基于异构哈希网络的跨模态人脸检索方法

董 震 裴明涛
(北京理工大学智能信息技术北京市重点实验室 北京 100081)
(北京理工大学计算机学院 北京 100081)

摘 要 该文提出一种基于异构哈希网络的跨模态人脸检索方法.异构哈希网络能够将位于不同空间的人脸图像和人脸视频映射到一个公共且有判别力的二值空间上,以获得有效的二值哈希表示.该网络包含图像分支、视频分支和哈希函数三个部分,首先图像和视频分支分别将人脸图像和人脸视频映射到一个公共空间,然后在公共空间中学习非线性哈希函数.网络的训练使用了三种损失函数:Fisher 损失、softmax 损失和三元排序损失(triplet ranking loss),其中的 Fisher 损失关注于公共空间的判别力,softmax 损失强调公共空间上表达的可分性,三元排序损失旨在提升最终的检索性能.在多个人脸视频数据集上的跨模态人检索实验结果表明了所提出方法的有效性.

关键词 异构哈希网络;跨模态;人脸检索;深度学习;损失函数
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2019.00073

Cross-Modality Face Retrieval Based on Heterogeneous Hashing Network

DONG Zhen PEI Ming-Tao
(*Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing* 100081)
(*School of Computer Science, Beijing Institute of Technology, Beijing* 100081)

Abstract Cross-modality face retrieval is to retrieve faces of a particular person in one modality given his/her face information in another modality, such as retrieving video shots containing particular person given one image of him/her (query-by-image video retrieval), or retrieving the face images of one person by using his/her video clip as query (query-by-video image retrieval). It is an important problem in computer vision with wide range of applications. Take the criminal investment as an example. The "query-by-image video retrieval" task plays an important role in rapid locating and tracking a suspect from masses of surveillance videos with the ID card, passport, or driver license photo of the suspect as query. The "query-by-video image retrieval" task helps to determine the identity of an unknown suspect by retrieving a huge mug-shot image database given his/her video shot taken by the surveillance cameras in the crime scene. In this paper, we present a cross-modality face retrieval method which uses the heterogeneous hashing network to generate effective and compact hash representations for both face images and face videos. The network contains an image branch and a video branch to project face images and videos into a common discriminative space, respectively. Each channel are equipped with two modules: feature extractor module and non-linear mapping module. The feature extractor modules aim to represent face images or videos via appropriate features, and the non-linear mapping modules are designed to transform the heterogenous image and video feature spaces into the common space. On the common space, the similarity between a face image and a face video can be measured through the distance of their corresponding discriminative features, but these features are still

收稿日期:2017-05-04;在线出版日期:2018-07-11.本课题得到国家自然科学基金(61472038)资助.**董 震**,男,1990 年生,博士,主要研究方向为计算机视觉、人脸识别、人脸检索.E-mail: dongzhen@bit.edu.cn.**裴明涛**(通信作者),男,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、人脸识别.E-mail: peimt@bit.edu.cn.

high-dimensional vectors of floating point numbers, which cannot satisfy the requirements of low computation and storage complexities in the retrieval task. The non-linear hash functions are thus learned in the common space to obtain the corresponding binary hash representations. To catch the compatability and the effectiveness of the branches and the hash functions, the heterogeneous hashing network is trained with three loss functions: Fisher loss, softmax loss, and triplet ranking loss. Our Fisher loss uses the difference form of the inter-class and the intra-class scatter where the mean vectors are learnable variables, which is feasible for the mini-batch based optimization method. The Fisher loss and the softmax loss are jointly exploited to enhance the discriminative power of the common space. The triplet ranking loss is enforced to the final binary space for the improvement of the retrieval performance. Experiments on a large-scale face video dataset and two challenging TV-series datasets demonstrate the effectiveness of the proposed method. The contributions of the paper are three-folds: (1) We propose an effective cross-modality face retrieval method based on the heterogeneous hashing network. Our network is able to generate isomorphic discriminative compact binary representations of both face images and videos. (2) The proposed heterogeneous hashing network provides a general framework for deep learning based cross-modality hashing methods, and can be easily adopted in many other cross-modality retrieval tasks. (3) The proposed method achieves excellent results of face retrieval across image and video modalities on a large-scale face video dataset and two challenging TV-series datasets.

