JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
07
VOL.26
ISSN1006-8961
CN11-3758/TB

中国多媒体大会(ChinaMM2020)专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第7期（总第303期）

2021年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

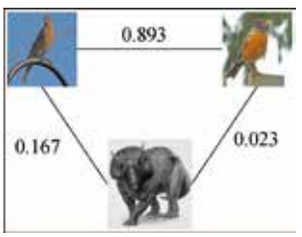
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年7月 第26卷 第7期（总第303期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



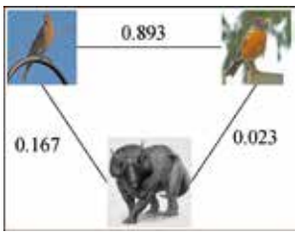
类别语义相似性监督的小样本图像识别(第1594页)

像素聚合和特征增强的任意形状场景文本检测(第1614页)

时序特征融合的视频实例分割(第1692页)

Volume 26, Number 7
Published July 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Few shot image recognition based on class semantic similarity supervision(P1594)

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement (P1614)

Video instance segmentation based on temporal feature fusion (P1692)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2021)07-1558-10

论文引用格式: Min K L, Zhang G B, Wang L and Li D P. 2021. Structure-preserving hashing with coupled projections for cross-modal retrieval. Journal of Image and Graphics,26(07):1558-1567(闵康凌,张国宾,王磊,李丹萍. 2021. 耦合保持投影哈希跨模态检索. 中国图象图形学报,26(07):1558-1567)[DOI:10.11834/jig.200526]

耦合保持投影哈希跨模态检索

闵康凌[1,2],张国宾[3],王磊[1,2],李丹萍[4*]

1. 西安电子科技大学电子工程学院, 西安　710071; 2. 上海交通大学海洋智能装备与系统教育部重点实验室, 上海　200240;
3. 中国电子科技集团公司第27研究所, 郑州　450047; 4. 西安电子科技大学通信工程学院, 西安　710071

摘　要: **目的**　基于哈希的跨模态检索方法因其检索速度快、消耗存储空间小等优势受到了广泛关注。但是由于这类算法大都将不同模态数据直接映射至共同的汉明空间,因此难以克服不同模态数据的特征表示及特征维度的较大差异性,也很难在汉明空间中同时保持原有数据的结构信息。针对上述问题,本文提出了耦合保持投影哈希跨模态检索算法。**方法**　为了解决跨模态数据间的异构性,先将不同模态的数据投影至各自子空间来减少模态"鸿沟",并在子空间学习中引入图模型来保持数据间的结构一致性;为了构建不同模态之间的语义关联,再将子空间特征映射至汉明空间以得到一致的哈希码;最后引入类标约束来提升哈希码的判别性。**结果**　实验在3个数据集上与主流的方法进行了比较,在Wikipedia数据集中,相比于性能第2的算法,在任务图像检索文本(I to T)和任务文本检索图像(T to I)上的平均检索精度(mean average precision,mAP)值分别提升了6%和3%左右;在MIR-Flickr数据集中,相比于性能第2的算法,优势分别为2%和5%左右;在Pascal Sentence数据集中,优势分别为10%和7%左右。**结论**　本文方法可适用于两个模态数据之间的相互检索任务,由于引入了耦合投影和图模型模块,有效提升了跨模态检索的精度。

关键词:跨模态检索;哈希;结构保持;图模型;耦合投影;子空间学习

Structure-preserving hashing with coupled projections for cross-modal retrieval

Min Kangling[1,2], Zhang Guobin[3], Wang Lei[1,2], Li Danping[4*]

1. School of Electronic Engineering, Xidian University, Xi'an　710071, China;　2. Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Jiao Tong University, Shanghai　200240, China;　3. The 27th Research Institute, China Electronics Technology Group Corporation, Zhengzhou　450047, China;
4. School of Telecommunications Engineering, Xidian University, Xi'an　710071, China

Abstract: **Objective** With the rapid development of multimedia technology, the scale of multimedia data has been growing rapidly. For example, people are used to describing the things they want to show with multimedia data such as texts, images, and videos. Obtaining the relevant results of one modality using another modality is a good objective. In this sense, how to effectively perform semantic correlation analysis and measure the similarity between the data has gradually become a

收稿日期:2020-08-28;**修回日期**:2021-02-02;**预印本日期**:2021-02-09

*通信作者:李丹萍　dpli@xidian.edu.cn

基金项目:国家重点研发计划项目(2016YFE0207000);国家自然科学基金项目(61203137, 61401328);陕西省自然科学基础研究计划项目(2014JQ8306, 2015JM6279)

Supported by: National Key Research and Development Program of China (2016YFE0207000); National Natural Science Foundation of China (61203137, 61401328); Natural Science Basic Research Program of Shaanxi (2014JQ8306, 2015JM6279)

