Jointly Sparse Hashing for Image Retrieval

Zhihui Lai, Yudong Chen, Jian Wu, Wai Keung Wong, and Fumin Shen

***Abstract*— Recently, hash learning attracts great attentions since it can obtain fast image retrieval on large-scale data sets by using a series of discriminative binary codes. The popular methods include manifold-based hashing methods, which aim to learn the binary codes by embedding the original high-dimensional data into low-dimensional intrinsic subspace. However, most of these methods tend to relax the discrete constraint to compute the final binary codes in an easier way. Therefore, the information loss will increase. In this paper, we propose a novel jointly sparse regression model to minimize the locality information loss and obtain jointly sparse hashing method. The proposed model integrates locality, joint sparsity, and rotation operation together with a seamless formulation. Thus, the drawback in previous methods using two separated and independent stages such as PCA-ITQ and the similar methods can be addressed. Moreover, since we introduce the joint sparsity, the feature extraction and jointly sparse feature selection can also be realized in a single projection operation, which has the potentials to select more important features. The convergence of the proposed algorithm is proved, and the essences of the iterative procedures are also revealed. The experimental results on large-scale data sets demonstrate the performance of the proposed method. The source code can be downloaded from http://www.scholat.com/laizhihui.**

***Index Terms*— Hash learning, regression model, feature selection.**

I. Introduction

THE retrieval on large-scale dataset is a critical problem in the fields of machine learning. A feasible way to achieve effective storage and fast retrieval is to utilize hash learning method to obtain the binary codes. Hash learning methods aim to represent the data by a set of binary codes so that the feature dimensionality and storage space is reduced simultaneously.

The data-independent hashing methods, such as Coherency Sensitive Hashing (CSH) [1] and Locality Sensitive Hashing (LSH) [2], used random projections to map the data and required long binary codes for good performance. In order to learn high quality binary codes for the training and query samples, some data-dependent hash learning methods were proposed. The supervised methods, such as Kernelize Supervised Hashing (KSH) [3], Asymmetric Discrete Graph Hashing (ADGH) [4] and Supervised Discrete Hashing (SDH) [5], took advantages of labeled information for discriminative analysis and binary coding. Due to their promising performance in image retrieval, they have been well studied and extended to various forms [6]–[8].

However, the labeled information is tedious to obtain and labeling the data will cost a lot in real-world big data applications. It is desirable to develop effective unsupervised hashing methods for binary codes learning. As a well-known unsupervised hash learning method, Iterative Quantization (ITQ) [9] imposed an orthogonal rotation matrix on the model to iteratively modify the learned binary bits. By using the rotation operation, ITQ greatly reduced the information loss between binary codes and low-dimensional features. However, ITQ separately performed dimensionality reduction and hash learning which led to sub-optimal solutions. Besides, ITQ mainly focused on the global information of dataset. In contrast, the manifold-based methods, such as Spectral Hashing (SH) [10], Anchor Graph Hashing (AGH) [11], Inductive Manifold Hashing (IMH) [12], [13] and Sparse Embedding and Least Variance Encoding (SELVE) [14] aimed at discovering the manifold structure of the high-dimensional dataset and preserving it in the low-dimensional binary codes space.

Previous research [15]–[18] showed that the images of an object lied on a low-dimensional manifold embedded in the high-dimensional space and the local geometric structure was of very importance in computer vision and pattern recognition. Thus, the manifold-based hash learning methods, such as SH and AGH, took advantages of the local structure to compute the binary representation. SH first learns the eigenvectors from the Laplacian graph of datasets and then performs binarization operation on the low-dimensional real values. The algorithm steps in SH lead to two problems. First, the time complexity of computing Laplacian graph is unacceptable for large-scale datasets. Second, the information loss between low-dimensional real value features and binary codes will increase because of the binarization operation. To save the computational cost for obtaining large-scale graph, AGH designs a

Manuscript received March 26, 2018; revised July 15, 2018; accepted August 12, 2018. Date of publication August 30, 2018; date of current version September 17, 2018. This work was supported in part by the Natural Science Foundation of China under Grant 61573248, Grant 61773328, Grant 61802267, and Grant 61732011, in part by the Natural Science Foundation of Guangdong Province under Grant 2017A030313367, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20170302153434048 and Grant JCYJ20160429182058044, and in part by the Hong Kong Polytechnic University under Project G-UC42. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Yudong Chen.)*

Z. Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong (e-mail: lai_zhi_hui@163.com).

Y. Chen is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: andrewlin7@qq.com).

J. Wu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: wujianhitsz@163.com).

W. K. Wong is with the Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong (e-mail: calvin.wong@polyu.edu.hk).

F. Shen is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: fshen@uestc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2867956

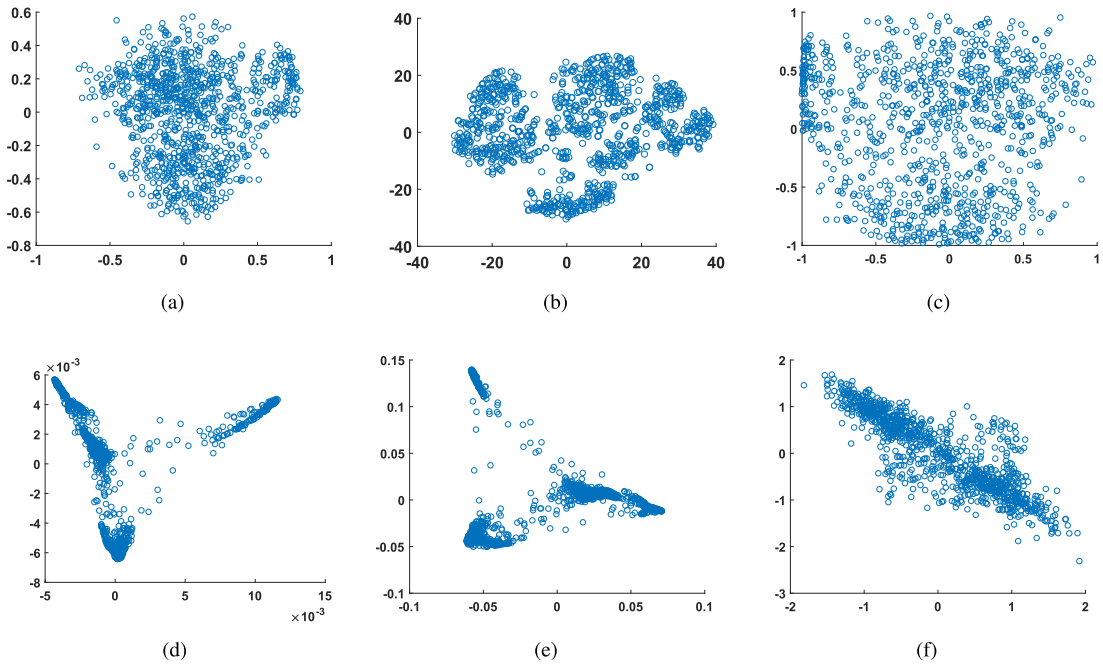


Fig. 1. The (a) original distribution and (b) t-SNE representation of testing samples on MNIST dataset on two-dimensional space. The real values of testing samples learned by (c) SH, (d) AGH, (e) IMH and (f) Ours on MNIST dataset with 2 bits length. The information loss of SH, AGH, IMH and Ours are 27.73, 31.32, 31.27 and 29.29 respectively. Although SH obtains the lowest information loss, the samples are projected evenly on the low-dimensional spaces and thus, the manifold structure of the data is destroyed in binarization. On the other hand, AGH and IMH relax the discrete constraint in binarization which increase the information loss. Our method takes the comprehensive strategy to balance the less information loss and manifold structure preservation in binarization.

modified version of SH by introducing the anchor graph. Since the number of anchors are far less than the training samples, the storage space and computational cost in constructing the graph is greatly reduced. Due to the efficiency of anchor-based graph construction, it is widely applied in subspace learning and hash learning methods to compute the low-dimensional representation [19]–[21]. However, AGH still fails to control the information loss between the real values of features and the binary codes in the learning steps. Although Large Graph Hashing with Spectral Rotation (LGHSR) [22] added an orthogonal rotation operation to minimize the accumulated error, it still used two separated steps to obtain the binary codes.

