Multi-view Collective Tensor Decomposition for Cross-modal Hashing

Limeng Cui
School of Computer and Control Engineering, University of Chinese Academy of Sciences
Beijing, China
lmcui932@163.com

Zhensong Chen
School of Economics and Management, University of Chinese Academy of Sciences
Beijing, China
wxzmczs@163.com

Jiawei Zhang
IFM Lab, Department of Computer Science, Florida State University
Tallahassee, FL
jzhang@cs.fsu.edu

Lifang He
Weill Cornell Medicine, Cornell University
New York, NY
lifanghescut@gmail.com

Yong Shi
Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences
Beijing, China
yshi@ucas.ac.cn

Philip S. Yu
Department of Computer Science, University of Illinois at Chicago
Chicago, IL

ABSTRACT

Multimedia data available in various disciplines are usually heterogeneous, containing representations in multi-views, where the cross-modal search techniques become necessary and useful. It is a challenging problem due to the heterogeneity of data with multiple modalities, multi-views in each modality and the diverse data categories. In this paper, we propose a novel multi-view cross-modal hashing method named Multi-view Collective Tensor Decomposition (MCTD) to fuse these data effectively, which can exploit the complementary feature extracted from multi-modality multi-view while simultaneously discovering multiple separated subspaces by leveraging the data categories as supervision information. Our contributions are summarized as follows: 1) we exploit tensor modeling to get better representation of the complementary features and redefine a latent representation space; 2) a block-diagonal loss is proposed to explicitly pursue a more discriminative latent tensor space by exploring supervision information; 3) we propose a new feature projection method to characterize the data and to generate the latent representation for incoming new queries. An optimization algorithm is proposed to solve the objective function designed for MCTD, which works under an iterative updating procedure. Experimental results prove the state-of-the-art precision of MCTD compared with competing methods.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**;

KEYWORDS

Cross-modal hashing; tensor factorization; metric learning; multi-view learning

ACM Reference Format:
Limeng Cui, Zhensong Chen, Jiawei Zhang, Lifang He, Yong Shi, and Philip S. Yu. 2018. Multi-view Collective Tensor Decomposition for Cross-modal Hashing. In *ICMR '18: 2018 International Conference on Multimedia Retrieval, June 11–14, 2018, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3206025.3206065

ICMR '18, June 11–14, 2018, Yokohama, Japan

ACM ISBN 978-1-4503-5046-4/18/06…$15.00
https://doi.org/10.1145/3206025.3206065

1 INTRODUCTION

With the prevalence of multimedia big data in social networks and search engines, data of multi-modality has becoming ubiquitous and abundant. What's more, in some scenarios, each modality can further have representations in multiple views. For instance, as shown in Figure 1, each news article has information in texts and images (i.e., multi-modality), and information in each modality can have different representations in the feature space (depending on the specific feature extraction emphases), which form different views of the article. Since these data presented in different modalities may have strong semantic correlations, cross-modal retrieval has attracted growing attentions, which aims at using one kind of modality to retrieve semantically relevant objects of different modality.

In the past few years, this has become a fundamental problem in several emerging applications including visual search [2], image annotation [6, 25], and object detection/recognition [4]. A promising solution to cross-modal retrieval is the hashing method, which embeds multi-modal data into a common latent representation space and generates similar binary codes for similar objects [28]. However, effective cross-modal hashing still remains a challenge due to the heterogeneous data: features obtained in multi-modality and multi-view, and complementary supervision information, e.g., data categories. In this paper, we focus on how to fuse these data properly to facilitate the cross-modal retrieval task.

Existing cross-modal retrieval methods fail to fully exploit the useful features, which may limit their performance. To be more specific, the features extracted through different views in each modality can provide complementary information. In this paper, we refer to the feature representations extracted with different emphases as "multi-view features". For example, hand-crafted features and deep-learnt features characterize the different aspects of image data [1, 9, 26]. Similarly, explicit and latent features play different roles for text data characterization. Since the relationship among these multi-view features in each modality can be highly nonlinear, simply concatenating the features may result in that dense views dominate the feature space and override the effects of others. An effective fusion strategy is needed to explore feature interactions across different views.

In terms of the supervision information, like news article categories as shown in Figure 1, it also has not been explored effectively in current methods. It represents the class label of the multi-modal data, such as the topic of a news coverage. Most existing methods enforce the same-class samples lie as close as possible in the representation space. However, as separating different subspaces that correspond to different classes is widely ignored, they lose interclass discriminant ability. Motivated by this, we propose to utilize the supervision information to learn a more discriminative representation space.

In this paper, we study the correlations of the multi-view features and construct a novel latent space with the help of supervision information, which is the first attempt to fuse the aforementioned data together for cross-modal retrieval. It is a non-trivial task due to the following problems:

- The features of each modality are available from multiple views and provide complementary information. How do we fuse the features and explore the potential correlations to facilitate the cross-modal retrieval task?
- In order to learn a more discriminative latent space, we should enforce different-class samples to be embedded far apart. How do we incorporate the class information into the learning process?
- For incoming new queries, how do we map their multi-view features into the latent representation space and obtain the hash code?

This paper presents a novel multi-view cross-modality hashing method, termed as Multi-view Collective Tensor Decomposition (MCTD), which addresses the above issues under a collective tensor decomposition framework, shown in Figure. 1. To our best knowledge, the tensor decomposition has not been studied in the problem of cross-modal retrieval so far. Motivated by the issues that early fusion strategies, i.e. concatenating views, would lead to the problem of dominant views, we first propose a fusion approach for modeling the multi-view data of different modalities, and collectively learn a novel latent tensor representation between them by using tensor decomposition technique. Next, we consider enforcing supervision information as maintaining the block-diagonal structure of the latent tensor representation between different modalities, which eventually boils down to a block-diagonal structure loss term in the objective function. Finally, for new queries, we map the original features into the latent representation space by proposing a group

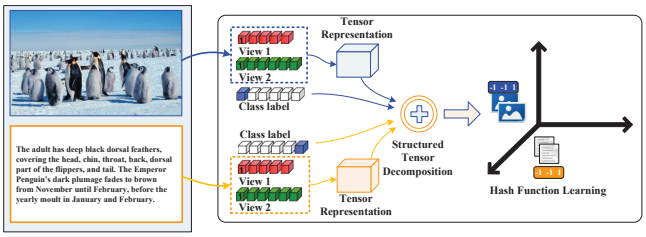


Figure 1: Multi-view Tensor Decomposition Hashing (MCTD) for cross-modal retrieval of images and text sentences.

of linear projections for each modality respectively. Extensive experiments show that MCTD yields state-of-the-art cross-modal retrieval performance.

2 RELATED WORK

Numerous papers have been published on cross-modal retrieval over the past decades [2, 8, 12, 14–16, 25, 27, 32, 35, 36]. Interested readers are referred to [29] for a comprehensive survey of various cross-modal retrieval methods. We now discuss related work in subspace learning and supervised cross-modal hashing.

