Discrete Cross-Modal Hashing for Efficient Multimedia Retrieval

Dekui Ma[a], Jian Liang[b,c†], Xiangwei Kong[a], Ran He[b,c,d] and Ying Li[a]
[a] *School of Information and Communication Engineering, Dalian University of Technology, China*
[b] *Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences (CAS), China*
[c] *University of Chinese Academy of Sciences, Beijing, China*
[d] *CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China*
Email: {madk, liying08}@mail.dlut.edu.cn, {jian.liang, rhe}@nlpr.ia.ac.cn, kongxw@dlut.edu.cn

***Abstract*—Hashing techniques have been widely adopted for cross-modal retrieval due to its low storage cost and fast query speed. Most existing cross-modal hashing methods aim to map heterogeneous data into the common low-dimensional hamming space and then threshold to obtain binary codes by relaxing the discrete constraint. However, this independent relaxation step also brings quantization errors, resulting in poor retrieval performances. Other cross-modal hashing methods try to directly optimize the challenging objective function with discrete binary constraints. Inspired by [1], we propose a novel supervised cross-modal hashing method called Discrete Cross-Modal Hashing (DCMH) to learn the discrete binary codes without relaxing them. DCMH is formulated through reconstructing the semantic similarity matrix and learning binary codes as ideal features for classification. Furthermore, DCMH alternately updates binary codes of each modality, and iteratively learns the discrete hashing codes bit by bit efficiently, which is quite promising for large-scale datasets. Extensive empirical results on three real-world datasets show that DCMH outperforms the baseline approaches significantly.**

***Keywords*-cross-modal hashing; cross-media retrieval; discrete binary codes;**

I. Introduction

Hashing is an effective technique for approximate nearest neighbor(ANN) search. Enjoying the low storage cost, hashing-based retrieval has drawn considerable attention in large data collections. Hence, numerous hashing methods were proposed in the last few years, most of which based on single-modal data. Since heterogeneity has been an increasingly important characteristic, numerous cross-modal retrieval methods [2–5] have been proposed.

Most previous work focus on the way of designing hashing functions that can preserve the similarities of data. Considering the cross-modal hashing methods, the key step is to learn hashing functions that map different modality features into a common binary space, while the similarities of both inter-modal and intra-modal are preserved simultaneously. Roughly speaking, existing cross-modal methods can be divided into two categories: supervised and unsupervised methods.

† The corresponding author (Email: jian.liang@nlpr.ia.ac.cn); the first two authors contributed equally and should be considered co-first authors.

One famous unsupervised method is [2], which extended [6] to the multimodal setting through minimizing the weighted distance. [7] utilized collective matrix factorization from different modalities of one instance to obtain the hashing functions with latent factor model. Besides, [8] captured the salient structures of images and learned latent concepts from texts through sparse coding and matrix factorization respectively. Supervised methods usually achieve better results because they make full use of provided semantic labels to learn discriminative hashing functions via some other criterion like label-similarity preserving. [9] was proposed to embed data from different feature space into a common metric space. Inter-media (IMH) [10] considered the differences between the modalities through exploring the single modality correlations and keeping the different modalities codes consistent. Semantic Correlation Maximization (SCM) [11] was also proposed to maximize the semantic correlation and learn the hashing functions greedily. [12] utilized neural network for cross-media hashing while [13] exploited matrix factorization for multi-view data. [14] proposed a simple two-step approach and obtained impressive retrieval performances on various benchmark datasets. They regard binary code obtaining via unimodal hashing methods as unified code.

Similar to unimodal hashing methods, cross-modal approaches have inevitable binary constraints, which make the discrete optimization process challenging. To make it feasible, most approaches adopt a two-step procedure: first learn real hashing functions to relax the constraints and then threshold it. For example, one can utilize Canonical Correlation Analysis (CCA) to map multimedia data into a common low-dimensional subspace, and then do threshold to obtain binary codes. However, this trick brings non-negligible quantization errors, thus it is suboptimal. More and more researches that aim to minimize the quantization errors have been proposed, among which, ITQ [15] is a classic iterative quantization method. By seeking a rotation, ITQ makes the learned code approach binary. [16] seeks to reconstruct the data from the binary code. And it learns the encoder and decoder separately with the help of auxiliary coordinates methods. By introducing an auxiliary variable

[1] reformulated the objective function and obtained discrete solution via cyclic coordinate descent. [17] is one of the pioneers that focus on binary quantization errors for cross-modal hashing. It seeks binary quantizers for each modality alternately through solving the problem of binary quantization and subspace learning simultaneously.

