an alternative, index refinement methods post-process detection scores obtained from individual detectors, allowing independent and specialized classification techniques to be leveraged for each concept. Context-Based Concept Fusion (CBCF) is an approach to refining the detection results for independent concepts by modeling relationships between them [15]. Concept correlations are either learned from annotation sets [15–17, 20, 34] or inferred from pre-constructed knowledge bases [18, 36, 41] such as WordNet. However, annotation sets are almost always inadequate for learning correlations due to their limited sizes and the annotation having being done with independent concepts rather than correlations in mind. The use of external knowledge networks also limits the flexibility of CBCF because they use a static lexicon which is costly to create. In [39], a training-free method was proposed to utilize concept correlations through global and local refinement. When pre-constructed ontology can be incorporated in the optimization procedure, the method can better adapt to this knowledge constraint. Similarly, [38] dealt with multi-label indexing problem through a tensor-factorization method by taking temporal semantics into account.

In bi-directional image and sentence retrieval, Hodosh et al. [13] proposed Kernel CCA (KCCA) in order to discover a shared feature space for both modalities of images and sentences, which is a powerful approach of extracting nonlinear features in machine learning area. KCCA increases the flexibility of the feature selection, which has been applied to map the hypotheses to a higher-dimensional feature space. KCCA has been applied in some previous work by Lai and Fyfe [21] and Vinokourov et al. [33] with improved results. [8] also uses KCCA to model correlation between web images and corresponding text captions. More recently, Yoshida et al. has proposed a novel method of two-stage kernel CCA to select appropriate kernels in the framework of multiple kernel learning [46]. Though highly non-linear inter-modal relations can be exploited by KCCA, multi-label semantics is not utilized in KCCA and how kernel method can be employed in multi-label cross-modal retrieval using CCA remains unaddressed. Sung Ju Hwang et al. [11] introduce a method for image retrieval based on KCCA that leverages the implicit information about object importance conveyed by the list of keyword labels. However, this type of labels is difficult to obtain. Thus we need a novel method to utilize label information more naturally and conveniently.

3 Method overview and notations

In this section, we present the overview of the proposed multi-label kernel Canonical Correlation Analysis (ml-KCCA) for multi-label settings in cross-modal retrieval tasks. In proposing ml-KCCA, we rely on CCA as the fundamental approach given its efficiency in learning a common subspace for different modalities. We restrict the discussion to multi-label entities containing image and text to simplify the notation and model description and our method can easily apply to any combination of content modalities. Before detailed description of the proposed ml-KCCA framework, a brief review of CCA is first presented for the purpose of completeness.

3.1 Brief on CCA

Provided with different views of data such as represented by two multidimensional variables, Canonical Correlation Analysis (CCA) is able to construct their common representation by analyzing the linear relationships between them. CCA uses data consisting of paired views to simultaneously find projections from each feature space such that correlation between projected features originating from the same instance is maximized [9]. Formally, given a set of N paired data samples $\{(t_1, p_1), ..., (t_N, p_N)\}$, where $t \in R^t$, and $p \in R^p$ denote textual and visual modal data respectively and are both normalized, the key is to seek two sets of vectors u and v to maximize the canonical correlation:

$$\rho = \max_{u,v} \frac{u^T C_{tp} v}{\sqrt{u^T C_{tt} u}\sqrt{v^T C_{pp} v}} \tag{1}$$

Where, $C_{tp} = \frac{1}{N}\sum_{i=1}^{N} t_i {p_i}^T$ denotes the between-sets covariance matrix, $C_{tt} = \frac{1}{N}\sum_{i=1}^{N} t_i t_i^T$ and $C_{pp} = \frac{1}{N}\sum_{i=1}^{N} p_i p_i^T$ denote the auto-covariance matrices for textual and visual data, respectively. The solution for (1) can be found via a generalized eigenvalue problem. As we can see, CCA cannot mine the non-linear correlations between different modalities as it is a linear method. Moreover, CCA cannot utilize high-level semantic information which further limits its performance. These disadvantages usually lead to the

common subspace learned by CCA which is not discriminative enough for cross-modal retrieval tasks.

3.2 Kernelizing CCA with multiple labels

In this section, we introduce multi-label kernel Canonical Correlation Analysis (ml-KCCA) to deal with the tasks of cross-modal retrieval involving multi-label scenarios. By optimizing kernel matrices with this approach, the similarity between corresponding multi-label vectors of paired data can be utilized to learn a more discriminative common subspace for different modalities which is more suitable for cross-modal retrieval tasks.

Figure 2 illustrates the schematic diagram of ml-KCCA in which triangles and squares denote images and texts. Different labels are represented by $+$, $-$, $\times$ and $\div$ in Fig. 2. As shown in Fig. 2a, semantic similarity matrix obtained from multi-label representation is employed to obtain the new form of kernel matrixes K_t^* and K_p^* for texts and images respectively. After solving the kernelized version of (1), a new feature space is constructed in Fig. 2b using optimized projection vectors α and β which have the same interpretation as u and v in (1). As shown in Fig. 2b, paired multi-modal instances with similar labels are semantically more close and then have small coordinate distance in this new projected common space. Bi-directional cross-modal retrieval is effectively performed such as retrieving images in response to a text query, and vice versa, after both texts and images are mapped to this common space using ml-KCCA.

