IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 28, 2020, accepted December 23, 2020, date of publication January 18, 2021, date of current version February 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052605

Cross-Modal Hashing by l_p-Norm Multiple Subgraph Combination

DONGXIAO REN[1,2], JUNWEI HUANG[2], ZHONGHUA WANG[3], AND FANG LU[2]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China
[3]Beijing GuoDianTong Network Technology Company Ltd., Beijing 100083, China

Corresponding author: Fang Lu (mengmeng2925@sohu.com)

This work was supported in part by the National Natural Science Foundation of China (No. 11801511).

ABSTRACT With the explosion of multi-modal Web data, effective and efficient techniques are in urgent need for cross-modal data retrieval with relevant semantics. Among all the possible solutions, the hashing techniques provide compact and measurable binary representation, thus gain much attention in related research domain. To better deal with diversified real world data, we propose MSC, a novel cross-modal hashing approach based on the generalized l_p-norm Multiple Subgraph Combination. Specifically, by jointly considering the content similarity, the correspondence and other weak correlation among cross-modal documents, we build the intra-modal similarity with multiple affinity subgraphs, and encode the inter-modal correlation with a bipartite subgraph. Then these subgraphs are combined into one multi-modal similarity graph for all the data from heterogeneous modalities, where the weights of multiple intra-modal visual similarity subgraphs are regularized by l_p-norm penalty. The optimal hash codes and the combination coefficients are learned simultaneously by efficient alternating optimization. The hash functions for different modalities are learned separately by utilizing nonlinear classification models, encoding the complicated semantic relations among cross-modal data. Experiments on challenging real world datasets demonstrate the advantage of our method over existing approaches.

INDEX TERMS Cross-modal hashing, feature combination, information fusion.

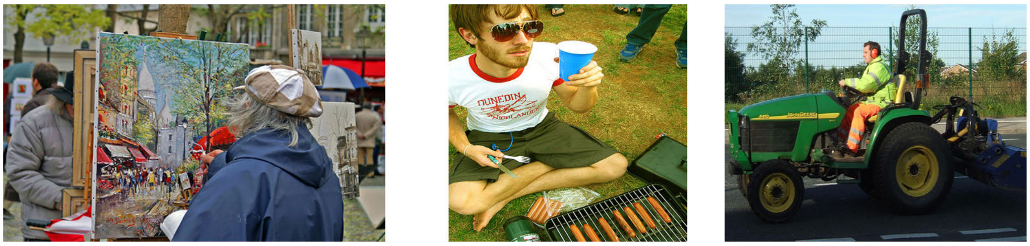
I. INTRODUCTION

Due to the development of Web and multimedia technology, the amount of Web multi-modal data is growing with an astonishing speed. Meanwhile, the diversified Web content is delivered by multiple information carriers. For example, the concept "*European Football Championship*" can be described by text, photographs and videos contributed by professional journalists or amateur users. For better understanding of interesting concepts or events, the user would search the images or videos by textual queries, or the textual descriptions by image queries. This new application demand is known as the cross-modal retrieval [1]–[3], which is very different from traditional single-modal image retrieval [4]–[7], where the queries and database documents are from the same modality. For cross-modal retrieval, how to effectively and efficiently retrieve multi-modal data with rich content and context becomes a very interesting yet challenging problem.

As great endeavors have been dedicated, a widely accepted paradigm for large scale retrieval is the neighborhood search. Instead of linear scan search, Locality Sensitive Hashing (LSH) [8] is an approximated neighborhood search method on high dimensional data where the collision probability of the hash codes is related to the similarity between the hashed data. Learning-based hash models, to name a few, Spectral Hashing [9], [10], Semantic Hashing [11] and task specific hashing [12] obtain the retrieved results with more semantic consistency. To exploit nonlinear similarity measure, Kulis *et al.* [13] construct LSH on some given kernel instead of original space (KLSH). Liu *et al.* [14] propose to learn the hash function based on the kernel and side information. They can only construct hashing system on single modality and are not applicable when the queries and database are from different modalities.

This paper studies the hash learning method for cross-modal retrieval, which corresponds to mapping heterogeneous modalities into a unified Hamming space. However, in real cross-modal retrieval applications, images and texts

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.



A person with long gray hair has a beret with beige and white wearing a blue raincoat is painting a marketplace scenery surrounded by other artists and paintings.

A guy in shorts and a white t-shirt sits on the ground in front of a grill with hotdogs on it.

A man in a neon green and orange uniform is driving on a green tractor.

FIGURE 1. Examples of real-world correlated image-text data. The words and phrases in red, green, orange, blue and purple represent the visual content descriptions from color, shape, object, action and person aspects.

are correlated from different aspects and from different levels. First, take an image as an example, the correlated text may describe the visual content from multiple and complementary aspects, *e.g.*, color, shape, object, scene and action. Some examples are shown in Fig. 1, where we mark the textual description from different aspects with distinguished colors. The visual content can be described by different visual features from complementary aspects. Accordingly, the visual-linguistic relation on feature level is also encoded by the correlation between different visual features and textual feature. Therefore, given that the textual features can be sufficiently represented by single feature extraction pipeline (*e.g.*, bag-of-word model, text CNN [15], LSTM [16], *etc.*), it is necessary to model the cross-modal correlation by considering multiple visual features for learning good cross-modal hash codes and functions.

Second, images and texts are correlated from different levels. For example, an image in a web page is strongly correlated with its surrounding textual description (as shown in in Fig. 1), while it can also be weakly related (or partially related) to those texts that are located in the web-pages that are linked to this web page, as discussed in [17].

To perform cross-modal hash learning, a good model should meet the following requirements. First, the hash codes of the semantically similar (dissimilar) data within the same modality should be similar (dissimilar), where the intra-modal relation can be described by local affinity [9], [10] or side information [18], [19]. Second, the hash codes of the correlated (uncorrelated) data from different modalities should be similar (dissimilar) [18]–[20].

In this paper, we study the cross-modal hashing method by addressing the following issues. First, beyond existing mainstream approaches which mainly learn to fit the given semantic relation among cross-modal data objects, we categorize the cross-modal correlation into strong correlation and weak correlation for hashing. The strong correlation corresponds to those image-text pairs discussing exactly the same topic. The weak correlation indicates that the topics of the two cross-modal documents are relevant, where the relevance can be reflected by their structure relation, *e.g.*, the hyperlinks, or semantic relevance of the two category words. By considering weak correlation among cross-modal data, we can utilize the correlation information more comprehensively to learn the cross-modal hash codes.

Second, considering that the visual information can be described by multiple features, combining their descriptive power is a potential way to obtain better intra-modal relation in visual modality, where the feature combination has been well studied in related research [21]–[25]. However, existing approaches consider feature combination for hash learning [23]–[25] using simple fusion schemes, *i.e.*, average weight or sparsity regularization. By jointly considering feature combination and diversified correlation among cross-modal documents, we construct a cross-modal hashing method based on l_p-norm multiple sub-graph combination (MSC), which projects each document from multiple modalities into a B-dimensional Hamming space. To regularize the combination of multiple features, we add the l_p-norm constraint on the weight vector corresponding to different features types, which encourages higher order nonlinearity on the feature combination, thus it leads to better model learning results and more semantically consistent hash codes.

The advantages of our method can be summarized as follows:

- We propose MSC, a cross-modal hash code learning with a generalized l_p-norm subgraph combination, which deals with content and correspondence diversity. By effectively combining information from multiple features and many-to-many correspondence information, our model is more capable of processing real world cross-modal data.
- We learn the hash function with nonlinear binary models, which better captures the complicated semantic correlation among different modalities. Compared with existing approaches, the semantic consistency of the learned hash codes is significantly enhanced. It can also be applied to real world unseen data to generate high quality hash codes.

- Experiments performed on challenging cross-modal datasets demonstrate the advantages of our method over existing approaches.

Section 2 discusses related work. Section 3 introduces our hashing approach. Section 4 presents the experiment and discussion. Section 5 concludes the paper.

II. RELATED WORK

Towards learning compact representation and hash codes for efficient cross-modal retrieval, relevant studies can be roughly categorized into "subspace learning" and "probabilistic models".

Subspace learning finds the subspace that maximizes the correlation of two modalities. Canonical Correlation Analysis (CCA) [26] and its variants [27] provide direct solutions, while they ignore the intra-modal similarity information. Based on the subspace learned by CCA, cross-modal topic classifiers [1] map heterogeneous modalities into a unified semantic space. Bronstein *et al.* [20] propose a boosting based hashing method where the "weak coders" and their weights are jointly learned for weighted cross-modal Hamming distance calculation. Gong *et al.* [28] propose an iterative quantization method which finds a rotation of zero-centered data to minimize the quantization error of mapping data to vertices of a zero-centered binary hypercube. Masci *et al.* extend [20] by multi-layered neuro-networks [18] which are trained with both the intra-modal similarity and inter-modal correlation. Wang *et al.* [29] propose a struture preserving image-text embdding approach which learns a pair of multi-layered neuro-networks by employing triplet loss function on hard negatives. A tailored feed-forward neuro-network approach [30] obtains sparse hash codes. Wu *et al.* [31] propose a quantized correlation hashing for fast cross-modal retrieval, which jointly optimize the quantization process and correlation learning process.

