Multimedia Tools and Applications (2019) 78:30895–30922
https://doi.org/10.1007/s11042-019-7192-5

Label guided correlation hashing for large-scale cross-modal retrieval

Guohua Dong[1,2] · Xiang Zhang[2,3] · Long Lan[2,3] · Shiwei Wang[1,2] · Zhigang Luo[1,2]

Received: 2 September 2018 / Revised: 10 December 2018 / Accepted: 9 January 2019 /
Published online: 6 February 2019

Abstract

With the explosive growth of multimedia data such as text and image, large-scale cross-modal retrieval has attracted more attention from vision community. But it still confronts the problems of the so-called "media gap" and search efficiency. Looking into the literature, we find that one leading type of existing cross-modal retrieval methods has been broadly investigated to alleviate the above problems by capturing the correlations across modalities as well as learning hashing codes. However, supervised label information is usually independently considered in the process of either generating hashing codes or learning hashing function. To this, we propose a label guided correlation cross-modal hashing method (LGCH), which investigates an alternative way to exploit label information for effective cross-modal retrieval from two aspects: 1) LGCH learns the discriminative common latent representation across modalities through joint generalized canonical correlation analysis (GCCA) and a linear classifier; 2) to simultaneously generate binary codes and hashing function, LGCH introduces an adaptive parameter to effectively fuse the common latent representation and the label guided representation for effective cross-modal retrieval. Moreover, each subproblem of LGCH has the elegant analytical solution. Experiments of cross-modal retrieval on three multi-media datasets show LGCH performs favorably against many well-established baselines.

Keywords Cross-modal retrieval · Cross-modal hashing · Generalized canonical correlation analysis · Supervised learning

1 Introduction

With the rapid growth of multimedia data, such as text, image, audio and video, etc., how to perform cross-modal retrieval [32–34, 55] on large-scale multimedia data is still a challenging yet interesting topic. This is because it is usually subject to feature representation

✉ Xiang Zhang
zhangxiang08@nudt.edu.cn

Extended author information available on the last page of the article.

across modalities [40, 45, 48, 49]. For efficiency, hashing methods have been broadly studied for large-scale cross-modal retrieval due to its low-storage cost and fast search speed. They seek to map the data points from original feature space to a Hamming space with binary codes as well as to preserve the neighborhood structure. Thereafter, hashing methods can be potentially applied to many vision tasks such as classification [43, 54], person re-identification [13, 53, 54] and information retrieval [41, 42, 44, 47, 51, 56, 58]. Thus far many hashing methods including Locally Sensitive Hashing (LSH) [10], Spectral Hashing (SH) [50] and Discrete Hashing via Affine Transformation (ADGH) [12], have been developed and mainly are applied for large-scale image retrieval. However, they are only designed for single-modal retrieval and neglect solving the media gap across modalities, thus they cannot directly be applied to cross-modal retrieval.

To this, much progress on hashing for cross-modal retrieval have been made to measure the similarities of heterogeneous modalities. This can be attributed to the fact that multi-media data belong to different attribution spaces and have the inconsistent semantic gap. Thus, we are eager to relieve such inconsistency across modalities before cross-modal retrieval. In the regard, current hashing methods for cross-modal retrieval might be roughly grouped into supervised hashing methods and unsupervised ones. As we know, unsupervised cross-modal hashing methods, such as Cross-View Hashing (CVH) [19], Inter-Media Hashing (IMH) [36], Composite Correlation Quantization (CCQ) [20] and Collaborative Subspace Graph Hashing (CSGH) [58], excel at working without supervised information such as semantic labels and pair-wised information. That is because they encode cross-modal data into hashing codes by looking into the data itself. Compared with unsupervised counterparts, cross-modal hashing methods with supervised information, such as Cross-Modal Sensitive Hashing (CMSSH) [6], Semantic Correlation Maximization (SCM) [57] and Quantized Correlation Hashing (QCH) [52], usually achieve better retrieval results. Our work is an alternative in this scope.

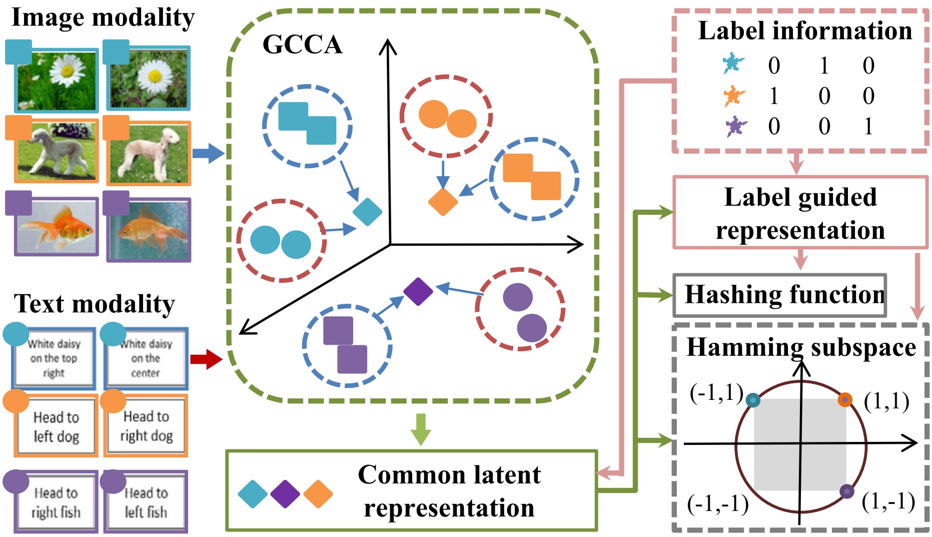


Fig. 1 The framework of LGCH, mainly containing two procedure: the common latent representation and the label guided representation learning procedure, and the hashing function and Hamming subspace learning procedure

Most of aforementioned cross-modal hashing methods usually maps data across modalities into a common Hamming subspace where the relationship across modalities can be unveiled. One leading type of such methods often captures the correlation between multiple modalities. These methods could mostly be inspired by canonical correlation analysis (CCA) [16] and its variants such as kernel canonical correlation analysis (KCCA) [1], cluster canonical correlation analysis (cluster-CCA) [31], generalized CCA (GCCA) [15], Deep Canonical Correlation Analysis (DCCA) [2] and Deep Generalized Canonical Correlation Analysis (DGCCA) [5]. Analogue with them, our work also uncovers the correlation across modalities in the frame of generalized CCA (GCCA) [15]. Compared to CCA, GCCA can be easily extended to more than two modalities in terms of model complexity. Moreover, it explicitly learns the common latent representation of all the modalities. Inspired by these two aspects of GCCA, this paper proposes a label guided correlation hashing method for cross-modal retrieval based on GCCA, dubbed as LGCH. Similar to GCCA, LGCH easily deals with multi-media data because it intends to maintain two merits mentioned above by imposing the extra constraints over the common latent representation. Different from GCCA, LGCH is a supervised cross-modal hashing method and takes GCCA as its special case. As illustrated in Fig. 1, without loss of generality, LGCH learns two individual projection matrices to respectively project text and image into a common latent subspace whilst to generate binary codes and hashing function on the basis of the learned common latent representation. To make the common latent representation more discriminative, we introduce a label guided linear classifier. More importantly, we devise a novel strategy of leveraging the semantic label to guide learning both binary codes and hashing function simultaneously. As we know, the semantic label is usually utilized to either generate binary codes or learn hashing function, with supervised joint learning both aspects to be rarely explored. Thus, LGCH can generate discriminative binary codes and meanwhile effectively fuse the common latent representation and the label guided representation by introducing an adaptive parameter. To highlight this work, the main contributions are summarized as follows:

1) LGCH can be easily extended for cross-modal retrieval that involves more than two types of multimedia data, because the involved formulation terms is linear with the number of media types.
2) To learn discriminative hashing codes, we use label information to guide the training process. On one hand, learning the classifier over the common latent representation is to make the learned common latent representation more discriminative. On the other hand, we try to learn the hashing function and encode the hashing codes with the label guided representation. More importantly, we introduce an adaptive parameter to jointly fuse the discriminative common latent representation and the label guided representation to simultaneously learn more discriminative hashing codes and hashing function.
3) As a byproduct, each sub-problem of our model has the elegant analytical solution.
4) Experiments of cross-modal retrieval on Wiki [30], NUS-WIDE [7] and Flickr25K [17] datasets, show the effectiveness of LGCH as compared to many well-established baselines.

2 Related work

Cross-modal retrieval is to deal with multimedia data search problem. It has been a popular research topic in machine learning and computer vision. The key point of cross-modal retrieval task is how to solve the “media gap” problem of multimedia data. Much progress has been made to address this issue. Considering whether using deep structures or not, existing cross-modal retrieval methods are roughly divided into traditional shallow structure based methods and deep neural network (DNN) based methods. For more details, we refer the readers to [28] for a comprehensive survey.

Traditional shallow structure based methods mostly adopt the hand-crafted features, mainly including Scale Invariant Feature Transform(SIFT) [21, 22], Histogram of Oriented Gradient (HOG) [9], Speeded-up Robust Features (SURF) [3, 4], Local Binary Pattern (LBP) [26] and Bag-of-words based features (BoW) [35], etc. One category of shallow structure based methods is to directly compute the similarities among different media types by analyzing the known data relationships without training an explicit common space. Since there is no common space in these methods, the cross-modal similarities cannot be measured directly by distance measuring or normal classifiers. One effective solution [39, 59] is to construct one or more graphs for multimedia data, and use the edge in graphs to represent the neighborhood structure among modalities. Then the retrieving process can be evaluated based on similarity propagation [59], constraint fusion [39] and so on. Another solution is neighbor analysis methods [8, 23], which are usually based on graph construction framework. The difference between these two solutions is that graph-based methods focus on the process of graph construction, while neighbor analysis methods concentrates on measuring the cross-modal similarities by using the neighbor relationships. The shallow structure based methods are easy to understand and implement.

Another category is to project different media features to a common space, then compute the similarities among different media types in this common space. The common space can be a continuous feature embedding space [25, 29, 38], or a binary Hamming space [6, 19, 20, 36, 52], where the latter is also dubbed hashing based methods. CCA based methods is one of the most representative common space methods. Many studies based on CCA have been made to perform cross-modal retrieval, such as three-view canonical correlation analysis (CCA (V+T+K)) [14], multi-label Canonical Correlation Analysis (ml-CCA) [29] and semantic correlation matching (SCM) [57]. CCA itself is an unsupervised method, while some of its variants make use of supervised information. For instance, ml-CCA can learn a shared subspace by taking the high-level semantic information into account. In Semantic matching (SM) procedure, SCM maps the image and text spaces into a pair of semantic spaces, which can be represented by a posterior probability distribution calculated by multiclass logistic regression. CCA (V+T+K) introduces a three-view kernel CCA formulation for learning a joint space for visual, textual, and semantic information. Our proposed method belongs to this scope and provides an alternative way to leverage label information. Different from previous CCA based retrieval methods, we exploit generalized CCA (GCCA) [15] to capture the correlation of over two types of media data as well as to learn hashing codes and function.