Inspired by the approaches above, we extend unimodal hashing method [1] to develop a discrete hashing method for cross-modal retrieval named *Discrete Cross-Modal Hashing* (DCMH). DCMH employs an iterative optimization method to learn hashing functions without relaxing the discrete constraints. We formulate the objective function through reconstructing the semantic inter-similarity matrix, and regarding the learned binary codes as an ideal features for intra-modal classification. To simplify the optimization, DCMH adopts linear regression to form both hashing functions and classification matrix. Regarding the NP-hard binary optimization problem, we use the *discrete cyclic coordinate descent* method proposed in [1]. Generally, the overall objective function mainly consists of two intra-modal hashing functions and one inter-similarity reconstructing term, and the intra-modal hashing function primarily relies on binary features classification error criterion. In details, DCMH alternately updates binary codes for each modality, which is efficiently solved. Here we summarize the main contributions as follows:

- By simultaneously preserving the similarity of inter- and intra- modality, we make full use of provided semantic labels (i.e., one in inter-modality and one in intra-modality).
- We optimize the formulation in an efficient discrete method, which minimizes the quantization efficiently.
- Extensive experiments on three datasets demonstrate that the proposed DCMH can significantly outperform the existing cross-modal hashing approaches.

II. PROPOSED METHOD

In this section, we explain the proposed method and show the associated optimization process in details.

A. Problem Definition

For simplicity, we assume that there are only two modalities, and it can be easily extended to more modalities.

Assume $X = \{x_i\}_{i=1}^n, x_i = \{x_i^1, x_i^2\}$ is the n data points of two different modalities, where $x_i^1 \in R^m$ is a m-dimensional image feature, and $x_i^2 \in R^d$ is a d-dimensional text feature. Given the code length k, our goal is to learn a pair of hashing function matrixes W_1, W_2 that map the original feature to binary code $h_i \in \{-1, 1\}^k$ for x_i. Such learned hashing codes can well preserve their semantic similarities. We denote a matrix $Y \in \{0, 1\}^{c \times n}$ to store the label, $y_i \in R^c$ denotes the i-th label vector, where c is the class number of the dataset.

B. Inter-Modality Similarity Preservation

Since we focus on cross-modal retrieval, the learned hashing codes should preserve the semantic similarity across different modalities. More specifically, we reconstruct the similarity affinity matrix S by the learned cross-modal hashing codes H_1 and H_2. Here, S is directly generated from Y, and $s_{i,j} = 1$ indicates that i-th and j-th objects belong to the same class, and otherwise $s_{i,j} = -1$. Hence, the basic object function about binary code is

$$\min \left\| H_1^T H_2 - cS \right\|_F^2, \tag{1}$$

where $H_1, H_2 \in \{-1, +1\}^{k \times n}$ are the learned hashing codes matrix, and c is a constant equaling to binary code length. Here, H_1 and H_2 are mapped from original features through hashing functions $f_1(x)$ and $f_2(x)$. There are many different kinds of functions to define $f(x)$, we adopt the most simple one, which is defined as $f(x) = sgn(P^T x)$ where $sgn(\cdot)$ is the sign function and $P_1 \in R^{m \times k}, P_2 \in R^{d \times k}$ are the projection matrices.

C. Intra-Modality Similarity Preservation

Similarity preservation indicates that similar objects should be mapped to similar codes in the Hamming space. In addition to the similarity preservation between modalities, we also aim to preserve the similarity within single-modality, which is the main problem of unimodal hashing method.

To leverage semantic labels in this step, we optimize the learned binary codes as a classification task. In another word, we consider the learned hashing codes can be well classified and the semantic label is ground truth.

