Modality-Dependent Cross-Media Retrieval

YUNCHAO WEI, YAO ZHAO, ZHENFENG ZHU, SHIKUI WEI, and YANHUI XIAO,
Beijing Jiaotong University
JIASHI FENG and SHUICHENG YAN, National University of Singapore

In this article, we investigate the cross-media retrieval between images and text, that is, using image to search text (I2T) and using text to search images (T2I). Existing cross-media retrieval methods usually learn one couple of projections, by which the original features of images and text can be projected into a common latent space to measure the content similarity. However, using the same projections for the two different retrieval tasks (I2T and T2I) may lead to a tradeoff between their respective performances, rather than their best performances. Different from previous works, we propose a modality-dependent cross-media retrieval (MDCR) model, where two couples of projections are learned for different cross-media retrieval tasks instead of one couple of projections. Specifically, by jointly optimizing the correlation between images and text and the linear regression from one modal space (image or text) to the semantic space, two couples of mappings are learned to project images and text from their original feature spaces into two common latent subspaces (one for I2T and the other for T2I). Extensive experiments show the superiority of the proposed MDCR compared with other methods. In particular, based on the 4,096-dimensional convolutional neural network (CNN) visual feature and 100-dimensional Latent Dirichlet Allocation (LDA) textual feature, the mAP of the proposed method achieves the mAP score of 41.5%, which is a new state-of-the-art performance on the Wikipedia dataset.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Cross-media retrieval, subspace learning, canonical correlation analysis

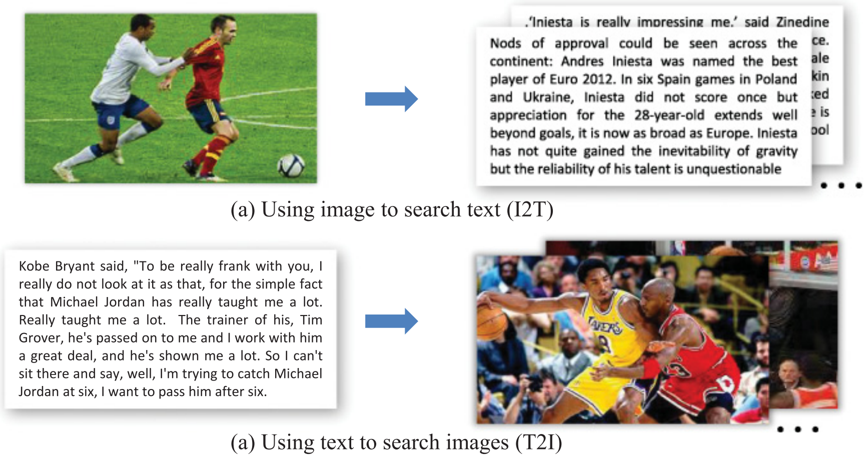
ACM Reference Format:
Yunchao Wei, Yao Zhao, Zhenfeng Zhu, Shikui Wei, Yanhui Xiao, Jiashi Feng, and Shuicheng Yan. 2016. Modality-dependent cross-media retrieval. ACM Trans. Intell. Syst. Technol. 7, 4, Article 57 (March 2016), 13 pages.
DOI: http://dx.doi.org/10.1145/2775109

1. INTRODUCTION

With the rapid development of information technology, multimodal data (e.g., image, text, video, or audio) have become widely available on the Internet. For example, an image often co-occurs with text on a web page to describe the same object or event.

This work is supported in part by the National Basic Research Program of China (No.2012CB316400) and Fundamental Scientific Research Project (No.K15JB00360).
Authors' addresses: Y. Wei, Y. Zhao, Z. Zhu, S. Wei, and Y. Xiao, Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China; email: wychao1987@gmail.com, {yzhao, zhfzhu, shkwei}@bjtu.edu.cn, xiaoyanhui@gmail.com; S. Wei, Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, Hubei 443002, China; J. Feng and S. Yan, Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore; emails: jshfeng@gmail.com, eleyans@nus.edu.sg.

DOI: http://dx.doi.org/10.1145/2775109



.'Iniesta is really impressing me.' said Zinedine

Nods of approval could be seen across the continent: Andres Iniesta was named the best player of Euro 2012. In six Spain games in Poland and Ukraine, Iniesta did not score once but appreciation for the 28-year-old extends well beyond goals, it is now as broad as Europe. Iniesta has not quite gained the inevitability of gravity but the reliability of his talent is unquestionable

(a) Using image to search text (I2T)

Kobe Bryant said, "To be really frank with you, I really do not look at it as that, for the simple fact that Michael Jordan has really taught me a lot. Really taught me a lot. The trainer of his, Tim Grover, he's passed on to me and I work with him a great deal, and he's shown me a lot. So I can't sit there and say, well, I'm trying to catch Michael Jordan at six, I want to pass him after six.

(a) Using text to search images (T2I)

Fig. 1. Cross-media retrieval tasks considered in this article. (a) Given an image of *Iniesta*, the task is to find some text reports related to this image. (b) Given a text document about *Kobe Bryant* and *Michael Jordan*, the task is to find some related images about them. Source images, ©ferhat_culfaz: https://goo.gl/of54g4; ©Basket Streaming: https://goo.gl/DfZLRs; ©Wikipedia: http://goo.gl/D6RYkt.

Related research has been conducted incrementally in recent decades, among which the retrieval across different modalities has attracted much attention and benefited many practical applications. However, multimodal data usually span different feature spaces. This heterogeneous characteristic poses a great challenge to cross-media retrieval tasks. In this work, we mainly focus on addressing the cross-media retrieval between text and images (Figure 1), that is, using image (text) to search text documents (images) with similar semantics.

