Fast Semantic Preserving Hashing for Large-Scale Cross-Modal Retrieval

Xingzhi Wang[1,2,3], Xin Liu[1,2,*], Shujuan Peng[1], Yiu-ming Cheung[4], Zhikai Hu[1], Nannan Wang[2]
[1]Dept. of Comput. Sci. & Fujian Key Lab. of Big Data Intelligence and Security, Huaqiao University, Xiamen, China
[2]State Key Lab. of Integrated Services Networks & School of Telecommun. Eng., Xidian University, Xi'an, China
[3]Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, China
[4]Dept. of Comput. Sci. and Institute of Research and Continuing Education, HK Baptist University, Hong Kong SAR, China
Email: {xzwang, xliu, pshujuan, zkhu}@hqu.edu.cn, ymc@comp.hkbu.edu.hk, nnwang@xidian.edu.cn

***Abstract*—Most Cross-modal hashing methods do not sufficiently exploit the discrimination power of semantic information when learning hash codes, while often involving time-consuming training procedures for large-scale dataset. To tackle these issues, we first formulate the learning of similarity-preserving hash codes in terms of orthogonally rotating the semantic data to hamming space, and then propose a novel Fast Semantic Preserving Hashing (FSePH) approach to large-scale cross-modal retrieval. Specifically, FSePH introduces an orthonormal basis to regress the targeted hash codes of training examples to their corresponding reasonably relaxed class labels, featuring significantly reducing the quantization error. Meanwhile, an effective optimization algorithm is derived for modality-specific projection function learning and an efficient closed-form solution for hash code learning, which are computationally tractable. Extensive experiments have shown that the proposed FSePH approach runs sufficiently fast, and also significantly improves the retrieval performances over the state-of-the-arts.**

***Index Terms*—Cross-modal hashing, fast semantic preserving, orthonormal basis, bi-Lipschitz continuity**

I. Introduction

With the tremendous explosion of multimedia data, cross-modal retrieval has attracted increasing attention to approximate nearest neighbors search across different modalities, such as using image to search the relevant text documents or using text to search the relevant images [1]. Nevertheless, the heterogeneous property within multi-modal data has been widely considered as a great challenge to cross-modal retrieval. To address this issue, early studies [2, 3] learn a common latent subspace to minimize their heterogeneity. However, these subspace methods are computationally expensive to deal with the large-scale and high dimensional media data.

Hashing, favored for its low storage cost and fast retrieval speed, has received considerable attention for indexing of large-scale multimedia data [4]. In recent years, various kinds of cross-modal hashing methods have been proposed in unsupervised fashion [5, 6, 7, 8] and supervised fashion [9, 10, 11, 12, 13, 14]. Since the label information is helpful to construct the correlations across different modalities, the supervised methods can well mitigate the semantic gap between heterogeneous modalities for better performances. In recent years, deep cross-modal hashing approaches [15, 16, 17], integrating feature learning and hashing code learning, have yielded outstanding performance, but they always involve computational complexity for learning optimum parameters.

In spite of some supervised methods have achieved promising retrieval performance, several intrinsic issues have not been well tackled. Firstly, most supervised methods only consider the semantic-preserving property provided by labels, but which fail to explore the discrimination power of labels. Secondly, some popular supervised methods convert the original discrete optimization problem into the continuous one by relaxing binary constraint into real number field, which may accumulate large quantization error as the code length increases. Thirdly, recent supervised discrete hashing methods attempt to learn the hash code bit by bit, which involves large iterations in learning process. Note that Fast Supervised Discrete Hashing (FSDH) [18] can update the whole hash code by a close form solution, however, it is a unimodal hashing work. Therefore, it is still desirable to study a fast cross-modal hashing method, while achieving high performance.

In this paper, we present a Fast Semantic Preserving Hashing (FSePH) to facilitate efficient retrieval across different modalities, and the main contributions are three-fold: 1) The learning of similarity-preserving hash codes is newly formulated in terms of orthogonally rotating the semantic data, whereby the quantization loss of mapping such data to hamming space can be significantly minimized; 2) The label values are reasonably relaxed to reduce the quantization error and speed up the learning process; 3) An effective optimization algorithm is proposed for projection function learning and a very efficient closed-form solution for hash code learning. The experiments have shown its outstanding performance.

The remaining part is organized as follows: Section II briefly surveys the related works, and Section III elaborates FSePH in detail. The experimental results and discussions are provided in Section IV. Finally, we draw a conclusion in Section V.

The work was supported by National Science Foundation of China (Nos. 61673185, 61672444, 61876142 and 61922066), the National Science Foundation of Fujian Province (Nos. 2017J01112), Quanzhou City Science&Technology Program of China (No. 2018C107R), State Key Laboratory of Integrated Services Networks of Xidian University (No. ISN20-11), Promotion Program for graduate student in Scientific research and innovation ability of Huaqiao University (No. 17013083010), the Initiation Grant for Faculty Niche Research Areas (IG-FNRA) of Hong Kong Baptist University (HKBU) with Grant: RC-FNRA-IG/18-19/SCI/03, and Interdisciplinary Research Clusters Matching Scheme (IRCMs) of HKBU with Grant: RC-IRCMs/18-19/SCI/01. Xin Liu is the corresponding author.