Moreover, the ability of sparse feature selection is ignored in the previous manifold-based hashing methods. In recent years, sparse subspace learning attracted great attentions since it could effectively improve the model's performance by introducing sparse feature selection. The sparse dimensionality reduction methods, such as Sparse Principal Component Analysis (SPCA) [23], [24] and Sparse Discriminant Analysis (SDA) [25], generated sparse projection matrix for sparse feature extraction by imposing elastic net constraint on the model. However, the L_1-norm regularized learning methods cannot provide consistent explanation to the selected features. Therefore, the jointly sparse feature selection methods were proposed [26]–[28]. Unlike the L_1-norm-based regularization term generates sparse projections independently, the $L_{2,1}$-regularized learning methods can shrink all the coefficients of corresponding unnecessary features and thus, provide row-sparsity projection matrix. Motivated by this desirable property, some dimensionality reduction methods performed subspace learning and feature selection simultaneously [29], [30] by utilizing $L_{2,1}$-norm regularization term. The sparse hashing methods proposed in [7], [31], and [32] obtained sparse coding matrix by using sparse constraint. However, all of them mainly focused on saving the storage space of coding matrix which cannot obtain jointly sparse feature selection.

In this paper, we propose a $L_{2,1}$-regularized regression formulation to minimize the information loss between binary codes and low-dimensional feature vectors. The original idea of the proposed method is to combine manifold learning and jointly sparse regression together for hash learning. We hope to do linear regression (i.e. rotation operation) in one subspace so that we can preserve the latent geometric structure described by the graph to learn the hash codes and obtain the optimal binary solutions directly. Since we utilize $L_{2,1}$-norm regularization term to obtain jointly sparse projection matrix, the model has better interpretability to the selected features [33]–[35]. Differing from the previous manifold-based methods to compute the eigenvectors at first and then to binarize the projective features using sign function, we integrate the low-dimensional feature learning and orthogonal rotation into a unified framework. Therefore, the optimal projection matrix and regression matrix (i.e. rotation matrix) can be jointly optimized, which further minimizes the information loss. Moreover, since the regression matrix provides an intrinsic connection between the binary codes and low-dimensional

features, we can directly use the learned matrix to obtain the binary codes of test samples. Therefore, the testing time can be reduced. Figure 1 shows the distribution of real values learned by SH, AGH, IMH and our algorithm on MNIST dataset with 2 bits length. Although SH obtains lowest information loss, it is obvious that SH cannot well preserve the manifold structure since the data is distributed evenly on the low-dimensional subspace. Even if AGH and IMH can preserve the manifold structure of the data, the information loss is larger than our method in binarization, which will greatly affect the effectiveness of the binary codes (for example, they map the 0.01 to be 1 in most cases). In total, our method gives more reasonable result with less information loss compared to AGH and IMH.

In summary, the problems of information loss and feature selection are still unsolved in the previous manifold-based methods. In order to address these drawbacks, we propose Jointly Sparse Hashing (JSH). The contributions of this paper are as follows:

1) The proposed method integrates the dimensionality reduction, jointly sparse feature selection and the rotation operation with a seamless formulation. And thus the drawback in previous methods using two separated and independent stages such as PCA-ITQ and the similar methods can be addressed. Moreover, the proposed model can control the information loss in a better way and directly obtain the discrete solutions without relaxation.

2) The convergence of the proposed iterative algorithm is proved. Furthermore, the essence of the iterative algorithm is also uncovered.

The remaining sections are organized as follows: In Section II, we review the unsupervised hash learning method SH and the construction of anchor graph. In Section III, we introduce the proposed method Jointly Sparse Hashing (JSH) and the designed iteratively reweighted algorithm. Convergence analysis, computational complexity and theoretical analysis of the algorithm will be discussed in Section VI. Section V gives the experimental results on different datasets and Section VI summarizes this paper.

II. Related Works

In this section, we first introduce the notations and definitions, and then review the traditional manifold-based hash learning method, i.e. Spectral Hashing (SH), and the construction of anchor graph.

A. Notations and Definitions

In this paper, the training samples are denoted by $X = [x_1, x_2, ..., x_i, ..., x_n]$, each column is a d-dimensional feature vector. The hash learning methods aim to obtain a set of binary codes $B = [b_1, b_2, ..., b_i, ..., b_n]$ to represent the original data, where $b_i \in R^L$ and $L << d$ is the number of bits. A common way to obtain the binary values of each sample is to utilize a sign function as follows:

$$b_i = sign(Fx_i), \tag{1}$$

where $F \in R^{L\times d}$ is coding matrix and $sign(\cdot)$ is binary function which automatically assigns 1 of -1 according to the real values. In order to obtain jointly sparse coding matrix for feature selection, we impose $L_{2,1}$-norm on the proposed model. The definition of $L_{2,1}$-norm is:

$$||X||_{2,1} = \sum_{i=1}^{d} ||x^i||_2, \tag{2}$$

where x^i is the i-th row of X and $||\cdot||_2$ is L_2-norm.

B. Spectral Hashing

To preserve the manifold structure in the binary codes space, SH aims to minimize the representation error of the binary codes between different neighbor points. With this purpose, the objective function of SH is defined as follows:

$$\begin{aligned} &\min_B \sum_{ij} ||b_i - b_j||_2^2 W_{ij} \\ &\text{s.t. } B \in \{-1, 1\}^{L\times n}, \quad B1 = 0, \ BB^T = nI_L, \end{aligned} \tag{3}$$

where I_L is identity matrix, $W_{ij} = \exp(-||x_i - x_j||^2/\varepsilon^2)$ and ε is parameter in SH. From problem (3), we can derive the following minimization problem:

$$\begin{aligned} &\min_B tr(B(D - W)B^T) \\ &\text{s.t. } B \in \{-1, 1\}^{L\times n}, \quad B1 = 0, \ BB^T = nI_L, \end{aligned} \tag{4}$$

where $D_{ii} = \sum_i W_{ij}$ and $D - W$ is Laplacian graph. Since it is difficult to solve problem (4) with the discrete constraint, SH obtains the binary codes by spectral relaxation. Therefore, the solutions of (4) are the eigenvectors of $D - W$ corresponding to the L minimal eigenvalues (excluding the zero eigenvalues). And the results of $sign(B)$ are the learned codes.

C. Anchor Graph Construction

Constructing the Laplacian graph used in the problem (3) requires $O(dn^2)$ in computational complexity, which is time-consuming when n is very large. To address this problem, AGH designs an anchor-based graph which greatly reduces the computation cost and achieves promising performance. Before constructing anchor graph, AGH performs clustering algorithm on the training dataset to obtain m anchor points $[u_1, u_2, ..., u_i, ..., u_m] \in R^d$. Then, the element of anchor graph Z is defined as:

$$Z_{ij} = \begin{cases} \dfrac{\exp(-||x_i - u_j||^2/\vartheta)}{\sum_{j'\in\{i\}} \exp(-||x_i - u_{j'}||^2/\vartheta)}, & \forall j \in \{i\}. \\ 0, & otherwise, \end{cases} \tag{5}$$

where ϑ is parameter defined in AGH and $\{i\}$ denotes the index set of k nearest neighbor anchors of x_i. There are some advantages by using anchor graph on the model. The merits related to the proposed method including three aspects. First, since $m << n$, the size of anchor graph is much smaller than the traditional Laplacian graph. Therefore, the storage space

can be saved. Second, the computational complexity in constructing the graph used in the model will be decreased. Third, anchor graph uses the k nearest neighbor anchors to compute the similarity, which also characterizes the manifold structure of high-dimensional dataset so that the anchor graph-based model can preserve the local structure in low-dimensional subspaces to a certain extent.