To address this issue, many approaches have been proposed by learning a common representation for the data of different modalities. We observe that most existing works [Hardoon et al. 2004; Rasiwasia et al. 2010; Sharma et al. 2012; Gong et al. 2013] focus on learning one couple of mapping matrices to project high-dimensional features from different modalities into a common latent space. By doing this, the correlations of two variables from different modalities can be maximized in the learned common latent subspace. However, only considering pairwise closeness [Hardoon et al. 2004] is not sufficient for cross-media retrieval tasks, since it is required that multimodal data from the same semantics should be united in the common latent subspace. Although Sharma et al. [2012] and Gong et al. [2013] have proposed to use supervised information to cluster the multimodal data with the same semantics, learning one couple of projections may only lead to compromised results for each retrieval task.

In this article, we propose a modality-dependent cross-media retrieval (MDCR) method, which recommends different treatments for different retrieval tasks, that is, I2T and T2I. Specifically, MDCR is a task-specific method, which learns two couples of projections for different retrieval tasks. The proposed method is illustrated in Figure 2. Figures 2(a) and 2(c) are two linear regression operations from the image and the text feature space to the semantic space, respectively. By doing this, multimodal data with the same semantics can be united in the common latent subspace. Figure 2(b) is a correlation analysis operation to keep pairwise closeness of multimodal data in the common space. We combine Figures 2(a) and 2(b) to learn a couple of projections for I2T, and a different couple of projections for T2I is jointly optimized by Figures 2(b) and 2(c). The reason we learn two couples of projections rather than one couple for different retrieval tasks can be explained as follows. For I2T, we argue that the accurate representation

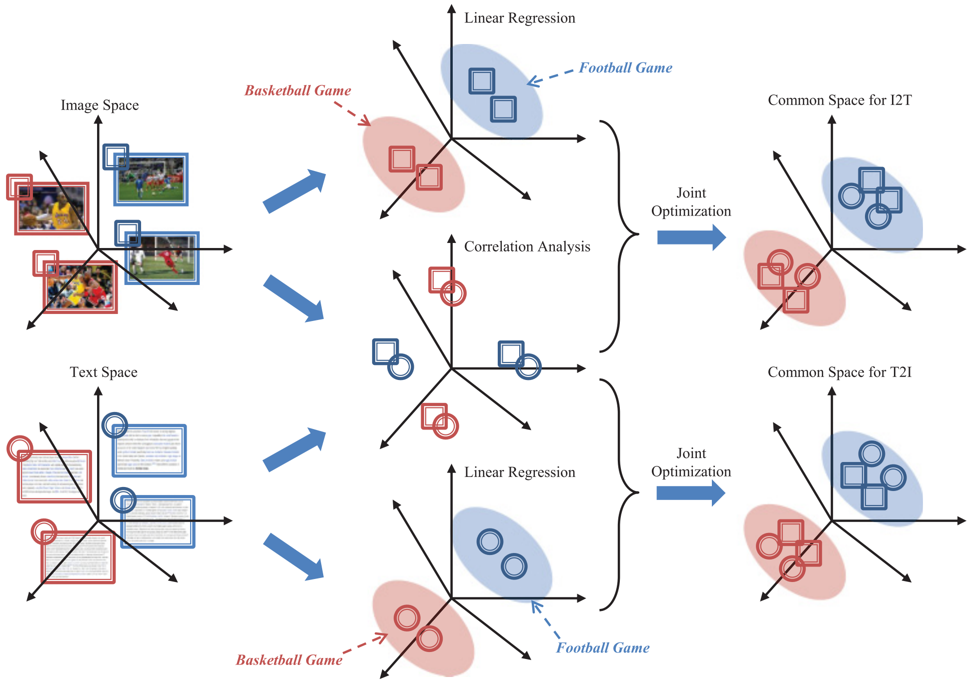


Fig. 2. Modality-dependent cross-media retrieval (MDCR) model proposed in this article. Images are represented by square icons, while text is represented by round icons; different colors indicate different classes. Ellipse fields with blue color and red color indicate semantic clusters of *FootballGame* and *BasketballGame*, respectively. (a) Linear regression from image feature space to semantic space to produce a better separation for images of different classes. (b) Correlation analysis between images and text to keep pairwise closeness. (c) Linear regression from text feature space to semantic space to produce a better separation for text of different classes. Source images, ©Basket Streaming: https://goo.gl/DfZLRs; ©Wikipedia: http://goo.gl/RqWL6O; ©Wikipedia: http://goo.gl/k3cPs8; ©Wikipedia: https://goo.gl/RdgsNL.

of the query (i.e., the image) in the semantic space is more important than that of the text to be retrieved. If the semantics of the query is misjudged, it will be even harder to retrieve the relevant text. Therefore, only the linear regression term from image feature to semantic label vector and the correlation analysis term are considered for optimizing the mapping matrices for I2T. For T2T, the reason is the same as that for I2T. The main contributions of this work are listed as follows:

- We propose a modality-dependent cross-media retrieval method, which projects data of different modalities into a common space so that a similarity measurement such as Euclidean distance can be applied for cross-media retrieval.
- To better validate the effectiveness of our proposed MDCR, we compare it with other state-of-the-art methods based on more powerful feature representations. In particular, with the 4,096-dimensional CNN visual feature and 100-dimensional Latent Dirichlet Allocation (LDA) textual feature, the mAP of the proposed method reaches 41.5%, which is a new state-of-the-art performance on the Wikipedia dataset as far as we know.
- Based on the INRIA-Websearch dataset [Krapac et al. 2010], we construct a new dataset for cross-media retrieval evaluation. In addition, all the features utilized in this article are publicly available.[1]

[1] https://sites.google.com/site/yunchaosite/mdcr.

