Hetero-manifold Regularisation for Cross-modal Hashing

Feng Zheng, Yi Tang, and Ling Shao, *Senior Member, IEEE*

Abstract—Recently, cross-modal search has attracted considerable attention but remains a very challenging task because of the integration complexity and heterogeneity of the multi-modal data. To address both challenges, in this paper, we propose a novel method termed hetero-manifold regularisation (HMR) to supervise the learning of hash functions for efficient cross-modal search. A hetero-manifold integrates multiple sub-manifolds defined by homogeneous data with the help of cross-modal supervision information. Taking advantages of the hetero-manifold, the similarity between each pair of heterogeneous data could be naturally measured by three order random walks on this hetero-manifold. Furthermore, a novel cumulative distance inequality defined on the hetero-manifold is introduced to avoid the computational difficulty induced by the discreteness of hash codes. By using the inequality, cross-modal hashing is transformed into a problem of hetero-manifold regularised support vector learning. Therefore, the performance of cross-modal search can be significantly improved by seamlessly combining the integrated information of the hetero-manifold and the strong generalisation of the support vector machine. Comprehensive experiments show that the proposed HMR achieve advantageous results over the state-of-the-art methods in several challenging cross-modal tasks.

Index Terms—Cross-modal hashing, Manifold regularisation, Information propagation, Hinge loss constraint, Cumulative distance inequality.

1 Introduction

SEARCHING is dramatically changed by the amount and the appearance of multi-modal data. Multi-modal data are heterogeneous and large-scale because of the advancement of digital technologies and the Internet. Both of these fundamental characteristics of multi-modal data require measuring the cross-modal similarity when developing any searching algorithms by hashing.

To bridge the gap between modalities, various straightforward strategies have been developed to learn the cross-modal similarity. Some methods focus on the supervision information including correspondences [1], semantic correlation [2], pairwise sets [3] and semantic affinities [4] between heterogeneous data, while others including composite multiple information sources [5], α-average technique [6], [7], Markov random field [8] and deep neural networks [9] emphasise the value of homogeneous manifold in the problem of multi-modal similarity learning in a common space.

However, despite the progress made by existing methods considering certain aspects of the problem, cross-modal search remains a very challenging task because of the integration complexity and heterogeneity of the multi-modal data. In fact, the nature of multi-modal data is a combination of heterogeneity and the homogeneity. Thus, in cross-modal search, the cross-modal and within-modal similarity information should be simultaneously considered. On the one hand, the methods developed based on supervision information mainly focus on the similarity information of heterogeneity without considering the homogeneous information, but it is obvious that the within-modal similarity benefits to capture the intrinsic geometric structure. On the other hand, the methods generated by emphasising within-modal similarity decompose multi-modal data into a set of uni-modal data, which means multi-modal similarity learning cannot be treated as a whole because more than one manifold are needed to represent both cross-modal and within-modal similarities. Therefore, it is necessary to **connect** and **integrate** all information from data in different modalities to describe the diversity of the world. To achieve this, the key of cross-modal search is to overcome the obstacle of multiple modalities by considering both the local geometric and global supervision information.

In this paper, by integrating the supervision information and the local structure of heterogeneous data, a novel method termed hetero-manifold regularisation (HMR) is proposed to learn hash functions for efficient cross-modal search. Three significant advantages are illustrated in the schematic diagram of a hetero-manifold shown in Fig. 1. Firstly, a hetero-manifold well describes the local information by representing homogeneous data on the sub-manifolds. In Fig. 1, the data in three different modalities are represented by three sub-manifolds which well model the relationship between homogeneous data. Secondly, the hetero-manifold emphasises the global information of multi-modal data as well, by modelling the *information propagation* across modalities with three-order random walks. It is clear in Fig. 1 that any pair of points could be connected via two steps on homogeneous sub-manifolds and one step crossing two different sub-manifolds. Thus, the samples across modalities could be compared by integrating the information from all related homogeneous sub-manifolds. Lastly, the hetero-manifold is flexible and can be extended to model any number of modalities. As far as we know, most of existing cross-modal searching algorithms either

- *Feng Zheng is with the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, S1 4DE, UK.*
- *Yi Tang is with the Key Laboratory of IOT Application Technology of Universities in Yunnan Province and the Department of Mathematics and Computer Science, Yunnan Minzu University, Kunming, Yunnan, 650500, P. R. China.*
- *Ling Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.*
  E-mail: ling.shao@ieee.org.

*Manuscript received * *, 20**; revised * *, 20**.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

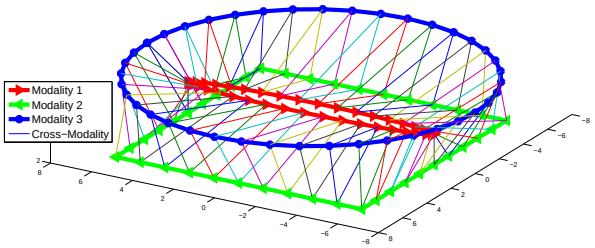

Fig. 1. A hetero-manifold with three modalities: the blue, red and green closed curves represent three uni-modal data sub-manifolds; the lines used to connect two uni-modal data sub-manifolds constitute a cross-modal sub-manifold; all uni- and cross-modal sub-manifolds constitute a hetero-manifold; any change of a uni- or cross-modal sub-manifold will result in a change of the hetero-manifold.

are limited to two modalities [2], [9], [10], [11], [12], [13] or strive to cope with more than two modalities but are still evaluated on the datasets with only two modalities [5], [14], [15].

Given a training set, the inherent similarity of multiple modalities on the hetero-manifold is represented by the hetero-Laplacian matrix. Thus, by minimising the regularisation item via the graph hetero-Laplacian, a set of cross-modal hash functions which are smooth on the hetero-graph can be learned to embed original data points into a Hamming space. In other words, the learned hash functions will preserve the geometrical structure and global supervision information of the hetero-manifold. Meanwhile, a novel weighted cumulative distance inequality on hetero-graph is introduced to cross the gap between Hamming distance and Euclidean distance. By using this novel distance inequality, the problem of learning hash functions is transformed into training a hetero-manifold regularised support vector machine.

In summary, our contributions are four-fold: (1) A novel hetero-manifold is firstly proposed as a well-defined platform to capture both local information of sub-manifolds corresponding to homogeneous data and global information of hetero-manifold corresponding to multi-modality data. (2) A weighted cumulative distance inequality on the hetero-manifold is provided to theoretically guarantee the reasonability of replacing Hamming distance by Euclidean distance during supervised learning. (3) A novel hetero-manifold regularised support vector machine, taking advantages of the hetero-manifold in representing the information of multi-modality data and the support vector machine in generalisation, is proposed based on the proposed weighted cumulative distance inequality for generating more efficient hash functions for cross-modality searching. (4) Extensive experiments on the multi-modality data with six modalities are reported for showing the flexibility of the hetero-manifold regularised support vector machine as more than two modalities are considered.

The rest of this paper is organised as follows. The related work is introduced in Sec. 2. In Sec. 3, constructing a Hetero-manifold for the multi-modal data is detailed. Next, based on this Hereo-manifold, learning a set of hash functions for cross-modal retrieval is presented in Sec. 4. Then, Sec. 5 provides a sequential strategy to solve a complicated objective function. Sec. 6 illustrates comprehensive experimental results for four datasets. Section 7 draws our conclusions.

2 Related work

The cross-modal similarity is generally established by mapping multi-modal data into a common space. The projection based method is motivated by the fact that multi-modal data are used to represent common objects. For example, in [1], a non-linear dimension reduction technique is introduced for cross-modal retrieval, where bimodal data are represented in a common low-dimensional Euclidean space and the cross-modal similarity is defined by using the Euclidean distance in the learned space. Mao *et al.* [2] propose a cross-modal retrieval algorithm based on parallel field alignment in which heterogeneous data are mapped into a common Euclidean space to measure the similarity between heterogeneous data. Deep learning [9], [14], [16] is also employed to learn a common feature space which could be shared by heterogeneous data. Similar to classical discriminant analysis methods, in [3], two pairwise sets (must-link and cannot-link) on the cross-modal samples are considered to learn a similarity function. More references can be found [12], [17], [18], [19], [20], [21], [22].

The Hamming space is more attractive than the Euclidean space because of its efficiency of searching in a large-scale multi-modal dataset [23]. Some existing cross-modal search algorithms, such as [5], [15], [24], adopt an ideal hash coding restriction that heterogeneous data representing common objects share the same hash coding. Others, such as [10], [11], [25], [26], [27], accept a more relaxed hash coding restriction that heterogeneous data representing common objects share similar binary codes which means the Hamming distance of their binary codes, should be small enough. Some other interesting methods could be found [23], [28], [29], [30], [31], [32].

Many works of cross-modal search adopt the manifold concept to model multi-modal data, however, the motivations of constructing the manifold are different. Firstly, multi-modal data are treated as an ensemble of homogeneous data, which are modeled as multiple homogeneous manifolds, such as [2], [5], [6], [33]. For example, Gao *et al.* [33] constructed a similarity graph matrix for each uni-modal feature or label feature, and then learned an optimal similarity graph matrix for the given multi-modal data by fusing the similarity information of uni-modal similarity graph matrices and the label information with semi-supervised learning. Secondly, a cross-modal manifold is constructed whereas uni-modal manifolds are omitted, such as [1]. In [1], Mahadevan *et al.*focused on using covariance between the labels of different modal data to measure the similarity between cross-modal data. Lastly, both uni- and cross-modal manifolds are adopted to model the similarity relation between multi-model data. For example, Masci *et al.* [34] use two uni-modal manifolds and one cross-modal manifold to represent bi-modal data; however, the information of these two uni-modal manifolds cannot be used at the same time because of the usage of gradient based optimization. Zoidi *et al.* [35] employed a high-order similarity matrix (similarity tensor) to represent the similarity information of uni- and cross-modal data. Amiri and Jamzad [36] modeled the similarity information of multi-modal data with a supergraph in which the similarity information of uni-modal data is represented by a subgrahp of the supergraph and the similarity information between cross-modal data is modeled by the connected weights between subgraphs.

Besides manifold-related methods, other techniques are also explored for cross-modal retrieval. For example, Masci *et al.* [34] proposed a novel deep learning framework to simultaneously learn multiple hash functions for preserving multi-modal similarity.