Graph-based methods [19], [32] encode the intra-modal similarity and inter-modal co-occurrence into a unified graph representation. Existing approaches require cross-modal data to be strictly aligned and organized into one-to-one correspondence. However, the relation among cross-modal documents are more complicated. For example, the correspondence of the real cross-modal data is not completely provided [17], [33]. Moreover, there may be other structure relation among cross-modal documents that may provide relevant description for a given topic, *i.e.*, the hyperlink between two documents indicates that one provides explanation to a certain concept of the topic discussed by the other. Similar to our study, Xie *et al.* [34] propose an unsupervised multi-graph hashing methods for large-scale multimedia retrieval, which assigns different weights to different modalities. However, it cannot deal with multiple features inside a modality. Zheng *et al.* [35] propose a hetero-manifold regularisation for cross-modal hashing, which integrates multiple sub-manifolds defined by homogeneous data with the help of cross-modal supervision information for hash learning.

As another possible solution, probabilistic models are constructed to describe how the multi-modal documents are correlated in a probabilistic way on either the feature level [36] or the latent topic level [37]–[41]. Correspondence LDA (Corr-LDA) [37] captures the topic-level relations between images and semantic annotations. Xiao *et al.* [38] combine LDA and Corr-LDA to link images and sounds via words. The model in [39] can be seen as Markov random field over LDA topic models which does not require the one-to-one data correspondence in [37]. Zhen *et al.* [40] develop a latent binary embedding approach, which learns the latent topics and the hash codes based on the observed intra-modal and inter-modal similarities. Chen *et al.* [41] propose a large margin multi-view latent subspace learning method. Xie *et al.* [42] propose a self-supervised cross-modal hashing methods based on hierarchical topic models. Although being successful in modeling the topic level inter-relation of multi-modal data, they are not flexible in processing real world cross-modal data with heterogeneous intra-modal and inter-modal relations.

Recently, deep learning has been widely used in study on hashing techniques. For example, Jin *et al.* [43] propose to learn ordinal representations to generate ranking-based hash codes by leveraging the ranking structure of feature space from both local and global views. Great endeavors have been devoted to deep learning method for image-sentence retrieval. Cao *et al.* [44] propose a deep visual-semantic hashing for cross-modal retrieval which consists of a visual-semantic fusion network to learn the joint embedding and two modality-specific networks for learning visual and textual representations. Jiang *et al.* [45] propose a deep cross-modal Hashing method which tries to use deep neural network to fit the intra-modal and inter-modal relation matrix. Yang *et al.* [46] train the deep cross-modal hashing model by fitting the pairwise relations. Deng *et al.* [47] propose to train the deep cross-modal hashing function by minimizing the triplet loss.

Similar as the cross-modal representation learning, Shu *et al.* [48] propose a deep transfer network which encourages the knowledge sharing and transferring between the network layers of source domain and target domain, which actually pursues the distribution alignment between different data domains. This idea is later developed into a more gneralized framework [49] which encourages more flexible distribution alignment between domains.

The above-mentioned cross-domain or cross-modal deep hashing networks can also be employed in our study. However, since we study the problem of feature combination via graph, this is beyond the scope of this paper. Besides, deep learning model usually involves intensive computational cost, which may bring about low efficiency in hash code learning and construction.

III. APPROACH

Given N_x data from $\mathbf{T}_1$ modality and N_y data from $\mathbf{T}_2$ modality, we denote them with $\mathbf{X}$ and $\mathbf{Y}$, respectively. Note that

N_x and N_y do not have to be equal in this paper. $X = \{X_1, \ldots, X_M\}$ represents the data with multiple feature representations of $\mathbf{T}_1$ modality, where X_m denotes the m-th feature. The aim is to learn a set of B dimensional binary codes for each data from both $\mathbf{T}_1$ and $\mathbf{T}_2$, denoted by $\overline{C}_x$ and $\overline{C}_y$, respectively. We denote the real value relaxation of $\overline{C}_x$ and $\overline{C}_y$ as C_x and C_y, respectively, and $C = [C_x; C_y] \in \mathbb{R}^{(N_x+N_y)\times B}$ is the matrix to be learned.

A. THE OVERALL FRAMEWORK

MSC provides a robust information fusion strategy than the existing hashing models, and are mainly includes the following key steps:

Step 1 (Multiple Subgraph Construction): By considering the content similarity in different feature channels, the visual similarities w.r.t. multiple features are represented with multiple visual subgraphs. The textual similarity is constructed by considering both the content similarity and structure relations among textual documents. These intra-modal subgraphs can be easily encoded with the label information, which makes them more semantically consistent. The inter-modal correlation are encoded with an asymmetric bipartite subgraph, where both the strong correlation (correspondence) and weak correlation (hyperlink and within-Webpage co-occurrence) are employed for cross-modal relation modeling. The inter-modal modeling method better deals with the diversified correlation information existing in Web cross-modal data, and allows certain level of correspondence information missing.

Step 2 (l_p-Norm Hash Code Learning): When the intra-modal subgraphs and inter-modal subgraph have been constructed, we combine them into a unified multi-modal similarity graph with a set of weight coefficients corresponding to different visual feature channels. Then we establish the cross-modal hash code learning with a generalized l_p-norm ($p \geq 1$) combination of multiple subgraphs. With different settings of p, our model automatically identifies different relative importance of the visual features. We design an efficient alternating optimization process to iteratively learn the hash codes and the weight coefficients. Compared to the existing approaches, our method better deals with the topic divergence among real world cross-modal data by incorporating complementary visual descriptions and complicated cross-modal correlation. Specifically, it is tolerant with correspondence missing by propagating the Hamming embedding along the neighborhood data.

Step 3 (Hash Function Learning): Based on the learned hash codes, we design a cross-modal self-taught hash function learning procedure. Given the cross-modal training data, the hash code learning is conducted on the cross-modal similarity graph. After that, two sets of linear / nonlinear hash functions are learned separately on different modalities by treating the learned hash codes as the binary labels as [10]. For any query, we obtain the hash codes by using the hash functions of the corresponding modality, and then feed them into the database for cross-modal data retrieval. Experiments performed on challenging cross-modal datasets demonstrate the advantage of our method over the existing approaches.

B. CORRELATION DEFINITION

Before introducing our method, let us first explain two types of correlations among cross-modal documents.

1) STRONG CORRELATION

The strong correlation indicates that the cross-modal documents are discussing exactly the same topic. If an image is co-occurred with a textual paragraph on a Webpage, and they are adjacent with each other, it is highly possible that they are describing the same topics. The one-to-one correspondence discussed in traditional cross-modal learning [18], [20], [32], [40] can be seen as the strong correlation.

2) WEAK CORRELATION

The weak correlation indicates that the topics of the cross-modal documents are relevant. For example, if an image and a textual document are located in the same Webpage, they may describe relevant topics even their positions are not close to each other. If a document is linked to other documents, they are likely to discuss relevant topics or concepts. Such structure relations (co-occurrence, hyperlink, etc.) reflect that two cross-modal documents are semantically related to a certain extent.

C. INTRA-MODAL SIMILARITY CONSTRUCTION

Given data with high-dimensional representation on single modality, a well studied paradigm for intra-modal similarity modeling is using the *graph Laplacian* [9], [10]. When the data is represented with multiple features, a straight way to model the intra-modal similarity is constructing a unified affinity graph on the concatenated feature using RBF or heat kernel. However, such a strategy suffers from several limitations. First, nearest neighbor search on extremely high dimensional space tends to be sensitive to noise. Second, different physical structures exist on different feature channels, making their dimensions incomparable. Moreover, relative importance of different features is ignored, leading to inappropriate estimation of the true affinity structure. To overcome these drawbacks, we construct M subgraphs on each feature channel using the domain specific similarity calculation, *i.e.*, RBF, χ^2 and histogram intersection kernels, and combine them to represent the visual similarity among images.

We denote the pairwise similarity matrix of $\mathbf{X}$ calculated with the m-th feature as W_m^x which satisfies:

$$W_m^x(i,j) = \begin{cases} K_m^x(i,j), & \textit{if } i \leftrightarrow j \\ 0, & \textit{else}, \end{cases} \tag{1}$$

where $i \leftrightarrow j$ denotes that the i-th and j-th data are mutually nearest neighbors, and $K_m^x(i,j)$ denotes the pairwise visual similarity (kernel) on m-th feature channel. Such a definition guarantees that $W_m, m = 1, ..M$ is positive semi-definite.