In recent years, since 2012, when Krizhevsky et al. [18] utilized the convolutional neural network (CNN) to achieve the state-of-the-art classification result on ILSRVC 2012, many studies based on deep learning have emerged. When processing the image data, some DNN based cross-modal retrieval methods directly pass the source RGB features through the network [11, 46]. Some methods also use traditional hand-crafted features as the input [27]. Existing DNN based cross-modal retrieval methods have two way to use the networks: the first one [25, 37] use a shared network, that is to say, inputs of different modalities pass through the same network layers. The second one [11, 46] use different subnetworks for different modalities, and introduce the correlation constraints to the code layers to capture the neighbourhood structure among different modalities. DNN based cross-modal retrieval methods usually yield better performance as compared to most shallow structure based methods, but their performance heavily relies on large-scale training data and it often costs expensive training time. In addition, most existing DNN retrieval methods usually consider two media types and hence might not be directly applied for over two media types.

3 Label guided correlation cross-modal hashing (LGCH)

In this section, we propose a label guided correlation cross-modal hashing method (LGCH) for cross-modal retrieval. Then we provide a toy example to clearly illustrate its effectiveness and extend it to multi-modal situation. In the end, we list the optimization procedure of LGCH.

3.1 Model

Why GCCA? As the correlation across modalities is always the focus of cross-modal retrieval, our model considers this point as well. Concretely, we prefer generalized CCA (GCCA) [15] to CCA in order to capture the correlation across two modalities. The reasons for this are twofold. Firstly, GCCA is closely akin to CCA and thus shares some analogue properties. Secondly, GCCA explicitly learns a common latent representation that simultaneously best reconstructs all of the view-specific representations. This has the advantage that, by using GCCA, we do not need to care for the number of media types. This is because, on the basis of the common latent representation, GCCA has the flexibility of coupling with some sophisticated regularization techniques by enforcing the regularized constraints over the common latent representation.

Without loss of generalization, given image and text descriptors $X_1 \in R^{m_1 \times n}$ and $X_2 \in R^{m_2 \times n}$, where n is the number of samples, and m_1 and m_2 are the dimensions of image and text features, respectively. Then the objective of GCCA is:

$$\begin{aligned} &\min_{G,W_1,W_2} O_1 = \left\|G - W_1^T X_1\right\|_F^2 + \left\|G - W_2^T X_2\right\|_F^2, \\ &s.t.\ \ GG^T = I \end{aligned} \tag{1}$$

where $W_1 \in R^{m_1 \times r}$ and $W_2 \in R^{m_2 \times r}$ are the transformation matrices of image and text, respectively, and $G \in R^{r \times n}$ is the common latent representation (latent rep. for short), wherein r is the dimension of the common latent subspace.

To avoid trivial solution, we impose the F-norm regularization over two transformation matrices as below:

$$\begin{aligned} &\min_{G,W_1,W_2} O_2 = \left\|G - W_1^T X_1\right\|_F^2 + \left\|G - W_2^T X_2\right\|_F^2 \\ &+\beta(\|W_1\|_F^2 + \|W_2\|_F^2). \\ &s.t.\ \ GG^T = I \end{aligned} \tag{2}$$

Learning discriminative common latent representation with linear classifier To make use of the label information and make the learned common subspace more discriminative, we intend to learn a linear classifier based on the common latent representation:

$$\min_{P,G} O_3 = \left\|Y - P^T G\right\|_F^2, \tag{3}$$

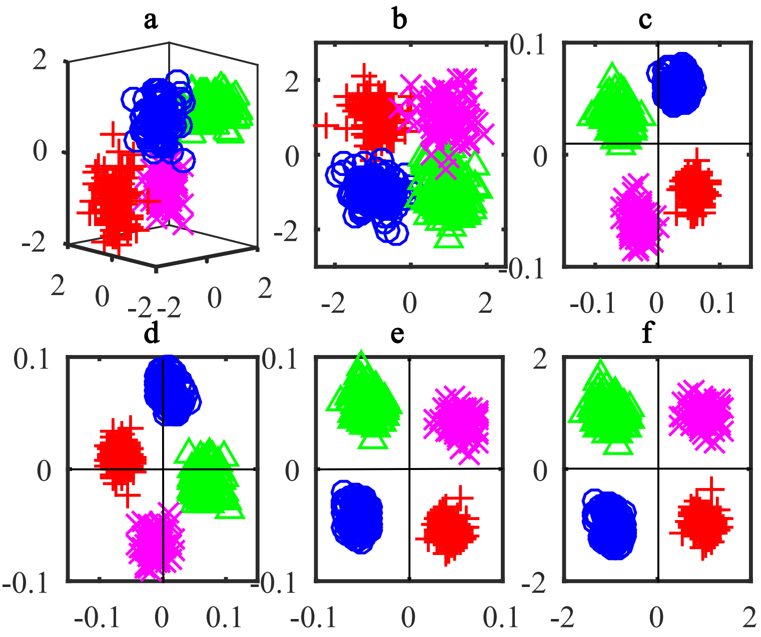


Fig. 2 Original cross-modal dataset is composed of 400 **a** 3D and **b** 2D samples; the common latent representation G learned by **c** $O_2 + \lambda \|B - G\|_F^2$, **d** $O_2 + O_3 + \lambda \|B - G\|_F^2$, and **e** LGCH, respectively; **f** the fused features $\sigma G + (1 - \sigma) Q P^T G$ learned by LGCH

where $Y \in R^{k \times n}$ is the label information shared by image and text, k is the number of classes, and $P \in R^{r \times k}$ is the transformation matrix mapping the common latent representation from the common latent subspace to the label subspace. Different from the similar strategy of the cross-modal counterparts, we base the classifier on the common latent representation.

Learning label guided binary codes and hashing function Most existing cross-modal hashing methods often focus on either directly generating the hashing code or learning hashing function. Different from them, we jointly consider both aspects based on the common latent representation in a unified way. Besides, we indirectly incorporate the label information into the unified process. Concretely, we minimize the quantization error loss:

$$\begin{aligned} \min_{P,G,Q,B} O_4 &= \left\| B - \sigma \underbrace{G}_{latent\ rep.} - (1-\sigma) \underbrace{Q \overbrace{P^T G}^{learned\ label\ info.}}_{new\ rep.} \right\|_F^2, \\ s.t.\ GG^T &= I, B \in \{-1, 1\}^{r \times n},\ B\mathbf{1} = 0, 0 \le \sigma \le 1 \end{aligned} \tag{4}$$

where $B \in R^{r \times n}$ is the binary codes shared by two modalities, and $\mathbf{1} = [1, \cdots, 1]^T \in R^n$. $Q \in R^{r \times k}$ transforms the learned label information (learned label info. for short) to the label guided representation (new rep. for short), and σ is an adaptive parameter to trade off the label guided representation and the common latent representation. By comparing Fig. 2c and d with Fig. 2e and f, we find that this proposed term greatly affects the discriminative property of the final common latent representation and the binary codes. The details about Fig. 2 are left in the paragraph after the final model part.

The final model Based on the above formula, we derive our final model dubbed label guided correlation cross-modal hashing (LGCH) as follows:

$$\begin{aligned}&\min_{P,G,W_1,W_2,B,Q,\sigma} O = O_2 + O_3 + O_4\\&= \left\|Y - P^T G\right\|_F^2 + \alpha(\left\|G - W_1^T X_1\right\|_F^2 + \left\|G - W_2^T X_2\right\|_F^2)\\&+\beta(\|W_1\|_F^2 + \|W_2\|_F^2) + \lambda \left\|B - \sigma G - (1-\sigma)QP^T G\right\|_F^2,\\&s.t.\ GG^T = I, B \in \{-1,1\}^{r\times n}, B\mathbf{1} = 0, 0 \le \sigma \le 1\end{aligned} \tag{5}$$

The first term O_2 in (5) aims to learn a label subspace to make full use of the discriminative label information; the second term O_3 in (5) can learn a common latent subspace through the GCCA procedure; the last term O_4 connect the learned label subspace and the common latent subspace with the Hamming space, in order to learn more discriminative binary hashing codes. It is easy to find that our model involves a cross-modal term O_3 which can learn a common latent representation, then all the constraints including O_2 and O_4 are imposed over the common latent representation. Thus, we easily extend (5) to multi-modal case as follows:

$$\begin{aligned}&\min_{P,G,W_1,W_2,B,Q,\sigma} \left\|Y - P^T G\right\|_F^2 + \sum_{i=1}^{m} (\alpha \left\|G - W_i^T X_i\right\|_F^2 + \beta \|W_i\|_F^2)\\&+\lambda \left\|B - \sigma G - (1-\sigma)QP^T G\right\|_F^2,\\&s.t.\ GG^T = I, B \in \{-1,1\}^{r\times n}, B\mathbf{1} = 0, 0 \le \sigma \le 1\end{aligned} \tag{6}$$

where m is the number of multimedia types. From (6), we can find that the multi-media formula of LGCH considers cross-modal relationships which is linear with the number of media types, thus can be easily extended to cross-modal retrieval that involves more than two types of multimedia data. More importantly, LGCH provides a novel alternative to leveraging the label information as the guidance for simultaneously learning the binary codes and hashing function.

To illustrate the function of label information in LGCH, we perform a toy example to show the common latent representation with and without label information guiding. We firstly generate a three-dimensional dataset with 400 samples and these samples belong to four different Gaussian distributions. Each Gaussian distribution stands for one category. Likewise, we generate a corresponding two-dimensional dataset again. Then we collect these two synthetic datasets to construct a cross-modal dataset. Based on such cross-modal dataset, we respectively show the common latent representation trained under different conditions. The learned features are displayed in Fig. 2. As shown in Fig. 2c, d and e, the common latent representation generated by LGCH shows the best discriminative ability for the four classes. By comparing Fig. 2e with f, we can find that LGCH can make the features of O_4 more discriminative by fusing the common latent representation and the label guided representation.

