Discriminant Cross-modal Hashing

Xing Xu
University of Electronic
Science and Technology of
China
Chengdu, China
xing.xu@uestc.edu.cn

Fumin Shen
University of Electronic
Science and Technology of
China
Chengdu, China
fumin.shen@gmail.com

Yang Yang
University of Electronic
Science and Technology of
China
Chengdu, China
dlyyang@gmail.com

Heng Tao Shen
University of Electronic
Science and Technology of
China
Chengdu, China
The University of Queensland
Brisbane, Australia
shenhengtao@hotmail.com

ABSTRACT

Hashing based methods have attracted considerable attention for efficient cross-modal retrieval on large-scale multimedia data. The core problem of cross-modal hashing is how to effectively integrate heterogeneous features from different modalities to learn hash functions using available supervising information, e.g., class labels. Existing hashing based methods generally project heterogeneous features to a common space for hash codes generation, and the supervising information is incrementally used for improving performance. However, these methods may produce ineffective hash codes, due to the failure to explore the discriminative property of supervising information and to effectively bridge the semantic gap between different modalities. To address these challenges, we propose a novel hashing based method in a linear classification framework, in which the proposed method learns modality-specific hash functions for generating unified binary codes, and these binary codes are viewed as representative features for discriminative classification with class labels. An effective optimization algorithm is developed for the proposed method to jointly learn the modality-specific hash function, the unified binary codes and a linear classifier. Extensive experiments on three benchmark datasets highlight the advantage of the proposed method and show that it achieves the state-of-the-art performance.

CCS Concepts

•Information systems → Multimedia and multimodal retrieval;

ICMR'16, June 06-09, 2016, New York, NY, USA
© 2016 ACM. ISBN 978-1-4503-4359-6/16/06...$15.00
DOI: http://dx.doi.org/10.1145/2911996.2912056

Keywords

Cross-modal retrieval; cross-modal hashing; discriminant analysis

1. INTRODUCTION

Nowadays, the fast growth of multimedia data on the Internet has significantly increased the demand for more sophisticated multimedia indexing [10] and retrieval technologies [11, 6]. Especially, the cross-modal retrieval problem [1, 9] that aims at matching one modal data to the other modal data, has gained much attentions due to the heterogeneous source of the multimedia data, i.e., texts, images and videos. However, different modal data cannot be directly matched due to the intrinsic diversity in different modalities.

To tackle this problem, several hashing-based approaches [7, 13, 15, 3] have garnered considerable interest and shown promising performance on cross-modal retrieval problem. These hashing based methods find the linear projections to embed the heterogeneous data into a common Hamming space, where the multi-modal features are indexed as binary codes. These methods can handle large-scale data with both low memory cost and computational cost, because the learned hash codes can be compactly stored and Hamming distance can be fast calculated with bit-wise XOR operations. Nevertheless, these methods are generally unsupervised since the valuable class label information has not been leveraged yet.

Indeed, utilizing the class label information has been found helpful to construct the correlations across different modalities, e.g., reducing the semantic gap between the low-level visual features and high level semantic descriptions [8]. Several recent studies [2, 14, 4] leverage supervising information like class labels or semantic affinities of training data incrementally to further improve retrieval performance. These supervised approaches can often achieve higher accuracy than the aforementioned unsupervised ones. However, one limitation of these supervised approaches is that the discrimination of supervising information is implicitly captured for hash function learning. It may inhibit these methods from learning effective hash codes that well capture the correlations of different modalities and incorporate the discrimination of supervising information.

To address the above challenges, we propose a novel cross-modal hashing method that learns unified binary codes from different modalities more effectively by explicitly capturing the discrimination of class labels. In the proposed method, the unified binary codes are produced by the modality-specific hash functions for one instance consisting of different modalities. The modality-specific hash functions are proposed to account the unique structure of each modality. With the help of class labels as the supervising information, the resulting binary codes are designed to be the optimal representation of the multi-modal data for classification. Unlike previous works [4, 14, 15] that utilize the class labels to generate semantic affinities for approximating hash codes or to cast independently learning each bit of hash codes to binary classification problem, the class label information is naturally leveraged in the unified binary codes learning process of the proposed method. Due to the NP-hard optimization problem in the proposed method, an efficient approximation algorithm is developed to iteratively obtain the quasi-optimal binary codes under the classification framework.