Song *et al.* [37] proposed another deep learning framework for integrating semi-supervised similarity learning and hash function learning. Lai *et al.* [38] proposed deep neural networks for simultaneous feature learning and hash functions learning. Zhu *et al.* [18] proposed a cross-modal dictionary learning framework for representing multi-modal features with common sparse codes. Pereira *et al.* [17] paid more attention on the role of semantic correlation matching in multi-modal retrieval. More references for similarity search on locality sensitive hashing and learning to hash can be found in [39], [40].

The methods, such as [2], [5], [15], [19], support our view that exploiting the manifold structure is very important for boosting the performance of cross-model retrieval. However, no general frameworks for multi-modalities are available, no higher-order relationships have been considered, and, except for CHMIS [5], most existing methods can hardly be extended to more complex multi-modalities. As stated before, in this paper, a general-purpose multi-modal graph embedding framework, which can preserve the uni-modal local structure and cross-modal similarity of high-order random walks, is proposed for cross-modal hashing.

3 Hetero-manifold of multi-modal data

Let $\mathcal{O} = \{O_1, O_2, \cdots, O_N\}$ be a set containing N objects. For the u-th modality, $\mathcal{O}$ is recorded as a $d_u \times N$ matrix X^u where the i-th column vector of X^u, x_i^u corresponds to O_i, $1 \leq u \leq M$, M is the number of modalities, and d_u is the dimension of x_i^u. Generally, the number of modalities is larger than 2, i.e., $M \geq 2$.

A hetero-manifold is an ensemble of uni- and cross-modal sub-manifolds. Uni-modal sub-manifolds are the manifolds whose elements corresponding to different objects share a common modality. For example, X^u is a dataset in which all samples are on the u-th uni-modal sub-manifold. It is clear that uni-modal sub-manifolds are used to represent the intra-structure of uni-modal data. In contrast, cross-modal sub-manifolds serve as bridges to connect different uni-modal data. Ideally, any pair of data points on different uni-modal sub-manifolds could be connected via a path on the cross-modal manifolds and the distance of the path could be used to represent the similarity between the cross-modal data.

Given training samples, the hetero-manifold could be represented as a hetero-graph $\mathbf{G} = (\mathbf{V}, \mathbf{S})$, where $\mathbf{V}$ is the set of vertices and $\mathbf{S}$ is the set of edges. In this paper, $\mathbf{V}$ contains all feature matrices $X^1, X^2, \cdots, X^M$, and the edge between two vertices is defined as the similarity measurement between these two vertices. Following the idea of the hetero-manifold, a hetero-graph could be decomposed into a set of sub-graphs on the homogeneous sub-manifolds and a set of sub-graphs on the cross-modal sub-manifolds. Generally, both sub-graphes could be defined as follows:

Definition 1. ***Uni-modal sub-graph.*** *$G^{uu} = (V^{uu}, S^{uu})$ is a uni-modal sub-graph, if all vertices in this graph come from X^u.*

Definition 2. ***Cross-modal sub-graph.*** *$G^{uv} = (V^{uv}, S^{uv})$ is a cross-modal sub-graph, if, for each edge of this graph, one vertex comes from X^u and the other vertex comes from X^v.*

Definition 3. ***Hetero-graph.*** *$\mathbf{G} = (\mathbf{V}, \mathbf{S})$ is a hetero-graph, if, its vertices correspond to all multi-modal data $X^1, X^2, \cdots, X^M$, and the similarity matrix $\mathbf{S}$ satisfies*

$$\mathbf{S} = \begin{pmatrix} S^{11} & S^{12} & \cdots & S^{1M} \\ S^{21} & S^{22} & \cdots & S^{2M} \\ \cdots & \cdots & \cdots & \cdots \\ S^{M1} & S^{M2} & \cdots & S^{MM} \end{pmatrix}, \tag{1}$$

where $\mathbf{S}(x_i^u, x_j^v) = S^{uv}(x_i^u, x_j^v)$.

Three-order random walks on the hetero-graph is used to model the information diffusion among the vertices on the hetero-graph. For each pair of vertices x_i^u, x_j^v on the hetero-graph, the connection between them consists of three steps: from the end x_i^u to a possible neighbour of x_i^u, from the neighbour of x_i^u to the neighbour of x_j^v, and from the neighbour of x_j^v to the end x_j^v, just like the path shown in Fig. 2. On the one hand, for the first and third steps, the neighbours of the end must be represented in a common modality. Thus, the similarity between x_i^u and its neighbour $x_{i'}^u$ is generally measured by a Gaussian kernel, such as

$$S^u(x_i^u, x_{i'}^u) = \exp\{-\frac{||x_i^u - x_{i'}^u||^2}{\sigma^2}\}, \tag{2}$$

where $\sigma \neq 0$ is a kernel parameter. Similarly, the similarity $S^v(x_j^v, x_{j'}^v)$ between x_j^v and its neighbour $x_{j'}^v$ for the third step can be also defined by the Gaussian kernel.

On the other hand, for the second step, the similarity between $x_{i'}^u$ and $x_{j'}^v$ should be defined according to the different situations of their modalities. If $x_{i'}^u$ and $x_{j'}^v$ share a same modality where $u = v$, the similarity between them could be defined according to their neighborhood relationship, such as:

$$P^{uv}(x_{i'}^u, x_{j'}^v) = \begin{cases} 1, & S^u(i', j') \leq \delta, \\ 0, & S^u(i', j') > \delta, \end{cases} \tag{3}$$

where $\delta \geq 1$ is a parameter for controlling the connection between two points on a uni-graph. Otherwise, if $x_{i'}^u$ and $x_{j'}^v$ are represented in different modalities, the similarity between them should be defined according to the credible priori. For example, the similarity $P^{uv}(x_{i'}^u, x_{j'}^v)$ between $x_{i'}^u$ and $x_{j'}^v$ could be set to be 1 if they correspond to a same object, and set to be 0 otherwise. More meaningful priori depending on a particular task can be used here, such as labels [41], semantic affinities and correlations.

Thus, all possible one-order similarities between the vertices on a uni- or cross-modal sub-graph could be respectively represented by the two kinds of matrices S^u and P^{uv}. Furthermore, in this paper, we assume that the priori matrix P^{uv} satisfies:

$$P^{uv} = (P^{vu})^T. \tag{4}$$

By combining these one-order similarities, the similarity information diffusion model could be defined by a three-order random walk as

$$S^{uv} = S^u P^{uv} S^v. \tag{5}$$

As a special case, the similarity matrix of a uni-modal sub-graph is $S^{uu} = S^u P^{uu} S^u$. The similarity matrix S^{uv} satisfies the following Lemmas.

Lemma 1. ***Non-negativity.*** *The elements of similarity matrix S^{uv} are non-negative.*

Lemma 2. ***Asymmetry.*** *In general, if two matrices S^{uu} and S^{vv} are unequal, S^{uv} is an asymmetric matrix.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

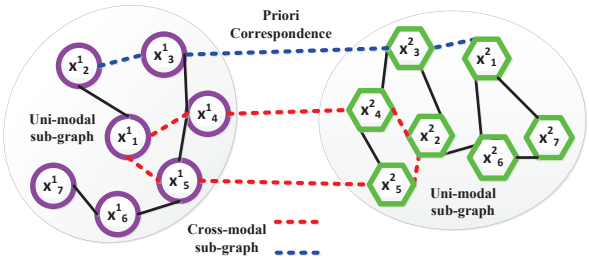

Fig. 2. Cross-modal similarities between features of two objects O_i and O_j captured in two modalities. The lines represent the similarity between two points. The longer the lines, the less similar the two points are. The black lines represent the uni-modal similarity while the dashed lines represent the similarity defined by three-order random walks from one modality to another modality. Among them, we can see that the features x_1^1 and x_2^2 are connected by two red dashed lines whilst the two features x_2^1 and x_1^2 are connected by only one dashed blue line. This point reflects the asymmetry of S^{uv} in Lemma 2.

Lemma 3. ***Equivalence.*** *Any pair of similarity matrices S^{uv} and S^{vu} satisfies the relationship:*

$$S^{uv} = (S^{vu})^T. \tag{6}$$

Therefore, the similarity matrix $\mathbf{S}$ on the hetero-graph satisfies $\mathbf{S} = \mathbf{S}^T$. Lemma 1 is a result of the non-negativeness of Gaussian kernel (2) and the definition of $p(x_{i'}^u, x_{j'}^v)$. Lemma 2 is the result of the definition of matrix multiplication. The proof of Lemma 3 can be found in Appendix A.

Lemma 1 is the theoretical base of learning hash functions on a hetero-manifold. Lemma 2 unveils the intrinsic barrier of treating a multi-modal problem in a cross-modal view because of the asymmetry of both similarity matrices S^{uv} and S^{vu}. Lemma 3 hints the advantages of the global view to understanding multi-modal data as the hetero-manifold because of the symmetry of the similarity matrix on the hetero-manifold $\mathbf{S}$. See Fig. 2 for more details.

4 Hash Function Learning on the Hetero-Manifold

A hetero-manifold integrates multi-modal data into a common manifold, however, a huge gap still exists for efficient cross-modal retrieval because of the difference of different modalities. To this end, a framework of hetero-manifold regularised hash function learning is introduced to embed multi-modal data into a common Hamming space and simultaneously preserve the cross-modal and within-modal similarities on the hetero-manifold.

For the u-th uni-modal data X^u, a set of functions $\mathcal{F}^u = \{f_k^u, 1 \le k \le K\}$ is used to generate the hash codes of X^u, where K is the length of codes. Using these functions $\mathcal{F}^u$, for each sample x_i^u, a vector of real values[1] $F(x_i^u) = (f_1^u(x_i^u), f_2^u(x_i^u), \cdots, f_K^u(x_i^u))^T \in R^K$ can be obtained. Then, a binary code vector y_i^u of x_i^u can be learned by using $y_i^u = (F(x_i^u))_+$, where $(\cdot)_+$ is an operator which sets all positive numbers to 1 and other numbers to 0. Specifically, we have the k-th element of y_i^u:

$$y_i^u(k) = (f_k^u(x_i^u))_+. \tag{7}$$

4.1 Distance inequality on a graph

In general, learning to hash tries to minimise a cumulative Hamming distance with some constraints. If the distance is defined on a manifold, then a weighted cumulative Hamming distance $\mathcal{L}_c^h(\mathbf{G})$ should be minimised.

$$\mathcal{L}_c^h(\mathbf{G}) = \sum_{u,v=1}^{M} \sum_{i,j=1}^{N} S^{uv}(x_i^u, x_j^v) \mathcal{D}_h(y_i^u, y_j^v), \tag{8}$$

where $\mathcal{D}_h(y_i^u, y_j^v)$ is the Hamming distance between y_i^u and y_j^v. Actually, the weights between the samples embody the intrinsic structures and useful information including local neighbourhood, prior semantic cues and affinities. By considering these weights, the original structure and information can be preserved in a new learned space. In this paper, the weights reflect the information contained in the hetero-manifold.