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00172

II. Related Work

A. Latent Subspace Learning

Due to the heterogeneous structure of multimodal data, it is difficult to implement cross-modal retrieval in their original feature space. To tackle this problem, Canonical Correlation Analysis (CCA) [2] learns a common latent subspace to maximize the correlation between heterogeneous modalities and achieves cross-modal retrieval. Similarly, Partial Least Square (PLS) [3] also learns a common latent subspace for cross-modal retrieval. Remarkably, these unsupervised methods ignore the semantic labels for discriminative learning. Therefore, the supervised extension of CCA, like GMA [19], cluster-CCA [20] and ml-CCA [21] are developed to improve the retrieval performance. To capture the nonlinear correlation between heterogeneous modalities, some deep methods, *e.g.*, deep CCA [22], have also been proposed for cross-modal retrieval. Remarkably, these methods are generally unsuitable for processing large-scale and high-dimensional data.

B. Cross-modal Hashing

Cross-modal hashing has recently attracted much attention in recent years due to its low storage cost and fast query speed, which mainly fall into unsupervised and supervised cases.

Unsupervised hashing methods directly learn the projection functions to map the original feature spaces into hamming spaces. Accordingly, Inter-media Hashing (IMH) [5] obtains a common hamming space by preserving the inter-view and intra-view consistency, while Collective Matrix Factorization Hashing (CMFH) [6] jointly learns unified hash codes and hash functions by collective matrix factorization. Similarly, Latent Semantic Sparse Hashing (LSSH) [7] first utilizes sparse coding and matrix factorization to extract latent semantic features, and then mapping the latent semantic feature to a joint abstraction space. Although these methods have achieve promising performance, the available class label information remains unexplored and the learned hash codes are not discriminative enough for high retrieval performance.

Supervised cross-modal hashing methods can well mitigate the semantic gap between heterogeneous modalities, and could produce more effective hash codes to improve the retrieval performance. For instance, Supervised Matrix Factorization Hashing (SMFH) [11] utilizes the label supervision to produce unified hash codes, while maintaining the label consistency and local geometric consistency. In addition, Semantic Correlation Maximization (SCM) [9] seamlessly integrates the label information into the hash code learning procedure, while Semantic Preserved Hashing (SePH) [10] and Generalized Semantic Preserving Hashing (GSePH) [13] construct an affinity matrix by label supervision to approximate hash codes. Similarly, Discrete Cross-modal Hashing (DCH) [14] directly updates hash codes bit by bit while retaining the discrete constraints for more compact hash codes. Although these supervised methods are able to achieve efficient cross-modal retrieval, they do not fully exploit the discrimination power of semantic information when learning hash codes, and often involve huge iterations in training procedures. With label embedding, recent deep cross-modal hashing works [15, 16] jointly learns the high-level features and hash code in an integrated way, but which are unsuitable for processing large-scale multi-modal database.

III. Fast Semantic Preserving Hashing

For ease of presentation, this section describes the proposed FSePH with only two modalities (*i.e.*, image and text), which can be easily extended to three or more modalities.

A. Problem Formulation

Let n be the number of training image-text pairs, denoted as $\mathbf{X}^{(t)} = \{\mathbf{x}_i^{(t)}\}_{i=1}^n, t = 1, 2$, where $\mathbf{x}_i^{(t)}$ is the i-th sample in t-th modality. In practice, the original features are assumed to be zero-centered. *i.e.*, $\sum_{i=1}^n \mathbf{x}_i^{(t)}$=0, and the provided training labels are $\mathbf{L}\in\{0,1\}^{c\times n}$, where c is the number of semantic categories, $\mathbf{L}_{i,j}$=1 indicates that the j-th sample fall into the i-th class (in general each sample belongs to no less than one class), otherwise, $\mathbf{L}_{i,j}$=0. The goal of cross-modal hashing is to learn binary codes matrix $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^n$ for training instances, and modality-specific projection matrices $\{\mathbf{P}_1, \mathbf{P}_2\}$ for respectively linking the original image&text feature spaces and the common hamming space, where $\mathbf{h}_i \in \{-1, 1\}^q$ is q bits hash code of the i-th sample.