III. Jointly Sparse Hashing

Motivated by the model of SH, the construction of anchor graph and the rotation operation in ITQ, we propose a novel regression model for hash learning in this section.

A. Motivation

The graph-based hash learning methods, such as SH, AGH and IMH, are limited to the discrete constraint. Their common property is to utilize spectral relaxation to compute the eigenspectral and then obtain the binary codes using sign function. The relaxation in optimization not only degrades the effectiveness of the learned codes, but also increases the information loss between the real value representation and the binary representation. Besides, since the relation between the learned binary codes and test samples is implicit, the coding step of these methods will be slower than the projection-based hashing methods, such as ITQ and SDH. In this paper, we consider how to solve these problems simultaneously. A feasible way to minimize the information loss and build the connection between binary codes and the original data is to formulate a regression model integrating the dimensionality reduction, feature selection and rotation operation. Motivated by the effectiveness of orthogonal rotation operation in ITQ, by designing an orthogonal matrix in regression to perform rotation, the information loss will be minimized. Since we impose $L_{2,1}$-norm on the model to obtain jointly sparse projection matrix, the model can select more discriminative features during the learning process and achieve feature selection on the binarization. Thus, dimensionality reduction, feature selection, minimum information loss will be integrated in the proposed method.

B. Objective Function

Aiming at directly learning the coding matrix and minimizing the information loss, we can easily obtain the following problem:

$$\min_{B,F} \sum_{i=1}^{n} ||b_i - Fx_i||_2^2 \quad \text{s.t. } B \in \{-1, 1\}^{L\times n}. \tag{6}$$

This idea can be found in the ITQ algorithm. However, ITQ mainly focuses on the global information and the training samples need to be preprocessed by dimensionality reduction algorithms, such as Principal Component Analysis (PCA) [36] and Canonical Correlation Analysis (CCA) [37]. Thus, ITQ obtains the suboptimal solutions since it uses two separated steps. In this paper, we aim to preserve the local structure in the binary codes space and build the direct connection between the low-dimensional features and binary codes. We first introduce anchor graph to the objective function for manifold learning. Then, in order to integrate the idea of rotation operation to minimize the quantization error (i.e. information loss), we further decompose coding matrix and derive the following optimization problem:

$$\begin{aligned} \min_{\tilde{B},A,P} & \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j - AP^T x_i||_2^2 Z_{ij} \\ \text{s.t. } & \tilde{B} \in \{-1, 1\}^{L\times m}, \quad A^T A = I_s, \end{aligned} \tag{7}$$

where I_s is identity matrix, $\tilde{B}$ is binary codes of anchor points and $\tilde{b}_j$ is the j-th column of $\tilde{B}$, $P \in R^{d\times s}$, $A \in R^{L\times s}$ and s is dimension of low-dimensional features. Different from the SH and AGH, which directly obtain the low-dimensional features before generating binary codes, in this paper, the matrix P can be viewed as a projective matrix for dimensionality reduction on the high-dimensional data. Then the orthogonal matrix A is imposed to regress the low-dimensional features to the nearest binary codes of anchors with an orthogonal rotation to minimize the quantization error. Minimizing (7) means to preserve the anchor graph such that the neighboring points of the anchors can be preserved.

Since model (7) cannot perform jointly sparse feature selection, we further impose $L_{2,1}$-norm regularization term to obtain jointly sparse projection matrix for feature extraction and selection. Finally, the objective function of the proposed Jointly Sparse Hashing (JSH) is defined as follows:

$$\begin{aligned} \min_{\tilde{B},A,P} & \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j - AP^T x_i||_2^2 Z_{ij} + \alpha||P||_{2,1} \\ \text{s.t. } & \tilde{B} \in \{-1, 1\}^{L\times m}, \quad A^T A = I_s, \end{aligned} \tag{8}$$

where $\alpha \geq 0$ is balance parameter. Since $L_{2,1}$-norm tends to generate row-sparsity matrix, then P will be jointly sparse in rows and $F = AP^T$ will also be jointly sparse in corresponding columns. Therefore, the proposed JSH can perform jointly sparse feature extraction and selection for hash learning.

C. Optimization

To solve the $L_{2,1}$-norm based model (8), we first introduce the following reweighted matrix:

$$G_{ii} = \frac{1}{2||p^i||_2}, \tag{9}$$

where G_{ii} is the i-th diagonal element of G and p^i is the i-th row of matrix P. Then, objective function (8) can be rewritten as:

$$\begin{aligned} \min_{\tilde{B},A,P} & \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j - AP^T x_i||_2^2 Z_{ij} + \alpha tr(P^T GP) \\ \text{s.t. } & \tilde{B} \in \{-1, 1\}^{L\times m}, \quad A^T A = I_s. \end{aligned} \tag{10}$$

We solve problem (10) by iteratively optimize P, A and $\tilde{B}$. Problem (10) can be written as:

$$\begin{aligned}&\min_{\tilde{B},A,P}\sum_{j=1}^{m}\sum_{i=1}^{n}||\tilde{b}_j - AP^T x_i||_2^2 Z_{ij} + \alpha tr(P^T GP)\\ &= \min_{\tilde{B},A,P}\sum_{j=1}^{m}\sum_{i=1}^{n} tr(\tilde{b}_j^T Z_{ij}\tilde{b}_j - 2\tilde{b}_j^T Z_{ij} AP^T x_i\\ &\quad + x_i^T PA^T Z_{ij} AP^T x_i) + \alpha tr(P^T GP). \end{aligned} \tag{11}$$

Since $\sum_i Z_{ij} = 1$, the weighted matrix between X and X^T becomes identity matrix and we omit it in (12). Due to the binary property of matrix $\tilde{B}$, the term $\tilde{b}_j^T Z_{ij}\tilde{b}_j$ becomes a constant. Then the optimization problem becomes:

$$\begin{aligned}&\min_{\tilde{B},A,P}\ tr(-2AP^T XZ\tilde{B}^T + P^T(\alpha G + XX^T)P)\\ &\text{s.t. } \tilde{B}\in\{-1,1\}^{L\times m},\quad A^T A = I_s. \end{aligned} \tag{12}$$

The following steps give the optimal solutions of the proposed model.

Step 1: To compute the optimal P. Taking the partial deviation with respect to P to be zero, we derive:

$$P = (\alpha G + XX^T)^{-1} XZ\tilde{B}^T A. \tag{13}$$

Step 2: To compute the optimal A. Given P and $\tilde{B}$, problem (12) is equal to the following maximization problem:

$$\begin{aligned}&\max_{A}\ tr\ (AP^T XZ\tilde{B}^T)\\ &\text{s.t.}\quad A^T A = I_s. \end{aligned} \tag{14}$$

Problem (14) has been studied in [22] and [23]. Denoted the Singular Value Decomposition of $P^T XZ\tilde{B}^T = \tilde{U}\tilde{D}\tilde{V}^T$, then the optimal solution of (14) is:

$$A = \tilde{V}\tilde{U}^T. \tag{15}$$

Step 3: To compute the optimal $\tilde{B}$. Given P and A, we can derive the following problem:

$$\begin{aligned}&\max_{\tilde{B}}\ tr(AP^T XZ\tilde{B}^T)\\ &\text{s.t. } \tilde{B}\in\{-1,1\}^{L\times m}. \end{aligned} \tag{16}$$

Thus, the solution of $\tilde{B}$ is as follows:

$$\tilde{B} = sign(AP^T XZ). \tag{17}$$

The details of the proposed algorithm are described in Algorithm 1.

Algorithm 1 Jointly Sparse Hashing

Input: Training set $X\in R^{d\times n}$, number of neighbor points, anchors, desired dimensions and bits (i.e. k, m, s, L), parameter α and iteration times T.
Output: Coding matrix F.
1: Use clustering algorithm to generate m anchors.
2: Construct anchor graph Z based on (5).
3: Initialize A as orthogonal matrix, G as identity matrix and $\tilde{B}$ as random binary matrix.
4: Begin iteration.
For $i = 1:T$ do
Step 1: $P = (\alpha G + XX^T)^{-1} XZ\tilde{B}^T A$.
Step 2: $A = \tilde{V}\tilde{U}^T$.
Step 3: $\tilde{B} = sign(AP^T XZ)$.
Step 4: update G using (9).
5: $F = AP^T$.