Meanwhile, besides the Hamming distance, for any pair of points x_i^u and x_j^v on graph $\mathbf{G}$, an accompanied Euclidean distance can be defined as $\mathcal{D}_e(F(x_i^u), F(x_j^v)) = ||F(x_i^u) - F(x_j^v)||_2^2$. Same as Hamming distance, a weighted cumulative Euclidean distance on graph $(\mathbf{G}, \mathbf{S})$ is given as:

$$\mathcal{L}_c^e(\mathbf{G}) = \sum_{u,v=1}^{M} \sum_{i,j=1}^{N} S^{uv}(x_i^u, x_j^v) \mathcal{D}_e(F(x_i^u), F(x_j^v)). \tag{9}$$

Normally, during the matching stage, the Hamming distance is far less computationally expensive than the Euclidean distance. However, despite the simplicity in Eq. 8, minimisation of the Hamming distance is generally intractable, because it is a concrete quantity. Thus, we seek to minimise an alternative item, which guarantees that the Hamming distance will be minimised simultaneously.

First, a constraint [42] will be given as follows:

Definition 4. ***Hinge loss constraint.*** *For a function f_k^u in the u-th modality, if any point x_i^u captured in this modality and its corresponding hash code defined in Eq. 7 satisfies*

$$y_i^u(k) f_k^u(x_i^u) \ge 1 - \xi_{ik}^u, \tag{10}$$

where ξ_{ik}^u is a minimal non-negative value, thus f_k^u is the hinge loss constraint-satisfied function in the u-th modality.

Next, under the above constraint, a distance inequality in the following can be obtained:

Lemma 4. ***Distance inequality.*** *If two sets of functions $\mathcal{F}^u$ and $\mathcal{F}^v$ are the hinge loss constraint-satisfied functions in modalities u and v respectively, for any two samples x_i^u and x_j^v, the two types of distance in the learned Hamming space and the Euclidean space have the following relationship, when satisfying $\forall k, \xi_{ik}^u + \xi_{jk}^v \le 1$:*

$$\mathcal{D}_h(y_i^u, y_j^v) \le \mathcal{D}_e(F(x_i^u), F(x_j^v)), \tag{11}$$

where $\mathcal{D}_h$ and $\mathcal{D}_e$ are defined in Eq. 8 and 9, respectively.

It is worth to point out that f_k^u is a hinge loss constraint-satisfied function only when all the samples in modality u satisfy condition 10. And Eq. 11 can be proved, when a condition $\forall k, \xi_{ik}^u + \xi_{jk}^v \le 1$ is given. We can see that ξ_{ik}^u and ξ_{jk}^v are

1. For simplicity, $F(x_i^u) = F^u(x_i^u)$ without confusion.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

two minimal non-negative values in the definition of the hinge loss constraint. If the two modalities are the same ($u = v$), the same inequality can be established for any two samples captured in the same modality.

Then, based on the condition 10, we can extend the inequality 11 to a weighted cumulative distance inequality on a graph.

Corollary 1. ***Weighted distance inequality.*** *For a graph $\mathbf{G} = (\mathbf{V}, \mathbf{S})$, if two sets of functions $\mathcal{F}^u$ and $\mathcal{F}^v$ satisfy the condition in Eq. 10, thus the following weighted cumulative distance inequality can be established, when $\mathbf{S}$ is a similarity matrix with non-negative members:*

$$\mathcal{L}_c^h(\mathbf{G}) \leq \mathcal{L}_c^e(\mathbf{G}). \tag{12}$$

Consequently, with the help of the inequality in the Corollary 1, a relaxed optimisation problem which will be introduced in the following section can be generated. In this paper, we will consider to learn linear hash functions via minimising the upper bound $\mathcal{L}_c^e(\mathbf{G})$ of the cumulative Hamming distance $\mathcal{L}_c^h(\mathbf{G})$. In fact, Corollary 1 is a direct result of Lemma 4. More proof details of Lemma 4 are provided in the Appendix B.

4.2 Objective function

Specifically, the binary codes of x_i^u are defined by linear functions as $y_i^u = (((w_1^u)^T x_i^u)_+, ((w_2^u)^T x_i^u)_+, \cdots, ((w_K^u)^T x_i^u)_+)^T = ((W^u)^T x_i^u)_+$, where W^u is a matrix whose k-th column vector is w_k^u. Then, for the u-th uni-modal dataset X^u, the corresponding binary code set is $Y^u = ((W^u)^T X^u)_+$, in which the i-th column y_i^u is the binary code vector of x_i^u.

Furthermore, denote projection matrix

$$\mathbf{W}^T = ((W^1)^T, (W^2)^T, \cdots, (W^M)^T), \tag{13}$$

and multi-modal data matrix

$$\mathbf{X} = \begin{pmatrix} X^1 & 0 & \cdots & 0 \\ 0 & X^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & X^M \end{pmatrix}. \tag{14}$$

Thus, the binary codes can be obtained:

$$\mathbf{Y} = (\mathbf{W}^T \mathbf{X})_+. \tag{15}$$

Using $Y^u = ((W^u)^T X^u)_+$, it is easy to prove that $\mathbf{Y} = (Y^1, \cdots, Y^u, \cdots, Y^M)$. Meanwhile, using Eq. 13 and 14, the cumulative Euclidean distance $\mathcal{L}_c^e(\mathbf{G})$ can be rewritten as

$$\mathcal{L}_c^e(\mathbf{G}) = 2\mathbf{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}), \tag{16}$$

where Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$, $\mathbf{D} = \mathbf{diag}(d_{11}, d_{12}, \cdots, d_{ui}, \cdots, d_{MN})$ and $d_{ui} = \sum_{v,j} \mathbf{S}(x_i^u, x_j^v)$. In this paper, $\mathbf{diag}$ is an operator to generate a diagonal matrix. The detailed proof of Eq. 16 is given in Appendix C.

With the hinge loss constraint, the problem of hash function learning on hetero-manifold (8) could be approximated by minimising its upper bound (16) with some constraint conditions:

$$\begin{aligned} \mathbf{W}^* &= \arg\min_{\mathbf{W}} \frac{1}{2}\mathbf{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}) \\ s.t. \quad & \forall u, i, k \\ & (i)\ y_i^u(k)(w_k^u)^T x_i^u \geq 1 - \xi_{ik}^u,\ \xi_{ik}^u \geq 0, \\ & (ii)\ \xi_{ik}^u + \xi_{jk}^u \leq 1, \\ & (iii)\ \mathbf{W}^T \mathbf{W} = I, \end{aligned} \tag{17}$$

where ξ_{ik}^u is a slack variable. The first and second constraint conditions which are from Lemma 4 ensure Euclidean distance based loss $\mathcal{L}_c^e(\mathbf{G})$ be the upper bound of the Hamming distance based loss $\mathcal{L}_c^h(\mathbf{G})$. The third constraint condition corresponds to the requirement of orthogonality between two hash functions.

To further simplify the optimisation problem (17), the last two constraint conditions are slightly relaxed and transferred into the objective function by using the Lagrangian principle. The constraint condition (iii) will be considered when the projections are learned using a sequential strategy. As for constraint condition (ii), the total number of pairs ξ_{ik}^u, ξ_{jk}^u is $\frac{M^2 N^2 K}{2}$ because of the structure of the hetero-graph, and each ξ_{ik}^u exists in MN constraint conditions. Thus all of these constraint conditions can be summed up and the conditions will be relaxed as

$$\sum_{u=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{ik}^u \leq \frac{MNK}{2}. \tag{18}$$

Therefore, the original optimisation problem (17) is transformed by replacing the constraint conditions (ii) with the relaxed constraint conditions (18) and using the Lagrangian principle into

$$\begin{aligned} \mathbf{W}^* &= \arg\min_{\mathbf{W}} \frac{1}{2}\mathbf{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}) \\ & \quad + C_1 \sum_{u=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{ik}^u \\ s.t. \quad & \forall u, i, k \\ & (i)\ y_i^u(k)(w_k^u)^T x_i^u \geq 1 - \xi_{ik}^u,\ \xi_{ik}^u \geq 0 \\ & (ii)\ \mathbf{W}^T \mathbf{W} = I, \end{aligned} \tag{19}$$

where $C_1 > 0$ is the regularisation parameter.

It should be noticed that the Laplacian matrix $\mathbf{L}$ depends on all uni- and cross-modal similarity matrices because any sole sub-matrix used to define the similarity matrix $\mathbf{S}$, for example S^{uv}, is not enough for defining the counterpart sub-matrix of $\mathbf{L}$. It implies that the Laplacian matrix contains the global information of the hetero-manifold. Therefore, the optimisation problem (19) is a hetero-manifold regularised hash function learning problem.

5 Sequential Optimisation

In order to solve the problem in Eq. 19, we first divide it into sub-problems, in each of which only one projection for the k-th code is considered. Thus, in Eq. 15, the k-th row vector $\mathbf{y}_k$ of $\mathbf{Y}$ is a binary vector which corresponds the k-th bits of all samples in all modalities while the corresponding k-th column vector of $\mathbf{W}$ is denoted as $\mathbf{w}_k$. Then, we have

$$\mathbf{y}_k = (\mathbf{w}_k^T \mathbf{X})_+, \tag{20}$$

where the vector $\mathbf{w}_k^T = ((w_k^1)^T, (w_k^2)^T, \cdots, (w_k^M)^T)$.

Although these sub-problems are not independent with each other, they are convex when all the other variables are fixed. The convexity will be reflected by the standard quadratic programming problems in the following Eq. 21 and 23. Hence, the optimisation problem (19) could be resolved bit by bit in a sequential way. A similar work of sequential learning could be found in [43], when the sub-problems can be solved by a direct eigen-decomposition. In this paper, more specifically, the local optimal solution $\mathbf{W}^*$ is learned by sequentially optimising each of its column vectors $\mathbf{w}_k^*, k = 1, 2, \cdots, K$. For distinguishing the iterations of optimisation, the τ-th $\mathbf{W}^*$ and $\mathbf{w}_k^*$ are denoted as $\mathbf{W}^{(\tau)}$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

$\mathbf{w}_k^{(\tau)}$, respectively. In round τ, before solving the sub-problem, the binary codes $\mathbf{y}_k^{(\tau-1)}$ should be initiated using codes in the last round or generated randomly.