4 Multi-label cross-modal retrieval with ml-KCCA

4.1 Embedding multi-label semantics with ml-KCCA

To formalize ml-KCCA, we denote N samples of paired images and texts with multi-label information as $\{(t_1, p_1, z_1), ..., (t_i, p_i, z_i), ..., (t_N, p_N, z_N)\}$, where z_i is the label vector

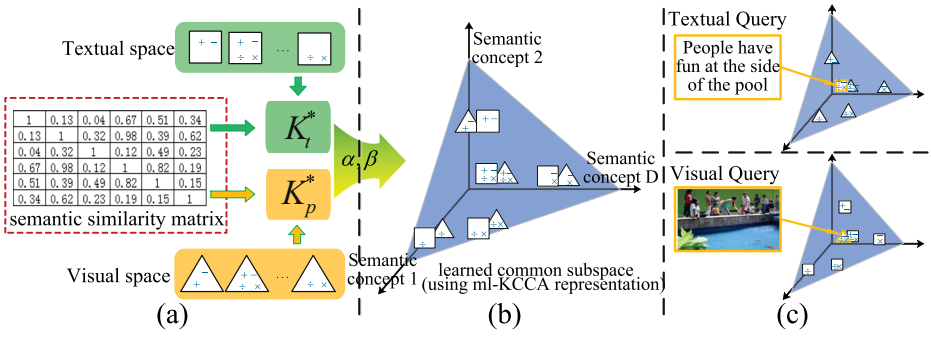


Fig. 2 Schematic diagram of multi-label kernel Canonical Correlation Analysis (ml-KCCA). Triangles and squares denote datapoints in visual and textual modalities respectively, and '+', '−', '×', ÷ denote different class labels. **a** Mapping of text and image instances from their respective feature spaces to a common subspace learned using ml-KCCA. **b** The distance of paired instances having similar labels are more closer in the common subspace learned by ml-KCCA. **c** An example of bi-directional cross-modal retrieval: After both texts and images are mapped to the learned subspace, images can be retrieved more accurately in response to a text query, and vice versa

of the i-th samples of paired data. $T = [t_1, t_2, ..., t_N] \in R^{dt \times N}$ is the matrix representation of textual samples, where dt is the dimension of the textual feature space. Similarly, $P = [p_1, p_2, ..., p_N] \in R^{dp \times N}$ is the matrix representation of the visual images, where dp denote the dimension of the visual feature space. $Z = [z_1, z_2, ..., z_N] \in R^{C \times N}$ is the label matrix whose columns are the label vectors and C is equal to the number of labels of interest. While only one single element in z_i is nonzero in single label problem, elements in z_i corresponding to multiple labels could be nonzero simultaneously in ml-KCCA. This introduces more complex semantic correlations which cannot be well settled with linear CCA.

Let $f(\cdot)$ be a function which gives similarity between any two label vectors. The semantic similarity matrix S can be calculated as $S = (f(z_i, z_j))_{N \times N}$, $1 \leq i, , j \leq N$. Given kernel functions (we use polynomial kernels, linear kernels and gaussian kernels in this work) for both feature spaces, $k_t(t_i, t_j) = \phi_t(t_i)^T \phi_t(t_j)$ and $k_p(p_i, p_j) = \phi_p(p_i)^T \phi_p(p_j)$, then original kernel matrices can be formalized as $K_t = (k_t(t_i, t_j))_{N \times N}$ and $K_p = (k_p(p_i, p_j))_{N \times N}$, $1 \leq i, j \leq N$. We further define $K_t^* = \eta S \cdot K_t$ and $K_p^* = \eta S \cdot K_p$ as the $N \times N$ multi-label kernel matrices which have multi-label embedding over N sample pairs, where '$\cdot$' denotes dot product and η is used to control the influence of semantic similarity matrix. By denoting $K_{tp}^* = K_t^* K_p^*$, the objective in the form of KCCA [13] can be extended using the above defined multi-label kernels in order to identify $\alpha, \beta \in R^N$ so as to maximize the canonical correlation:

$$\rho^* = \max_{\alpha, \beta} \frac{\alpha^T K_{tp}^* \beta}{\sqrt{\alpha^T K_t^{*2} \alpha \beta^T K_p^{*2} \beta}} \tag{2}$$

4.2 Constructing common subspace

Similar as CCA problem defined by (1), (2) can also be reduced to an eigenvalue problem (See Hardoon et al. [8] for more details of the solution). Therefore, computational cost of our method is similar to KCCA because the main cost lies in eigenvalue problem. α, β can be obtained in a similar manner as that in the case of kernel CCA as:

$$B^{-1} A w = \lambda w, \tag{3}$$

where,

$$A = \begin{bmatrix} 0 & K_t^* K_p^* \\ K_p^* K_t^* & 0 \end{bmatrix}, B = \begin{bmatrix} K_t^* K_t^* & 0 \\ 0 & K_p^* K_p^* \end{bmatrix},$$
$$w = \begin{bmatrix} \alpha & \beta \end{bmatrix}^T.$$