3.2 Optimization

For convenience, we consider the two-modality cross-modal case to show the optimization of our method. Although (5) is jointly non-convex with respect to all the variables, we can solve it through the use of an iterative optimization method, where G, W_1, W_2, P, Q , B

and σ are solved alternatively. It is interesting to find that each subproblem can yield a closed-form solution. The details of our optimization procedure are listed as follows:

***G*-subproblem** When W_1, W_2, P, Q , B and σ are fixed, we rewrite (5) as

$$\begin{aligned} &\max_{G} tr(GH), \\ &s.t. \quad GG^T = I \end{aligned} \tag{7}$$

where $H = Y^T P^T + \alpha X_1^T W_1 + \alpha X_2^T W_2 + \lambda\sigma B^T + \lambda(1-\sigma)B^T Q P^T$. Then we can get an analytical solution of G via Theorem 1.

Theorem 1 *$G = UV^T$ is the optimal solution of the problem in (7), where U and V are the left- and right-part of the Singular Value Decomposition (SVD) of H^T, respectively.*

Proof Suppose that the singular value decomposition (SVD) form of H^T is $H^T = U\Sigma V^T$, and substitute this into (7), we can obtain:

$$tr(GH) = tr(GV\Sigma U^T) = tr(\Phi\Lambda), \tag{8}$$

where $\Phi = U^T GV$. According to the von Neumann's trace inequality [24],

$$tr(\Phi\Lambda) \le \sum_{i=1}^{r} \phi_i \eta_i, \tag{9}$$

where $\phi_1 \ge \cdots \ge \phi_i$ and $\eta_1 \ge \cdots \ge \eta_i$ are the singular values of ϕ and η, respectively. As $\Phi\Phi^T = I$, the singular values $\phi_i = 1$ and (9) becomes $tr(\Phi\Lambda) \le \sum\limits_{i=1}^{r} \eta_i$. The equality holds when $\Phi = I$, then the solution of G is

$$G = UV^T \tag{10}$$

This completes our proof. □

***P*-subproblem** When G, W_1, W_2, Q , B and σ are fixed, setting the gradient of (5) over P to zero, we can obtain

$$GY^T + \lambda(1-\sigma)GB^T Q - \lambda\sigma(1-\sigma)Q = P + \lambda(1-\sigma)^2 P Q^T Q. \tag{11}$$

Then the solution of P is

$$P = (GY^T + \lambda(1-\sigma)GB^T Q - \lambda\sigma(1-\sigma)Q)(I_k + \lambda(1-\sigma)^2 Q^T Q)^{-1}. \tag{12}$$

W_1(W_2)-subproblem When G, W_2, P, Q , B and σ are fixed, setting the gradient of (5) over W_1 to zero, we can obtain

$$(\alpha X_1 X_1^T + \beta I)W_1 = \alpha X_1 G^T. \tag{13}$$

The solution of W_1 is

$$W_1 = \alpha(\alpha X_1 X_1^T + \beta I)^{-1} X_1 G^T. \tag{14}$$

Likewise, when G, W_1, P, Q , B and σ are fixed, setting the gradient of (5) over W_2 to zero, we can obtain the solution of W_2:

$$W_2 = \alpha(\alpha X_2 X_2^T + \beta I)^{-1} X_2 G^T. \tag{15}$$

***Q*-subproblem** Note that only when $\sigma \neq 1$, we need solve this problem. When G, W_1, W_2, P , B and σ are fixed and $\sigma \neq 1$, setting the gradient of (5) over Q to zero, we can obtain

$$(1-\sigma)^2 QP^T P = (1-\sigma)BG^T P - \sigma(1-\sigma)P. \tag{16}$$

The solution of Q is

$$Q = \frac{(BG^T P - \sigma P)(P^T P)^{-1}}{(1-\sigma)}. \tag{17}$$

***B*-subproblem** When G, W_1, W_2, P , Q and σ are fixed, (5) can be written as

$$\begin{aligned} &\max_B B(\sigma G^T + (1-\sigma)G^T P Q^T). \\ &s.t.\ B \in \{-1, 1\}, B\mathbf{1} = 0 \end{aligned} \tag{18}$$

According to [12, 58], the solution of B is

$$B(\mathbf{rank}(i,j), j) = \begin{cases} 1, & i \leq \frac{n}{2} \\ -1, & otherwise, \end{cases} \tag{19}$$

where **rank** is the sorted index of $\sigma G + (1-\sigma)QP^T G)$ in descending order, and i and j are the indices of the matrix **rank**.

***σ*-subproblem** When G, W_1, W_2, P , Q and B are fixed, we can rewrite (5) as

$$\begin{aligned} &\min_\sigma -2\lambda\sigma tr(BG^T) - 2\lambda(1-\sigma)tr(BG^T PQ^T) + \sigma^2\lambda r + 2\lambda\sigma(1-\sigma)tr(PQ^T) \\ &+(1-\sigma)^2\lambda tr(QP^T PQ^T) \\ &= \min_\sigma -2\sigma tr(BG^T) + 2\sigma tr(BG^T PQ^T) + \sigma^2 r + 2\sigma tr(PQ^T) - 2\sigma^2 tr(PQ^T) \\ &-2\sigma tr(QP^T PQ^T) + \sigma^2 tr(QP^T PQ^T) \\ &= \min_\sigma \tfrac{1}{2}\sigma^2(2r - 4tr(PQ^T) + 2tr(QP^T PQ^T)) + \sigma(-2tr(BG^T) \\ &+2tr(BG^T PQ^T) + 2tr(PQ^T) - 2tr(QP^T PQ^T)). \\ &s.t.\ 0 \leq \sigma \leq 1 \end{aligned} \tag{20}$$

Let $h = 2r - 4tr(PQ^T) + 2tr(QP^T PQ^T)$ and $f = -2tr(BG^T) + 2tr(BG^T PQ^T) + 2tr(PQ^T) - 2tr(QP^T PQ^T)$, then (20) becomes

$$\begin{aligned} &\min_\sigma \tfrac{1}{2}\sigma^2 h + \sigma f. \\ &s.t.\quad 0 \leq \sigma \leq 1 \end{aligned} \tag{21}$$

Since $h = 2r - 4tr(PQ^T) + 2tr(QP^T PQ^T) = 2\left\|I - PQ^T\right\|_F^2 \geq 0$, (21) is convex and has a closed-form solution. By simple algebra, we can obtain the solution to (21) as follows:

$$\sigma = \begin{cases} 0, & if\ -\frac{f}{h} \notin [0,1]\ and\ \frac{1}{2}\sigma^2 h + \sigma f\,|_{\sigma=1} > \frac{1}{2}\sigma^2 h + \sigma f\,|_{\sigma=0}, \\ 1, & if\ -\frac{f}{h} \notin [0,1]\ and\ \frac{1}{2}\sigma^2 h + \sigma f\,|_{\sigma=1} < \frac{1}{2}\sigma^2 h + \sigma f\,|_{\sigma=0}, \\ -\frac{f}{h}, & if\ -\frac{f}{h} \in [0,1]. \end{cases} \tag{22}$$

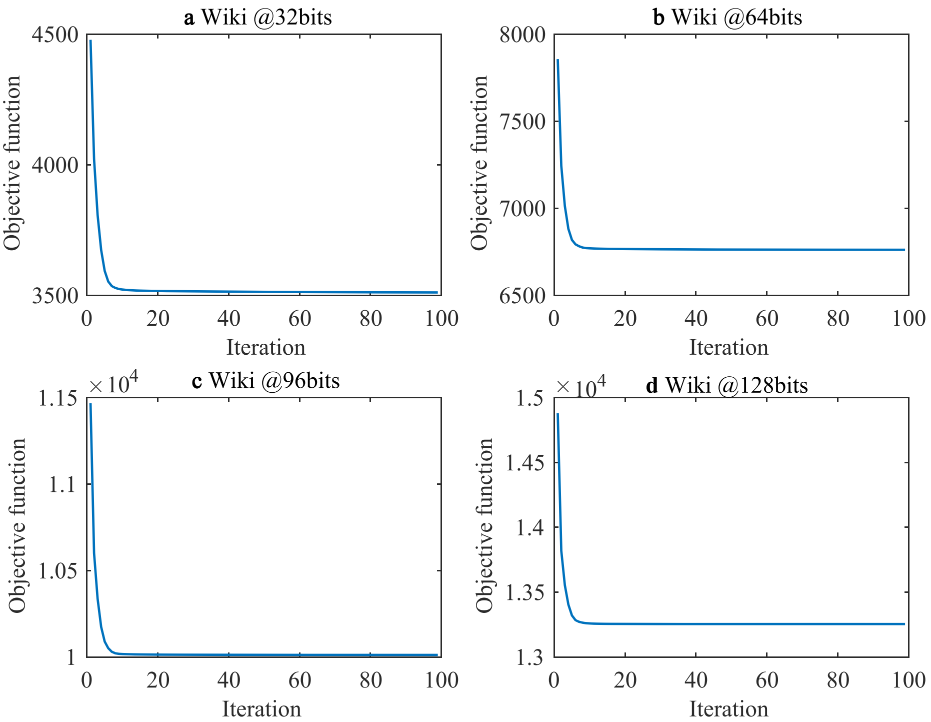


Fig. 3 The objective value of LGCH for **a** 32 bits, **b** 64bits, **c** 96 bits and **d** 128 bits cross-modal retrieval versus different iterations on Wiki dataset

Algorithm 1 Label guided correlation cross-modal hashing.

```
   Input: Code length $r$, parameter $\alpha$, $\beta$ and $\lambda$, number of iterations $N$,training data
          $X_1 \in R^{m_1 \times n}$ and $X_2 \in R^{m_2 \times n}$, label information $Y \in R^{k \times n}$.
   Output: The hashing codes $B \in R^{n \times r}$. The transformation matrices: $W_1 \in R^{m_1 \times r}$,
          $W_2 \in R^{m_2 \times r}$, $Q \in R^{r \times k}$ and $P \in R^{r \times k}$. The adaptive parameter $\sigma$.
 1 Initialize $\sigma \in (0, 1)$ , randomly initialize $W_1$ ,$W_2$ and $Q$. Compute G = ${W_1}^T X_1$,
   $B = sign(G)$;
 2 repeat
 3     Update $P$ via (12);
 4     Update $W_1$ via (14);
 5     Update $W_2$ via (15);
 6     Update $G$ via (10);
 7     Update $B$ via (19);
 8     If $\sigma = 1$;
 9         Jump to step 3, ;
10     else
11         Update $Q$ via (17);
12         Update $\sigma$ via (22);
13     end
14 until Convergence;
```