5.1 The first hash function learning

To train the hash functions, the hash codes $\mathbf{y}_1^{(0)}$ will be randomly initialised in the first round when $\tau = 1$. Then, $\mathbf{w}_1^{(1)}$ could be learned from the optimisation problem

$$\begin{aligned} \mathbf{w}_1^{(1)} &= \arg\min_{\mathbf{w}_1} \frac{1}{2}\mathbf{w}_1^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}_1 \\ &+C_1\sum_{u=1}^{M}\sum_{i=1}^{N}\xi_{i1}^u \\ s.t. \quad & \forall u,i,\ y_i^u(1)(w_1^u)^T x_i^u \geq 1-\xi_{i1}^u,\ \xi_{i1}^u \geq 0. \end{aligned} \tag{21}$$

The optimisation problem (21) is derived from the problem (19) where the orthogonal constraint condition becomes zero because it is assumed that $\mathbf{w}_1^{(1)}$ is orthogonal with the other projection directions $\mathbf{w}_k^{(1)}, k = 2, 3, \cdots, K$ without any information about $\mathbf{w}_k^{(1)}, k = 2, 3, \cdots, K$.

It is clear that the optimisation problem (21) is convex. Meanwhile, the Lagrange dual of the optimisation problem (21) is a problem of quadratic programming. Therefore, the optimal $\mathbf{w}_1^{(1)}$ could be defined as

$$\mathbf{w}_1^{(1)} = (\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{X}_{\mathbf{y}_1}^{(0)}\alpha_1^{(1)}, \tag{22}$$

where $\mathbf{X}_{\mathbf{y}_1}^{(0)} = \mathbf{diag}(X_{\mathbf{y}_1}^1, \cdots, X_{\mathbf{y}_1}^M)$[2], the matrix $X_{\mathbf{y}_1}^u = (y_1^u(1)x_1^u, \cdots, y_N^u(1)x_N^u)$, and $\alpha_1^{(1)}$ is the result of the Lagrange dual problem of (21). $y_i^u(1)$ is the initial bit from $\mathbf{y}_1^{(0)}$ for object O_i in the u-th modality.

5.2 The following hash function learning

Given $\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \cdots, \mathbf{w}_{k-1}^{(1)}$, the next optimal projection $\mathbf{w}_k^{(1)}$ could be defined via the following optimisation problem

$$\begin{aligned} \mathbf{w}_k^{(1)} &= \arg\min_{\mathbf{w}_k} \frac{1}{2}\mathbf{w}_k^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}_k \\ &+\frac{C_2}{2}\mathbf{w}_k^T\mathbf{Q}_k^{(1)}\mathbf{w}_k + C_1\sum_{u=1}^{M}\sum_{i=1}^{N}\xi_{i,k}^u \\ s.t. \quad & \forall u,i,\ y_i^u(k)(w_k^u)^T x_i^u \geq 1-\xi_{i,k}^u,\ \xi_{i,k}^u \geq 0, \end{aligned} \tag{23}$$

where $C_2 > 0$ is a regularisation parameter, and $\mathbf{Q}_k^{(1)} = \sum_{l=1}^{k-1}\mathbf{w}_l\mathbf{w}_l^T$ which is used to measure the orthogonality between $\mathbf{w}_k$ and the other learned $\mathbf{w}_l,\ l = 1, 2, \cdots, k-1$. It is clear that

$$\mathbf{w}_k^T\mathbf{Q}_k^{(1)}\mathbf{w}_k = \sum_{l=1}^{k-1}(\mathbf{w}_k^T\mathbf{w}_l)^2, \tag{24}$$

where $\mathbf{w}_k^T\mathbf{w}_l$ defines the linear correlation between $\mathbf{w}_k$ and $\mathbf{w}_l$. By minimising the term $\mathbf{w}_k^T\mathbf{Q}_k^{(1)}\mathbf{w}_k$, the learned projection direction $\mathbf{w}_k^{(1)}$ will be approximatively orthogonal to all of the other learned projection directions. Similar to formula (22), the optimisation problem (23) could also be resolved by using the Lagrange dual method

$$\mathbf{w}_k^{(1)} = (\mathbf{X}\mathbf{L}\mathbf{X}^T + C_2\mathbf{Q}_k^{(1)})^{-1}\mathbf{X}_{\mathbf{y}_k}^{(0)}\alpha_k^{(1)}, \tag{25}$$

where $\alpha_k^{(1)}$ is the result of Lagrange dual of optimisation problem (23) and $\mathbf{X}_{\mathbf{y}_k}^{(0)}$ will be updated according to the binary vector $\mathbf{y}_k^{(0)}$.

When $\mathbf{W}^{(1)}$ is learned according to the formulas (21) and (25), the following $\mathbf{W}^{(\tau)}, \tau = 2, 3, \cdots, t$ could be learned by using a similar objective function. The differences to problem (21) are the definition of the orthogonal item:

$$\mathbf{Q}_k^{(\tau)} = \sum_{l\neq k}\mathbf{w}_l^{(\tau-1)}(\mathbf{w}_l^{(\tau-1)})^T - \mathbf{w}_k^{(\tau-1)}(\mathbf{w}_k^{(\tau-1)})^T,$$

and, according to the bits learned in the last round $\mathbf{y}_k^{(\tau-1)}$, the quantity $\mathbf{X}_{\mathbf{y}_k}^{(\tau-1)}$ should be also updated. Similarly, the optimal result $\mathbf{w}_k^{(\tau)}$ could be represented as

$$\mathbf{w}_k^{(\tau)} = (\mathbf{X}\mathbf{L}\mathbf{X}^T + C_2\mathbf{Q}_k^{(\tau)})^{-1}\mathbf{X}_{\mathbf{y}_k}^{(\tau-1)}\alpha_k^{(\tau)}. \tag{26}$$

The objective functions in Eq. 21 and 23 can be considered as a general dual problem[3], when we define $H = \mathbf{X}\mathbf{L}\mathbf{X}^T + C_2\mathbf{Q}^{(1)}$. Thus, the optimal solution can be obtained by a Representation Theory in Appendix D. Therefore, all of these steps of optimising the original optimisation problem (19) can be summarised in Algorithm 1.

Algorithm 1 Hetero-manifold Regularised Hashing (HMR)

Input: Dataset $\{X^1, \cdots, X^M\}$, parameters C_1, C_2, the number of iterations t and the length of hash coding vector K.
Output: $\mathbf{W}^t$.
Initialisation
(0) Construct matrix $\mathbf{S}$ according to Eqs. (2), (5), and (1).
(1) Construct Laplacian graph $\mathbf{L}$ according to Eq. (16).
(2) Randomly initiate the binary codes $\mathbf{y}_1^{(0)}$ and calculate $\mathbf{X}_{\mathbf{y}_1}^{(0)}$.
(3) Generate the first projection $\mathbf{w}_1^{(1)}$ according to Eq. (22).
For $k = 2, \cdots, K$
  (4) Randomly initiate the binary codes $\mathbf{y}_k^{(0)}$.
  (5) Calculate $\mathbf{Q}_k^{(1)}$ and $\mathbf{X}_{\mathbf{y}_k}^{(0)}$.
  (6) Generate $\mathbf{w}_k^{(1)}$ according to Eq. (25).
  (7) Update $\mathbf{W}^{(1)}$ and $\mathbf{y}_k^{(1)}$ using Eq. 20.
End
For $\tau = 2, \cdots, t$
  For $k = 1, \cdots, K$
    (8) Calculate $\mathbf{Q}_k^{(\tau)}$ and $\mathbf{X}_{\mathbf{y}_k}^{(\tau-1)}$.
    (9) Generate the k-th projection $\mathbf{w}_k^{(\tau)}$ according to Eq. (26).
    (10) Update $\mathbf{W}^{(\tau)}$ and $\mathbf{y}_k^{\tau}$ using Eq. 20.
  End
End
Return

6 Experiments

The proposed HMR is validated on four recent public datasets: the VIPeR [44] and CUHK01 [45] datasets for cross-camera person re-identification, the Wiki dataset [17] for cross-modal retrieval and the FG-NET ageing dataset [46] for cross-age face image retrieval where the number of modalities is 6. Four state-of-the-art cross-modal binary code learning methods, including PDH [10], CVH [15], CMSSH [24] and CMFH [11], are mainly compared with and some other area-specific methods are also used for comparative analysis in our experiments.

Evaluation Metrics: On the one hand, for identification systems, the Cumulated Matching Characteristics (CMC) [47] are commonly used for performance evaluation and measuring how well an identification system ranks the identities in the gallery

2. Without confusion, the subscript $\mathbf{X}_{\mathbf{y}_1^{(0)}}$ will be simplified as $\mathbf{X}_{\mathbf{y}_1}^{(0)}$.

3. In the case of Eq. 21, the parameter can be set to $C_2 = 0$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

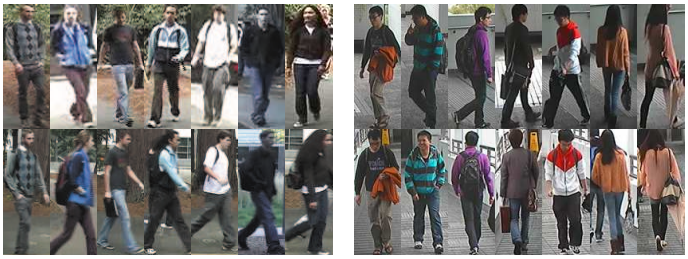

Fig. 3. Some image examples of the two person re-identification datasets: VIPeR (left) and CUHK01 (right).

Method	R1	R5	R10	R15	R20	AUC
HMR	**0.299**	**0.590**	**0.729**	**0.826**	**0.880**	**0.897**
CMFH	0.247	0.528	0.712	0.766	0.816	0.871
PDH	0.171	0.449	0.604	0.693	0.778	0.822
CMSSH	0.190	0.437	0.639	0.725	0.791	0.831
CCA	0.168	0.427	0.551	0.633	0.693	0.776
CVH	0.085	0.209	0.294	0.345	0.399	0.551

TABLE 1
Ranking accuracy comparison at ranks 1, 5, 10, 15 and 20 and overall AUC performance comparison when 512 dimensional binary codes are learned. R1 denotes Rank 1.

