Signal Processing: Image Communication 65 (2018) 221–230

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Discriminative correlation hashing for supervised cross-modal retrieval

Xu Lu [a], Huaxiang Zhang [a,b,*], Jiande Sun [a,b], Zhenhua Wang [a,b], Peilian Guo [a,b], Wenbo Wan [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
Hashing
Subspace learning
Discriminant analysis

ABSTRACT

Due to their storage and calculational efficiency, hashing techniques have been used for cross-modal retrieval on large-scale multi-modal data. Cross-modal hashing methods retrieve relevant items of one modality for the query of the other modality by mapping heterogeneous data of different modalities into a common Hamming space, where the binary codes are generated. However, the existing cross-modal hashing methods pay little attention to the discriminative property of the binary codes. In this paper, we propose a novel supervised cross-modal hashing method, named Discriminative Correlation Hashing (DCH), which integrates discriminative property into the hashing learning procedure. DCH introduces the Linear Discriminant Analysis (LDA) to preserve the discriminative property of textual modality and transfers it to the corresponding image modality by the learned unified binary code, thus making data in the common Hamming space much more discriminative. Extensive experimental results demonstrate that DCH outperforms state-of-the-art cross-modal hashing methods.

1. Introduction

With the development of information technology, a large amount of multi-modal data has been uploaded in the Internet, making it a large-scale database. In order to search for useful data from the Internet, many information retrieval techniques have been proposed. Traditional information retrieval techniques commonly retrieve data of one modality by inputting a query of the same modality and the query is commonly represented by keywords. However, these traditional information retrieval techniques are quite limited because they cannot be applied to multi-modal data. Thus, cross-modal retrieval [1–5], which aims at finding the best matches for a different modal query, has attracted much attention. For example, given some photos of the Great Wall, we expect to utilize these photos to retrieve some related textual descriptions [6]. However, cross-modal retrieval still faces some challenging issues. The first one is that heterogeneous data of different modalities cannot be matched directly. The most common methods to tackle this issue are subspace learning approaches [7–17]. This kind of approach aims at projecting heterogeneous data into a common space, where the similarity of multi-modal data can be measured directly. The other issue of cross-modal retrieval is that it is difficult to conduct on a large-scale dataset because of the scalability issue.

Fortunately, hashing based methods [18–27] help to tackle the efficiency limitation. This kind of method maps data into binary codes and measures the similarity by bit-wise XOR operations, which effectively saves the storage and computational costs. Specifically, cross-modal hashing methods map heterogeneous data into a common Hamming space, where the binary codes are generated for cross-modal retrieval. The framework of cross-modal hashing methods is illustrated in Fig. 1. Cross-modal hashing methods can be grouped into two types: unsupervised methods and supervised methods. Unsupervised methods project heterogeneous data into binary codes with cross-correlation between data, while supervised methods incorporate semantic information extracted from class labels into binary code learning procedure.

Moreover, taking cross-correlation and supervised information into account helps to improve the performance of cross-modal retrieval. Specifically, a discriminative distribution is helpful for classification. That is to say, the same class data lie as close as possible, and different class data stay as far as possible. In this paper, we propose a novel cross-modal hashing method, named Discriminant Correlation Hashing (DCH), with full consideration of the discriminative property of multi-modal data. The proposed DCH learns modality-specific hashing functions for two modalities, and maps different modal data into a common Hamming space to generate binary codes. In the common Hamming space, we learn the unified hash code to construct cross-correlation between different modalities. Additionally, we introduce Linear Discriminative Analysis (LDA) to preserve the discriminative property of textual modality and transfer it to the corresponding image modality by the learned unified binary code. Thus, the learned binary

* Corresponding author.
E-mail address: huaxzhang@hotmail.com (H. Zhang).

https://doi.org/10.1016/j.image.2018.04.009
Received 5 June 2017; Received in revised form 10 April 2018; Accepted 11 April 2018
Available online 17 April 2018

