语义耦合相关的判别式跨模态哈希学习算法

严双咏 刘长红 江爱文 叶继华 王明文
(江西师范大学计算机信息工程学院 南昌 330022)

摘 要 基于哈希的跨模态检索以其存储消耗低、查询速度快等优点受到广泛的关注.跨模态哈希学习的核心问题是如何对不同模态数据进行有效地共享语义空间嵌入学习.大多数算法在对多模态数据进行共享空间嵌入的过程中忽略了特征表示的语义判别性,从而导致哈希码表示的类别区分性不强,降低了最近邻搜索的准确性和鲁棒性.该文提出了基于语义耦合相关的判别式跨模态哈希特征表示学习算法.算法在模型的优化目标函数设计上综合了线性判别分类器的思想和跨模态相关性最大化思路,通过引入线性分类器,使得各模态都能够分别学习到各自具有判别性的二进制哈希码.同时利用耦合哈希表示在嵌入语义空间中最大化不同模态之间的相关性,不仅克服了把多种数据投影到一个共同嵌入语义空间的缺陷,而且能够捕捉到不同模态之间的语义相关性.算法在 Wiki、LabelMe 以及 NUS_WID 三个基准数据集上与最近相关的算法进行了实验比较.实验结果表明该文提出的方法在检索精度和计算效率上有明显的优势.

关键词 跨模态检索;跨模态哈希;线性分类器;语义相关性;共享子空间;多模态
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00164

Discriminative Cross-Modal Hashing with Coupled Semantic Correlation

YAN Shuang-Yong LIU Chang-Hong JIANG Ai-Wen YE Ji-Hua WANG Ming-Wen
(*School of Computer and Information Engineering, Jiangxi Normal University, Nanchang* 330022)

Abstract A variety of multimedia data on the network have increased exponentially in recent years including multi-modal data, such as video, picture, audio, text, etc. Different modal data are often interrelated. For example, in WeChat's moments, voice and short videos are often given when publishing pictures. When searching a topic, users expect to get rich and comprehensive retrieval results which include different media data, so how to achieve the cross-modal retrieval between different modal data has become a research hotspot in the multimedia field. The cross-modal retrieval methods based on hashing has attracted much attention for their low storage cost and fast query speed. The core problem of cross-modal hashing learning is how to learn efficiently the shared embedding semantic space of different modal data. There are two categories of approaches to handle the problem. The first category is the unsupervised methods, trying to learn the hashing function from the underlying structure, distribution, and topology information of the data in order to maintain the original data space structure. The second category is the supervised methods to combine the semantic label information in the process of the hashing learning. However, Most of algorithms neglect the semantic discrimination of feature representation in the

收稿日期:2017-04-13;在线出版日期:2018-05-15.本课题得到国家自然科学基金(61662030,61365002,61462042,61462045)、江西省自然科学基金(20171BAB202016)、江西省教育厅科技项目(GJJ150350)资助.**严双咏**,男,1990 年生,硕士研究生,主要研究方向为信息检索、计算机视觉.E-mail:13170884058@163.com.**刘长红**(通信作者),女,1977 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、机器学习、高光谱图像处理.E-mail:liuch@jxnu.edu.cn.**江爱文**,男,1984 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为模式识别、图像分析与检索、机器学习.**叶继华**,男,1966 年生,硕士,教授,中国计算机学会(CCF)会员,主要研究领域为数据融合、模式识别与物联网技术.**王明文**,男,1965 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索.

process of embedding the multi-modal data into the shared space, which leads to weaken the classification discrimination of the hash codes from different classes and reduce accuracy and robustness of the nearest neighbor search. In this paper, a linear discriminative cross-modal hashing learning algorithm with coupled semantic correlation is proposed, which integrates linear discriminative classifier and maximizing the correlation between cross-modals in the objective function of the model. First, we apply the linear classifier into modeling the supervised hashing learning so that each modal can learn respectively the discriminative binary hash code with high classification performance. Second, we project data from different modes into their embedding spaces to get their respective hash codes, and then the correlations between different modalities are maximized in the embedding spaces by joint coupled-hashing representation, so not only the defects of projecting a variety of data into a common embedding semantic space are overcome, but also the semantic relevance between different modal data can be captured. In the experiments, three kinds of performance evaluation indexes were employed, including the mean average precision (MAP) for ten times, the precision recall curve (PR) which implies the retrieval accuracy under different recall rates and the top N precision that indicates the change of accuracy relative to the number of the retrieval instances. In order to show the effectiveness of this algorithm, we compared it with six current relevant algorithms on three benchmark datasets including two cross-modal retrieval tasks: 1) the retrieving pictures with text; 2) the retrieving text with pictures. The experimental results show that the proposed method achieves obvious advantages on the retrieval accuracy and the computational efficiency. Additionally, the influence of the algorithm's parameters on its performance was also investigated by changing one parameter while fixing other parameters. The investigation demonstrates the proposed method is insensitive to the parameters varieties in a wide range and obtained good results.

Keywords　cross-modal retrieval; cross-modal hashing; linear classifier; semantic correlation; shared subspace; multi-modal

1　引　言

近年来各种多媒体数据呈指数增长，而且不同模态数据之间是互相相关的. 例如在网络相册中，图片经常是和一段相关性的文字描述配套出现；微信朋友圈里，在发布图片时还经常配有语音和短视频. 用户在搜索某个主题信息时，期望返回的结果列表是尽可能丰富的、主题相关的不同媒体数据，以获得全面的检索结果. 因此，如何实现不同模态间的跨模态检索已成为多媒体领域的研究热点[1-4].

近年来，由于数据的哈希特征表示具有存储空间小和检索速度快、通讯开销低等优点，因此在大规模信息检索领域逐渐得到广泛的关注和重视[5-8]. 跨模态哈希(Cross-Modal Hashing)技术的主要目的是通过哈希码映射函数的学习，实现不同模态数据的有效共享空间嵌入及相似性匹配与对齐[9-11]. 目前主流的跨模态哈希学习大致可分为无监督的方法和有监督的方法.

传统的无监督跨模态哈希方法试图从数据的底层结构、分布以及拓扑信息[12-13]来学习哈希函数，以保持原始数据空间结构，如 Cross-View Hashing (CVH)[9]、Semi-Paired Discrete Hashing(SPDH)[14]、Inter-Media Hashing (IMH)[10] 和 Co-Regularized Hashing (CRH)[11]等. 通过学习不同模态的哈希函数，使不同模态的数据能够映射成相似的二进制哈希码，但是这类方法搜索和存储代价都非常大. 为了克服这些方法的缺点，统一哈希码(Unified Hash Codes)表示的方法赢得了广大关注[15-16]. 它们将不同模态表示为统一的哈希码. Collective Matrix Factorization Hashing(CMFH)[15] 通过协同矩阵分解的方法将不同模态数据映射到共同的潜在语义空间，根据共同的潜在语义表示求解得到统一的哈希码. LSSH[16] 则采用稀疏编码和矩阵分解的方法分别学习图像和文本两种模态各自的潜在语义特征，并将所学到的潜在语义特征映射到联合概念空间.

统一哈希码表示的方法确实能够更有效地节省多模态数据的存储空间,但这种方法强制迫使不同模态的数据投影到共同潜在空间以得到一致的表示.然而不同模态下的数据特性的差异非常大,这种表示不能很好地捕捉各模态下原始数据的特性.为了更好地让不同模态数据在各自潜在空间进行投影,又保持模态间的语义相关性,在文献[17]中则提出了一种结合耦合哈希表示(Joint Coupled-Hashing,JCH)的跨模态检索方法.这种检索方法采用矩阵分解的方法分别学习两种模态各自的潜在语义特征,并结合耦合哈希表示最大化模态间的相关关系,从而实现跨模态哈希检索.

有监督的跨模态哈希方法的主要思路是在哈希学习的过程中结合语义类标信息.目前大多数方法是通过图约束的方式来保持模态间的语义相似性或模态内的结构一致性,如 Semantic Correlation Maximization(SCM)[18]和 Supervised Matrix Factorization Hashing(SMFH)[19-20].SCM[18]将语义类标结合到哈希学习过程,保持模态间所学哈希码的语义相似性.文献[19-20]在 CMFH[15]模型的基础上从不同角度引入了语义类标信息.其中,文献[19]在协同矩阵分解过程中结合语义类标信息保持共同潜在语义特征的一致性,同时在哈希学习过程保持所学哈希码的结构一致性,而文献[20]则在潜在语义特征学习过程中既保持了模态间的语义相似性又保持了模态内的结构一致性.然而,这些方法并没有直接捕获不同类别之间的判别信息.因此,所学到的哈希码缺少类间数据的区分能力.最近的Discrete Cross-modal Hashing(DCH)[21],将不同模态的哈希函数、统一哈希表示和线性分类器进行联合学习,基于 Supervised Discrete Hashing(SDH)[22]能够直接学习得到具有一定判别力的哈希码.