hot research topic. As the representation of different modal data is heterogeneous, it poses a great challenge to the cross-modal retrieval task. Hashing-based methods have received great attention in cross-modal retrieval because of its fast retrieval speed and low storage consumption. To solve the problem of heterogeneity between different modalities of the data, most of the current supervised hashing algorithms directly map different modal data into the Hamming space. However, these methods have the following limitations: 1) The data from each modality have different feature representations, and the dimensions of their feature spaces vary greatly. Therefore, it is difficult for these methods to obtain a consistent hash code by directly mapping the data from different modalities into the same Hamming space. 2) Although label information has been considered for these hashing methods, the structural information of the original data is ignored, which could result in a less-representative hash code to encode the original structural information in each modality. To solve these issues, a novel hashing algorithm called structure-preserving hashing with coupled projections (SPHCP) is proposed in this paper for cross-modal retrieval. **Method** Considering the heterogeneity between the cross-modal data, this algorithm first projects the data from different modalities into their respective subspaces to reduce the modal difference. A local graph model is also designed in the subspace learning to maintain the structural consistency between the samples. Then, to build a semantic relationship between different modalities, the algorithm maps the subspace features to the Hamming space to obtain a consistent hash code. At the same time, the label constraint is exploited to improve the discriminant power of the obtained hash codes. Finally, the algorithm measures the similarity of different modal data in terms of the Hamming distance. **Result** We compared our model with several state-of-the-art methods on three public datasets, namely, Wikipedia, MIRFlickr, and Pascal Sentence. The mean average precision (mAP) is used as the quantitative evaluation metric. We first test our method on two benchmark datasets, Wikipedia and MIRFlickr. To evaluate the impact of hash-code length on the performance of the algorithm, this experiment set the hash code length to 16, 32, 64, and 128 bits. The experimental results show that for both the text-retrieving image task and image-retrieving text task, our proposed method outperforms the existing methods in each length setting. To further measure the performance of our proposed method on the dataset with deep features, we test the algorithm on the Pascal Sentence dataset. The experimental results show that our SPHCP algorithm can also achieve higher mAP on such dataset with deep features. In general, cross-modal retrieval methods based on deep networks can handle nonlinear features well, so their retrieval accuracy is supposed to be higher than that of traditional methods, but they need much more computational power. As a "shallow" method, the proposed SPHCP algorithm is competitive with deep methods in terms of mAP. Therefore, as an interesting direction, our framework can be used in conjunction with the deep learning method in the future, i. e. , using deep learning to extract the features of images and text offline, and using the SPHCP algorithm for fast retrieval. Furthermore, we analyze the parameter sensitivity of the proposed algorithm. As this algorithm has 7 parameters, a controlled variable method is used for evaluation. The experimental results show that the proposed algorithm is not sensitive to parameters, which means that the training process does not require much optimization time, making it suitable for practical application. **Conclusion** In this study, a novel method called SPHCP is proposed to solve the problems mentioned. First, aiming at the "modal gap" between cross-modal data, the scheme of coupled projections is applied to gradually reduce the modal difference of multimedia data. In this way, a more consistent hash code can be obtained. Second, considering the structural information and semantic discrimination of the original data, the algorithm introduces the graph model in subspace learning, which can maintain the intra-class and inter-class relationship of the samples. Finally, a label constraint is introduced to improve the discriminability of the hash code. The experiments on the benchmark datasets verify the effectiveness of the proposed algorithm. Specifically, compared with the second-best method, SPHCP achieves an improvement by 6% and 3% on Wikipedia for two retrieval tasks. On MIRFlickr, SPHCP achieves an improvement by 2% and 5%. On Pascal Sentence, the improvement is approximately 10% and 7%. However, the proposed method requires a large amount of computing power when dealing with large-scale data, because SPHCP introduces a graph model to maintain the structural information between the data. The calculation of the structural information between each sample leads to a larger computing complexity. In future research, we will introduce nonlinear feature mapping into our SPHCP framework to improve its scalability when dealing with nonlinear feature data. Furthermore, we can extend the SPHCP from a cross-modal retrieval algorithm to a multi-modal version.

Key words: cross-modal retrieval; hashing; structure-preserving; graph model; coupled projections; subspace learning

0 引 言

大数据时代,不仅数据量呈爆炸式增长,而且数据的模态种类繁多,多模态数据已经广泛存在于日常生活中。因此,对跨模态检索方法的研究在检索领域变得愈发重要。但是,不同模态数据采用不同的特征进行表示,这导致了在跨模态检索中很难将不同模态的数据直接进行相似性衡量,如何处理数据间的异构特征是跨模态检索研究中无法回避的问题。

为了克服数据间的异构性,学者们采用子空间学习的方法,即通过不同模态的数据来学习一个共同的语义子空间,从而消除数据间的异构性,然后再进行相似性衡量就可以完成检索任务。Sharma 等人(2012)和 Wu 等人(2017)的研究表明子空间学习方法有较高的检索精度。但这种方法仅适用于规模较小和特征维数较小的数据,当处理大规模数据时,一方面很难学习一个共同的语义空间,另一方面,随着数据规模的增大,算法计算复杂度也大大增加。因此,传统子空间方法很难满足现实需求。

由于用哈希码对数据进行特征表示检索速度快,所需存储空间小,哈希跨模态检索方法受到了广泛的关注,其主要思路是将不同模态的数据映射至一个共同的汉明空间,然后用汉明距离进行相似性衡量。一般地,根据是否利用标签信息,哈希跨模态检索方法可以分为无监督和有监督两类。无监督方法试图直接从数据的底层结构和数据分布来学习哈希码,希望生成具有原始数据结构信息的哈希码。实际上,给数据添加标签信息非常耗时耗力,因此无监督哈希跨模态检索有着重要的现实意义。但是无监督方法生成的哈希码判别性有限,与之相对,有监督哈希方法的思路则是在哈希码的学习过程中结合标签信息来提高所得哈希码的判别性。

大多数有监督哈希跨模态检索方法在原理上有几点共性:首先,将不同模态的数据直接映射至汉明空间或者将不同模态数据先投影至共同的子空间,再映射至汉明空间;其次,为了增强哈希码的判别性,通常会将类标信息引入到哈希码的学习中。但是,通过理论分析和实际数据验证,仍能看出这些方法的一些缺陷:1)对于不同模态的数据,无论是表示特征还是特征维数都有很大的差异,将这些模态不同的数据直接映射至汉明空间,很难获得一致的哈希码,进而影响检索精度;2)尽管引入类标信息可以提升哈希码的判别性,但是这些方法大都没有考虑数据间的结构信息,另一方面,在实际应用中,标签信息缺失和错误是比较常见的问题,仅仅依赖标签信息会降低算法的鲁棒性,甚至可能会破坏结构信息。

为了解决上述问题,本文提出耦合保持投影哈希(structure-preserving hashing with coupled projections,SPHCP)跨模态检索算法,一方面,为了减少模态"鸿沟",算法将不同模态的数据投影至各自的子空间,并在各自子空间学习时,引入图模型来保持数据间的结构信息。另一方面,算法再将各自子空间特征映射至汉明空间以得到一致的哈希码,并引入类标约束使所学到的哈希码具有更好的判别性。图 1 为 SPHCP 算法原理图,其模型构建思路将在第 2 节中详细阐述。

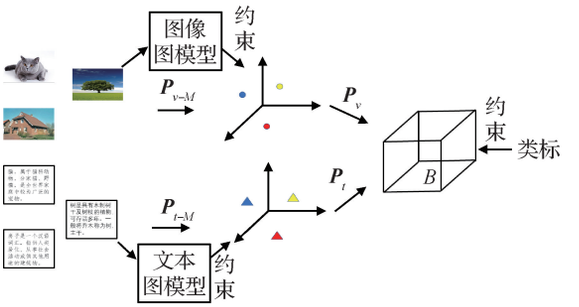


图 1 耦合保持投影哈希跨模态检索算法原理图

Fig. 1 Framework of the proposed SPHCP method

1 相关工作

基于哈希的跨模态检索可以大致分成有监督和无监督两类。无监督方法主要从数据分布上学习哈希码。Kumar 和 Udupa(2011)提出了 CVH(learning hash functions for cross-view similarity search)算法,CVH 将传统的单模态检索扩展到多模态检索,将同一目标映射到相似的二进制代码,从而提高了检索速度。Ding 等人(2014)提出了协同矩阵分解哈希算法(collective matrix factorization hashing for multimodal data,CMFH),该算法通过矩阵分解的方式得到共同的潜在语义空间,然后将潜在的语义映射至汉明空间以求得一致的哈希码。Zhou 等人(2014)提出 LSSH(latent semantic sparse hashing for cross-modal similarity search)算法,其主要思想是使用稀

