Journal of Visual Communication and Image Representation 54 (2018) 213–222

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Joint feature selection and graph regularization for modality-dependent cross-modal retrieval[☆]

Li Wang[a], Lei Zhu[a,*], Xiao Dong[a], Li Liu[a], Jiande Sun[a,b,*], Huaxiang Zhang[a,b,*]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China
[b] Institute of Data Science and Technology, Shandong Normal University, Jinan 250014, Shandong Province, China

ARTICLE INFO

Keywords:
Cross-modal retrieval
Feature selection
Subspace learning
Graph regularization

ABSTRACT

Most existing cross-modal retrieval methods ignore the discriminative semantics embedded in multi-modal data and the unique characteristics of different sub-retrieval tasks. To address the problem, we propose a novel approach in this paper, which is named Joint Feature selection and Graph regularization for Modality-dependent cross-modal retrieval (JFGM). The key idea of JFGM is learning modality-dependent subspaces for different sub-retrieval tasks while simultaneously preserving the semantic consistency of multi-modal data. Specifically, besides to the shared subspace learning between different modalities, a linear regression term is introduced to further correlate the discovered modality-dependent subspace with the explicit semantic space. Furthermore, a multi-model graph regularization term is formulated to preserve the inter-modality and intra-modality semantic consistency. In order to avoid over-fitting problems and select the discriminative features, $l_{2,1}$-norm is imposed on the projection matrices. Experimental results on several publicly available datasets demonstrate the superiority of the proposed method compared with several state-of-the-art approaches.

1. Introduction

Multimedia data has been explosively generated in the Internet during the past few years. These data typically appear in the form of multi-modal pairs (e.g. image-text pair, video-image pair, etc.) on a web page to represent the same object [1]. Fig. 1 illustrates several co-occurrence multimedia documents collected from Pascal Sentence [2]. Cross-modal retrieval, which retrieves semantically relevant results in other modalities for a given query, has attracted great attention in the field of multimedia research [3–9]. In this paper, we mainly investigate two common cross-modal sub-retrieval tasks: image retrieves text (I2T) and text retrieves image (T2I). Different from the existing unimodal data processing technologies [10–14], the similarity measurement among multi-modal data is challenging due to the heterogeneous property of different modalities.

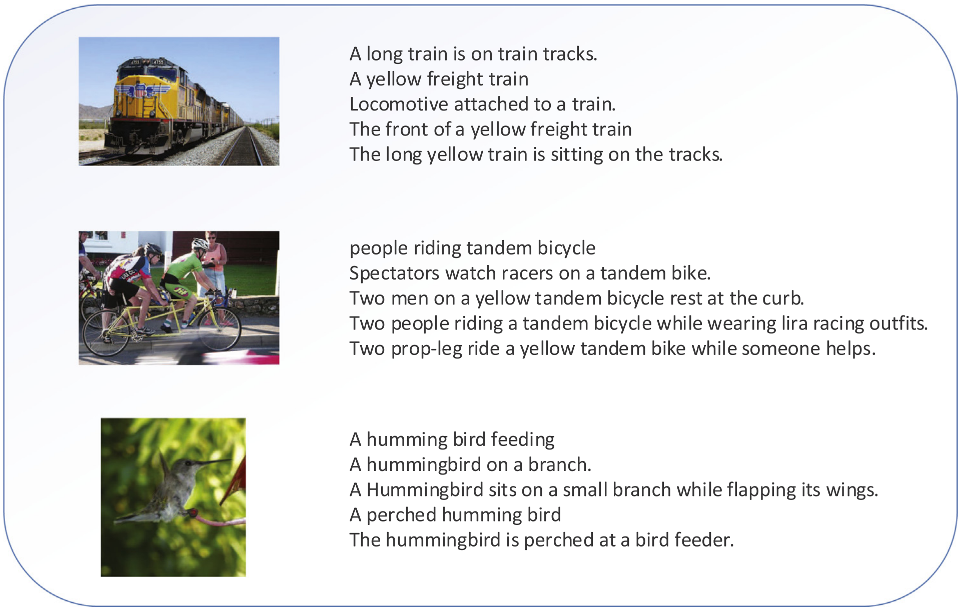
To tackle the above mentioned problem, many solutions have been developed to model the relationships of different modalities by learning a common subspace [15]. Thereby, the content similarity across different modalities can be possibly measured. Generally, subspace-based cross-modal retrieval methods can be divided into two categories: *unsupervised methods and supervised methods* [16]. Unsupervised methods merely leverage the paired multimedia documents from different modalities to learn the shared subspace. Under this circumstance, the explicit high-level semantics cannot be captured during the subspace learning. In contrast, supervised cross-modal retrieval methods learn the discriminative subspace with supervised semantic labels. Since labels can reflect the high-level semantics of multimedia data, the correlation of different modalities can be directly and effectively modeled [17]. Motivated by this advantage, supervised cross-modal retrieval methods have received increasing attentions in recent years [18–20].