As indicated by (3), once K_t^* and K_p^* are computed, the returned top D eigenvectors yield a series of bases $(\alpha^1, \beta^1), ..., (\alpha^D, \beta^D)$ with which to compute the D-dimensional projections for an arbitrary modal input t or p. For example, an unseen textual input t_x can be projected to the common space as a single coordinate specified by α, i.e., evaluating the weighted kernel function between the t_x and the N sampled training points, formalized as:

$$\sum_{i=1}^{N} \alpha_i \phi_t(t_i)^T \phi_t(t_x) = \sum_{i=1}^{N} \alpha_i k_t(t_i, t_x) \tag{4}$$

Then, the final projection of t_x onto the D-dimensional common subspace is formed as:

$$\left[\sum_{i=1}^{N} \alpha_i^1 k_t(t_i, t_x) \ \ ... \ \ \sum_{i=1}^{N} \alpha_i^D k_t(t_i, t_x) \right] \tag{5}$$

Similarly, an unseen image input p_x can be represented in the common subspace as:

$$\left[\sum_{i=1}^{N} \beta_i^1 k_p(p_i, p_x) \ \ldots \ \sum_{i=1}^{N} \beta_i^D k_p(p_i, p_x) \right] \tag{6}$$

After projecting all image and text instances into this learned common subspace, various tasks like image annotation and image search can be performed based on this semantic representation. Because the data points in this D-dimensional common subspace are more correlated semantically, vector distance can be utilized precisely to measure the distance between instances of different modalities.

4.3 Similarity function

The role of $f(\cdot)$ is to measure the semantic relationship of two multi-label vectors, as introduced in Section 4.1. While different forms of similarity can be employed in $f(\cdot)$, we employ the following similarity functions investigated in [25] since they have been demonstrated to be effective in assigning a higher value to the label pair (z_i, z_j) when the labels are more similar, as reported in [25]:

Dot-product based similarity:

$$f(z_i, z_j) = \frac{\langle z_i, z_j \rangle}{||z_i|| ||z_j||} \tag{7}$$

Squared exponential distance based similarity:

$$f(z_i, z_j) = e^{(-||z_i - z_j||_2^2/\sigma)}, \tag{8}$$

where σ is a constant factor for scaling the sample-wise distance.

5 Experiments

In this section, the proposed ml-KCCA method is applied to three public datasets and experimental results are reported on two tasks of image annotation (image query text) and image retrieval (text query image). While performing image annotation and image retrieval, the query and the test points are projected to the common subspace using (5) or (6), and the retrieval performance is measured by comparing the label vector of the query with the label vectors of retrieved test points. Gaussian kernels are used for all component features. We empirically fix the number of selected top eigenvectors returned by (3) as $D = 20$ for all experiments reported in the following sections since the overall performance is shown to be insensitive to the dimensionality of the common subspace constructed by ml-KCCA. To tackle the computational issue that can arise when using large data sets, we applied incomplete Cholesky decomposition to accelerate solving the eigenvalue problem for ml-KCCA.

5.1 Experiment setup

5.1.1 Datasets

Three datasets of NUS-WIDE [3], PASCAL VOC 2007 [5] and LabelMe [12] are employed to evaluate the proposed method and all datasets contain two modalities of images and

texts annotated with multi-label information. The details of these three datasets can be summarized as:

NUS-WIDE consists of 269,648 Flickr documents and we randomly select 20K for training and 20K for testing from this dataset. Each document consists of an image and its corresponding textual tags which are selected from a vocabulary of 81 semantic concepts. In this experiment, we employ the widely-used bag of visual words (BoVW) as the visual feature and 1,000 dimensional bag-of-word (BoW) tag features as textual feature.

Pascal VOC 2007 consists of 5,011 training and 4,952 testing images and this split is directly used in our experiemnts. Similar as in NUS-WIDE, we use the publicly available BoVW, together with gist [23] and color histogram features as visual features. Convolutional features extracted by VGG 16 layers model [31] are also used in experiments of Section 5.2.3. For textual feature, we use the 399 dimensional absolute tag rank features provided by [11]. Groundtruth annotations of the images in this dataset are used as multi-label information.

LabelMe consists of a total of 3,825 images and we select 3,000 samples randomly for training and the rest 825 samples for testing in our experiments. We use the publicly available bag of visual words, gist and color histogram features for image representation. For text representation, we use the 209 dimensional absolute tag rank features provided by [11]. For multi-label information representation, we use the groundtruth annotation of the images.

5.1.2 Evaluation metrics

To evaluate the proposed retrieval method, the following evaluation metrics are adopted:

- Precision@K: Precision@K (P@K) measures the precision at top-K results of the retrieved list and is employed in our evaluation.
- NDCG@K: Performances are also evaluated using normalized discounted cumulative gain at top-K (NDCG@K) [14], a measure commonly used in information retrieval. It gives graded relevance to retrieved results instead of binary relevance and more strongly emphasizes the accuracy of the higher ranked items. The score ranges from 0 to 1 and 1 indicates perfect agreement. NDCG@K can be calculated as:

$$\text{NDCG@K} = \frac{\sum_{i=1}^{K} 2^{rel_i - 1/\log_2(i+1)}}{\sum_{j=1}^{K} 2^{rel_j - 1}/\log_2(j+1)} \tag{9}$$

  where rel_i denotes the degree of relevance of the i-th document in the result while rel_j is judged according to the groundtruth ranking.
- MAP: As a widely adopted metrics in evaluating the retrieval performance, mean average precision (MAP) criterion is used in our experiment which is formalized as

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \left(\frac{1}{R} \times \sum_{r=1}^{R} \frac{r}{position(r)} \right)}{Q} \tag{10}$$

  where Q denotes the number of all queries, R indicates the number of relevant documents in the result returned for a query q, $position(r)$ indicates the position of the r-th relevant document in the result list.