针对统一哈希码表示所存在的缺陷和有监督模态哈希方法的语义判别性优点,本文提出了一种结合线性分类器和耦合哈希表示的跨模态检索方法.这个检索方法通过引入线性分类器,使得各模态都能够分别学习到各自具有判别性的二进制哈希码,同时通过耦合哈希表示在嵌入语义空间中对不同模态之间进行交叉表示.因此这个检索方法不仅克服了把多种数据投影到一个共同嵌入语义空间的缺陷,而且能够捕捉到不同模态之间的语义相关性,使所学到哈希码具有更好的判别力.

2 相关工作

随着网络多媒体数据的不断增加,基于哈希的最近邻搜索方法由于低存储量和高效性在许多应用场景中都取得了巨大成功.目前已有的哈希方法大致可分为两类:单模态哈希和多模态哈希.

2.1 单模态哈希

单模态哈希的目的是对单模态数据的单一特征学习紧凑的哈希码.主要有两类方法:数据独立的哈希方法和数据依赖的哈希方法.基本的单模态哈希方法是数据独立的哈希方法,它利用随机投影去重建哈希函数而不是利用训练数据.这类方法中代表性的有局部敏感哈希(Locality-Sensitive Hashing,LSH)[23]和基于核的局部敏感哈希(Kernelized Locality-Sensitive Hashing,KLSH)[24].后来数据依赖哈希方法逐渐受到关注,它是在数据独立的哈希方法的基础上应用机器学习方法产生更加紧凑的数据相关哈希函数.在这方面最早的研究工作是谱哈希(Spectral Hashing)[25],他利用数据分布和一个图拉普拉斯矩阵将问题转化为特征值分解问题.为了消除二值化过程中的量化损失,Gong 等人[26]提出了迭代量化的方法(Iterative Quantization,ITQ),其目的是为了找到一个最优的旋转矩阵使最后得到的二元码和原始数据之间的差异最小.在利用数据之间相关性所生成的哈希函数上,数据独立的哈希方法比数据依赖的哈希方法可获得更好的效果.为了使所产生的二元码具有判别性,Strecha 等人[27]提出了 LDA 哈希,该方法利用线性判别分析(LDA)把原始数据嵌入到一个低维的空间中,然后将嵌入空间的数据点进行阈值化来,产生最后的二元码.在此基础之上,文献[28-29]分别提出了局部保持的判别式哈希(Locality preserving discriminative hashing,LPDH)和有监督的判别式哈希(Supervised Discriminative Hashing,SDH),它们分别从局部和全局上保持了邻域几何相似性和判别性.与 LDA 哈希方法不同的 Self-Taught Hashing(STH)[30]则利用线性分类器学习具有判别性的二元码的特性,通过最小化加权平均汉明距离得到训练数据的最优二元码,然后以此训练线性分类器去学习测试数据的哈希函数.

2.2 多模态哈希

多模态学习也叫做多视觉学习,在图像分类中被使用广泛,同时被更多地应用于跨模态检索.多模态哈希方法是在多模态检索中结合哈希方法的一种多模态检索方法.多模态哈希的目的是促使来自不同模态之间的数据实现快速的检索.多模态哈希方法大致可分为有监督多模态哈希和无监督多模态哈希两种.

无监督多模态哈希是从数据分布上去学习哈希函数.通常将原始空间的多模态数据投影到一个共同的潜在语义空间中,同时保持原始空间数据间的几何结构.IMH[10]在投影到共同潜在语义空间的同时维持模态内和模态间的相似性,用线性回归方法学习哈希函数.为了避免在 IMH 中进行大规模的图重建,线性跨模态哈希(Linear Cross-Modal Hashing,LCMH)[31]通过用少量的聚类中心来表示训练数据,然后计算每个训练数据点到聚类中心之间的距离.而协同矩阵分解哈希(CMFH)[15]是第一个在每个模态上使用协同矩阵分解的方法来获得统一的哈希码从而实现跨模态检索的方法.潜在语义稀疏哈希(LSSH)[16]则假设同一个实例的不同模态的哈希码是一样的,然后对图像和文本分别采用稀疏编码和矩阵分解的方法,然后经过投影、量化得到最后的哈希码.

有监督多模态哈希方法的目的是从训练数据的标签中获得语义信息,与无监督方法比可以得到更高的精度.CVH[9]把传统的谱哈希方法扩展到多模态中,最终的目标函数优化问题转化为一个一般的特征值求解问题.Semantics-Preserving Hashing (SePH)[32]将原始空间中训练数据间的语义仿射矩阵看成一种概率分布,Hamming 空间中的哈希码之间的关系看成另一种概率分布,并通过最小化两概率分布间的 KL 散度(Kullback-Leibler Divergence)来保持原始空间和 Hamming 空间的语义结构相似性.在该方法中,将训练数据间的语义仿射关系作为监督信息,指导哈希码的学习.然而在训练数据较多的情况下,仿射矩阵的运算将导致复杂度高和存储消耗大等问题.SCM[18]为了避免这种矩阵运算,将语义类标集成到哈希学习过程当中.然而以上这些方法都忽略了类别信息.很显然来自同一类别的多模态数据通常具有某些共享属性,而来自不同类别的数据应该有一些判别属性.通过维持这些共享属性和判别属性能够提高哈希码的辨别力,从而提高搜索的精度.因此,利用类别信息学习具有判别性的哈希码是非常重要的.

本文提出一种新的跨模态检索模型,它能够充分考虑到类别信息,使得学习到的哈希码具有判别性,同时也形成了一个算法用于实现我们提出的模型.

3　基于线性判别分类和耦合哈希的表示

本文所提出的基于线性判别分类和耦合哈希的表示方法的框架示意图见图 1 所示.受 DCH[21]的跨模态哈希模型的启发,通过引入线性分类器使哈希码学习的过程是有监督的,但本文的方法不是将不同模态的数据投影到共同的潜在语义空间得到统一哈希码,而是将不同模态的数据投影到各自的潜在空间,得到不同模态的哈希码.同时通过耦合哈希表示在潜在语义空间中对不同模态之间进行交叉耦合表示,使不同模态的数据在潜在语义空间中保持语义相关性,从而不仅使得各模态都能够分别学习到各自具有判别性的二进制哈希码,而且又保持了它们之间的语义相关性.

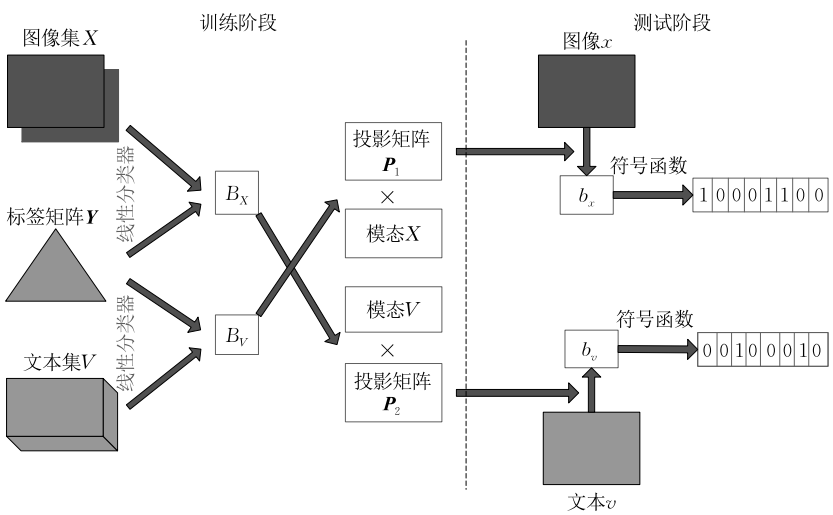


图 1　本文方法的框架示意图

3.1　基于线性判别分类的哈希表示

与 DCH[21]相类似,本文亦采用简单的线性分类器来建模有监督的哈希码学习,使所学到的哈希码表示具备较强的类别可区分能力,从而实现较好

的数据分类性能.

本文首先从图像和文本两个模态上来进行讨论.设 $\boldsymbol{X}=[x_1,\cdots,x_n]\in \mathrm{R}^{d\times n}$,$\boldsymbol{V}=[v_1,\cdots,v_n]\in \mathrm{R}^{m\times n}$分别为图像和文本数据矩阵,$d$ 和 m 分别表示图像和文本数据的特征维数,n 为样本个数.$\boldsymbol{Y}=[y_1,\cdots,y_n]\in \mathrm{R}^{c\times n}$为标签矩阵,其中 c 为总类别数,$\boldsymbol{B}=[b_1,\cdots,b_n]\in \mathrm{R}^{k\times n}$表示所学习到的哈希码矩阵,$k$ 为哈希码的位数.