Given the binary codes h_i, we consider only one modality and the objection function of classification can be written as:

$$\begin{aligned} &\min_{W,H} \sum_{i=1}^{n} L(y_i, W^T h_i) + \lambda \|W\|_F^2 \\ &s.t.\ h_i = f(x_i) = sgn(P^T x_i), \end{aligned} \tag{2}$$

where $L(\cdot)$ is the loss function of classification, λ is the regularization parameter, and y_i is the ground truth of i-th object. We can select any appropriate loss function for Eq.2. Here the l_2 loss function is chosen due to its simplicity. By introducing matrix expression, the problem can be rewritten as:

$$\begin{aligned} &\min_{H,W,P} \left\| Y - W^T H \right\|^2 + \eta \left\| H - P^T X \right\|^2 + \lambda(\|W\|^2 + \|P\|^2) \\ &\quad s.t.\ H \in \{-1, 1\}^{k \times n}. \end{aligned} \tag{3}$$

D. Overall Formulation and Optimization

Combining the inter- and intra- modality similarity preservation terms in Eq.1 and Eq.3 together, the overall objective function of DCMH is:

$$\begin{aligned}\min_{\mathbb{H},\mathbb{W},\mathbb{P}} G = &\sum_{i=1,2} \left\|Y - {W_i}^T H_i\right\|^2 + \eta\left\|H_i - {P_i}^T X_i\right\|^2 \\ &+ \lambda R(W_i, P_i) + \gamma\left\|H_1^T H_2 - cS\right\|^2 \\ &s.t.\ \ H_i \in \{-1,1\}^{k\times n}.\end{aligned} \tag{4}$$

Here η, λ and γ are tradeoff parameters, and we define $R(\cdot) = \|\cdot\|_F^2$ as regularization term to avoid overfitting.

Besides, nonlinear embedding beforehand can boost the performances of linear methods and it is scalable for high-dimensional data matrices. Hence, we adopt a simple yet effective non-linear technique [18, 19] as follows:

$$F(x) = sgn(P^T\phi(x)), \tag{5}$$

where

$$\phi(x) = [\exp\left(\|x - z_1\|^2/\sigma\right), \cdots, \exp\left(\|x - z_l\|^2/\sigma\right)]. \tag{6}$$

Here $\{z_j\}_{j=1}^l$ are the randomly selected l landmark points, σ is the kernel width. Then X in Eq. 4 is replaced by $\phi(X)$.

P-Step Fix H and W, let $\frac{\partial G}{\partial P_i} = 0$, then we can obtain:

$$P_i = \left(\phi(X_i)\phi(X_i)^T + \lambda I\right)^{-1}\phi(X_i)H_i^T, \tag{7}$$

where I is a diagonal matrix. This step is about least-square linear regression.

W-Step Fix H and P, let $\frac{\partial G}{\partial W_i} = 0$, then obtain:

$$W_i = \left(H_i H_i^T + \lambda I\right)^{-1} H_i Y^T \tag{8}$$

The same as Eq.7, we can get a closed-form solution by regression.

H-Step Fix W and P, Eq.4 can be rewritten as:

$$\begin{aligned}G(H_i) = &\left\|Y - {W_i}^T H_i\right\|^2 + \eta\left\|H_i - {P_i}^T\phi(X_i)\right\|^2 \\ &+ \gamma\left\|H_1^T H_2 - cS\right\|^2 \\ &s.t.\ \ H_i \in \{-1,1\}^{k\times n}\end{aligned} \tag{9}$$

Due to the discrete constraints, solving H is a NP-hard problem. Most existing methods relaxed this constraint while some try to optimize it by introducing sigmoid function. In this paper we attempt to learn the binary hashing codes taking along with the discrete constraints. One naive approach is enumeration which is uncomputable. However, we can solve it through parallel processing. To illustrate it, we expand Eq.9 as,

$$\begin{aligned}G(H_1) = &\|W_1 H_1\|^2 - 2Tr(H_1^T W_1 Y) + cons \\ &+ \eta(cons - 2Tr(H_1^T P_1^T\phi(X_1))) \\ &+ \gamma(cons - 2cTr(H_1^T H_2)) \\ &s.t.\ \ H_1 \in \{-1,1\}^{k\times n},\end{aligned} \tag{10}$$

Algorithm 1 Discrete Cross-Modal Hashing (DCMH)

Input: Data matrices $X^{(t)}$, $t = 1, 2$, semantic label matrix Y and hash code length k.
Output: Hash projection matrices P_i , $i = 1, 2$.
Procedure:
1. Randomly select l objects to get the nonlinear embedding data $\phi(X)$ via the RBF kernel function.
2. Initialize H as $\{-1, 1\}^{k\times n}$ randomly;
3.Repeat:
1. Obtain P_1 and P_2 via Eq.7;
2. Obtain W_1 and W_2 via Eq.8;
3. Iteratively solve B_1 and B_2 via Eq.11 with the help of DCC;

Until converge or reach maximum iterations.

it can be rewritten as:

$$\begin{aligned}\min_{H_1} \|W_1 H_1\|^2 - 2Tr(H_1^T(W_1 Y + \eta P_1^T\phi(X_1)) + \gamma c H_2)) \\ s.t.\ \ H_1 \in \{-1,1\}^{k\times n}.\end{aligned} \tag{11}$$

Here we take a measure that h^i is updated while the remains are fixed, where h^i denotes the i-th column of H. Actually, it is the *discrete cyclic coordinate descent (DCC)* method proposed in [1]. We adopt DCC to optimize Eq.11 with iterating $5 \sim 10$ times each column.