IV. Theoretical Analysis

In this section, we discuss the convergence and computational complexity of Algorithm 1. The essence of the algorithm is also presented.

A. Convergence

We establish a week convergence of the proposed algorithm. To prove the convergence, we need two lemmas as follows.

Lemma 1 [38]: For nonzero vectors $\tilde{a}$ and $\tilde{b}$, the following inequality holds:

$$||\tilde{a}||_2 - \frac{||\tilde{a}||_2^2}{2||\tilde{b}||_2} \le ||\tilde{b}||_2 - \frac{||\tilde{b}||_2^2}{2||\tilde{b}||_2}. \tag{18}$$

Lemma 2: The objective function value is monotonically non-increasing in each iteration.

Proof: The proof is given in Appendix section.

Lemma 2 indicates that the objective value is monotonically non-increasing in each iteration. Then we need to prove that the sequence $\tilde{B}^t, A^t, P^t$ generated by Algorithm 1 is bounded and has limit points so that the convergence of the algorithm can be established. We point out that the proposed model (8) is a combinational optimization problem since the variable $\tilde{B}$ is not continuous. Hence the problem (8) is not a convex problem, but we can partially discuss its convexity. With respect to the variable A and P for fixed $\tilde{B}$, the model (8) is proper and convex.

Theorem 1: The sequence $\tilde{B}^t, A^t, P^t$ generated by Algorithm 1 is bounded and has limit points.

Proof: The domain of the variable $\tilde{B}$ is finite and hence is bounded. On one hand, the objective function with respect to the variables A and P is proper and convex, and on the other hand, the objective function is monotonically non-increasing on the sequence A^t, P^t for fixed $\tilde{B}$, therefore, the sequence A^t, P^t is bounded. Based on the above discussions, we point out that the sequence $\tilde{B}^t, A^t, P^t$ is bounded and (according to the Weistrass Theorem [39]) has limit points. ■

Theorem 2: Each limit point of the sequence $\tilde{B}^t, A^t, P^t$ is a solution of the problem (8).

Proof: The proof is given in Appendix section.

B. Computational Complexity

The main computation cost of the iterative algorithm is from the construction of anchor graph Z and optimization of

projection matrix P. Compared to the traditional Laplacian graph, the usage of anchor graph greatly reduces the storage space and computation cost. The same as AGH and IMH, it totally costs $O(t_c dnm + dnm)$ to construct anchor graph if we utilize clustering algorithm to generate anchors, where t_c is the iteration times for clustering algorithm. As for the optimization of matrix P, it first costs $O(d^2 n)$ and $O(dnm)$ to compute XX^T and XZ, respectively. Then, in each iteration, it costs $O(d^3)$ to compute $(\alpha G + XX^T)^{-1}$. The computational complexity of iterative part is $O(d^2 n + dnm + Td^3)$, where T is the number of iterations. Thus, the total computational complexity is $O(t_c dnm + 2dnm + d^2 n + Td^3)$, which is linear with the number of the data. Although the training time of the proposed algorithm is higher than AGH and IMH, the coding time of testing samples can be reduced since we directly obtain the jointly sparse coding matrix.

C. Essence of The Algorithm

In this subsection, we analyze the algorithm essence or the algorithm mechanism. At first, in order to discover the model essence, we discard the binary constraint $\tilde{B} \in \{-1, 1\}^{L \times m}$. Then we get the following optimization problem:

$$\begin{aligned} &\min_{\tilde{B}, A, P} \; tr(\tilde{B} Q \tilde{B}^T - 2AP^T XZ\tilde{B}^T + P^T(\alpha G + XX^T)P) \\ &\text{s.t. } A^T A = I_s, \end{aligned} \tag{19}$$

where $Q_{ii} = \sum_i Z_{ji}$. Taking the partial deviation with respect to $\tilde{B}$ to be zero, we derive:

$$2\tilde{B}Q - 2AP^T XZ = 0 \Rightarrow \tilde{B} = AP^T XZQ^{-1}. \tag{20}$$

Comparing above equation with $\tilde{B} = sign(AP^T XZ)$, we find that the optimal $\tilde{B}$ is in essence the locally linear combiner of the features projected to the subspace P plus a rotation using orthogonal matrix A.

Suppose $\tilde{B}$ is given. Let us look back to the variable P. The same as the formulation on (13), we also have $P = (\alpha G + XX^T)^{-1} XZ\tilde{B}^T A$. Substituting it back to the optimization problem (19), we have:

$$\max_{A^T A = I_s} tr(A^T \tilde{B} Z^T X^T (\alpha G + XX^T)^{-1} XZ\tilde{B}^T A). \tag{21}$$

From (21) we can find that the optimization problem is in essence a weighted PCA, where $Z^T X^T (\alpha G + XX^T)^{-1} XZ$ is the weight matrix. In other words, the rotation matrix is in essence the eigenvectors of the weighted PCA.

In short, the above analysis shows that the essence of the proposed algorithm is to learn a regular subspace with joint sparsity for feature extraction and selection and then perform a rotation to match the binary codes with the lowest information loss. Therefore, the proposed model integrates the dimensionality reduction, sparse feature selection and the rotation operation with a seamless formulation.

V. Experiments and Results

In this paper, seven state-of-the-art unsupervised hashing methods were selected for comparison on MNIST,[1] CIFAR-10,[2] NUS-WIDE,[3] SUN397 [40] and ImageNet [41] datasets. The data-independent method LSH was selected for testing the performance. Four manifold-based hash learning methods, including SH, AGH, IMH and LGHSR, were used for comparison. Different from the manifold-based methods, PCA-ITQ mainly focuses on global information of datasets. Thus, PCA-ITQ was also compared in the experiments. As the sparse extension of ITQ, Sparse Projections (SP) [31] was added for comparing the performance of using sparse coding matrix. The code of JSH can be downloaded from http://www.scholat.com/laizhihui.

A. Details of Datasets

We selected five large-scale datasets including MNIST, CIFAR-10, NUS-WIDE, SUN397 and ImageNet to evaluate the performance of different methods. The MNIST dataset contains 70,000 handwritten digits images in total. The digits include number 0 to number 9. Each image was reshaped to a 784-dimensional vector. On MNIST dataset, we randomly selected 6900 images from each class as training samples. The rest 1000 images formed the testing set. The CIFAR-10 dataset contains 60,000 object images collected from 10 labeled classes. Each image on this dataset was transformed to GIST feature vector [42]. In this experiment, 5900 images were randomly selected from each class to form the training samples. The remaining 1000 images formed the testing samples. The NUS-WIDE dataset involves over 200,000 images and the labels of each image are related to the ground truth concept. We first picked 21 labels following the rule in [11]. The testing set contained 100 images per label and the rest images related to these 21 labels formed the training set. The SUN397 dataset comprises of 106,953 training images and 1800 testing images from 397 different classes. Each image was resized to 1600-dimensional feature vector. The ImageNet dataset contains over 1.2 million training images extracting from 1000 objectives. In this experiment, we randomly selected 10,000 images from validation images to form the testing set. Similar to [5], we first learned deep features of each image using Convolution Neural Networks (CNN) [43] and then, performed PCA to obtain 1000-dimensional feature vectors for hash learning. We analysis the experimental results in the following sections.