However, most existing cross-modal retrieval methods learn the same couple of projection matrices for different sub-retrieval tasks. They unfortunately ignore the important differences of sub-retrieval tasks (I2T and T2I) when learning the common subspace [21]. The projected representation of query cannot be guaranteed to involve enough semantics in the common subspace, and thus certain deterioration of the retrieval performance may be brought. Recently, modality-dependent cross-modal retrieval (MDCR) [22] has been proposed to mitigate this limitation by learning two couples of projections for different sub-retrieval tasks. However, little attention has been paid to preserve the semantic consistency of multi-modal data in the shared subspace. Furthermore, these approaches generally focus on preserving the pairwise similarities, without seriously considering the relations between the embedded feature and the class label [23–25]. Under the

[☆] This paper has been recommended for acceptance by Yue Gao.

[*] Corresponding authors at: School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China.
E-mail addresses: leizhu0608@gmail.com (L. Zhu), jiandesun@hotmail.com (J. Sun), huaxzhang@163.com (H. Zhang).

https://doi.org/10.1016/j.jvcir.2018.05.006
Received 8 August 2017; Received in revised form 26 March 2018; Accepted 9 May 2018
Available online 19 May 2018



A long train is on train tracks.
A yellow freight train
Locomotive attached to a train.
The front of a yellow freight train
The long yellow train is sitting on the tracks.

people riding tandem bicycle
Spectators watch racers on a tandem bike.
Two men on a yellow tandem bicycle rest at the curb.
Two people riding a tandem bicycle while wearing lira racing outfits.
Two prop-leg ride a yellow tandem bike while someone helps.

A humming bird feeding
A hummingbird on a branch.
A Hummingbird sits on a small branch while flapping its wings.
A perched humming bird
The hummingbird is perched at a bird feeder.

Fig. 1. Typical co-occurrence multimedia documents collected from *Pascal Sentence* dataset [2].

circumstance, the cross-modal retrieval system may suffer from suboptimal performance as the shared subspace may involve limited high-level semantics.

To alleviate the aforementioned limitations, in this paper, we propose a novel Joint Feature selection and Graph regularization for Modality-dependent cross-modal retrieval (JFGM). Different from existing techniques, we simultaneously learn different couples of projection matrices for different sub-retrieval tasks, and employ the discriminative semantics latently involved in multi-modal features. Specifically, besides to a correlation analysis term that preserves the pair-wise semantic consistency, a linear regression term is formulated to further correlate the discovered modality-specific subspace with the explicit semantic labels. To avoid noises and over-fitting, $l_{2,1}$-norm is imposed on the projection matrices. Furthermore, a multi-modal graph regularization term is designed to comprehensively preserve the inter-modality and intra-modality similarity relationships. An iterative algorithm is presented to effectively solve the formulated optimization problem. At the stage of online retrieval, the modality-dependent projection matrices of a query are adaptively determined by considering the specific sub-retrieval task. The retrieval system can return more semantically relevant results for queries from different modalities.

The main contributions of our work are summarized as follows:

- We simultaneously consider the important differences of sub-retrieval tasks and the discriminative semantics latently involved in multi-modal features when learning the shared subspace for cross-modal retrieval. To the best of our knowledge, there is no similar work.
- We seamlessly integrate the linear regression term, correlation analysis term, graph regularization term and feature selection term into a joint cross-modal learning framework. These terms interact with each other and embed more semantics in the shared cross-modal retrieval subspace. An iterative algorithm guaranteed with convergence is proposed to solve the formulated optimization problem.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work on cross-modal retrieval. Section 3 describes the details of the proposed approach. In Section 4, we introduce the experiments. Section 5 finally concludes the paper.

2. Related work

With the rapid growth and widespread application of heterogeneous multimedia data, cross-modal retrieval has attracted more and more attention [26,27]. In the past few years, various methods have been developed to improve retrieval performance, such as subspace learning [18,28], graph learning [29,30], etc. Due to the limited space, we mainly review the most relevant work of this paper in this section.

Many subspace learning methods have been proposed to project multi-modal data into a common subspace, so that the data similarity across different modalities can be measured directly. Unsupervised methods learn the common subspace with the paired training samples from different modalities. Canonical correlation analysis (CCA) [28] is widely used method of this kind. It attempts to find a couple of projections to maximize the feature correlation between two different modalities. As a kernelized extension of CCA, kernel canonical correlation analysis (KCCA) [31] maps the original feature into a common subspace via a nonlinear mapping. In [32], CCA is extended further with deep learning framework as Deep CCA. In addition, partial least squares (PLS) [33] is also exploited for cross-modal retrieval by linearly mapping heterogeneous data into a common subspace, where the directions of maximum covariance are identified. The main drawback of these unsupervised cross-modal retrieval methods is that the discovered common subspace cannot capture any explicit high-level semantics.

Supervised cross-modal retrieval methods explore semantic labels to learn more discriminative common subspace. In [18], generalized multiview analysis (GMA) is proposed as a supervised extension of CCA [28]. Wang et al. [19] propose a coupled linear regression framework to learn coupled feature spaces, where cross-modal data matching can be performed. In [29], learning coupled feature spaces (LCFS) [19] is extended as joint feature selection and subspace learning (JFSSL) by preserving the neighbourhood relationships during regression. Considering that real-world data is generally annotated with multiple labels, multi-label CCA (ml-CCA) [34] is proposed to learn the shared subspace with the modality correspondence established by multi-label annotations. Besides, joint latent subspace learning and regression (JLSLR) [20] is developed for cross-modal retrieval by jointly employing label information to preserve semantic structure and minimizing the regression error for different modalities. Recently, many deep learning models [35,36] and hashing methods [37–40] are also developed to project multi-modal data into a common space with