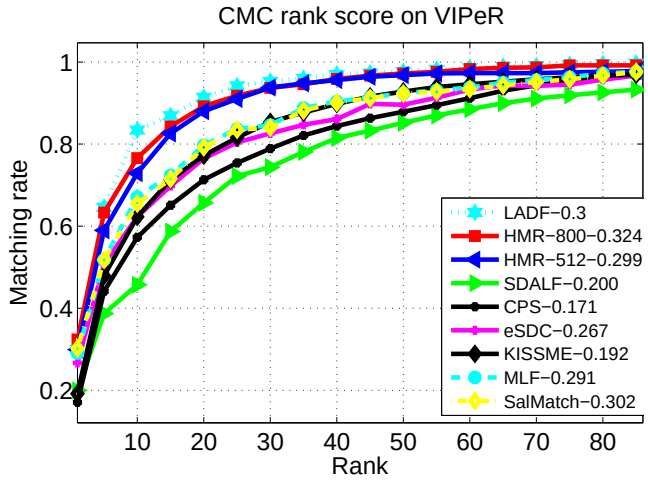

Fig. 4. The CMC rankings of the compared methods on the VIPeR dataset with #316 test persons. Numbers in legend are the Rank-1 accuracies and HMR-512 means the length of learned codes of HMR is 512.

with respect to a probe sample. Moreover, the Area Under Curve (AUC) corresponding to the CMC curves is also reported to show the overall performance at ranks from 1 to a fixed maximum. A larger AUC score means the corresponding method is more robust. On the other hand, for the ranking cases of multiple feedbacks, the precision and recall are normally calculated:

$$precision = \frac{|\Im \cap \Re|_{\#}}{|\Re|_{\#}}, \quad recall = \frac{|\Im \cap \Re|_{\#}}{|\Im|_{\#}},$$

where $\Re$ is a set of retrieved samples, $\Im$ is a set of relevant samples and $|\cdot|_{\#}$ denotes the size of the set. Precision-Recall (PR) curves [48] which are often used in information retrieval are used to measure performance in cross-modal retrieval. By varying the similarity measurement between the pair of retrieved samples (Hamming distance in this paper) and evaluating the precision, recall and the number of retrieved points accordingly, PR curves can be obtained. Furthermore, Mean Average Precision (MAP) [11], which is the average precision at the ranks where recall changes, is generally used to evaluate a ranking system.

6.1 Cross-camera re-identification

Cross-camera person re-identification is a very challenging task because of the variation of camera views and the environment. Given a probe image containing a person, the most popular method of recognising the person is to rank the similarities between the probe image and the images in the gallery (captured by other cameras). In this experiment, the similarity is calculated in the learned Hamming space across the cameras and the maximum rank of AUC is 85.

VIPeR: This dataset contains 632 pedestrian image pairs in an outdoor environment. Each pair contains two images of the same individual taken from two different camera views. Changes of viewpoint, illumination and pose are the most significant causes of appearance change. Each image has been scaled to be 128×48 pixels. Some example images in VIPeR are shown in Fig. 3 (Left). The experimental setting is the same as [49]. Half of the dataset including 316 images for each view is used for training the algorithms and the remaining (316 pedestrian) is used for testing.

CUHK01: Two cameras setting in different places of a campus environment are used to collect the samples. Camera A captures the frontal view or back view of pedestrians, while camera B captures the side view. This dataset contains 971 persons, each of which has two images. Some example images in CUHK01 are shown in Fig. 3 (Right). All the images are normalised to 160×60 for evaluations. The experimental setting is the same as [50] where 486 persons are chosen for testing and the remaining persons for training.

In this experiment, the Local Maximal Occurrence Feature (LOMO) which was proposed in [51] is used. The original dimension of the LOMO feature is 26960 and then is reduced to 70 as suggested by [51]. In this experiment, the parameters C_1 and C_2 of Algorithm 1 are set to 20 and 2, respectively. All the results are reported by averaging 10 runs.

To compare the performance with the state-of-the-art person re-identification methods, we evaluate the proposed HMR and the recently published algorithms on the VIPeR dataset including: SDALF [52], CPS [53], KISSME [54], eSDC [55], SalMatch [56], MLF [50] and LADF [57]. For the proposed HMR, two lengths of binary codes 512 and 800 have been learned and the experimental results corresponded to both code lengths are denoted as HMR-512 and HMR-800, respectively. The comparison results are shown in Fig. 4. Firstly, we can see that, except for LADF, HMR (HMR-512 and -800) significantly outperforms other methods and the advantages are more obvious especially at higher ranks (from 5 to 60). It is worth to point out that HMR is the only hashing method among the compared ones and still achieves comparative results to a non-hashing metric learning method LADF. In fact, due to quantisation loss, the performance of hashing methods is normally lower than that of non-hashing methods in many applications. Secondly, HMR-512 achieves similar results as HMR-800 and this demonstrates that the performance keeps stable when the code length is above a certain threshold. Finally, we also compare with other hashing methods on the VIPeR dataset when the binary code length is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

fixed at 512[4] and the comparison results are illustrated in Table 1. We can see that, both from the perspectives of ranks 1, 5, 10, 15 and 20 and the overall performance AUC, HMR achieves much better results than state-of-the-art hashing methods.

To further compare with other hashing methods, binary codes of shorter lengths (32, 64 and 128) are learned on the CUHK01 dataset. The results are shown in Fig. 5 and Table 2. We can observe that, as the code length increases, the performance of eigenvalue decomposition based methods such as CVH decreases since the first few projection directions occupy most of variances. However, it is reasonable that our HMR can achieve better when the code length increases. More information can be kept because HMR considers both the orthogonality and the cross-modal intrinsic structure. We can see that HMR achieves best results at all code lengths. Specifically, the advantages of HMR are more obvious, when the length of learned codes increases. The rank 1 scores of the five methods are also shown in the legend of Fig. 5 and HMR obtains at least 0.024 higher scores than other methods.

Method	CVH	CMFH	PDH	CMSSH	HMR
32 bits	45.66	65.39	58.96	54.21	**67.99**
64 bits	38.47	65.37	66.59	54.78	**69.36**
128 bits	30.13	67.03	69.29	55.15	**72.14**

TABLE 2
AUC Comparison on CUHK01 corresponding to the curves in Fig. 5.

6.2 Cross-modal retrieval

Images and texts are the two popular modalities for testing cross-modal retrieval methods. There are several datasets available but Wiki is the most popular one. Thus, in this experiment, the Wiki [17] dataset is used for our evaluations.

Wiki: It is generated from the "Wikipedia featured articles" and consists of 2866 image-text pairs in 10 most populated categories. The texts are represented by 10 dimensional latent Dirichlet allocation model and each image has a 128 dimensional SIFT histogram feature. We follow the data partition adopted in [17] to split the dataset into a training set of 2173 pairs and a test set of 693 pairs. In our setting, both gallery and query samples are from the test set which is different to the setting in [11]. In [11], the gallery samples are from the training set and thus their retrieval results are better than ours. If the query comes from the test set, then the samples in the text test set will be considered as the database and vice versa. In this experiment, the parameters C_1 and C_2 of Algorithm 1 are set as 30 and 1.2, respectively. The number of retrieved instances is set to $|\Re|_{\#} = 50$.

The MAP results on the test set are shown in Table 3. The same phenomenon of performance reduction as the code length increases for the eigenvalue decomposition based methods can be also observed on Wiki. From Table 3, we can see that HMR outperforms the state-of-the-art methods at code lengths 32 and 64, and achieves very close scores to the best method at code length 16. Moreover, the Precision-Recall (PR) curves on the Wiki dataset, which are obtained by varying the Hamming distance between the query points and the retrieved points, are reported in Fig. 6. HMR can obtain higher scores for almost all the Hamming radii from 1 to the maximum at code lengths 32 and 64 and get a similar PR curve to the best one at code length 16. Finally, MAP performance on each category is shown in Fig. 7. The retrieval difficulties of the 10 categories to the five methods are similar and three of them, i.e., Biology, Geography and Warfare, seem to be more easily classified. From Fig. 7, we can see that HMR is more robust on different categories over other methods. Very recently, deep neural networks were also exploited for multi-modal hashing [16] or cross-modal hashing [14] and achieved more advanced results than some other types of methods. However, the complexity of code generation in deep neural networks is generally much higher than that in linear functions. Take the model of layers $100-256-128-64-32-32$ in [16] for example, the number of multiplications is 68608 times of that in the corresponding linear function. Nevertheless, the capacity of hypothesis space and the non-linearity exploited in deep learning make it feasible to perform better in most cases. Thus, this motivates us to, besides exploring the structure of the hetero-manifold, rewrite the proposed framework in Reproduced Kernel Hilbert Space [58] in the future to capture the non-linearity as well.

Task	Method	16 bits	32 bits	64 bits
Image Query	CVH	0.2021	0.1668	0.1723
	CMSSH	0.2276	0.1940	0.1982
	PDH	0.1885	0.1796	0.2086
	CMFH	**0.2583**	0.2567	0.2691
	HMR	0.2503	**0.2621**	**0.2833**
Text Query	CVH	0.2560	0.1902	0.2019
	CMSSH	0.2483	0.2431	0.2505
	PDH	0.2309	0.2278	0.2279
	CMFH	**0.3192**	0.3347	0.3351
	HMR	0.3151	**0.3408**	**0.3511**

TABLE 3
MAP Comparison on Wiki.

6.3 Cross-age face retrieval

In this section, we validate the proposed HMR on a more challenging task: cross-age face retrieval. Given a probe face image, we need to search for the face images of the same person but captured in different age stages. This task is derived from age estimation [59] but it is more difficult and novel because: 1) The principal characteristics of the face appearance of a same person vary hugely along with the variation of his or her age. 2) The capturing conditions of images are quite diverse in different places and years. 3) As far as we know, the cross-age face retrieval is the first multi-modal experiment, in which 6 modalities are considered. Intuitively, the ages of faces can be considered as modalities in our setting, in which faces of different persons with the same age range share similar characteristics including smoothness, wrinkles and hair.