5.2 Results and discussion

5.2.1 Effects of multi-label semantics

As formalized in Section 4, parameters of η and σ measure how multiple labels affect kernel matrices and the scaling of semantic similarity based on squared exponential distance, respectively. To evaluate the effects of multi-label semantics in ml-KCCA, we train the value of parameters η and σ of ml-KCCA on NUS-WIDE training set as mentioned in Section 5.1.1. Dot-product based similarity and gaussian kernels are adopted while we seek the influence of η in order to eliminate the effects of σ. After stable performance is achieved, we fix η and evaluate the influence of σ on ml-KCCA using squared exponential distance based similarity function as introduced in Section 4.3.

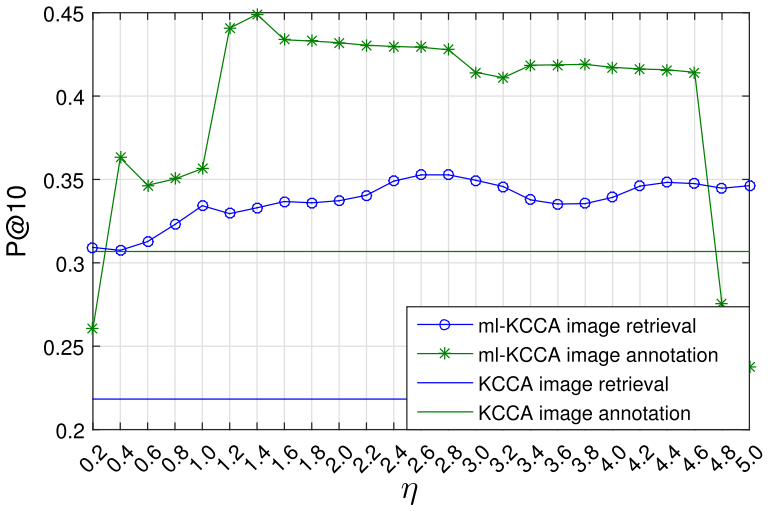
We use KCCA as baseline method in the same settings with our method except ml-KCCA utilizes the semantic similarity matrix derived from multi-label information, and experimental results are shown in Fig. 3. As we can see from Fig. 3a, the proposed ml-KCCA method performs much better than KCCA [13] for most η settings in bi-directional information retrieval. The robustness over η values indicates that utilizing multi-label information using ml-KCCA is valuable in finding a more distinctive common subspace for cross-modal tasks. Only at two extremes of very small and large η values, ml-KCCA has less satisfactory performances for image annotation task. This makes sense because small η value imposes less effects of multi-label semantics while large value will force multi-label semantics to dominate the correlation learning.

The influence of σ is shown in Fig. 3b which is generated by fixing $\eta = 2.0$. While the curve of ml-KCCA image retrieval is stable in Fig. 3b, the fluctuation of ml-KCCA image annotation curve shows that σ has more influence on image annotation than image retrieval task. By comparing two figures of Fig. 3a and b, we can find that: 1) Squared exponential distance based similarity function performs better than dot-product based similarity function because the former uses Gaussian formula which is smoothed by hyper parameter σ, hence is more effective in measuring the similarity of multi-label label pair (z_i, z_j). 2) ml-KCCA is more sensitive in image annotation than in image retrieval while ml-KCCA performs better in image annotation than in image retrieval for most cases.

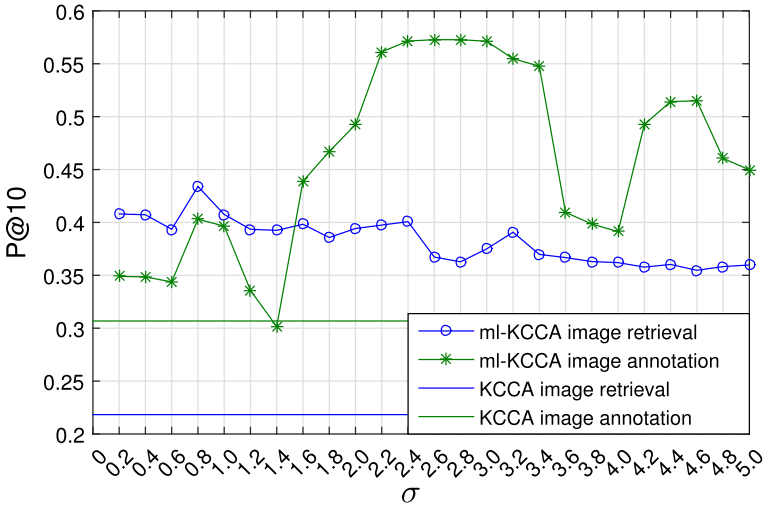
5.2.2 Effects of label quality

As we previously introduced, the utilization of multi-label semantics can enhance the performance of cross-modal retrieval, which has been validated in our experiment as discussed in Section 5.2.1. Because the concept correlations are an important part of semantics in bridging two modalities, the performance of ml-KCCA might be degraded if the inherent label correlations are destroyed. In contrast to the experiment in Section 5.2.1 in which the number of labels are fixed as they are originally annotated, in this section we evaluate the effects of multi-label label correlations by controlling the quantity of labels. We first categorize the samples into different sets according to the quantity of labels and further evaluate our method in these sets respectively in order to compare the performances of the proposed method on datasets with different quantities of labels.