The remainder of this article is organized as follows. We briefly review the related work of cross-media retrieval in Section 2. In Section 3, the proposed modality-dependent cross-media retrieval method is described in detail. Then, in Section 4, experimental results are reported and analyzed. Finally, Section 5 presents the conclusions.

2. RELATED WORK

During the past few years, numerous methods have been proposed to address cross-media retrieval. Some works [Hardoon et al. 2004; Tenenbaum and Freeman 2000; Rosipal and Krämer 2006; Yang et al. 2008; Sharma and Jacobs 2011; Hwang and Grauman 2010; Rasiwasia et al. 2010; Sharma et al. 2012; Gong et al. 2013; Wei et al. 2014; Zhang et al. 2014] try to learn an optimal common latent subspace for multimodal data. This kind of method projects representations of multiple modalities into an isomorphic space, such that similarity measurement can be directly applied between multimodal data. Two popular approaches, Canonical Correlation Analysis (CCA) [Hardoon et al. 2004] and Partial Least Squares (PLS) [Rosipal and Krämer 2006; Sharma and Jacobs 2011], are usually employed to find a couple of mappings to maximize the correlations between two variables. Based on CCA, a number of successful algorithms have been developed for cross-media retrieval tasks [Rashtchian et al. 2010; Hwang and Grauman 2010; Sharma et al. 2012; Gong et al. 2013]. Rashtchian et al. [2010] investigated the cross-media retrieval problem in terms of correlation hypothesis and abstraction hypothesis. Based on the isomorphic feature space obtained from CCA, a multiclass logistic regression is applied to generate a common semantic space for cross-media retrieval tasks. Hwang and Grauman [2010] used KCCA to develop a cross-media retrieval method by modeling the correlation between visual features and textual features. Sharma et al. [2012] presented a generic framework for multimodal feature extraction techniques, called Generalized Multiview Analysis (GMA). More recently, Gong et al. [2013] proposed a three-view CCA model by introducing a semantic view to produce a better separation for multimodal data of different classes in the learned latent subspace.

To address the problem of prohibitively expensive nearest-neighbor search, some hashing-based approaches [Kumar and Udupa 2011; Wu et al. 2014] to large-scale similarity search have drawn much interest from the cross-media retrieval community. In particular, Kumar and Udupa [2011] proposed a cross-view hashing method to generate hash codes by minimizing the distance of hash codes for the similar data and maximizing the distance for the dissimilar data. Recently, Wu et al. [2014] proposed a sparse multimodal hashing method, which can obtain sparse codes for the data across different modalities via joint multimodal dictionary learning to address cross-modal retrieval. Besides, with the development of deep learning, some deep models [Frome et al. 2013; Wang et al. 2014; Lu et al. 2014; Zhuang et al. 2014] have also been proposed to address cross-media problems. Specifically, Frome et al. [2013] presented a deep visual-semantic embedding model to identify visual objects using both labeled image data and semantic information obtained from unannotated text documents. Wang et al. [2014] proposed an effective mapping mechanism, which can capture both intramodal and intermodal semantic relationships of multimodal data from heterogeneous sources, based on the stacked auto-encoders deep model.

Beyond the aforementioned models, some other works [Yang et al. 2009, 2010, 2012; Wu et al. 2013; Zhai et al. 2013; Kang et al. 2014] have also been proposed to address cross-media problems. In particular, Wu et al. [2013] presented a bidirectional cross-media semantic representation model by optimizing the bidirectional listwise ranking loss with a latent space embedding. In Zhai et al. [2013], both the intramedia and the intermedia correlation are explored for cross-media retrieval. Most recently, Kang et al. [2014] presented a heterogeneous similarity learning approach based on metric learning for cross-media retrieval. With the convolutional neural network (CNN) visual feature, some new state-of-the-art cross-media retrieval results have been achieved in Kang et al. [2014].

3. MODALITY-DEPENDENT CROSS-MEDIA RETRIEVAL

In this section, we detail the proposed supervised cross-media retrieval method, which we call modality-dependent cross-media retrieval. Each pair of image and text in the training set is accompanied with semantic information (e.g., class labels). Different from the work of Gong et al. [2013], which incorporates the semantic information as a third view, in this article, semantic information is employed to determine a common latent space with a fixed dimension where samples with the same label can be clustered.

Suppose we are given a dataset of n data instances, that is, $\mathcal{G} = \{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{t}_i \in \mathbb{R}^q$ are original low-level features of image and text document, respectively. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ be the feature matrix of image data, and $T = [\mathbf{t}_1, \ldots, \mathbf{t}_n]^T \in \mathbb{R}^{n \times q}$ be the feature matrix of text data. Assume that there are c classes in $\mathcal{G}$. $S = [\mathbf{s}_1, \ldots, \mathbf{s}_n]^T \in \mathbb{R}^{n \times c}$ is the semantic matrix with the ith row being the semantic vector corresponding to $\mathbf{x}_i$ and $\mathbf{t}_i$. In particular, we set the jth element of $\mathbf{s}_i$ as 1 if $\mathbf{x}_i$ and $\mathbf{t}_i$ belong to the jth class.