FG-NET: Some examples of an ageing dataset [46], which contains 82 people with age ranges from 0 to 69, are shown in Fig. 8. The images of a same person distribute unevenly and most of the images are captured in the early ages. Thus, we divide the ages into 6 stages including $0-4$, $5-9$, $10-14$, $15-19$, $20-30$ and $31-69$ which correspond to 6 modalities in our method. In this experiment, the parameters C_1 and C_2 of Algorithm 1 are set to 10 and 0.1, respectively. 10-fold cross validation is used and, in each fold, 90% persons will be chosen as training and the remaining as for testing. In this experiment, the maximum value for AUC is set to 50.

4. Because of the limitation of covariance, CVH and CCA cannot learn functions with a number exceeding the rank of the matrix. Thus, best results are reported at a certain length.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

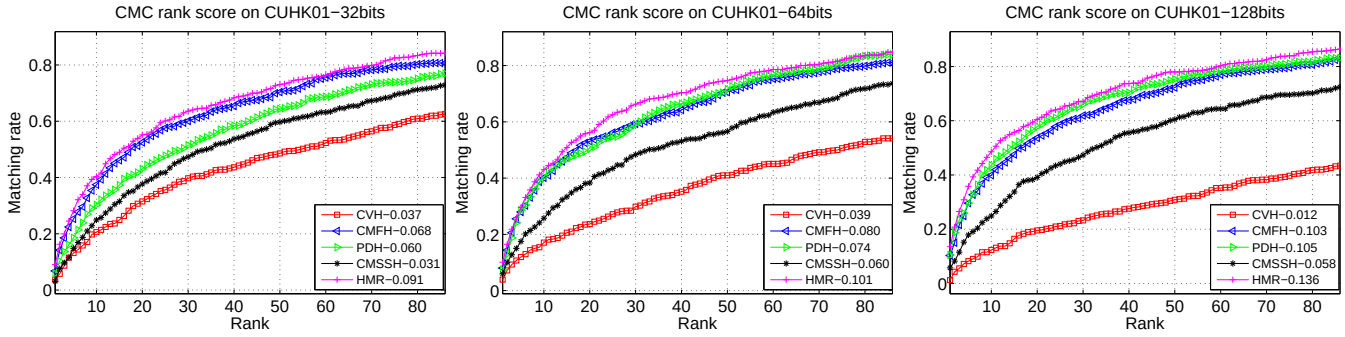

Fig. 5. The CMC rankings of five methods on the CUHK01 dataset at code lengths 32, 64 and 128 with 486 test persons.

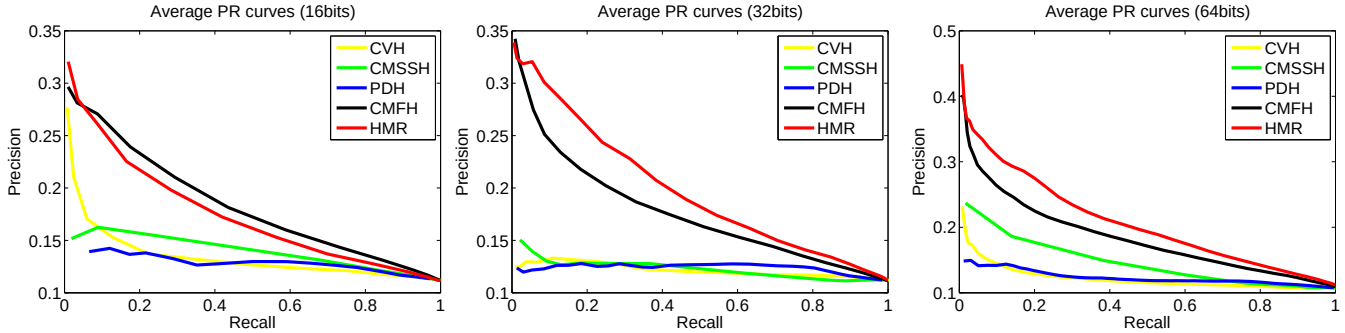

Fig. 6. Precision recall curves on Wiki by varying the Hamming distance.

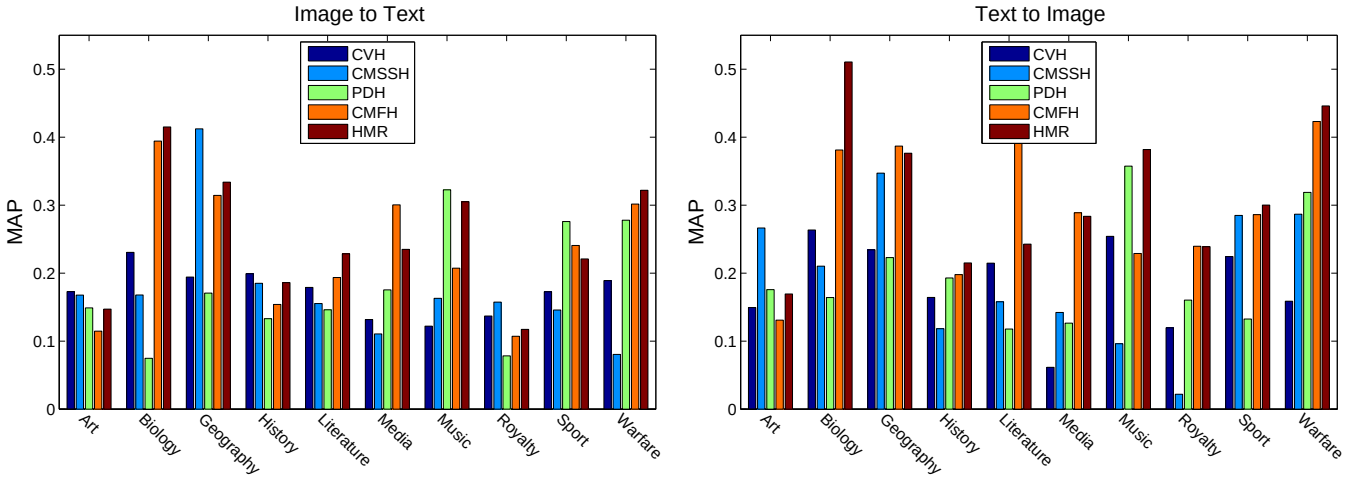

Fig. 7. MAP performance for each category at 32 bits.

Firstly, same as most age estimation works, features are directly extracted based on the 64 landmarks offered by the FG-NET dataset. For each landmark, a simple descriptor GIST [60] is used for representing a fixed rectangle (19×19) around it and then a feature for a face image can be constructed by concatenating the features of all landmarks. Principal Component Analysis (PCA) is adopted to reduce the feature into a space with 255 dimensions. Secondly, it is worth to point out that the number of images of a same person differs significantly for different age stages. Thus, compared to person re-identification and cross-modal retrieval, the task becomes more difficult because the correspondence matrix between two modalities is not diagonal. For some methods such as CMFH, the optimisation is not even technically correct. By duplicating the samples of a same person, a diagonal correspondence matrix can be obtained. Moreover, except for our HMR, no existing methods can directly tackle multiple modalities with an inconsistent number of samples or features. To compare with these methods, any two modalities will be considered as the input of the two-modality methods. Taking the PDH, CMFH, CCA, CMSSH and CVH as examples, these methods will be trained 15 times for cross-age face retrieval and, for each modality, 5 different groups of projections will be obtained. This demonstrates that our proposed HMR is very powerful and flexible to deal with different tasks without particular limitations and the hash functions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

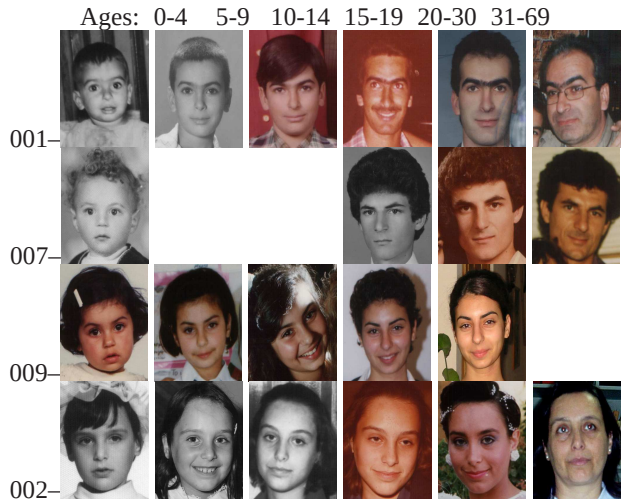

Fig. 8. Some image examples of the FG-NET dataset. For person 007, the dataset contains no image samples with age range 5-14.

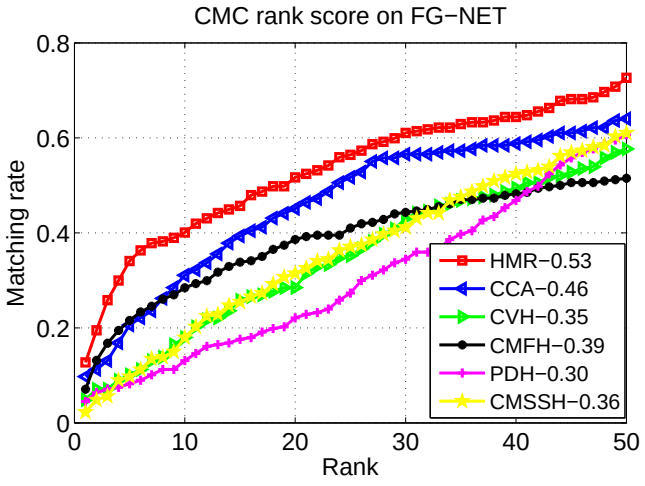

Fig. 9. Overall performance comparison between the proposed HMR, CCA and other state-of-the-art methods. The number in the legend is the Area Under Curve (AUC) and the possible largest AUC can be up to 1.

for different modalities can be obtained simultaneously by one optimisation.

Modalities	0-4	5-9	10-14	15-19	20-30	31-69
0-4	–	0.108	0.102	0.059	0.043	0.000
5-9	0.248	–	0.216	0.179	0.050	0.050
10-14	0.220	0.265	–	0.134	0.125	0.163
15-19	0.096	0.220	0.162	–	0.149	0.304
20-30	0.120	0.102	0.055	0.141	–	0.322
31-69	0.033	0.113	0.132	0.125	0.103	–

TABLE 4
Rank 1 performance of cross-age retrieval on the FG-NET face dataset with 6 modalities.