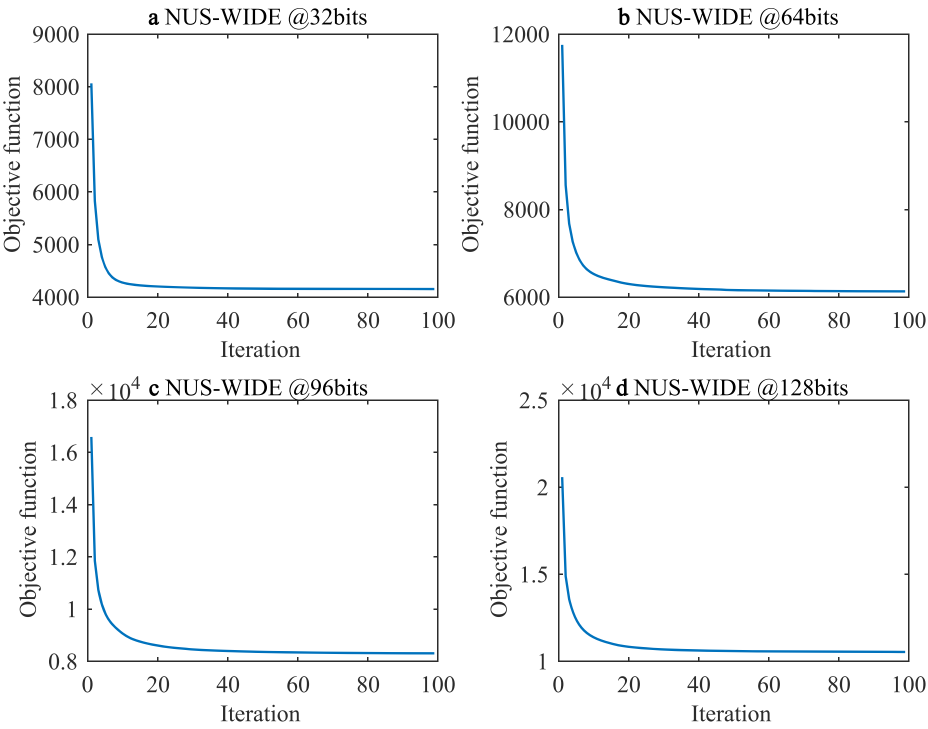


Fig. 4 The objective value of LGCH for **a** 32 bits, **b** 64bits, **c** 96 bits and **d** 128 bits cross-modal retrieval versus different iterations on NUS-WIDE dataset

The overall optimization procedure for LGCH is summarized in Algorithm 1. Since each sub-problem of our model is minimized via the update rules mentioned above, the objective value of (5) is no-increasing. Thus, the convergence of our optimization algorithm meets:

$$O(G^t, P^t, {W_1}^t, {W_2}^t, B^t, Q^t, \sigma^t) \leq O(G^{t+1}, P^{t+1}, {W_1}^{t+1}, {W_2}^{t+1}, B^{t+1}, Q^{t+1}, \sigma^{t+1}), \tag{23}$$

where $\cdot^t$ indicates the matrix '$\cdot$' at the t^{th} iteration wherein '$\cdot$' is a matrix placeholder. To accelerate the convergence of LGCH, we treat (24) as the termination condition of Algorithm 1.

$$\frac{|O_t - Q_{t+1}|}{|O_t|} \leq \zeta, \tag{24}$$

where the tolerance $\zeta = 10^{-4}$ is used in all the experiments. The objective curve is shown in Figs. 3, 4 and 5. As Figs. 3–5 shows, the objective value curve of LGCH takes on the monotone decreasing tendency, and respectively converges at around 10 iterations on Wiki dataset and 20 iterations on both NUS-WIDE and Flickr25K datasets.

Computational complexity The total number of iterations and the time cost per iteration are two aspects of the time complexity of iterative algorithms for our LGCH. As in Figs. 3–5, LGCH converges with relatively small total number of iterations, thus we only discuss the computational cost per iteration. For LGCH, the total time cost lies in seven steps (3-7 and 11-12) of Algorithm 1. In particular, solving B involves the sort algorithm and matrix multiplication, which respectively take $O(rnlogn)$ and $O(r^2(k+n))$ in time. For P, the

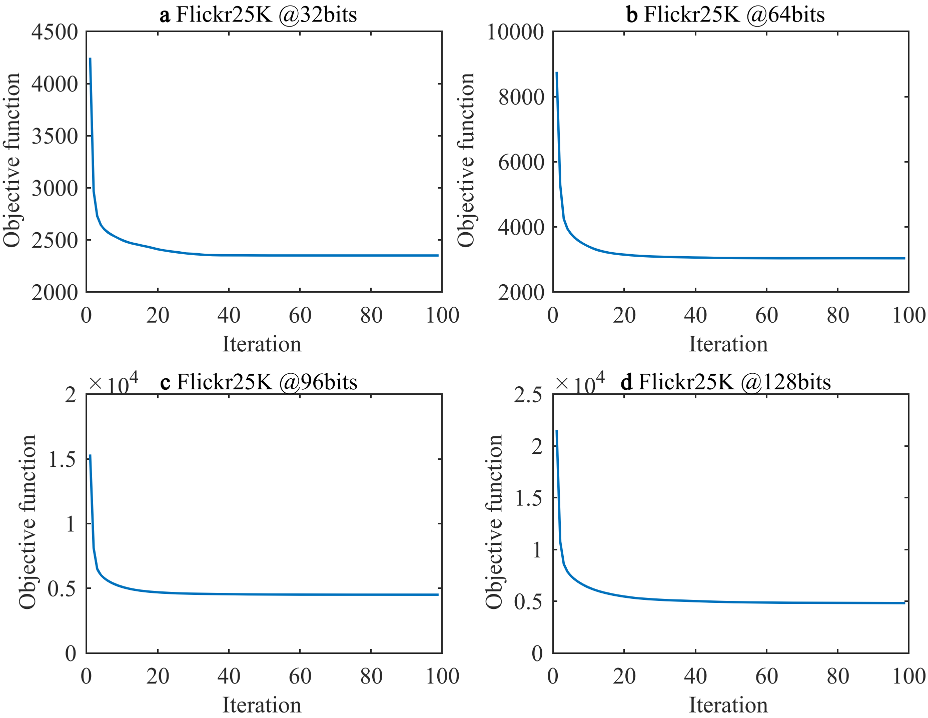


Fig. 5 The objective value of LGCH for **a** 32 bits, **b** 64bits, **c** 96 bits and **d** 128 bits cross-modal retrieval versus different iterations on Flickr25K dataset

time cost of step 3 mainly depends on the matrix multiplication and inverse operations, thus the total time cost of this step is $O(nkr + (n+k)r^2 + k^2r + k^3)$. Likewise, updating W_1, W_2, Q and G at most cost $O(n{m_1}^2 + {m_1}^3 + nm_1r)$, $O(n{m_2}^2 + {m_2}^3 + nm_2r)$, $O(nr^2 + kr^2 + k^2r + k^3)$ and $O(nkr + nm_1r + nm_2r + nr^2 + rn\min(r, n))$, respectively. For solving σ, since the involved matrix multiplication QP^T can be calculated in step 7, the time cost is relatively very small and can be omitted here. In summary, the total time cost per iteration of Algorithm 1 is $O(T(n(k+m_1+m_2)r + rnlogn + (n+k)r^2 + k^2r + k^3 + n{m_1}^2 + n{m_2}^2 + {m_1}^3 + {m_2}^3 + rnmin(r, n)))$. In practice, m_1, m_2, r and k are far less than n. Then, the main time cost of LGCH depends on $O(rnlogn)$. For comparison, two typical counterparts such as IMH [36] and CVH [19] are analyzed here. Particularly, the time cost per iteration is $O(n^3)$ for IMH, and at least $O(n^2)$ for CVH, respectively. In a nutshell, both IMH and CVH take higher time cost per iteration against LGCH.

New coming sample When a new sample $x \in R^{m_1}$ comes, if it is an image, we firstly map its image descriptor to the common subspace to gain the common latent representation $g = {W_1}^T x$, and then we solve the following problem:

$$\begin{aligned} &\min_{b} \left\| b - \sigma g - (1-\sigma) QP^T g \right\|_F^2 . \\ &s.t.\ b \in \{-1, 1\}^{r\times 1} \end{aligned} \tag{25}$$

Then, the solution of b is:

$$b = sign(\sigma g + (1-\sigma) QP^T g). \tag{26}$$

Table 1 The mAP results on Wiki, NUS-WIDE and Flickr25K dataset, respectively

Tasks	Methods	Dataset											
		Wiki				NUS-WIDE				Flickr25K			
		32 bits	64 bits	96 bits	128 bits	32 bits	64 bits	96 bits	128 bits	32 bits	64 bits	96 bits	128 bits
I→T	CMSSH	0.1759	0.1560	0.1581	0.1764	0.2641	0.2799	0.2981	0.3075	0.5039	0.5509	0.4946	0.5521
	SCM	0.1839	0.1648	0.1786	0.1657	0.4633	0.4430	0.4305	0.4255	0.6229	0.6288	0.6089	0.6118
	QCH	0.2153	0.2170	0.2019	0.1842	0.3783	0.4025	0.4033	0.4044	0.6285	0.6431	0.6398	0.6610
	CVH	0.1806	0.1582	0.1590	0.1687	0.4417	0.4302	0.4221	0.4195	0.6514	0.6459	0.6372	0.6334
	IMH	0.2007	0.1658	0.1577	0.1493	0.4781	0.4726	0.4695	0.4666	0.6941	0.6822	0.6744	0.6714
	CCQ	0.2366	0.2366	0.2366	0.2290	0.5212	0.5275	0.5304	0.5222	0.7320	0.7154	0.7181	0.6989
	CSGH	0.2658	**0.2682**	**0.2716**	0.2655	0.5108	0.5331	0.5405	0.5426	0.7214	0.7338	0.7434	0.7493
	LGCH	**0.2669**	0.2640	0.2689	**0.2671**	**0.5394**	**0.5358**	**0.5501**	**0.5483**	**0.7492**	**0.7631**	**0.7739**	**0.7792**
T→I	CMSSH	0.1762	0.2011	0.2386	0.2454	0.2827	0.2830	0.2840	0.2746	0.4758	0.5148	0.4193	0.6044
	SCM	0.2376	0.2053	0.2027	0.2011	0.4754	0.4477	0.4442	0.4353	0.6635	0.6695	0.6579	0.6560
	QCH	0.3970	0.3198	0.3034	0.2442	0.4198	0.4073	0.4041	0.3994	0.6971	0.7100	0.7081	0.7177
	CVH	0.2517	0.2231	0.2237	0.2178	0.4333	0.4222	0.4193	0.4179	0.6984	0.6954	0.7039	0.6925
	IMH	0.2784	0.2384	0.2344	0.2062	0.4832	0.4758	0.4685	0.4661	0.6928	0.6837	0.6628	0.6667
	CCQ	0.4077	0.4130	0.4171	0.3545	0.5428	0.5524	0.5532	0.5502	0.7288	0.7378	0.7479	0.7565
	CSGH	0.4830	0.4923	0.4848	0.5051	0.5209	0.5496	0.5637	0.5692	0.7309	0.7415	0.7490	0.7575
	LGCH	**0.4891**	**0.5023**	**0.4925**	**0.5104**	**0.6033**	**0.6065**	**0.6180**	**0.6390**	**0.7678**	**0.7614**	**0.7679**	**0.7811**

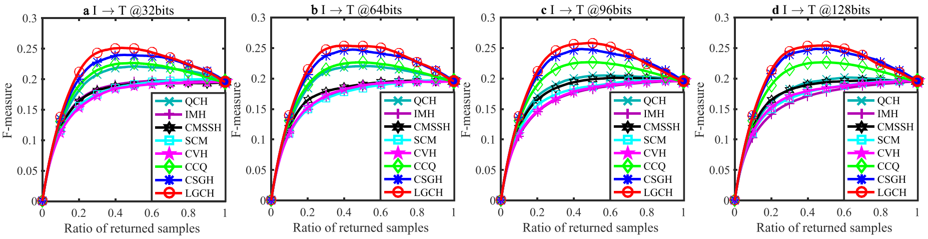


Fig. 6 The F-measure results of different bits for I → T task on Wiki dataset

Likewise, if a new sample belonged to the text data, let $g = {W_2}^T x$, and we can compute its binary codes using (26).