疏编码来获取图像特征,用矩阵分解来学习文本中的潜在语义,然后映射至汉明空间。Hu 等人(2019)提出 CRE(collective reconstructive embeddings for cross-modal hashing)算法,使用余弦相似度对文本数据进行建模,而利用欧几里德距离对图像数据进行建模,以期同时解决多模态数据的异质性和集成复杂性。

本文研究重点为有监督的哈希跨模态检索,这类方法则是将标签信息考虑在内,以期增强哈希码的判别性。Zhang 和 Li(2014)提出 SCM(large-scale supervised multimodal hashing with semantic correlation maximization)算法,该算法将语义类标引入哈希学习中,可以较好地保持模态间的语义相似性。Lin 等人(2015)提出 SePH(semantics-preserving hashing for cross-view retrieval)算法,其主要思想是通过最小化原始空间和汉明空间的散度距离以期保持哈希码的语义相似性。Xu 等人(2017)提出了离散跨模态哈希(learning discriminative binary codes for large-scale cross-modal retrieval,DCH)算法,不仅在哈希码学习过程中引入了标签信息,还考虑了离散约束,因而获得了较高的检索准确率。严双咏等人(2019)针对将不同模态数据投影到一个共同空间的问题,提出了 DHCSC(discriminative cross-modal hashing with coupled semantic correlation)算法,一方面利用耦合哈希表示技术缓解了模态的异构性,另一方面利用标签信息挖掘了不同模态数据间的语义相关性,并可以直接学习二进制哈希码。Zhang 等人(2020)为了生成更具判别性的哈希码,提出了 DSPH(learning latent hash codes with discriminative structure preserving for cross-modal retrieval)方法,该方法在学习共同子空间时考虑了类内类间关系,并通过矩阵分解的思想得到一个共同的哈希码。

值得一提的是,随着深度学习的发展,研究人员已经注意到深度学习在对数据建模和表示方面的强大能力。深度学习可以在跨模态数据之间建立更为复杂的非线性关系,从而可以进一步弥合不同模态数据间的异质性差距。在哈希跨模态检索领域中的代表工作有由 Jiang 和 Li(2017)提出的深度跨模态哈希(deep cross-modal hashing,DCMH),由 Jin 等人(2019)提出的深度语义保留序数哈希(deep semantic-preserving ordinal hashing for cross-modal similarity search,DSPOH)和由 Li 等人(2018)提出的自监督对抗哈希(self-supervised adversarial hashing, SSAH)等。这些方法在公开数据集上的平均检索精度都很高,但值得注意的是,深度哈希跨模态检索方法不仅模型复杂,而且计算资源消耗很大。深度网络的复杂结构与跨模态检索的实时性需求形成了矛盾,因而限制了其使用发展。

2　耦合保持投影哈希跨模态检索算法

2.1　问题描述

假设有 n 个训练样本,定义为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,每个样本点 $\boldsymbol{x}_i$ 由一对有着相同语义的图像和文本组成,即可以表示成 $\boldsymbol{x}_i=(\boldsymbol{v}_i,\boldsymbol{t}_i)$,其中 $\boldsymbol{v}_i\in\mathbf{R}^{1\times m_v}$ 表示图像的特征矢量,$\boldsymbol{t}_i\in\mathbf{R}^{1\times m_t}$ 表示文本的特征矢量,m_v 和 m_t 分别为图像特征维数和文本特征维数。定义样本类标为 $\boldsymbol{y}_i\in\mathbf{R}^{1\times c}$,其中 c 为样本类别数,当样本 $\boldsymbol{x}_i$ 属于 k 类时,那么 $\boldsymbol{y}_i$ 中对应列 $\boldsymbol{y}_{ik}=1$,其余列置 0。此外,假设输入图像样本为 $\boldsymbol{V}$,那么 $\boldsymbol{V}=\{\boldsymbol{v}_i\}_{i=1}^{n}\in\mathbf{R}^{n\times m_v}$,同样,假设输入文本样本为 $\boldsymbol{T}$,则 $\boldsymbol{T}=\{\boldsymbol{t}_i\}_{i=1}^{n}\in\mathbf{R}^{n\times m_t}$。

2.2　模型构建

不同模态的数据通常可用不同的特征空间进行表示。由于这些空间里的数据往往具有不同的物理意义,特征空间维度也具有很大差异,哈希跨模态检索方法通常直接将原始数据或者原始特征映射至汉明空间,这忽视了不同模态之间的“模态鸿沟”,会导致检索精度下降。而本文所提 SPHCP 算法首先将不同模态的数据通过两个合适的投影矩阵 $\boldsymbol{P}_{v-M}$ 和 $\boldsymbol{P}_{t-M}$ 分别投影至各自的子空间 $\boldsymbol{M}_v$ 和 $\boldsymbol{M}_t$,这可以缩小不同模态数据间的差异,其过程表示为

$$\boldsymbol{P}_{v-m}^{\mathrm{T}}\boldsymbol{V}\rightarrow\boldsymbol{M}_v,\quad \boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{T}\rightarrow\boldsymbol{M}_t$$

式中,$\boldsymbol{M}_v\in\mathbf{R}^{n\times d_v}$,$\boldsymbol{M}_t\in\mathbf{R}^{n\times d_t}$,$d_v$ 和 d_t 表示图像嵌入空间维数和文本嵌入空间维数。

为了在各自子空间中保持数据原有的结构信息,SPHCP 在文本和图像子空间中引入了一个图模型。首先定义权值矩阵 $\boldsymbol{W}$ 为

$$\boldsymbol{W}_{ij}=\begin{cases}1/n_c & \boldsymbol{y}_i=\boldsymbol{y}_j\in c\\ 0 & \text{其他}\end{cases}$$