Definition 1. The cross-media retrieval problem is to learn two optimal mapping matrices $V \in \mathbb{R}^{c \times p}$ and $W \in \mathbb{R}^{c \times q}$ from the multimodal dataset $\mathcal{G}$, which can be formally formulated into the following optimization framework:

$$\min_{V,W} \; f(V, W) = \mathcal{C}(V, W) + \mathcal{L}(V, W) + \mathcal{R}(V, W), \tag{1}$$

where f is the objective function consisting of three terms. In particular, $\mathcal{C}(V, W)$ is a correlation analysis term used to keep pairwise closeness of multimodal data in the common latent subspace. $\mathcal{L}(V, W)$ is a linear regression term from one modal feature space (image or text) to the semantic space, used to centralize the multimodal data with the same semantics in the common latent subspace. $\mathcal{R}(V, W)$ is the regularization term to control the complexity of the mapping matrices V and W.

In the following subsections, we will detail the two algorithms for I2T and T2I based on the optimization framework in Equation (1).

3.1. Algorithm for I2T

This section addresses the cross-media retrieval problem of using an image to retrieve its related text documents. Denote the two optimal mapping matrices for images and text as $V_1 \in \mathbb{R}^{c \times p}$ and $W_1 \in \mathbb{R}^{c \times q}$, respectively. Based on the optimization framework in Equation (1), the objective function of I2T is defined as follows:

$$\begin{aligned}\min_{V_1, W_1} \; f(V_1, W_1) = {} & \lambda \left\| X V_1^T - T W_1^T \right\|_F^2 + (1 - \lambda) \left\| X V_1^T - S \right\|_F^2 \\ & + R(V_1, W_1),\end{aligned} \tag{2}$$

where $0 \leq \lambda \leq 1$ is a tradeoff parameter to balance the importance of the correlation analysis term and the linear regression term, $\| \cdot \|_F$ denotes the Frobenius norm of the matrix, and $R(V_1, W_1)$ is the regularization function used to regularize the mapping matrices. In this article, the regularization function is defined as

$$R(V_1, W_1) = \eta_1 \left\| V_1 \right\|_F^2 + \eta_2 \left\| W_1 \right\|_F^2 ,$$

where η_1 and η_2 are nonnegative parameters to balance these two regularization terms.

ALGORITHM 1: Optimization for Modality-Dependent Cross-Media Retrieval

```
Input: The feature matrix of image data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$, the feature matrix of text
        data $T = [\mathbf{t}_1, \ldots, \mathbf{t}_n]^T \in \mathbb{R}^{n \times q}$, the semantic matrix corresponding to images and text
        $S = [\mathbf{s}_1, \ldots, \mathbf{s}_n]^T \in \mathbb{R}^{n \times c}$.
Initialize $V_1^{(\upsilon)}$, $W_1^{(\omega)}$, $\upsilon \leftarrow 0$, and $\omega \leftarrow 0$. Set the parameters $\lambda$, $\eta_1$, $\eta_2$, $\mu$, and $\epsilon$. $\mu$ is the step size
in the alternating updating process and $\epsilon$ is the convergence condition.
repeat
    Alternative optimization process for I2T (Algorithm 2).
until Convergence or maximum iteration number achieves.;
Output: $V_1^{(\upsilon)}$, $W_1^{(\omega)}$.
```

3.2. Algorithm for T2I

This section addresses the cross-media retrieval problem of using text to retrieve its related images. Different from the objective function of I2T, the linear regression term for T2I is a regression operation from the textual space to the semantic space. Denote the two optimal mapping matrices for images and text in T2I as $V_2 \in \mathbb{R}^{c \times p}$ and $W_2 \in \mathbb{R}^{c \times q}$, respectively. Based on the optimization framework in Equation (1), the objective function of T2I is defined as follows:

$$\begin{aligned}\min_{V_2, W_2} \; f(V_2, W_2) = {} & \lambda \left\| X V_2^T - T W_2^T \right\|_F^2 + (1 - \lambda) \left\| T W_2^T - S \right\|_F^2 \\ & + R(V_2, W_2),\end{aligned} \tag{3}$$

where the setting of the tradeoff parameter λ and the regularization function $R(V_2, W_2)$ are consistent with the setting presented in Section 3.1.

3.3. Optimization

The optimization problems for I2T and T2I are unconstrained optimizations with respect to two matrices. Hence, both Equation (2) and Equation (3) are nonconvex optimization problems and only have many local optimal solutions. For the nonconvex problem, we usually design algorithms to seek stationary points. We note that Equation (2) is convex with respect to either V_1 or W_1 while fixing the other. Similarly, Equation (3) is also convex with respect to either V_2 or W_2 while fixing the other. Specifically, by fixing $V_1(V_2)$ or $W_1(W_2)$, the minimization over the other can be finished with the gradient descent method.

The partial derivatives of V_1 or W_1 over Equation (2) are given as follows:

$$\nabla_{V_1} f(V_1, W_1) = V_1 X^T X + 2\left[\eta_1 V_1 - \lambda W_1 T^T X - (1 - \lambda) S^T X\right], \tag{4}$$

$$\nabla_{W_1} f(V_1, W_1) = 2\left[\eta_2 W_1 + \lambda \left(W_1 T^T T - V_1 X^T T\right)\right]. \tag{5}$$

Similarly, the partial derivatives of V_2 or W_2 over Equation (3) are given as follows:

$$\nabla_{V_2} f(V_2, W_2) = 2\left[\eta_1 V_2 + \lambda \left(V_2 X^T X - W_2 T^T X\right)\right], \tag{6}$$

$$\nabla_{W_2} f(V_2, W_2) = W T^T T + 2\left[\eta_2 W_2 - \lambda V_2 X^T T - (1 - \lambda) S^T T\right]. \tag{7}$$

A common way to solve this kind of optimization problem is an alternating updating process until the result converges. Algorithm 1 summarizes the optimization procedure of the proposed MDCR method for I2T, which can be easily extended for T2I.

