Signal Processing 131 (2017) 249–260

Contents lists available at ScienceDirect

Signal Processing

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

Review

Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval

Leiquan Wang [a,b,*], Weichen Sun [a], Zhicheng Zhao [a], Fei Su [a]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[b] College of Computer and Communication Engineering, China University of Petroleum, Qingdao, China

ARTICLE INFO

Article history:
Received 17 March 2016
Received in revised form
29 June 2016
Accepted 10 August 2016
Available online 11 August 2016

Keywords:
Cross-modal retrieval
Heterogeneous high-order preserving
Correlation learning
Kernel

ABSTRACT

Cross modal (e.g., text-to-image or image-to-text) retrieval has received great attention with the flushed multi-modal social media data. It is of considerable challenge to stride across the heterogeneous gap between modalities. Existing methods project different modalities into a common space by minimizing the distance within the heterogeneous pairs (intra-pair) of the new latent space. However, the relationship among these multi-modal pairs (inter-pair) are neglected, which are beneficial to eliminate the heterogeneity. In this paper, we propose a novel algorithm based on canonical correlation analysis by considering the high-order relationship among pairs (HCCA) for cross-modal retrieval. Supervised with additional semantic labels and unsupervised without semantic labels are simultaneously considered by treating the intra- and inter-pair correlation discriminatively. Moreover, kernel tricks are also performed on HCCA to learn a non-linear projection, termed HKCCA. Extensive experiments conducted on three public datasets demonstrate the superiority of the proposed methods compared with the state-of-the-art approaches in cross modal retrieval.

Contents

* Corresponding author at: School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China.
E-mail addresses: richiewlq@gmail.com (L. Wang), weichen.sun@hotmail.com (W. Sun), zhaozc@bupt.edu.cn (Z. Zhao), sufei@bupt.edu.cn (F. Su).

http://dx.doi.org/10.1016/j.sigpro.2016.08.012
0165-1684/

1. Introduction

Recent years have witnessed the rapid proliferation and widespread adoption of social media. Usually, these data appear in the form of multi-modal pairs (e.g., text–image pair, text–video pair, etc.). Semantic, visual and auditory information are interwoven to describe relevant topics or events from different aspects. As a result, a flexible way has been brought to users for information search and knowledge acquisition. Users input queries for searching complementary information of other heterogeneous modalities, the aim of which is to acquire a more comprehensive understanding about certain topic. Consequently, it has become a meaningful yet challenging issue to retrieve information among heterogeneous modalities, which is known as cross-modal retrieval.

The difficulty for cross-modal retrieval is that we cannot measure cross-modal distance directly with low-level content features which are noisy and heterogeneous [1]. A straight way is to transform the heterogeneous modalities into a shared space, where the similarities of heterogeneous modalities can be measured directly. Canonical correlation analysis (CCA) [2] is the most popular approach, which seeks the optimal basic vectors for two sets of variables to model the correlation. Based on CCA, many variants [3–5] are used to perform cross-modal retrieval. Other algorithms are also proposed to deal with cross-modal problems, such as partial least square (PLS) [6], Bilinear Model (BLM) [7,8], etc. Most of them learn the projection by minimizing the distance of different modalities within the paired samples (intra-pair) during the training phase (see Fig. 3(a)). However, the relationship across the paired samples (inter-pair) are neglected. Particularly for social media (see Fig. 1), the user-contributed tags are incomplete and not always truthful [9]. The one-to-one alignment in intra-pair makes the correlation model sensitive to noise [10]. The many-to-many inter-pair correlation schema is a complement of intra-pair correlation. Inter-pair correlation enriches the diversity of paired samples, which enlarges the scale of training samples indirectly. Existing works (see Fig. 2(b)) [11–14] compensate the defect of only modeling intra-pair correlation by encoding the intra-modal consistency implicitly. However, intra-modal consistency cannot always guarantee inter-modal consistency. The different types of modalities in social media complement each other to express the semantic information. How to construct the relationship among the multi-modal pairs for correlation learning is worth pondering. Hypergraph, which describes the relationship not only between two vertices, but also among three or more vertices containing local grouping information, has been widely used to facilitate image retrieval, classification, segmentation, etc. It should be also helpful to encode the higher-order relationships among multi-modal pairs with a hypergraph for inter-pair correlation.

Semantic gap is an important ingredient that should be considered in eliminating the heterogeneity of multi-modalities. In [15], Zhou et al. proposed a supervised hashing for cross-modal retrieval by preserving inter-modal similarity. Both intra-pair and inter-pair correlation are considered simultaneously and equally. As a result, it achieves satisfactory effects when semantic label information is known. However, in most cases, semantic label information is unavailable. In this situation, many other information should be utilized to construct the relationships among multi-modal pairs for modeling inter-pair correlation, such as context information (e.g., temporal or geographic information) or unsupervised local manifold. Therefore, a scalable algorithm that can adapt to many situations (both supervised, with semantic labels and unsupervised, without semantic labels) is desired. However, the importance of intra-pair and inter-pair correlation needs a thorough consideration and should not be treated equally in all cases. Furthermore, due to the complexity of original features, linear projection may not be adequate for studying correlation among variables. Detecting nonlinear correlation [3,16–18] among data should be also considered.

In this paper, an effective approach based on CCA for cross-modal retrieval is proposed. Different from the view of keeping intra-modal and inter-modal consistency, our proposed method models the correlation with the perspective of intra-pair and inter-pair. The intra-pair and the inter-pair correlation are simultaneously modeled for correlation learning. A hypergraph is

Fig. 1. An example of social image. The text of (a) can be used to describe the image of (b), while the text of (b) can also be used to describe the image of (a). They complement each other. Not only intra-pair but also inter-pair should be considered to narrow the gap of heterogeneous modalities, hence, to improve the adaptability of learned latent space.