Keywords heterogeneous hashing network; cross-modality; face retrieval; deep learning; loss function

1 引 言

现实世界中采集到的人脸数据可能存在多种模态,按照载体类型,可以分为人脸图像和人脸视频;按照具体表现形式,可以分为照片,肖像画,素描画,漫画等;按照采集设备类型,可以分为可见光人脸数据和红外人脸数据等.给定一个特定人物某种模态的人脸信息,跨模态人脸检索任务是指从包含另一种模态人脸信息的数据库中找寻到该人,例如给定某人的人脸图像,在人脸视频数据库中检索到包含该人的所有人脸视频(基于图像的人脸视频检索);或者根据某人的人脸视频,在人脸图像数据库中检索含有该人的所有人脸图像(基于视频的人脸图像检索).伴随着互联网和自媒体技术的不断发展,互联网上多种模态的人脸数据越来越丰富,这使得跨模态人脸检索具有非常广泛的应用前景.以罪犯追踪和身份分析为例,基于图像的人脸视频检索的应用场景是:根据嫌疑犯的身份证、护照或者驾驶证等证件上的人脸图像信息,从含有大量监控视频的数据库中快速定位和追踪该嫌疑犯;基于视频的人脸图像检索的应用场景是:根据案发现场监控拍摄到的含有罪犯的视频,对公安机关备案的人脸图像数据集进行检索以确定罪犯的身份.

以人脸图像和人脸视频为例,跨模态人脸检索需要度量一张人脸图像和一个人脸视频的相似性.最直接的度量方法是将人脸视频看成帧数据的无序集合,利用人脸图像与视频所有帧的相似性的最大值或平均值作为该人脸图像和视频的相似性.然而,这种度量方法没有考虑人脸视频各帧之间的相关性,浪费了宝贵的相关信息,同时在人脸视频包含的帧数较多时还面临着较大的计算和存储消耗问题.对一个人脸视频片段中的所有帧联合建模可以充分利用各帧之间的相关信息,并生成一个与帧数无关的人脸视频表示,然而,对人脸视频所有帧的联合建模将导致视频特征和图像特征处于异构的特征空间,使得跨模态人脸检索任务变得非常困难.

本文构造了一类称为异构哈希网络(Heterogeneous Hashing Network,HHN)的深度神经网络,将位于异构空间的特征映射到一个公共的且有足够判别力的二值空间上,实现不同模态数据之间的度量,从而实现跨模态的人脸检索任务.异构哈希网络的结构如图 1 所示.网络包含图像和视频两个分支,分别用来处理人脸图像数据和人脸视频数据,将它

们从异构特征空间映射到一个公共空间来度量相似性.每个分支包含两个模块:特征提取模块和同构表示生成模块.特征提取模块完成人脸图像或人脸视频的特征提取和表示,例如人脸图像的卷积神经网络特征[1-2]、人脸视频的 3D 卷积网络特征[3]或者神经聚合网络特征[4]等;同构表示生成模块将位于两个异构空间的人脸图像特征和人脸视频特征映射到一个公共且有判别力的空间.在公共空间上,人脸图像和人脸视频之间的相似性可以通过它们在公共空间上的相应特征表示之间的距离来度量.然而,公共空间上的特征表示仍然是高维的浮点数向量,不能满足检索任务对检索算法低时间和空间复杂度的要求.为了解决这个问题,在公共空间上进一步学习非线性的哈希函数来获得人脸图像和视频数据的二值哈希表示.哈希表示的二值属性,可以以极低的时间复杂度和极少的存储消耗实现人脸图像和视频之间的相似性计算,从而减少检索过程的时间和空间花费.

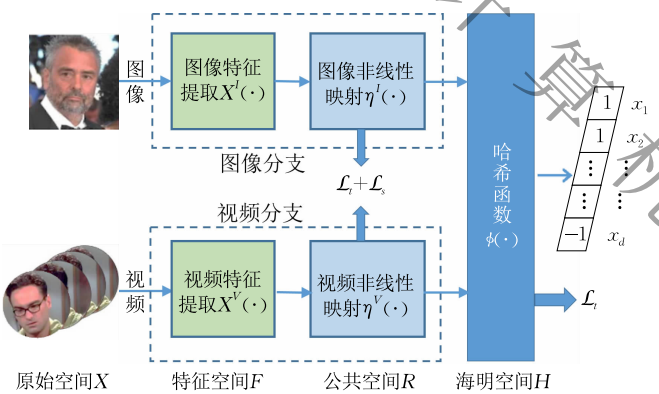

图 1　异构哈希网络结构示意图

Li 等人[5]提出了一种跨欧式空间和黎曼流形的哈希方法,用于学习人脸图像和视频的统一哈希表示.该方法首先利用核函数将异构特征空间映射至高维再生希尔伯特空间,然后在再生希尔伯特空间上学习哈希函数进而获得原始数据的哈希表达,并利用学得的哈希表达直接度量人脸图像和视频数据之间的距离.尽管该方法取得了较好的检索结果,但该方法将特征提取和哈希函数的学习视作两个独立的过程,这可能会导致特征表示与哈希函数的不兼容,从而对最终的检索性能造成影响.与之不同,本文方法将特征学习和哈希函数学习融合到一个统一的学习框架中联合进行,使得两个部分可以相互促进,从而更好地为最终的检索任务服务.