对于一个样本点 x_i,假设其最终的二元码表示为 b_i和标签向量为 y_i,根据线性分类器模型可表示为式(1)所示.

$$\min_{\boldsymbol{B},\boldsymbol{W}}\sum_{i=1}^{n}\|y_i-\boldsymbol{W}b_i\|_F^2+\gamma\|\boldsymbol{W}\|_F^2 \tag{1}$$

其中 $\boldsymbol{W}\in \mathrm{R}^{c\times k}$是线性分类器的超平面权值系数,$\|\cdot\|$是 L_2范数,γ 是正则化参数.式(1)也可表示为式(2)所示:

$$\min_{\boldsymbol{B},\boldsymbol{W}}\|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{B}\|_F^2+\gamma\|\boldsymbol{W}\|_F^2 \tag{2}$$

将上述线性分类模型分别应用于图像和文本数据,得到相应的模型分别表示为式(3)和(4)所示:

$$\min_{\boldsymbol{B}_1,\boldsymbol{W}_{\boldsymbol{X}}}\|\boldsymbol{Y}-\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{B}_1\|_F^2+\gamma\|\boldsymbol{W}_{\boldsymbol{X}}\|_F^2 \tag{3}$$

$$\min_{\boldsymbol{B}_2,\boldsymbol{W}_{\boldsymbol{V}}}\|\boldsymbol{Y}-\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{B}_2\|_F^2+\gamma\|\boldsymbol{W}_{\boldsymbol{V}}\|_F^2 \tag{4}$$

其中矩阵 $\boldsymbol{W}_{\boldsymbol{X}}\in \mathrm{R}^{c\times k}$,$\boldsymbol{W}_{\boldsymbol{V}}\in \mathrm{R}^{c\times k}$分别为图像和文本的线性分类器的超平面权值系数,$\boldsymbol{B}_1\in \mathrm{R}^{k\times n}$,$\boldsymbol{B}_2\in \mathrm{R}^{k\times n}$分别为图像和文本对应的哈希码矩阵.

3.2 线性投影及耦合交叉最大化模态间相似性

对于图像和文本这两种模态数据,假设存在两个线性哈希函数能将它们分别投影到各自对应的潜在语义空间中,分别定义为

$$F_X(x_i)=\boldsymbol{P}_1x_i,\ F_V(v_i)=\boldsymbol{P}_2v_i \tag{5}$$

其中,$\boldsymbol{P}_1\in \mathrm{R}^{k\times d}$,$\boldsymbol{P}_2\in \mathrm{R}^{k\times m}$分别为两个投影矩阵.则通过投影之后,能得到两种模态对应的潜在空间的哈希码 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,表示为

$$\boldsymbol{B}_1=\boldsymbol{P}_1\boldsymbol{X},\ \boldsymbol{B}_2=\boldsymbol{P}_2\boldsymbol{V} \tag{6}$$

在跨模态检索中,一般认为同一个实例的不同模态之间是语义相关的,那么所学习到的哈希码 $\boldsymbol{B}_1$和 $\boldsymbol{B}_2$也应该具有良好的语义相关性.为了使不同模态的数据在潜在语义空间中保持语义相关性,我们通过耦合哈希表示的方法,在嵌在语义空间中对不同模态之间进行交叉耦合表示,使

$$\boldsymbol{B}_2\approx\boldsymbol{P}_1\boldsymbol{X},\ \boldsymbol{B}_1\approx\boldsymbol{P}_2\boldsymbol{V} \tag{7}$$

即它们之间应该满足下列两个交叉耦合的最小化目标形式.

$$\min_{\boldsymbol{B}_2,\boldsymbol{P}_1}\|\boldsymbol{B}_2-\boldsymbol{P}_1\boldsymbol{X}\|_F^2+\gamma\|\boldsymbol{P}_1\|_F^2 \tag{8}$$

$$\min_{\boldsymbol{B}_1,\boldsymbol{P}_2}\|\boldsymbol{B}_1-\boldsymbol{P}_2\boldsymbol{V}\|_F^2+\gamma\|\boldsymbol{P}_2\|_F^2 \tag{9}$$

为了使得各模态能够分别学习到具有判别性的哈希码同时又保持模态间语义相关性,结合式(3)、(4)、(8)和式(9),得到本文的目标函数为式(10)所示:

$$\begin{aligned}F=\min_{\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{W}_{\boldsymbol{X}},\boldsymbol{W}_{\boldsymbol{V}}}&\lambda\|\boldsymbol{Y}-\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{B}_1\|_F^2+\\&(1-\lambda)\|\boldsymbol{Y}-\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{B}_2\|_F^2+\alpha\|\boldsymbol{B}_2-\boldsymbol{P}_1\boldsymbol{X}\|_F^2+\\&\beta\|\boldsymbol{B}_1-\boldsymbol{P}_2\boldsymbol{V}\|_F^2+\gamma(\|\boldsymbol{W}_{\boldsymbol{X}}\|_F^2+\\&\|\boldsymbol{W}_{\boldsymbol{V}}\|_F^2+\|\boldsymbol{P}_1\|_F^2+\|\boldsymbol{P}_2\|_F^2)\end{aligned} \tag{10}$$

式中 λ 是权重系数,α,β 为平衡参数,γ 为正则化参数以用来防止过拟合.该目标函数综合考虑了所要学习的哈希码的语义鉴别能力和跨模态的有效交叉检索性能.

3.3 优化算法

对于含有 6 个矩阵变量 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,$\boldsymbol{P}_1$,$\boldsymbol{P}_2$,$\boldsymbol{W}_{\boldsymbol{X}}$,$\boldsymbol{W}_{\boldsymbol{V}}$来说,优化目标函数式(10)是非凸的,直接求解非常困难.但是我们不难发现,当固定其中任意 5 个变量,求解剩下的一个变量时,目标函数(10)均是凸的.因此,我们采取交替迭代的方法进行优化求解,具体的优化过程如下所述.

(1)固定 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,$\boldsymbol{P}_1$,$\boldsymbol{P}_2$和 $\boldsymbol{W}_{\boldsymbol{V}}$,式(10)对 $\boldsymbol{W}_{\boldsymbol{X}}$求偏导:

$$\frac{\partial F}{\partial \boldsymbol{W}_{\boldsymbol{X}}}=-2\lambda\boldsymbol{Y}\boldsymbol{B}_1^{\mathrm{T}}+2\lambda\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{B}_1\boldsymbol{B}_1^{\mathrm{T}}+2\gamma\boldsymbol{W}_{\boldsymbol{X}}=0 \tag{11}$$

可以求解得到

$$\boldsymbol{W}_{\boldsymbol{X}}=\boldsymbol{Y}\boldsymbol{B}_1^{\mathrm{T}}\left(\boldsymbol{B}_1\boldsymbol{B}_1^{\mathrm{T}}+\frac{\gamma}{\lambda}\boldsymbol{I}\right)^{-1} \tag{12}$$

其中 $\boldsymbol{I}$ 为单位矩阵.

(2)固定 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,$\boldsymbol{P}_1$,$\boldsymbol{P}_2$和 $\boldsymbol{W}_{\boldsymbol{X}}$,式(10)对 $\boldsymbol{W}_{\boldsymbol{V}}$求偏导:

$$\frac{\partial F}{\partial \boldsymbol{W}_{\boldsymbol{V}}}=-2(1-\lambda)\boldsymbol{Y}\boldsymbol{B}_2^{\mathrm{T}}+2(1-\lambda)\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{T}}+2\gamma\boldsymbol{W}_{\boldsymbol{V}}=0 \tag{13}$$

可以求解得到

$$\boldsymbol{W}_{\boldsymbol{V}}=\boldsymbol{Y}\boldsymbol{B}_2^{\mathrm{T}}\left(\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{T}}+\frac{\gamma}{1-\lambda}\boldsymbol{I}\right)^{-1} \tag{14}$$

(3)固定 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,$\boldsymbol{P}_2$,$\boldsymbol{W}_{\boldsymbol{X}}$和 $\boldsymbol{W}_{\boldsymbol{V}}$,式(10)对 $\boldsymbol{P}_1$求偏导:

$$\frac{\partial F}{\partial \boldsymbol{P}_1}=-2\alpha\boldsymbol{B}_2\boldsymbol{X}^{\mathrm{T}}+2\alpha\boldsymbol{P}_1\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+2\gamma\boldsymbol{P}_1=0 \tag{15}$$

可以求解得到

$$\boldsymbol{P}_1=\boldsymbol{B}_2\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}+\frac{\gamma}{\alpha}\boldsymbol{I}\right)^{-1} \tag{16}$$

