HDMFH: HYPERGRAPH BASED DISCRETE MATRIX FACTORIZATION HASHING FOR MULTIMODAL RETRIEVAL

*Jing Gao, Wenjun Zhang, Zhikui Chen, Fangming Zhong**

School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116620, China
gaojing@dlut.edu.cn, junlian95@mail.dlut.edu.cn, zkchen@dlut.edu.cn, fmzhong@dlut.edu.cn

ABSTRACT

In recent years, hashing based cross-modal retrieval methods have attracted considerable attention for the high retrieval efficiency and low storage cost. However, most of the existing methods neglect the high-order relationship among data samples. In addition, most of them can only deal with two modalities, e.g., image and text, without discussing the scenario of multiple modalities. To address these issues, in this paper, we propose a novel cross-modal hashing method, named Hypergraph Based Discrete Matrix Factorization Hashing (HDMFH), for multimodal retrieval. Different from most previous approaches, our method based on hypergraph regularization and matrix factorization can handle the cross-modal retrieval of more than two modalities, which is known as multimodal retrieval. Extensive experiments demonstrate that HDMFH outperforms the state-of-the-art cross-modal hashing methods.

Index Terms— Cross-modal retrieval, multimodal retrieval, hashing, hypergraph learning

1. INTRODUCTION

Cross-modal retrieval has attracted considerable attention due to the massive growth of multimedia data such as text, images, audios, and videos [1, 2, 3, 4]. Taking text and image modalities as an example, cross-modal retrieval uses one modality (e.g., texts) as a query to search another modality (e.g., images) that shares the similar semantics with the query item. It has been widely investigated and applied in computer vision [5], text mining [6], and information retrieval [7], and how to effectively perform cross-modal similarity search has become a hot research topic.

Recently, the hashing-based cross-modal retrieval methods have been widely studied due to the low storage overhead and fast query speed [8, 9, 10, 11]. Most of the previous cross-modal hashing methods can be briefly divided into two categories: unsupervised and supervised methods. The unsupervised methods learn the hashing function and the binary codes by maximizing the intra- and inter-modality similarity of training data without supervised labels. Representative examples include Inter-Media Hashing (IMH) [12], Collective Matrix Factorization Hashing (CMFH) [10], and Unsupervised Deep Cross-Modal Hashing (UDCMH) [13]. While, the supervised ones enhance the common semantic relationship by utilizing the label information of training data. Typical supervised methods include Cross-View Hashing (CVH) [14], Supervised Matrix Factorization Hashing (SMFH) [11], Semantic Preserved Hashing (SePH) [15], and Cross-Modal Discrete Hashing (CMDH) [9].

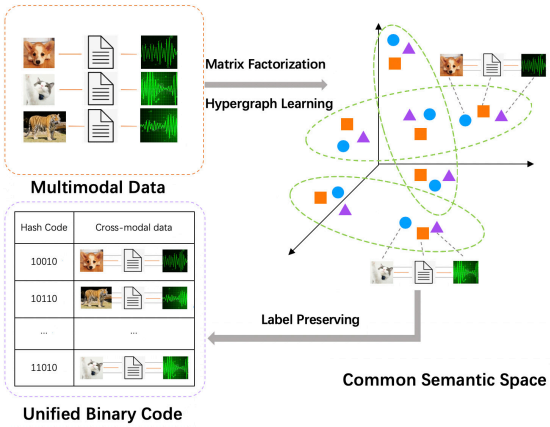

Fig. 1. Framework of the proposed HDMFH.

Although great progress has been made in recent years, most of the existing cross-modal hashing methods [7, 11, 15, 16] ignore the high-order relationship among data samples. They only consider the pairwise relationship between two samples, which can not fully describe the semantics of the modality, thus decreasing the discriminative property of representations. In addition, most of them can handle the cross-modal retrieval with only two modalities. The scenario of cross-modal retrieval with more than two modalities, i.e. multimodal retrieval has not yet been investigated well. Several methods such as Fusion Similarity Hashing (FSH) [7] and Scalable Discrete Matrix Factorization Hashing (SCRATCH)

*Corresponding author. This work is supported by the National Key Research and Development Program of China (No. 2018YFC0831305), the Nature Science Foundation of China (No. 61672123, No. 61602083), and the Doctoral Scientific Research Foundation of Liaoning Province (No. 20170520425).

[17] claim that they can be easily extended to multimodal scenario. However, none of them have discussed the difference and validated via experiments.