The main contributions of our work can be summarized as follows: 1) We propose a novel hashing method for cross-modal retrieval, in which the unified binary codes and a linear classifier are jointly learned with the modality-specific hash functions. 2) An efficient algorithm is developed for the proposed method to iteratively obtain the approximate optimized solution. 3) The proposed method is extensively evaluated on three benchmark dataset and the results show its superiority to several other state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 presents the details of the proposed method. Extensive experiments on three cross-modal datasets are depicted in Section 3. Lastly, the conclusions are made in Section 4.

2. PROPOSED METHOD

2.1 Problem Formulation

Suppose that we have a set of n multi-modal training instances $\mathcal{X} = \{x_i\}_{i=1}^n$, and each instance $x_i = (v_i, t_i)$ has a class label vector $y_i \in \mathbb{R}^{c\times 1}$; $v_i \in \mathbb{R}^{m\times 1}$ is the image feature vector, and $t_i \in \mathbb{R}^{d\times 1}$ is the text feature vector. Here we denote $V = \{v_i\}_{n=1}^m \in \mathbb{R}^{m\times n}$, $T = \{t_i\}_{i=1}^d \in \mathbb{R}^{d\times n}$ and $Y = \{y_i\}_{i=1}^c \in \mathbb{R}^{c\times n}$ as the image, text and label matrices of $\mathcal{X}$. We aim to learn binary codes matrix $B = \{b_i\}_{i=1}^n \in \{-1,1\}^{L\times n}$, where b_i is the unified L-bits hash codes for instance x_i.

In our work, we introduce the modality-specific hash functions $F_V(\cdot)$ and $F_T(\cdot)$ for image and text modalities, respectively. Specifically, given the multi-modal feature vectors v_i and t_i of the i-th instance, these two hash functions are correspondingly defined as:

$$F_V(v_i) = P_V^\top v_i, \quad F_T(t_i) = P_T^\top t_i, \tag{1}$$

where $P_V \in \mathbb{R}^{m\times L}$ and $P_T \in \mathbb{R}^{m\times L}$ are the linear projection matrices that map the original feature representations of v_i and t_i onto the low-dimensional latent spaces.

Furthermore, we consider that one instance x_i consisting of two modalities has unified binary codes b_i. Similar to [5], to leverage the supervised label information, b_i is expected to be ideal feature vector to represent original multi-modal data for classification. Given the binary codes b_i and label vector y_i for x_i , we import the simple linear classification form with binary constraints on b_i as in [5]:

$$\min_B \sum_{i=1}^n \|y_i - W^\top b_i\|^2 + \lambda\|W\|^2, \quad s.t. \;\; b_i \in \{-1,1\}^L. \tag{2}$$

Here $W \in \mathbb{R}^{L\times c}$ is the linear classifier, $\|\cdot\|$ is the L_2 norm for vectors and Frobenius norm for matrices, and λ is the regularization parameter.

Moreover, b_i is also desired to bridge the semantic gap between original data v_i and t_i of instance x_i. Thus we assume that the projected spaces of $F_V(v_i)$ and $F_T(t_i)$ are commonly shared as the Hamming space, where the unified binary codes b_i can be generated. We follow the discrete optimization scheme in [5] to preserve the binary constraints on b_i and to fit the errors of binary codes b_i with $F_V(v_i)$ as $\|b_i - F_V(v_i)\|$, with $F_T(t_i)$ as $\|b_i - F_T(t_i)\|$, respectively.

Finally, the optimization problem in the proposed method is formulated by jointly learning the modality-specific hash functions, unified binary codes and a linear classifiers, as:

$$\begin{aligned}\min_{B,W,F_V,F_T} \sum_{i=1}^n \|y_i - W^\top b_i\|^2 + \mu_V \sum_{i=1}^n \|b_i - F_V(v_i)\|^2 \quad (3)\\ + \mu_T \sum_{i=1}^n \|b_i - F_T(t_i)\|^2 + \lambda\|W\|^2, \quad s.t. \;\; b_i \in \{-1,1\}^L,\end{aligned}$$

where μ_V and μ_T are the penalty parameters. From Eq. 3 it can be learned that the first term demands good binary codes for supervised classification to preserve the discrimination, and the last two terms ensure the modality-specific hash functions are simultaneously optimized to fit the binary codes with less errors.