E. Computational Complexity Analysis

DCMH adopts a iterative optimization, **P-Step** and **W-Step** are classical linear regression solutions, which occupy $O\left(nl^2k\right)$ and $O\left(nd^2k\right)$. **H-Step** occupies $O\left(tk^2n + tk^2c\right)$ each iteration, where t is the number of the internal iteration. The overall computational complexity is $O\left(T(nk^2)\right)$, where T is the number of the external iterations. In testing phase, the complexity of generating hashing codes is constant with $O\left(mk\right)$ for an image query and $O\left(dk\right)$ for a text query. Generally, DCMH have a linear complexity to n and is flexible for large-scale datasets.

F. Convergence Analysis

To seek an optimal solution, **P**, **W** and **B** are alternately learned for several iterations. The objective function in Equation (4) is minimized in each step and we show the convergence analysis of DCMH as follows:

$$\begin{aligned}G\left(P^{(t)}, W^{(t)}, B^{(t)}\right) \geq G\left(P^{(t+1)}, W^{(t)}, B^{(t)}\right) \geq \\ G\left(P^{(t+1)}, W^{(t+1)}, B^{(t)}\right) \geq G\left(P^{(t+1)}, W^{(t+1)}, B^{(t+1)}\right)\end{aligned} \tag{12}$$

where $P^{(t)}, W^{(t)}, B^{(t)}$ are matrices in the t-th iteration. In summary, the whole procedure of the proposed method DCMH is shown in Algorithm 1.

III. EXPERIMENTS

We compare our DCMH with baseline methods on three different dataset: Wiki, Labelme and VOC. The experiment results illustrated that DCMH can significantly outperform the baseline methods.

A. Datasets and Setting

The **Wiki** [20] dataset consists of 2,866 (2,173 training and 693 test) text-image documents which were collected from 'Wikipedia' and labeled by one of 10 semantic categories. Each image is detailed with 128-dimensional SIFT feature vector, while the text is depicted with 10-dimensional LDA topic features. The **Wiki++** shares the same setting as the Wiki dataset, except for the 4,096-dimensional CNN features for images extracted by Caffe[1] and 5,000-dimensional feature vectors for texts extracted by using the Bag-of-Words representation with the TF-IDF weighting scheme.

The **LabelMe** [21] outdoor dataset consists of 2,686 fully annotated outdoor images from 8 scene categories. For the text modality, we generate the object account vector via the LabelMe[2] toolbox. We randomly split the dataset into training/testing set as 3:1. The image and text features are 512-dimensional Gist features and 470-dimensional word frequency features, respectively.

The **VOC+** [22] dataset includes 2,808 training and 2,841 testing data. The images are associated with only single label as the way in [23]. Here, we also extract the 4,096-dimensional CNN features instead of original 512-dimensional Gist features for image representation.

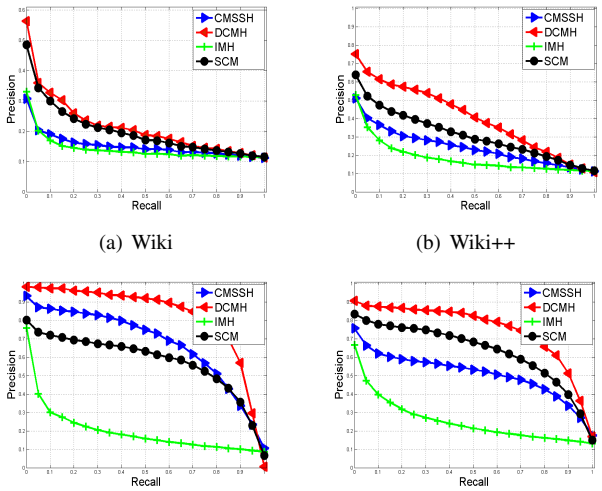
B. Experiment Setting

Baseline Methods: CMMSH[9], IMH[10] and SCM[11]. All the source codes are available publicly, and all the parameters set as consistent with their paper presented. Note that, we carry IMH as a supervised way by training all instances. All the results are averaged over 4 runs, to eliminate the influence of random initialization. All our experiments are run on a workstation with a 2.60GHZ Intel Xeon E5-2650 CPU and 32.0GB RAM.