To address the aforementioned challenges, in this paper, we present a novel cross-modal hashing method for multimodal retrieval, namely, **H**ypergraph Based **D**iscrete **M**atrix **F**actorization **H**ashing (**HDMFH**). Fig. 1 depicts the working flow of the proposed HDMFH. The goal of HDMFH is to push multimodal data to a common semantic space and to maintain the high-order relationship among samples in each modality simultaneously. Because hypergraph can model the high-order relationship among instances, we employ a hypergraph regularization term for each modality to capture the high-order relationship of samples, which can help enforce the discriminative property of learned common semantic representations. In addition, we use matrix factorization and supervised semantic labels to bridge the semantic gap across different modalities. Such connection without constraints of any two modalities makes our method scalable to multiple modalities. Moreover, HDMFH is a two-step hashing method that consists of unified binary codes learning and hashing function learning resulting in a high degree of flexibility of our method. A large number of experiments on four widely used datasets show that our method is superior to the state-of-the-art cross-modal hashing methods in both cross-modal retrieval and multimodal retrieval.

The main contributions are summarized as follows:

- We propose a novel hypergraph based discrete matrix factorization hashing method for multimodal retrieval, which can efficiently capture the correlations across different modalities to handle cross-modal retrieval with more than two modalities.
- Different from the previous cross-modal hashing methods, we employ hypergraph learning to model the high-order relationship among different samples in each modality, which further improves the discriminative property of learned common sematic representations.
- Extensive experiments are conducted on three cross-modal and one multimodal dataset. The results demonstrate that HDMFH is superior to the state-of-the-art cross-modal hashing methods.

The rest of this paper is organized as follows. Section 2 introduces the details of our proposed HDMFH. Section 3 includes the experiments of cross-modal retrieval conducted on four datasets with settings, results, and analysis. Finally, the conclusion is presented in Section 4.

2. PROPOSED METHOD

2.1. Notations

In this paper, we use bold uppercase letters to represent matrices and bold lowercase letters to represent vectors. Given m different modalities data $\mathbf{X}^{(t)} = \{\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}, \cdots, \mathbf{x}_n^{(t)}\} \in R^{d_t \times n}, t = 1, 2, \cdots, m$, where d_t represents the feature space dimensionality corresponding to the t-th modality, and n is the number of training instances. Without loss of generality, we assume that the data are zero-centered in each modality, i.e., $\sum_{i=1}^{n} \mathbf{x}_i^{(t)} = 0$. Let $\mathbf{Y} \in \{0,1\}^{c \times n}$ be the label matrix, where c is the number of classes and $\mathbf{Y}_{ki} = 1$ if $\mathbf{x}_i$ belongs to class k and 0 otherwise. $\mathbf{B} \in \{-1,1\}^{r \times n}$ is the to-be-learned binary codes matrix, where r is the length of the hash codes. $\|\cdot\|_F$ denotes the Frobenius norm, $Tr(\cdot)$ is the trace operation, and $sgn(\cdot)$ is an element-wise sign function defined as follows,

$$sgn(x) = \begin{cases} 1 & x > 0; \\ -1 & x \leq 0. \end{cases} \tag{1}$$

2.2. Objective Function of HDMFH

Latent Semantic Representation Learning. In order to obtain the latent semantic representation of different modalities, here we borrow the idea of collective matrix factorization [18]. As for multimodal data, they are different descriptions of the same objects. Hence, they usually share the same common semantics. Thus, collective matrix factorization can be used to extract the latent semantic representation, which can be stated as follows,

$$\min_{\mathbf{U_t}, \mathbf{V}} \left(\sum_{t=1}^{m} \lambda_t \left\| \mathbf{X}^{(t)} - \mathbf{U}_t \mathbf{V} \right\|_F^2 + \gamma \left\| \mathbf{U}_t \right\|_F^2 \right) + \gamma \left\| \mathbf{V} \right\|_F^2, \tag{2}$$

where $\mathbf{V} \in R^{r \times n}$ is the common latent semantic representation of different modalities, $\mathbf{U}_t \in R^{d_t \times r}$ is the basis matrix, λ_t and γ are the balance parameters, where $\sum_{t=1}^{m} \lambda_t = 1$.