4 Experiments

In this section, we verify the effectiveness of LGCH by conducting experiments of cross-modal retrieval on three popular cross-modal datasets including Wiki, NUS-WIDE and Flickr25K.

4.1 Datasets

- The Wiki dataset contains 2,866 images and corresponding texts collected from Wikipedia's featured articles. All of the image-text pairs are labelled by a 10-dimensional vector of 10 categories. Each image is represented by a 128-dimensional bag-of-words feature vector based on SIFT, while each text sample is represented by a 10-dimensional feature vector. Following [20, 30, 58], we set the released 2173 image-text pairs as training set and the 693 pairs as testing set. We treat the training dataset as the gallery database.
- The NUS-WIDE dataset comprises 269,648 images and the associated tags from Flickr, with the 81ground-truth concepts used for evaluation. Following [20, 58], we select the most 21 frequently-used concepts to construct a new dataset of 195,834 image-text pairs. Each image is represented by a 500-dimensional bag of words feature vector based on SIFT descriptors, and meanwhile each text is encoded as a 1000-dimensional tag occurrence feature vector. From the new dataset, we randomly sample 2,000 image-text pairs as the testing set, and the rest image-text pairs are used as the gallery database.

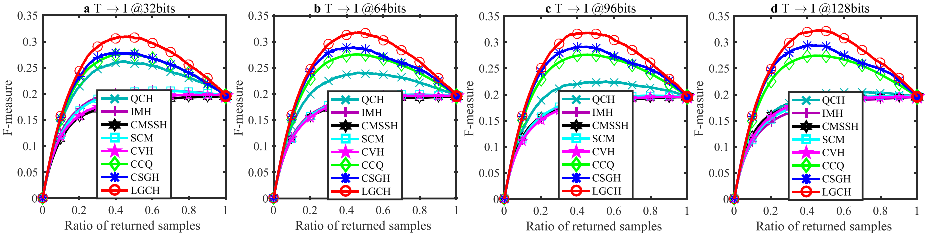


Fig. 7 The F-measure results of different bits for T → I task on Wiki dataset

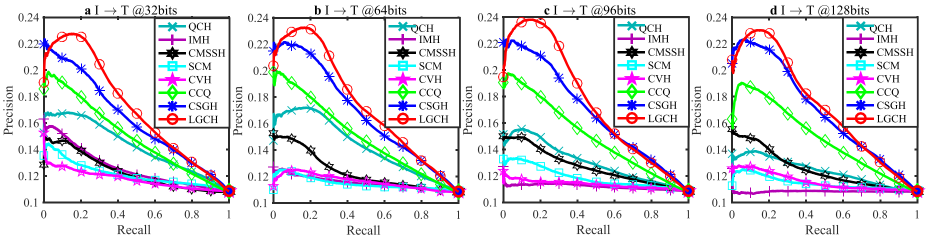


Fig. 8 The precision-recall curves of different bits for I → T task on Wiki dataset

From the gallery database, we randomly select 10,000 image-text pairs as the training set.

- The Flickr25K dataset is the labelled subset of Flicke1M dataset which contains 1 million Flickr image-text pairs. There are 38 categories in Flickr25K dataset. Each image is represented by a 3,857-dimensional feature vector and each text is represented by a 2,000-dimensional feature vector generated by tag occurrences [38]. We use 1,000 randomly selected image-text pairs as the testing set and the remaining 2, 4000 pairs as training set. Like Wiki dataset, we also treat the training set as the gallery database.

4.2 Evaluation protocols

As LGCH is an CCA based method, we should compare it with CCA based methods. Among those CCA based methods, SCM is a representative state-of-the-art one. We first choose to compare our method with the supervised method SCM [57]. Then we compare it with other six state-of-the-art cross-modal hashing methods including the supervised ones such as CMSSH [6] and QCH [52], and the unsupervised ones like CVH [19], IMH [36], CCQ [20] and CSGH [58]. Two types of cross-modal retrieval tasks are employed: 1) image to text task (I → T), we treat images as query and texts as the returned results; 2) text to image (T → I) task, we treat texts as query and images as the returned results. We employ the mean average precision to evaluate the retrieval effect of LGCH. Following [58], we report the MAP@50 results. That is to say, we calculate the mean average precision of top 50 retrieved samples. In addition, we also display the F-measure curves, the precision-recall curves and the precision@top-R curves results of the two retrieval tasks in figure, where R represent the number of retrieved samples.

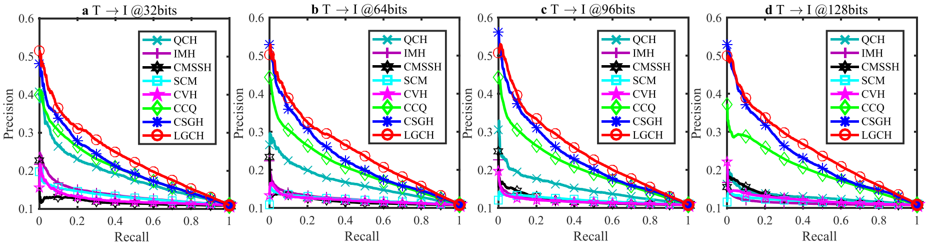


Fig. 9 The precision-recall curves of different bits for T → I task on Wiki dataset

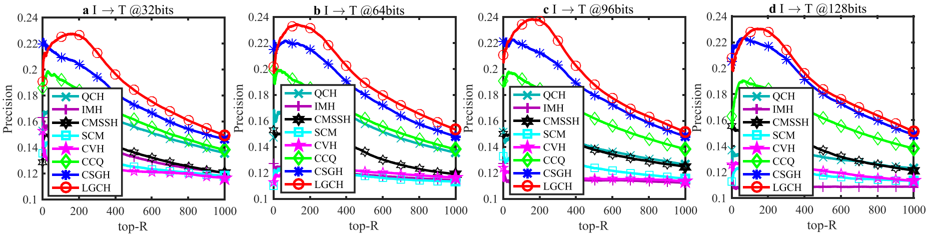


Fig. 10 The precision@topR curves of different bits for I → T task on Wiki dataset

LGCH has three parameters: α , β and λ. We explored the impact of these three parameters on LGCH for the above two cross-modal retrieval tasks. And finally, for both Wiki and NUS-WIDE dataset, we set $\alpha = 100$, $\beta = 0.001$ and $\lambda = 0.01$. For Flickr25K dataset, we set $\alpha = 0.001$, $\beta = 1$ and $\lambda = 0.01$.

4.3 Experiment results

For all datasets, we report the 32-bit, 64-bit, 96-bit and 128-bit of: the F-measure results, the mAP results, the precision-recall curves and the precision@topR curves of two cross-modal retrieval tasks: I → T and T → I, respectively. Particularly, the best mAP results are bold emphasized in Table 1.

Results on Wiki dataset The F-measure curves of LGCH comparing with other seven state-of-the-art methods on Wiki dataset are shown in Figs. 6 and 7. As Figs. 6–7 shows, LGCH is superior to the other seven compared models for two cross-modal retrieval tasks on Wiki dataset. In addition, the T → I task results in Fig. 7 shows that LGCH performs better compared with other methods when the bit is growing.

As shown in Table 1, the mAP results of Wiki dataset show the effectiveness of LGCH in most cases except for the T → I task at 64 bits and 96 bits, but there is little gap between LGCH and the best model in the table, also shows that LGCH is competitive.

The precision-recall curves results on Wiki dataset are shown in Figs. 8 and 9. Figure 8 shows the I → T task performance and Fig. 9 shows the T → I performance of LGCH and other seven state-of-the-art methods, and the results show the effectiveness of LGCH comparing with other methods.

The precision@topR results on Wiki dataset is shown in Figs. 10 and 11, and these results are similar to the above precision-recall results, which show that LGCH is superior to other methods in most cases. The main reason of this similarity may be that Wiki dataset is a

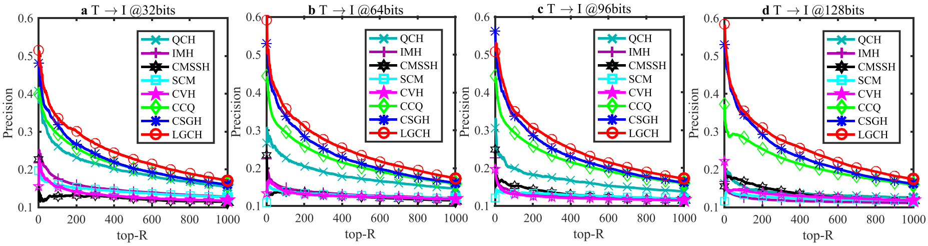


Fig. 11 The precision@topR curves of different bits for T → I task on Wiki dataset

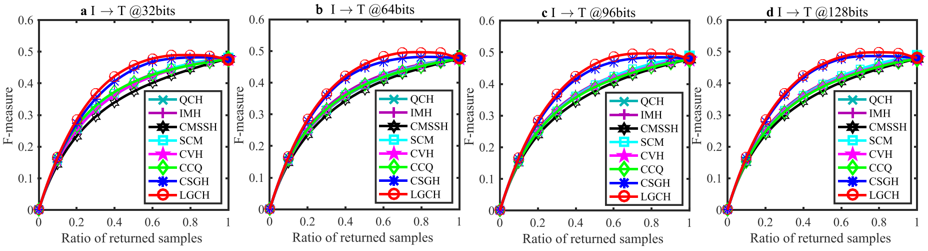


Fig. 12 The F-measure results of different bits for I → T task on NUS-WIDE dataset