2.2 Optimization Algorithm

To make the derivation more intuitive, Eq. 3 can be rewritten in matrix form as follows:

$$\begin{aligned}\min_{B,W,F_V,F_T} \|Y - W^\top B\|^2 + \mu_V\|B - F_V(V)\|^2 \quad (4)\\ + \mu_T\|B - F_T(T)\|^2 + \lambda\|W\|^2, \quad s.t. \;\; B \in \{-1,1\}^{L\times n}.\end{aligned}$$

The optimization problem in Eq. 4 is non-convex with respect to matrices variables B, W, F_V and F_T and difficult to solve. However, it is convex with respect to any one of the four variable while fixing the other three. Therefore, Eq. 4 can be solved by an iterative framework with the following steps until convergency is reached.

Step1: Learn projection matrices P_V and P_T by fixing other variables, the problem in Eq. 4 becomes

$$\begin{aligned}\min_{F_V} \mu_V\|B - F_V(V)\|^2 = \min_{P_V} \mu_V\|B - P_V V\|^2,\\ \min_{F_T} \mu_T\|B - F_T(T)\|^2 = \min_{P_T} \mu_T\|B - P_T T\|^2,\end{aligned} \tag{5}$$

respectively, which can be computed by regression as:

$$\begin{aligned}P_V = (VV^\top)^{-1}VB^\top,\\ P_T = (TT^\top)^{-1}TB^\top.\end{aligned} \tag{6}$$

Step2: Learn classification matrix W by fixing other variables, the problem in Eq. 4 becomes

$$\min_W \|Y - W^\top B\|^2 + \lambda\|W\|^2, \; s.t. \; B \in \{-1,1\}^{L\times n}, \tag{7}$$

which is a simple regularized least squares problem. Thus the closed-form solution of W can be derived as:

$$W = (BB^{\top} + \lambda I)^{-1} BY^{\top}. \tag{8}$$

Step3: Learn binary codes matrix B by fixing other variables, the problem in Eq. 4 becomes:

$$\min_{W} \|Y - W^{\top}B\|^2 + \mu_V \|B - F_V(V)\|^2 + \mu_T \|B - F_T(T)\|^2,$$
$$s.t. \quad B \in \{-1, 1\}^{L \times n}. \tag{9}$$

The problem in Eq. 9 is NP-hard for directly optimizing the optimal binary codes B. However, we can solve a relaxed problem through discarding the discrete constraints in Eq. 9, then the gradient regarding to B can be derived as:

$$g(B) = W(Y - W^{\top}B) + \mu_V(B - P_V V) + \mu_T(B - P_T T). \tag{10}$$

After setting $g(B)$ to zero, we can get the continuous closed-form solution of B, which is:

$$B = [WW^{\top} + (\mu_V + \mu_T)I]^{-1}(WY + \mu_V P_V V + \mu_T P_T T). \tag{11}$$

Finally, the approximate binary codes for the training instances can be obtained by quantization, as $sgn(B)$.

2.3 Discussion

In each iteration of the training phase described above, the time complexity is about $\mathcal{O}(k^2 n)$, where $k = \max\{m, d, L\}$. Thus the training time of the proposed method is linear to the size of training set, hence is very efficient and scalable for large-scale training data.

During the test phase, given a new query instance x', generating its hash codes b' relying on its components. Specifically, when x' only contains data of one modality, i.e. $x' = (v',)$ or $x' = (t',)$, its hash codes can be computed as $b' = sgn(F_V(v'))$ or $b' = sgn(F_T(t'))$; when x' consists of data of two modalities, i.e. $x' = (v', t')$, its hash codes is computed as the quantization of Eq. 11.

3. EXPERIMENT

3.1 Experimental Settings

To verify the effectiveness of the proposed method, we conduct experiments on three real-world datasets: Wiki, MIRFlickr and NUS-WIDE, for cross-modal retrieval problem. All these datasets consist of both image and text modalities, for cross-modal retrieval. Some general statistics of these datasets and the query/retrieval splits in our experiments are illustrated in Table 1. dataset Regarding the feature representations of image and text modalities, for each instance on Wiki dataset, its image is represented by a 128-dimension SIFT feature vector and its text as a 10-dimension topic vector generated by latent Dirichlet allocation (LDA); on MIRFlickr dataset, its image is represented by 150-dimensional edge histogram, and its text as a 500-dimensional feature vector generated from PCA on its index tagging vector; on NUS-WIDE dataset, its image is represented by a 500-dimension SIFT feature vector and its text as an index vector of selected top 1,000 tags.

	Wiki	MIRFlickr	NUS-WIDE
Retrieval Size	2,173	15,902	184,711
Training Set	2,173	5,000	5,000
Query Set	693	836	1,866
Labels	10	24	10

Table 1: General statistics of three datasets.

