Digital Signal Processing ••• (••••) •••–•••

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Supervised Coarse-to-Fine Semantic Hashing for cross-media retrieval ☆

Tao Yao [a,b], Xiangwei Kong [a,*], Haiyan Fu [a], Qi Tian [c]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116023, China
[b] Department of Information and Electrical Engineering, LuDong University, Yantai, 264025, China
[c] Department of Computer Science, University of Texas at San Antonio, San Antonio, 78249, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Cross-modal retrieval
Coarse-to-fine semantic information
Hashing
Inter-category and intra-category

ABSTRACT

Due to its storage efficiency and fast query speed, cross-media hashing methods have attracted much attention for retrieving semantically similar data over heterogeneous datasets. Supervised hashing methods, which utilize the labeled information to promote the quality of hashing functions, achieve promising performance. However, the existing supervised methods generally focus on utilizing coarse semantic information between samples (*e.g.* similar or dissimilar), and ignore fine semantic information between samples which may degrade the quality of hashing functions. Accordingly, in this paper, we propose a supervised hashing method for cross-media retrieval which utilizes the coarse-to-fine semantic similarity to learn a sharing space. The inter-category and intra-category semantic similarity are effectively preserved in the sharing space. Then an iterative descent scheme is proposed to achieve an optimal relaxed solution, and hashing codes can be generated by quantizing the relaxed solution. At last, to further improve the discrimination of hashing codes, an orthogonal rotation matrix is learned by minimizing the quantization loss while preserving the optimality of the relaxed solution. Extensive experiments on widely used Wiki and NUS-WIDE datasets demonstrate that the proposed method outperforms the existing methods.

1. Introduction

With the rapid development of social media (*e.g.* blogs, social networks), large amounts of heterogeneous data (*e.g.* images, texts, videos) are generated on the web every day. For example, a microblog may contain an image and corresponding texts; videos on YouTube are often associated with related descriptions or labels. Accordingly, heterogeneous data may have the same semantic concepts. It is desirable to support similarity retrieval cross different modalities, *e.g.* using a text to query images and vice versa. However, with the sharp increase of multimedia, the traditional similarity retrieval methods, which take linear time cost, are impractical to be directly applied to large-scale dataset. The hashing method is a favorable way to address this problem [1–16]. Cross-media hashing methods aim at designing compact binary codes by preserving the semantic correlations between different modalities. That is, semantically similar samples are transformed to similar hashing codes. In the retrieval phase, firstly, the query is mapped to hashing code by the learned hashing functions. Then the similarity between the query and samples in the database can be simply calculated by Hamming distances. At last, the retrieved samples will be ranked by Hamming distance in response to query, and the top H ranked samples will be returned.

Recently, it was shown that the performance could be improved by incorporating supervised information into hashing functions learning, *e.g.* labels [12,17,18]. Most of the existing supervised hashing methods simply utilize the coarse semantic information between samples to learn a sharing low-dimensional Hamming space. However, the fine semantic information between samples has not received special attention. More specifically, although bird and cat belong to the category of animal, the semantic correlation between cat and tiger is different from that of bird and tiger to some extent. Hence, the existing supervised cross-media hashing methods do not fully exploit the diversity of semantic information between samples, which may result in poor indexing performance. In this paper, to further leverage the semantic similarity between samples, we formulate to construct a coarse-to-fine semantic similarity matrix to learn a better sharing Hamming space, named Supervised Coarse-to-Fine Semantic Hashing (SCFSH). Then the overall optimization problem can be effectively solved by an iterative algorithm. Furthermore, we propose to learn an orthog-

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail address: kongxw@dlut.edu.cn (X. Kong).

http://dx.doi.org/10.1016/j.dsp.2017.01.003
1051-2004/

onal rotation matrix to improve the quality of hashing codes by minimizing the quantization loss.

The rest of this paper is organized as follows. Section 2 presents the related work. In Section 3, we introduce the proposed supervised coarse-to-fine semantic hashing method and the optimal algorithm. Section 4 shows the experiments on two real world datasets and we conclude the paper in Section 5.

2. Related work

There have been many recent works focusing on cross-media retrieval [12,18–25]. According to whether supervised information is used or not, cross-media hashing methods can be divided into unsupervised cross-media hashing methods and supervised cross-media hashing methods.

Unsupervised hashing methods generally utilize pair-wise feature information of data to learn hashing functions. Canonical Correlation Analysis (CCA) hashing method maximize the correlation between two modalities to learn hashing functions [19]. Latent Semantic Sparse Hashing (LSSH) algorithm embeds texts and images into a latent semantic space by sparse coding and matrix factorization respectively [20]. Collective Matrix Factorization Hashing (CMFH), which utilizes collective matrix factorization to learn an identical representation for pair-wise samples, learns a sharing space by preserve inter-modal similarity [21]. Semantic Topic Multimodal Hashing (STMH) [22] proposes to learn latent topics and hashing codes from texts, and learn a semantic subspace for images. Since supervised information is not used in these methods, in fact the learned sharing space is less discriminative.