4. EXPERIMENTAL RESULTS

To evaluate the proposed MDCR algorithm, we systematically compare it with other state-of-the-art methods on three datasets: Wikipedia [Rasiwasia et al. 2010], Pascal

ALGORITHM 2: Alternative Optimization Process for I2T

```
repeat
    Set $value1 = f\left(V_1^{(\upsilon)}, W_1^{(\omega)}\right)$;
    Update $V_1^{(\upsilon+1)} = V_1^{(\upsilon)} - \mu\nabla_{V_1^{(\upsilon)}} f\left(V_1^{(\upsilon)}, W_1^{(\omega)}\right)$;
    Set $value2 = f\left(V_1^{(\upsilon+1)}, W_1^{(\omega)}\right)$, $\upsilon \leftarrow \upsilon + 1$;
until $value1 - value2 \leq \epsilon$;
repeat
    Set $value1 = f\left(V_1^{(\upsilon)}, W_1^{(\omega)}\right)$;
    Update $W_1^{(\omega+1)} = W_1^{(\omega)} - \mu\nabla_{W_1^{(\omega)}} f\left(V_1^{(\upsilon)}, W_1^{(\omega)}\right)$;
    Set $value2 = f\left(V_1^{(\upsilon)}, W_1^{(\omega+1)}\right)$, $\omega \leftarrow \omega + 1$;
until $value1 - value2 \leq \epsilon$;
```

Sentence [Rashtchian et al. 2010], and a subset of INRIA-Websearch [Krapac et al. 2010].

4.1. Datasets

Wikipedia:[2] This dataset contains in total 2,866 image-text pairs from 10 categories. The whole dataset is randomly split into a training set and a test set with 2,173 and 693 pairs. We utilize the publicly available features provided by Rasiwasia et al. [2010] (i.e., 128-dimensional SIFT BoVW for images and 10-dimensional LDA for text) to compare directly with existing results. Besides, we also present the cross-media retrieval results based on the 4,096-dimensional CNN visual features[3] and the 100-dimensional LDA [Blei et al. 2003] textual features (we first obtain the textual feature vector based on 500 tokens and then the LDA model is used to compute the probability of each document under 100 topics).

Pascal Sentence:[4] This dataset contains 1,000 pairs of image and text descriptions from 20 categories (50 for each category). We randomly select 30 pairs from each category as the training set and the rest are taken as the testing set. We utilize the 4,096-dimensional CNN visual feature for image representation. For textual features, we first extract the feature vector based on the 300 most frequent tokens (with stop words removed) and then utilize the LDA to compute the probability of each document under 100 topics. The 100-dimensional probability vector is used for textual representation.

INRIA-Websearch: This dataset contains 71,478 pairs of image and text annotations from 353 categories. We remove those pairs that are marked as *irrelevant* and select those pairs that belong to any one of the 100 largest categories. Then, we get a subset of 14,698 pairs for evaluation. We randomly select 70% pairs from each category as the training set (10,332 pairs), and the rest are treated as the testing set (4,366 pairs). We utilize the 4,096-dimensional CNN visual feature for image representation. For textual features, we first obtain the feature vector based on the 25,000 most

[2] http://www.svcl.ucsd.edu/projects/crossmodal/.

[3] The CNN model is pretrained on ImageNet. We utilize the outputs from the second fully connected layer as the CNN visual feature in this article. For more details, please refer to Krizhevsky et al. [2012].

[4] http://vision.cs.uiuc.edu/pascal-sentences/.

Table I. Map Scores for Image and Text Query on the Wikipedia Dataset Based on the Publicly Available Features

Query	PLS	BLM	CCA	SM	SCM	GMMFA	GMLDA	T-V CCA	MDCR
Image	0.207	0.237	0.182	0.225	0.277	0.264	0.272	0.228	**0.287**
Text	0.192	0.144	0.209	0.223	0.226	0.231	**0.232**	0.205	0.225
Average	0.199	0.191	0.196	0.224	0.252	0.248	0.253	0.217	**0.256**

frequent tokens (with stop words removed) and then employ the LDA to compute the probability of each document under 1,000 topics.

For semantic representation, the ground-truth labels of each dataset are employed to construct semantic vectors (10 dimensions for Wikipedia dataset, 20 dimensions for Pascal Sentence dataset, and 100 dimensions for INRIA-Websearch dataset) for pairs of image and text.

4.2. Experimental Settings

In the experiment, Euclidean distance is used to measure the similarity between features in the embedding latent subspace. Retrieval performance is evaluated by mean average precision (mAP), which is one of the standard information retrieval metrics. Specifically, given a set of queries, the average precision (AP) of each query is defined as

$$AP = \frac{\sum_{k=1}^{R} P(k)rel(k)}{\sum_{i=1}^{R} rel(k)},$$

where R is the size of the test dataset. $rel(k) = 1$ if the item at rank k is relevant, and $rel(k) = 0$ otherwise. $P(k)$ denotes the precision of the result ranked at k. We can get the mAP score by averaging the AP for all queries.

4.3. Results

In the experiments, we mainly compare the proposed MDCR with six algorithms, including CCA, Semantic Matching (SM) [Rasiwasia et al. 2010], Semantic Correlation Matching (SCM) [Rasiwasia et al. 2010], Three-View CCA (T-V CCA) [Gong et al. 2013], Generalized Multiview Marginal Fisher Analysis (GMMFA) [Sharma et al. 2012], and Generalized Multiview Linear Discriminant Analysis (GMLDA) [Sharma et al. 2012].