Evaluation Scheme: We adopt the mean average precision (MAP) which is widely used for retrieval task to measure the performance of all methods. In this paper, we take a test-test In addition, we also plot precision-recall curves to further study the overall retrieval performance. For our DCMH, the parameter l is fixed as 500 for each dataset.

C. Experimental Results and Discussion

We compared DCMH with other baseline methods on the Wiki, Wiki++, VOC+ and LabelMe datasets. The MAP values are presented in Table I with the hashing bits in the range of $\{16, 24, 32, 64\}$. Obviously, DCMH significantly outperforms other methods in all datasets about both text query task and image query task. For example, the values of DCMH increase over 12% and 13% on average compered with SCM on LableMe for image query and text query, respectively. Compared with the second best method (e.g., SCM) on Wiki++ dataset, the maximum gains of DCMH reach 29.7% for image query and 25.9% for text query with 32bits.



(a) Wiki (b) Wiki++

(c) VOC+ (d) Labelme

Figure 1. Precision-recall curves with 32 bits code length for text query. (best viewed in colors)

Due to its dramatic performance, deep feature has been increasingly popular. Especially on VOC+, all of the approaches achieve perfect performances, DCMH attains nearly 100% MAP value with 64bits. Compared with Wiki, Wiki++ allow all the methods a better result. Note that our DCMH increases 50.9% for image query and 69.4% for text query with 24bits, while SCM increases 32.4% and 51.0% accordingly. This illustrates that DCMH preferably utilizes deep feature.

From Figure 1 and Figure 2, we can draw the conclusion that our proposed DCMH performs best for both text query and image query at all datasets. The advantage of DCMH can be distinctly seen via precision-recall curves. DCMH is close to SCM in Figure 1.(a) in Wiki dataset while the gap between them are widen in Wiki++ dataset.

datasets	**Wiki**	**Wiki++**	**VOC+**	**Labelme**
IMH	9.05	10.78	19.94	11.68
CMMSH	0.69	13.42	44.70	7.66
SCM	0.20	862.72	798.67	67.88
DCMH	3.59	4.42	5.19	2.26

Table II
TRAINING TIME (IN SECONDS) ON THREE DATASETS.

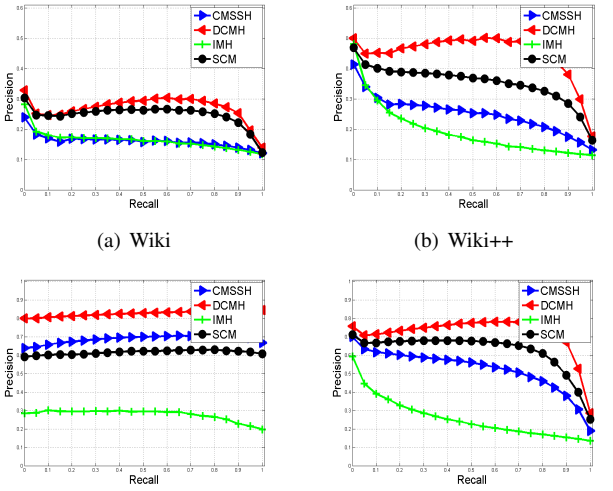
To evaluate the simplicity of time complexity, we al-

[1]http://caffe.berkeleyvision.org/.
[2]http://labelme.csail.mit.edu/Release3.0/browserTools/php/.