式中,n_c 为 c 类的样本数量,$\boldsymbol{y}_i,\boldsymbol{y}_j$ 为图像和文本的标签。为了保持文本和图像子空间中数据的结构信

息,该算法引入了如下模型

$$\min \frac{1}{2}\sum_{i,j} \boldsymbol{W}_{ij}(\boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{t}_i - \boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{t}_j)^2 = \mathrm{tr}(\boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{L}_t\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_{t-m})$$

$$\min \frac{1}{2}\sum_{i,j} \boldsymbol{W}_{ij}(\boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{v}_i - \boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{v}_j)^2 = \mathrm{tr}(\boldsymbol{P}_{t-m}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{L}_v\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_{t-m})$$

因为跨模态数据之间是有高层语义关联的,因此可将不同模态的数据用同一个汉明空间来表示,通过投影矩阵 $\boldsymbol{P}_v$ 和 $\boldsymbol{P}_t$ 将子空间中的数据再映射至汉明空间 $\boldsymbol{B}$,这一过程可以表示成

$$\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v \rightarrow \boldsymbol{B},\quad \boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t \rightarrow \boldsymbol{B}$$

最后,为了使所学到的哈希码较之以往监督方法更具判别性,借鉴回归思想,希望通过哈希码学到的类别与原始数据的类别趋于一致,因此引入类标约束动态地学习分类器 $\boldsymbol{W}$,即

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{b}_i \rightarrow \boldsymbol{y}_i$$

综合以上过程,所提算法的完整目标函数为

$$\begin{aligned}\min_{\substack{\boldsymbol{W},\boldsymbol{B},\boldsymbol{M}_t,\boldsymbol{M}_v,\\ \boldsymbol{P}_{t-M},\boldsymbol{P}_{v-M},\boldsymbol{P}_t,\boldsymbol{P}_v}} & \|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\|_{\mathrm{F}}^2 + \mu_T\|\boldsymbol{M}_t - \boldsymbol{P}_{t-M}^{\mathrm{T}}\boldsymbol{T}\|_{\mathrm{F}}^2 + \\ & \mu_V\|\boldsymbol{M}_v - \boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V}\|_{\mathrm{F}}^2 + \lambda_T\|\boldsymbol{B} - \boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t\|_{\mathrm{F}}^2 + \\ & \lambda_V\|\boldsymbol{B} - \boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v\|_{\mathrm{F}}^2 + \gamma_T\mathrm{tr}(\boldsymbol{P}_{t-M}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{L}_T\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_{t-M}) + \\ & \gamma_V\mathrm{tr}(\boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{L}_V\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_{v-M}) + \\ & \alpha_1(\|\boldsymbol{W}\|_{\mathrm{F}}^2 + \|\boldsymbol{P}_{v-M}\|_{\mathrm{F}}^2 + \|\boldsymbol{P}_{t-M}\|_{\mathrm{F}}^2 + \|\boldsymbol{P}_v\|_{\mathrm{F}}^2 + \|\boldsymbol{P}_t\|_{\mathrm{F}}^2) \\ \text{s.t.}\quad & \boldsymbol{b}_t \in \{-1,1\}^l,\ \boldsymbol{P}_{t-M}^{\mathrm{T}}\boldsymbol{P}_{t-M} = \boldsymbol{I},\ \boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{P}_{v-M} = \boldsymbol{I}\end{aligned} \tag{1}$$

式中,μ_V,μ_T,λ_V,λ_T,γ_V,γ_T 为惩罚参数, α_1 是正则项参数, $\|\cdot\|_{\mathrm{F}}^2$ 表示 F 范数,通过优化可以最终得到投影矩阵 $\boldsymbol{P}_{v-M}$,$\boldsymbol{P}_{t-M}$,$\boldsymbol{P}_v$,$\boldsymbol{P}_t$ 、统一的哈希码矩阵 $\boldsymbol{B}$ 、嵌入空间 $\boldsymbol{M}_v$ 和 $\boldsymbol{M}_t$ 以及线性分类器 $\boldsymbol{W}$ 。从目标函数式(1)各项的作用看,第 1 项引入标签信息,提升了哈希码的判别性;第 2、3 项将原始数据投影至各自子空间,一定程度上避免了直接将异质数据映射至共同汉明空间的困难;第 4、5 项将嵌入空间再映射至汉明空间,所学得的哈希码不仅为具有异质特性的不同模态数据搭建起了共同桥梁,而且可以通过汉明距离衡量数据间的相似性;第 6、7 项分别为图像子空间和文本子空间的图模型,用来在各自子空间中保持数据间的结构信息;最后一项可以有效抑制投影矩阵更新时发生过拟合问题。

2.3 算法推导

式(1)是一个带有 8 个变量的非凸优化问题,很难直接求解。但当固定另外 7 个变量时,这个非凸优化问题就变成了凸优化问题,因此可以循环迭代求解。

1)求解分类器 $\boldsymbol{W}$ 时,固定其他变量,则式(1)等价于

$$\min_{\boldsymbol{W}}\|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\|_{\mathrm{F}}^2 + \alpha_1\|\boldsymbol{W}\|_{\mathrm{F}}^2 \tag{2}$$

式(2)是一个正则最小二乘问题,因此只需对 $\boldsymbol{W}$ 求偏导,然后令偏导等于 0,可得

$$\boldsymbol{W} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \alpha_1\boldsymbol{I})^{-1}\boldsymbol{B}\boldsymbol{Y}^{\mathrm{T}} \tag{3}$$

2)求解投影矩阵 $\boldsymbol{P}_{v-M}$ 时,固定其他变量,则式(1)等价于

$$\begin{aligned}\min_{\boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{P}_{v-M}=\boldsymbol{I}} & \mu_V\|\boldsymbol{M}_v - \boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V}\|_{\mathrm{F}}^2 + \\ & \gamma_V\mathrm{tr}(\boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{L}_V\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_{v-M}) + \alpha_1\|\boldsymbol{P}_{v-M}\|_{\mathrm{F}}^2\end{aligned} \tag{4}$$

式(4)有正交约束,这是一个非凸优化问题,这里采用 Wen 和 Yin(2013)所提的算法对 $\boldsymbol{P}_{v-M}$ 进行优化,同理可以优化 $\boldsymbol{P}_{t-M}$ 。