For the Wikipedia dataset, we first compare the proposed MDCR with other methods based on the publicly available features [Rasiwasia et al. 2010], that is, 128-SIFT BoVW for images and 10-LDA for text. We fix $\mu = 0.02$ and $\epsilon = 10^{-4}$ and experimentally set $\lambda = 0.1$, $\eta_1 = 0.5$, and $\eta_2 = 0.5$ for the optimization of I2T, and the parameters for T2I are set as $\lambda = 0.5$, $\eta_1 = 0.5$, and $\eta_2 = 0.5$. The mAP scores for each method are shown in Table I. It can be seen that our method is more effective compared with other common space learning methods. To further validate the necessity to be task specific for cross-media retrieval, we evaluate the proposed method in terms of training a unified V and W by incorporating both linear regression terms in Equation (2) and Equation (3) into a single optimization objective. As shown in Table II, the learned subspaces for I2T and T2I could not be used interchangeably and the unified scheme can only achieve compromised performance for each retrieval task, which cannot compare to the proposed modality-dependent scheme.

As a very popular dataset, Wikipedia has been employed by many other works for cross-media retrieval evaluation. With a different *train*/*test* division, Wu et al. [2014] achieved an average mAP score of 0.226 (Image Query: 0.227, Text Query: 0.224) through a sparse hash model, and Wang et al. [2014] achieved an average mAP score of 0.183 (Image Query: 0.187, Text Query: 0.179) through a deep autoencoder model. Besides, some other works utilized their own extracted features (both for images and

Table II. Comparison Between MDCR and Its Unified Scheme for Cross-Media Retrieval on the Wikipedia Dataset

Wikipedia	MDCR-Equation (2)	MDCR-Equation (3)	Unified Scheme
I2T	**0.287**	0.165	0.236
T2I	0.146	**0.225**	0.216