本文提出的基于异构哈希网络的跨模态人脸检索方法是首个利用深度神经网络来实现跨模态人脸检索的方法,所提出的异构哈希网络能够为位于异构空间的人脸图像和视频数据生成同一空间上的紧致二值哈希表示;异构哈希网络为跨模态检索问题提供了一套深度网络的通用框架,本文设计的网络不仅可以完成跨图像和视频模态的检索任务,还可以扩展到其他跨模态的检索任务上;在多个人脸视频数据集上进行的两类跨模态人脸检索实验,包括基于图像的人脸视频检索和基于视频的人脸图像检索,均取得了非常好的检索结果.

2　相关工作

2.1　单模态哈希方法

Gionis 等人[6]提出的局部敏感哈希(LSH)方法是数据无关的哈希方法的代表.LSH 方法使用随机投影作为哈希函数将高维特征表示投影到低维二值空间,哈希函数的设计原则是:原始特征空间中距离近的点的哈希表示发生碰撞的概率大于距离远的点的碰撞概率,所有满足这一设计原则的函数构成 LSH 哈希函数族.Weiss 等人[7]提出的谱哈希方法期望在原始特征空间中欧式距离近的数据点在二值空间中海明距离依然近,针对此要求,谱哈希方法将哈希函数的求解建模为图的分割过程,其中图的顶点表示数据点,顶点之间边的权值表示数据点间的欧式距离,之后可利用该图的拉普拉斯矩阵的特征向量作为求解的哈希函数.Gong 等人[8]提出的迭代量化哈希方法首先对中心化后的数据进行 PCA 降维,然后迭代执行数据集合的旋转和量化误差的最小化,最终获得最优解.Wang 等人[9]提出了半监督哈希方法,该方法在带标签的数据上最小化经验误差,同时在所有数据上包括有标签数据和无标签数据最大化哈希码的方差和均衡性,既可以建模度量信息,也可以建模语义相似性信息.Liu 等人[10]探究了二值空间的海明距离与哈希码内积表示的相似性度量之间的关系,并提出了有监督的核哈希方法,将目标函数中海明距离巧妙地转化为二值哈希编码的内积形式,从而可以方便地进行优化.

2.2　多模态哈希方法

单模态的哈希方法在很多领域取得了成功,但是只能检索同一个模态的数据,而不能处理多模态数据检索的问题.多模态检索问题在近年来受到了很大的关注[11-22].已有的多模态哈希方法可以分为两类:无监督方法和有监督方法.代表性的无监督多模态哈希方法包括对应自编码器哈希方法[11]、跨视角哈希方法[12]和可预测对偶视角哈希方法[13].这三

类方法的学习目标分别是重构误差最小化、基于图的相似性保持和可预测性的维持.典型的有监督多模态哈希方法有跨模态相似敏感哈希方法[14]、多模态隐二值化嵌入[15]、多模态神经网络哈希方法[16]、参数化局部多模态哈希方法[17]、语义相关最大化哈希方法[18]和语义保持哈希方法[19].与无监督的多模态哈希方法相比,有监督的多模态哈希方法能够充分地利用训练数据的语义信息来降低多模态数据之间的差异化程度,并减小不同模态数据之间的语义鸿沟,有监督的多模态哈希方法也因此比无监督的方法取得了更好的性能.Jiang 和 Li[21]提出了深度跨模态哈希方法,并将其应用在了图像和文本之间的检索问题上.Cao 等人[22]提出了深度视觉-语义哈希方法,该方法利用复杂深度神经网络来挖掘文本数据和图像数据之间的异构相关信息,并利用学习到的信息提升跨模态检索的性能.这两个工作利用深度神经网络在跨图像域和文本域的检索问题中取得了较好的效果.与之不同,本文的异构哈希网络提供了一个多模态检索的深度学习框架,并基于此框架实现了人脸的跨图像域和视频域的检索,取得了很好的检索性能.

3 异构哈希网络

3.1 网络结构

如图 1 所示,异构哈希网络由五部分组成:图像分支的特征提取模块 $\chi^I(\cdot)$、图像分支的同构表示生成模块 $\eta^I(\cdot)$、视频分支的特征提取模块 $\chi^V(\cdot)$、视频分支的同构表示生成模块 $\eta^V(\cdot)$以及哈希编码生成模块 $\phi(\cdot)$.这 5 个部分都可以利用神经网络实现,因此 HHN 建立了一套用于跨模态人脸检索且可以进行端到端学习的深度网络结构.HHN 的应用范围不仅仅局限于跨模态人脸检索,只要能够设计出合适的特征提取网络,HHN 可以很容易的用于图像与视频之外的其它模态数据的检索任务.