To avoid the influence of textual and visual features of different samples, we select 3,500 image-text pairs with 4 labels from NUS-WIDE dataset and use this subset (3,000 pairs for training and 500 pairs for testing) to repeat the experiment as implemented in Section 5.2.1 four times using randomly selected 1–3 labels and 4 labels of each sample respectively. Experimental results are shown in Fig. 4. Similar to Section 5.2.1, dot-product based



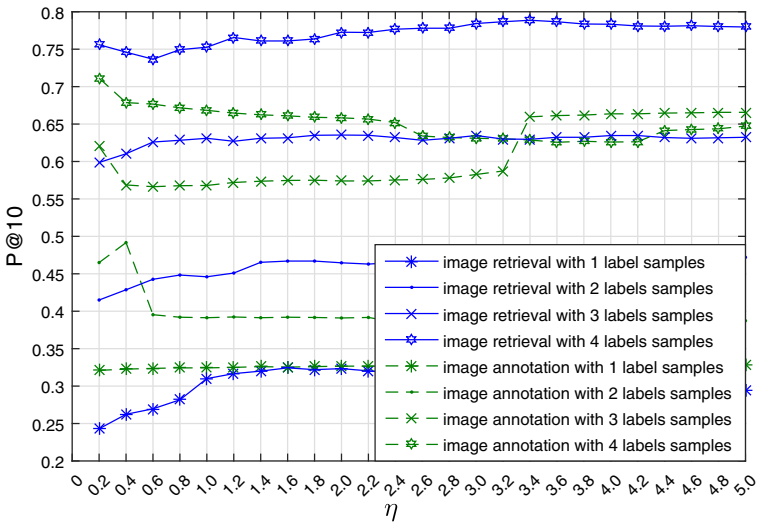
(a)



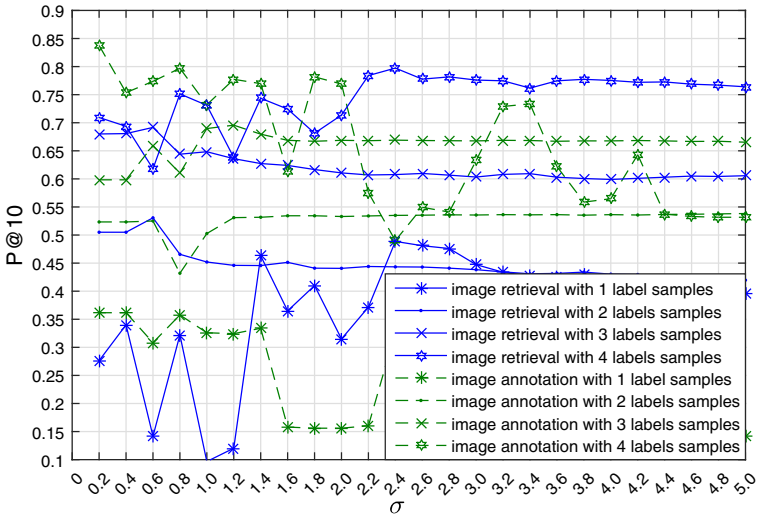
(b)

Fig. 3 Experimental results. **a** The influence of η with dot-product based similarity function(horizontal axis: η). **b** The influence of σ with squared exponential distance based similarity function and $\eta = 2.0$ (horizontal axis: σ). Precision@10 is used as the performance measure

similarity and gaussian kernels are adopted while we seek the influence of η and we fix $\eta = 2.0$ and evaluate the influence of σ on ml-KCCA using squared exponential distance. From Fig. 4, we can find the quantity of sample labels has a obvious influence on the performance of ml-KCCA indicating the importance of multi-label semantics in cross-modal



(a)



(b)

Fig. 4 Effects of label quality. **a** The effects of label quality using dot-product based similarity function (horizontal axis: η). **b** The effects of label quality using squared exponential distance based similarity function with $\eta = 2.0$ (horizontal axis: σ). Precision@10 is used as the performance measure

retrieval. When label quality is improved, i.e. more labels are correctly annotated, both tasks of image annotation and retrieval can be further enhanced. This also validates that our proposed ml-KCCA can make full use of such semantics and effectively embed semantic correlations in enhancing the final retrieval tasks.