B. The Proposed FSePH Methodology

For cross-modal hashing, it is necessary to produce an efficient code in which the bits are pairwise uncorrelated. As pointed in [23], the learning of similarity-preserving binary codes can be successfully formulated in terms of orthogonally rotating zero-centered PCA-projected data, so as to minimize the quantization error of mapping that data to the vertices of a zero-centered binary hypercube. Geometrically, it is not difficult to find that hamming space is consistent with the vertices of unit hypercube. Accordingly, we introduce an orthonormal basis $\mathbf{C}=\{\xi_i\}_{i=1}^c$ to seamlessly joint the semantic preserving and quantization error reduction. More specifically, we propose to orthogonally rotate $\mathbf{L}$ to reduce the quantization error and ensure the semantic subspace $\mathbf{CL}$ to be as close as possible to the vertices of unit hypercube, whereby the optimal orthogonal basis $\mathbf{C}$ can be obtained by minimizing the following quantization error:

$$\min_{\mathbf{C}} \sum_{i=1}^{n} \|\mathrm{sgn}(\mathbf{CL}_i) - \mathbf{CL}_i\|_2^2 \;\; s.t.\; \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_c \qquad (1)$$

where $\mathbf{I}_c$ is c-order identity matrix. Fig. 1 is a geometric example for semantic subspace illustration, where $\vec{s_1}$ and $\vec{s_2}$ are semantic vectors formed by an orthonormal orthogonal basis $\mathbf{C} = \{\xi_1, \xi_2, \xi_3\}$ and label vectors $\mathbf{L}_1 = (1, 1, 0)^{\mathrm{T}}$, $\mathbf{L}_2 = (1, 0, 1)^{\mathrm{T}}$. The minimum quantization error is attainted by computing the sum of the length of $\vec{e_1}, \vec{e_2}$. Accordingly, the learning of semantic-preserving binary codes can be formulated in terms of orthogonally rotating the semantic data to minimize the quantization loss, and we can rewrite Eq. (1) as:

$$\min_{\mathbf{H},\mathbf{C}} \|\mathbf{H} - \mathbf{CL}\|_F^2, \; s.t.\; \mathbf{H}\in\{-1,1\}^{q\times n}, \; \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_c \qquad (2)$$

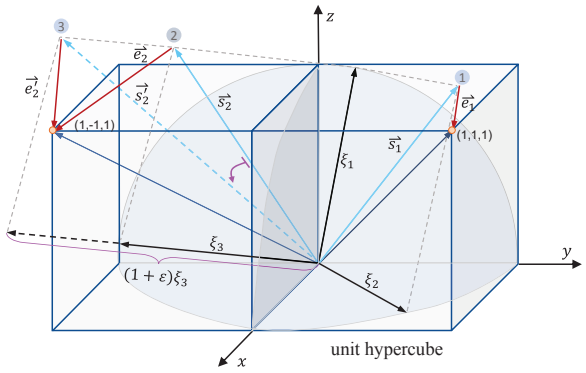


Fig. 1: A geometric example of the semantic subspace.

According to the orthogonal constraint, Eq. (2) can be equivalently transformed into following form.

$$\min_{\mathbf{H},\mathbf{C}} \left\|\mathbf{L}-\mathbf{C}^{\mathrm{T}}\mathbf{H}\right\|_F^2,\ s.t.\ \mathbf{H}\in\{-1,1\}^{q\times n},\ \mathbf{C}^{\mathrm{T}}\mathbf{C}=\mathbf{I}_c \quad (3)$$

The above formulation is a typical regression problem, which regress $\mathbf{H}$ to $\mathbf{L}$. However, it is difficult to regress $\mathbf{L}$ accurately due to the binary variable $\mathbf{H}$ and orthogonal constraint. Besides, for one-hot label matrix $\mathbf{L}$, the margins between the true labels (belong to the classes) and wrong labels (not belong to the classes) are very small (only 1), which may cause false positive and false negative. To tackle these problems, we relax the true label and wrong label respectively onto $[1,+\infty)$ and $(-\infty,0]$ to produce a large margin. For simplicity, we utilize $\mathbf{Y}$ (initial $\mathbf{Y}=\mathbf{L}$) instead of $\mathbf{L}$ as the regression target to learn hash codes $\mathbf{H}$:

$$\begin{aligned}&\min_{\mathbf{Y},\mathbf{C},\mathbf{H}} \left\|\mathbf{Y}-\mathbf{C}^{\mathrm{T}}\mathbf{H}\right\|_F^2+\delta\left\|\mathbf{Y}\right\|_F^2\\ &s.t.\ \mathbf{H}\in\{-1,1\}^{q\times n},\ \ \mathbf{C}^{\mathrm{T}}\mathbf{C}=\mathbf{I}_c\\ &\forall i\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=0\}}\le 0,\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=1\}}\ge 1\end{aligned} \quad (4)$$

where δ is weight coefficient to control the degree of relaxation. For instance, as shown in Fig. 1, the label $\mathbf{L}_2=(1,0,1)$ is relaxed as $\mathbf{L}_2=(1,0,1+\varepsilon)$, and the semantic vector $\vec{s_2}$ is updated to $\vec{s_2'}$. Accordingly, the resulted length of $\vec{e_2'}$ is smaller than the original one of $\vec{e_2}$, and the total quantization error is reduced by $|\vec{e_2}|-|\vec{e_2'}|$. Considering the extension for out-of-sample data, we impose the projection item in Eq. (4), and the final objective function is defined as following:

$$\begin{aligned}&\min_{\mathbf{H},\mathbf{C},\mathbf{Y},\mathbf{P}_1,\mathbf{P}_2} G(\mathbf{H},\mathbf{C},\mathbf{Y},\mathbf{P}_1,\mathbf{P}_2)\\ &s.t.\ \mathbf{H}\in\{-1,1\}^{q\times n},\ \ \mathbf{C}^{\mathrm{T}}\mathbf{C}=\mathbf{I}_c\\ &\forall i\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=0\}}\le 0,\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=1\}}\ge 1\end{aligned} \quad (5)$$

where

$$\begin{aligned}G=&\left\|\mathbf{Y}-\mathbf{C}^{\mathrm{T}}\mathbf{H}\right\|_F^2+\lambda\sum_{t=1,2}\left\|\mathbf{H}-\mathbf{P}_t\phi(\mathbf{X}^{(t)})\right\|_F^2\\ &+\delta\left\|\mathbf{Y}\right\|_F^2+\gamma\Re(\mathbf{P}_1,\mathbf{P}_2)\end{aligned} \quad (6)$$

where λ and γ are trade-off parameters, $\Re(\cdot)=\|\cdot\|_F^2$ is the regularization term to avoid overfitting, $\phi(\cdot)$ is the RBF kernel [24] which could better capture the underlying nonlinear information in feature space.

C. Optimization for FSePH

To solve the non-convex problem in Eq. (5), we propose an iterative approach that solving any one variable while fixing the others, and the optimization procedures are shown below.

Update $\mathbf{P_1},\mathbf{P_2}$: removing the terms that are irrelevant to $\mathbf{P_1},\mathbf{P_2}$, Eq. (5) can be rewritten as following:

$$\min_{\mathbf{P}_t} G(\mathbf{P}_t)=\lambda\left\|\mathbf{H}-\mathbf{P}_t\phi(\mathbf{X}^{(t)})\right\|_F^2+\gamma\left\|\mathbf{P}_t\right\|_F^2, t=1,2 \quad (7)$$

It is easy to verify that $G(\mathbf{P}_t), t=1,2$ is a convex function, and the optimal solution of $\mathbf{P}_t$ can be computed by:

$$\mathbf{P}_t=\mathbf{H}\phi(\mathbf{X}^{(t)})^T(\phi(\mathbf{X}^{(t)})\phi(\mathbf{X}^{(t)})^T+\gamma/\lambda\mathbf{I})^{-1},\ t=1,2 \quad (8)$$

Update $\mathbf{C}$: removing the terms that are irrelevant to $\mathbf{C}$, Eq. (5) can be rewritten as follows:

$$\min_{\mathbf{C}} G(\mathbf{C})=\left\|\mathbf{Y}-\mathbf{C}^{\mathrm{T}}\mathbf{H}\right\|_F^2,\ \ s.t.\ \mathbf{C}^{\mathrm{T}}\mathbf{C}=\mathbf{I}_c \quad (9)$$

The objective function in Eq. (9) is a typical Orthogonal Procrustes Problem, which can be well solved by computing the Singular Value Decomposition (SVD) of the $q\times q$ matrix $\mathbf{H}\mathbf{Y}^{\mathrm{T}}$, *i.e.*, $\mathbf{H}\mathbf{Y}^{\mathrm{T}}=\mathbf{U}\Sigma\mathbf{V}$, then, updating $\mathbf{C}=\mathbf{U}\mathbf{V}^{\mathrm{T}}$.

Update $\mathbf{H}$: removing the terms that are irrelevant to $\mathbf{H}$, Eq. (5) can be simplified as:

$$\begin{aligned}\min_{\mathbf{H}} G(\mathbf{H})&=\left\|\mathbf{Y}-\mathbf{C}^{\mathrm{T}}\mathbf{H}\right\|_F^2+\lambda\sum_{t=1,2}\left\|\mathbf{H}-\mathbf{P}_t\phi(\mathbf{X}^{(t)})\right\|_F^2\\ &=\underbrace{(1+2\lambda)\left\|\mathbf{H}\right\|_F^2}_{\mathbf{const}}-2tr(\mathbf{H}^{\mathrm{T}}(\mathbf{C}\mathbf{Y}+\lambda\sum_{t=1,2}\mathbf{P}_t\phi(\mathbf{X}^{(t)})))\end{aligned} \quad (10)$$

The discrete solution of $\mathbf{H}$ can be computed through an efficient close-form solution:

$$\mathbf{H}=\mathrm{sgn}(\mathbf{C}\mathbf{Y}+\lambda(\mathbf{P}_1\phi(\mathbf{X}^{(1)})+\mathbf{P}_2\phi(\mathbf{X}^{(2)}))) \quad (11)$$