In many cases, data can be assigned with certain degree of side information. Intuitively, the hash codes of similar and dissimilar labeled data is required to be similar and dissimilar, respectively, which tends to enhance semantic consistency of the learned hash codes. To encode the side information, especially the dissimilar information, we adopt the dissimilarity graph construction method in [50], and the intra-modal similarity of m-th feature is given as:

$$W_m^x(i,j) = \begin{cases} I(\delta_i^x, \delta_j^x) \cdot K_m^x(i,j), & \textit{if } i \leftrightarrow j \\ 0, & \textit{else}, \end{cases} \quad (2)$$

where δ_i^x denotes the label information of the i-th data in T_1 modality. $I(\cdot,\cdot)$ denotes the indicator function, where $I(\delta_i^x, \delta_j^x) = 1$ when δ_i^x is identical with δ_j^x, otherwise $I(\delta_i^x, \delta_j^x) = -1$. Such a definition enforces that the hash codes of data from different categories have opposite signs as $c_i = -c_j$. Note that if there is no dissimilarity information, W_m^x is identical to the original similarity matrix. According to [50], W_m^x is also positive semi-definite. The overall neighborhood similarity between data i and j within visual modality can be represented by $W^x(\alpha)$, which is a weighted combination of M similarity subgraphs as:

$$W^x(i,j;\alpha) = \begin{cases} \sum_{m=1}^{M} \alpha_m W_m^x(i,j), & \textit{if } i \leftrightarrow j \\ 0, & \textit{else}, \end{cases} \quad (3)$$

where $\alpha_m \geq 0, \forall m$ denote the non-negative weight parameters for all the feature channels. The definition is similar with the kernel combination in Multiple Kernel Learning [22], where the weights need to be learned towards the optimality of some objective function.

Similar as the visual modality, we construct the intra-modal similarity matrix W^y for the textual modality $\mathbf{T}_2$ using the domain specific representation and similarity, *e.g.*, TF-IDF and cosine similarity. Moreover, as there are weak correlation among textual documents (they appear on the same Webpage, or there is hyperlink between i-th and j-th document from $\mathbf{T}_2$), the definition of the intra-modal relation of textual modality can be represented as:

$$W^y(i,j) = \begin{cases} I(\delta_i^y, \delta_j^y) \cdot (K^y(i,j) + \tau), & \textit{if } i \leftrightarrow j \\ 0, & \textit{else}, \end{cases} \quad (4)$$

where K^y indicates the content similarity of two textual documents, and $\tau \in (0, 1]$ denotes a predefined gain that measures the impact of the weak correlation. The value of τ is dataset dependent.

Note that we do not consider the weak correlation in intra-modal visual similarity modeling, since the weak correlation is more subtle among images. For example, on the Wiki page "*Zurich*", the visual contents cover many aspects of the city, include art, satellite photo, street views, sports event, etc. Identifying the weak correlation requires the guidance of certain knowledge base or extra image understanding process, otherwise, it will be misleading.

D. INTER-MODAL CORRELATION MODELING

The inter-modal correlation is crucial to describe the semantic relation of cross-modal documents. We encode the inter-modal correlation information with a correlation matrix $S^{xy} \in \mathbb{R}^{N_x \times N_y}$. $S^{xy}(i,j) = 1$ if the two documents from heterogeneous modalities are co-occurred or they are adjacent on the same Webpage, they describe the same topic and are semantically related; otherwise, $S^{xy}(i,j) = 0$. Note that there can be multiple non-zero elements, or there can be all zeroes in one row or one column. With such a definition, the commonly existing one-to-many or many-to-many correlation among real world cross-modal documents can be described. The requirement in previous study [1], [19], [26], [40], *i.e.*, cross-modal documents should be organized into strict one-to-one correspondence pairs, can be treated as a special case of S^{xy}.

We further consider the weak correlation among cross-modal documents. Specifically, on some Web cross-modal dataset with complicated structure information (*e.g.*, the WIKI-CMR dataset [33]), if there is a hyperlink between i-th document from $\mathbf{T}_1$ and j-th document from $\mathbf{T}_2$, or they occur in the same Webpage but their positions are far away, there is weak correlation between the two document. In this case, $S^{xy}(i,j) = \tau$, where τ is a predefined gain which has been explained in Section 2.1.

E. CROSS-MODAL HASH CODE LEARNING

Based on the intra-modal similarity matrix ($W^x(\alpha)$ and W^y) and the inter-modal correlation matrix S^{xy}, we define a symmetric multi-modal similarity graph as:

$$\Theta(\alpha) = \begin{bmatrix} W^x(\alpha) & S^{xy} \\ (S^{xy})^\top & W^y \end{bmatrix}. \quad (5)$$

Based on $\Theta(\alpha)$, we learn the cross-modal hash code matrix C for all the data with the orthogonality constraints, where the weights of different visual subgraphs are regularized by squared l_p-norm penalty [22]:

$$\begin{aligned} &\min_{C,\alpha} tr\left(C^\top \left(I - D^{-\frac{1}{2}} \Theta_\alpha D^{-\frac{1}{2}}\right) C\right), \\ &s.t.\ ||\alpha||_p^2 = s_0, \quad C^\top C = I, \quad \alpha_m \geq 0, \end{aligned} \quad (6)$$

where $p \in \mathbb{R}^+$ and $p \geq 1$. D denotes the diagonal matrix where $D(i,i) = \sum_j |\Theta_\alpha(i,j)|$. The definition is slightly different from traditional approaches because we model the label information by performing Hadamard product between the label similarity (dissimilarity) indicator matrix and the pairwise similarity matrix. Note that $\left(I - D^{-\frac{1}{2}} \Theta_\alpha D^{-\frac{1}{2}}\right)$ is the normalized *graph Laplacian*. Since $C^\top C = I$, the term $tr\left(C^\top C\right)$ in the objective function can be omitted. Moreover, the orthogonality constraint makes the problem hard to solve. By relaxing this constraint and relaxing $||\alpha||_p^2 = s_0$ into $||\alpha||_p^2 \leq s_0$, we obtain the following equivalent objective function:

$$\begin{aligned} &\min_{C,\alpha} -tr\left(C^\top D^{-\frac{1}{2}} \Theta_\alpha D^{-\frac{1}{2}} C\right) + \rho ||C^\top C - I||_F^2 + \lambda ||\alpha||_p^2, \\ &s.t.\ \alpha_m \geq 0, \end{aligned} \quad (7)$$

where ρ and λ denote positive coefficients. The new objective function penalizes the large values of α by imposing l_p-norm penalty, where the regularization level is controlled by λ (*i.e.*, $\frac{1}{s_0}$). Equation (7) has certain tolerance to nonorthogonality, where the tolerance is controlled by ρ.

In fact, the l_p-norm can be considered as adding some prior information on the kernel weight. If the proposed method ignores the lp-norm, the method will degrade to the traditional multiple kernel learning setting, *i.e.*, the mechanism like the canonical MKL in [51]. Using different values of p equals to applying different norms on the weight vectors, resulting in different shapes of the loss contour. The l_p-norm penelty brings about significant influence on the model structure.

To efficiently learn the hash codes and the weight parameters, we develop the following alternating optimization process in which C and α are optimized iteratively until a local optimal solution is achieved.

Step 1 (Fixing α, Optimize C:) When α is fixed, the equivalent subproblem in Eq.(7) is:

$$J(C) = -tr\left(C^\top D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}C\right) + \rho||C^\top C - I||_F^2. \quad (8)$$

By calculating the derivative with respect to C, we have:

$$\frac{\partial J(C)}{\partial C} = 0 \Rightarrow CC^\top C = \left(I + \frac{1}{\rho}D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}\right)C. \quad (9)$$

According to [52], the matrix $I + \frac{1}{\rho}D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}$ is positive definite if $\rho > \max\left(0, -\overline{\lambda}_{\min}\right)$, where $\overline{\lambda}_{\min}$ is the smallest eigenvalue of $D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}$. If the positive definiteness is satisfied, the solution is $C = LU_B$ according to [52], where L denotes the Cholesky decomposition of $I + \frac{1}{\rho}D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}$ and U_B denotes the top B eigenvectors of $D^{-\frac{1}{2}}\Theta_\alpha D^{-\frac{1}{2}}$. If the positive definiteness is not always guaranteed, a small diagonal matrix ϵI to Θ_α can be added to avoid the ill-posed solution.

Step 2 (Fixing C, Optimize α:) When C is fixed, we obtain the following equivalent problem in Eq.(7) as:

$$J(\alpha) = -\sum_{m=1}^{M}\alpha_m tr\left(C_x^\top P_m C_x\right) + \lambda||\alpha||_p^2 \ s.t. \ \alpha_m \geq 0, \quad (10)$$

where $P_m = (D^x)^{-\frac{1}{2}}W_m^x(D^x)^{-\frac{1}{2}}$. The *Lagrangian* is:

$$J'(\alpha) = -\sum_{m=1}^{M}\alpha_m tr\left(C_x^\top P_m C_x\right) + \lambda||\alpha||_p^2 - \sum \pi_m\alpha_m, \quad s.t. \ \pi_m \geq 0. \quad (11)$$

By setting $\frac{\partial J'(\alpha)}{\partial \alpha_m} = 0$, we have:

$$2\lambda(\sum_{m=1}^{M}(\alpha_m)^p)^{\frac{2}{p}-1}(\alpha_m)^{p-1} = tr\left(C_x^\top P_m C_x\right) + \pi_m. \quad (12)$$

When $p = 1$, we directly obtain the solution as:

$$\alpha_m = \frac{1}{2\lambda}tr\left(C_x^\top P_m C_x\right), \quad \alpha_{m'} = 0, \ \forall m' \neq m, \quad (13)$$

where m denotes the feature channel index with the largest $\frac{1}{2\lambda}tr\left(C_x^\top P_m C_x\right)$. When $p > 1$, by considering the KKT condition, we obtain the following equation:

$$\alpha_m = \frac{1}{\lambda}b_m^{\frac{1}{p-1}}\left(\sum_{m'=1}^{M}(b_{m'})^{\frac{p}{p-1}}\right), \quad m = 1, \ldots, M, \quad (14)$$

where $b_m = \frac{1}{2}tr\left(C_x^\top P_m C_x\right)$. The weight α is determined by the relative performance of different feature channels w.r.t. b_m. When p is 1 or close to 1, α is sparse and only the feature channel with the best performance is selected by the model. When p is large, all α_m tend to be identical, and the model will be equivalent to the average combination. The setting of p is application dependent.