The overall performance comparison of cross-age face retrieval is given in Fig. 9 and the different methods are ranked according to the Area Under Curve (AUC). From this figure, we can see that the proposed method consistently outperforms other methods at all ranks. Moreover, we can conclude that non-hashing method CCA achieves better results than other hashing-based methods. Furthermore, compared to the above experiments of two modalities, the advantages of the proposed HMR are more obvious in this experiment. The substantial reason is that the information can be propagated on the proposed Hetero-manifold and then supervises the learning of hash functions. However, most state-of-the-art methods are specially designed for two modalities and, in the multi-modal cases ($M > 2$), to some extent, the global information is ignored.

Modalities	0-4	5-9	10-14	15-19	20-30	31-69
0-4	–	0.284	0.216	0.151	0.085	0.111
5-9	0.537	–	0.437	0.358	0.400	0.250
10-14	0.515	0.565	–	0.387	0.328	0.490
15-19	0.346	0.488	0.414	–	0.460	0.536
20-30	0.337	0.367	0.233	0.424	–	0.589
31-69	0.333	0.340	0.374	0.347	0.370	–

TABLE 5
Rank 10 performance of cross-age retrieval on the FG-NET face dataset with 6 modalities.

Modalities	0-4	5-9	10-14	15-19	20-30	31-69
0-4	–	0.319	0.282	0.168	0.106	0.148
5-9	0.578	–	0.477	0.421	0.475	0.350
10-14	0.556	0.604	–	0.465	0.391	0.571
15-19	0.394	0.549	0.485	–	0.506	0.565
20-30	0.361	0.408	0.301	0.515	–	0.633
31-69	0.400	0.453	0.396	0.403	0.495	–

TABLE 6
Rank 20 performance of cross-age retrieval on the FG-NET face dataset with 6 modalities.

To investigate the details of cross-age retrieval, the performance at ranks 1, 10 and 20 between any modalities is shown in Tables 4, 5 and 6, respectively. On the one hand, we can see that, in general, the performance of cross-age retrieval between two adjacent modalities is higher than that of non-adjacent modalities. In essence, the appearance changes between adjacent modalities will be smaller than those between large age gaps. On the other hand, it is interesting that the retrieval performance when the probe image comes from older age stages and the gallery consists of images from earlier ages normally will be better than the opposite conditions. We think this is because the appearance variation trend in the later age stages becomes smaller and some important identification characteristics remain as age increases.

Two probe samples with first 3 matches are shown in Fig. 10. The two persons have images from the $0-4$ modality to the $15-19$ modality. The left probe comes from the $5-9$ modality while the right one comes from the $0-4$ modality. We can see that several images with a same person have been successfully matched in different age stages by cross-age retrieval.

7 Conclusion

In this paper, the concept of hetero-manifold was introduced for integrating the uni- and cross-modal similarities of multi-modal data in a global view. Both types of similarity are represented in the Laplacian matrix L corresponding to the hetero-manifold. The Laplacian matrix L appears smoothly when the Hamming distance in Eq. (8) is replaced by the Euclidean distance in Eq. (17), which hints that no hash functions could be learned without all uni- and cross-modal similarities being defined on the hetero-manifold. Therefore, the proposed framework of hetero-manifold regularised hash function learning (Eq. (17)) could benefit from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2645565, IEEE Transactions on Pattern Analysis and Machine Intelligence

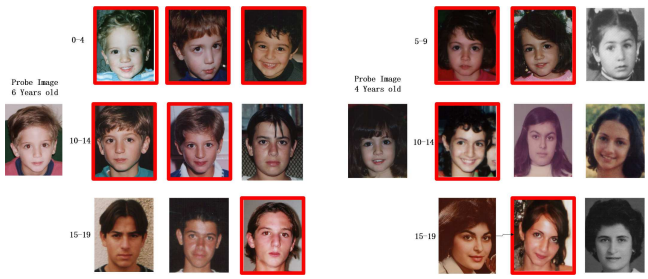

Fig. 10. The cross-modal first three matching results of two probe images. The red rectangles demonstrate the correctly matched images in the gallery of a same person.

the view of treating multi-modal data as a whole. The experimental results demonstrate that the proposed HMR outperforms the state-of-the-art methods on four popular datasets.

The hetero-manifold also offers some interesting problems in the field of cross-modal hashing. Firstly, it is interesting to consider a kernel extension of the proposed HMR. It is clear that the proposed hetero-manifold regularised framework (Eq. (17)) can be rewritten in Reproducing Kernel Hilbert Space (RKHS). By using RKHS, nonlinear hash functions could be learned, which may improve the performance of HMR. However, to achieve this, an induced problem needs to be considered for multi-modalities. For a common reproduced space or several individually reproduced spaces, which case is more reasonable? Moreover, what is the relationship between the reproduced spaces and the kernels? Secondly, it would be interesting to consider the proposed framework (Eq. (17)) in semi-supervised settings.

Appendix A
Proof of Lemma 3

Proof. We have $S^{uv} = S^{uu}P^{uv}S^{vv}$ and $S^{vu} = S^{vv}P^{vu}S^{uu}$. The transposition of S^{vu} is:

$$\begin{aligned}(S^{vu})^T &= (S^{vv}P^{vu}S^{uu})^T \\ &= (S^{uu})^T(P^{vu})^T(S^{vv})^T \\ &= S^{uu}P^{uv}S^{vv} \\ &= S^{uv}.\end{aligned}$$

The third equation holds because matrices S^{uu}, S^{vv} and $P^{uv} = (P^{vu})^T$ are symmetric.

According to the definition of similarity matrix $\mathbf{S}$, the symmetry of $\mathbf{S}$ could be proved by using the fact of $(S^{vu})^T = S^{uv}$. □

Appendix B
Proof of Lemma 4

Proof. The Hamming distance between two binary codes y_i^u and y_j^v is defined by:

$$\begin{aligned}\mathcal{D}_h(y_i^u, y_j^v) &= \sum_k y_i^u(k) \oplus y_j^v(k) \\ &= \sum_k \mathbf{1}((f_k^u(x_i^u))_+ \neq (f_k^v(x_j^v))_+),\end{aligned}$$

where $\mathbf{1}(\cdot)$ is an indicator function. Thus, for any k, we consider two conditions:

(1) If $(f_k^u(x_i^u))_+ = (f_k^v(x_j^v))_+$, it is obvious that

$$y_i^u(k) \oplus y_j^v(k) = 0 \leq |f_k^u(x_i^u) - f_k^v(x_j^v)|.$$

(2) If $(f_k^u(x_i^u))_+ \neq (f_k^v(x_j^v))_+$, we assume that $(f_k^u(x_i^u))_+ = 1$ (Otherwise, same conclusion can be also obtained). There must be $(f_k^v(x_j^v))_+ = -1$. Since the two linear projections are both hinge loss constraint-satisfied functions, we have:

$$\begin{aligned}f_k^u(x_i^u) &\geq 1 - \xi_{ik}^u, \\ f_k^v(x_j^v) &\leq -1 + \xi_{jk}^v.\end{aligned}$$

So, there is $2 - \xi_{ik}^u - \xi_{jk}^v \leq |f_k^u(x_i^u) - f_k^v(x_j^v)|$. Provided that $\xi_{ik}^u + \xi_{jk}^v \leq 1$, the following inequality is true:

$$y_i^u(k) \oplus y_j^v(k) = 1 \leq 2 - \xi_{ik}^u - \xi_{jk}^v \leq |f_k^u(x_i^u) - f_k^v(x_j^v)|.$$

In total, we obtain the following conclusion by satisfying $\forall k, \xi_{ik}^u + \xi_{jk}^v \leq 1$:

$$\begin{aligned}\mathcal{D}_h(y_i^u, y_j^v) &= \sum_k y_i^u(k) \oplus y_j^v(k) \\ &= \sum_k \mathbf{1}((f_k^u(x_i^u))_+ \neq (f_k^v(x_j^v))_+), \\ &= \sum_k \mathbf{1}^2((f_k^u(x_i^u))_+ \neq (f_k^v(x_j^v))_+), \\ &\leq \sum_k (f_k^u(x_i^u) - f_k^v(x_j^v))^2.\end{aligned}$$

The third equation holds due to that $0^2 = 0$ and $1^2 = 1$. Therefore, we have:

$$\begin{aligned}\mathcal{D}_h(y_i^u, y_j^v) &\leq ||F(x_i^u) - F(x_j^v)||_2^2 \\ &= \mathcal{D}_e(F(x_i^u), F(x_j^v)).\end{aligned}$$

□

Appendix C
Proof of Equation (16)

Proof. According to the definition of $\mathbf{W}^T$ (13) and the definition of $\mathbf{X}$ (14), it is clear that

$$\begin{aligned}&\mathbf{W}^T\mathbf{X} \\ =\ & ((W^1)^T, \cdots, (W^M)^T)\begin{pmatrix} X^1 & 0 & \cdots & 0 \\ 0 & X^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & X^M \end{pmatrix} \\ =\ & ((W^1)^T X^1, (W^2)^T X^2, \cdots, (W^M)^T X^M). \qquad (27)\end{aligned}$$

Then

$$\begin{aligned}&\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W}) \\ =\ & \mathbf{tr}\Big(((W^u)^T X^u)_{u=1}^M (S^{uv})_{u,v=1}^M ((X^v)^T W^v)_{v=1}^M\Big) \\ =\ & \sum_{u,v} \mathbf{tr}((W^u)^T X^u S^{uv} (X^v)^T W^v) \qquad (28)\end{aligned}$$

Notice the definition of $F(x_i^u)$ and $X^u = (x_1^u, \cdots, x_N^u)$, we have

$$\begin{aligned}&\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W}) \\ =\ & \sum_{u,v} \mathbf{tr}((F(x_i^u))_{i=1}^N S^{uv} ((F(x_j^v))_{j=1}^N)^T) \\ =\ & \sum_{u,v} \sum_{i,j} S^{uv}(x_i^u, x_j^v) \langle F(x_i^u), F(x_j^v) \rangle_2. \qquad (29)\end{aligned}$$

Meanwhile, we have the following equations

$$\begin{aligned}&\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}) = \sum_{u,i} d_{ui}||F(x_i^u)||_2^2\\ =\ &\sum_{u,i}||F(x_i^u)||_2^2\sum_{v,j}\mathbf{S}(x_i^u,x_j^v)\\ =\ &\sum_{u,v}\sum_{i,j}\mathbf{S}(x_i^u,x_j^v)||F(x_i^u)||_2^2 \end{aligned} \tag{30}$$

where $\mathbf{D} = \mathbf{diag}(d_{11}, d_{12}, \cdots, d_{ui}, \cdots, d_{MN})$ and $d_{ui} = \sum_{v,j}\mathbf{S}(x_i^u, x_j^v)$. Similarly, the following equation is true.