Update $\mathbf{Y}$: removing the terms that are irrelevant to $\mathbf{Y}$, Eq. (5) can be rewritten as:

$$\begin{aligned}\min_{\mathbf{Y}} G(\mathbf{Y})&=tr\left((\mathbf{H}^{\mathrm{T}}-\mathbf{Y}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}})(\mathbf{H}-\mathbf{C}\mathbf{Y})+\delta\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right)\\ &=\underbrace{\left\|\mathbf{H}\right\|_F^2}_{\mathbf{const}}+tr\left(\mathbf{Y}^{\mathrm{T}}\left((1+\delta)\mathbf{Y}-2\mathbf{C}^{\mathrm{T}}\mathbf{H}\right)\right)\\ &\forall i\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=0\}}\le 0,\ \mathbf{Y}_{i,j\in\{\mathbf{L}_i=1\}}\ge 1\end{aligned} \quad (12)$$

For any wrong label $\mathbf{Y}_{ij}$, we aim to relax it onto $(-\infty,0]$, thus the subproblem of Eq. (12) can be simplified as:

$$\min_{\mathbf{Y}_{ij}\in(-\infty,0]} G(\mathbf{Y}_{ij})=\mathbf{const}+(1+\delta)\mathbf{Y}_{ij}^2-2\mathbf{Y}_{ij}\mathbf{C}_i^{\mathrm{T}}\mathbf{H}_j \quad (13)$$

The subproblem of Eq. (13) is a convex problem, which involves the minimization of a quadratic function over an interval. Let the gradient of $G(\mathbf{Y}_{ij})$ equal to zero, its unconstrained minimum is attained at $\widehat{\mathbf{Y}}_{ij}=\frac{1}{1+\delta}\mathbf{C}_i^{\mathrm{T}}\mathbf{H}_j$. Since the quadratic coefficient is a positive, the unique point of the

minimum of Eq. (13) is attained by mapping $\widehat{\mathbf{Y}}_{ij}$ to $(-\infty, 0]$, and the updating scheme is regularized as:

$$\mathbf{Y}_{ij} = \begin{cases} \mathbf{0}, & \text{if } \widehat{\mathbf{Y}}_{ij} > 0 \\ \widehat{\mathbf{Y}}_{ij}, & \text{if } \widehat{\mathbf{Y}}_{ij} \leq 0 \end{cases} \tag{14}$$

For any true label $\mathbf{Y}_{ij}$, we aim to relax it onto $[1, +\infty)$, thus the subproblem of Eq. (12) can be rewritten as:

$$\min_{\mathbf{Y}_{ij} \in [1, +\infty)} G(\mathbf{Y}_{ij}) = \mathbf{const} + (1+\delta)\mathbf{Y}_{ij}^2 - 2\mathbf{Y}_{ij}\mathbf{C}_i^{\mathrm{T}}\mathbf{H}_j \tag{15}$$

Similarly, the subproblem of true labels is also a convex problem. By setting the gradient of $G(\mathbf{Y}_{ij})$ equal to zero, the unconstrained minimum of Eq. (15) is attained at $\widehat{\mathbf{Y}}_{ij} = \frac{1}{1+\delta}\mathbf{C}_i^{\mathrm{T}}\mathbf{H}_j$ with the mapping interval at $[1, +\infty)$. Therefore, the updating scheme is regularized as:

$$\mathbf{Y}_{ij} = \begin{cases} \widehat{\mathbf{Y}}_{ij}, & \text{if } \widehat{\mathbf{Y}}_{ij} > 1 \\ \mathbf{1}, & \text{if } \widehat{\mathbf{Y}}_{ij} \leq 1 \end{cases} \tag{16}$$

D. Out-of-Sample Extension

According to Eq. (5), we can directly generate hash code for any unseen query samples via the learned modality-specific projections, and the formula as follows:

$$\mathbf{h}_q = \mathrm{sgn}(\mathbf{P}_t \phi(\mathbf{x}_i^{(t)})) \tag{17}$$

where, $\mathbf{x}_i^{(t)}$ is the t-th modality of query sample, and $\mathbf{h}_q$ is the corresponding hash code.

E. Theoretical Analysis

Efficiency of Complexity Analysis: The computational complexity of FSePH mainly involves RBF mapping and the optimization. For RBF mapping, whose complexity is $\mathcal{O}(m^2 + kdn)$, where $d = \max(d_1, d_2)$, m is the number of instances selected to compute the kernel width, and k is the number of anchor points. For solving Eq. (5), whose complexity is $\mathcal{O}(n(q+c+qd+qc+d^2+q^2)+q^3+qd^2+d^3)$. Since $c \leq q < d \ll n$, the former complexity can be simplified as $\mathcal{O}(n(q+c+d^2)+d^3)$. Let t be the iterative number to converge, the overall complexity is approximated as $\mathrm{O}(m^2+kdn+((q+c+d^2)n+d^3)t)$, which is linear to n and competitive to existing methods. The iteration t is always less than 30 in experiments.