3.2 损失函数

同构表示生成模块将原本属于不同特征空间的人脸数据映射到同一公共空间.为了增强映射后的特征表示的判别力,对公共空间添加判别性约束,则引入 Softmax 损失函数和 Fisher 损失函数,来提升属于同一个人不同模态的人脸数据之间的关联程度.同时,在哈希函数的学习过程中利用监督信息对哈希函数添加约束,即引入三元排序损失函数,使得同一个人的不同模态人脸数据的特征表示之间的差异进一步减小.

3.2.1 Softmax 损失函数

为了保证人脸数据在公共空间上的特征表示具有较强的可分性,引入 softmax 损失:

$$\mathcal{L}_s=-\sum_{i=1}^{c}\sum_{K\in\{I,V\}}\sum_{j=1}^{n_i^K}\log\frac{e^{\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{r}_{i,j}^K+b_i}}{\sum_{k=1}^{c}e^{\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{r}_{i,j}^K+b_k}} \tag{1}$$

其中 $\boldsymbol{W}_i$ 和 b_i 是 softmax 损失的参数,c 是要处理的个体的总数(类别总数),n_i^K 为第 i 个类别中图像($K=I$)或视频($K=V$)的数量,$\boldsymbol{r}_{i,j}^K$ 为第 i 个个体的第 j 个图像或视频对应的公共空间中的特征表示.

3.2.2 Fisher 损失函数

为了在增强公共空间判别力的同时减小同一个人的人脸视频与人脸图像数据特征表示的不一致性,则需要对共同空间添加判别性约束.Fisher 损失可以在最小化类内差异的同时最大化类间差异,是一个合适的选择.将人脸图像和视频的公共空间上的特征放在一起组成一个集合,该集合上的 Fisher 损失通过最小化类内散度矩阵 $\boldsymbol{S}_W$ 和类间散度矩阵 $\boldsymbol{S}_B$ 的 Rayleigh 熵(Rayleigh quotient)来实现:

$$\min\frac{tr(\boldsymbol{S}_W)}{tr(\boldsymbol{S}_B)} \tag{2}$$

其中 $tr(\cdot)$表示矩阵的迹.

为了保证 $\boldsymbol{S}_W$ 和 $\boldsymbol{S}_B$ 的非奇异性以获得最优解,式(2)所显示的 Rayleigh 熵需要满足特征表示的维度小于个体数量.然而这在实际应用中通常是难以满足的,因为用于求解哈希表示的特征向量的维度通常是比较高的.另外,由于均值向量是在整个训练集上计算的,Rayleigh 熵形式的 Fisher 损失不能利用基于批次(batch)的算法进行优化.为了解决上述的问题,使用差分形式的 Fisher 损失[23]来替代 Rayleigh 熵形式:

$$tr(\boldsymbol{S}_W)-tr(\boldsymbol{S}_B)+\|\boldsymbol{R}\|_F \tag{3}$$

其中 $\|\cdot\|_F$ 表示矩阵的 Frobenius 范数,加入此范数项的目的是保证式(3)所示的目标函数是凸的.进一步地,为了使差分形式的 Fisher 损失可以利用基于批次的算法进行优化,将原始 Fisher 损失中作为定值的均值向量转变为可学习的变量,保持均值向量的可学习性.这样,在每一次迭代中,均值向量都在训练数据的每个批次上进行更新.因此,公共空间上的判别损失表示为

$$\mathcal{L}_d=\|\boldsymbol{R}\|_F^2+\frac{1}{2n}\sum_{i=1}^{c}\sum_{K\in\{I,V\}}\sum_{j=1}^{n_i^K}\|\boldsymbol{r}_{i,j}^K-\boldsymbol{\mu}_i\|_2^2-\frac{1}{2c}\sum_{i=1}^{c}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}\|_2^2 \tag{4}$$

其中 $\boldsymbol{\mu}_i$ 表示第 i 个个体的所有样本(无论人脸视频还是图像)在公共空间上特征表示的均值向量，$\boldsymbol{\mu}$ 表示所有 c 个个体的所有样本在公共空间上特征表示的均值向量.