(4)固定 $\boldsymbol{B}_1$,$\boldsymbol{B}_2$,$\boldsymbol{P}_1$,$\boldsymbol{W}_{\boldsymbol{X}}$和 $\boldsymbol{W}_{\boldsymbol{V}}$,式(10)对 $\boldsymbol{P}_2$求偏导:

$$\frac{\partial F}{\partial \boldsymbol{P}_2}=-2\beta\boldsymbol{B}_1\boldsymbol{V}^{\mathrm{T}}+2\beta\boldsymbol{P}_2\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+2\gamma\boldsymbol{P}_2=0 \quad (17)$$

可以求解得到

$$\boldsymbol{P}_2=\boldsymbol{B}_1\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\frac{\gamma}{\beta}\boldsymbol{I}\right)^{-1} \quad (18)$$

(5) 固定 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{W}_{\boldsymbol{X}},\boldsymbol{W}_{\boldsymbol{V}}$ 和 $\boldsymbol{B}_2$，式(10)对 $\boldsymbol{B}_1$ 求偏导：

$$\frac{\partial F}{\partial \boldsymbol{B}_1}=-2\lambda\boldsymbol{W}_{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{Y}+2\lambda\boldsymbol{W}_{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{B}_1+2\beta\boldsymbol{B}_1-2\beta\boldsymbol{P}_2\boldsymbol{V}=0 \quad (19)$$

可以求解得到

$$\boldsymbol{B}_1=(\boldsymbol{W}_{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{X}}+(\beta/\lambda)\boldsymbol{I})^{-1}(\boldsymbol{W}_{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{Y}+(\beta/\lambda)\boldsymbol{P}_2\boldsymbol{V}) \quad (20)$$

(6) 固定 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{W}_{\boldsymbol{X}},\boldsymbol{W}_{\boldsymbol{V}}$ 和 $\boldsymbol{B}_1$，式(10)对 $\boldsymbol{B}_2$ 求偏导：

$$\frac{\partial F}{\partial \boldsymbol{B}_2}=-2(1-\lambda)\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{Y}+2(1-\lambda)\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{B}_2+2\alpha\boldsymbol{B}_2-2\alpha\boldsymbol{P}_1\boldsymbol{X}=0 \quad (21)$$

可以求解得到

$$\boldsymbol{B}_2=\left(\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{V}}+\left(\frac{\alpha}{1-\lambda}\right)\boldsymbol{I}\right)^{-1}\left(\boldsymbol{W}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{Y}+\left(\frac{\alpha}{1-\lambda}\right)\boldsymbol{P}_1\boldsymbol{X}\right) \quad (22)$$

因此，通过不断地更新 $\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{W}_{\boldsymbol{X}},\boldsymbol{W}_{\boldsymbol{V}}$ 直到满足收敛阈值或者达到最大的迭代次数来优化目标函数式(10)的各参数. 具体的优化过程见算法 1.

算法 1.　本文的优化算法.

输入：图像矩阵 $\boldsymbol{X}$ 和文本矩阵 $\boldsymbol{V}$，语义标签矩阵 $\boldsymbol{Y}$，哈希码长度 k，参数 λ,α,β 和 γ

输出：线性投影矩阵 $\boldsymbol{P}_1,\boldsymbol{P}_2$ 和哈希码 $\boldsymbol{H}_{\boldsymbol{X}},\boldsymbol{H}_{\boldsymbol{V}}$

1. 用随机矩阵初始化 $\boldsymbol{B}_1,\boldsymbol{B}_2$，通过均值中心化 $\boldsymbol{X},\boldsymbol{V}$
2. Repeat
3. 用式(12)和(14)分别更新 $\boldsymbol{W}_{\boldsymbol{X}},\boldsymbol{W}_{\boldsymbol{V}}$；
4. 用式(16)和(18)分别更新 $\boldsymbol{P}_1,\boldsymbol{P}_2$；
5. 用式(20)和(22)分别更新 $\boldsymbol{B}_1,\boldsymbol{B}_2$；
6. Until 收敛阈值或迭代次数；
7. $\boldsymbol{H}_{\boldsymbol{X}}=\mathrm{sign}(\boldsymbol{B}_1),\boldsymbol{H}_{\boldsymbol{V}}=\mathrm{sign}(\boldsymbol{B}_2)$.

3.4　多模态扩展

本文所提出的方法也能够扩展应用于多模态(3 个以上的模态)的情况. 设 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_M$ 表示 M 个模态的数据矩阵，$\boldsymbol{B}_1,\cdots,\boldsymbol{B}_M$ 表示 M 个模态下所学习到的哈希码，$\boldsymbol{P}_1,\cdots,\boldsymbol{P}_M$ 表示 M 个模态下的投影矩阵，则对于基于线性分类的哈希表示部分和语义交叉耦合部分，我们可以定义如下表示形式.

对于 M 个模态的线性分类模型可表示为如下式(23)的形式.

$$L=\sum_{i=1}^{M}\lambda_i\|\boldsymbol{Y}-\boldsymbol{W}_i\boldsymbol{B}_i\|_F^2 \quad (23)$$

其中 $\sum_{i=1}^{M}\lambda_i=1$，$\boldsymbol{W}_i$ 是第 i 个线性分类器的超平面权值系数.

假设对于 2 个模态下的语义交叉耦合部分我们定义 T_2：

$$T_2=\alpha_1\|\boldsymbol{B}_2-\boldsymbol{P}_1\boldsymbol{X}_1\|_F^2+\beta_1\|\boldsymbol{B}_1-\boldsymbol{P}_2\boldsymbol{X}_2\|_F^2 \quad (24)$$

则 3 个模态下的语义交叉耦合部分可表示为

$$\begin{aligned}T_3&=\alpha_1\|\boldsymbol{B}_2-\boldsymbol{P}_1\boldsymbol{X}_1\|_F^2+\beta_1\|\boldsymbol{B}_1-\boldsymbol{P}_2\boldsymbol{X}_2\|_F^2+\\&\quad\alpha_2\|\boldsymbol{B}_3-\boldsymbol{P}_2\boldsymbol{X}_2\|_F^2+\beta_2\|\boldsymbol{B}_2-\boldsymbol{P}_3\boldsymbol{X}_3\|_F^2\\&=T_2+\alpha_2\|\boldsymbol{B}_3-\boldsymbol{P}_2\boldsymbol{X}_2\|_F^2+\beta_2\|\boldsymbol{B}_2-\boldsymbol{P}_3\boldsymbol{X}_3\|_F^2 \quad (25)\end{aligned}$$

以此类推，则当有 $i(i\geq 3)$ 个模态时，可以得到

$$T_i=T_{i-1}+\alpha_{i-1}\|\boldsymbol{B}_i-\boldsymbol{P}_{i-1}\boldsymbol{X}_{i-1}\|_F^2+\beta_{i-1}\|\boldsymbol{B}_{i-1}-\boldsymbol{P}_i\boldsymbol{X}_i\|_F^2 \quad (26)$$

则可以得到 M 个模态的优化目标函数为

$$F=\min_{\{\boldsymbol{B}_i,\boldsymbol{P}_i,\boldsymbol{W}_i\}_{i=1}^{M}}\left\{L+T_M+\gamma\sum_{i=1}^{M}(\|\boldsymbol{P}_i\|_F^2+\|\boldsymbol{W}_i\|_F^2)\right\} \quad (27)$$

对于多模态下的目标函数式(27)的优化同样可参照算法 1 的方法进行求解，即在求解 B_i 时，可固定其它参数 $B_1,B_2,\cdots,B_{i-1},B_{i+1},\cdots,B_M,\{\boldsymbol{P}_i\}_{i=1}^{M}$ 以及 $\{\boldsymbol{W}_i\}_{i=1}^{M}$，通过式(27)对 B_i 进行优化估计. 同样，其它的参数的求解方法可采用类似的方法，通过迭代更新能够分别求解出相应参数 $\{\boldsymbol{B}_i\}_{i=1}^{M}$，$\{\boldsymbol{P}_i\}_{i=1}^{M}$，$\{\boldsymbol{W}_i\}_{i=1}^{M}$.

4　实验与分析

本文在 Wiki、LabelMe 以及 NUS_WIDE 三个基准数据集上进行了实验验证，并与最近的无监督的跨模态哈希方法 CMFH[15]、LSSH[16] 和有监督的跨模态哈希方法 SMFCH_Liu[19]、SMFH_Tang[20]、耦合哈希方法 JCH[17] 和基于线性分类器的 DCH[21] 在两个跨模态检索任务上进行了性能比较和分析：(1) 用文本搜图片；(2) 用图片搜文本.