Subspace Learning: In [37], authors proposed to perform cross-modal similarity search by employing Sparse Coding and Matrix Factorization (SCMF) to bridge the semantic gap and capture high-level latent semantic information. In [15], the Non-negative Matrix Factorization (NMF) was applied across the different modalities to tackle the multi-modal problem. In [31], authors proposed Semantic Consistency Hashing (SCH) method by learning a shared semantic space. A cross-modality hashing method based on matrix factorization (SMFH) [27] was proposed to consider the label consistency across different modalities. In [13], authors proposed a ranking-based method which constructs a common Hamming space where the cross-modal similarity can be measured by using Hamming distance.

Our collective tensor decomposition differs from these methods. On one hand, we explore the correlations of the input multi-view features. Different from traditional multi-view methods [24], MCTD considers multi-view features. On the other hand, inspired by the idea of collective matrix factorization, we propose to use tensor decomposition to learn the latent space which captures a broader view of features.

Supervised Cross-modal Hashing: The prior cross-modal hashing methods can be roughly divided into three categories including unsupervised, semi-supervised [33] and supervised cross-modal hashing. Here we only discuss several techniques that have explored the use of supervision information. In [32], Semantic Correlation Maximization (SCM) was proposed to seamlessly integrate semantic labels into the hashing learning procedure for large-scale data modeling. Semantics-Preserving Hashing (SePH) was proposed in [14], which can transform the given semantic affinities of training data into a probability distribution and approximate it with the hash codes in Hamming space. Semi-paired Discrete Hashing (SPDH) [23] jointly learns the latent features and hash codes with a factorization-based coding scheme. Discrete Cross-modal Hashing

(DCH) was proposed in [30], which directly learns discriminative binary codes while retaining the discrete constraints. As for deep learning methods, [8] presented a deep hashing model to capture the cross-modal correspondences between visual data and natural language. [34] used GAN for unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. [19] proposed a hierarchical network with multi-grained fusion for cross-modal correlation learning. Cross-media multiple deep network (CMDN) was proposed in [18] to exploit the cross-modal correlation by hierarchical learning.

In contrast to previous work, we maintain the intraclass similarity and the interclass dissimilarity at the same time. This allows us to learn more subtle variations in data structure and leads to a more accurate and efficient algorithm.

3 MULTI-VIEW COLLECTIVE TENSOR DECOMPOSITION

Suppose that we have training data with n instances drawn from two modalities $\mathcal{I}$ and $\mathcal{T}$, where the data from each modality are composed with $V \in \mathbb{R}$ views. Specifically, $\mathbf{X}_I = [\mathbf{X}_I^{(1)}; \mathbf{X}_I^{(2)}; \cdots; \mathbf{X}_I^{(V)}] \in \mathbb{R}^{(m_I^1+\cdots+m_I^V)\times n}$ and $\mathbf{X}_T = [\mathbf{X}_T^{(1)}; \mathbf{X}_T^{(2)}; \cdots; \mathbf{X}_T^{(V)}] \in \mathbb{R}^{(m_T^1+\cdots+m_T^V)\times n}$ are the training data matrices drawn from modality $\mathcal{I}$ and modality $\mathcal{T}$ respectively, where $\mathbf{X}_I^{(v)} \in \mathbb{R}^{m_I^v \times n}$ and $\mathbf{X}_T^{(v)} \in \mathbb{R}^{m_T^v \times n}$ are the data matrix for the v-th view, m_I^v and m_T^v are the corresponding dimensions of view $v \in [1:V]$. The goal of MCTD is to learn two groups of hash functions for the data from each modality that are able to generate unified hash codes.

Multi-view Collective Tensor Decomposition (MCTD) is a unified framework with three main components for supervised learning to hash, as shown in Figure 2. The framework accepts input in an image-text pairwise form and processes them through latent representation learning: (1) collective tensor decomposition to generate a common latent representation space between two modalities represented in full-order tensor form; (2) a block-diagonal loss for exploiting supervision information; and (3) two groups of linear projections for mapping the new queries into the latent space.

3.1 Collective Tensor Decomposition

Most cross-modal hashing methods are built upon a reasonable assumption that heterogeneous data with the same semantic label share a common subspace [15, 27, 38], called latent representation space. In the latent space, the semantic representations of relevant data from different modalities are close to each other. We follow this idea and pursue a more general framework. In this part, we explore the correlations on multi-views across different modalities and propose a novel latent representation space learning method by using collective tensor decomposition.

Modeling Correlations on Multi-view: In order to capture interactions among the features across multiple views on two modalities, here we propose a fusion strategy by exploring the concept of Factorization Machines [21] to capture the second-order interactions as well as the concept of Multi-view Machines [3] to capture higher order interactions.

Hence, to fully utilize the complementary information provided by multiple views, we use the full-order interactions among all the V views to represent each data instead of the direct concatenating. Specifically, for each instance $\mathbf{x}_I = [\mathbf{x}_I^{(1)}; \cdots; \mathbf{x}_I^{(V)}]$ from modality $\mathcal{I}$, we can compose the full-order interactions among different views through the outer product of the feature vectors from different views as follows:

$$\begin{aligned} &1^{st}\ order: \mathbf{x}_I^{(v)} \quad \forall v \in [1:V] \\ &2^{nd}\ order: \mathbf{x}_I^{(v_1)} \circ \mathbf{x}_I^{(v_2)} \quad \forall v_1, v_2 \in [1:V], v_1 \neq v_2 \\ &\cdots \\ &V^{th}\ order: \mathbf{x}_I^{(1)} \circ \cdots \circ \mathbf{x}_I^{(V)} \end{aligned} \tag{1}$$

It is easy to integrate all the interactions into a unified tensor representation by adding a constant value "1" to each feature vector $\mathbf{x}_I^{(v)}$, $v \in [1:V]$. Let $\mathbf{k}_I^{(v)} = [1; \mathbf{x}_I^{(v)}]$, we have the tensor representation for each instance as $\mathcal{K}_I = \mathbf{k}_I^{(1)} \circ \cdots \circ \mathbf{k}_I^{(V)} \in \mathbb{R}^{d_I^1 \times \cdots \times d_I^V}$, where $d_I^v = m_I^v + 1$ for all $v \in [1:V]$. Different from directly modeling the interactions of feature $\mathbf{x}_I^{(v)}$, now we can get feature interactions with different orders which reflects complementary insights.

Then, the data matrix from modality $\mathcal{I}$ can be transformed into the data tensor $\mathcal{X}_I = [\mathcal{K}_{I1}, \mathcal{K}_{I2}, \cdots, \mathcal{K}_{In}] \in \mathbb{R}^{d_I^1 \times \cdots \times d_I^V \times n}$. Similarly, we can get the tensor representation for the data matrix of modality $\mathcal{T}$: $\mathcal{X}_T \in \mathbb{R}^{d_T^1 \times \cdots \times d_T^V \times n}$, where $d_T^v = m_T^v + 1$ for all $v \in [1:V]$.