The learning process is shown in Algorithm 1. The time complexity of Step 1 is $O(s^2N^3)$, where s denotes the sparse degree of the multi-modal similarity matrix, which is about [0.005, 0.015] in this paper, N denotes the number of cross-modal data. The time complexity of Step 2 is $O(sMN^2)$ in computing b_m. The total time complexity of hash code learning is $O(s^2N^3T)$, where $T \leq 10$ denotes the number of iterations.

F. HASH FUNCTION LEARNING

After learning and binarizing the hash codes of the training data on both modalities, B hash functions are learned for each data modality separately by mapping the features to the learned hash codes. There are mainly three types of hashing functions. The first choice to obtain the hash functions is training linear SVM on the original features by using each column of C as the label vectors similar as [10]. Second, a multi-layer neuro-network trained with back propagation can be trained based on the learned codes C for each modality, similar as [18], [29]. To deal with the complicated distribution of real world multi-modal data, we employ Multiple Kernel Learning [21], [22] to learn the hash functions for visual modality and kernel SVM for textual modality, where the parameter $p \geq 1$ used in the l_p-norm weight regularization in MKL is identical to the regularization of the weight α. The following hash functions are obtained for X and Y, respectively:

$$\begin{cases} h_k^x(z) = \mathrm{sgn}(\sum_{i=1}^{N_x}\theta_{i,k}^x c_{i,k}\sum_{m=1}^{M}\alpha_m^x K_m^x(x_{i,m}, z_m) + \gamma_k) \\ h_k^y(z) = \mathrm{sgn}(\sum_{j=1}^{N_y}\theta_{j,k}^y c_{j,k}K^y(y_j, z) + \upsilon_k), \end{cases} \quad (15)$$

where $k = 1, \ldots, B$, $\theta_{\cdot,k}$ and $c_{\cdot,k}$ denote the support vectors of the k-th hash function and the k-th dimensional binarized hash codes of the training data, α_m^x denotes the learned kernel weight by using [22], γ_k and υ_k represent the bias term of the k-th hash functions. Compared with other hashing strategies which either learn a set of linear functions [10] or directly learn a set of linear projection during hash code learning [19], [23], [32], our hash code learning strategy fully

Algorithm 1 Cross-Modal Hash Code Learning

```
$\alpha_m(0) = \frac{1}{M}$, $t = 0$.
while $t <= t_{max}$ or not converged do
    Get $\Theta_\alpha$ using $\alpha(t)$
    Get $D$ using $\Theta_\alpha$
    Optimize Eq.(8) with $C(t) = L(t)U_B(t)$
    Optimize Eq.(10) with Eq.(13) or Eq. (14)
    $t = t + 1$
end while
```

discovers the nonlinear semantic relation across modalities, and encodes unknown data with more semantically consistent binary codes. For convenience, in the consequent sections we denote our model using linear, neuro-network and multiple kernel functions as MSC-l, MSC-n and MSC-k, respectively.

G. RELATION TO OTHER APPROACHES

Our model can be recognized as the random walk among heterogeneous nodes, where each node represents a visual or textual document. When the intra-modal feature weights are fixed, it can be seen as Spectral Hashing [9] on the complicated multi-modal similarity graph Θ_α. Our method can be viewed as a generalized version of two other graph based cross-modal hashing approaches [19], [32] since it is capable of dealing with multiple features and represent the inter-modal relation in a more reasonable way. It is related with [52] in using label information and orthogonality relaxation. Compared with other multi-feature hashing approaches [23], [24], we provide a more generalized l_p-norm weighted combination on multiple features and information fusion strategy for processing cross-modal data.

IV. EXPERIMENTS

Datasets. We conduct experiments on two cross-modal datasets: (1) **NUS-WIDE** [53] consists of 269,648 images and the associated tags collected from Flickr. Six types of low level visual features are provided. The 1000-dim tag vectors of images are treated as the textual description and 81-dim tag vectors are treated as ground-truth class labels. (2) **WIKI-CMR** is a collection of 6382 Wikipedia webpages constructed by [33]. Each Webpage is categorized into 11 topic categories by Wikipedia. Each page is split into several paragraphs, and each image in the page is associated with the paragraph where it was originally placed. For this dataset, multiple cross-modal documents belong to one original Webpage, and each cross-modal document may have several hyperlinks pointing to or pointed by other cross-modal documents. Consequently, the whole dataset contains 74961 paragraphs (textual documents), 35149 images, their category labels and cross-modal correlation. The textual paragraphs are represented by TF-IDF (70K-dim) after a stop word removal. Eight types of visual features are calculated for each image, including color, texture and bag-of-words.

Training/retrieving data partition. For NUS-WIDE, we randomly choose 10K (image,text) pairs as the training data, and the rest are treated as the retrieval database. Note that such setting is different from traditional data partition scheme of using NUS-WIDE. Our scheme is more suitable for evaluating the model generality of the hash code and function learning using small number of training data and database with larger size, where similar setting has been adopted by IMH [19]. Since the structure information on NUS-WIDE is missing, we model the intra-modal relation with only the content similarity, and the inter-modal relation with only the correspondence information. For WIKI-CMR, We select 20% of the data from both modalities as the training data, and the rest are treated as the retrieval database for cross-modal retrieval.

Compared Approaches: Our approach includes three versions, *i.e.*, MSC-l, MSC-n and MSC-k. For the hash code learning, the number of nearest neighbors is set to be 71 for constructing intra-modal affinity. For hash function learning, we use hinge loss and $C = 10$ for MSC-l and MSC-k, and use the optimal setting for MSC-n.

To better show the performance of our approach, we compare our model with the following approaches which are models with hand-crafted features: (1) SSH [20]: Cross-modal similarity-sensitive hashing method. (2) MMNN [18]: the multi-layered neuro-network method. (3) MLBE [40]: a probabilistic multi-modal hashing approach. (4) CVH [32]: the graph based cross view hashing method. (5) IMH [19]: an inter-media linear hashing method. (6) SVG: cross-modal hashing with the best visual kernel and linear hash functions, which is the simplified version of our model. On WIKI-CMR, the compared methods are working on a 500-dim visual feature using dimension reduction on the concatenated visual feature (22K-dim), and 1000-dim reduced feature using SVD.

Evaluation Criteria: For NUS-WIDE, we evaluate the Mean Average Precision (MAP) using the 81-dim tag set as the ground-truth. We treat each data in the retrieval database as query, and the other as the database. Then the results of all the queries are averaged. For WIKI-CMR, each text document may either have more than one corresponding image or no image. Therefore, we randomly choose the text documents which have at least one corresponding image as the queries. Each retrieving process for a query input is considered to be successful if any one of the ground-truth corresponding image appears in the top 10 returned documents. We record the average success rate at top 10 results (ASR@10).

The experiments are conducted using Matlab on a desktop computer with Intel i5 3.1GHZ dual core CPU and 12G RAM.

A. IMAGE→TEXT RETRIEVAL

The results of Image→Text retrieval w.r.t. the code length on both datasets are recorded in Table 1. The bold numbers indicate the highest performance among all the compared methods, and * denotes the higher performance among each version of our approach with/without weak correlation. For

TABLE 1. Performance of Image→Text w.r.t. the code length.

Method	**NUS-WIDE**(MAP)				**WIKI-CMR**(ASR@10)			
	Code Length (B)				**Code Length** (B)			
	16	32	64	128	16	32	64	128
SSH	0.2521	0.2617	0.2723	0.2742	0.0952	0.1096	0.1145	0.1157
MMNN	0.2989	0.3143	0.3232	0.3235	0.1137	0.1291	0.1386	0.1388
MLBE	0.2335	0.2411	0.2496	0.2513	0.0910	0.1136	0.1239	0.1254
CVH	0.2499	0.2536	0.2687	0.2695	0.0938	0.1094	0.1137	0.1143
IMH	0.2938	0.3101	0.3198	0.3205	0.1159	0.1283	0.1392	0.1393
SVG	0.2563	0.2601	0.2798	0.2804	0.1045	0.1179	0.1331	0.1338
MSC-l (wo. WC)	0.3037	0.3156	0.3246	0.3248	0.1188*	0.1309	0.1393	0.1395
MSC-l (wt. WC)	-	-	-	-	0.1184	0.1317*	0.1408*	0.1424*
MSC-n (wo. WC)	0.3226	**0.3456**	0.3610	0.3612	0.1230	0.1343*	0.1428	0.1429
MSC-n (wt. WC)	–	-	-	-	0.1234*	0.1339	0.1437*	0.1443*
MSC-k (wo. WC)	**0.3237**	0.3453	**0.3616**	**0.3619**	0.1244	0.1377	0.1458	0.1461
MSC-k (wt. WC)	-	-	-	-	**0.1249***	**0.1386***	**0.1467***	**0.1479***

example, when $B = 32$, the ASRs of MSC-l (wt. WC) (learning with weak correlation) and MSC-l (wo. WC) (learning without weak correlation) are 0.1311 and 0.1309, respectively. We mark 0.1311 with * as it outperforms the other. There is no weak correlation on NUS-WIDE, so we learn our hash codes on the setting of (wo. WC), and perform both versions on WIKI-CMR.