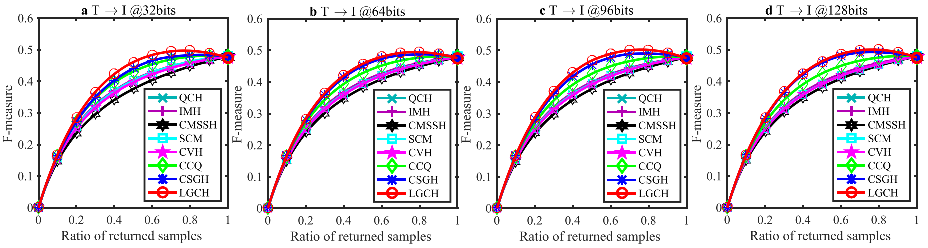


Fig. 13 The F-measure results of different bits for T → I task on NUS-WIDE dataset

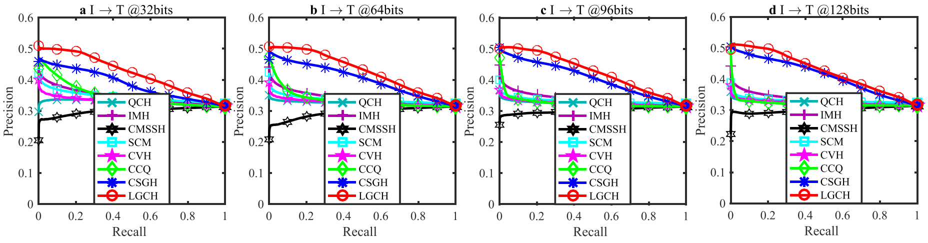


Fig. 14 The precision-recall curves of different bits for I → T task on NUS-WIDE dataset

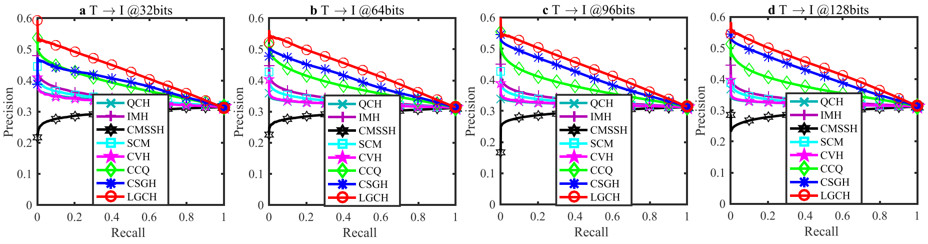


Fig. 15 The precision-recall curves of different bits for T → I task on NUS-WIDE dataset

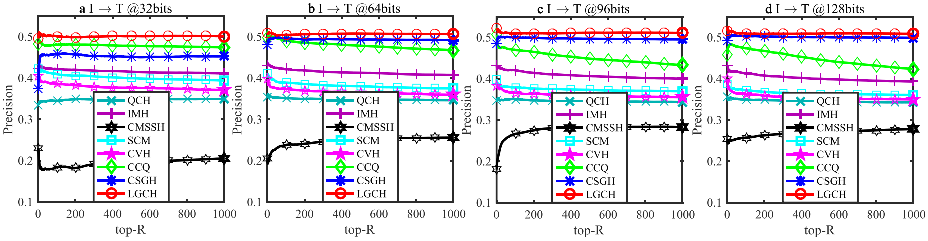


Fig. 16 The precision@topR curves of different bits for I → T task on NUS-WIDE dataset

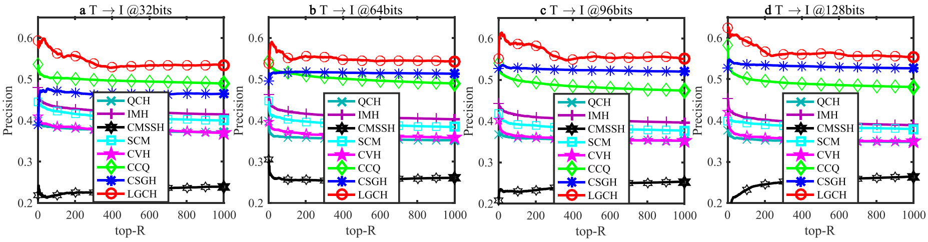


Fig. 17 The precision@topR curves of different bits for T → I task on NUS-WIDE dataset

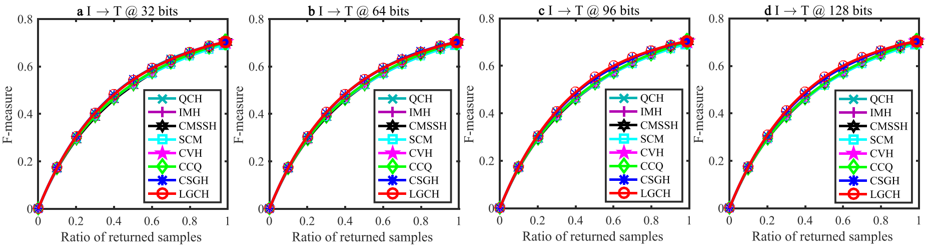


Fig. 18 The F-measure results of different bits for I → T task on Flickr25K dataset

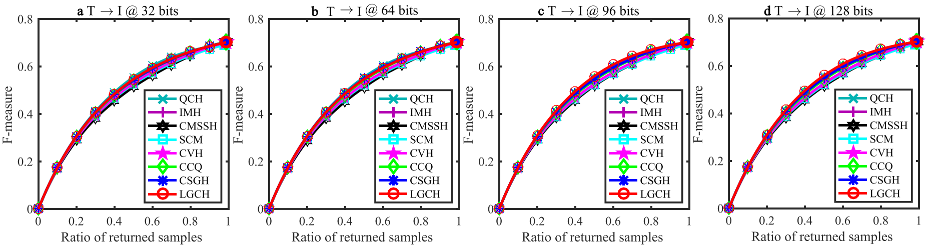


Fig. 19 The F-measure results of different bits for T → I task on Flickr25K dataset

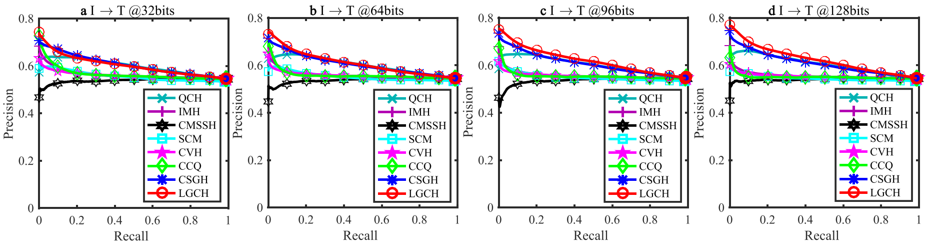


Fig. 20 The precision-recall curves of different bits for I → T task on Flickr25K dataset

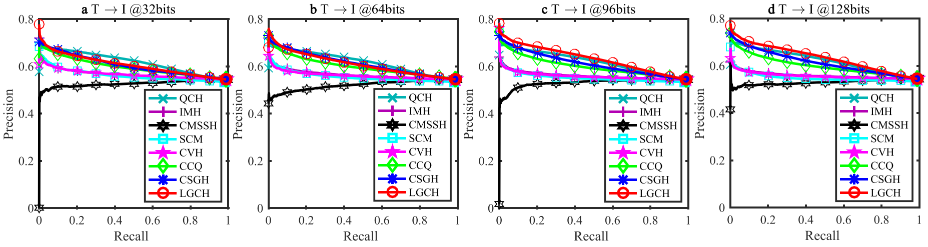


Fig. 21 The precision-recall curves of different bits for T → I task on Flickr25K dataset

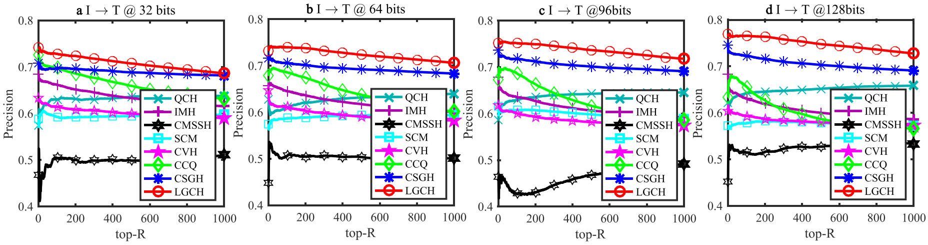


Fig. 22 The precision@topR curves of different bits for I → T task on Flickr25K dataset

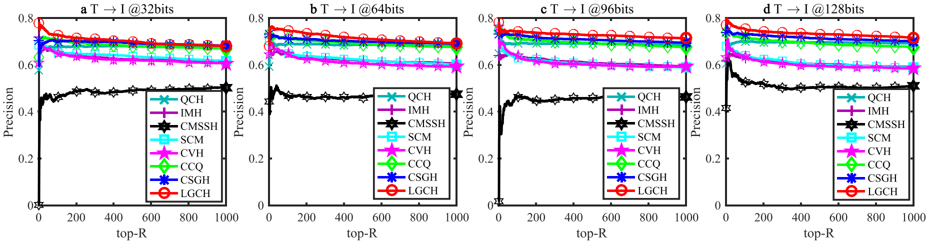


Fig. 23 The precision@topR curves of different bits for T → I task on Flickr25K dataset

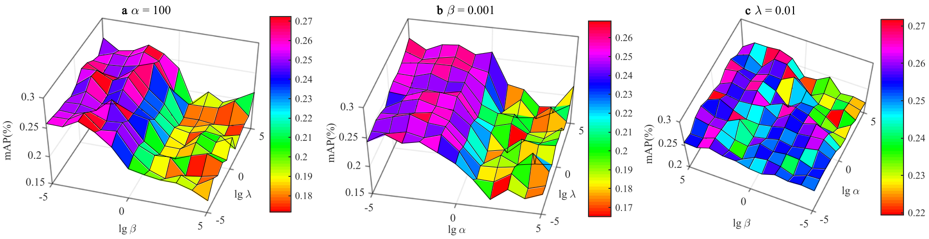


Fig. 24 The impact of α, β and λ for 32 bits I → T task on Wiki dataset: **a** the mAP results of β and λ when fixing $\alpha = 100$, **b** the mAP results of α and λ when set $\beta = 0.001$ and **c** the mAP results of α and β when set $\lambda = 0.01$