Learning Latent Representation Space: In cross-modal hashing, heterogenous data are mapped into a unified latent representation space so that the similarity can be directly compared. Learning such latent space is of great importance. In this section, we propose a method called Collective Tensor Decomposition (CTD) to obtain the common representation. We apply Tucker tensor decomposition, which can be considered as a higher-order generalization of Principal Component Analysis (PCA). It decomposes a tensor into a core tensor multiplied by a matrix along each mode [10].

Suppose that we are given two heterogeneous data tensors $\mathcal{X}_I \in \mathbb{R}^{d_I^1 \times \cdots \times d_I^V \times n}$ and $\mathcal{X}_T \in \mathbb{R}^{d_T^1 \times \cdots \times d_T^V \times n}$. According to [17], the results of CTD on $\mathcal{X}_I$ and $\mathcal{X}_T$ can be expressed by

$$\begin{cases} \mathcal{X}_I \approx \mathcal{V} \times_1 \mathbf{U}_I^{(1)} \times_2 \mathbf{U}_I^{(2)} \cdots \times_V \mathbf{U}_I^{(V)} \\ \mathcal{X}_T \approx \mathcal{V} \times_1 \mathbf{U}_T^{(1)} \times_2 \mathbf{U}_T^{(2)} \cdots \times_V \mathbf{U}_T^{(V)} \end{cases} \tag{2}$$

where $\{\mathbf{U}_I^{(v)} \in \mathbb{R}^{d_I^v \times R}\}_{v=1}^V$, $\{\mathbf{U}_T^{(v)} \in \mathbb{R}^{d_T^v \times R}\}_{v=1}^V$ are the factor matrices (which are usually orthogonal) and can be thought of as the *principal components in each view*, $\mathcal{V} \in \mathbb{R}^{R \times \cdots \times R \times n}$ is the *core tensor* and its entries show the level of interaction between the different components. The $(V+1)$-th order tensor $\mathcal{V}$ is the common latent representation of $\mathcal{X}_I$ and $\mathcal{X}_T$.

The average decomposition loss for CTD is defined as

$$\begin{aligned} \mathcal{L}_{ctd} = \alpha ||\mathcal{X}_I - \mathcal{V} \times_1 \mathbf{U}_I^{(1)} \cdots \times_V \mathbf{U}_I^{(V)}||^2 \\ + (1-\alpha) ||\mathcal{X}_T - \mathcal{V} \times_1 \mathbf{U}_T^{(1)} \cdots \times_V \mathbf{U}_T^{(V)}||^2 \end{aligned} \tag{3}$$

where α is a trade-off parameter.

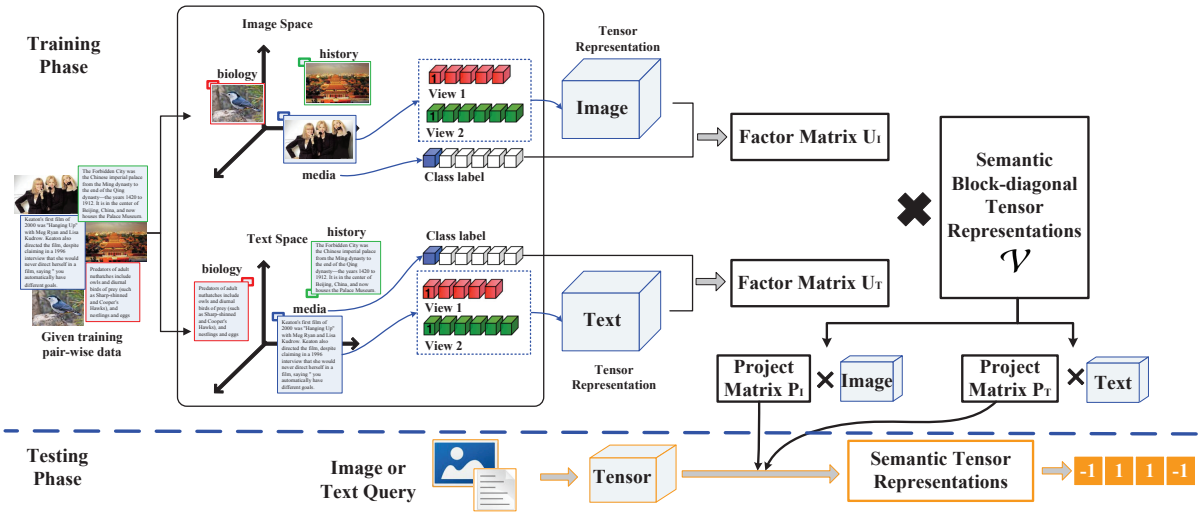


Figure 2: MCTD constitutes: (1) collective tensor decomposition to generate a common latent representation space between two modalities represented in full-order tensor form; (2) a block-diagonal loss for exploiting supervision information; and (3) two groups of linear projections for mapping the new queries into the latent space.

3.2 Block-diagonal Structure Loss

It is natural to assume that the intrinsic representations of data points from the same class are embedded in the same subspace and that these subspaces are separated. Therefore, it is straightforward to explicitly purse the block-diagonal structure of the latent tensor representation by exploring the labeled data. A novel loss named block-diagonal structure loss is proposed in this part.

Assume that these n data points are sampled from C classes and each instance is labeled with one class label. To better illustrate the block-diagonal structure, the labeled data instances are arranged according to their labels. For the tensor instances that belong to c class $\mathcal{X}_{Ic}$ and $\mathcal{X}_{Tc}$, their ideal common representation is denoted by $\mathcal{V}_c^* \in \mathbb{R}^{r\times\cdots\times r\times n_c}$, where r is the dimensionality of each subspace, n_c is the instance number of class c and $c \in [1:C]$. Since then, the ideal block-diagonal structured tensor representation $\mathcal{V}^*$ of data tensors $\mathcal{X}_I$ and $\mathcal{X}_T$ is shown as follows:

$$\mathcal{V}^* = diag(\mathcal{V}_1^*, \mathcal{V}_2^*, \cdots, \mathcal{V}_C^*) \tag{4}$$