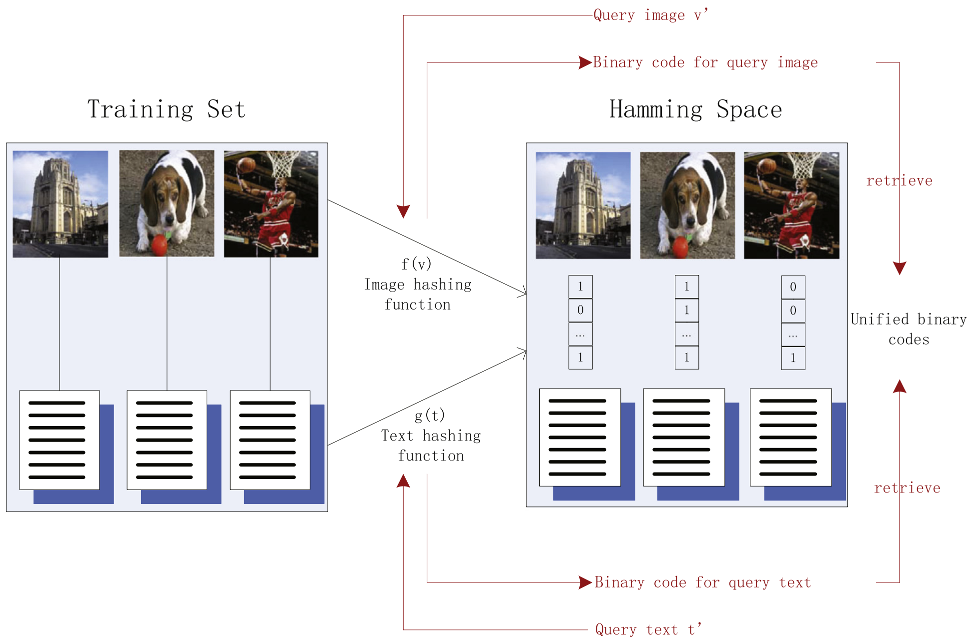

Fig. 1. Framework of cross-modal hashing methods.

codes are more discriminative for classification. The contributions of the proposed DCH are summarized as below:

1. We propose a novel cross-modal hashing framework DCH, which integrates discriminative property into the hashing learning procedure. Based on Linear Discriminant Analysis (LDA), DCH preserves the discriminative property of textual modality and transfers it to the corresponding image modality by the learned unified binary codes.

2. The learned binary codes in the common Hamming space are discriminative and helpful for cross-modal retrieval.

3. Extensive experiments on four popular datasets show that DCH outperforms several state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related cross-modal works, and Section 3 details our Discriminative Correlation Hashing method. Section 4 describes the optimization algorithm for DCH, and Section 5 reports the experimental results on common cross-modal datasets and describes the comparative analysis. The conclusion is presented in Section 6.

2. Related work

Cross-modal retrieval aims at finding the best matches from one modality for a query data from the other modality [28]. The breakthrough point to solve this problem is to establish correlation between heterogeneous features. Hashing based cross-modal methods map heterogeneous features from different modalities into a common Hamming space, where binary codes are generated. Liu et al. [29] proposed a method named CH to learn compact binary codes by simultaneously preserving the entity similarities in each modality and the relationship between different views, so binary codes can be used for cross-modal retrieval tasks. This kind of method has gained much attention in recent years for the great efficiency on large-scale datasets. CH pays more attention on similarity preservation of data, but ignores the quantization quality. Irie et al. [30] gave minimal quantization errors while preserving data similarity, which can boost the retrieval performance in the Hamming space. Considering the semantic gap, Ding et al. [24] proposed a novel model named CMFH, which learns unified binary codes for different modalities in a common semantic space where the correlation between different modalities is established. However the above mentioned methods ignore class labels, which are helpful for classification. Liu et al. [31] proposed a cross-modality hashing algorithm, named Supervised Matrix Factorization Hashing (SMFH), which incorporates semantic labels into the binary code learning procedure, and the learned binary codes preserve similarities among multi-modal data with the help of a graph regularization. Xu et al. [32] proposed a cross-modal hashing framework based on linear classification to learn modality-specific hashing functions for generating unified binary codes, and the learned binary codes represent the features for discriminative classification with class labels. Feng et al. [33] proposed a model involving correspondence autoencoder (Corr-AE), which is constructed by correlating hidden representations of two uni-modal autoencoders. Rafailidis et al. [34] presented a hashing method to exploit the discriminative power of the Dimensions' Value Cardinalities (DVC) of image descriptor. Additionally, Rafailidis et al. [35] proposed a cross-modal hashing method based on a cluster-based joint matrix factorization strategy. Motivated by the fact that optimizing the top of ranking is much applicable in practice, Zhang et al. [36] focused on the low-rank optimization framework named Pairwise–Listwise ranking (PL-ranking).

The distribution of data influences the retrieval performance. Linear Discriminant Analysis (LDA) [37] aims at mapping high-dimensional data into an optimal discriminative space. After projection, the distance of data from different categories becomes large, while data from the same class become much closer. Wei et al. [30] utilized the discriminant analysis on texts to enhance the discriminative semantic characteristic of textual features. This characteristic of textual features is transferred to their corresponding visual features via the correlation analysis process, thus the visual understanding can be enhanced simultaneously [38–42].

Based on the above analysis, we propose a novel cross-modal hashing method named DCH, which incorporates discriminative property into binary code learning procedure. By modality-specific hashing functions, heterogeneous data of different modalities are mapped into a common Hamming space to generate binary codes. In this common Hamming space, we learn unified binary code for different modalities to construct cross-correlation. By utilizing LDA [37], the discriminative characteristic of text modality is maintained and transferred into the data of image modality by the learned unified binary code, which effectively improves the retrieval performance.

The variables are listed in Table 1.