3)求解嵌入空间 $\boldsymbol{M}_v$ 时,固定其他变量,则式(1)等价于

$$\min_{\boldsymbol{M}_v}\mu_V\|\boldsymbol{M}_v - \boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V}\|_{\mathrm{F}}^2 + \lambda_V\|\boldsymbol{B} - \boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v\|_{\mathrm{F}}^2 \tag{5}$$

对式(5)求关于 $\boldsymbol{M}_v$ 的偏导,并使偏导数等于 0,可得

$$\boldsymbol{M}_v = (\mu_V\boldsymbol{I} + \lambda_V\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{P}_v)^{-1}(\mu_V\boldsymbol{P}_{v-M}^{\mathrm{T}}\boldsymbol{V} + \lambda_V\boldsymbol{P}_v\boldsymbol{B}) \tag{6}$$

同理可得

$$\boldsymbol{M}_t = (\mu_T\boldsymbol{I} + \lambda_T\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{P}_t)^{-1}(\mu_T\boldsymbol{P}_{t-M}^{\mathrm{T}}\boldsymbol{T} + \lambda_T\boldsymbol{P}_t\boldsymbol{B}) \tag{7}$$

4)求解投影矩阵 $\boldsymbol{P}_v$ 时,固定其他变量,则式(1)等价于

$$\min_{\boldsymbol{P}_v}\lambda_V\|\boldsymbol{B} - \boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v\|_{\mathrm{F}}^2 + \alpha_1\|\boldsymbol{P}_v\|_{\mathrm{F}}^2 \tag{8}$$

对式(8)求关于 $\boldsymbol{P}_v$ 的偏导,并使偏导数等于 0,可得

$$\boldsymbol{P}_v = (\lambda_V\boldsymbol{M}_v^{\mathrm{T}}\boldsymbol{M}_v + \alpha\boldsymbol{I})^{-1}\lambda_V\boldsymbol{M}_v\boldsymbol{B}^{\mathrm{T}} \tag{9}$$

同理可得

$$\boldsymbol{P}_t = (\lambda_T\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{M}_t + \alpha\boldsymbol{I})^{-1}\lambda_T\boldsymbol{M}_t\boldsymbol{B}^{\mathrm{T}} \tag{10}$$

5)求解哈希码矩阵 $\boldsymbol{B}$ 时,固定其他变量,则式(1)等价于

$$\begin{aligned}\min_{\boldsymbol{B}} & \|\boldsymbol{Y} - \boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda_T\|\boldsymbol{B} - \boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t\|_{\mathrm{F}}^2 + \\ & \lambda_V\|\boldsymbol{B} - \boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v\|_{\mathrm{F}}^2 \\ \text{s.t.}\quad & \boldsymbol{B} \in \{-1,1\}^l\end{aligned} \tag{11}$$

对于式(11),直接求哈希码矩阵 $\boldsymbol{B}$ 是一个 NP 难题,但是 $\boldsymbol{B}$ 的每一行的闭式解都可以通过固定其

他行得到,因此式(11)等价于

$$\min_{\boldsymbol{B}} \|\boldsymbol{Y}\|_{\mathrm{F}}^2 - 2\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}) + \|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda_T(\|\boldsymbol{B}\|_{\mathrm{F}}^2 - 2\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t) + \|\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t\|_{\mathrm{F}}^2) + \lambda_V(\|\boldsymbol{B}\|_{\mathrm{F}}^2 - 2\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v) + \|\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v\|_{\mathrm{F}}^2)$$
$$\text{s.t.} \quad \boldsymbol{B} \in \{-1,1\}^l \tag{12}$$

显然 $\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}) = l$ 为常数,所以上式可以化简为

$$\min_{\boldsymbol{B}} -2\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}) + \|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\|_{\mathrm{F}}^2 - 2\lambda_T\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t) - 2\lambda_V\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v)$$
$$\text{s.t.} \quad \boldsymbol{B} \in \{-1,1\}^l \tag{13}$$

将式(13)进一步化简可得

$$\min_{\boldsymbol{B}} |\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}|_F^2 - 2\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Q})$$
$$\text{s.t.} \quad \boldsymbol{B} \in \{-1,1\}^l,\ \boldsymbol{Q} = \boldsymbol{W}\boldsymbol{Y} + \lambda_T\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{M}_t + \lambda_V\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{M}_v \tag{14}$$

可以利用 DCC(discrete cyclic coordinate descent)算法迭代求解 $\boldsymbol{B}$ 的每一行。令 $\boldsymbol{z}^{\mathrm{T}}$ 表示 $\boldsymbol{B}$ 的第 l 行,这里 $l = 1,\cdots,L$。$\boldsymbol{B}'$ 表示不包括 $\boldsymbol{z}$ 的矩阵 $\boldsymbol{B}$。令 $\boldsymbol{q}^{\mathrm{T}}$ 表示 $\boldsymbol{Q}$ 的第 l 行,同样,$\boldsymbol{Q}'$ 表示不包括 $\boldsymbol{q}$ 的矩阵 $\boldsymbol{Q}$。令 $\boldsymbol{w}^{\mathrm{T}}$ 表示 $\boldsymbol{W}$ 的第 l 行,同样,$\boldsymbol{W}'$ 表示不包括 $\boldsymbol{w}$ 的矩阵 $\boldsymbol{W}$,那么可以得到

$$\boldsymbol{z} = \mathrm{sgn}(\boldsymbol{q} - \boldsymbol{B}'\boldsymbol{W}') \tag{15}$$

将所提算法的实现步骤总结如下。

输入:图像数据 $\boldsymbol{V}$;文本数据 $\boldsymbol{T}$;类标矩阵 $\boldsymbol{Y}$;哈希码长度 L;模型参数 $\mu_V,\mu_T,\lambda_V,\lambda_T,\gamma_V,\gamma_T$。

1) 将图像数据 $\boldsymbol{V}$ 和文本数据 $\boldsymbol{T}$ 的每一行进行零均值化;

2) 使用随机矩阵初始化 $\boldsymbol{W},\boldsymbol{M}_v,\boldsymbol{M}_t,\boldsymbol{P}_{v-M},\boldsymbol{P}_{t-M},\boldsymbol{P}_v,\boldsymbol{P}_t$,用一个随机 $\{-1,1\}^{n\times L}$ 初始化 $\boldsymbol{B}$;

3) 开始循环:

(1) 根据式(3)计算分类器 $\boldsymbol{W}$;

(2) 根据步骤 2)计算投影矩阵 $\boldsymbol{P}_{v-M},\boldsymbol{P}_{t-M}$;