IV. Experiments

A. Experimental Settings

1) *Data sets*: In the experiment, the popular MIRFlickr [25] and NUS-WIDE [26] datasets are selected for testing. The features are kept as the same as in work [27]. For MIRFlickr, we keep 20015 image-text pairs whose textual tags appear more than 20 times, and randomly select 2000 instances as a test set and the rest as training set. For NUS-WIDE, we remain 186577 image-text pairs which belong to most frequent concepts, and randomly select 1866 instances as a test set, while the rest is selected as training set.

2) *Baseline methods*: We compare the proposed FSePH with state-of-the-art unsupervised methods (*i.e.*, CMFH [6], IMH [5]) and supervised methods (SCM [9], SePH [10], GSePH [13], DCH [14]). It is noted that CMFH, GSePH and SePH are too computationally intensive in training, especially when the dataset is large (*i.e.*, NUS-WIDE). Therefore, to avoid highly computational cost on NUS-WIDE, following the setting of literature [10, 14], we randomly select 10000 instances from its retrieval set to form the new training sets for these methods, and set the number of RBF anchor as 800 for SePH and GSePH. For all the baselines, we utilize the source codes kindly provided by the respective authors.

3) *Evaluation Metric*: Mean of average precision (mAP@R) and topK-precision are utilized to evaluate the retrieval performance, including retrieving text with given image (I→T) and retrieving image with given text (T→I).

B. Results and Discussions

1) *Results of retrieval tasks*: The MAP scores and top50 precisions tested with different datasets are shown in Table I, while the comparisons of topK-precision curves on different datasets are shown in Fig. 2, respectively. Meanwhile, we perform the ablation studies of FSePH without relaxation for label value (abbreviated as FSePH_WR) and without RBF mapping for input data (abbreviated as FSePH_NR).

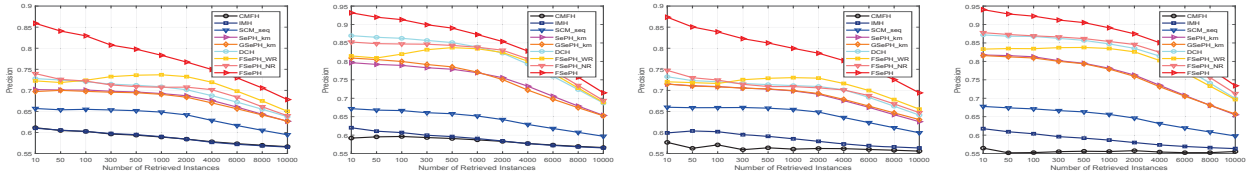
As shown in Table I, it can be observed that FSePH always outperforms the competing baselines. Evidently, the retrieval performances obtained by the proposed FSePH are much better than the results generated by unsupervised methods, *i.e.*, CMFH and IMH. The main reason lies that CMFH and IMH learn the hash code by preserving the feature similarities. By contrast, the proposed FSePH, SCM, SePH_km, GSePH_km and DCH produce the hash code by preserving the semantic similarities. It means that the semantic information is really helpful to improve the cross-modal retrieval performance. Compared with the best semantic-preserving baselines, the FSePH has also significantly improved the cross-modal retrieval performance. As shown in Fig. 2, FSePH always yields the highest precision scores than the baselines at different number of retrieved instances. That is, the proposed FSePH approach is able to search much more similar samples in the beginning, which is very important for a practical retrieval system [28].

2) *Results of ablation studies*: We heuristically evaluate the effectiveness of the proposed FSePH with different learning modules. As illustrated in Table I, the MAP scores attained by FSePH_WR and FSePH_NR have also delivered very competitive performances, especially tested on NUS-WIDE dataset. That is, the reasonable relaxation of label values is able to reduce the quantization error, while the utilization of RBF mapping can capture the nonlinear structure of data to improve retrieval performance. Importantly, the MAP scores obtained by FSePH are higher than that produced by FSePH_WR and FSePH_NR in most cases. That is, the integration of relaxed label value and RBF mapping could yield more discriminative hash codes for performance improvements.

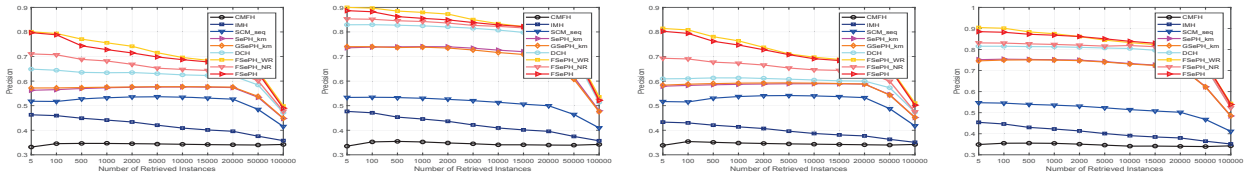
3) *Results of training time*: We perform all experiments on different subsets of NUS-WIDE dataset and record the training time when the code length is set at 128 bits, as shown in Table II. It is noted that the recent competing methods, i.e.,

TABLE I: The MAP scores and top50 precisions tested on handcrafted features.