In many real applications, not only feature information but also supervised information (*e.g.* similar or dissimilar pair-wise samples, labels) are available. The supervised information generally provided by people contains discriminative semantic information, which is unquestionably beneficial to learn hashing functions. Cross-Modality Metric learning using Similarity-sensitive Hashing (CMMSH) utilizes labels to construct a set of similar and dissimilar pair-wise samples, and then models hashing functions learning problem as binary classification problem with boosting algorithm [18]. Co-Regularized Hashing (CRH) proposes a boosting co-regularization framework to minimize the loss of similar samples, dissimilar samples, and quantization [12]. Iterative Multi-View Hashing (IMVH), which formulates to preserve inter-view and intra-view similarity, minimizes the Hamming distance of similar samples and punishes the dissimilar samples if they have similar binary codes [24]. Semantic Correlation Maximization (SCM) formulates to utilize the semantic labels to construct semantic similarity matrix [17]. Supervised Matrix Factorization Hashing (SMFH), which employs collective matrix factorization to learn a sharing space, incorporates semantic labels into the hashing functions learning [25]. However, only the coarse semantic similarity (*e.g.* similar or dissimilar pair-wise samples, labels) is considered in these methods, ignoring the fine semantic similarity which may degrade the retrieval performance.

Furthermore, due to the discrete constraint on hashing codes, it is hard to achieve optimal solution. Most existing cross-media hashing methods generally drop the discrete constraint of hashing codes to obtain an optimal relaxed continuous solution [1,2,7,10,12,17,18,20,21,26]. After obtaining the relaxed solution, hashing codes can be generated by a simple way, *e.g.* thresholding operation. However, the binary quantization error, which typically limits the performance of the learned hashing functions, is often neglected. Iterative quantization (ITQ) proposed in [19], which aims at learning an orthogonal rotation matrix by minimizing the quantization loss between the Hamming space and the learned continuous sharing space, achieves significant performance. However, this method focuses on single modality and unsupervised hashing method.

Table 1
Terms and notations.

Symbol	Definition
$X^{\mathcal{I}}, X^{\mathcal{T}}$	Images and texts
n	Number of samples
c	The code length of hashing codes
l	The labels of samples
$W_{\mathcal{I}}, W_{\mathcal{T}}$	Hashing functions for images and texts respectively
B,	The hashing codes of samples
S	The fine similarity matrix
O_q^p	The qth subcategory's cluster center in category p
O^k	The cluster center of category k

In this paper, we propose a supervised cross-media hashing method to learn a better Hamming space by the coarse-to-fine semantic similarity matrix. Not only the inter-category but also the intra-category semantic similarity is effectively preserved in the learned Hamming space. After that, an orthogonal rotation matrix is learned to further improve the quality of hashing functions by minimizing binary quantization loss.

3. Supervised coarse-to-fine semantic hashing

In this section, we first introduce the formulation of our method whose flowchart is illustrated in Fig. 1. Then an iterative optimization algorithm is proposed. At last, an orthogonal rotation matrix is learned by minimizing quantization loss to improve the quality of hashing functions.

3.1. Formulation

For the convenience of discussion, we focus on two modalities (*e.g.* images and texts), which can be easily extended to more than two modalities. The terms and notations used in this paper are listed in Table 1.

$X^* \in R^{d_* \times n}$, where $*$ is a placeholder for space $\mathcal{I}$ or $\mathcal{T}$, and d_* is the dimension of feature space (generally $d_{\mathcal{I}} \neq d_{\mathcal{T}}$), and the ith sample from images (texts) is denoted as $X_i^{\mathcal{I}}$ ($X_i^{\mathcal{T}}$). $l \in \{0,1\}^{g \times n}$, where g is the total number of classes. Specifically, $l_{i,j} = 1$ denotes that the jth sample belongs to the ith class, otherwise $l_{i,j} = 0$.

Cross-media hashing methods aim at learning a group of hashing functions to map images and texts to a sharing Hamming space. In this paper, the linear mappings are adopted to act as hashing functions which are defined as follows:

$$h^{\mathcal{I}}(X_i^{\mathcal{I}}) = sgn(W_{\mathcal{I}}^T X_i^{\mathcal{I}}) \tag{1}$$

$$h^{\mathcal{T}}(X_i^{\mathcal{T}}) = sgn(W_{\mathcal{T}}^T X_i^{\mathcal{T}}) \tag{2}$$

where $sgn(\cdot)$ denotes the element-wise sign function, and $W_{\mathcal{I}} \in R^{d_{\mathcal{I}} \times c}$, $W_{\mathcal{T}} \in R^{d_{\mathcal{T}} \times c}$. Moreover, the learned hashing functions, which map the original feature vectors to a sharing Hamming space, should preserve the semantic similarity of samples as much as possible. The label, one of the most important supervised semantic information, is the widely utilized in supervised learning methods [17,27]. In this paper, to leverage the semantic similarity between samples, besides the labeled information, feature information is utilized to further refine the semantic similarity matrix. The fine semantic information has been applied in many fields achieving promising performance, such as activity recognition [28], natural language processing [29], classification [30,31].

In the real world, the differences of inter-category and intra-category are divergent. Occasionally, the intra-category divergence is larger than that of inter-category. How to model the problem is

Please cite this article in press as: T. Yao et al., Supervised Coarse-to-Fine Semantic Hashing for cross-media retrieval, Digit. Signal Process. (2017), http://dx.doi.org/10.1016/j.dsp.2017.01.003