For our method, we set $\rho = 0.1$. $p = 1.6$ and $\lambda = 0.8$ for NUS-WIDE, while $p = 2.5$ and $\lambda = 1$ are used for WIKI-CMR. When the code length is increased, the performance of all the methods is enhanced. On both datasets, the MAP and ASR@10 become stable with 64-bit or longer Hamming codes. On NUS-WIDE, the textual and visual data are organized with one-to-one correspondence and the data has been well annotated with 81-dim labels, so they can be fully exploited by all the compared approaches. On WIKI-CMR, Image→Text is a very challenging task, because the textual database contains many data items without corresponding images, and the category information of the cross-modal documents, *i.e.*, the topic domain (such as *politics* and *history*), does not provide direct guidance towards the evaluation criteria of ASR@k. In this case, methods without intra-modal similarity modeling (SSH) or without good inter-modal modeling (*e.g.*, CVH) tend to perform poorly. Under different settings, MSC consistently outperforms other approaches. Moreover, our model fully utilizes both the intra-modal similarity and inter-modal correlation information, so that more semantically consistent codes are learned. The results show that both intra-modality similarity (dissimilarity) and inter-modal correlation are of equivalent importance for constructing effective cross-modal retrieval model. Finally, by incorporating the weak correlation on WIKI-CMR, we observe certain improvement on all the versions of our method under nearly all the settings of B. Such a performance gain explains the rationality of using the structure context information of real world cross-modal data.

B. Text→Image RETRIEVAL

The results of Text→Image retrieval are recorded in Table 2. The parameter setting of our method is identical to Section IV-A. Compared with Image→Text retrieval, our approach outperforms the other approaches more significantly, especially on WIKI-CMR. Similar as the Image→Text retrieval, MMNN tends to have the best performance among all the benchmark approaches, and its performance is consistently increasing even when the code length is longer than 64. The performance of MSC are slightly increased when the number of bits is larger than 64. In general, the performance of all the approaches tends to become stable with 64-bit or longer codes. MSC-l outperforms other baselines. MSC-k performs the best. The results once again show that our approach achieves better generalization power by combining heterogeneous information.

To be more specific, as shown Table 1 and 2, our method MSC achieves the highest MAP among the compared approaches, which is at least relatively 10% higher than other approaches. On NUS-WIDE dataset, due to the data diversity, it appears that the multi-kernel and neural network hash functions used in different versions of MSC (*i.e.*, MSC-k and MSC-n, respectively) perform up and down over each other under different code lengths. The phenomenon demonstrates that different hash functions have different data fitting abilities on real-world large-scale data.

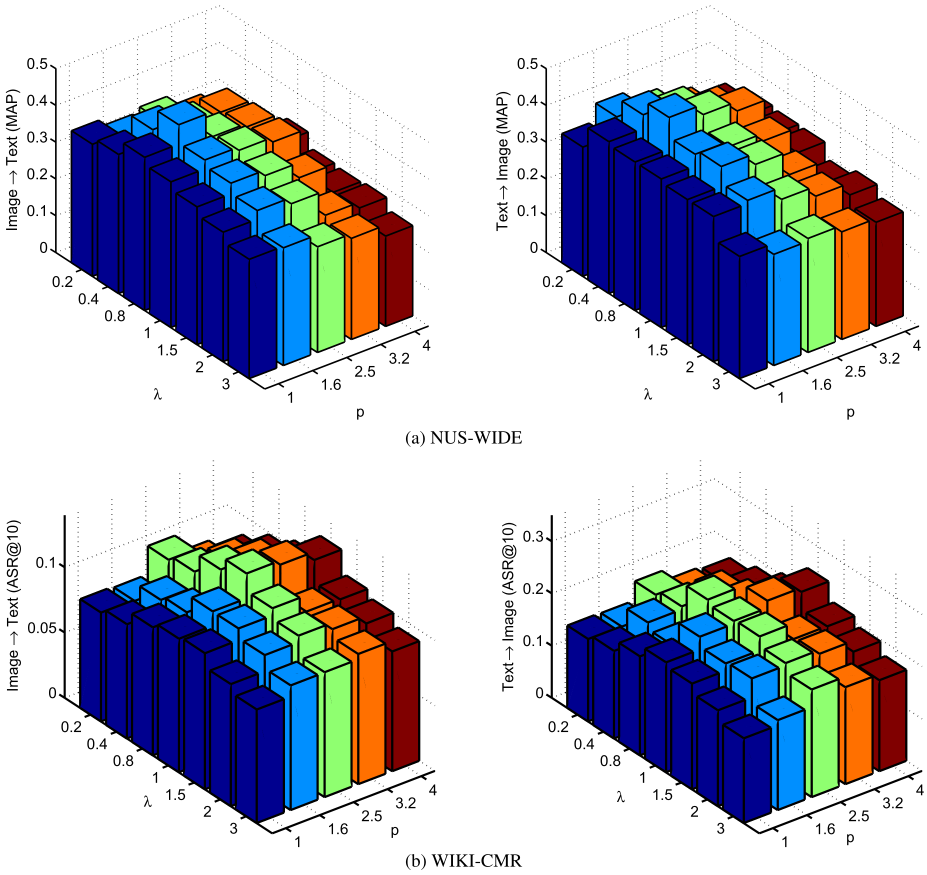
C. SENSITIVITY ON PARAMETERS

Our method has four parameters: the weak correlation influence $\tau \in (0, 1]$, the tolerance ρ of nonorthogonality, the structure of the feature weights p and the penalty λ on α. According to our empirical observation, τ is dataset dependent and cannot be too large, otherwise the solution tends to be unstable. For experiments on WIKI-CMR, we perform an empirical validation on τ, and find that $\tau = 0.3$ is the optimal setting.

When ρ becomes large, the hash codes will be more orthogonal to each other, leading to better hash coding performance [52]. However, large ρ also brings in more difficulty on the convergence since the solution will be pushed more tightly towards the identity matrix. We conduct experiments on $\rho = [10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$. The results show that when ρ is large, *e.g.*, 10 or 100, the training time is at least 2 times longer than small ρ, while the performance

TABLE 2. Performance of Text→Image w.r.t. the code length.

Method	**NUS-WIDE**(MAP)				**WIKI-CMR**(ASR@10)			
	Code Length (B)				**Code Length** (B)			
	16	32	64	128	16	32	64	128
SSH	0.2616	0.2698	0.2785	0.2789	0.1234	0.1287	0.1463	0.1479
MMNN	0.2976	0.3204	0.3238	0.3246	0.1379	0.1498	0.1556	0.1567
MLBE	0.2601	0.2712	0.2766	0.2803	0.1254	0.1358	0.1429	0.1467
CVH	0.2507	0.2586	0.2711	0.2768	0.1262	0.1341	0.1405	0.1409
IMH	0.3010	0.3198	0.3224	0.3226	0.1363	0.1499	0.1564	0.1576
SVG	0.2703	0.2811	0.2899	0.2906	0.1242	0.1349	0.1451	0.1462
MSC-l (wo. WC)	0.3045	0.3218	0.3286	0.3289	0.1393	0.1502	0.1574	0.1581
MSC-l (wt. WC)	-	-	-	-	0.1399*	0.1511*	0.1589*	0.1596*
MSC-n (wo. WC)	**0.3349**	0.3688	**0.3779**	0.3781	0.1635	0.1736*	0.1844	0.1849
MSC-n (wt. WC)	-	-	-	-	**0.1642***	0.1734	0.1857*	0.1861*
MSC-k (wo. WC)	0.3345	**0.3698**	0.3776	**0.3801**	0.1639*	0.1742	0.1854	0.1861
MSC-k (wt. WC)	-	-	-	-	0.1638	**0.1756***	**0.1867***	**0.1879***



(a) NUS-WIDE

(b) WIKI-CMR

FIGURE 2. Sensitivity analysis on different *p* and λ on NUS-WIDE (a) and WIKI-CMR (b).

improvement is not statistically significant. On the other hand, the generalization power of the nonlinear hash function may be a good compensation of nonorthogonality. By consideration of both efficiency and effectiveness, we set $\rho = 10^{-1}$.

To ensure the structure balance among different modalities, the penalty λ should not be too large or too small. By fixing the code length as $B = 64$, we test different settings of λ and p, the experimental results are demonstrated in Fig. 2. When λ is small, we see from Eqs. (13) and (14) that less penalty is imposed on α, thus the model variance is likely to increase, leading to unstable or even ill-posed solution. When λ is large, α would fail to capture the relative feature importance adequately, thus over-smooth solution is likely to occur.