However, the dimension of $\mathcal{V}$ is defined by hash code length, not by r. So we introduce a group of auxiliary matrices $\mathbf{Z}^{(v)}$ to change the mode of $\mathcal{V}$ into $\mathcal{V}^*$ with arbitrary dimension:

$$\mathcal{V}^* = \mathcal{V} \times_1 \mathbf{Z}^{(1)} \cdots \times_V \mathbf{Z}^{(V)} \tag{5}$$

where $\mathbf{Z}^{(v)} \in \mathbb{R}^{rC\times R}$ and $v \in [1:V]$. To enforce the block-diagonal structure of $\mathcal{V}$, we propose a loss function. In detail, let $\mathcal{E}_0 \in \mathbb{R}^{rC\times\cdots\times rC\times n}$ and $\mathcal{E}_c^* \in \mathbb{R}^{r\times\cdots\times r\times n_c}$ ($c \in [1:C]$) be the tensors with all elements equal "1". We first define an indicator tensor as

$$\mathcal{E} = \mathcal{E}_0 - diag(\mathcal{E}_1^*, \mathcal{E}_2^*, \cdots, \mathcal{E}_C^*) \tag{6}$$

Then, we have the loss of Block-diagonal Structure (BDS) as follows:

$$\mathcal{L}_{bds} = \frac{1}{2}||\mathcal{E} * (\mathcal{V} \times_1 \mathbf{Z}^{(1)} \cdots \times_V \mathbf{Z}^{(V)})||^2 \tag{7}$$

in which $*$ is the *Hadamard product*, which denotes the element-wise multiplication operator.

In fact, the block-diagonal structure loss can be seen as a global form of structural regularization that can influence the representations of all the classes. In this step, pursuing block-diagonal representations of the latent space guarantees that the representations of data points from the same class will be embedded in the same subspace and that different subspaces can be easily separated.

3.3 New Query Projection

For new queries, we can map the original feature interactions into the latent representation space by two groups of linear projections respectively:

$$\begin{cases} \mathcal{V}_I = \mathcal{X}_I \times_1 \mathbf{P}_I^{(1)} \times_2 \mathbf{P}_I^{(2)} \cdots \times_V \mathbf{P}_I^{(V)} \\ \mathcal{V}_T = \mathcal{X}_T \times_1 \mathbf{P}_T^{(1)} \times_2 \mathbf{P}_T^{(2)} \cdots \times_V \mathbf{P}_T^{(V)} \end{cases} \tag{8}$$

where $\mathbf{P}_I^{(v)} \in \mathbb{R}^{R\times d_I^v}$ and $\mathbf{P}_T^{(v)} \in \mathbb{R}^{R\times d_T^v}$ are the projecting matrix groups for all $v \in [1:V]$.

Since the tensors from different modalities that describing the same objects have the same semantic representations, we can present the loss for linear projections as

$$\begin{aligned} \mathcal{L}_{lp} &= ||\mathcal{V} - \mathcal{V}_I||^2 + ||\mathcal{V} - \mathcal{V}_T||^2 \\ &= ||\mathcal{V} - \mathcal{X}_I \times_1 \mathbf{P}_I^{(1)} \cdots \times_V \mathbf{P}_I^{(V)}||^2 \\ &\quad + ||\mathcal{V} - \mathcal{X}_T \times_1 \mathbf{P}_T^{(1)} \cdots \times_V \mathbf{P}_T^{(V)}||^2 \end{aligned} \tag{9}$$

3.4 Overall Objective Function

The overall objective function, consisting of the collective tensor decomposition term $\mathcal{L}_{ctd}$ in Eq. (3), the block-diagonal structure term $\mathcal{L}_{bds}$ in Eq. (7), the linear projection term $\mathcal{L}_{lp}$ in Eq. (9) and a regularization term, is given as follows:

$$
\begin{aligned}
\min \mathcal{L} &= \mathcal{L}_{ctd} + \mu\mathcal{L}_{bds} + \beta\mathcal{L}_{lp} \\
&\quad + \lambda\Psi(\{\mathbf{U}_I^{(v)}\},\{\mathbf{U}_T^{(v)}\},\{\mathbf{P}_I^{(v)}\},\{\mathbf{P}_T^{(v)}\},\{\mathbf{Z}^{(v)}\},\mathcal{V}) \\
&= \alpha||\mathcal{X}_I - \mathcal{V}\times_1\mathbf{U}_I^{(1)}\cdots\times_V\mathbf{U}_I^{(V)}||^2 \\
&\quad + (1-\alpha)||\mathcal{X}_T - \mathcal{V}\times_1\mathbf{U}_T^{(1)}\cdots\times_V\mathbf{U}_T^{(V)}||^2 \\
&\quad + \frac{\mu}{2}||\mathcal{E}*(\mathcal{V}\times_1\mathbf{Z}^{(1)}\cdots\times_V\mathbf{Z}^{(V)})||^2 \\
&\quad + \beta(||\mathcal{V} - \mathcal{X}_I\times_1\mathbf{P}_I^{(1)}\cdots\times_V\mathbf{P}_I^{(V)}||^2 \\
&\quad + ||\mathcal{V} - \mathcal{X}_T\times_1\mathbf{P}_T^{(1)}\cdots\times_V\mathbf{P}_T^{(V)}||^2) \\
&\quad + \lambda\Psi(\{\mathbf{U}_I^{(v)}\},\{\mathbf{U}_T^{(v)}\},\{\mathbf{P}_I^{(v)}\},\{\mathbf{P}_T^{(v)}\},\{\mathbf{Z}^{(v)}\},\mathcal{V})
\end{aligned} \tag{10}
$$

where μ, β and λ are the trade-off parameters of the corresponding terms, and the regularization term $\Psi(\cdot)$ is used to prevent overfitting.

The proposed formulation in (10) is hard to be directly solved since it is not convex or smooth with matrices $\{\mathbf{U}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{U}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ and tensor $\mathcal{V}$. Therefore, we adopt an iterative multiplicative strategy. Specifically, the optimization procedure can be divided into the following steps:

Step 1: With $\{\mathbf{P}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ and $\mathcal{V}$ fixed, the minimization over $\{\mathbf{U}_I^{(v)}\}$ and $\{\mathbf{U}_T^{(v)}\}$ are given by

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{U}_I^{(v)}} &= -2\alpha\Big(\mathbf{X}_{I(v)} - \mathbf{U}_I^{(v)}\mathbf{V}_{(v)}\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{U}_I^{(v')}\big)^T\Big) \\
&\quad\cdot\Big(\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{U}_I^{(v')}\big)\mathbf{V}_{(v)}^T\Big) + \lambda\frac{\partial\Psi\big(\mathbf{U}_I^{(v)}\big)}{\partial\mathbf{U}_I^{(v)}}
\end{aligned} \tag{11}
$$

and

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{U}_T^{(v)}} &= -2(1-\alpha)\Big(\mathbf{X}_{T(v)} - \mathbf{U}_T^{(v)}\mathbf{V}_{(v)}\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{U}_T^{(v')}\big)^T\Big) \\
&\quad\cdot\Big(\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{U}_T^{(v')}\big)\mathbf{V}_{(v)}^T\Big) + \lambda\frac{\partial\Psi\big(\mathbf{U}_T^{(v)}\big)}{\partial\mathbf{U}_T^{(v)}}
\end{aligned} \tag{12}
$$

where $\mathbf{X}_{I(v)}$, $\mathbf{X}_{T(v)}$ and $\mathbf{V}_{(v)}$ are the mode-v matricization of tensor $\mathcal{X}_I$, $\mathcal{X}_T$ and $\mathcal{V}$ respectively, $\otimes$ is the *Kronecker product* of matrices, and $\mathbf{U}_I^{(V+1)} = \mathbf{U}_T^{(V+1)} = \mathbf{E} \in \mathbb{R}^{n\times n}$ is the identity matrix.