Hypergraph Learning. To effectively preserve the intra-modality similarity, we introduce a hypergraph to maintain the high-order relationship among samples. Let $\mathbf{H}$ be an incidence matrix that indicates whether a vertex v is contained in a hyperedge e. $\mathbf{H}(v,e) = 1$ if $v \in e$, otherwise $\mathbf{H}(v,e) = 0$. Thus, the vertex degree of v can be computed as $d(v) = \sum_{e \in E} w(e)\mathbf{H}(v,e)$, where $w(e)$ is the weights corresponding to hyperedge e. The edge degree of a hyperedge e is defined as $\delta(e) = \sum_{e \in E} \mathbf{H}(v,e)$. $\mathbf{D}_v$ is a diagonal matrix where the diagonal element is the degree of each vertex. Similarly, $\mathbf{D}_e$ and $\mathbf{W}_e$ are also diagonal matrices corresponding to the hyperedge degrees and the edge weights, respectively. Then, we have the un-normalized hypergraph Laplacian matrix $\mathbf{L} = \mathbf{D}_v - \mathbf{S}$, where $\mathbf{S} = \mathbf{H}\mathbf{W}_e\mathbf{D}_e^{-1}\mathbf{H}^T$. More information please refer to [19].

Since hypergraph can capture the high-order relationship among data samples, we impose a hypergraph regularization term to constrain the learning of latent common semantic representation.The hypergraph regularization term for each modality can be formulated as follows,

$$Tr(\mathbf{V}\mathbf{L}^{(t)}\mathbf{V}^T), \tag{3}$$

where $\mathbf{L}^{(t)}$ is the hypergraph Laplacian matrix of computed based on the t-th modality data.

Discrete Hash Codes Learning. After obtaining the latent semantic representation, we then need to learn the binary codes. Different from most previous methods taking the signs of $\mathbf{V}$ directly, we introduce an orthogonal rotation matrix $\mathbf{R} \in R^{r\times r}$ which can ensure that the bits in $\mathbf{B}$ are orthogonal to each other. This can also contribute to avoid the large quantization error generated by the relaxation scheme. We formulate the binary codes learning based on orthogonal rotation matrix $\mathbf{R}$ as follows,

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{R}} \alpha \left\|\mathbf{B}-\mathbf{R}\mathbf{V}\right\|_F^2, \\ &s.t.\ \mathbf{B}\in\{-1,1\}^{r\times n}, \mathbf{R}\mathbf{R}^T=\mathbf{I}. \end{aligned} \tag{4}$$

Label Preserving. In addition, we also make full use of the supervised information by inversely regressing it to the binary codes to enhance the discriminative property, which can be stated as follows,

$$\min_{\mathbf{B},\mathbf{G}} \beta \left\|\mathbf{B}-\mathbf{G}\mathbf{Y}\right\|_F^2 + \gamma \left\|\mathbf{G}\right\|_F^2, \tag{5}$$

where $\mathbf{G}\in R^{r\times c}$ is the projection matrix.

Overall Objective Function. Combining Eqs. (2), (3), (4), and (5), we obtain the overall objective function of HDMFH as,

$$\begin{aligned} &\min_{\mathbf{U_t},\mathbf{V},\mathbf{G}} \sum_{t=1}^{m}\lambda_t \left\|\mathbf{X}^{(t)}-\mathbf{U}_t\mathbf{V}\right\|_F^2 + \mu\sum_{t=1}^{m} Tr(\mathbf{V}\mathbf{L}^{(t)}\mathbf{V}^T) \\ &+\alpha\left\|\mathbf{B}-\mathbf{R}\mathbf{V}\right\|_F^2 + \beta\left\|\mathbf{B}-\mathbf{G}\mathbf{Y}\right\|_F^2 + \gamma R(\mathbf{U}_t,\mathbf{V},\mathbf{G}) \\ &s.t.\ \mathbf{B}\in\{-1,1\}^{r\times n}, \sum_{t=1}^{m}\lambda_t=1, \mathbf{R}\mathbf{R}^T=\mathbf{I}, \end{aligned} \tag{6}$$

where $\lambda_t, \mu, \alpha, \beta$ and γ are the tradeoff parameters, and $R=(\mathbf{U}_t,\mathbf{V},\mathbf{G})$ is a regularization term to avoid overfitting. The objective can be solved using an alternative optimization method.