B. Experimental Settings

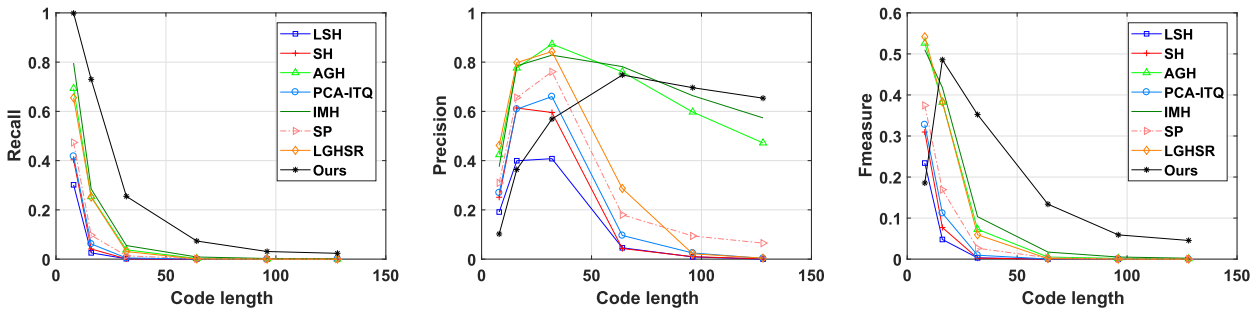
The iteration times and parameters of different algorithms were appropriately set according to the recommendation of the authors and our experimental observations. For the proposed method, we set $\alpha = 10$ and $s = L$. As for the construction of anchor graph, we set $m = 500$, $k = 5$ on MNIST, CIFAR-10 and SUN397 datasets while $m = 500$, $k = 10$ on NUS-WIDE dataset and $m = 1000$, $k = 10$ on ImageNet dataset. For testing the performance with different length of codes, the bits of binary codes were ranged in the set [8, 16, 32, 64, 96, 128].

After obtaining the binary codes of training samples and testing samples, we tested the retrieval performance. The

[1] http://yann.lecun.com/exdb/mnist/

[2] http://www.cs.toronto.edu/Ëœkriz/cifar.html

[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm



(a) (b) (c)

Fig. 2. Results of (a) Recall, (b) Precision and (c) F-measure of different algorithms on MNIST dataset.

TABLE I

BEST RESULTS OF (A)MAP(BITS) AND (B) F-MEASURE(BITS) OF DIFFERENT ALGORITHMS ON FOUR DATASETS (THE TOP 2 ARE HIGHLIGHTED)

	MNIST		CIFAR-10		NUS-WIDE		SUN397	
	MAP	F-measure	MAP	F-measure	MAP	F-measure	MAP	F-measure
LSH	0.3807 (128)	0.2341 (8)	0.1411 (128)	0.1810 (8)	**0.4024 (128)**	0.3073 (8)	0.0659 (128)	0.0256 (8)
SH	0.2661 (16)	0.3097 (8)	0.1266 (16)	0.1815 (8)	0.3580 (128)	**0.5299 (16)**	0.0944 (128)	0.0863 (16)
AGH	0.5194 (8)	**0.5263 (8)**	0.1666 (8)	0.2202 (8)	0.3633 (16)	0.2042 (8)	0.1711 (64)	0.1703 (16)
PCA-ITQ	0.4032 (128)	0.3276 (8)	0.1504 (128)	0.2148 (8)	0.3903 (96)	0.5029 (8)	0.1239 (128)	0.1038 (16)
IMH	0.5554 (16)	0.5101 (8)	**0.1900 (96)**	**0.2576 (16)**	0.3913 (128)	0.5295 (8)	**0.1799 (64)**	0.1808 (32)
SP	0.4676 (128)	0.3746 (8)	0.1581 (128)	0.2203 (8)	0.3940 (96)	0.5138 (8)	0.1526 (128)	0.1360 (16)
LGHSR	**0.5589 (16)**	**0.5417 (8)**	0.1645 (16)	0.2123 (8)	0.3631 (64)	0.2205 (8)	**0.1820 (128)**	**0.1824 (16)**
Ours	**0.5725 (128)**	0.4857 (16)	**0.1913 (96)**	**0.2510 (32)**	**0.3972 (96)**	**0.5315 (8)**	0.1542 (128)	**0.1895 (32)**

TABLE II

RESULTS OF MAP AND F-MEASURE VERSUS THE NUMBER OF ANCHORS (m) AND NEAREST NEIGHBOR POINTS (k) ON MNIST DATASET WITH 32 bits LENGTH

	m	100	200	300	400	500	600	700	800	900	1000	2000
MAP	$k=5$	0.5074	0.5228	0.4717	0.4917	0.4750	0.5534	0.4687	0.4597	0.4827	0.4889	0.4895
	$k=10$	0.3134	0.4335	0.3957	0.5115	0.4511	0.5022	0.5042	0.5525	0.4762	0.5195	0.4607
	$k=20$	0.1791	0.2610	0.3286	0.4076	0.4585	0.4614	0.4785	0.4650	0.4845	0.4834	0.4420
F-measure	$k=5$	0.5164	0.4582	0.4334	0.4197	0.3480	0.3123	0.3907	0.2834	0.3281	0.2711	0.3360
	$k=10$	0.3810	0.4552	0.4550	0.4594	0.4355	0.4207	0.4508	0.5183	0.3414	0.4022	0.3900
	$k=20$	0.1827	0.2665	0.3317	0.4872	0.5191	0.4721	0.5202	0.4296	0.4362	0.5046	0.4557

results of precision, recall and F-measure with Hamming radius 2 and the results of mean average precision (MAP) according to Hamming ranking are presented.

C. MNIST Dataset

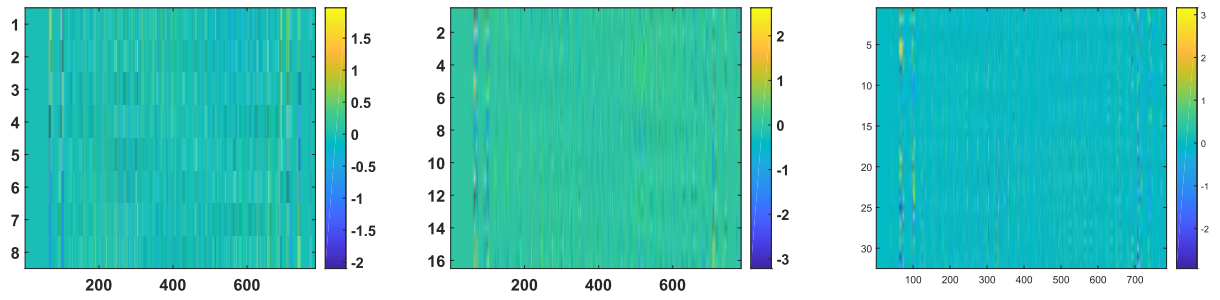
Figure 2 shows the results about Recall, Precision and F-measure of different methods on MNIST dataset. In Figure 2(a) and 2(c), it is obvious that the proposed JSH performs better than other methods when the length of binary codes are longer than 16 bits. This is because the proposed method better preserves the manifold structure by integrating the dimensionality reduction, sparse feature selection and binary regression together. Besides, the best results on MAP and F-measure are listed in Table I. The proposed method performs better than the other methods on MAP at least 2%. According to the F-measure results of AGH, IMH, LGHSR and JSH, we can find that the anchor graph preserve enough manifold structure information on high-dimensional dataset so that the anchor graph based methods can outperform SH.

Table II lists the results of MAP and F-measure versus the number of anchors and nearest neighbor points on MNIST dataset with 32 bits codes. When $k = 10$ and 20, the proposed algorithm tends to obtain better performance as m increases. On the contrary, when $k = 5$, the performance of proposed algorithm gradually degrades as m increases. According to the experimental results, we can find that it is important to control the proportion of m and k. If m is large, the model requires larger k to characterize the local structure of dataset. However, if m is small, a larger k may lead to low accuracy.

In Figure 3(a), 3(b) and 3(c), we show the visualization of the coding matrix with size 8×784, 16×784 and 32×784. It is shown that the elements of some columns are equal to zeros. Since we introduce the joint sparsity, the function of consistent feature selection can be realized in a single projection operation. For example, the first column of coding matrix is zero vector, thus, the first feature of data can be ignored for binarization. Although both SP and JSH obtain sparse matrix for coding, JSH still outperforms SP since JSH considers the manifold structure of datasets and selects the more discriminative features for binarization.