Step 2: With $\{\mathbf{U}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{U}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_T^{(v)}\}_{v=1}^V$ and $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ fixed, we have

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathcal{V}} &= -2\alpha\big(\mathcal{X}_I\times_1\mathbf{U}_I^{(1)T}\cdots\times_V\mathbf{U}_I^{(V)T} - \mathcal{V}\big) \\
&\quad - 2(1-\alpha)\big(\mathcal{X}_T\times_1\mathbf{U}_T^{(1)T}\cdots\times_V\mathbf{U}_T^{(V)T} - \mathcal{V}\big) \\
&\quad + 2\beta\big(\mathcal{V} - \mathcal{X}_I\times_1\mathbf{P}_I^{(1)}\cdots\times_V\mathbf{P}_I^{(V)}) \\
&\quad + (\mathcal{V} - \mathcal{X}_T\times_1\mathbf{P}_T^{(1)}\cdots\times_V\mathbf{P}_T^{(V)})\big) \\
&\quad + \mu(\mathcal{E}\times_1\mathbf{Z}^{(1)T}\cdots\times_V\mathbf{Z}^{(V)T})*\mathcal{V} + \lambda\frac{\partial\Psi\big(\mathcal{V}\big)}{\partial\mathcal{V}}
\end{aligned} \tag{13}
$$

Step 3: With $\{\mathbf{U}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{U}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{P}_T^{(v)}\}_{v=1}^V$ and $\mathcal{V}$ fixed, the gradient w.r.t $\{\mathbf{Z}^{(v)}\}$ is shown as

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{Z}^{(v)}} &= \mu(\mathbf{E}_{(v)} * \mathbf{Z}^{(v)}\mathbf{V}_{(v)}\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{Z}^{(v')}\big)^T)) \\
&\quad\cdot\Big(\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{Z}^{(v')}\big)\mathbf{V}_{(v)}^T\Big) + \lambda\frac{\partial\Psi\big(\mathbf{Z}^{(v)}\big)}{\partial\mathbf{Z}^{(v)}}
\end{aligned} \tag{14}
$$

in which $\mathbf{E}_{(v)}$ is the mode-v matricization of tensor $\mathcal{E}$ and $\mathbf{Z}^{(V+1)} = \mathbf{E} \in \mathbb{R}^{n\times n}$ is the identity matrix.

Step 4: Similarly, with all the $\{\mathbf{U}_I^{(v)}\}_{v=1}^V$, $\{\mathbf{U}_T^{(v)}\}_{v=1}^V$, $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ and $\mathcal{V}$ fixed, we can obtain

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{P}_I^{(v)}} &= -2\beta\Big(\mathbf{V}_{(v)} - \mathbf{P}_I^{(v)}\mathbf{X}_{I(v)}\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{P}_I^{(v')}\big)^T\Big) \\
&\quad\cdot\Big(\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{P}_I^{(v')}\big)\mathbf{X}_{I(v)}^T\Big) + \lambda\frac{\partial\Psi\big(\mathbf{P}_I^{(v)}\big)}{\partial\mathbf{P}_I^{(v)}}
\end{aligned} \tag{15}
$$

and

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{P}_T^{(v)}} &= -2\beta\Big(\mathbf{V}_{(v)} - \mathbf{P}_T^{(v)}\mathbf{X}_{T(v)}\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{P}_T^{(v')}\big)^T\Big) \\
&\quad\cdot\Big(\big(\otimes\prod_{\substack{v'=V+1\\v'\neq v}}^{1}\mathbf{P}_T^{(v')}\big)\mathbf{X}_{T(v)}^T\Big) + \lambda\frac{\partial\Psi\big(\mathbf{P}_T^{(v)}\big)}{\partial\mathbf{P}_T^{(v)}}
\end{aligned} \tag{16}
$$

in which $\mathbf{P}_I^{(V+1)} = \mathbf{P}_T^{(V+1)} = \mathbf{E} \in \mathbb{R}^{n\times n}$ is the identity matrix. The optimization procedure of MCTD is summarized in Algorithm 1.

Overall, for any new instance $\mathbf{x}_I = [\mathbf{x}_I^{(1)};\cdots;\mathbf{x}_I^{(V)}]$ and $\mathbf{x}_T = [\mathbf{x}_T^{(1)};\cdots;\mathbf{x}_T^{(V)}]$ drawn from each modality, we first transfer them into the full-order interactions presented in the form of tensor representations $\mathcal{K}_I$ and $\mathcal{K}_T$ according to Eq. (1). Then, the MCTD is to learn two groups of hash functions for the data from each modality that are able to generate unified hash codes, i.e., $f(\mathcal{K}_I) = sign(\mathcal{K}_I\times_1\mathbf{P}_I^{(1)}\cdots\times_V\mathbf{P}_I^{(V)}) : \mathbb{R}^{d_I^1\times\cdots\times d_I^V} \to \{-1,+1\}^{R^V}$ and $g(\mathcal{K}_T) = sign(\mathcal{K}_T\times_1\mathbf{P}_T^{(1)}\cdots\times_V\mathbf{P}_T^{(V)}) : \mathbb{R}^{d_T^1\times\cdots\times d_T^V} \to \{-1,+1\}^{R^V}$, where d_I^v and d_T^v are the dimensions of mode-v fiber of tensor $\mathcal{K}_I$ and $\mathcal{K}_T$, and R^V is the length of binary codes.

3.5 Complexity Analysis

In the application, MCTD firstly generates the latent representation for a new query based on the achieved projection matrix groups $\{\mathbf{P}_I^{(v)}\}$ and $\{\mathbf{P}_T^{(v)}\}$, and then the hash codes can be obtained. The main time consumption of the proposed MCTD is the tensor decomposition, its complexity is $O(\Pi_{v=1}^V(d_I^v + d_T^v)R^{V-1}n^2)$. The parameters in Algorithm 1 are updated simultaneously, which indicates that the computation procedure can be paralleled. Therefore, the complexity caused by the interaction across V views is ameliorated. The convergence criterion used in our experiments is that the number of iterations is greater than a threshold (e.g. 200) or the decrease of the objective function value is smaller than a threshold.

Algorithm 1 MCTD

Require: Image feature matrix $\mathbf{X}_I$ and text feature matrix $\mathbf{X}_T$ both in V views, the length of hash codes R, the category C, and the model parameters α, β, μ and λ.