Table 1 Performances of ml-KCCA and other CCA extensions are compared on Pascal dataset using NDCG@30 for cross-modal retrieval task

Model	BoVW	Color	Gist	Combination	CNN
	Image annotation				
CCA [9]	0.3015	0.2389	0.3162	0.3673	0.6310
3-view CCA [7]	0.2426	0.2900	0.3034	0.2465	0.5967
cluster CCA [27]	0.2730	0.2429	0.2771	0.2941	0.5931
KCCA [13]	0.3126	0.2742	0.3281	0.3927	0.6369
ml-CCA [25]	0.3213	0.2618	**0.3400**	0.3942	0.6474
our method	**0.3672**	**0.3127**	0.3314	**0.4236**	**0.6821**
	Image retrieval				
CCA [9]	0.4362	0.2997	0.4272	0.5110	0.7720
3-view CCA [7]	0.3520	0.2693	0.3982	0.4555	0.6985
cluster CCA [27]	0.3763	0.2886	0.3630	0.4123	0.6596
KCCA [13]	0.4413	0.3028	0.4172	0.5310	0.7839
ml-CCA [25]	**0.4723**	0.3236	0.4344	0.5568	0.8004
Our method	0.4512	**0.3527**	**0.4462**	**0.5714**	**0.8117**

The best results are highlighted in boldface

5.2.3 Image annotation and image retrieval

In this section, more comprehensive comparison is given to evaluate how well the textual or visual information can be retrieved to describe the content of the given image or sentence. In the implementation, we fix $\eta = 1.8$ and use squared exponential distance based similarity function with $\sigma = 2$ on Pascal dataset. Tables 1 and 2 list the perfor-

Table 2 Performances of ml-KCCA and other CCA extensions are compared on LabelMe dataset using NDCG@30 for cross-modal retrieval task

Model	BoVW	Color	Gist	Combination	CNN
	Image annotation				
CCA [9]	0.5522	0.4765	0.5352	0.5791	0.5586
3-view CCA [7]	0.4223	0.4573	0.4773	0.5618	0.5482
cluster CCA [27]	0.5021	0.4225	0.4368	0.5097	0.5748
KCCA [13]	0.5613	0.4618	0.5526	0.5961	0.5867
ml-CCA [25]	0.5813	0.4789	0.5404	0.6212	0.5885
our method	**0.6117**	**0.5064**	**0.6019**	**0.6573**	**0.6794**
	Image retrieval				
CCA [9]	0.5509	0.5129	0.5519	0.5995	0.5685
3-view CCA [7]	0.5828	0.5615	0.5461	0.6322	0.6171
cluster CCA [27]	0.5319	0.4825	0.4723	0.5561	0.5695
KCCA [13]	0.5713	0.5318	0.5827	0.6139	0.5983
ml-CCA [25]	0.6160	0.5536	0.5819	0.6534	0.6185
Our method	**0.6513**	**0.6004**	**0.6268**	**0.7041**	**0.7284**

The best results are highlighted in boldface

Table 3 Comparison of ml-KCCA with various state-of-the-arts using MAP

Model	Image annotation	Image retrieval
KGMMFA [30]	0.421	0.328
KGMLDA [30]	0.427	0.339
LCFS [35]	0.344	0.267
LGCFL [19]	0.378	0.329
KCCA [13]	0.412	0.357
ml-CCA [25]	0.484	0.380
Our method	**0.5091**	**0.4117**

The best results are highlighted in boldface

mance of ml-KCCA and other typical CCA extensions on Pascal dataset and LableMe dataset, using BoVW, color histogram, gist features, combination of them, and Convolutional Neural Network (CNN) features as visual features. According to Table 1 and 2, it is obvious that ml-KCCA outperforms all the other approaches on two datasets across most features. This demonstrates the advantage of ml-KCCA in utilizing multi-label information and exploiting non-linear inter-modal relations effectively. By taking multi-label semantics into account, both ml-CCA and our proposed ml-KCCA methods can outperform the other methods significantly. Moreover, our proposed ml-KCCA method more flexibly models the underling non-linear correspondence in multiple labels, and outperforms ml-CCA on most of features (BoVW, Color histogram, Combination and CNN) on two datasets. When using gist for image annotation and BoVW for image retrieval, the proposed ml-KCCA still obtains comparable performance on NDCG@30 matric with ml-CCA without degrading the performance.

In Table 3, the proposed ml-KCCA is also compared to other state-of-the-art cross-modal retrieval baselines used in [25] on Pascal dataset, using publicly available image and text features provided by [11]. Similar as in Table 1, with the use of multi-label information, the MAP of ml-KCCA method outperforms other methods including kernelized methods like KGMMFA and KGMLDA [30] in fulfilling the tasks. Our method outperforms ml-CCA which is the CCA extension with best performance before because our method can exploit more complex non-linear relations of different modal as a kernelized method.

6 Conclusions and future work

In this work, we propose Multi-Label Kernel Canonical Correlation Analysis (ml-KCCA), a novel kernelized method for cross-modal retrieval, which can effectively utilize multi-label information while learning the common subspace of multiple modalities. Experimental results on public datasets show that ml-KCCA achieves state-of-the-art performance in bi-directional retrieval. Though multi-label cross-modal retrieval can be semantically enhanced with the proposed ml-KCCA methods, there are still some space to improve the work which can be regarded as part of future work. For example, it is helpful in improving the proposed model by designing better similarity functions between multi-label vectors, during which pairwise label semantic correlation can be further considered. More extensive experiments

on large-scale datasets like ImageNet can be carried out as another future work by taking multi-label annotations into consideration.