4.1　数据集

Wiki：Wiki 数据集来源于维基百科，共包含 2866 个文档. 每个文档包含一幅图片和至少 70 个单词组成的文字说明，其中每幅图片被表示成 128 维的 SIFT 特征，而文本则表述为 10 维的典型主题特征. 该数据集总共包含 10 个类别，每个文档分属其中一个类别. 如果两个文档的标签相同，则认为这两个文档相似.

LabelMe：LabelMe 数据集包含 2688 张图片.该数据集被划分为 8 个独特的室外场景，例如："海岸"、"森林"、"公路"等.每幅图像都属于其中的一个场景.每幅图片用 512 维的 GIST 特征进行表示，而每个文本由选中标签的索引向量来表示.如果图片文本对的场景类别相同，则认为它们是相似的.

NUS_WIDE：NUS_WIDE 数据集包含 269 648 张图片和 81 个语义信息.本文挑选其中的 10 个最大的语义信息，总共 186 577 个标记好的图像文本对用于本文的实验.在该数据集中图像表示为 500 维的视觉特征，文本表示为 1000 维的 BOW 特征，本文选择 4000 个图像文本对用于测试集，剩下的作为训练集，见表 1 所示.

表 1 实验所使用数据集的统计信息

数据集	Wiki	LabelMe	NUS_WIDE
数据集大小	2866	2688	186577
训练集大小	2173	2016	182577
测试集大小	693	672	4000
类别数	10	8	10

4.2 实验设置

在实验中，检索实例个数设置为 $R=1000$，参数 λ 控制两个模态之间的权重，一般设置 $\lambda=0.5$，平衡参数 α 和 β 分别被设置为 0.001 和 0.005，正则化参数 γ 设置为 0.01.模型的性能评价指标采用平均精度均值(Mean Average Precision，MAP)，为重复 10 次检索的结果平均值.同时，本文也给出了每个数据集上当哈希码位数分别为 32 和 64 时的两种性能曲线，一种是精度召回率曲线(Precision Recall curve，PR)，它显示了在不同召回率下的检索精度；另一种是 topN_precision 精度曲线，它反映了精度相对于检索实例个数的变化.所有的实验结果都是在 Intel(R)Core(TM)CPU i7-4790@3.60 GHz 16 GB RAM 的机器上运行得到的.

4.3 实验结果及分析

表 2 和表 3 给出了所有方法在 Wiki、LabelMe 以及 NUS_WIDE 三个数据集上的两个跨模态检索任务和哈希码从 16 到 128 位的 MAP 值.在表 3 中没有给出 SMFH_Tang 的实验 MAP 值，因为在该方法中需要建立一个混合图，这是一个时间复杂度极高的约束项，不适用于大数据集.在该论文中的大数据集上实验是从数据集中取一部分来分别做训练集和测试集.

表 2 在 Wiki 和 LabelMe 上的 MAP 值比较

任务	方法	Wiki				LabelMe			
		16	32	64	128	16	32	64	128
Image To Text	CMFH	0.2060	0.2215	0.2309	0.2352	0.4231	0.4418	0.4754	0.4763
	LSSH	0.2101	0.2145	0.2166	0.2092	0.6389	0.6810	0.6784	0.6919
	SMFH_Liu	0.1776	0.2208	0.2442	0.2441	0.3400	0.4399	0.5541	0.6340
	SMFH_Tang	0.2548	0.2679	0.2757	0.2797	0.6172	0.6620	0.6900	0.7030
	JCH	0.1829	0.1887	0.1942	0.1973	0.5126	0.5645	0.6106	0.6414
	DCH	0.2366	0.2780	0.3144	0.3213	0.6480	0.6952	0.7410	0.7560
	Ours	**0.2756**	**0.2906**	**0.3165**	**0.3228**	**0.6695**	**0.7651**	**0.8102**	**0.8293**
Text To Image	CMFH	0.5112	0.5331	0.5503	0.5565	0.5453	0.5778	0.6144	0.6217
	LSSH	0.5212	0.5421	0.5485	0.5525	0.6302	0.6503	0.6592	0.6678
	SMFH_Liu	0.4412	0.5547	0.6022	0.6023	0.4341	0.5675	0.6929	0.7783
	SMFH_Tang	0.6052	0.6288	0.6367	0.6434	0.7336	0.7772	0.7998	0.8077
	JCH	0.4742	0.5097	0.5250	0.5349	0.6482	0.6958	0.7340	0.7547
	DCH	0.5757	0.6771	0.6972	0.7058	0.8171	0.8612	0.8738	0.8782
	Ours	**0.6626**	**0.6904**	**0.7092**	**0.7150**	**0.8510**	**0.8861**	**0.8991**	**0.9055**

表 3 在 NUS_WIDE 上的 MAP 值比较

任务	方法	NUS_WIDE			
		16	32	64	128
Image To Text	CMFH	0.4803	0.5059	0.5105	0.5211
	LSSH	0.4847	0.4949	0.5086	0.5322
	SMFH_Liu	0.5307	0.5848	0.6183	0.6418
	JCH	0.4787	0.5046	0.5306	0.5471
	DCH	0.5706	0.5908	0.6051	0.6574
	Ours	**0.6632**	**0.6918**	**0.7005**	**0.7058**
Text To Image	CMFH	0.6310	0.6372	0.6912	0.7160
	LSSH	0.5897	0.6175	0.6623	0.6839
	SMFH_Liu	0.5553	0.6454	0.6725	0.6972
	JCH	0.5640	0.6407	0.6984	0.7193
	DCH	0.7291	0.7312	0.7480	0.7730
	Ours	**0.7974**	**0.8141**	**0.8152**	**0.8192**

从表 2 和表 3 的总体上来看，本文提出的方法在所有码长上都比其它方法取得的效果好，这也很好的说明了本文提出的方法在跨模态检索中的有效性.与位居第二的基于线性分类器的 DCH 方法相比，在图片搜文本的任务中，本文方法的最好结果在 Wiki，LabelMe 和 NUS_WIDE 数据集上分别比其大约高出了 4%，7%和 10%；而在文本搜图片的任务中，则分别大约高出了 9%，4%和 8%.这证明了本文对不同模态数据采用交叉耦合哈希表示的有效性，比将不同模态数据嵌入到共同的子空间学习统

一哈希码的方法具有更好的效果. 与没有基于线性分类器的方法 CMFH[15]、LSSH[16]、SMFCH_Liu[19]、SMFH_Tang[20] 和 JCH[17] 方法相比较，基于线性分类器的 DCH[21] 和本文的方法均获得了更优的性能，这证明了基于线性分类器所学习到的哈希码具有更强的语义判别能力. 另外，从表中也可看出，本文提出的方法的 MAP 值随着哈希码位数的增加而增加，这表明本文提出的方法能够利用比较长的哈希码来编码更多的判别信息，从而提高检索性能. 并且通过比较发现，在所有方法中，文本作为查询来检索图片时的性能比图片作为查询来检索文本的效果更好，这表明图片很难检索出语义相似性的文本，因为文本所含的语义信息比图像要多.

图 2、图 3 和图 4 分别给出了所有比较的方法在 Wiki、LabelMe 和 NUS_WIDE 三个数据集上哈希码位数为 32 和 64 时两个跨模态检索任务的 PR