3.2.3　三元排序损失函数

为了学得对检索更有效的二值哈希表示，鼓励哈希函数能够在学得的正样本对的哈希表示距离和负样本对的哈希表示距离之间保持一个较大的间隔，我们采用三元排序损失[24](triplet ranking loss). 给定一个三元组 $(\boldsymbol{h},\tilde{\boldsymbol{h}},\hat{\boldsymbol{h}})$，其中 $(\boldsymbol{h},\tilde{\boldsymbol{h}})$ 构成正样本对，$(\boldsymbol{h},\hat{\boldsymbol{h}})$ 构成负样本对. 三元排序损失函数的优化目标是使得人脸哈希表示之间的相对相似性满足“与人脸 $\hat{\boldsymbol{h}}$ 相比，人脸 $\tilde{\boldsymbol{h}}$ 与人脸更相似”. 在本文方法中，正样本对与负样本对通过样本的语义信息来进行划分，即属于同一个个体的人脸数据(无论是人脸图像数据还是视频数据)可以构成正样本对，属于不同个体的人脸数据可以构成负样本对. 三元组 $(\boldsymbol{h},\tilde{\boldsymbol{h}},\hat{\boldsymbol{h}})$ 的损失表示为

$$t(\boldsymbol{h},\tilde{\boldsymbol{h}},\hat{\boldsymbol{h}})=\max(d(\boldsymbol{h},\tilde{\boldsymbol{h}})-d(\boldsymbol{h},\hat{\boldsymbol{h}})+\zeta,0) \tag{5}$$

其中，$d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ 表示二值空间中的海明距离，定义为

$$d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\frac{l-\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_2}{2} \tag{6}$$

l 是哈希编码的长度，ζ 表示学得的正样本对的哈希表达距离和负样本对的哈希表达距离之间所要保持的间隔. 将所有批次中选取的三元组的损失加和，最终的三元排序损失表示为

$$\mathcal{L}_t=\sum_{\substack{i_{1/2}=1,\\ i_1\neq i_2}}^{c}\ \sum_{K_{1/2/3}\in\{I,V\}}\ \sum_{\substack{j_{1/2/3}=1,\\ (K_2,j_2)\neq(K_1,j_1)}}^{n_{i_{1/1/2}}^{K_{1/2/3}}} t(\boldsymbol{h}_{i_1,j_1}^{K_1},\boldsymbol{h}_{i_1,j_2}^{K_2},\boldsymbol{h}_{i_2,j_3}^{K_3}) \tag{7}$$

其中的 $\boldsymbol{h}_{i,j}^{K}$ 表示的是公共空间特征 $\boldsymbol{r}_{i,j}^{K}$ 对应的二值哈希表示，即第 i 个个体的第 j 个图像或视频(由 K 的取值来决定)对应的二值哈希表示.

同时考虑公共特征空间的判别性约束、可分性约束和二值空间的三元排序损失，异构哈希网络损失函数的最终形式为

$$\min_{\Gamma,\Theta,\Lambda}\mathcal{L}=\alpha\mathcal{L}_d+\beta\mathcal{L}_s+\mathcal{L}_t \tag{8}$$

其中 $\Gamma=\{\chi^I(\cdot),\eta^I(\cdot),\chi^V(\cdot),\eta^V(\cdot),\phi(\cdot)\}$ 表示异构哈希网络的 5 个部分所包含的参数的，$\Theta=\{\boldsymbol{W}_i,b_i\mid i=1,2,\cdots,c\}$ 表示 softmax 损失层的参数，$\Lambda=\{\boldsymbol{\mu}_i\mid i=1,2,\cdots,c\}$ 是 Fisher 损失层的参数，α 和 β 是用于平衡各项重要性的超参数.

3.3　优　化

利用经典的反向传播算法来训练异构哈希网络模型，反向传播算法通过利用随机梯度下降算法优化式(8)来实现. 其中 $\mathcal{L}_s$ 关于参数 Θ 和输入特征 $\boldsymbol{r}_{i,j}^{K}$ ($K\in\{I,V\}$)的梯度可以较为容易地计算出来，许多深度学习的开源工具，如 Caffe[25] 等，已经提供了该层的前向和 $\boldsymbol{h}$ 反向计算过程，这里不再赘述. $\mathcal{L}_d$ 对 $\boldsymbol{\mu}_i$ 和 $\boldsymbol{r}_{i,j}^{K}$ 的梯度为

$$\begin{aligned}\frac{\partial\mathcal{L}_d}{\partial\boldsymbol{\mu}_i}&=\frac{1}{n}\sum_{K\in\{I,V\}}\sum_{j=1}^{n_i^K}(\boldsymbol{\mu}_i-\boldsymbol{r}_{i,j}^{K})-\frac{1}{c}(\boldsymbol{\mu}_i-\boldsymbol{\mu}),\\ \frac{\partial\mathcal{L}_d}{\partial\boldsymbol{r}_{i,j}^{K}}&=\frac{1}{n}(\boldsymbol{r}_{i,j}^{K}-\boldsymbol{\mu}_i)+2\boldsymbol{r}_{i,j}^{K}\end{aligned} \tag{9}$$