Image query	**Wiki**				**Wiki++**				**VOC+**				**Labelme**			
# of bits	16	24	32	64	16	24	32	64	16	24	32	64	16	24	32	64
CMMSH	28.29	22.56	20.20	22.40	36.25	33.79	34.21	29.35	82.58	86.15	86.79	85.68	56.63	59.06	62.03	60.53
IMH	23.99	23.55	23.33	21.43	33.19	33.13	32.49	30.88	64.03	62.99	61.29	58.70	46.14	43.01	40.41	35.57
SCM	34.28	35.24	34.57	36.23	42.26	46.66	46.66	48.59	83.68	88.91	90.42	91.74	67.10	68.56	70.48	72.53
DCMH	**36.81**	**38.71**	**41.49**	**43.44**	**53.39**	**58.43**	**60.52**	**61.16**	**90.89**	**97.13**	**98.94**	**99.11**	**76.00**	**78.36**	**78.88**	**79.66**

Text query	**Wiki**				**Wiki++**				**VOC+**				**Labelme**			
# of bits	16	24	32	64	16	24	32	64	16	24	32	64	16	24	32	64
CMMSH	24.03	25.72	21.36	23.63	41.67	34.59	36.21	31.98	83.64	85.20	86.48	86.37	56.85	61.64	60.71	60.35
IMH	24.36	22.91	21.62	20.40	33.40	33.87	32.99	31.37	54.95	49.89	43.79	34.98	48.64	44.81	42.09	35.90
SCM	31.37	32.24	32.41	33.67	45.75	48.67	47.86	51.95	74.48	76.55	75.61	75.36	74.56	75.11	76.79	80.28
DCMH	**37.88**	**34.24**	**36.22**	**36.72**	**55.48**	**58.01**	**60.27**	**61.25**	**87.58**	**93.24**	**95.83**	**96.43**	**85.57**	**87.31**	**86.10**	**88.03**

Table I
MAP@50 RESULTS ON THREE DATASETS FOR DIFFERENT TASKS. THE BEST VALUE IS SHOWN IN BOLDFACE.



(a) Wiki (b) Wiki++

(c) VOC+ (d) Labelme

Figure 2. Precision-recall curves with 32 bits code length for image query. (best viewed in colors)

so compare the training time with baselines in Table II. Generally, all the methods cost relatively less time at the low-dimensional datasets. SCM always achieves the second performance, while the training time significantly increased at high-dimensional data. By contrast, DCMH has a strong ability to adapt to high-dimensional data.

IV. CONCLUSIONS

In this paper, we proposed a supervised discrete approach named DCMH for cross-modal hashing, which focused on obtaining the binary codes with a discrete approach in cross-media search. To leverage the semantic labels, this method explored similarity preservation terms based on classification criterion, and introduced a inter-similarity reconstruction term. We further depict a efficient and effective solution for DCMH. Extensive experimental results illustrated the huge advantages of our DCMH over other existing cross-modal hashing methods.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (Grant No. 61502073,61473289), the Open Projects Program of National Laboratory of Pattern Recognition (No. 201407349), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02070000).

REFERENCES

[1] F. Shen, C. Shen, W. Liu, and H. Tao Shen, "Supervised discrete hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 37–45.

[2] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proceedings of the 22nd international joint conference on Artificial Intelligence*. AAAI, 2011, pp. 1360–1365.

[3] J. Liang, D. Cao, R. He, Z. Sun, and T. Tan, "Principal affinity based cross-modal retrieval," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 126–130.

[4] J. Liang, R. He, Z. Sun, and T. Tan, "Group-invariant cross-modal subspace learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. AAAI, 2016, pp. 1739–1745.

[5] J. Liang, Z. Li, D. Cao, R. He, and J. Wang, "Self-paced cross-modal subspace matching," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 569–578.

[6] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*. Curran Associates, Inc., 2009, pp. 1753–1760.

[7] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2075–2082.

[8] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. AAAI, 2014, pp. 415–424.

[9] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in

Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3594–3601.

[10] 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*. ACM, 2013, pp. 785–796.

[11] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization." in *Proceedings of the 25nd international joint conference on Artificial Intelligence*, vol. 1, no. 2, 2014, p. 7.

[12] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 901–904.

[13] X. Shen, F. Shen, Q.-S. Sun, and Y.-H. Yuan, "Multi-view latent hashing for efficient multimedia search," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 831–834.

[14] D. Ma, J. Liang, X. Kong, and R. He, "Frustratingly easy cross-modal hashing," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 237–241.

[15] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2013.

[16] M. A. Carreira-Perpinán and R. Raziperchikolaei, "Hashing with binary autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 557–566.

[17] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1886–1894.

[18] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. IEEE, 2011, pp. 1–8.

[19] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE transactions on cybernetics*, vol. 45, no. 8, pp. 1669–1680, 2015.

[20] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 251–260.

[21] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[22] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 6, pp. 1145–1158, 2012.

[23] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167.