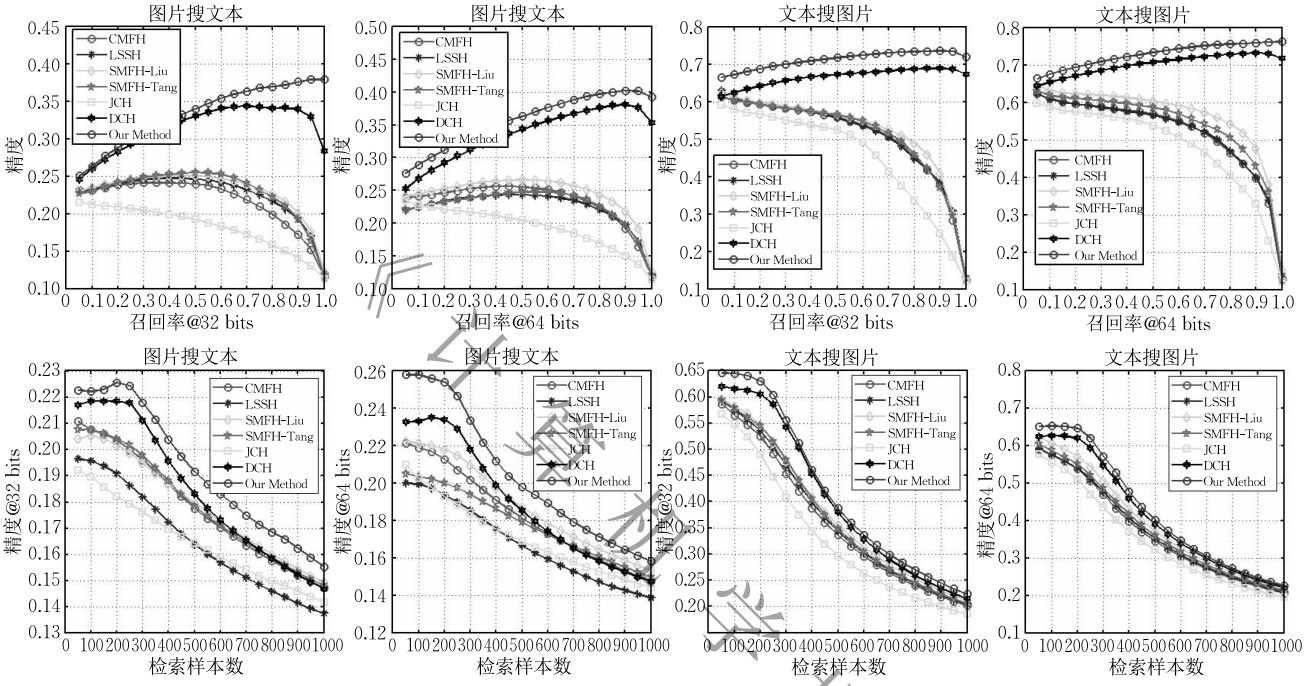


图 2　在 Wiki 上哈希码位数为 32、64 时的 PR 和 topN_precision 曲线

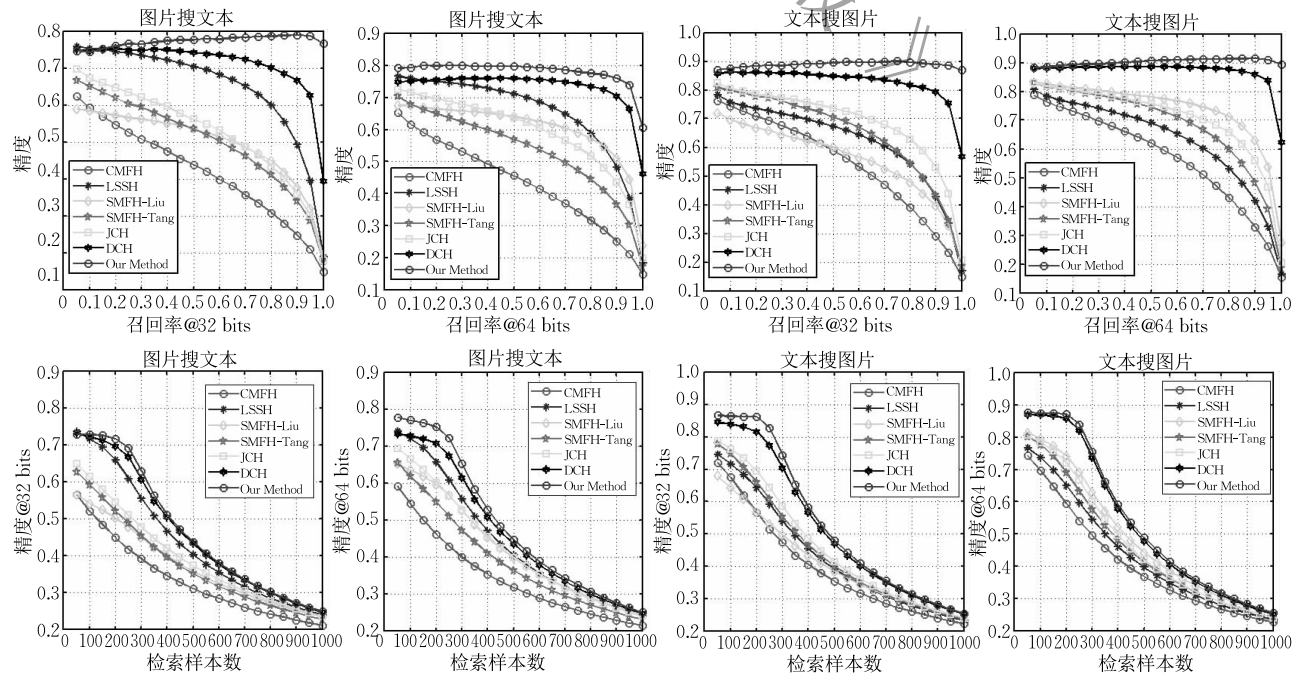


图 3　在 LabelMe 上哈希码位数为 32、64 时的 PR 和 topN_precision 曲线

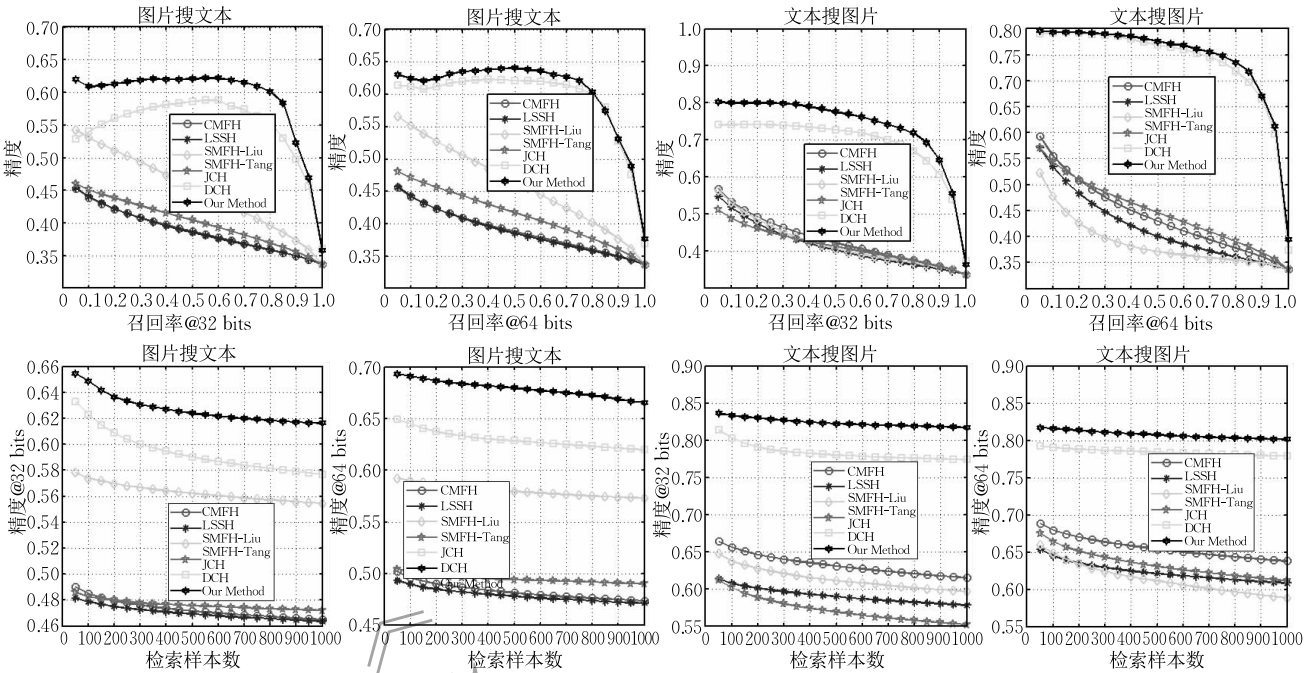


图 4 在 NUS_WIDE 上哈希码位数为 32、64 时的 PR 和 topN_precision 曲线

和 topN_precision 曲线. 从各图的 topN_precision 曲线可以看出,本文提出的方法在检索实例数目不断变化的情况下均比其它方法取得了更高的精度. 并且从各图的 PR 曲线可以看出,本文的方法在不同召回率下也均获得了较其它方法更高的精度. 这也证明了本文结合线性分类器和不同模态间交叉耦合哈希表示进行跨模态检索的有效性.

MAP 值和 topN_precision 都是依据于汉明排序产生的,通过分析可以看出本文提出的方法在汉明排序上同样具有较好的优势.

4.4 算法训练时间对比分析

本文对算法复杂度做了进一步的分析和比较,见表 4 所示,给出了除 SMFH_Tang 之外的其它比较方法在 Wiki、LabelMe 和 NUS_WIDE 三个数据集上哈希码位数为 32 和 64 的训练时间. 由于 SMFH_Tang 不适用于大数据集,所以在表中没有给出在 NUS_WIDE 上的训练时间.