$\mathcal{L}_t$ 对输入 $(\boldsymbol{h},\tilde{\boldsymbol{h}},\hat{\boldsymbol{h}})$ 的梯度可通过如下公式计算：

$$\begin{aligned}\frac{\partial t}{\partial\boldsymbol{h}}&=\frac{1}{2}(\hat{\boldsymbol{h}}-\tilde{\boldsymbol{h}})\times\boldsymbol{I}\{A\},\\ \frac{\partial t}{\partial\tilde{\boldsymbol{h}}}&=-\frac{1}{2}\boldsymbol{h}\times\boldsymbol{I}\{A\},\\ \frac{\partial t}{\partial\hat{\boldsymbol{h}}}&=\frac{1}{2}\boldsymbol{h}\times\boldsymbol{I}\{A\},\\ A&\triangleq d(\boldsymbol{h},\hat{\boldsymbol{h}})-d(\boldsymbol{h},\tilde{\boldsymbol{h}})+\zeta>0,\end{aligned} \tag{10}$$

其中，$\boldsymbol{I}\{\cdot\}$ 是一个指示函数，当其内部的条件为真时返回 1，反之则返回 0. 根据式(10)可以计算式(7)所示的损失 $\mathcal{L}_t$ 对 $\boldsymbol{r}_{i,j}^{K}$ 的梯度.

如图 1 所示，由于 $\mathcal{L}_d$ 和 $\mathcal{L}_s$ 是添加在公共空间的特征表示上的损失，$\mathcal{L}_t$ 是添加在最终的二值哈希表示上的损失，所以采用先优化 $\mathcal{L}_d$ 和 $\mathcal{L}_s$ 再联合优化三者的训练策略，即整个优化方法分为两个步骤：预训练与微调.

在预训练步骤中，对异构哈希网络中除哈希函数以外的部分，即图像分支和视频分支进行优化，以便为下一步网络整体微调提供一个较好的初值. 预训练步骤之后的异构哈希网络可以提供两个有效的特征表示生成器 $\chi^I(\cdot)+\eta^I(\cdot)$ 和 $\chi^V(\cdot)+\eta^V(\cdot)$，它们不仅能够分别为人脸图像和人脸视频数据提取合适的特征表示，而且能够将位于异构空间中的数据映射到一个具有较强判别力的公共空间中，从而实现异构数据的特征表示同构化. 由于在此过程中，只有 $\chi^I(\cdot)$、$\eta^I(\cdot)$、$\chi^V(\cdot)$ 和 $\eta^V(\cdot)$ 是可学习的，而 $\phi(\cdot)$ 是固定不动的，因此目标函数中只利用 $\mathcal{L}_d$ 和 $\mathcal{L}_s$ 即可. 在预训练过程中使用了 softmax 和 Fisher 损失，所以利用预训练之后的异构哈希网络计算得到的人脸图像与人脸视频的特征表示是可分且具有足够判别力的，这就显著减少了同一个体不

同模态数据间的差异.

微调步骤利用预训练步骤获得的初始值对整个异构哈希网络进行优化,这一步骤的目的是将图像分支、视频分支与哈希函数融合至一个统一的优化框架中,使网络的各组成部分能够通力协作、相互促进,且都能很好地为最终的检索任务服务.图像与视频分支对原始数据处理的情况能够影响后面哈希函数的性能,反过来,哈希函数的性能也能够指导这两个分支的学习.因此,微调过程在预训练过程的基础之上,进一步减少同一个体不同模态数据间的差异,从而实现跨模态人脸检索任务.

算法 1 总结了异构哈希网络的训练过程,其中,$\Phi=\{\chi^I(\cdot),\eta^I(\cdot),\chi^V(\cdot),\eta^V(\cdot)\}$表示图像与视频分支的所有参数,$\Gamma=\Phi\cup\{\phi(\cdot)\}$表示整个网络的参数,$\gamma$ 和 λ 是网络更新时的学习率.

算法 1. 异构哈希网络训练算法.

输入:人脸图像数据 $\boldsymbol{X}^I$、人脸图像数据 $\boldsymbol{X}^V$ 以及它们相对应的标签信息、超参数 α 和 β

输出:HHN 的所有参数 Γ

步骤 1. 预训练.

$t\leftarrow 0$.

随机初始化网络和损失层的参数:$\Psi^t=\Phi^t\cup\Theta^t\cup\Lambda^t$.

迭代.

前向过程:计算 Fisher 和 softmax 的联合损失 $\mathcal{L}=\alpha\mathcal{L}_d+\beta\mathcal{L}_s$.

反向过程:根据式(9)逐层计算联合损失对输入的梯度以及对网络和损失层参数的梯度 $\partial\mathcal{L}/\partial\Psi^t$.

更新过程:$\Psi^t=\Psi^{t+1}-\gamma^t(\partial\mathcal{L}/\partial\Psi^t)$.

$t\leftarrow t+1$.

至收敛

步骤 2. 微调.