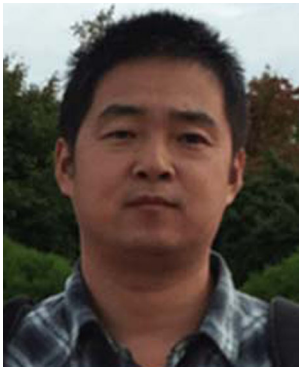
Acknowledgments This work is supported by the Natural Science Foundation of China under Grant No. 61571453, No. 61502264, and No. 61405252, Natural Science Foundation of Hunan Province, China under Grant No. 14JJ3010, Research Funding of National University of Defense Technology under grant No. ZK16-03-37.

References

1. Akaho S (2006) A kernel method for canonical correlation analysis. In: Proceedings of the international meeting of the psychometric society, vol 40, pp 263–269
2. Bekkerman R, Jeon J (2007) Multi-modal clustering for multimedia collections. In: IEEE conference on computer vision and pattern recognition, pp 1–8
3. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: a real-world web image database from National University of Singapore. In: ACM international conference on image and video retrieval, p 48
4. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition
5. Everingham M, Gool L, Williams CK, Winn J, Zisserman A (2010) The Pascal Visual Object Classes (VOC) challenge. Int J Comput Vis 88(2):303–338
6. Gong Y, Lazebnik S, Gordo A et al (2013) Iterative quantization: a Procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans Pattern Anal Mach Intell 35(12):2916
7. Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233
8. Hardoon D, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
9. Hotelling H (1992) Relations between two sets of variates. In: Breakthroughs in statistics, pp 321–377
10. Huyn N (2001) Data analysis and mining in the life sciences. In: ACM
11. Hwang SJ, Grauman K (2010) Accounting for the relative importance of objects in image retrieval. In: British machine vision conference, pp 1–12
12. Hwang SJ, Grauman K (2010) Reading between the lines: object localization using implicit cues from image tags. In: IEEE conference on computer vision and pattern recognition, pp 2971–2978
13. Hwang SJ, Grauman K (2012) Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int J Comput Vis 100(2):134–153
14. Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446
15. Jiang W, Chang S-F, Loui AC (2007) Context-based concept fusion with boosted conditional random fields. In: IEEE international conference on acoustics, speech and signal processing
16. Jiang Y-G, Wang J, Chang S-F, Ngo C-W (2009) Domain adaptive semantic diffusion for large scale context-based video annotation. In: IEEE 12th international conference on computer vision, pp 1420–1427
17. Jiang Y-G, Dai Q, Wang J, Ngo C-W, Xue X, Chang S-F (2012) Fast semantic diffusion for large-scale context-based image and video annotation. IEEE Trans Image Process 21(6):3080–3091
18. Jin Y, Khan L, Wang L, Awad M (2005) Image annotations by combining multiple evidence & WordNet. In: ACM international conference on multimedia, pp 706–715
19. Kang C, Xiang S, Liao S, Xu C, Pan C (2015) Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans Multimed 17(3):370–381
20. Kennedy LS, Chang S-F (2007) A reranking approach for context-based concept fusion in video indexing and retrieval. In: Proceedings of the 6th ACM international conference on image and video retrieval, pp 333–340