Ensure: Unified hash codes $\mathbf{H}$, and the projection matrix groups $\{\mathbf{P}_I^{(v)}\}_{v=1}^V$ and $\{\mathbf{P}_T^{(v)}\}_{v=1}^V$.

Transforming the data matrix $\mathbf{X}_I$ and $\mathbf{X}_T$ into the tensor representations $\mathcal{X}_I$ and $\mathcal{X}_T$.
Randomly initializing $\{\mathbf{U}_I^{(v)}\}$, $\{\mathbf{U}_T^{(v)}\}$, $\{\mathbf{P}_I^{(v)}\}$, $\{\mathbf{P}_T^{(v)}\}$, $\{\mathbf{Z}^{(v)}\}$ and $\mathcal{V}$ respectively.
while not converged **do**
  for $v := 1$ to V **do**
    Fixing $\{\mathbf{P}_I^{(v)}\}$, $\{\mathbf{P}_T^{(v)}\}$, $\{\mathbf{Z}^{(v)}\}$ and $\mathcal{V}$, update $\mathbf{U}_I^{(v)}$ and $\mathbf{U}_T^{(v)}$.
  end for
  Fixing $\{\mathbf{U}_I^{(v)}\}$, $\{\mathbf{U}_T^{(v)}\}$, $\{\mathbf{P}_I^{(v)}\}$, $\{\mathbf{P}_T^{(v)}\}$ and $\{\mathbf{Z}^{(v)}\}$, update $\mathcal{V}$.
  for $v := 1$ to V **do**
    Fixing $\{\mathbf{U}_I^{(v)}\}$, $\{\mathbf{U}_T^{(v)}\}$, $\{\mathbf{P}_I^{(v)}\}$, $\{\mathbf{P}_T^{(v)}\}$ and $\mathcal{V}$, update $\mathbf{Z}^{(v)}$.
  end for
  for $v := 1$ to V **do**
    Fixing $\{\mathbf{U}_I^{(v)}\}$, $\{\mathbf{U}_T^{(v)}\}$, $\{\mathbf{Z}^{(v)}\}$ and $\mathcal{V}$, update $\mathbf{P}_I^{(v)}$ and $\mathbf{P}_T^{(v)}$.
  end for
end while
Generating the hash codes by $\mathbf{H} = sign(\mathbf{V}_{(V+1)})$.

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed method MCTD comparing with several state-of-the-art hashing methods on two public cross-modal datasets.

4.1 Datasets

Experiments are conducted to validate the advantages of the proposed cross-modality hashing method on two real-world datasets.

Wiki[1]: Wiki dataset is collected from Wikipedia consisting of 2173/693 (training/testing) multimedia documents. Each document contains a single image and at least 70 words. Totally 10 categories are considered in this dataset and each image-text pair is labeled by one of them. Documents are considered to be similar if they belong to the same category.

Pascal VOC[2]: The data set [7] consists of 5011/4952 (training/testing) image-tag pairs, which can be categorized into 20 different classes. Since some images are multi-labeled, researchers usually select images with only one object as the way in [22], resulting in 1865 training and 1905 testing data. The image features include histograms of bag-of-visual-words, GIST and color. The text features are 399-D tag occurrence features.

4.2 Compared Methods

We compare the performance of our method with several state-of-the-art hashing based cross-modal retrieval methods including **CMFH**[3] [5], **LSSH**[4] [37], **SCM** [32], **SePH**[5] [14], **SMFH**[6] [27] and **DCMH**[7] [8], which can be organized into three categories:

- **Unsupervised hashing: LSSH** is an unsupervised method, which learns a joint abstraction space for image and text by using sparse coding and matrix factorization.
- **Supervised hashing (with shallow architecture): SCM** is a representative supervised method for cross-modal hashing, which is proposed to seamlessly integrate semantic labels into the hashing learning procedure. **CMFH** and **SMFH** are two methods based on matrix factorization, which both lean a common latent space for image and text. **SePH** uses the semantic affinities of training instances into a probability distribution and aims to approximate it in Hamming space. In the experiments, we use RBF kernel and take 500 as sampling size as advised in [14].
- **Supervised hashing (with deep architecture): DCMH** is the most recent work on deep cross-modal hashing, which integrates feature learning and hash-code learning into the same framework.

As existing cross-modal hashing methods can not deal with the multiple views, we concatenate the features to fit the model.

4.3 Evaluation Protocols

For Wiki dataset, each image is represented by a 128-D SIFT histogram and a 128-D CNN feature. We use the output of layer *fc*8 in the Alexnet [11], which is pretrained on ImageNet. Each text is represented by a 200-D bag-of-words feature and a 10-D topics' vector generated by Latent Dirichlet Allocation (LDA) model [20]. For Pascal VOC dataset, each image is also represented by both hand-crafted features and deep features. The ground-truth neighbors are defined as those image-text pairs which share category label.

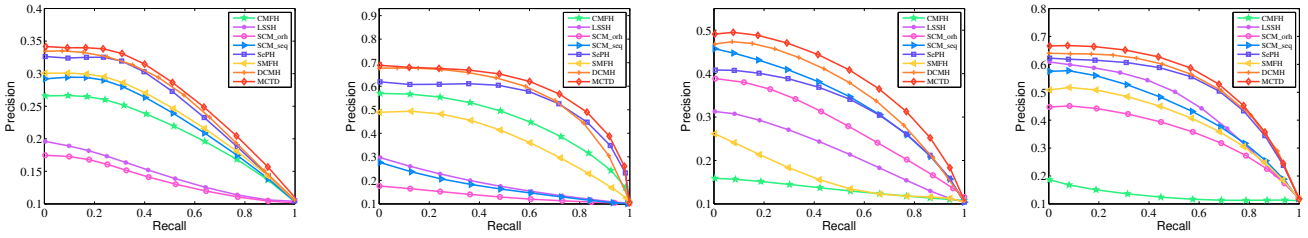
We perform two cross-modal retrieval tasks: using image queries to search relevant text ($I \rightarrow T$) and text query on image databases ($T \rightarrow I$). Following [14, 27, 32], we evaluate the retrieval performance based on two metrics: Mean Average Precision (MAP) and precision-recall curves. In our experiments, we repeat ten times for each group of parameters and report the mean MAP score. The results of numerical experiments are summarized in Table 1.

For our method, based on the rule of thumb, we set the parameters $r = 2$, $\alpha = 0.5$, and $\lambda = 0.05$ throughout the paper. The grid searching is applied to identify optimal values for the parameters from $\mu \in [0.001, 10]$ and $\beta \in [1, 200]$.