$s\leftarrow 0$.

根据步骤 1 结果初始化 Ψ,并随机初始化哈希函数 $\varphi(\cdot)$,即 $\Omega^s=\Psi^t\cup\{\phi^s(\cdot)\}$.

迭代:

前向过程:计算 Fisher、softmax 和三元排序的联合损失 $\mathcal{L}=\alpha\mathcal{L}_d+\beta\mathcal{L}_s+\mathcal{L}_t$.

后向过程:根据式(9)和式(10)逐层计算联合损失对输入的梯度以及对网络和损失层参数的梯度 $\partial\mathcal{L}/\partial\Omega^s$.

更新过程:$\Omega^{s+1}=\Omega^s-\lambda^s(\partial\mathcal{L}/\partial\Omega^s)$.

$s\leftarrow s+1$.

至收敛

4 实 验

4.1 数据集与实验设置

为了评价所提出方法的性能,我们在 ICT-TV[26] 和 Celebrity-1000[27] 两个公共数据集上进行了实验.ICT-TV 中的人脸视频片段均收集于两部美剧的第一季:《生活大爆炸》(the Big Bang Theory,BBT)和《越狱》(Prison Break,PB).BBT 数据集共包含了 14 个人的共 4458 段人脸视频片段,每个人的视频片段数量最小为 11,最大为 1528.PB 数据集共包含了 19 个人的共 7500 段人脸视频片段,每个人的视频片段数量最小为 49,最大为 1965.两部美剧所提供的人脸视频片段都是经过标注仅包含人脸区域,且以逐帧方式存储的每帧大小为 150×150 像素的人脸图像.ICT-TV 数据集是根据美剧建立的,其中包含的人物数量较少.与 ICT-TV 数据集不同,Celebrity-1000(Celeb1K)是一个大规模的人脸视频数据集,包含 1000 个人的 159726 个人脸视频片段,共含有大约 240 万帧数据(每个人脸视频片段约包含 15 帧),这些人脸视频片段都是在真实场景中拍摄的.为了消除真实场景中的背景干扰,人脸视频的所有帧都经过了人脸检测和人脸对齐的操作,并归一化为大小为 64×48 像素的人脸灰度图像.

由于数据集中每个人的视频数量相差较大,且有些视频帧数较少,所以在 BBT、PB、Celeb1K 数据集上都分别选择一个子集作为实验数据.具体来讲,先根据数据集中每个个体包含人脸视频的数量由大到小对个体进行排序,之后选取含有视频较多的个体的数据用来进行实验,其中 BBT 数据集选取视频数最多的 5 个个体,PB 数据集选取视频数最多的 7 个个体,Celeb1K 选取视频数最多的 15 个个体.对于每个个体,从其视频中随机选取 200 个作为训练数据.由于多模态检索方法在训练时需要提供各个模态的数据,所以,对于每个个体选取出的 200 个视频片段,每个视频片段中随机选取一帧作为训练图像,即每个个体提供 200 个训练视频和 200 个训练图像.因此,BBT 数据集的训练数据为 1000 张人脸图像和 1000 个人脸视频,PB 数据集的训练数据为 1400 张人脸图像和 1400 个人脸视频,Celeb1K 数据集的训练数据为 3000 张人脸图像和 3000 个人脸视频.训练集选取之后,将数据子集中剩余的人脸视频作为人脸视频测试集,仍然从每个视频中随机选取一帧组成人脸图像测试集.针对不同的检索任务,测试集中对查询集和待查询集的划分略有不同.对于"基于图像的人脸视频检索"任务,从每个个体的测试人脸图像中随机选取 5 个作为查询集,整个人脸视频测试集作为待查询集;对于"基于视频的人脸图像检索"任务,从每个个体的测试人脸视频中随

机选取 5 个作为查询集，整个人脸图像测试集作为待查询集.

4.2　实现细节

为了与其它哈希方法进行公平的比较，本文方法与对比方法针对人脸图像和视频使用相同的特征表示. 具体来讲，人脸图像用一个 p 维的向量表示，这个向量是由原始图像的像素特征经过主成分分析(Principal Component Analysis，PCA)获得的，称为灰度 PCA 特征；对于一个人脸视频片段，将该视频片段看成各帧数据的无序集合，首先对每一帧提取和人脸图像一样的 p 维灰度 PCA 特征，然后利用这些帧的灰度 PCA 特征计算一个大小为 $p\times p$ 的 RBF 核矩阵特征 $\boldsymbol{K}$，称为灰度 RBF 核矩阵特征：

$$\begin{aligned}&\boldsymbol{K}(i,j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2}{2\sigma^2}\right),\\&\sigma=\frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\end{aligned}\tag{11}$$

其中 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 分别表示矩阵 $\boldsymbol{X}\in\mathbb{R}^{p\times q}$ 的第 i 和第 j 行，$\boldsymbol{X}$ 是一个含有 q 帧的人脸视频的各帧 p 维灰度 PCA 特征矩阵，$\boldsymbol{K}(i,j)$ 表示矩阵 $\boldsymbol{K}$ 的第 i 行第 j 列的元素.