(3) 根据式(6)和式(7)计算嵌入空间 $\boldsymbol{M}_v,\boldsymbol{M}_t$;

(4) 根据式(9)和式(10)计算投影矩阵 $\boldsymbol{P}_v,\boldsymbol{P}_t$;

(5) 根据式(15)计算哈希码矩阵 $\boldsymbol{B}$。

4) 直到式(1)收敛;

输出:哈希码矩阵 $\boldsymbol{B}$;投影矩阵 $\boldsymbol{P}_{v-M},\boldsymbol{P}_{t-M}$;投影矩阵 $\boldsymbol{P}_v,\boldsymbol{P}_t$。

3 实验结果与分析

采用 3 个在跨模态检索领域常用的数据集来测试所提算法的性能,分别为 Wikipedia 数据集、MIRFlickr 数据集和 Pascal Sentence 数据集,其中最后一个数据集采用了深度特征。对比算法包括 LSSH(Zhou 等, 2014)、CMFH(Ding 等, 2014)、SCM(Zhang 和 Li,2014)、SePH(Lin 等,2015)、DCH(Xu 等,2017)、DSPH(Zhang 等,2020)和 DHCSC(严双咏 等,2019)。评价指标为平均检索精度(mean average precision, mAP),本文采用与 Ding 等人(2014)相同的评价指标。

1) Wikipedia 数据集。该数据集包含 10 类数据共 2 866 个图像文本对。图像数据采用 128 维 SIFT(scale-invariant feature transform)特征进行表示,文本数据采用 500 维 BoW(bag of words)特征进行表示,然后用 LDA(latent dirichlet allocation)模型构建 10 维文本特征。将其中的 2 173 对数据划分成训练集,其余 693 对构成测试集。

2) MIRFlickr 数据集。该数据集有 16 738 个样本,共 24 类,全部来源于 Flickr 网站。其中图像用 150 维柱状图表示,文本则通过 PCA(principal component analysis)方法构建的 500 维特征进行表示。将其中 15 902 对数据划分成训练集,其余 836 对构成测试集。

3) Pascal Sentence 数据集。该数据集包含 1 000 个图像文本数据对,共 20 类。其中图像数据采用1 024 维CNN(convolutional neural networks)视觉特征进行表示,文本数据则是先提取 300 维 BoW 特征,然后用 LDA 模型构建 100 维文本特征。将其中 600 对数据划分成训练集,剩下 400 对构成测试集。

3.1 实验结果与分析

3.1.1 浅度特征数据集实验

本文采用浅度特征数据集 Wikipedia 和 MIRFlickr 测试 SPHCP 算法的性能,表 1 为 SPHCP 算法以及对比方法在这两个数据集上 I to T 和 T to I 的检索 mAP 值。为衡量哈希码位数对算法性能的影响,实验将哈希码分别设置成 16 bit、32 bit、64 bit 和 128 bit。

从表 1 可以看到:

1)对于数据集 Wikipedia 和 MIRFlickr,本文所提算法在任务图像检索文本(I to T)和任务文本检索图像(T to I)上 mAP 值都超过了其他算法,这表明 SPHCP 算法可以有效地进行跨模态检索。

表 1　在 Wikipedia 和 MIRFlickr 数据集上各算法的平均检索精度(mAP)

Table 1　Cross-modal retrieval performance(mAP) of different methods on Wikipedia and MIRFlickr datasets with various code lengths

/%

方法		Wikipedia				MIRFlickr			
		16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
I to T	LSSH	20.83	22.09	22.14	22.07	56.84	57.05	56.99	57.23
	CMFH	21.32	22.59	23.62	24.19	57.99	58.32	58.97	58.99
	SCM	22.13	25.89	27.88	28.33	62.37	63.43	64.48	64.89
	SePH	23.91	26.50	27.30	27.73	66.99	67.19	67.71	67.83
	DCH	27.16	29.01	31.44	32.31	67.43	67.74	68.01	69.31
	DSPH	27.42	28.36	29.52	31.28	67.52	68.31	68.65	69.04
	DHCSC	27.33	28.56	30.61	32.51	67.89	68.32	68.45	68.13
	SPHCP	**33.87**	**36.03**	**37.32**	**38.08**	**68.82**	**69.38**	**70.53**	**71.06**
T to I	LSSH	50.21	52.08	52.77	53.25	57.16	58.32	59.24	59.33
	CMFH	48.84	51.32	52.69	53.75	59.37	59.19	59.31	59.24
	SCM	21.34	23.66	24.79	25.73	61.55	62.12	63.04	63.40
	SePH	55.75	57.12	57.96	58.40	72.01	72.75	73.33	73.80
	DCH	65.63	69.57	71.06	71.21	75.43	75.84	76.89	76.68
	DSPH	56.25	57.46	58.91	59.03	66.72	67.15	69.82	70.15
	DHCSC	66.28	69.64	70.41	71.12	74.31	75.23	76.13	76.11
	SPHCP	**70.06**	**72.05**	**72.33**	**72.53**	**78.72**	**80.50**	**81.67**	**83.49**

注:加粗字体为每列最优值。

2)SPHCP 算法的 mAP 值比性能第 2 的 DHCSC 算法在 I to T 任务上高了约 3%,而在 MIRFlickr 数据集上,T to I 任务中,SPHCP 算法的 mAP 值高了约 5%,这证明了采用耦合投影和图模型能够有效地提升算法性能。

3)随着哈希码位数的增加,所有算法的 mAP 值都在提升,这是因为哈希码位数越多,哈希码所能表示的信息也越多。但实验对比算法在哈希码位数从 64 bit 升到 128 bit 时,mAP 值提升很小,甚至出现了下降趋势,与之相比,本文算法的 mAP 值仍在稳步提升,这表明通过类标所获取的信息是有限的,而通过图模型可以进一步挖掘文本和图像原始数据的结构信息,从而确保检索精度。

3.1.2　深度特征数据库扩展实验

进一步将所提算法在 Pascal Sentence 数据集的深度特征库上进行实验,这里将 T-V CCA(three view canonical correlation analysis)(Gong 等,2012)、Deep-SM(deep semantic matching)(Wei 等,2017)、DCH、DSPH 和 DHCSC 作为对比方法,表 2 为各算法在该数据集上的 mAP 值。

表 2　在 Pascal Sentence 数据集上各算法的平均检索精度(mAP)

Table 2　Cross-modal retrieval performance(mAP) of different methods on Pascal Sentence dataset