表 4 三个数据集上不同算法训练时间的比较(单位:s)

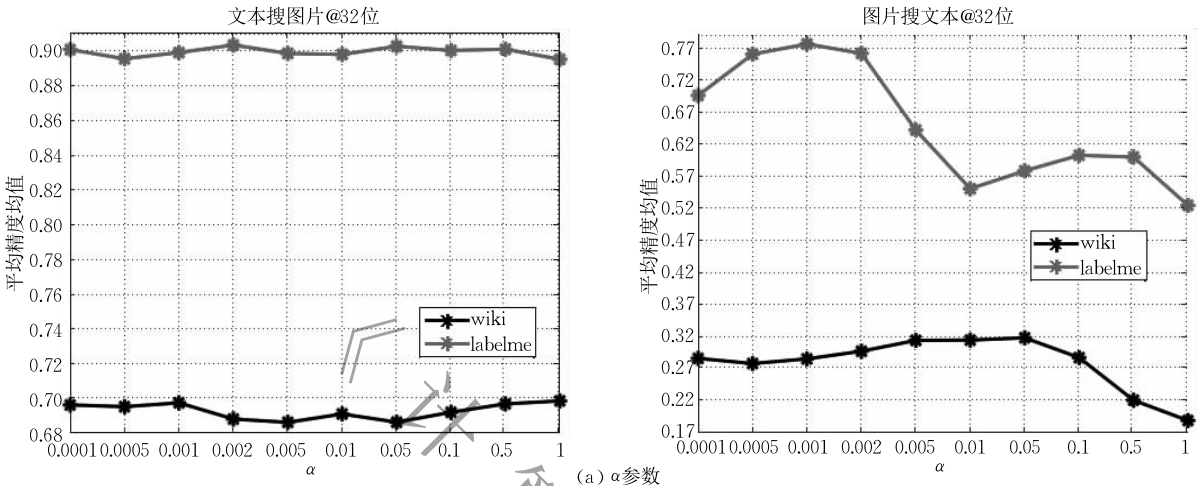
方法	Wiki		LabelMe		NUS_WIDE	
	32	64	32	64	32	64
CMFH	0.35	0.43	5.12	4.62	571.29	592.35
LSSH	121.33	137.28	157.72	178.46	593.13	659.67
SMFH_Liu	1.19	2.05	3.09	3.77	73.94	88.04
SMFH_Tang	148.50	167.91	547.23	620.83	—	—
JCH	0.53	1.07	4.06	5.93	192.24	269.42
DCH	1.23	1.95	4.58	7.10	246.50	394.38
Ours	1.74	2.77	8.31	10.00	384.12	497.41

从表 4 可以看出,在小数据集 Wiki 和 LabelMe 上,无监督方法 JCH 和 CMFH 训练时间最短,而有监督的方法 SMFH_Liu、DCH 和本文的方法训练时间略长些,但都在 10 s 以下. 在大数据集 NUS_WIDE 上, SMFH_Liu 获得最好的效率,这是由于该方法在训练过程使用了采样的方法. LSSH 由于需要进行稀疏编码和矩阵分解,这两种运算复杂度都比较高,因此在各数据集上的训练时间比较长. 而 SMFH_Tang 方法需要构建混合图,这是一个时间复杂度很高的项,即使在小数据集 Wiki 和 LabelMe 上的训练时间都较长. 由于 CMFH 采用了矩阵分解的方法,因此在大数据 NUS_WIDE 上的训练时间也很长. JCH、DCH 和本文的方法在 NUS_WIDE 上的训练时间都比在小数据集上的训练时间更长. DCH 和本文方法由于采用了线性分类器的表示,线性分类器的超平面权值系数的训练将增加了它们的运算复杂度. 从以上结果可以看出,本文的方法在训练时间上与基于线性分类器的方法 DCH 相差不大,但在性能上获得更好的结果. 综上所述,相比较其他方法,本文的方法在时间复杂度和性能上均获得了较好的结果.

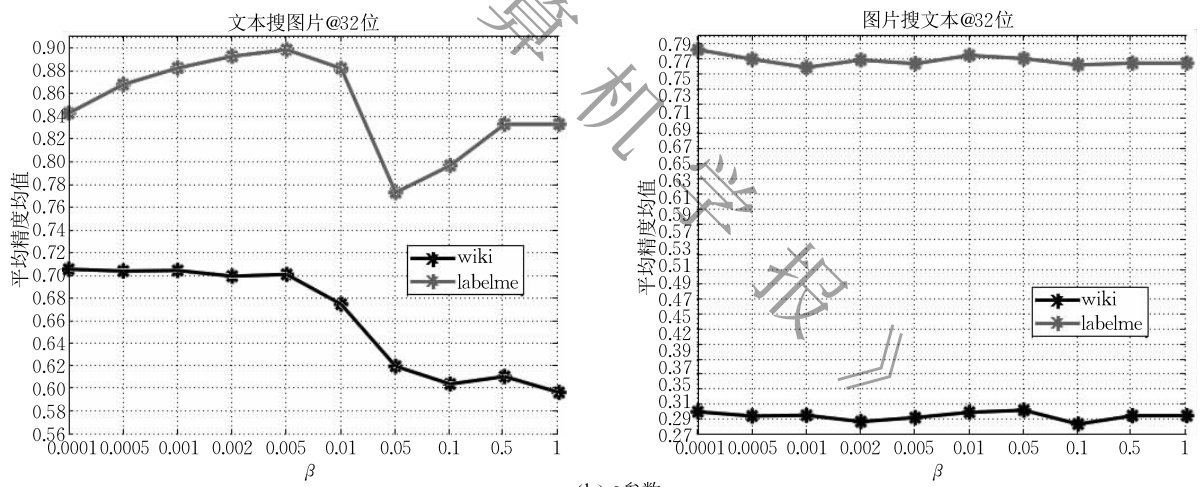
4.5 参数分析

下面进一步分析了参数对实验结果的影响和敏感性,主要分析了参数 α、β 和 γ 的影响. 参数 λ 一般设置为 0.5,使两种模态权重相当,不做进一步分析. 本文分别对 α、β 和 γ 在{0.0001,0.0005,0.001,0.005,0.01,0.05,0.1,0.5,1.0}的取值范围上进行了实验,固定其它参数. 其它参数采用 3.2 节实验设置中的取值,哈希码的长度设置为 32,MAP 作为参数分析的评价标准.

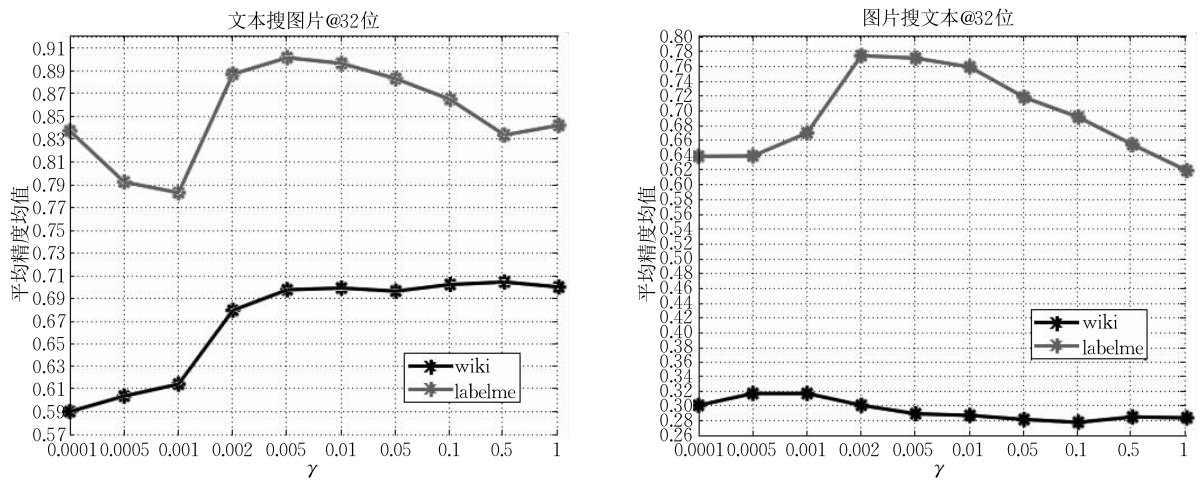
图 5 显示了各参数在两个数据集上的实验结果. 从图 5 中可以看出平衡参数 α,β 分别在[0.0005, 0.003]和[0.002,0.01]范围内,获得了最好效果;正则化参数 γ 控制着整个模型的复杂度,当它取值太小时,导致模型容易过拟合;而当取值较大时,容易引起欠拟合. 从其曲线变化图上我们可以看出,在[0.007,0.03]的取值范围内,其对应的 MAP 值最大. 由于文本和图像所含有的语义信息不一样,参数 α,β 是用来对图像和文本所含的比重进行一个匹配. 通过图 5 我们可以看到参数 α 在用图片搜文本时的波动比用文本搜图片时的波动要大. 通过以上分析我们能够发现本文方法中的参数在一个比较宽的范围内,它们是不敏感的,取得了较好的结果,在其它范围内,也能够获得相当满意的结果.