$$\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}) = \sum_{u,v}\sum_{i,j}\mathbf{S}(x_i^u,x_j^v)||F(x_j^v)||_2^2 \tag{31}$$

Combining the equations (29), (30) and (31) and considering $\mathbf{S}(x_i^u, x_j^v) = S^{uv}(x_i^u, x_j^v)$, we have

$$\begin{aligned}&2\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})\\ =\ &2\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}) - 2\mathbf{tr}(\mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W})\\ =\ &\sum_{u,v}\sum_{i,j}S^{uv}(x_i^u,x_j^v)\|F(x_i^u) - F(x_j^v)\|_2^2\\ =\ &\mathcal{L}_c^e(\mathbf{G}) \end{aligned} \tag{32}$$

□

APPENDIX D
PROOF OF THE FORMULA (25 AND 22)

Proof. For simplicity, we delete the index of projections and, then the objective function in 23 become similar to the function in 21. The only difference between them is Eq. 23 has a orthogonal item. Thus, if further define $H = \mathbf{X}\mathbf{L}\mathbf{X}^T + C_2\mathbf{Q}$, we obtain:

$$\begin{aligned}&\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T H\mathbf{w} + C_1\sum_{u=1}^{M}\sum_{i=1}^{N}\xi_i^u\\ &s.t. \forall u, i,\ y_i^u(w^u)^T x_i^u \geq 1 - \xi_i^u,\ \xi_i^u \geq 0, \end{aligned} \tag{33}$$

where the element y_i^u of $\mathbf{y}$ is the bit of initial or learned in the last round. In case of solving the problem in 21, the parameter C_2 can be directly set to 0. The Lagrange function of the problem 33 is

$$\begin{aligned}&L(\mathbf{w},\xi,\alpha,\gamma)\\ =\ &\frac{1}{2}\mathbf{w}^T H\mathbf{w} + C_1 e^T\xi\\ &-\mathbf{w}^T\mathbf{X}_\mathbf{y}\alpha + e^T\alpha - \alpha^T\xi - \gamma^T\xi, \end{aligned} \tag{34}$$

where $\alpha = (\alpha_1, \cdots, \alpha_i, \cdots, \alpha_{NM})^T$ and $\mathbf{X}_\mathbf{y} = \mathbf{diag}(X_\mathbf{y}^1, \cdots, X_\mathbf{y}^u, \cdots, X_\mathbf{y}^M)$, the matrix $X_\mathbf{y}^u = (y_1^u x_1^u, \cdots, y_N^u x_N^u)$. The gradients with respect to the parameters are:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{w}} &= H\mathbf{w} - \mathbf{X}_\mathbf{y}\alpha;\\ \frac{\partial L}{\partial \xi} &= C_1 e - \alpha - \gamma. \end{aligned}$$

Thus, the optimal values should satisfy the following conditions:

$$\begin{aligned}\mathbf{w}^* &= H^{-1}\mathbf{X}_\mathbf{y}\alpha;\\ \gamma &= C_1 e - \alpha. \end{aligned}$$

Substituting the above equations into the original Lagrange function (34), we obtain the dual problem:

$$\begin{aligned}\alpha^* &= \arg\min_{\alpha} -e^T\alpha + \frac{1}{2}\alpha^T\mathbf{X}_\mathbf{y}^T H^{-1}\mathbf{X}_\mathbf{y}\alpha\\ &s.t.\quad 0 \leq \alpha_i \leq C_1. \end{aligned} \tag{35}$$

The problem (35) is a standard quadratic programming problem. Therefore, if α^* is the solution of (35), the optimal projection direction can be obtained as:

$$\mathbf{w}^* = (\mathbf{X}\mathbf{L}\mathbf{X}^T + C_2\mathbf{Q})^{-1}\mathbf{X}_\mathbf{y}\alpha^*.$$

□

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant 61528106, and in part by the Newton International Exchanges Scheme. Ling Shao is the corresponding author. Yi Tang is partly supported by the National Natural Science Foundation of China (Grant nos. 61462096) and the Science and Technology Plan Project of Yunnan Province (Grant no. 2014FB148).

REFERENCES

[1] V. Mahadevan, C. WahWong, J. C. Pereira, T. T. Liu, N. Vasconcelos, and L. K. Saul, "Maximum covariance unfolding: Manifold learning for bimodal data," in *Proc. NIPS*, 2011.
[2] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *Proc. MM*, 2013.
[3] C. Kang, S. Liao, Y. He, J. Wang, W. Niu, and S. Xiang, "Cross-modal similarity learning : A low rank bilinear formulation," in *Proc. CIKM*, 2015.
[4] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. CVPR*, 2015.
[5] D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," in *Proc. SIGIR*, 2011.
[6] S. Kim, Y. Kang, and S. Choi, "Sequential spectral learning to hash with multiple representations," in *Proc. ECCV*, 2012.
[7] S. ichi Amari, "Integration of stochastic models by minimizing alpha - divergence," *Neural Computation*, vol. 19, pp. 2780–2796, 2007.
[8] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. ICCV*, 2011.
[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011.
[10] M. Rastegari, J. Choi, S. Fakhraei, H. D. III, and L. S. Davis, "Predictable dual-view hashing," in *Proc. ICML*, 2013.
[11] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. CVPR*, 2014.
[12] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. ICCV*, 2013.
[13] M. Yu, L. Liu, and L. Shao, "Binary set embedding for cross-modal retrieval," *IEEE Transactions on NNLS*, 2016.
[14] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *arXiv:1602.02255v2*, 2016.
[15] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. IJCAI*, 2011.
[16] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. IJCAI*, 2015.
[17] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on PAMI*, vol. 36, no. 3, pp. 521–535, Aug. 2014.
[18] F. Zhu, L. Shao, and M. Yu, "Cross-modality submodular dictionary learning for information retrieval," in *Proc. CIKM*, 2014.
[19] Y. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, Jan. 2008.
[20] Y. Yang, Y. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, Mar. 2008.

[21] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. MM*, 2009.
[22] D. Hu, X. Lu, and X. Li:, "Multimodal learning via exploring deep semantic similarity," *ACM Multimedia*, 2016.
[23] L. Liu and L. Shao, "Sequential compact code learning for unsupervised image hashing," *IEEE Transactions on NNLS*, vol. 27, no. 12, pp. 2526–2536, Dec. 2016.
[24] M. M. Bronstein and A. M. Bronstein, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. CVPR*, 2010.
[25] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3157–3166, July 2016.
[26] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in *Proc. MM*, 2014.
[27] 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. SIGMOD*, 2013.
[28] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 956–966, Mar. 2015.
[29] ——, "Unsupervised local feature hashing for image similarity search," *IEEE Transactions on Cybernetics*, vol. 46, no. 11, pp. 2548–2558, Nov. 2016.
[30] ——, "Projection bank: From high-dimensional data to medium-length binary codes," in *Proc. ICCV*, 2015.
[31] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on PAMI*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.
[32] L. Shao, M. Yu, and L. Liu, "Kernelized multiview projection for robust action recognition," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 115–129, Jun. 2016.
[33] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. CVPR*, 2015, pp. 4371–4379.
[34] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Transactions on PAMI*, vol. 36, no. 4, pp. 824–830, 2014.
[35] O. Zoidi, N. Nikolaidis, A. Tefas, and I. Pitas, "Multi-modal label propagation based on a higher order similarity matrix," in *Workshop on MLSP*, 2015.
[36] S. H. Amiri and M. Jamzad, "Efficient multi-modal fusion on supergraph for scalable image annotation," *Pattern Recognition*, vol. 48, pp. 2241–2253, 2015.
[37] J. Song, L. Gao, F. Zou, Y. Yan, N. Sebe, and J. Wang, "Deep and fast: Deep learning hashing with semi-supervised graph construction," *Journal of Image and Vision Computing*, vol. 55, no. 2, pp. 101–108, Nov. 2016.
[38] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. CVPR*, 2015, pp. 3270–3278.
[39] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *CoRR*, 2014.
[40] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *arXiv:1606.00185v1*, 2016.
[41] F. Zheng, Z. Song, L. Shao, R. Chung, K. Jia, and X. Wu, "A semi-supervised approach for dimensionality reduction with distributional similarity," *Neurocomputing*, vol. 103, pp. 210–221, 2013.
[42] F. Zheng and L. Shao, "Learning cross-view binary identities for fast person re-identification," in *Proc. IJCAI*, USA, July 2016.
[43] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI*, 2014.
[44] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008.
[45] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014.
[46] Face and G. R. W. group., "Fg-net aging database," in *http://www-prima.inrialpes.fr/FGnet/*, 2000.
[47] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. CVPR*, 2007.
[48] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proc. ICML*, 2006.
[49] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Transactions on PAMI*, 2012.
[50] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. CVPR*, 2014.
[51] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015.
[52] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. CVPR*, 2010.
[53] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, 2011.
[54] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, 2012.
[55] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, 2013.
[56] ——, "Person re-identification by salience matching," in *Proc. ICCV*, 2013.
[57] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *In Proc. CVPR*, 2013.
[58] X. You, W. Guo, S. Yu, K. Li, J. C. Prłncipe, and D. Tao, "Kernel learning for dynamic texture synthesis," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4782–4795, Aug. 2016.
[59] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, May 2007.
[60] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

Feng Zheng is currently working toward the Ph.D. degree in the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K. Previously, he received the B.S. and M.S. degrees in Applied Mathematics from Hubei University, Wuhan, China, in 2006 and 2009 respectively. From 2009 to 2012, he worked as an Assistant Research Professor in Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), China. His research interests include computer vision, machine learning and human-computer interaction.

Yi Tang received the B.S., the M.S. and Ph.D. degrees in mathematics from Hubei University, Wuhan, China, in 2001, 2006, and 2010, respectively. He is currently an Associate Professor with Key Laboratory of IOT Application Technology of Universities in Yunnan Province and department of mathematics and computer Science, Yunnan Minzu University, Kunming, China. His research interests include machine learning, statistical learning theory, image processing, and pattern recognition.

Ling Shao (M09-SM10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with the University of Sheffield and a senior scientist (2005- 2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of the IEEE.