Method/Dataset		MAP						Top50 precision					
		MIRFlickr			NUS-WIDE			MIRFlickr			NUS-WIDE		
		32 bits	64 bits	128 bits	32 bits	64 bits	128 bits	32 bits	64 bits	128 bits	32 bits	64 bits	128 bits
I→T	CMFH	0.5722	0.5582	0.5581	0.3429	0.3433	0.3431	0.6361	0.5924	0.5879	0.4201	0.4235	0.4240
	IMH	0.5718	0.5685	0.5650	0.3738	0.3609	0.3548	0.6383	0.6321	0.6288	0.5086	0.4770	0.4632
	SCM	0.6201	0.6287	0.5774	0.5023	0.5097	0.5083	0.6784	0.6847	0.6478	0.5552	0.5528	0.5483
	SePH_km	0.6679	0.6713	0.6710	0.5628	0.5752	0.5811	0.7240	0.7344	0.7312	0.5908	0.6115	0.6217
	GSePH_km	0.6570	0.6649	0.6694	0.5565	0.5710	0.5769	0.7204	0.7325	0.7391	0.6045	0.6129	0.6202
	DCH	0.6906	0.7017	0.6974	0.6398	0.6443	0.6626	0.7523	0.7523	0.7598	0.6633	0.6209	0.6515
	FSePH_WR	0.7097	0.7101	0.7042	0.6876	**0.7066**	0.7060	0.7387	0.7524	0.7228	0.8073	0.8259	0.8298
	FSePH_NR	0.6981	0.7070	0.7130	0.6630	0.6659	0.6677	0.7602	0.7659	0.7645	0.7437	0.7133	0.7510
	FSePH	**0.7416**	**0.7597**	**0.7717**	**0.6890**	0.7038	**0.7084**	**0.8668**	**0.8825**	**0.8968**	**0.8163**	**0.8292**	**0.8370**
T→I	CMFH	0.5718	0.5562	0.5560	0.3418	0.3422	0.3421	0.6231	0.5891	0.5960	0.4101	0.4116	0.4107
	IMH	0.5710	0.5685	0.5651	0.3705	0.3605	0.3530	0.6441	0.6422	0.6350	0.5224	0.5017	0.4698
	SCM	0.6107	0.6129	0.5822	0.4516	0.4541	0.4543	0.6918	0.6994	0.6584	0.5689	0.5802	0.5857
	SePH_km	0.7102	0.7158	0.7206	0.6670	0.6738	0.6705	0.8122	0.8295	0.8392	0.7528	0.7685	0.7701
	GSePH_km	0.7004	0.7100	0.7166	0.6523	0.6700	0.6776	0.8210	0.8289	0.8405	0.7632	0.7623	0.7792
	DCH	0.7760	0.7963	0.7923	0.7822	0.8018	0.8172	0.8779	0.8787	0.8914	0.8339	0.8185	0.8219
	FSePH_WR	0.7775	0.7828	0.7711	0.8052	0.8202	0.8173	0.8229	0.8437	0.8318	0.8842	**0.9088**	0.9034
	FSePH_NR	0.7887	0.8121	0.8204	0.8035	0.8088	0.8195	0.8634	0.8832	0.8777	0.8593	0.8342	0.8588
	FSePH	**0.8064**	**0.8296**	**0.8448**	**0.8091**	**0.8231**	**0.8269**	**0.9327**	**0.9435**	**0.9476**	**0.8960**	0.8938	**0.9084**



(a) I→T@32 bits on MIRFlickr (b) T→I@32 bits on MIRFlickr (c) I→T@64 bits on MIRFlickr (d) T→I@64 bits on MIRFlickr



(e) I→T@32 bits on NUS-WIDE (f) T→I@32 bits on NUS-WIDE (g) I→T@64 bits on NUS-WIDE (h) T→I@64 bits on NUS-WIDE

Fig. 2: TopK-Precision curves tested on different datasets.

CMFH, IMH, SePH and GSePH, only sample a small training subset from NUS-WIDE dataset, for reason that their overall complexities are infeasible to train on a large-scale dataset. The kernel time is the another part contributed to the training time in SePH, GSePH and the proposed FSePH method. In general, the computation time of kernel mapping is a constant.

TABLE II: Training time (in second) on subsets of NUS-WIDE.