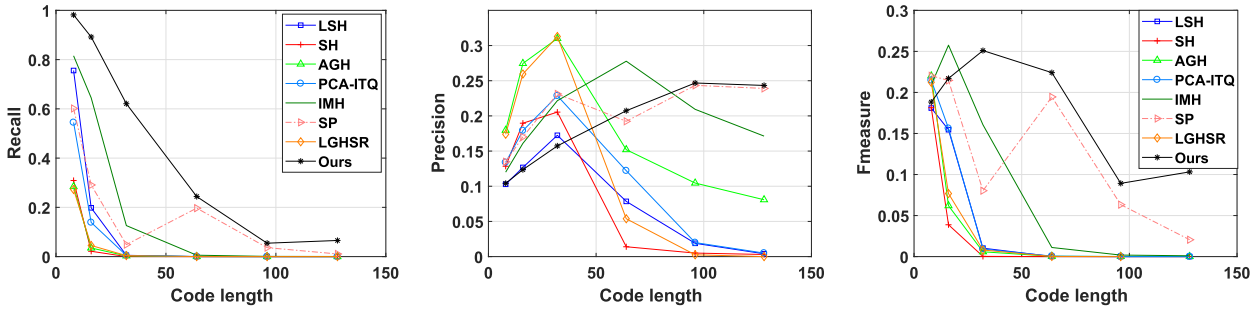
D. CIFAR-10 Dataset

As shown in Figure 4(a), the proposed method obtains the highest recall rate on CIFAR-10 dataset. Besides, as the



(a) (b) (c)

Fig. 3. Visualization of coding matrix with different bits (a) 8 bits, (b) 16 bits and (c) 32 bits on MNIST dataset.



(a) (b) (c)

Fig. 4. Results of (a) Recall, (b) Precision and (c) F-measure of different algorithms on CIFAR-10 dataset.

TABLE III

RESULTS OF MAP AND F-MEASURE VERSUS THE NUMBER OF ANCHORS (m) AND NEAREST NEIGHBOR POINTS (k) ON CIFAR-10 DATASET WITH 32 bits LENGTH

	m	100	200	300	400	500	600	700	800	900	1000	2000
MAP	$k=5$	0.1709	0.1524	0.1690	0.1674	0.1785	0.1728	0.1685	0.1723	0.1562	0.1547	0.1552
	$k=10$	0.1513	0.1498	0.1598	0.1601	0.1617	0.1722	0.1678	0.1603	0.1564	0.1576	0.1618
	$k=20$	0.1285	0.1486	0.1487	0.1572	0.1318	0.1475	0.1465	0.1494	0.1641	0.1508	0.1662
F-measure	$k=5$	0.2534	0.2222	0.2636	0.2555	0.2513	0.1954	0.2490	0.2540	0.2453	0.2546	0.2487
	$k=10$	0.2204	0.2339	0.2460	0.2301	0.2538	0.2578	0.2509	0.2439	0.2468	0.2407	0.2533
	$k=20$	0.2190	0.2255	0.2194	0.2364	0.2264	0.2343	0.2397	0.2377	0.2534	0.2124	0.2576

length of codes increase, the performance of LSH, SH, AGH, PCA-ITQ and LGHSR decrease rapidly. The hamming distances between different samples are close when the lengths of binary codes are short. Thus, most of hashing methods can easily recall the related data points on the dataset. However, as the lengths of codes increase, the hamming distances between data points will greatly increase and it is difficult to retrieve the related data points. That is why the recall rates drop as the code's length increases. Since the proposed method integrates dimensionality reduction, local structure preservation and orthogonal rotation seamlessly, the hamming distance of the same class data points can be reduced. Therefore, the proposed method can perform better than the other methods in long binary codes cases. By comparing the results between PCA-ITQ and SP, JSH and AGH, IMH, we can find that the sparse coding matrix is beneficial to improve the performance of methods. In Table I, the proposed method ranks at least top-2 on MAP and F-measure on CIFAR-10 dataset. Besides, as shown in Figure 4(b) and 4(c), our method can achieve highest Precisions and F-measures for longer codes on all the datasets. Table III lists the results of MAP and F-measure versus the values of m and k on CIFAR-10 dataset. Similar to the conclusions on MNIST dataset, the values of m and k need to be appropriately set for obtaining good performance.

E. NUS-WIDE Dataset

The performance of different methods on NUS-WIDE dataset is displayed in Table I. It is shown that JSH performs better than the other methods on F-measure. In Table IV, we display the training times and testing times of different methods on NUS-WIDE dataset with 16 and 64 bits length. Since JSH jointly learns projection matrix and regression matrix, the training times of JSH are longer than the other methods when $L = 16$. However, compared to the other manifold-based methods, such as AGH, IMH and LGHSR, the testing time of JSH is faster. Based on the experimental

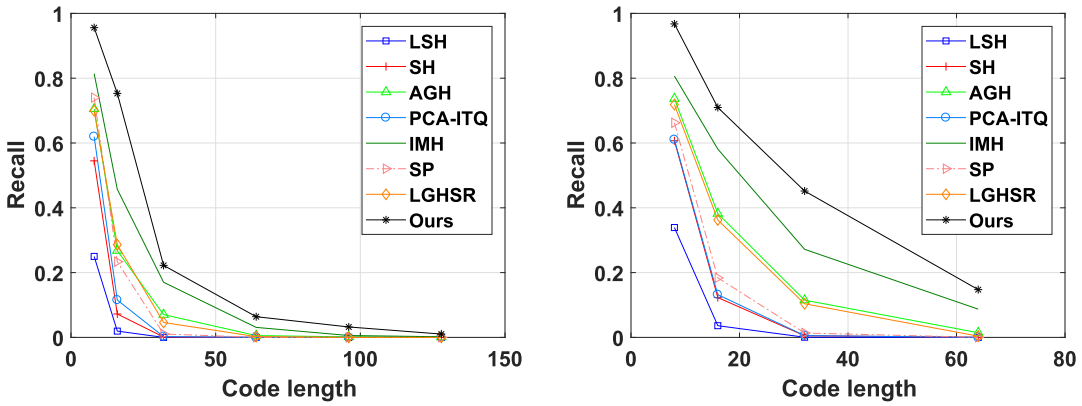


Fig. 5. Recall rates on (a) SUN397 and (b) ImageNet datasets.

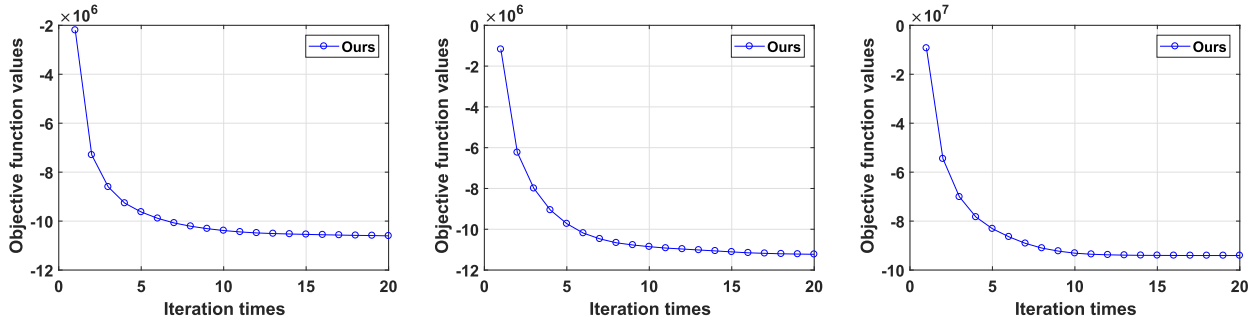


Fig. 6. Convergences of Algorithm 1 on (a) MNIST, (b) CIFAR-10 and (c) NUS-WIDE datasets with 64 bits.