In our method, the value of p determines the whole inter-modal and intra-modal correlation structure captured by the model. As discussed in previous sections, when p is small, α would be sparse, where the discriminating power of different

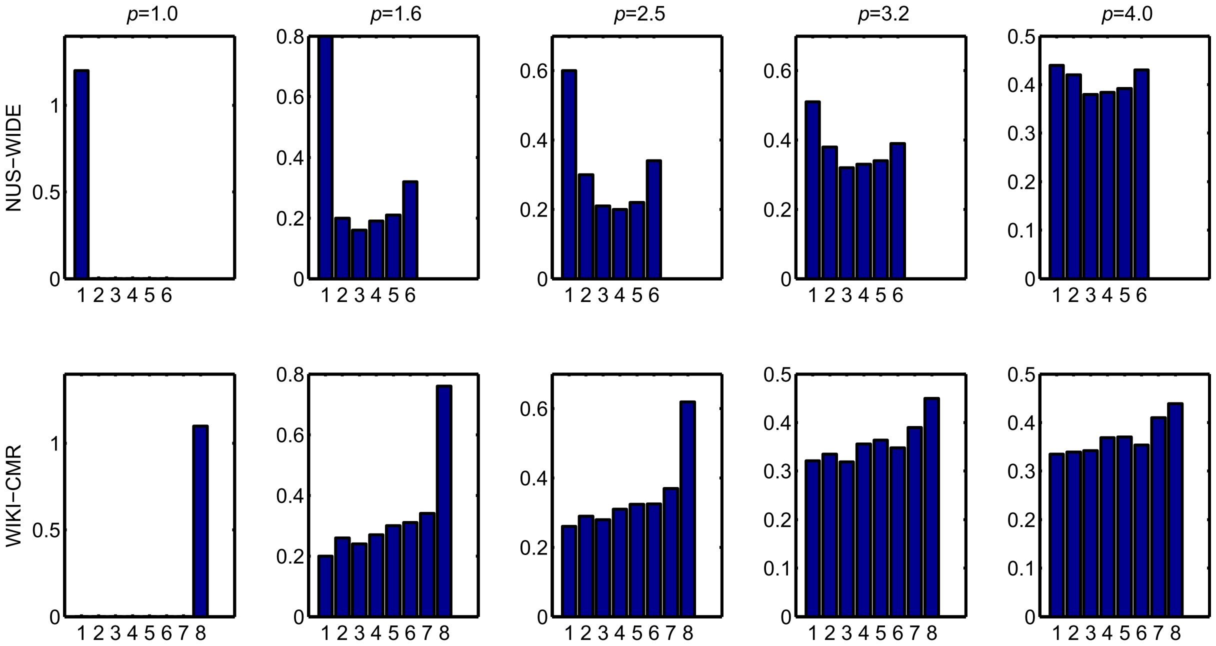


FIGURE 3. The learned α on NUS-WIDE (top) and WIKI-CMR (bottom). The bars from left to right on the top row are BOW, CH, CM, CORR, EDH and WT, respectively. The bars from left to right on the bottom row are PHOG180, PHOG360, CM, GIST, GIST4 × 4, LBP, SS and BOW, respectively.

features cannot be effectively combined. When p is too large, α would be over-dense, leading to inappropriate feature combination. To experimentally verify the influence of different p, we conduct experiments under different values of p to report the results on a randomly sampled validation data from NUS-WIDE training data by fixing other parameters. The results indicate that an appropriate value of p is neccessary to guarantee the optimal performance. Also, according to the results in Fig.2 by checking different combinations of p and λ, following similar validation process, we set $p = 1.6$ and $\lambda = 0.8$ for NUS-WIDE, and $p = 2.5$ and $\lambda = 1$ for WIKI-CMR.

Based on the fixed λ, we record the learned α with different p on both datasets in Fig. 3. We see that when $p = 1$, only the BOW feature is selected on both datasets, which means that the intra-modal similarity provided by BOW is more semantically consistent. When p becomes large, the learned feature weights tend to be uniform. The weights for good visual features are consistently larger than other features under all the settings of p.

D. EFFICIENCY

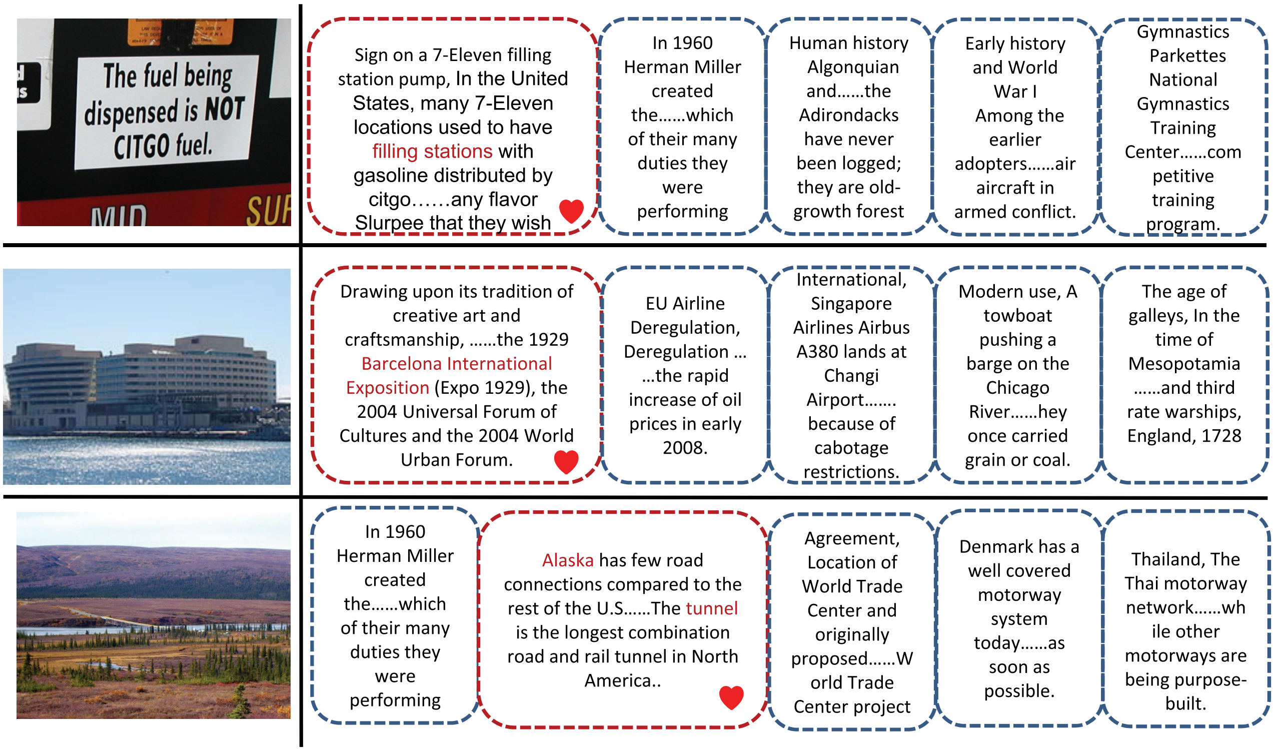
We record the training time under optimal parameter setting in Table 5 on WIKI-CMR. We use partial generalized Schur decomposition for optimizing CVH. We implement SSH where each weak coder in the boosting learning is learned by SVD. For MMNN, we use stochastic gradient as in [18]. We see that our method is efficient among the compared approaches except CVH, as our method involves an iterative process for both hash code learning and kernel weight learning. SSH and MMNN are time consuming because the dimensions of visual feature are very high, leading to time consuming SVD calculation and slow convergence rate using stochastic gradient. IMH is slow because it requires two $O(N^3)$ operations, namely, the inverse matrix calculation and eigen decomposition. Our model training is composed two steps, *i.e.*, hash code learning and hash function learning. The hash code learning is efficient according to the analysis in Section III-E. For hash function learning, the SMO algorithm we use in MSC-k has $O(N^3)$ complexity in the worst case, and is generally close to $O(N^2)$. In fact, an early stopping criterion for training any type of hash functions of our method is enough to guarantee good performance.

TABLE 3. Sensitivity on p on NUS-WIDE dataset (in MAP, $\lambda = 1$, $B = 128$).

p	1	1.4	1.6	1.8	2.0	2.2
MSC-n	0.2825	0.2847	**0.3102**	0.3023	0.2928	0.2899
MSC-k	0.2931	0.2949	**0.3134**	0.3108	0.2997	0.2933
p	2.5	2.8	3.0	3.2	3.6	4.0
MSC-n	0.2886	0.2897	0.2883	0.2871	0.2812	0.2790
MSC-k	0.2949	0.2901	0.2856	0.2832	0.2801	0.2799

E. PERFORMANCE EVALUATION ON MSCOCO

We conduct experiment on MSCOCO dataset which is a challenging cross-modal retrieval dataset. This dataset contains 82,783 images with 80 labels. Each image is also annotated by 5 independent sentences via Amazon Mechanical Turk. Following [54], after removing images without labels, we randomly select 10,000 image-text pairs for testing, and the remaining 72,081 pairs are used for training. For image representation, we use the multi-layer CNN features extracted



Sign on a 7-Eleven filling station pump, In the United States, many 7-Eleven locations used to have filling stations with gasoline distributed by citgo……any flavor Slurpee that they wish

In 1960 Herman Miller created the……which of their many duties they were performing

Human history Algonquian and……the Adirondacks have never been logged; they are old-growth forest

Early history and World War I Among the earlier adopters……air aircraft in armed conflict.