2.3. Hashing Functions Learning

As mentioned previously, HDMFH is a two-step hashing method. After getting the unified binary codes $\mathbf{B}$, we need to learn a hashing function for each modality. Here, we adopt the linear regression as hashing function to transform the original features into compact binary codes. Specifically, we obtain the hashing function for the *t*-th modality, i.e., $\mathbf{W}_t$ by optimizing the following problem,

$$\min_{\mathbf{W_t}} \left\|\mathbf{B}-\mathbf{W}_t\mathbf{X}^{(t)}\right\|_F^2 + \theta\left\|\mathbf{W}_t\right\|_F^2, \tag{7}$$

where $\left\|\mathbf{W}_t\right\|_F^2$ is a regularization term, and θ is a balance parameter. We can easily obtain the solution,

$$\mathbf{W}_t = \mathbf{B}\mathbf{X}^{(t)^T}(\mathbf{X}^{(t)}\mathbf{X}^{(t)^T}+\theta\mathbf{I})^{-1}. \tag{8}$$

Then, for the *t*-th modality query data $\mathbf{X}^{(t)}_{query}$, the hash codes can be generated as follows,

$$\mathbf{B}_{query} = sgn(\mathbf{W}_t\mathbf{X}^{(t)}_{query}). \tag{9}$$

3. EXPERIMENTS

In this section, to validate the effectiveness of the proposed HDMFH, the details of the experiments are presented. We firstly conducte the experiments on three widely used cross-modal datasets consisting of images and text, i.e., Pascal VOC [20], MIRFlickr [21], and NUS-WIDE [22]. Then, we discusse the experiments of the proposed HDMFH on PKU XMedia [23, 24] dataset which containing multiple modalities. The settings of datasets are presented in Table 1.

Table 1. The Settings of Datasets.

Datasets	#Instances	#Training	#Testing	#Modality
Pascal VOC	9963	2808	2481	2
MIR Flickr	25000	5000	836	2
NUS-WIDE	269648	5000	1000	2
PKU XMedia	12000	800	200	3

3.1. Baselines and Implementation Details

To evaluate the effectiveness of our model, we compare the proposed HDMFH with eight state-of-the-art cross-modal hashing methods, i.e., CVH [14], CMFH [10], SMFH [11], LSSH [16], FSH [7], IISPH [8], CMDH [9], and SCRATCH [17]. In the experiments, all the parameters in these competitors are carefully set based on the original papers. We adopt a widely used metric for cross-modal retrieval to evaluate our method, namely mean average precision (MAP).

We firstly evaluate the performance of all methods on two common cross-modal retrieval tasks: 1) Text-to-Image and 2) Image-to-Text. Then, we conducte six cross-modal retrieval tasks or called multimodal retrieval on the multimodal dataset: 1) Text-to-Image. 2) Image-to-Text. 3) Text-to-Audio. 4) Audio-to-Text. 5) Image-to-Audio. 6) Audio-to-Image.

For HDMFH, in the two modalities retrieval tasks, its parameters are set to $\lambda_1=0.3$, $\lambda_2=0.7$, $\mu=1000$, $\alpha=0.1$, $\beta=0.1$, $\gamma=100$, and $\theta=5$. In the multiple modalities retrieval tasks, its parameters are changed to $\lambda_1=0.3$, $\lambda_2=0.4$, $\lambda_3=0.3$, $\mu=1000$, $\alpha=0.1$, $\beta=1$, $\gamma=0.01$, and $\theta=10^{-4}$. All parameters are selected by a validation procedure.

3.2. Experiment Results

We present the results from two aspects. The first one is about the traditional cross-modal retrieval with two modalities. The second one is multimodal retrieval, i.e., cross-modal retrieval with more than two modalities.

3.2.1. Results on two modalities

The MAP results of HDMFH and all baselines on Pascal VOC, MIR Flickr, and NUS-WIDE datasets are reported in

Table 2. The MAP Results of all methods on Pascal VOC, MIR Flickr and NUS-WIDE datasets with various code lengths.