Method	1K	5K	10K	50K	184K	Kernel time
CMFH	43.03	345.14	977.17	-	-	0.00
IMH	0.59	21.33	140.26	-	-	0.00
SCM	13.16	13.94	13.83	18.18	25.79	0.00
SePH_km	2.27	64.32	250.36	-	-	323.50
GSePH_km	135.23	1343.94	4182.59	-	-	352.82
DCH	1.08	2.15	4.45	40.21	242.84	0.00
FSePH	0.15	0.42	0.81	4.58	11.29	45.05

Although the training time obtained by SCM seems to be faster than other methods on a large-scale training size, yet its retrieval performance is not desirable, as shown in Table I. By contrast, it can be found that the proposed FSePH method not only significantly reduces the training time, but also can achieve best retrieval performance. The main reason contributed to such faster learning performance are two-fold: 1) The orthogonal constraint is embedded to learn the hash codes, which can well reduce the quantization error and speed up the learning process. 2) The relaxed label values can make the optimization converge within limited iterations. 3) FSePH has a close-form solution to hash code learning and only requires a single step to update the whole hash codes, instead of iteratively updating hash codes bit by bit (*e.g.*, DCH). Therefore, FSePH could significantly speed up the cross-modal retrieval on large-scale database.

V. Conclusion

This paper has proposed a novel Fast Semantic Preserving Hashing approach for large-scale cross-modal retrieval, which can well preserve the semantic similarities from original data to a shared Hamming space. The proposed FSePH introduces

an orthonormal basis to regress the targeted hash codes of training examples to their corresponding reasonably relaxed class label values, which has the provable large margin property to efficiently reduce the quantization error. Meanwhile, an effective optimization algorithm is derived for orthonormal basis, projection function and relaxed label value learning, meanwhile an efficient closed-form solution is exploited for hash code learning. Experiment results have shown its superior performance. Our future work will consider more complex multi-modal data, e.g., audio-visual data [29].

REFERENCES

[1] Y. Hua, S. Wang, S. Liu, Q. Huang, and A. Cai, "Tina: Cross-modal correlation learning by adaptive hierarchical semantic aggregation," in *Proc. IEEE ICDM*, 2014, pp. 190–199.

[2] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM MM*, 2010, pp. 251–260.

[3] A. Li, S. Shan, X. Chen, and G. Wen, "Cross-pose face recognition based on partial least squares," *Pattern Recognit. Lett.*, vol. 32, no. 15, pp. 1948–1955, 2011.

[4] Y. Cao, H. Qi, W. Zhou, J. Kato, K. Li, X. Liu, and J. Gui, "Binary hashing for approximate nearest neighbor search on big data: A survey," *IEEE Access*, vol. 6, pp. 2039–2054, 2017.

[5] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *ACM SIGMOD*, 2013, pp. 785–796.

[6] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. CVPR*, 2014, pp. 2075–2082.

[7] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. ACM SIGIR*, 2014, pp. 415–424.

[8] M. Long, Y. Cao, J. Wang, and P. S. Yu, "Composite correlation quantization for efficient multimodal retrieval," *Computer Science*, pp. 579–588, 2016.

[9] D. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI*, 2014, pp. 2177–2183.

[10] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. CVPR*, 2015, p. 3864–3872.

[11] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, 2016.

[12] L. Hong, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. CVPR*, 2017, pp. 6345–6353.

[13] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proc. CVPR*, 2017, p. 4076–4084.

[14] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, p. 2494–2507, 2017.

[15] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proc. CVPR*, 2017, p. 3132–3240.

[16] Y. Cao, M. Long, J. Wang, and S. Liu, "Collective deep quantization for efficient cross-modal retrieval," in *Pro. AAAI*, 2017, pp. 3974–3980.

[17] X. Gong, L. Huang, and F. Wang, "Deep semantic correlation learning based hashing for multimedia cross-modal retrieval," in *Proc. ICDM*, 2018, pp. 117–126.

[18] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Fast supervised discrete hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 490–496, 2018.

[19] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. CVPR*, 2012, pp. 2160–2167.

[20] N. Rasiwasia, D. Mahajan, V. Mahadevan, Aggarwal, and Gaurav, "Cluster canonical correlation analysis," in *Proc. IJACI*, 2014, pp. 823–831.

[21] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. ICCV*, 2015, pp. 4094–4102.

[22] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.

[23] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, 2013.

[24] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. ICCV*, 2010, pp. 2130–2137.

[25] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative," in *Proc. ICMIR*, 2010, pp. 527–536.

[26] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. CIVR*, 2009, pp. 48:1–48:9.

[27] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 11, pp. 5292–5303, 2018.

[28] Y. Chen, X. Hu, W. Fan, L. Shen, Z. Zhang, X. Liu, J. Du, H. Li, Y. Chen, and H. Li, "Fast density peak clustering for large scale data based on knn," *Knowledge-Based Systems*, 2019.

[29] X. Liu, J. Geng, H. Ling, and Y. M. Cheung, "Attention guided deep audio-face fusion for efficient speaker naming," *Pattern Recognition*, vol. 88, pp. 557–568, 2019.