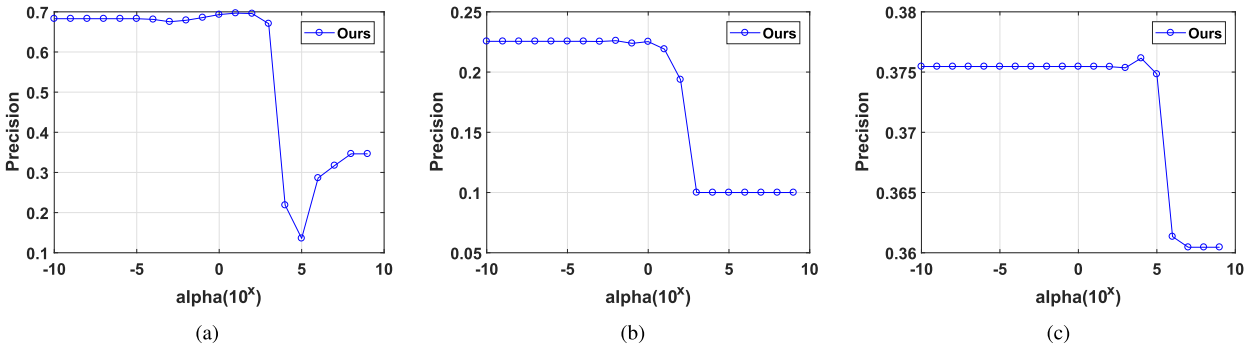


Fig. 7. Precisions on (a) MNIST, (b) CIFAR-10 and (b) NUS-WIDE datasets with different values of parameter α.

results, we can find that the testing time of SP is the fastest since it obtains sparse and linear projections for coding.

F. SUN397 and ImageNet Datasets

We show the recall rates of different algorithms on SUN397 in Figure 5(a). It is obvious that the proposed method obtains much higher recall rates than the other methods. Besides, in Table I, our method obtains the best F-measure rate on SUN397. Figure 5(b) displays the recall rates on ImageNet dataset. It is shown that the global-information based methods SH, PCA-ITQ and SP cannot obtain promising performance on such large-scale dataset. The manifold-based methods AGH, IMH, LGHSR and the proposed method outperform them by considering local structure.

G. Parameter Sensitivity

The variations of objective function values versus iteration times are shown in Figure 6 to illustrate the convergences of Algorithm 1. We find that the algorithm converges within about 20 iteration steps. However, since the objective function values are small enough after 5 iterations, we can reduce the iteration times according to different situations.

For the regularization coefficient α, Figure 7(a), 7(b) and 7(c) display the results of precision versus α on MNIST, CIFAR-10 and NUS-WIDE datasets, respectively. In these experiments, we set $L = 64$ and the parameter α ranged from $[10^{-10}, 10^{-9}, ..., 10^{8}, 10^{9}]$. It is clear that the algorithm performs well when α ranged from 10^{-10} to 10^{0} on three datasets. As α increases, the performance of algorithm

TABLE IV
TRAINING TIMES AND TESTING TIMES (SECONDS) ON NUS-WIDE DATASET WITH 16 AND 64 bits

	16 bits		64 bits	
	Train	Test	Train	Test
LSH	0.2	7.1×10^{-6}	0.3	7.1×10^{-6}
SH	3.1	4.2×10^{-6}	10.4	3.7×10^{-5}
AGH	20.6	3.7×10^{-5}	20.9	3.7×10^{-5}
PCA-ITQ	5.2	4.7×10^{-6}	17.7	7.6×10^{-6}
IMH	20.5	2.9×10^{-5}	20.8	3.7×10^{-5}
SP	15.7	1.4×10^{-6}	22.5	1.4×10^{-6}
LGHSR	25.1	3.5×10^{-5}	40.3	3.5×10^{-5}
Ours	36.1	7.6×10^{-6}	39.9	7.1×10^{-6}

gradually decreases. The possible reason is that the sparsity of the coding matrix is large so that the information loss increases. That is, in the extremely sparse case, the proposed method cannot be used since it might select far less feature to represent the data.

H. Limitations and Potential Improvements

As shown in Figure 2(b) and Figure 4(b), the performance of our method is unpromising in short codes cases. The potential reason is that the proposed method fails to separate different neighbors in such low-dimensional hamming spaces. Therefore, the distances between different neighbors are close which leads to low precisions. To improve the quality of the learned codes, the distances between different neighbors needs to be enlarged. Besides, due to the inverse operation, our method requires $O(d^3)$ to obtain projection matrix P in each iteration. It will be time-consuming for very high-dimensional datasets. In the future, it is desirable to propose a more effective and efficient method to solve these problems.

VI. CONCLUSION

In this paper, we propose a jointly sparse regression model to perform hash learning. By adding $L_{2,1}$-norm regularization term on the model, the proposed JSH can learn the optimal jointly sparse projection matrix for low-dimensional feature extraction and selection. With the orthogonal constraint as a rotation operation between the low-dimensional features and the binary codes, the proposed model can further minimize the information loss and directly obtain the binary solutions. The convergence analysis and computational complexity are also presented. Moreover, theoretical analysis reveals the essence of the proposed model. That is, the proposed model integrates the dimensionality reduction, sparse feature selection and the rotation operation with a seamless formulation. Therefore, the information loss can be minimized in learning the optimal projections and binary codes.

APPENDIX

Proof of Lemma 2: Let $\Psi(\tilde{B}, A, P, G)$ denotes the objective function value of (10). That is

$$\Psi(\tilde{B}, A, P, G) = \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j - AP^T x_i||_2^2 Z_{ij} + \alpha tr(P^T G P)$$

Since the solution derived by partial deviation provides the optimal P in t-th iteration, we have

$$\Psi(\tilde{B}^{t-1}, A^{t-1}, P^t, G^{t-1}) \leq \Psi(\tilde{B}^{t-1}, A^{t-1}, P^{t-1}, G^{t-1}), \tag{22}$$

where $\tilde{B}^t$, A^t, P^t, G^t denote matrix $\tilde{B}$, A, P, G in t-th iteration, respectively. The SVD also give the optimal A to continuously reduce the objective function value in t-th iteration. Therefore

$$\Psi(\tilde{B}^{t-1}, A^t, P^t, G^{t-1}) \leq \Psi(\tilde{B}^{t-1}, A^{t-1}, P^t, G^{t-1}). \tag{23}$$

Then, we obatin the optimal $\tilde{B}^t$ via solving (16) which further reduces the objective function value. That is

$$\Psi(\tilde{B}^t, A^t, P^t, G^{t-1}) \leq \Psi(\tilde{B}^{t-1}, A^t, P^t, G^{t-1}). \tag{24}$$

Therefore, we have

$$\begin{aligned}
&\sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j^t - A^t(P^t)^T x_i||_2^2 Z_{ij} + \alpha tr((P^t)^T G^{t-1} P^t) \\
&\quad \leq \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j^{t-1} - A^{t-1}(P^{t-1})^T x_i||_2^2 Z_{ij} \\
&\quad\quad + \alpha tr((P^{t-1})^T G^{t-1} P^{t-1}).
\end{aligned} \tag{25}$$

According to the definition of G, we rewrite the regularization term of (25) as

$$\alpha \sum_{i=1}^{d} \frac{||p_t^i||_2^2}{2||p_{t-1}^i||_2} \leq \alpha \sum_{i=1}^{d} \frac{||p_{t-1}^i||_2^2}{2||p_{t-1}^i||_2}, \tag{26}$$

where p_t^i is the i-th row of matrix P^t. Based on Lemma 1, we have following inequality

$$\alpha \sum_{i=1}^{d} (||p_t^i||_2 - \frac{||p_t^i||_2^2}{2||p_{t-1}^i||_2}) \leq \alpha \sum_{i=1}^{d} (||p_{t-1}^i||_2 - \frac{||p_{t-1}^i||_2^2}{2||p_{t-1}^i||_2}). \tag{27}$$

(27) indicates that

$$\alpha ||P^t||_{2,1} \leq \alpha ||P^{t-1}||_{2,1}. \tag{28}$$

Combining (28) and (25), we obtain

$$\begin{aligned}
&\sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j^t - A^t(P^t)^T x_i||_2^2 Z_{ij} + \alpha ||P^t||_{2,1} \\
&\quad \leq \sum_{j=1}^{m}\sum_{i=1}^{n} ||\tilde{b}_j^{t-1} - A^{t-1}(P^{t-1})^T x_i||_2^2 Z_{ij} + \alpha ||P^{t-1}||_{2,1}
\end{aligned} \tag{29}$$

which completes the proof. ■

Proof of Theorem 2: A solution of the problem (8) needs to satisfy two conditions as follows:

First, the point is a feasible point. Apparently, the condition is fulfilled.