Task	Methods	Pascal VOC				MIR Flickr				NUS-WIDE			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
Text-to-Image	CVH	0.1253	0.1298	0.1363	0.1482	0.5786	0.5736	0.5699	0.5674	0.3718	0.3644	0.3587	0.3561
	CMFH	0.3685	0.4663	0.4558	0.4304	0.5919	0.5929	0.5943	0.5957	0.3803	0.3842	0.3903	0.3971
	SMFH	0.3943	0.5493	0.6856	0.6618	0.5981	0.6055	0.6117	0.6208	0.3780	0.3866	0.3887	0.3947
	LSSH	0.4554	0.5403	0.6013	0.6182	0.5922	0.5893	0.5910	0.5934	0.4116	0.4215	0.4200	0.4253
	FSH	0.1723	0.2161	0.3229	0.3901	0.5781	0.5825	0.5861	0.5904	0.3692	0.3829	0.3797	0.3893
	IISPH	0.2971	0.4352	0.5004	0.5915	0.5966	0.5983	0.5952	0.5919	0.3871	0.4002	0.4081	0.4094
	CMDH	0.8498	0.8938	0.9063	0.9067	0.6308	0.6366	0.6387	0.6934	0.6571	0.7014	0.7261	0.7317
	SCRATCH	0.6301	0.8235	0.8993	0.9151	0.6904	0.7253	0.7165	0.7660	0.6387	0.5971	0.6786	0.7342
	HDMFH	**0.8743**	**0.9069**	**0.9237**	**0.9283**	**0.7302**	**0.7796**	**0.7818**	**0.7827**	**0.6867**	**0.7247**	**0.7556**	**0.7643**
Image-to-Text	CVH	0.1253	0.1253	0.1256	0.1231	0.5785	0.5739	0.5699	0.5673	0.3767	0.3670	0.3602	0.3555
	CMFH	0.1717	0.1826	0.1646	0.1639	0.5837	0.5815	0.5862	0.5859	0.3814	0.3826	0.3885	0.3938
	SMFH	0.1808	0.2129	0.2404	0.2395	0.5854	0.5916	0.5945	0.5972	0.3768	0.3889	0.3849	0.3909
	LSSH	0.2325	0.2498	0.2659	0.2792	0.5761	0.5768	0.5783	0.5799	0.3926	0.3932	0.3952	0.3969
	FSH	0.1549	0.1658	0.2337	0.2541	0.5835	0.5905	0.5901	0.6004	0.3691	0.3859	0.3801	0.3894
	IISPH	0.1724	0.2046	0.1989	0.2066	0.5853	0.5852	0.5832	0.5801	0.3807	0.3913	0.3976	0.3982
	CMDH	0.2111	0.2842	0.307	0.3276	0.5821	0.5859	0.6043	0.6379	0.5459	0.6006	0.6039	0.6284
	SCRATCH	0.2575	0.309	0.3625	0.3942	0.6533	0.6787	0.6674	0.7046	0.5529	0.5197	0.5811	0.6214
	HDMFH	**0.3502**	**0.3847**	**0.4173**	**0.4380**	**0.6714**	**0.7012**	**0.7098**	**0.7080**	**0.5691**	**0.6017**	**0.6167**	**0.6287**

Table 3. The MAP Results of FSH, SCRATCH and HDMFH on PKU XMedia.

Task	Methods	PKU XMedia			
		8 bits	16 bits	32 bits	64 bits
Text-to-Image	FSH	0.0886	0.0951	0.1047	0.1120
	SCRATCH	0.0797	0.0894	0.0940	0.0938
	HDMFH	**0.1152**	**0.1260**	**0.1551**	**0.1705**
Image-to-Text	FSH	0.0820	0.0871	0.0932	0.0978
	SCRATCH	0.0717	0.0822	0.0811	0.0716
	HDMFH	**0.0844**	**0.0983**	**0.1177**	**0.1320**
Text-to-Audio	FSH	0.0844	0.0898	0.1015	0.1084
	SCRATCH	0.0878	0.0971	0.1031	0.1064
	HDMFH	**0.1018**	**0.1151**	**0.1267**	**0.1371**
Audio-to-Text	FSH	0.0847	0.0867	0.0938	0.1002
	SCRATCH	0.0797	0.0908	0.0815	0.0763
	HDMFH	**0.0905**	**0.0985**	**0.1044**	**0.1154**
Image-to-Audio	FSH	0.0885	0.0924	0.1078	0.1116
	SCRATCH	0.0780	0.0885	0.0867	0.0969
	HDMFH	**0.1063**	**0.1267**	**0.1515**	**0.1594**
Audio-to-Image	FSH	0.0971	0.1017	0.1092	0.1211
	SCRATCH	0.0777	0.0836	0.0889	0.0856
	HDMFH	**0.1342**	**0.1429**	**0.1777**	**0.1992**

Table 2. It presents the performance of Text-to-Image task and Image-to-Text task when hash code length is 8 bits, 16 bits, 32 bits, and 64 bits, respectively. From this table, we can conclude the following observations: 1) HDMFH obtains the best results on both tasks with various code lengths and significantly outperforms the baselines in some cases, which can demonstrate the effectiveness of our method. The superiority of HDMFH can be mainly attributed to its capability of modeling complex data, which can better preserve the high-order relationship among data samples. 2) Most supervised methods, e.g., CMDH and SCRATCH, are superior to the unsupervised ones, such as CMFH, LSSH, and FSH demonstrating the advantage of utilizing the semantic information. 3) HDMFH performs much better than the baselines when the length of hash codes is small. This depicts that HDMFH can capture the high-order relationship better with short hash codes, which is significant in a search task.