Gymnastics Parkettes National Gymnastics Training Center……competitive training program.

Drawing upon its tradition of creative art and craftsmanship, ……the 1929 Barcelona International Exposition (Expo 1929), the 2004 Universal Forum of Cultures and the 2004 World Urban Forum.

EU Airline Deregulation, Deregulation … …the rapid increase of oil prices in early 2008.

International, Singapore Airlines Airbus A380 lands at Changi Airport…… because of cabotage restrictions.

Modern use, A towboat pushing a barge on the Chicago River……hey once carried grain or coal.

The age of galleys, In the time of Mesopotamia ……and third rate warships, England, 1728

In 1960 Herman Miller created the……which of their many duties they were performing

Alaska has few road connections compared to the rest of the U.S……The tunnel is the longest combination road and rail tunnel in North America..

Agreement, Location of World Trade Center and originally proposed……World Trade Center project

Denmark has a well covered motorway system today……as soon as possible.

Thailand, The Thai motorway network……while other motorways are being purpose-built.

FIGURE 4. Examples of Image→Text retrieval on WIKI-CMR, where the ground-truth documents are denoted with red dots.

from conv$_4$, conv$_5$, fc$_6$ and fc$_7$ of a pre-trained VGG-19 model. For representing text instances, we use 1,000-dim BoW vector with the TF-IDF weighting scheme. We compare our method with several recent methods using the same visual and textual feature extractors. For deep learning based method, we use the original image and text as the input. We compare with some recent competitors including:

(1) CMOS [3]: a cross-modal retrieval method which performs the multi-layer CNN feature aggregation for deriving an asymmetric image-text similarity.

(2) Harmonized GPLVM [54]: a harmonized multi-modal GPLVM which performs topological alignment between the hyperparameter space of multi-modal GPLVM and the kernel matrix of the joint latent space. We report the results using the best variants hm-SimGP (tr) and hm-RSimGP (tr) with trace-norm kernel alignment.

(3) DCCAE [55]: a deep extension of the popular CCA for deep multimodal representation learning.

(4) ml-CCA [56]: a multi-label CCA-based method to perform cross-modal retrieval.

(5) 3V-CCA [57]: a three-view CCA-based method which treats the label space as the third modality.

(6) PRGDH [46]: a pair-wise relation guided deep cross-modal hashing method.

Note that for shallow models, we reproduce the results of the above mentioned methods based on features of the same network. Therefore, the results may be different from the original paper. For all the compared methods, we used the optimal parameter settings to produce nearly optimal performance. Code length of our methods are set to 128 for optimal performance, and p is set to 1.6. We report the mean Average Precision (mAP) for all the methods, as shown in Table 4.

TABLE 4. Cross-modal retrieval comparison in terms of mAP on MSCOCO dataset.

Methods	I→T	T→I	Average
3V-CCA	0.6216	0.6204	0.6210
ml-CCA	0.6297	0.6329	0.6313
DCCAE	0.6249	0.6257	0.6253
CMOS	0.6346	0.6778	0.6562
hm-SimGP (tr)	0.6237	0.6479	0.6358
hm-RSimGP (tr)	0.6540	0.6872	0.6706
PRGDH	0.6567	0.6908	0.6738
MSC-n	0.6500	0.6748	0.6624
MSC-k	0.6546	0.6891	0.6719

From the table we can see that, our method MSC-k ranks the second best compared to many recent state-of-the-art approaches. MSC-k is only outperformed by the end-to-end trained deep cross-modal hashing method PRGDH. The merit of our method comes from the ability of aggregating different visual features, self-supervised hashing function learning mechanism, which better fits to the diversed data distribution.

F. DISCUSSIONS ON THE EXPERIMENT RESULTS

1) ON PERFORMANCE GAIN

According to the experiment results in Tables 1 and 2, the performance gain depends on the following factors, *i.e.*, feature

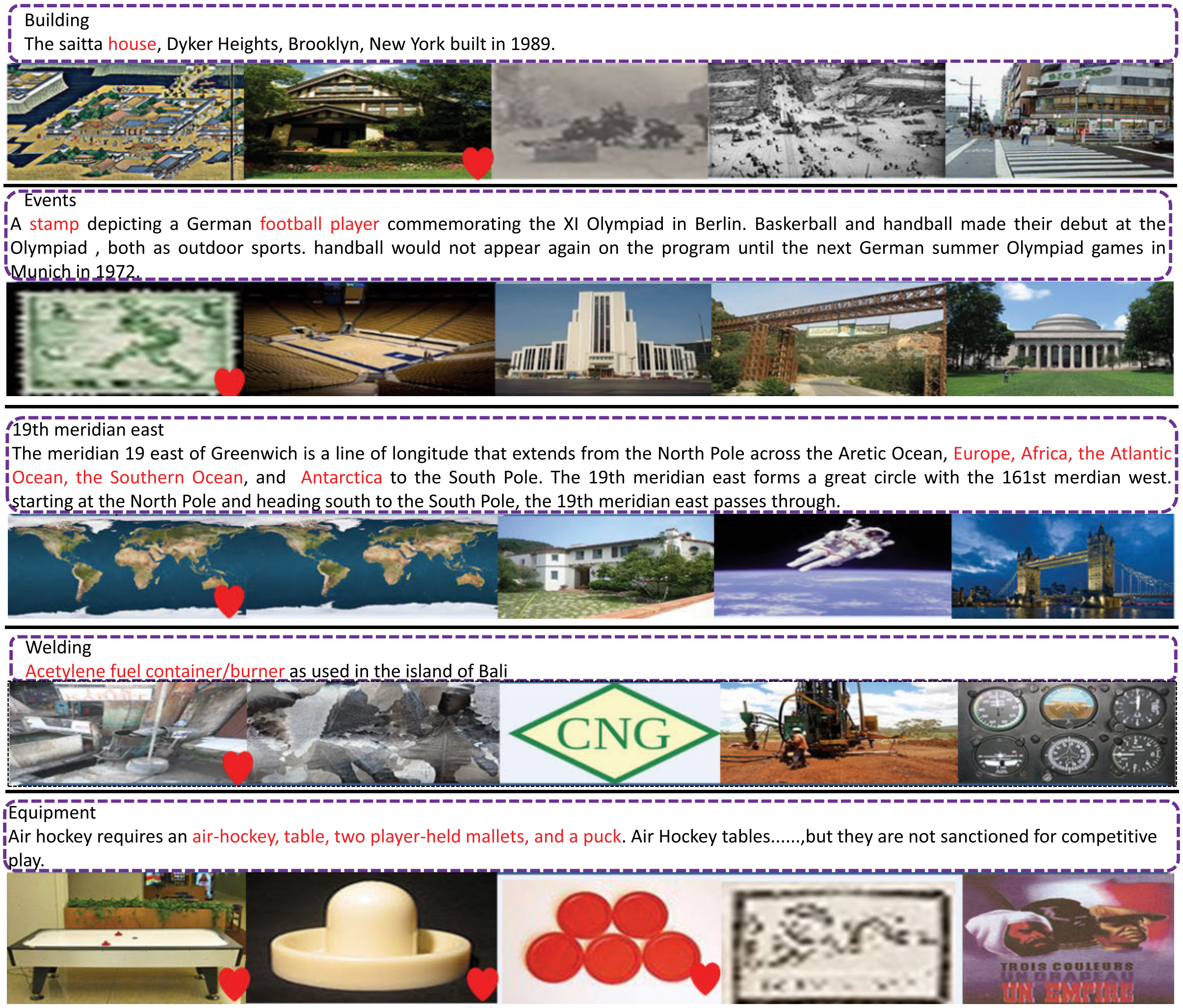

Building
The saitta house, Dyker Heights, Brooklyn, New York built in 1989.
Events
A stamp depicting a German football player commemorating the XI Olympiad in Berlin. Baskerball and handball made their debut at the Olympiad , both as outdoor sports. handball would not appear again on the program until the next German summer Olympiad games in Munich in 1972.
19th meridian east
The meridian 19 east of Greenwich is a line of longitude that extends from the North Pole across the Aretic Ocean, Europe, Africa, the Atlantic Ocean, the Southern Ocean, and Antarctica to the South Pole. The 19th meridian east forms a great circle with the 161st merdian west. starting at the North Pole and heading south to the South Pole, the 19th meridian east passes through.
Welding
Acetylene fuel container/burner as used in the island of Bali
CNG
Equipment
Air hockey requires an air-hockey, table, two player-held mallets, and a puck. Air Hockey tables......,but they are not sanctioned for competitive play.
TROIS COULEURS
UN DRAPEAU
UN EMPIRE

FIGURE 5. Examples of Text→Image retrieval on WIKI-CMR, where the ground-truth documents are denoted with red dots.

TABLE 5. The training time statistics on WIKI-CMR.