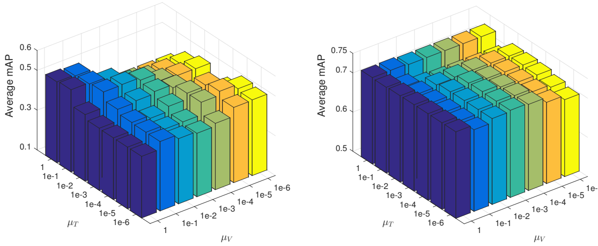
Furthermore, like the setting in [4], to reduce the computational cost on large-scale datasets MIRFlickr and NUS-WIDE, we randomly select 5,000 instances from their retrieval database to train hash functions, and then these hash functions are applied to the other instances in the databases to generate hash codes for them; for Wiki dataset, we take the entire retrieval set as the training data because of its small size. Moreover, for query instances, we follow the scheme in Section 2.3 to generate unified hash codes for these out-of-sample instances.

We compare the proposed methods with various state-of-the-art cross-modal hashing methods, including unsupervised ones IMH [7], LSSH [15], CMFH [3] and supervised ones CMSSH [2], CRH [14], SCM$_{Seq}$ [12], SePH$_{km}$ [4]. We use the codes and suggested parameters published by the respective authors. We perform 10 runs for all these methods and take the average performance for comparison. All the experiments are performed on Matlab platform installed in a desktop which has 4-core 3.2GHz CPUs with 16GB RAM.

Two types of cross-modal retrieval tasks are conducted in our experiments: 1) *Text2Img* that uses text as query to search similar images and 2) *Img2Text* that uses images as query to search similar texts. The retrieval performance is evaluated by the widely-used *mean average precision* (mAP), which is an overall metric on a set of queries. Note that a larger mAP indicates better performance that relevant ones have high rank in the retrieved instances.

3.2 Parameter Sensitive Analysis

The parameters μ_V and μ_T in the proposed method control the error of the learned binary codes of different modalities, in this subsection we analyze their effects on the qualities of the learned hash codes during training. By prefixing the code length as 16 bits, we vary both μ_V and μ_T in $\{10^{-6}, 10^{-5}, ..., 1\}$ and assess their qualities with two retrieval tasks on the training set of Wiki and MIRFlickr datasets. Figure 1 shows the average mAP of *Img2Text* and *Text2Img* tasks under different settings of μ_V and μ_T on the two datasets. It can be observed that the proposed method can achieve stable performance under a wide range of values of μ_V and μ_T when they are considerably large and comparable, e.g., μ_V and μ_T are from the range $[10^{-2}, 1]$. Therefore, to make comprehensive comparison with other baselines, for the proposed method, we empirically set λ to 1, μ_V and μ_T both to 10^{-1} in the following experiments.



(a) On Wiki (b) On MIRFlickr

Figure 1: The average mAP under different settings of the parameters μ_V and μ_T.

Table 2: Overall comparison of mAP values on the three datasets. The top panel is the performance for Img2Text task and the bottom panel is for Text2Img task.