3.2.2. Results on multiple modalities

Table 3 demonstrates the MAP results of our proposed HDMFH and the compared methods on the multimodal retrieval tasks on PKU XMedia dataset. From this table, we can observe that the proposed HDMFH achieves the best MAP results on all the six tasks, which further demonstrates the advantage of our proposed approach on multimodal retrieval. This is mainly because our HDMFH models the intra-modality semantic similarity by the hypergraph learning, which maintains the high-order relationship. In addition, the accuracy of the retrieval can be further improved by inversely regressing label to the unified binary codes.

4. CONCLUSION

In this paper, we have proposed a novel cross-modal hashing method HDMFH for multimodal retrieval. Combined with the hypergraph learning, our method can capture the high-order relationship among samples in each modality. Our HDMFH can handle the cross-modal retrieval with more than two modalities. Extensive experiments on three cross-modal benchmark datasets and one public multimodal dataset demonstrated that HDMFH outperforms the state-of-the-art cross-modal hashing methods.We plan to extend our HDMFH into a deep framework in the future.

5. REFERENCES

[1] Xingbo Liu, Xiushan Nie, Haoliang Sun, Chaoran Cui, and Yilong Yin, “Modality-specific structure preserving

hashing for cross-modal retrieval," in *ICASSP*. IEEE, 2018, pp. 1678–1682.

[2] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *ICASSP*. IEEE, 2019, pp. 4095–4099.

[3] Fangming Zhong, Zhikui Chen, and Geyong Min, "Deep discrete cross-modal hashing for cross-media retrieval," *Pattern Recognition*, vol. 83, pp. 64–77, 2018.

[4] Wenming Cao, Qiubin Lin, Zhihai He, and Zhiquan He, "Hybrid representation learning for cross-modal retrieval," *Neurocomputing*, vol. 345, pp. 45–57, 2019.

[5] Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua, "Harvesting visual concepts for image search with complex queries," in *ACM MM*. ACM, 2012, pp. 59–68.

[6] Zufan Zhang, Yang Zou, and Chenquan Gan, "Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression," *Neurocomputing*, vol. 275, pp. 1407–1415, 2018.

[7] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *CVPR*, 2017, pp. 7380–7388.

[8] Zhikui Chen, Fangming Zhong, Geyong Min, Yonglin Leng, and Yiming Ying, "Supervised intra-and inter-modality similarity preserving hashing for cross-modal retrieval," *IEEE Access*, vol. 6, pp. 27796–27808, 2018.

[9] Venice Erin Liong, Jiwen Lu, and Yap-Peng Tan, "Cross-modal discrete hashing," *Pattern Recognition*, vol. 79, pp. 114–129, 2018.

[10] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *CVPR*, 2014, pp. 2075–2082.

[11] Jun Tang, Ke Wang, and Ling Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *TIP*, vol. 25, no. 7, pp. 3157–3166, 2016.

[12] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *ACM SIGMOD*. ACM, 2013, pp. 785–796.

[13] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval.," in *IJCAI*, 2018, pp. 2854–2860.

[14] Shaishav Kumar and Raghavendra Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011.

[15] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang, "Semantics-preserving hashing for cross-view retrieval," in *CVPR*, 2015, pp. 3864–3872.

[16] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *ACM SIGIR*. ACM, 2014, pp. 415–424.

[17] Chuan-Xiang Li, Zhen-Duo Chen, Peng-Fei Zhang, Xin Luo, Liqiang Nie, Wei Zhang, and Xin-Shun Xu, "Scratch: A scalable discrete matrix factorization hashing for cross-modal retrieval," in *ACM MM*. ACM, 2018, pp. 1–9.

[18] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[19] Bin Qian, Xiaobo Shen, Zhenqiu Shu, Xiguang Gu, Jin Huang, and Jiabin Hu, "Hyper-graph regularized multi-view matrix factorization for vehicle identification," in *International Conference on Cloud Computing and Security*. Springer, 2018, pp. 543–554.

[20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[21] Mark J Huiskes and Michael S Lew, "The mir flickr retrieval evaluation," in *ACM MM*. ACM, 2008, pp. 39–43.

[22] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *CIVR*. ACM, 2009, p. 48.

[23] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 583–596, 2015.

[24] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *TCSV*, vol. 24, no. 6, pp. 965–978, 2013.