Second, the point is a minimum point of the objective function. Suppose $\tilde{B}^\infty$, A^∞, P^∞ is a limit point of $\tilde{B}^t$, A^t, P^t. The two constraints in the problem (8) are surely satisfied since

the constraints are fulfilled in each iteration. For the variable P, we have the following relation

$$(\alpha G^{t-1} + XX^T)P^t = XZ(\tilde{B}^{t-1})^T A^{t-1} \tag{30}$$

Take the limits of both sides of the equation, we obtain

$$(\alpha G^{\infty} + XX^T)P^{\infty} = XZ(\tilde{B}^{\infty})^T A^{\infty} \tag{31}$$

which is the optimal condition concerning the variable P. Similarly, the optimality of the variable A can be verified. Finally, since both the feasibility and the optimality is fulfilled, the limit point $\tilde{B}^{\infty}, A^{\infty}, P^{\infty}$ are solutions of the problem (8). ■

REFERENCES

[1] S. Korman and S. Avidan, "Coherency sensitive hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1099–1112, Jun. 2016.

[2] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.

[3] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[4] X. Shi, F. Xing, K. Xu, M. Sapkota, and L. Yang, "Asymmetric discrete graph hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2541–2547.

[5] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 37–45.

[6] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Fast supervised discrete hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 490–496, Feb. 2018.

[7] Y. Xu, F. Shen, X. Xu, L. Gao, Y. Wang, and X. Tan, "Large-scale image retrieval with supervised sparse hashing," *Neurocomputing*, vol. 229, pp. 45–53, Mar. 2017.

[8] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Supervised discrete hashing with relaxation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 608–617, Mar. 2018.

[9] 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, Dec. 2013.

[10] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1753–1760.

[11] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[12] F. Shen, C. Shen, Q. Shi, A. V. D. Hengel, Z. Tang, and H. T. Shen, "Hashing on nonlinear manifolds," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1839–1851, Jun. 2015.

[13] F. Shen, C. Shen, Q. Shi, A. van den Hengel, and Z. Tang, "Inductive hashing on manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1562–1569.

[14] X. Zhu, L. Zhang, and Z. Huang, "A sparse embedding and least variance encoding approach to hashing," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3737–3750, Sep. 2014.

[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[16] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[17] Y. Lu, Z. Lai, X. Li, W. K. Wong, C. Yuan, and D. Zhang, "Low-rank 2-D neighborhood preserving projection for enhanced robust image representation," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2018.2815559.

[18] Y. Lu, C. Yuan, X. Li, Z. Lai, D. Zhang, and L. Shen, "Structurally incoherent low-rank 2DLPP for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2018.2849757.

[19] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2018.2789887.

[20] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3419–3427.

[21] F. Nie, W. Zhu, and X. Li, "Unsupervised large graph embedding," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2422–2428.

[22] X. Li, D. Hu, and F. Nie, "Large graph hashing with spectral rotation," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2203–2209.

[23] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, Jun. 2006.

[24] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear sparse principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1942–1950, Oct. 2014.

[25] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[26] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2,1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 22, 2011, no. 1, pp. 1589–1594.

[27] J. Wen, Z. Lai, Y. Zhan, and J. Cui, "The $L_{2,1}$-norm-based unsupervised optimal feature selection with applications to action recognition," *Pattern Recognit.*, vol. 60, pp. 515–530, Dec. 2016.

[28] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by nuclear/$L_{2,1}$-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, Apr. 2017.

[29] W. K. Wong, Z. Lai, J. Wen, X. Fang, and Y. Lu, "Low-rank embedding for robust image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2905–2917, Jun. 2017.

[30] S. Yi, Z. Lai, Z. He, Y.-M. Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognit.*, vol. 61, pp. 524–536, Jan. 2017.

[31] Y. Xia, K. He, P. Kohli, and J. Sun, "Sparse projections for high-dimensional binary codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3332–3339.

[32] X. Zhu, Z. Huang, H. Cheng, J. Cui, and H. T. Shen, "Sparse hashing for fast multimedia search," *ACM Trans. Inf. Syst.*, vol. 31, no. 2, p. 9, 2013.

[33] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[34] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 22, no. 1, 2011, pp. 1294–1299.

[35] Z. Chen, J. Lu, J. Feng, and J. Zhou, "Nonlinear sparse hashing," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1996–2009, Sep. 2017.

[36] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[37] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[38] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[39] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*. Hoboken, NJ, USA: Wiley, 2011.

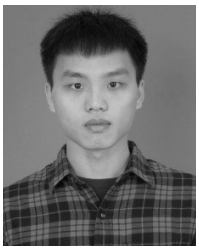
[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

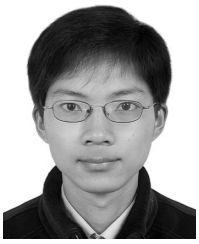
[42] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

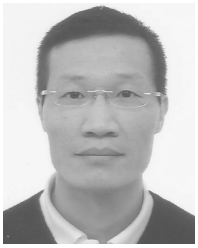
Zhihui Lai received the B.S. degree in mathematics from South China Normal University in 2002, the M.S. degree from Jinan University in 2007, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, China, in 2011. He was a Research Associate, a Post-doctoral Fellow, and a Research Fellow with The Hong Kong Polytechnic University. He has published over 60 scientific articles. His research interests include face recognition, image processing, and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the fields of intelligent robot research. He is currently an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.



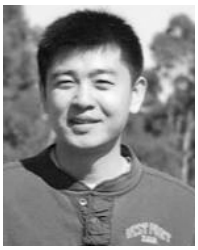
Yudong Chen is currently pursuing the M.S. degree in computer science and technology at Shenzhen University, Shenzhen, China. He is currently with the College of Computer Science and Software Engineering, Shenzhen University. His research interests include image processing and pattern recognition.



Jian Wu received the B.S. degree in mathematics from Liaoning Normal University, Dalian, China, in 2010, and the M.S. degree in mathematics from Gannan Normal University, Ganzhou, China, in 2014. He is currently pursuing the Ph.D. degree in computer science with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His research interests focus on medical biometrics, pattern recognition, and image processing.



Wai Keung Wong received the Ph.D. degree from The Hong Kong Polytechnic University, Kowloon, Hong Kong. He is currently a Professor with The Hong Kong Polytechnic University. He has authored or co-authored over 100 papers in refereed journals and conferences, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-Part B: Cybernetics, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-Part C: Applications and Reviews, *Pattern Recognition*, *CHAOS*, the *European Journal of Operational Research*, *Neural Networks*, *Applied Soft Computing*, *Information Science*, and *Decision Support Systems*, among others. His current research interests include pattern recognition and feature extraction.



Fumin Shen received the bachelor's degree from Shandong University in 2007 and the Ph.D. degree from the Nanjing University of Science and Technology, China, in 2014. He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His major research interests include computer vision and machine learning. He was a recipient of the Best Paper Award Honorable Mention at ACM SIGIR 2016, ACM SIGIR 2017 and the World's FIRST 10K Best Paper Award—Platinum Award at the IEEE ICME 2017.