4.4 Quantitative Results

We evaluate all methods with different lengths of hash codes, i.e. 16, 32, 64 and 128 bits, and report their MAP results in Table 1, where the best results are presented in bold figures. From the experimental results, we can observe that the proposed MCTD method substantially outperforms all compared methods for cross-modal retrieval tasks which well demonstrates its effectiveness. Specifically,

[1] http://www.svcl.ucsd.edu/projects/crossmodal/
[2] htt p://www.cs.utexas.edu/~grauman/research/datasets.html
[3] http://ise.thss.tsinghua.edu.cn/MIG/code_data_cm.zip
[4] http://ise.thss.tsinghua.edu.cn/MIG/LSSH_code.rar
[5] https://bitbucket.org/linzijia72/
[6] We thank the authors for kindly providing the codes.
[7] https://github.com/jiangqy/DCMH-CVPR2017

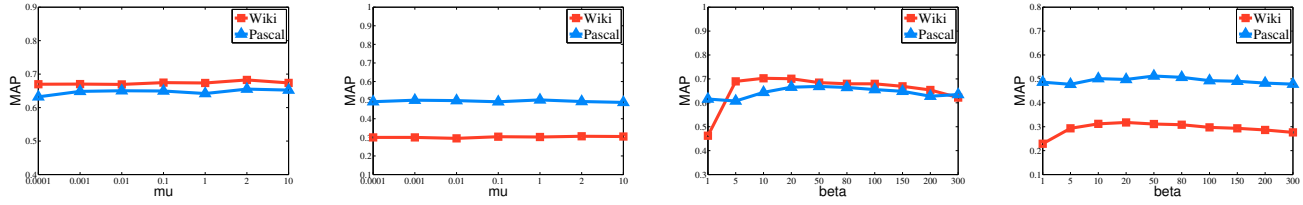


(a) Wiki (I→T) (b) Wiki (T→I) (c) Pascal (I→T) (d) Pascal (T→I)

Figure 3: Precision-recall curves of cross-modal retrieval on Wiki and Pascal.

Table 1: Mean Average Precision (MAP) for cross-modal retrieval tasks on two datasets. Items in bold indicate the best performance.

Task	Method	Wiki				Pascal VOC			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
I → T	CMFH	0.2115	0.2230	0.2238	0.2351	0.1575	0.1508	0.1490	0.1429
	LSSH	0.1541	0.1546	0.1544	0.1521	0.2988	0.3083	0.3194	0.3166
	SCM_orth	0.1527	0.1331	0.1216	0.1172	0.4063	0.4040	0.4067	0.4144
	SCM_seq	0.2257	0.2459	0.2461	0.2510	0.3842	0.4868	0.3972	0.4115
	SePH	0.2562	0.2654	0.2793	0.2823	0.4356	0.4424	0.4242	0.4245
	SMFH	0.2507	0.2646	0.2715	0.2787	0.2291	0.2477	0.2586	0.2500
	DCMH	0.2798	0.2809	0.2910	0.2993	0.4564	0.4613	0.4793	0.4801
	MCTD	**0.2919**	**0.3048**	**0.3068**	**0.3138**	**0.4921**	**0.4927**	**0.5194**	**0.5072**
T → I	CMFH	0.5351	0.5445	0.5586	0.5616	0.1576	0.1550	0.1523	0.1463
	LSSH	0.2641	0.2723	0.2795	0.2803	0.6145	0.6177	0.6042	0.5906
	SCM_orth	0.1532	0.1393	0.1297	0.1273	0.4791	0.4526	0.4962	0.4721
	SCM_seq	0.2341	0.2410	0.2445	0.2554	0.4816	0.5455	0.4526	0.4866
	SePH	0.6276	0.6324	0.6513	0.6514	0.6476	0.6524	0.6153	0.6571
	SMFH	0.4481	0.4827	0.4920	0.5038	0.4189	0.4942	0.6035	0.7388
	DCMH	0.6292	0.6524	0.6674	0.6720	0.6513	0.6504	0.6638	0.6708
	MCTD	**0.6482**	**0.6832**	**0.6898**	**0.6972**	**0.6567**	**0.6553**	**0.7074**	**0.7464**



(a) μ (T→I @32 bits) (b) μ (I→T @32 bits) (c) β (T→I @32 bits) (d) β (I→T @32 bits)

Figure 4: Parameter sensitive analysis: the MAP score @32 bits on both the cross-modal retrieval tasks.

compared to the best results of CMFH, LSSH, SCM_orth, SCM_seq, SePH and SMFH, MCTD achieves absolute increases of 1.66%/2.44% and 3.36%/3.24% in average MAP score for two cross-modal tasks $I \rightarrow T$ and $T \rightarrow I$ on Wiki and Pascal VOC datasets. This indicates that MCTD can not only effectively leverage the complementary information provided by the interactions among multi-view features, but also make full use of the category information to purse the block-diagonal representations that can boost the discriminant ability among different classes.

An interesting observation is that our method performs better than deep method DCMH. We assume that our model can use both hand-crafted features and deep-learned features from multiple views and exploit the complex feature correlations effectively. To confirm our assumption, we further test the effect of correlations on multi-views in Section 4.5.

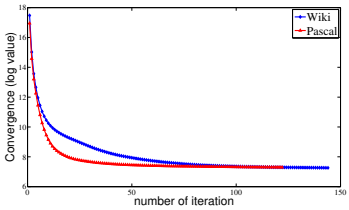


Figure 5: The curves of convergence validation: objective function value @32 bits on both datasets.

Table 2: The training time (seconds) of each method on Wiki.

Method	Dataset	
	Wiki	Pascal VOC
CMFH	14.02	16.38
LSSH	432.22	361.86
SCM_orth	2.90	29.72
SCM_seq	2.11	8.76
SePH	189.85	154.40
SMFH	39.02	60.78
MCTD	45.33	73.24

The precision-recall curves with 32 bits for the two cross-modal tasks $I \rightarrow T$ and $T \rightarrow I$ on these two datasets are presented in Figure 3, respectively. As it is shown, MCTD achieves the highly competitive results. Specifically, it achieves the second best precision at some lower recall values and obtains the best performance at the other recall levels on both datasets.

We then validate the convergence with 32 bits on these two datasets during each iteration. For making a clear show, we adjust the objective value to the log value. As shown in Figure 5, we can see that our method provides a satisfactory convergence rate in the optimization procedure on both datasets.

Finally, we investigate the training time of all these methods. The experiments are conducted on the Wiki dataset and run on a PC with 2.5 GHz Intel Core i7 CPU and 16GB RAM. As for experiments on DCMH, we use a server with NVIDIA GeForce GTX 1080 Ti GPU, so the training time is not included. Here we only evaluate the case that the code length is 32 bits. The results are reported in Table 2. We can observe that the time consumption of MCTD is of the same order of magnitude as that of CMFH and SMFH, both of which involve the computation of matrix inversion. The time cost is acceptable in comparison with that of LSSH and SePH. As MCTD needs to solve the object function in an iterative way, it spends a little more time than others in the training phase.