(a) α参数



(b) β参数



(c) γ参数

图 5　参数分析

5 总结与展望

本文提出了一种有效的基于交叉相似性最大化的判别式跨模态哈希特征表示学习算法.算法考虑到语义判别性在跨模态检索的特征表示上的重要性,将线性分类器引入模型的目标优化过程,使得最终学习到的哈希码兼顾了跨模态语义相关和类别判别力.并在三大公开的数据集上与最近无监督和有监督的跨模态哈希算法进行了实验比较.实验结果表明,本文的算法在算法效率和检索性能方面均具有较好的优越性.

下一步工作我们将考虑在模型中引入非线性特征映射关系,使得学习到的哈希码不局限于简单的线性投影,进一步提升哈希码的鉴别力和紧凑性.另一方面,各模态的特征表示将考虑引入深度表示,使各模态的特征更丰富和具有更好的鲁棒性,并将进一步考虑模态间特征表示的语义层级相关性.

参 考 文 献

[1] Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3594-3601

[2] Masci J, Bronstein M M, Bronstein A M, et al. Multimodal similarity-preserving hashing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(4): 824-830

[3] Pereira J C, Coviello E, Doyle G, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 521-535

[4] Zhang Hong, Wu Fei, Zhuang Yue-Ting. Research on cross media correlation reasoning and retrieval. Journal of Computer Research and Development, 2008, 45(5): 869-876(in Chinese)
(张鸿, 吴飞, 庄越挺. 跨媒体相关性推理与检索研究. 计算机研究与发展, 2008, 45(5): 869-876)

[5] Wang Kai-Ye, Yin Qi-Yue, Wang Wei, et al. A comprehensive survey on cross-modal retrieval. CoRRabs/1607.06215, 2016

[6] Cao Yue, Long Ming-Sheng, Wang Jian-Min, et al. Deep visual-semantic hashing for cross-modal retrieval//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1445-1454

[7] Peng Yu-Xin, Zhu Wen-Wu, Zhao Yao, et al. Cross-media analysis and reasoning: advances and directions. Frontiers of IT&EE, 2017, 18(1): 44-57

[8] Xu Yan, Shen Fu-Min, Xu Xing, et al. Large-scale image retrieval with supervised sparse hashing. Neurocomputing, 2017, 229: 45-53

[9] Kumar S, Udupa R. Learning hash functions for cross-view similarity search//Proceedings of the International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1360-1365

[10] Song Jing-Kuan, Yang Yang, Yang Yi, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, USA, 2013: 785-796

[11] Zhen Yi, Yeung Dit-Yan. Co-regularized hashing for multimodal data//Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2012: 1753-1760

[12] Zhang Hong, Wu Fei, Zhuang Yue-Ting, Chen Jian-Xun. Cross-media retrieval method based on content correlations. Chinese Journal of Computers, 2008, 31(5): 820-826(in Chinese)
(张鸿, 吴飞, 庄越挺, 陈建勋. 一种基于内容相关性的跨媒体检索方法. 计算机学报, 2008, 31(5): 820-826)

[13] Li Zhi-Xin, Shi Zhi-Ping, Chen Hong-Chao, et al. Multimodal image retrieval based on semantic learning. Computer Engineering, 2013, 39(3): 258-263(in Chinese)
(李志欣, 施智平, 陈宏朝等. 基于语义学习的图像多模态检索. 计算机工程, 2013, 39(3): 258-263)

[14] Shen Xiao-Bo, Shen Fu-Min, Sun Quan-Sen, et al. Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval. IEEE Transactions on Cybernetics, 2017, 47(12): 4275-4288

[15] Ding Gui-Guang, Guo Yu-Chen, Zhou Ji-Le. Collective matrix factorization hashing for multi-modal data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2083-2090

[16] Zhou Ji-Le, Ding Gui-Guang, Guo Yu-Chen. Latent semantic sparse hashing for cross-modal similarity search//Proceedings of the International ACM SIGIR Conference. Gold Coast, Australia, 2014: 415-424

[17] Liu Yi-Han, Chen Zhao-Jia, Deng Cheng, et al. Joint coupled-hashing representation for cross-modal retrieval//Proceedings of the International Conference on Internet Multimedia Computing and Service. Xi'an, China, 2016: 35-38

[18] Zhang D, Li W-J. Large-scale supervised multimodal hashing with semantic correlation maximization//Proceedings of the 28th Association for the Advancement of Artificial Intelligence. Québec City, Canada, 2014: 2177-2183

[19] Liu Hong, Ji Rong-Rong, Wu Yong-Jian, et al. Supervised matrix factorization for cross-modality hashing//Proceedings of the Association for the Advancement of Artificial Intelligence. Phoenix, USA, 2016: 1767-1773

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

[21] Xu Xing, Shen Fu-Min, Yang Yang, et al. Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Transactions on Image Processing, 2017, 26(5): 2494-2507

[22] Shen Fu-Min, Shen Chun-Hua, Liu Wei, et al. Supervised discrete hashing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 37-45

[23] Datar M-Y, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions//Proceedings of the Twentieth Annual Symposium on Computational Geometry. Brooklyn, USA, 2004: 253-262

[24] Kulis B, Grauman K. Kernelized locality-sensitive hashing for scalable image search//Proceedings of the International Conference on Computer Vision. Kyoto, Japan, 2009: 2130-2137

[25] Weiss Y, Torralba A, Fergus R. Spectral hashing//Proceedings of the 22nd Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1753-1760

[26] Gong Yun-Chao, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916-2929

[27] Strecha C, Bronstein A M, Bronstein M M, et al. LDAHash: Improved matching with smaller descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(1): 66-78

[28] Zhao Kang, Lu Hong-Tao, He Yang-Cheng, et al. Locality preserving discriminative hashing//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 1089-1092

[29] Nguyen V A, Lu Ji-Wen, Do M N. Supervised discriminative hashing for compact binary codes//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 989-992

[30] Zhang Dell, Wang Jun, Cai Deng, et al. Self-taught hashing for fast similarity search//Proceedings of the 33rd Annual International ACM SIGIR. Geneva, Switzerland, 2010: 18-25

[31] Zhu Xiao-Feng, Huang Zi, Shen Heng-Tao, et al. Linear cross-modal hashing for efficient multimedia search//Proceedings of the ACM International Conference on Multimedia. Barcelona, Spain, 2013: 143-152

[32] Lin Zi-Jia, Ding Gui-Guang, Hu Ming-Qing, et al. Semantics-preserving hashing for cross-view retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3864-3872

YAN Shuang-Yong, born in 1990, M. S. candidate. His research interests include information retrieval and computer vision.

LIU Chang-Hong, born in 1977, Ph. D., associate professor. Her research interests include computer vision, machine learning and hyperspectral image processing.

JIANG Ai-Wen, born in 1984, Ph. D., associate professor. His research interests include pattern recognition, image analysis and retrieval, and machine learning.

YE Ji-Hua, born in 1966, M. S., professor. His research interests include data fusion, pattern recognition and Internet of Things technology.

WANG Ming-Wen, born in 1965, Ph. D., professor. His research interests include natural language processing and information retrieval.

Background

With the rapid growth of the multimedia data in the network, the cross-modal retrieval has been one of the hot research topics and has been attached great importance in large-scale information retrieval field. Cross-modal retrieval based on hashing has attracted much attention for its low storage cost and fast query speed. The key problem of cross-modal hashing is how to use the data from different modes to learn the hash codes effectively. At present, the cross-modal hash learning methods can be roughly divided into the unsupervised methods and the supervised methods. In the unsupervised methods, the unified hash codes based approaches have been recently attracted much attention for saving the storage space of multi-modal data more effectively, which enforce different modal data to be projected into a shared subspace for a consistent representation. However, the difference between the characteristics of different modal data is very large and the obtained representation cannot capture the characteristics of different modal data in original space. The supervised methods combine the label information in the hashing learning. However, Most of methods in embedding multi-modal data into the common semantic space neglect the semantic discriminative feature representation.

This paper proposes a discriminative cross-modal hashing learning with coupled semantic correlation algorithm. In this algorithm, linear discriminative classifier is used to the supervised hashing learning and at the same time the correlations between different modalities are maximized in the embedding spaces by joint coupled-hashing representation. This approach can learn respectively the discriminative binary hash code for each modal but also solve the defects of projecting a variety of data into a common embedding semantic space and capture the semantic relevance between multi-modal data. The experimental results show that the proposed approach obtains great improvement on the retrieval accuracy and the computational efficiency by comparing it with six current relevant algorithms on three benchmark datasets.

This work was supported by the National Natural Science Foundation of China (Nos. 61662030, 61365002, 61462042, 61462045), the Jiangxi Provincial Natural Science Foundation (No. 20171BAB202016) and the Science Research Fund of Jiangxi Educational Department (No. GJJ150350).