/%

方法	I to T/ T to I
T-V CCA	31.12/34.64
Deep-SM	39.80/35.42
DCH	44.14/65.81
DSPH	46.58/68.27
DHCSC	37.63/70.22
SPHCP	**54.01/77.63**

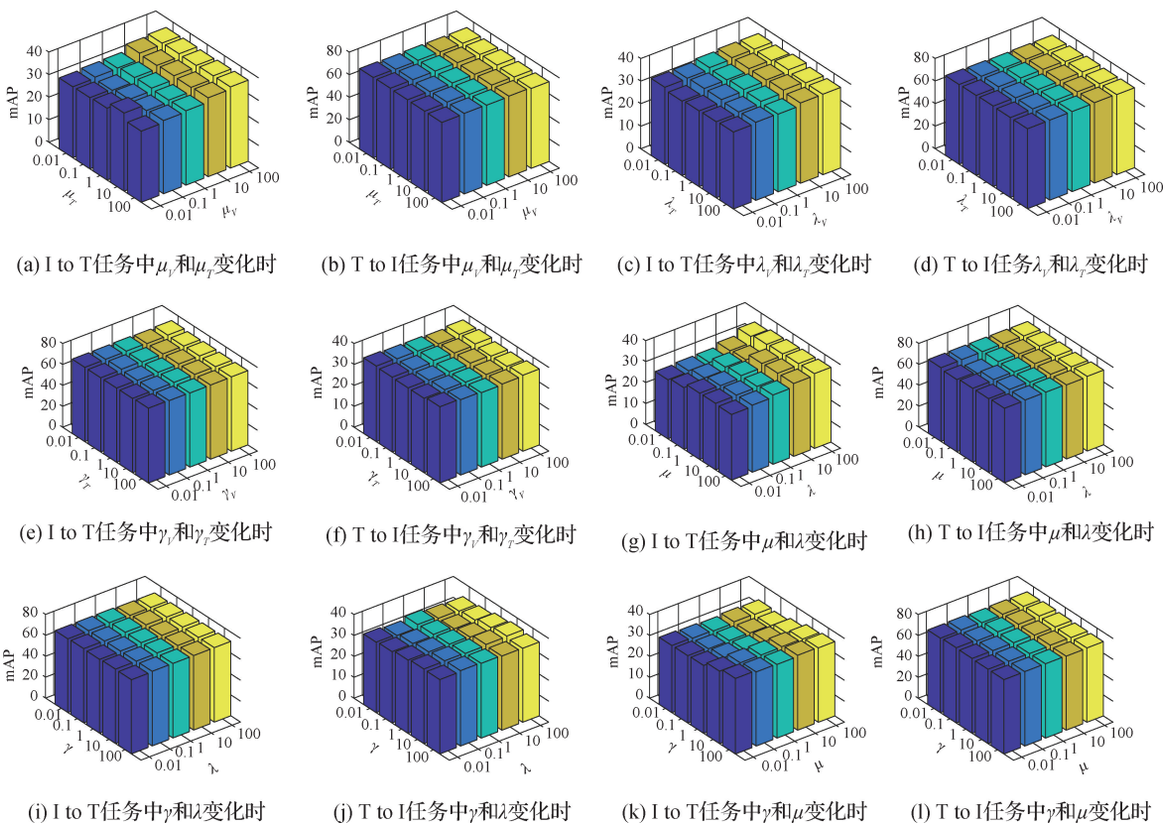
注:加粗字体为最优值。

从实验结果中可以发现，在任务 I to T 和 T to I 上，SPHCP 算法的 mAP 分别达到了 54.01% 和 77.63%，明显高于其他算法，这表明 SPHCP 也可以很好地处理具有深度特征的数据。需要注意的是，虽然本文算法提升效果明显，但各算法在任务 I to T 上的性能均非常有限，因此 Pascal Sentence 数据集也是目前跨模态检索的挑战难点。

一般来说，基于深度学习的跨模态检索方法可以较好地处理非线性特征，因此检索精度可能会比传统的方法高，但这是以损耗时间和计算资源为代价的。而作为一种“浅度”的方法，所提 SPHCP 算法精度提升明显，消耗的时间却较短。因此是一个有意义的方向，未来可以把该算法和深度特征提取方法配合使用：用深度学习离线提取图像和文本的特征，用 SPHCP 算法进行快速检索。

3.2　参数敏感性分析

在 Wikipedia 数据集上测试所提算法的参数敏感性，其中哈希码长度设为 16 bit。SPHCP 算法中有7 个参数，其中 α_1 为正则项参数，其作用为防止过拟合，按经验将其设为 10，另外 6 个参数可以采用控制变量法进行分析，参数范围设置为 $\{10^{-2}, 10^{-1}, \cdots, 10^{2}\}$。首先固定 λ_V，λ_T，γ_V 和 γ_T 测试 μ_V，μ_T，实验结果如图 2(a)(b)所示，很明显 μ_V 和 μ_T 的取值对算法的 mAP 值影响不大；然后固定 μ_V，μ_T，γ_V 和 γ_T 测试 λ_V，λ_T，实验结果如图 2(c)(d)所示，容易发现 λ_V，λ_T 的取值对算法的 mAP 值



(a) I to T任务中μ_V和μ_T变化时　(b) T to I任务中μ_V和μ_T变化时　(c) I to T任务中λ_V和λ_T变化时　(d) T to I任务λ_V和λ_T变化时

(e) I to T任务中γ_V和γ_T变化时　(f) T to I任务中γ_V和γ_T变化时　(g) I to T任务中μ和λ变化时　(h) T to I任务中μ和λ变化时

(i) I to T任务中γ和λ变化时　(j) T to I任务中γ和λ变化时　(k) I to T任务中γ和μ变化时　(l) T to I任务中γ和μ变化时

图 2　在 Wikipedia 数据集上 SPHCP 算法在不同参数下的平均检索精度

Fig. 2　The mAP of SPHCP with different values of model parameters on Wikipedia dataset

((a) mAP of I to T with different μ_V and μ_T; (b) mAP of T to I with different μ_V and μ_T; (c) mAP of I to T with different λ_V and λ_T; (d) mAP of T to I with different λ_V and λ_T; (e) mAP of I to T with different γ_V and γ_T; (f) mAP of T to I with different γ_V and γ_T; (g) mAP of I to T with different μ and λ; (h) mAP of T to I with different μ and λ; (i) mAP of I to T with different γ and λ; (j) mAP of T to I with different γ and λ; (k) mAP of I to T with different γ and μ; (l) mAP of T to I with different γ and μ)