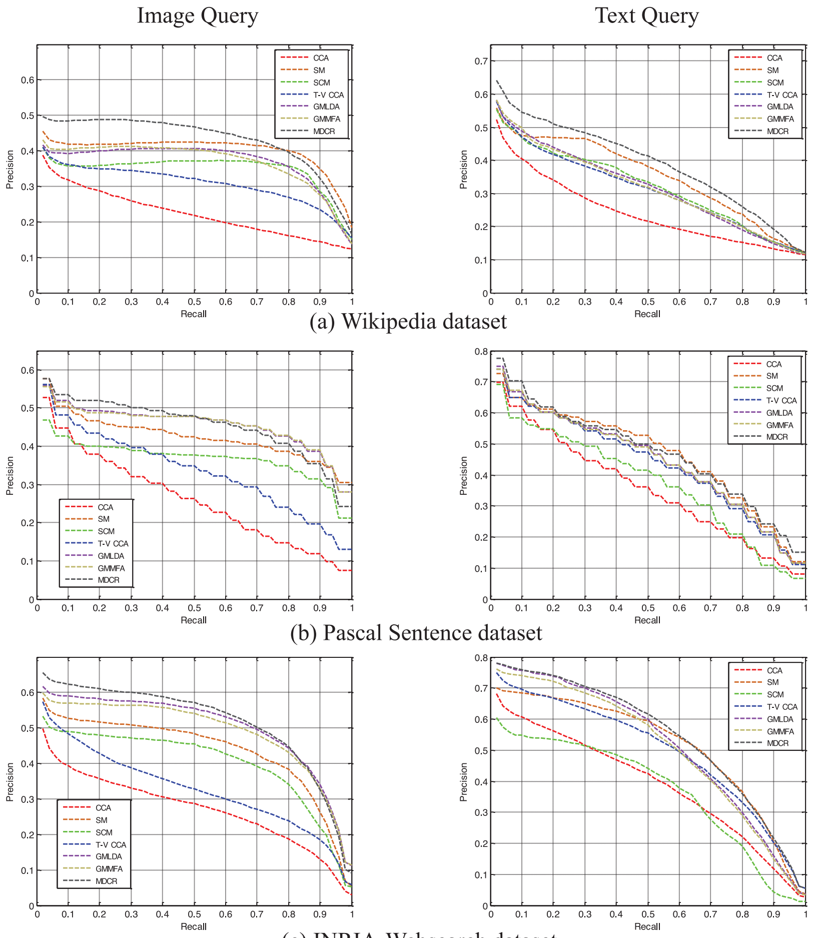


Image Query

Text Query

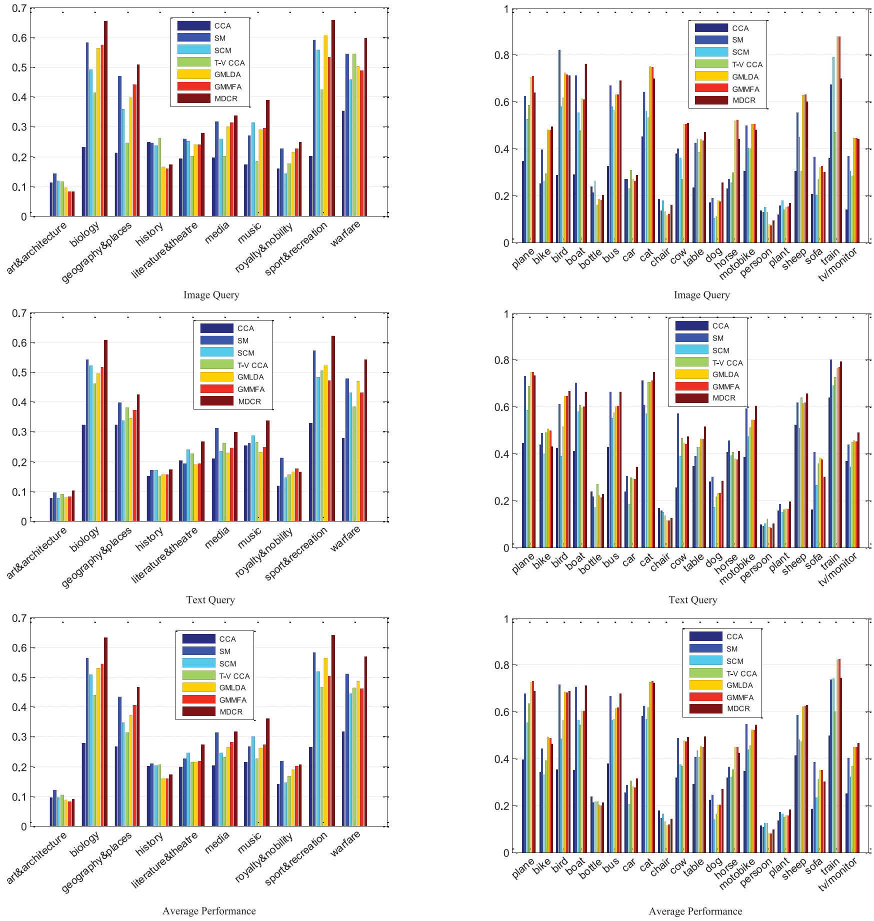
(a) Wikipedia dataset

(b) Pascal Sentence dataset

(c) INRIA-Websearch dataset

Fig. 3. Precision-Recall curves of the proposed MDCR and compared methods.

text) for cross-media retrieval evaluation. To further validate the effectiveness of the proposed method, we also compare MDCR with other methods based on more powerful features, that is, 4,096-CNN for images and 100-LDA for text. We fix $\mu = 0.02$ and $\epsilon = 10^{-4}$, and experimentally set $\lambda = 0.1$, $\eta_1 = 0.5$, and $\eta_2 = 0.5$ for the optimization of I2T and T2I. The comparison results are shown in Table IV. It can be seen that some new state-of-the-art performances are achieved by these methods based on the new feature representations, and the proposed MDCR can also outperform others. In addition, we also compare our method with the recent work of Kang et al. [2014], which utilizes 4,096-CNN for images and 200-LDA for text, in Table III. We can see that the proposed MDCR reaches a new state-of-the-art performance on the Wikipedia dataset. Figure 3 gives the comparisons of Precision-Recall curves, and Figure 4 gives the mAP score of



(a) Wikipeida dataset

(b) Pascal Sentence dataset

Fig. 4. mAP performance for each class on the Wikipedia dataset and the Pascal Sentence dataset.

each category. Figure 5 gives some successful and failure cases of our method. For the image query (the second row), although the query image is categorized into *Art*, it is prevailingly characterized by the human figure (i.e., a strong man), which has been captured by our method and thus leads to the failure results shown. For the text query (the fourth row), there exist many *Warfare* descriptions in the document such as *war, army*, and *troops*, which can be hardly related to the label of the query text, that is, *Art*.

For the Pascal Sentence dataset and the INRIA-Websearch dataset, we experimentally set $\lambda = 0.5$, $\eta_1 = 0.5$, $\eta_2 = 0.5$, $\mu = 0.02$, and $\epsilon = 10^{-4}$ during the alternative optimization process for I2T and T2T. The comparison results can be found in Table IV. It can be seen that our method is more effective compared with others even on a more challenging dataset, that is, INRIA-Websearch (with 14,698 pairs of multimedia data and 100 categories). Figure 3 gives the comparisons of Precision-Recall curves for these two datasets, and Figure 4 gives the mAP score of each category on the Pascal Sentence dataset.

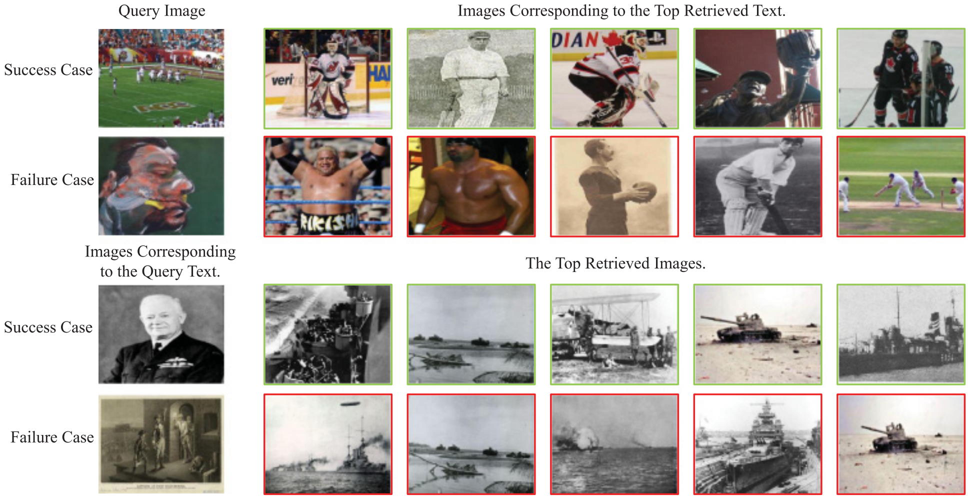

Query Image
Images Corresponding to the Top Retrieved Text.
Success Case
Failure Case
Images Corresponding to the Query Text.
The Top Retrieved Images.
Success Case
Failure Case

Fig. 5. Some successful and failure cases of our method on the Wikipedia dataset. Green and red borders indicate true and false retrieval results, respectively. All the images in this figure are from the Wikipedia dataset [Rasiwasia et al. 2010].