Method	SSH [20]	MMNN [18]	CVH [32]	IMH [19]	MLBE [40]	SVG	MSC-l	MSC-n	MSC-k
Time(s)	342	457	91	310	497	104	193	298	289

combination, better correlation modeling and the nonlinear hash functions. By effective feature combination and correlation modeling, we obtain hash codes with high semantic consistency. Therefore, promising performance is achieved even with linear mapping function. The neuro-network functions capture certain level of complicated nonlinear relations of cross-modal data, but they suffer from local solutions and high dimensionality.

2) ILLUSTRATIVE EXAMPLES

Some examples of Image→Text and Text→Image retrieval on WIKI-CMR are illustrated in Figures 4 and 5, respectively, where the top 5 retrieved documents are shown and the ground-truth corresponding documents are denoted with red dots. We mark the relevant textual words with red colors, and we see that the retrieved documents are semantically relevant with the query with respect to these keywords. Specially, in the bottom example of Fig. 5, all the ground-truth corresponding images are ranked as the top 3 results.

3) MODALITY IMBALANCE

The performance of Image→Text is consistently worse than Text→Image for all the methods. The reason can be attributed to the modality imbalance. For example, on NUS-WIDE, the textual information is very sparse. On WIKI-CMR, the number of textual documents is larger than that of visual documents. For Image→Text retrieval, the semantics in single visual query cannot be directly inferred, leading to large semantic gap between the visual query and textual database. For Text→Image retrieval, the semantic gap between queries and visual data can be alleviated by modeling the intra-modal visual affinity using feature combinations of database documents.

V. CONCLUSION

We propose MSC, a cross-modal hashing approach based on multiple subgraph combination. By jointly considering the content similarity and structure relation among cross-modal documents, we encode the intra-modal multi-feature similarity and inter-modal correlation with multiple subgraphs. Then they are combined into one similarity graph among all the data from heterogeneous modalities with an l_p-norm regularized weight coefficients on visual modality. The optimal hash codes and the weight coefficients are simultaneously learned in an alternating optimization process. The hash functions for different modalities can be separately constructed by utilizing linear or nonlinear binary classification models, which captures the complicated semantic relations among different modalities. Experiments on two challenging cross-modal datasets demonstrate the advantages of our approach over existing approaches.

REFERENCES

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

[2] Y. Wu, S. Wang, and Q. Huang, "Online fast adaptive low-rank similarity learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1310–1322, May 2020.

[3] Y. Wu, S. Wang, G. Song, and Q. Huang, "Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4299–4312, Sep. 2019.

[4] L. Chu, S. Jiang, S. Wang, Y. Zhang, and Q. Huang, "Robust spatial consistency graph model for partial duplicate image retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1982–1996, Dec. 2013.

[5] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.

[6] H. Liu, Z. Li, and X. Shu, "Image retrieval based on optimized visual dictionary and adaptive soft assignment," in *Proc. 9th Int. Conf. Internet Multimedia Comput. Service, (ICIMCS)*, Qingdao, China, Aug. 2017, pp. 180–190.

[7] G. Sun, S. Wang, X. Liu, Q. Huang, Y. Chen, and E. Wu, "Accurate and efficient cross-domain visual matching leveraging multiple feature representations," *Vis. Comput.*, vol. 29, nos. 6–8, pp. 565–575, Jun. 2013.

[8] 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 - SCG*, 2004, pp. 253–262.

[9] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. NIPS*, 2008, pp. 1753–1760.

[10] D. Zhang, J. Wang, D. Cai, and J. Lu, "Self-taught hashing for fast similarity search," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. - SIGIR*, 2010, pp. 18–25.

[11] R. Salakhutdinov and G. Hinton, "Semantic hashing," *Int. J. Approx. Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[12] G. Shakhnarovich, "Learning task specific similarity," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Feb. 2006.

[13] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2130–2137.

[14] 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.

[15] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[16] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," in *Proc. 26th Int. Conf. Comput. Linguistics, Conf., Tech. Papers*, Osaka, Japan, Dec. 2016, pp. 3485–3495.

[17] S. Wang, Y. Wu, and Q. Huang, "Improving cross-modal correlation learning with hyperlinks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[18] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.

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

[20] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3594–3601.

[21] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Jan. 2004.

[22] S. Viswanathan, N. Ampornpunt, M. Varma, and S. Vishwanathan, "Multiple kernel learning and the SMO algorithm," in *Proc. NIPS*, 2010, pp. 2361–2369.

[23] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *Proc. 19th ACM Int. Conf. Multimedia - MM*, 2011, pp. 423–432.

[24] D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. - SIGIR*, 2011, pp. 225–234.

[25] S. Kim, Y. Kang, and S. Choi, "Sequential spectral learning to hash with multiple representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, 2012, pp. 538–551.

[26] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 34, pp. 321–372, 1936.

[27] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. AISTATS*, 2012, pp. 199–207.

[28] 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.

[29] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[30] J. Masci, A. M. Bronstein, M. M. Bronstein, P. Sprechmann, and G. Sapiro, "Sparse similarity-preserving hashing," 2013, *arXiv:1312.5479*. [Online]. Available: http://arxiv.org/abs/1312.5479

[31] B. Wu, Q. Yang, W. S. Zheng, Y. Wang, and J. Wang, "Quantized correlation hashing for fast cross-modal search," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3946–3952.

[32] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. IJCAI*, 2011, pp. 1–6.

[33] W. Xiong, S. Wang, C. Zhang, and Q. Huang, "WIKI-CMR: A Web cross modality dataset for studying and evaluation of cross modality retrieval models," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[34] L. Xie, L. Zhu, and G. Chen, "Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 9185–9204, Aug. 2016.

[35] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1059–1071, May 2018.

[36] G. Irie, D. Liu, Z. Li, and S.-F. Chang, "A Bayesian approach to multimodal visual dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 329–336.

[37] D. Blei and M. Jordan, "Modeling annotated data," in *Proc. SIGIR*, 2003, pp. 127–134.

[38] H. Xiao and T. Stibor, "Toward artificial synesthesia: Linking images and sounds via words," in *Proc. NIPS Workshop Mach. Learn. Next Gener. Comput. Vis. Challenges*, 2010, pp. 1–20.

[39] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2407–2414.

[40] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2012, pp. 940–948.

[41] N. Chen, J. Zhu, F. Sun, and E. P. Xing, "Large-margin predictive latent subspace learning for multiview data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2365–2378, Dec. 2012.

[42] L. Xie, L. Zhu, P. Pan, and Y. Lu, "Cross-modal self-taught hashing for large-scale image retrieval," *Signal Process.*, vol. 124, pp. 81–92, Jul. 2016.

[43] L. Jin, X. Shu, K. Li, Z. Li, G.-J. Qi, and J. Tang, "Deep ordinal hashing with spatial attention," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2173–2186, May 2019.

[44] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1445–1454.

[45] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3232–3240.

[46] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.

[47] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[48] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 35–44.

[49] J. Tang, X. Shu, Z. Li, G. J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4s, pp. 68:1–68:22, 2016.

[50] A. B. Goldberg, X. Zhu, and S. Wright, "Dissimilarity in graph-based semi-supervised classification," in *Proc. AISTAT*, 2007, pp. 155–162.

[51] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. 21st Int. Conf. ICML Mach. Learn.*, Jul. 2004, p. 6.

[52] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3424–3431.

[53] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from national Uiversity of Sngapore," in *Proc. ACM Int. Conf. Image Video Retr. - CIVR*, 2009, pp. 1–9.

[54] G. Song, S. Wang, Q. Huang, and Q. Tian, "Harmonized multimodal learning with Gaussian process latent variable models," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 17, 2019, doi: 10.1109/TPAMI.2019.2942028.

[55] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[56] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4094–4102.

[57] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.

DONGXIAO REN was born in Nanyang, Henan, China, in 1982. She received the B.S. degree in computer science and technology from Henan University in 2005, the M.S. degree in computer software and theory from Southwest Jiaotong University in 2008, and the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology in 2012. From 2012 to 2017, she was a Senior Engineer with the State Grid Ningxia Electric Power Company and successfully completed dozens of projects. Since 2017, she has been a teacher with the School of Science, Zhejiang University of Science and Technology, Hangzhou, China. She is the author of more than ten articles. Her research interests include information processing, data science, big data technology, and image processing.

JUNWEI HUANG was born in Wenzhou, Zhejiang, China, in 1999. He is currently pursuing the bachelor's degree with Zhejiang University of Science and Technology, Hangzhou, China. During the period of school, he participated in the research and development of many projects, accumulated certain project experience, and had strong hands-on ability. His research interests include machine learning and quantitative analysis.

ZHONGHUA WANG was born in Zhoukou, Henan, China, in 1980. He received the B.S. degree in computer science and technology from Henan University in 2005, and the M.S. degree in business administration from Ningxia University in 2015. He has worked in industry for decades and accumulated rich project experience. He is currently a project manager with Beijing GuoDianTong Network Technology Company Ltd., and has finished more than ten projects. His research interests include information processing and machine learning.

FANG LU received the B.S. degree in information and computing science and the M.S. degree in applied mathematics from Zhengzhou University, China, in 2009 and 2012, respectively, and the Ph.D. degree in applied mathematics from Zhejiang University, China, in 2016. Since 2017, she has been a teacher with the School of Science, Zhejiang University of Science and Technology, Hangzhou, China. Her research interests include medical image processing and machine learning.