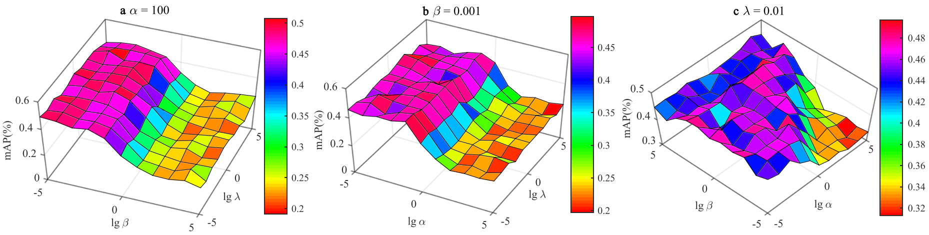


Fig. 25 The impact of α, β and λ for 32 bits T → I task on Wiki dataset: **a** the mAP results of β and λ when fixing $\alpha = 100$, **b** the mAP results of α and λ when set $\beta = 0.001$ and **c** the mAP results of α and β when set $\lambda = 0.01$

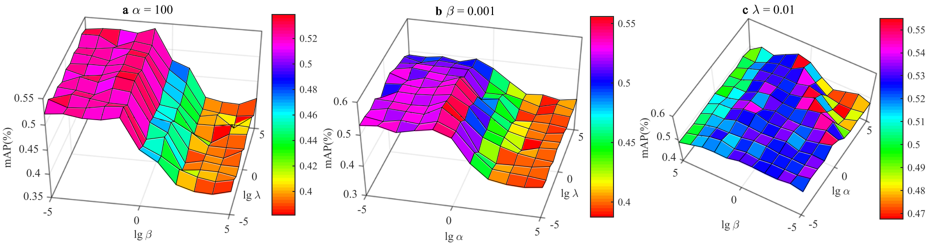


Fig. 26 The impact of α, β and λ for 32 bits I → T task on NUS-WIDE dataset: **a** the mAP results of β and λ when fixing $\alpha = 100$, **b** the mAP results of α and λ when set $\beta = 0.001$ and **c** the mAP results of α and β when set $\lambda = 0.01$

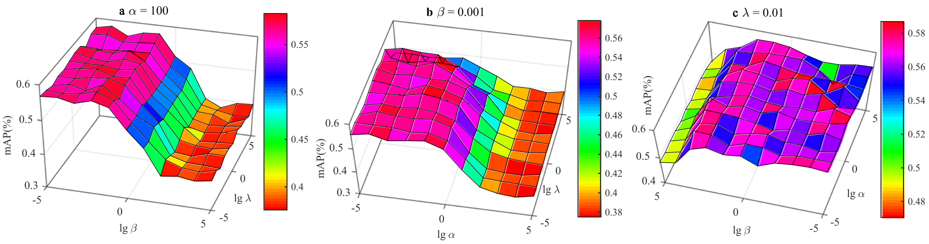


Fig. 27 The impact of α, β and λ for 32 bits T → I task on NUS-WIDE dataset: **a** the mAP results of β and λ when fixing $\alpha = 100$, **b** the mAP results of α and λ when set $\beta = 0.001$ and **c** the mAP results of α and β when set $\lambda = 0.01$

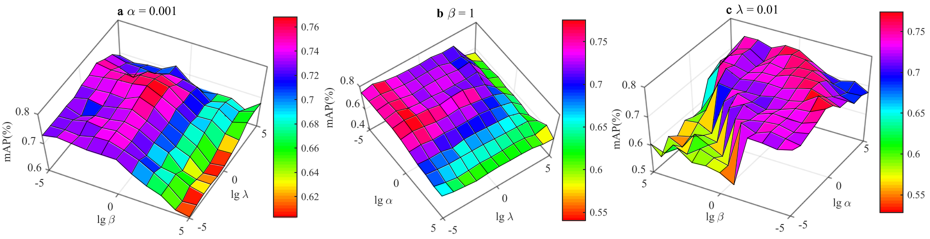


Fig. 28 The impact of α, β and λ for 32 bits I → T task on Flickr25K dataset: **a** the mAP results of β and λ when fixing $\alpha = 0.001$, **b** the mAP results of α and λ when set $\beta = 1$ and **c** the mAP results of α and β when set $\lambda = 0.01$

small dataset, thus the top1000 returned samples occupy a large proportion of the database and the precision@topR curves are in line with the trend of the precision-recall curves.

Results on NUS-WIDE dataset The F-measure results on NUS-WIDE dataset are shown in Figs. 12 and 13. From Figs. 12–13, we can easily draw the conclusion that LGCH is superior than all other methods for all compared bits at the two retrieval tasks.

The mAP results on NUS-WIDE dataset in Table 1 also show that LGCH can get better performance than all other compared methods, especially for the T → I task. The precision-recall curves on NUS-WIDE dataset are shown in Figs. 14 and 15 and the precision@topR curves are shown in Figs. 16 and 17. These results also show the effectiveness of LGCH: it is superior than all other compared methods.

Results on Flickr25K dataset The F-measure results on Flickr25K dataset are displayed in Figs. 18 and 19. All methods have not much difference in performance for both I → T and T → I tasks. Thus showing that LGCH is competitive with other compared methods.

The mAP results on Flickr25K dataset in Table 1 show that LGCH has better performance than other compared methods in I → T and T → I tasks, respectively. The precision-recall curves of different bits for both tasks on Flickr25K dataset are shown in Figs. 20 and 21, respectively. For 96 and 128 bits in Figs. 20 and 21, LGCH is superior to the other compared methods. For 32 and 64 bits in Figs. 20 and 21, LGCH outweighs the compared methods except CSGH and QCH: LGCH is competitive with these two methods, as in Fig. 20; it is competitive with CSGH and is slightly inferior to QCH, as shown in Fig. 21.

Although the precision-recall results on Flickr25K are not all better than other methods. The precision@top results in Figs. 22 and 23 show the superiority of LGCH comparing

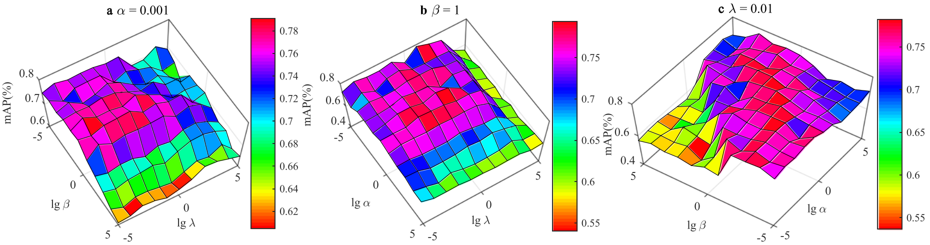


Fig. 29 The impact of α, β and λ for 32 bits T → I task on Flickr25K dataset: **a** the mAP results of β and λ when fixing $\alpha = 0.001$, **b** the mAP results of α and λ when set $\beta = 1$ and **c** the mAP results of α and β when set $\lambda = 0.01$

with all other methods, containing CSGH and QCH. In fact, real situations unnecessarily retrieve a large number of query results. For such situations, LGCH might be the better choice because it can work better than the other compared methods.

Impact of parameters LGCH has three free parameters α, β and λ, so we should analyze the impact of these parameters. We research the impact of $\alpha, \beta, \lambda \in \left\{10^i\right\}_{i=-5}^{5}$ for both I → T and T → I retrieval tasks of 32-bits on Wiki, NUS-WIDE and Flickr25K dataset, respectively. To visualize this point, we show the mAP results versus various values of two parameters with another parameter fixed.

The impact of three parameters on Wiki dataset is shown in Figs. 24 and 25. Figures 24a and 25a show that when α is fixed, the small β where $\beta \in \left\{10^i, lg\beta \leq 0\right\}_{i=-5}^{5}$ can induce better mAP results for both 32-bits I → T and T → I retrieval tasks on Wiki dataset. Likewise, when β is fixed, Figs. 24b and 25b show that the small α where $\alpha \in \left\{10^i, lg\beta \leq 0\right\}_{i=-5}^{5}$ have better mAP results for both 32-bits I → T and T → I retrieval tasks on Wiki dataset. When λ is fixed, Fig. 24c shows that the mAP results change slightly for 32 bits I → T task on Wiki dataset, but for T → I task in Fig. 25c, when $\alpha, \beta \in \left\{10^i, lg\beta \geq \lg\alpha - 4\right\}_{i=-5}^{5}$, LGCH has better mAP results.

The impact of three parameters on NUS-WIDE dataset is shown in Figs. 26 and 27. Like the results on Wiki dataset, Figs. 26a and 27a show that when α is fixed, the small β where $\beta \in \left\{10^i, lg\beta \leq 0\right\}_{i=-5}^{5}$ have better mAP results for both 32 bits I → T and T → I retrieval tasks on NUS-WIDE dataset. Likewise, when β is fixed, Figs. 26b and 27b show that the small α where $\alpha \in \left\{10^i, lg\beta \leq 0\right\}_{i=-5}^{5}$ have better mAP results for both 32 bits I → T and T → I retrieval tasks on NUS-WIDE dataset. When λ is fixed, Fig. 26c shows that $\alpha, \beta \in \left\{10^i, lg\beta \geq \lg\alpha - 4, lg\beta \leq 3\right\}_{i=-5}^{5}$ have better mAP results for 32 bits I → T task on NUS-WIDE dataset, and for T → I task in Fig. 27c, $\alpha, \beta \in \left\{10^i, lg\beta \leq 4\right\}_{i=-5}^{5}$ have better mAP results.

The impact of three parameters on Flickr25K dataset is shown in Figs. 28 and 29. When α is fixed, Fig. 28a shows that the small β where $\beta \in \left\{10^i, lg\beta \leq 1\right\}_{i=-5}^{5}$ have better mAP results for 32 bits I → T retrieval task on Flickr25K dataset, and Fig. 29a shows that $\beta, \lambda \in \left\{10^i, lg\beta \leq 2, lg\lambda \leq 2\right\}_{i=-5}^{5}$ have better mAP results for 32 bits T → I retrieval task on Flickr25K dataset. When β is fixed, Figs. 28b and 29b show that $\alpha, \lambda \in \left\{10^i, lg\alpha \leq 2, lg\lambda \leq 4\right\}_{i=-5}^{5}$ have better mAP results for both 32 bits I → T and T → I retrieval tasks on Flickr25K dataset. When λ is fixed, Figs. 28c and 29c show that $\alpha, \beta \in \left\{10^i, lg\alpha \geq \lg\beta - 4\right\}_{i=-5}^{5}$ have better mAP results for both 32 bits I → T and T → I tasks on Flickr25K dataset.