21. Lai PL, Fyfe C (2000) Kernel and nonlinear canonical correlation analysis. Int J Neural Syst 10(5):365
22. Miller GA (1995) WordNet: a lexical database for english. Commun ACM 38(11):39–41
23. Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36
24. Qi G-J, Hua X-S, Rui Y, Tang J, Mei T, Zhang H-J (2007) Correlative multi-label video annotation. In: ACM international conference on multimedia, pp 17–26
25. Ranjan V, Rasiwasia N, Jawahar CV (2015) Multi-label cross-modal retrieval. In: IEEE international conference on computer vision, pp 4094–4102
26. Rasiwasia N, Pereira JC, Coviello E et al (2010) A new approach to cross-modal multimedia retrieval. In: ACM international conference on multimedia, pp 251–260
27. Rasiwasia N, Mahajan D, Mahadevan V, Aggarwal G (2014) Cluster canonical correlation analysis. In: Proceedings of international conference on artificial intelligence and statistics
28. Sang J, Xu C, Liu J (2012) User-aware image tag refinement via ternary semantic analysis. IEEE Trans Multimed 14(3):883–895
29. Sang J, Fang Q, Xu C (2017) Exploiting social-mobile information for location visualization. ACM TIST 8(3):39:1–39:19
30. Sharma A (2012) Generalized multiview analysis: a discriminative latent space. In: IEEE conference on computer vision and pattern recognition, pp 2160–2167
31. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computer Science
32. Srivastava N, Salakhutdinov R (2014) Multimodal learning with deep Boltzmann machines. J Mach Learn Res 15(8):1967–2006
33. Vinokourov A, Shawe-Taylor J, Cristianini N (2002) Inferring a semantic representation of text via cross-language correlation analysis. In: Advances of neural information processing systems, pp 1497–1504
34. Wang C, Jing F, Zhang L, Zhang H-J (2006) Image annotation refinement using random walk with restarts. In: ACM international conference on multimedia, pp 647–650
35. Wang K, He R, Wang W, Wang L, Tan T (2013) Learning coupled feature spaces for cross-modal matching. In: IEEE international conference on computer vision, pp 2088–2095
36. Wang P, Sun LF, Yang SQ, Smeaton AF (2016) Semantically smoothed refinement for everyday concept indexing. In: Pacific rim conference on multimedia (PCM)
37. Wang P, Sun LF, Yang SQ, Smeaton AF (2016) Towards training-free refinement for semantic indexing of visual media. In: International conference on multimedia modeling, pp 251–263
38. Wang P, Sun LF, Yang SQ, Smeaton AF, Gurrin C (2016) Characterizing everyday activities from visual lifelogs based on enhancing concept representation. Comput Vis Image Underst 148:181–192
39. Wang P, Sun LF, Yang SQ, Smeaton AF (2017) Training-free indexing refinement for visual media via multi-semantics. Neurocomputing 236:39–47
40. Wang H, Wu X, Jia Y (2017) Heterogeneous domain adaptation method for video annotation. IET Comput Vis 11(2):181–187
41. Wu Y, Tseng B, Smith JR (2004) Ontology-based multi-classification learning for video concept detection. In: IEEE international conference on multimedia and expo
42. Wu F, Zhang H, Zhuang Y (2007) Learning semantic correlations for cross-media retrieval. In: IEEE international conference on image processing. IEEE, pp 1465–1468
43. Wu F, Lu X, Zhang Z, Yan S, Rui Y, Zhuang Y (2013) Cross-media semantic representation via bi-directional learning to rank. In: ACM international conference on multimedia, pp 877–886
44. Xue X, Zhang W, Zhang J, Wu B, Fan J, Lu Y (2011) Correlative multi-label multi-instance image annotation. In: ICCV, pp 651–658
45. Yao T, Mei T, Ngo CW (2015) Learning query and image similarities with ranking canonical correlation analysis. In: IEEE international conference on computer vision, pp 28–36
46. Youshida K, Yoshimoto J, Doya K (2017) Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. BMC Bioinf 18(1):108
47. Yu J, Rui Y, Tao D (2014) Click Prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process 23(5):2019–2032
48. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. IEEE Trans Cybern 45(4):767–779
49. Yu J, Yang X, Gao F, Tao D (2016) Deep multimodal distance metric learning using click constraints for image ranking. IEEE Trans Cybern PP(99):1–11

Yuhua Jia received his B.E. degree from National University of Defense Technology. He is currently pursuing the M.E. degree in the College of Information System and Management at the National University of Defense Technology in China. He is also a member of Science and Technology on Information Systems Engineering Laboratory. His research interests include deep learning, cross-media retrieval.



Liang Bai is currently an associate professor in the School of Information System and Management, National University of Defense Technology. He received the B.E. (Bachelor Degree of Engineering) and B.M. (Bachelor Degree of Management) degrees in 2002 from Xi'an Jiao Tong University, and M.E. degree in 2005, Ph.D. degree in 2008 both from National University of Defense Technology. He has published about 60 peer-reviewed papers. He is also a regularly reviewer for a number of top international or Chinese journals. He was awarded the Second Prize for Science and Technology Development of the Ministry of Education, P.R. China in 2011. His research interest includes multimedia content analysis and access, particularly for video and image. Big Multimedia Data is also his research focus now. He is a member of ACM and IEEE Computer Society, and a member of the Chinese Computer Society.

Shuang Liu received her B.E. degree from National University of Defense Technology. She is currently pursuing the M.E. degree in the College of Information System and Management at the National University of Defense Technology in China. She is also a member of Science and Technology on Information Systems Engineering Laboratory. Her research interest is cross-media retrieval and content-based multimedia analysis and retrieval.

Peng Wang is now a postdoctoral researcher in Department of Computer Science and Technology, Tsinghua University. He received his PhD degree in Computing from CLARITY: Center for Sensor Web Technologies, Dublin City University on the topic of semantic mining and enhancement of lifelogging events. During 2010, he visited the Digital Enterprize Research Institute (DERI), National University of Ireland, Galway to work on applying Semantic Web technologies to the application of topic-related concept selection and semantic event enhancement. Since 2014, he has worked as a post doctor at the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include human behavior analysis, semantic mining from multimedia. He has published more than 20 peer-reviewed papers in journals, conferences and workshops, mainly on the topic of multimedia content analysis, event detection and recognition, human behavior mining from large scale mobilities, etc.

Jinlin Guo is currently a lecturer in the School of Information System and Management, National University of Defense Technology. He received his PhD degree in Computing from CLARITY: Center for Sensor Web Technologies, Dublin City University on the topic of content analysis for user-generated videos. His research interest includes multimedia information processing and retrieval, machine learning.

Yuxiang Xie received her B.S., M.S. and Ph.D degrees from National University of Defense Technology in 1998, 2001 and 2004 respectively, all in Systems Engineering. She is currently an associate professor in the School of Information System and Management, National University of Defense Technology. Her current research interests include image and video analysis, classification and retrieval.