Method	Wiki				MIRFlickr				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMSSH [2]	0.1976	0.1999	0.1889	0.1907	0.5520	0.5539	0.5506	0.5559	0.4686	0.4768	0.4741	0.4637
IMH [7]	0.1869	0.1938	0.1873	0.1834	0.6088	0.6063	0.5977	0.5857	0.4543	0.4469	0.4371	0.4285
LSSH [15]	0.2141	0.2216	0.2218	0.2211	0.5784	0.5804	0.5797	0.5816	0.3900	0.3924	0.3962	0.3966
CMFH [3]	0.2403	0.2474	0.2533	0.2572	0.6273	0.6343	0.6410	0.6451	0.5150	0.5208	0.5256	0.5285
CRH [14]	0.2031	0.1966	0.1982	0.1943	0.5826	0.5745	0.5726	0.5718	0.5136	0.5079	0.4996	0.5013
SCM$_{Seq}$ [12]	0.2210	0.2337	0.2442	0.2596	0.6237	0.6343	0.6448	0.6489	0.4842	0.4941	0.4947	0.4965
SePH$_{km}$ [4]	0.2787	0.2956	0.3064	0.3134	**0.6723**	0.6771	0.6783	0.6817	**0.5421**	0.5499	0.5537	0.5601
Proposed	**0.3253**	**0.3340**	**0.3443**	**0.3558**	0.6687	**0.6911**	**0.7015**	**0.6992**	0.5292	**0.5595**	**0.5622**	**0.5643**
CMSSH [2]	0.2495	0.2360	0.2348	0.2382	0.6010	0.6048	0.6029	0.6041	0.4635	0.4685	0.4594	0.4556
IMH [7]	0.3731	0.3967	0.3720	0.3519	0.5996	0.5999	0.5914	0.5824	0.4546	0.4497	0.4415	0.4304
LSSH [15]	0.5031	0.5224	0.5293	0.5346	0.5898	0.5927	0.5932	0.5932	0.4286	0.4248	0.4248	0.4175
CMFH [3]	0.5705	0.5909	0.6022	0.6110	0.6095	0.6134	0.6184	0.6199	0.5952	0.6133	0.6282	0.6283
CRH [14]	0.2634	0.2622	0.2631	0.2625	0.5944	0.5913	0.5838	0.5811	0.5273	0.5114	0.5033	0.4977
SCM$_{Seq}$ [12]	0.2134	0.2366	0.2479	0.2573	0.6133	0.6209	0.6295	0.6340	0.4536	0.4620	0.4630	0.4644
SePH$_{km}$ [4]	0.6318	0.6577	0.6646	0.6709	0.7197	0.7271	0.7309	0.7354	0.6302	0.6425	0.6506	0.6580
Proposed	**0.7014**	**0.7002**	**0.7165**	**0.7231**	**0.7542**	**0.7753**	**0.7968**	**0.7991**	**0.6555**	**0.6742**	**0.6773**	**0.6796**

3.3 Overall Comparison with Baselines

For the proposed two methods and all baselines, we report their cross-modal retrieval performance on the three dataset in Table 2 with code length ranging from 16-bit to 128-bit. Note that since we use the same experimental settings as in SePH$_{km}$ [4], here we refer to the experimental results of LSSH, SCM$_{Seq}$ and SePH$_{km}$ in [4]. From Table 2, we have the following observations: 1) On all three datasets, in most cases the proposed method significantly outperforms all baselines with different hash code length on different retrieval tasks and achieves the state-of-the-art performance, which well demonstrates its effectiveness. The superiority of the proposed method can be attributed to its advantaged hashing framework that integrates unified hash code generation and classifier training. The generated hash codes well incorporate the correlation between different modalities and preserve the discrimination of the class label information; 2) Generally, the mAP scores of the proposed method consistently increase when the code length becomes longer. It also indicates that they can encode more discriminative information and better preserve the semantic relations between different modalities by utilizing longer hash codes.

4. CONCLUSION

In this paper, we proposed a novel supervised hashing method for cross-modal retrieval. The proposed method was formulated in a linear classification framework, where the supervised label information was naturally leveraged to preserve the discrimination during the unified binary code learning process. In this framework, the modality-specific hash functions, the unified binary codes and a linear classifier were learned in a joint optimization problem. Due to the intractability of the optimization problem in the proposed method, we proposed an efficient approximate algorithm to obtain the quasi-optimal solution. The proposed method was validated on three benchmark datasets, where the state-of-the-art results validated its efficacy.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Project 61502081 and Project 61572108, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007.

5. REFERENCES

[1] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *ACM International Conference on Multimedia Retrieval*, 2014.

[2] M. Bronstein, A. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2010.

[3] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] Z. Lin, G. , M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] F. Shen, C. Shen, W. Liu, and H. T. Shen. Supervised Discrete Hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[6] F. Shen, C. Shen, Q. Shi, A. Van Den Hengel, and Z. Tang. Inductive hashing on manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[7] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 785–796, 2013.

[8] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for scalable image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3424–3431, 2010.

[9] X. Xu, Y. Yang, A. Shimada, R.-i. Taniguchi, and L. He. Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 847–850, 2015.

[10] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Transation on Multimedia*, 16:1677–1689, 2014.

[11] Y. Yang, H. Zhang, M. Zhang, F. Shen, and X. Li. Visual coding in a semantic hierarchy. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 59–68, 2015.

[12] D. Zhang and W.-J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI Conference on Artificial Intelligence*, 2014.

[13] D. Zhang, F. Wang, and L. Si. Composite hashing with multiple information sources. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR '11, pages 225–234, 2011.

[14] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In *Advances in neural information processing systems*, pages 1385–1393, 2012.

[15] J. Zhou. Latent Semantic Sparse Hashing for Cross-Modal Similarity Search. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 415–424, 2014.