5 Conclusion

This paper proposes the label guided correlation cross-modal hashing method (LGCH), which can learn more discriminative common latent representation and binary codes by leveraging the label information as the guidance for simultaneously learning the binary codes and hashing function. This model is based on generalized CCA and can be easily extended to multi-media retrieval that contains more than two media types. When solving LGCH, it has an elegant analytical solution for each subproblem. The toy example illustrates the efficacy of label guided strategy, and experiments of cross-modal retrieval on three popular cross-modal datasets show that LGCH performs favorably as compared to several state-of-the-art cross-modal retrieval methods.

Acknowledgements This work was supported by the National Natural Science Foundation of China [61806213, U1435222]

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

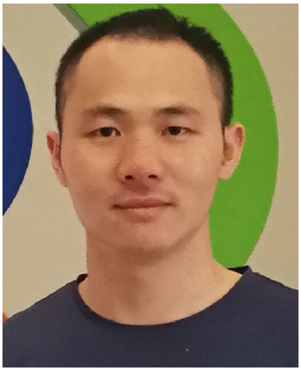
1. Akaho S (2006) A kernel method for canonical correlation analysis. arXiv:cs/0609071
2. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: International conference on machine learning, pp III–1247
3. Bay H, Tuytelaars T, Gool LJV (2006) SURF: speeded up robust features. In; European conference on computer vision, pp 404–417
4. Bay H, Ess A, Tuytelaars T, Gool LJV (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359
5. Benton A, Khayrallah H, Gujral B, Reisinger D, Zhang S, Arora R (2017) Deep generalized canonical correlation analysis. arXiv:1702.02519
6. Bronstein MM, Bronstein AM, Michel F, Paragios N (2010) Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: IEEE Conference on computer vision and pattern recognition, pp 3594–3601
7. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng YT (2009) Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval, p 48
8. Clinchant S, Ah-Pine J, Csurka G (2011) Semantic combination of textual and visual information in multimedia retrieval. In: International conference on multimedia retrieval, p 44
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on computer vision and pattern recognition, pp 886–893
10. Datar M, Immorlica N, Indyk P, Mirrokni VS (2004) Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the 20th ACM symposium on computational geometry, pp 253–262
11. Deng C, Chen Z, Liu X, Gao X, Tao D (2018) Triplet-based deep hashing network for cross-modal retrieval. IEEE Trans Image Process 27(8):3893–3903
12. Dong G, Zhang X, Lan L, Huang X, Luo Z (2018) Discrete graph hashing via affine transformation. In: IEEE International conference on multimedia and expo
13. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang TS (2018) Horizontal pyramid matching for person re-identification. In: AAAI Conference on artificial intelligence
14. Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233
15. Horst P (1961) Generalized canonical correlations and their applications to experimental data. J Clin Psychol 17(4):331–347
16. Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321–377
17. Huiskes MJ, Lew MS (2008) The mir flickr retrieval evaluation. In: Proceedings of the ACM SIGMM international conference on multimedia information retrieval, pp 39–43
18. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Neural information processing systems, pp 1106–1114
19. Kumar S, Udupa R (2011) Learning hash functions for cross-view similarity search. In: International joint conference on artificial intelligence, pp 1360–1365
20. Long M, Cao Y, Wang J, Yu PS (2016) Composite correlation quantization for efficient multimodal retrieval. In: Annual International ACM SIGIR conference on research and development in information retrieval, pp 579–588
21. Lowe DG (1999) Object recognition from local scale-invariant features. In: IEEE International conference on computer vision, pp 1150–1157
22. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

23. Ma D, Zhai X, Peng Y (2013) Cross-media retrieval by cluster-based correlation analysis. In: IEEE International conference on image processing, pp 3986–3990
24. Mirsky L (1975) A trace inequality of john von neumann. Monatshefte Fu̇r Mathematik 79(4):303–306
25. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: International conference on machine learning, pp 689–696
26. Ojala T, Pietikȧinen M, Harwood D (1994) Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: International conference on pattern recognition, pp 582–585
27. Peng Y, Huang X, Qi J (2016) Cross-media shared representation by hierarchical learning with multiple deep networks. In: International joint conference on artificial intelligence, pp 3846–3853
28. Peng Y, Huang X, Zhao Y (2017) An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges. IEEE Transactions on circuits and systems for video technology
29. Ranjan V, Rasiwasia N, Jawahar CV (2015) Multi-label cross-modal retrieval. In: IEEE International conference on computer vision, pp 4094–4102
30. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet G, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the ACM international conference on multimedia, pp 251–260
31. Rasiwasia N, Mahajan D, Mahadevan V, Aggarwal G (2014) Cluster canonical correlation analysis. In: Proceedings of the seventeenth international conference on artificial intelligence and statistics, pp 823–831
32. Shen X, Shen F, Sun Q, Yang Y, Yuan Y, Shen HT (2017) Semi-paired discrete hashing: learning latent hash codes for semi-paired cross-view retrieval. IEEE Trans Cybern 47(12):4275–4288
33. Shen X, Liu W, Tsang IW, Sun Q, Ong Y (2018) Multilabel prediction via cross-view search. IEEE Trans Neural Netw Learn Syst 29(9):4324–4338
34. Shen X, Shen F, Liu L, Yuan Y, Liu W, Sun Q (2018) Multiview discrete hashing for scalable multimedia search. ACM Trans Intell Syst Technol 9(5):53:1–53:21
35. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: IEEE International conference on computer vision, pp 1470–1477
36. Song J, Yang Y, Yang Y, Huang Z, Shen HT (2013) Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the ACM SIGMOD international conference on management of data, pp 785–796
37. Srivastava N, Salakhutdinov R (2012) Multimodal learning with deep Boltzmann machines. In: Neural information processing systems, pp 2231–2239
38. Srivastava N, Salakhutdinov RR (2014) Multimodal learning with deep Boltzmann machines. J Mach Learn Res 15(1):2949–2980
39. Tong H, He J, Li M, Zhang C, Ma W (2005) Graph based multi-modality learning. In: Proceedings of the ACM international conference on multimedia, pp 862–871
40. Wang Y, Wu L (2018) Beyond low-rank representations: orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Netw 103:1–8
41. Wang X, Li Z, Tao D (2011) Subspaces indexing model on grassmann manifold for image search. IEEE Trans Image Process 20(9):2627–2635
42. Wang X, Li Z, Zhang L, Yuan J (2011) Grassmann hashing for approximate nearest neighbor search in high dimensional space. In: IEEE International conference on multimedia and expo, pp 1–6
43. Wang X, Bian W, Tao D (2013) Grassmannian regularized structured multi-view embedding for image classification. IEEE Trans Image Process 22(7):2646–2660
44. Wang Y, Lin X, Wu L, Zhang W, Zhang Q (2015) LBMCH: learning bridging mapping for cross-modal hashing. In: Annual international ACM SIGIR conference on research and development in information retrieval, pp 999–1002
45. Wang Y, Lin X, Wu L, Zhang W, Zhang Q, Huang X (2015) Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Trans Image Process 24(11):3939–3949
46. Wang B, Yang Y, Xu X, Hanjalic A, Shen HT (2017) Adversarial cross-modal retrieval. In: Proceedings of the ACM international conference on multimedia, pp 154–162
47. Wang Y, Lin X, Wu L, Zhang W (2017) Effective multi-query expansions: collaborative deep networks for robust landmark retrieval. IEEE Trans Image Process 26(3):1393–1404
48. Wang Y, Zhang W, Wu L, Lin X, Zhao X (2017) Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. IEEE Trans Neural Netw Learn Syst 28(1):57–70
49. Wang Y, Wu L, Lin X, Gao J (2018) Multiview spectral clustering via structured low-rank matrix factorization. IEEE Trans Neural Netw Learn Syst 29(10):4833–4843
50. Weiss Y, Torralba A, Fergus R (2009) Spectral hashing. In: Neural information processing systems, pp 1753–1760

51. Wu L, Wang Y (2017) Robust hashing for multi-view data: jointly learning low-rank kernelized similarity consensus and hash functions. Image Vis Comput 57:58–66
52. Wu B, Yang Q, Zheng WS, Wang Y, Wang J (2015) Quantized correlation hashing for fast cross-modal search. In: International joint conference on artificial intelligence, pp 3946–3952
53. Wu L, Wang Y, Ge Z, Hu Q, Li X (2018) Structured deep hashing with convolutional neural networks for fast person re-identification. Comput Vis Image Underst 167:63–73
54. Wu L, Wang Y, Li X, Gao J (2018) Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Trans Cybern, 1–12
55. Wu L, Wang Y, Shao L (2019) Cycle-consistent deep generative hashing for cross-modal retrieval. IEEE Trans Image Process 28(4):1602–1612
56. Yang E, Deng C, Liu T, Liu W, Tao D (2018) Semantic structure-based unsupervised deep hashing. In: International joint conference on artificial intelligence, pp 1064–1070
57. Zhang D, Li W (2014) Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI Conference on artificial intelligence, pp 2177–2183
58. Zhang X, Dong G, Du Y, Wu C, Luo Z, Yang C (2018) Collaborative subspace graph hashing for cross-modal retrieval. In: International conference on multimedia retrieval, pp 213–221
59. Zhuang Y, Yang Y, Wu F (2008) Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Trans Multimed 10(2):221–229

Guohua Dong pursuing the Ph.D. degree from College of Computer, National University of Defense Technology. Her current research interests include image retrieval, cross-modal retrieval, hashing methods and graph model.

Xiang Zhang received the M.S., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 2010 and 2015, respectively. He is currently a research assistant with the Institute for Quantum Information & State Key Laboratory of High Performance Computing, College of Computer, National University of Defense Technology. His current research interests include computer vision and machine learning.



Long Lan is currently a lecturer with College of Computer, National University of Defense Technology. He received the Ph.D. degree in computer science from National University of Defense Technology 2017. He was a visiting Ph.D. student in University of Technology, Sydney from 2015 to 2017. His research interests include multi-object tracking, computer vision and discrete optimization.

Shiwei Wang pursuing the Ph.D. degree from College of Computer, National University of Defense Technology. Her current research interests include computer vision and machine learning.

Zhigang Luo received the B.S., M.S., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1981, 1993, and 2000, respectively. He is currently a Professor with the College of Computer, National University of Defense Technology. His current research interests include machine learning, computer vision, and bioinformatics.

Affiliations

Guohua Dong[1,2] · Xiang Zhang[2,3] · Long Lan[2,3] · Shiwei Wang[1,2] · Zhigang Luo[1,2]

✉ Long Lan
long.lan@nudt.edu.cn

✉ Zhigang Luo
zgluo@nudt.edu.cn

[1] Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha, Hunan, 410073, People's Republic of China

[2] College of computer, National University of Defense Technology,
Changsha, Hunan, 410073, People's Republic of China

[3] Institute for Quantum Information & State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, Hunan, 410073, People's Republic of China