4.5 Effect of Correlations on Multi-views

To validate the advantage of correlations on multi-views, we present two variants of the proposed algorithm for comparison. The first one only uses the directly concatenated features and the second one only uses the highest-order correlations among all the views, where these two methods are referred as MCTD_c and MCTD_h. We evaluate their performance on the dataset Wiki with different lengths of hash codes and the MAP scores are summarized in Table 3. From this table, we can learn that the correlations on multi-views produce positive results and the full-order correlations can lead to the better performance by providing complementary information, which verifies the effectiveness of the proposed strategy.

Table 3: Effect of Correlations on Multi-views. Items in bold indicate the best performance.

Task	Method	Wiki			
		16bits	32bits	64bits	128bits
I → T	MCTD_c	0.2594	0.2783	0.2817	0.2909
	MCTD_h	0.2708	0.2865	0.2902	0.2944
	MCTD	**0.2919**	**0.3048**	**0.3068**	**0.3138**
T → I	MCTD_c	0.5536	0.6054	0.6231	0.6504
	MCTD_h	0.6024	0.6365	0.6592	0.6467
	MCTD	**0.6482**	**0.6832**	**0.6898**	**0.6972**

4.6 Parameter Sensitivity

We further discuss the performance of MCTD w.r.t the model parameters to analyze the effects of different parameter settings. The experiments are performed by varying the value of one parameter while fixing the others. Due to the space limit, here we only compute the MAP score @ 32 bits on both the cross-modal retrieval tasks, and the results with two important parameters β and μ are shown in Figure 4. From the figures, we can see that these two parameters are not sensitive and MCTD can yield satisfactory results in a wide range of parameter values.

5 CONCLUSION

In this paper, we propose an effective cross-modal hashing method called MCTD. Our framework builds upon complementary features from multi-views and combines this representation with tensor decomposition. More importantly, our method can discover multiple separated subspaces by leveraging the supervision information. Our innovations are shown as follows: Firstly, we propose to use collective tensor decomposition to capture the latent representation space between different modalities. Secondly, introducing a block-diagonal structure loss makes it possible to exploit the supervision information and maintain the global structure of the subspace. Thirdly, a group of linear projections for each modality, is proposed to map the original features of new queries into the latent representation space. In addition, we propose an optimization algorithm to solve the objective function, which can solve the problem effectively and is able to update multiple parameters simultaneously. Experimental results prove the effectiveness of our method in cross-modal retrieval compared to several competing methods.

Source code will be online shortly.

ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China under Grant No.: 61672313, 61503253, 91546201 and 71331005, the National Science Foundation under Grant No.: IIS-1526499 and CNS-1626432, and Natural Science Foundation of Guangdong Province under Grant No.: 2017A030313339.

REFERENCES

[1] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay. 2015. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1263–1266.

[2] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3594–3601.

[3] Bokai Cao, Hucheng Zhou, Guoqiang Li, and Philip S Yu. 2016. Multi-view machines. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 427–436.

[4] Changxing Ding and Dacheng Tao. 2015. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia* 17, 11 (2015), 2049–2058.

[5] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2075–2082.

[6] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision* 106, 2 (2014), 210–233.

[7] Sung Ju Hwang and Kristen Grauman. 2012. Reading between the lines: Object localization using implicit cues from image tags. *IEEE transactions on pattern analysis and machine intelligence* 34, 6 (2012), 1145–1158.

[8] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep Cross-Modal Hashing. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 3270–3278.

[9] Lu Jin, Shenghua Gao, Zechao Li, and Jinhui Tang. 2014. Hand-crafted features or machine learnt features? together they improve RGB-D object recognition. In *Multimedia (ISM), 2014 IEEE International Symposium on*. IEEE, 311–319.

[10] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[12] Shaishav Kumar and Raghavendra Udupa. 2011. Learning hash functions for cross-view similarity search. In *IJCAI proceedings-international joint conference on artificial intelligence*, Vol. 22. 1360.

[13] Kai Li, Guo-Jun Qi, Jun Ye, and Kien A Hua. 2017. Linear subspace ranking hashing for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence* 39, 9 (2017), 1825–1838.

[14] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang. 2017. Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE transactions on cybernetics* 47, 12 (2017), 4342–4355.

[15] Hong Liu, Rongrong Ji, Yongjian Wu, and Gang Hua. 2016. Supervised matrix factorization for cross-modality hashing. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 1767–1773.

[16] Sean Moran and Victor Lavrenko. 2015. Regularised cross-modal hashing. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 907–910.

[17] Morten Mørup, Lars Kai Hansen, and Sidse M Arnfred. 2008. Algorithms for sparse nonnegative Tucker decompositions. *Neural computation* 20, 8 (2008), 2112–2131.

[18] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks.. In *IJCAI*. 3846–3853.

[19] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia* 20, 2 (2018), 405–420.

[20] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 251–260.

[21] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.

[22] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2160–2167.

[23] Xiaobo Shen, Fumin Shen, Quan-Sen Sun, Yang Yang, Yun-Hao Yuan, and Heng Tao Shen. 2017. Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval. *IEEE transactions on cybernetics* 47, 12 (2017), 4275–4288.

[24] Xiaobo Shen, Fumin Shen, Quan-Sen Sun, and Yun-Hao Yuan. 2015. Multi-view latent hashing for efficient multimedia search. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 831–834.

[25] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 785–796.

[26] Jinhui Tang, Lu Jin, Zechao Li, and Shenghua Gao. 2015. RGB-D object recognition via incorporating latent data structure and prior knowledge. *IEEE Transactions on Multimedia* 17, 11 (2015), 1899–1908.

[27] Jun Tang, Ke Wang, and Ling Shao. 2016. Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Transactions on Image Processing* 25, 7 (2016), 3157–3166.

[28] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. 2014. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927* (2014).

[29] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A Comprehensive Survey on Cross-modal Retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[30] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* 26, 5 (2017), 2494–2507.

[31] Tao Yao, Xiangwei Kong, Haiyan Fu, and Qi Tian. 2016. Semantic consistency hashing for cross-modal retrieval. *Neurocomputing* 193 (2016), 250–259.

[32] Dongqing Zhang and Wu-Jun Li. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization.. In *AAAI*, Vol. 1. 7.

[33] Jian Zhang and Yuxin Peng. 2017. SSDH: semi-supervised deep hashing for large scale image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[34] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. 2018. Unsupervised Generative Adversarial Cross-modal Hashing. (2018).

[35] Yi Zhen and Dit-Yan Yeung. 2012. Co-regularized hashing for multimodal data. In *Advances in neural information processing systems*. 1376–1384.

[36] Yi Zhen and Dit-Yan Yeung. 2012. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 940–948.

[37] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 415–424.

[38] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 143–152.