影响不大;同理,图2(e)(f)表明 γ_T 和 γ_V 的取值对算法的 mAP 值影响也不大。为了进一步研究不同模态的影响,不妨令 $\mu_V = \mu_T = \mu$, $\lambda_V = \lambda_T = \lambda$ 和 $\gamma_V = \gamma_T = \gamma$,图2(g)(h),图2(i)(j)和图2(k)(l)分别为固定 γ,测试 μ,λ 对算法的影响;固定 μ,测试 γ, λ 对算法的影响;以及固定 λ,测试 μ,γ 对算法的影响。显然,综合图2来看,所提算法对参数不敏感,适合实际应用。

4 结 论

本文提出一种耦合保持投影哈希跨模态检索算法(SPHCP)。为了处理跨模态数据之间的模态"鸿沟",算法采用子空间耦合投影的方式,达到了逐步减少不同数据模态差异的目的,从而可以得到较为一致的哈希码。为了让哈希码兼顾原始数据的结构信息和语义判别性,算法在子空间学习中引入了图模型来保持数据的类内类间关系。在哈希码生成的过程中引入了类标回归项来提升哈希码的判别性。

公开数据集上的实验结果验证了所提算法的优越性能:一方面,在浅度特征数据集上,SPHCP 的平均检索精度优于主流的哈希方法;另一方面,面对深度特征,SPHCP 的检索精度也令人满意。但是需要注意的是,由于 SPHCP 引入了图模型来保持数据间的结构信息,当处理大规模数据集时,计算各个样本点之间的相似性需要较大的算力支持。

未来研究中拟进一步引入非线性特征映射,提升 SPHCP 算法对具有非线性特征数据的处理能力;另外还可将 SPHCP 跨模态检索算法进一步拓展成为多模态数据检索算法。

参考文献(References)

Ding G G, Guo Y C and Zhou J L. 2014. Collective matrix factorization hashing for multimodal data//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 2083-2090 [DOI: 10.1109/CVPR.2014.267]

Gong Y C, Ke Q F, Isard M and Lazebnik S. 2012. A multi-view embedding space for modeling internet images, tags, and their semantics. International Journal of Computer Vision, 106(2): 210-233 [DOI: 10.1007/s11263-013-0658-4]

Hu M Q, Yang Y, Shen F M, Xie N, Hong R C and Shen H T. 2019. Collective reconstructive embeddings for cross-modal hashing. IEEE Transactions on Image Processing, 28(6): 2770-2784 [DOI: 10.1109/TIP.2018.2890144]

Jiang Q Y and Li W J. 2017. Deep cross-modal hashing//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3270-3278 [DOI: 10.1109/CVPR.2017.348]

Jin L, Li K, Li Z C, Xiao F, Qi G J and Tang J H. 2019. Deep semantic-preserving ordinal hashing for cross-modal similarity search. IEEE Transactions on Neural Networks and Learning Systems, 30(5): 1429-1440 [DOI: 10.1109/TNNLS.2018.2869601]

Kumar S and Udupa R. 2011. Learning hash functions for cross-view similarity search//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Catalonia, Spain: IJCAI/AAAI: 1360-1365 [DOI: 10.5591/978-1-57735-516-8/IJCAI11-230]

Li C, Deng C, Li N, Liu W, Gao X B and Tao D C. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval //Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4242-4251 [DOI 10.1109/CVPR.2018.00446]

Lin Z J, Ding G G, Hu M Q and Wang J M. 2015. Semantics-preserving hashing for cross-view retrieval//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3864-3872 [DOI: 10.1109/CVPR.2015.7299011]

Sharma A, Kumar A, Daume H and Jacobs D W. 2012. Generalized multiview analysis: a discriminative latent space//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2160-2167 [DOI: 10.1109/CVPR.2012.6247923]

Wei Y C, Zhao Y, Lu C Y, Wei S K, Liu L Q, Zhu Z F and Yan S C. 2017. Cross-modal retrieval with CNN visual features: a new baseline. IEEE Transactions on Cybernetics, 47(2): 449-460 [DOI: 10.1109/TCYB.2016.2519449]

Wen Z W and Yin W T. 2013. A feasible method for optimization with orthogonality constraints. Mathematical Programming, 142(1-2): 397-434 [DOI: 10.1007/s10107-012-0584-1]

Wu J L, Lin Z C and Zha H B. 2017. Joint latent subspace learning and regression for cross-modal retrieval//Proceedings of the 40th International ACM Sigir Conference. Shinjuku, Japan: ACM: 917-920 [DOI: 10.1145/3077136.3080678]

Xu X, Shen F M, Yang Y, Shen H T and Li X L. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Transactions on Image Processing, 26(5): 2494-2507 [DOI: 10.1109/TIP.2017.2676345]

Yan S Y, Liu C H, Jiang A W, Ye J H and Wang M W. 2019. Discriminative cross-modal hashing with coupled semantic correlation. Chinese Journal of Computers, 42(1): 164-175 (严双咏, 刘长红, 江爱文, 叶继华, 王明文. 2019. 语义耦合相关的判别式跨模态哈希学习算法. 计算机学报, 42(1): 164-175) [DOI: 10.

11897/SP. J. 1016. 2019. 00164］

Zhang D Q and Li W J. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec City, Canada: AAAI Press: 2177-2183［DOI: 10. 5555/2892753. 2892854］

Zhang D, Wu X J and Yu J. 2020. Learning latent hash codes with discriminative structure preserving for cross-modal retrieval. Pattern Analysis and Applications, 24(4):283-297［DOI: 10. 1007/s10044-020-00893-6］

Zhou J, Ding G G and Guo Y C. 2014. Latent semantic sparse hashing for cross-modal similarity search//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia: ACM: 415-424［DOI: 10. 1145/2600428. 2609610］

作者简介

闵康凌，1996 年生，男，硕士研究生，主要研究方向为多媒体信息检索。
E-mail：1213447021@ qq. com

李丹萍，通信作者，女，讲师，主要研究方向为通信信号处理、模式识别、多模态数据融合。
E-mail：dpli@ xidian. edu. cn

张国宾，男，高级工程师，主要研究方向为多模态信号处理。
E-mail：zgb115@ 126. com

王磊，男，副教授，主要研究方向为信号处理、机器学习。
E-mail：leiwang@ mail. xidian. edu. cn