Table III. Cross-Media Retrieval Comparison with Results of Four Methods Reported by Kang et al. [2014] on the Wikipedia Dataset

Query	*GMLDA*	*GMMFA*	*MsAlg*	*LRBS*	MDCR
Image	0.368	0.387	0.373	**0.445**	0.435
Text	0.297	0.311	0.327	0.377	**0.394**
Average	0.332	0.349	0.350	0.411	**0.415**

Table IV. Comparisons of Cross-Media Retrieval Performance

Dataset	Query	CCA	SM	SCM	T-V CCA	GMLDA	GMMFA	MDCR
Wikipedia	Image	0.226	0.403	0.351	0.310	0.372	0.371	**0.435**
	Text	0.246	0.357	0.324	0.316	0.322	0.322	**0.394**
	Average	0.236	0.380	0.337	0.313	0.347	0.346	**0.415**
Pascal Sentence	Image	0.261	0.426	0.369	0.337	**0.456**	0.455	0.455
	Text	0.356	0.467	0.375	0.439	0.448	0.447	**0.471**
	Average	0.309	0.446	0.372	0.388	0.452	0.451	**0.463**
INRIA-Websearch	Image	0.274	0.439	0.403	0.329	0.505	0.492	**0.520**
	Text	0.392	0.517	0.372	0.500	0.522	0.510	**0.551**
	Average	0.333	0.478	0.387	0.415	0.514	0.501	**0.535**

5. CONCLUSIONS

Cross-media retrieval has long been a challenge. In this article, we focus on designing an effective cross-media retrieval model for images and text, that is, using image to search text (I2T) and using text to search images (T2I). Different from traditional common space learning algorithms, we propose a modality-dependent scheme that recommends different treatments for I2T and T2I by learning two couples of projections for different cross-media retrieval tasks. Specifically, by jointly optimizing a correlation term (between images and text) and a linear regression term (from one modal space, i.e., image or text to the semantic space), two couples of mappings are gained for different retrieval tasks. Extensive experiments on the Wikipedia dataset, the Pascal Sentence dataset, and the INRIA-Websearch dataset show the superiority of the proposed method compared with state-of-the-art methods.

REFERENCES

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. 2121–2129.

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. 2013. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106 (2013), 1–24.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664.

S. J. Hwang and K. Grauman. 2010. Accounting for the relative importance of objects in image retrieval. In *Proceedings of the British Machine Vision Conference*. 1–12.

C. Kang, S. Liao, Y. He, J. Wang, S. Xiang, and C. Pan. 2014. Cross-modal similarity learning: A low rank bilinear formulation. *Arxiv Preprint Arxiv:1411.4738* (2014).

J. Krapac, M. Allan, J. Verbeek, and F. Jurie. 2010. Improving web-image search results using query-relative classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1094–1101.

A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1106–1114.

S. Kumar and R. Udupa. 2011. Learning hash functions for cross-view similarity search. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'11)*, Vol. 22. 1360.

X. Lu, F. Wu, X. Li, Y. Zhang, W. Lu, D. Wang, and Y. Zhuang. 2014. Learning multimodal neural network with ranking examples. In *Proceedings of the International Conference on Multimedia*. 985–988.

C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 139–147.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*. 251–260.

R. Rosipal and N. Krämer. 2006. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*. Springer, 34–51.

A. Sharma and D. W. Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 593–600.

A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2160–2167.

J. B. Tenenbaum and W. T. Freeman. 2000. Separating style and content with bilinear models. *Neural Computation* 12, 6 (2000), 1247–1283.

W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang. 2014. Effective multi-modal retrieval based on stacked auto-encoders. In *Proceedings of the International Conference on Very Large Data Bases* 7, 8 (2014), 649–660.

Y. Wei, Y. Zhao, Z. Zhu, Y. Xiao, and S. Wei. 2014. Learning a mid-level feature space for cross-media regularization. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 1–6.

F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. 2013. Cross-media semantic representation via bi-directional learning to rank. In *Proceedings of the International Conference on Multimedia*. 877–886.

F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang. 2014. Sparse multi-modal hashing. *IEEE Transactions on Multimedia* 16, 2 (2014), 427–439.

Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4 (2012), 723–742.

Y. Yang, F. Wu, D. Xu, Y. Zhuang, and L.-T. Chia. 2010. Cross-media retrieval using query dependent search methods. *Pattern Recognition* 43, 8 (2010), 2927–2936.

Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. 2009. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the International Conference on Multimedia*. 175–184.

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

X. Zhai, Y. Peng, and J. Xiao. 2013. Cross-media retrieval by intra-media and inter-media correlation mining. *Multimedia Systems* 19, 5 (2013), 395–406.

L. Zhang, Y. Zhao, Z. Zhu, S. Wei, and X. Wu. 2014. Mining semantically consistent patterns for cross-view data. *IEEE Transactions on Knowledge and Data Engineering* 26, 11 (2014), 2745–2758.

Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao. 2014. Cross-media hashing with neural networks. In *Proceedings of the International Conference on Multimedia*. 901–904.

Received April 2014; revised March 2015; accepted May 2015