在本文方法的实现中，图像分支由全连接层和 ReLU 激活层相互堆叠构成，该分支的大小为“100－512－1024－100”，这里的数字代表输入层和全连接层的所含有的神经元的个数，每个全连接层后面跟随着一个 ReLU 激活层；视频分支首先使用一个向量化操作层将核矩阵转化为向量，之后是全连接层和 ReLU 激活层的堆叠，因此视频分支的大小为“100×100－5050－100－512－1024－100”. 向量化操作层[28]执行的操作是

$$\begin{aligned}vec(\boldsymbol{K})=&[\boldsymbol{V}_{1,1},\sqrt{2}\boldsymbol{V}_{1,2},\cdots,\sqrt{2}\boldsymbol{V}_{1,p},\boldsymbol{V}_{2,2},\sqrt{2}\boldsymbol{V}_{2,3},\cdots,\\&\sqrt{2}\boldsymbol{V}_{2,p},\cdots,\sqrt{2}\boldsymbol{V}_{p-1,p},\boldsymbol{V}_{p,p}]^{\mathrm{T}}\end{aligned}\tag{12}$$

通过这样的方式，向量化操作层输出 $p(p+1)/2$ 维的向量，其中 $\boldsymbol{V}=\log(\boldsymbol{K})$ 是 RBF 核矩阵 $\boldsymbol{K}$ 的矩阵对数. 哈希函数也是由全连接层和 ReLU 激活层组成，大小为“100－100－l”，每个全连接层后面都接着一个 ReLU 激活层，其中 l 表示哈希编码的长度. 最后一个全连接层之后，添加一个双曲正切层来近似 sgn(•)符号函数来实现哈希码的量化. 以上的网络结构是在 ICT-TV 数据集上(包括 BBT 和 PB)的. Celebrity 数据集包含的个体更多，数据集更复杂，因此网络的结构也设计得更大，图像分支大小为“100－256－512－512－1024－2048－100”，视频分支大小为“100×100－5050－100－256－512－512－1024－2048－100”，哈希函数的大小为“100－100－l”.

预处理与调优过程都是在 Caffe 框架[25]下实现的. 预训练过程通过随机梯度下降方法进行，其中 momentum 和 weight decay 分别设置为 0.9 和 5×10^{-4}，学习率初始化为 0.01 并按照多项式规则(polynomial policy)进行递减，递减的参数 power value 设置为 0.8. 训练集的批次大小设置为 256，迭代的总次数设置为 10^5. 微调过程中的 momentum 和 weight decay 分别设置为 0.8 和 5×10^{-5}，学习率初始化为 0.001 并按照多项式规则进行递减，递减的参数 power value 设置为 0.8，全部的迭代次数为 5×10^4.

4.3　对比方法和评价指标

我们与 7 个单模态哈希方法和 6 个多模态哈希方法进行了对比：

(1) 单模态哈希方法：LSH[6]、SH[7]、ITQ[8]、SITQ[8]、RR[8]、SSH[9]和 KSH[10]；

(2) 多模态哈希方法：CCA[8]、PDH[13]、CMSSH[14]、MMNN[16]、SCM[18]和 HER[5].

为实现公平的比较，我们的方法和所有对比方法都采用相同的数据集划分策略和相同的数据特征表示，即人脸图像的 100 维灰度 PCA 特征和人脸视频的大小为 100×100 的灰度 RBF 核矩阵特征. 由于一些对比的哈希方法仅能生成长度比输入特征小的哈希编码，因此，在实验中，哈希码的长度分别设置为 8、16、32 和 64 来显示哈希方法性能随哈希编码位数不同的变化情况. 为了公平的比较，所有对比方法中的重要参数都是根据原始参考文献中的建议设置的.

我们使用准确率-召回率(Precision Recall，PR)曲线、准确率-优先返回样本数量(Precision curve w. r. t. Number of top returned samples，PN)曲线、召回率-优先返回样本数量(Recall curve w. r. t. Number of top returned samples，RN)曲线和平均正确率均值(mean Average Precision，mAP)进行评价. 同时，也列出了 mAP 对哈希码位数的变化曲线以探究不同哈希方法中哈希码长对检索性能的影响. 实验中报告的各性能指标均为 300 次测试的平均值. 受篇幅所限，我们仅给出哈希码长为 64 位时的 PR、PN 和 RN 曲线.

4.4　实验结果与分析

严格来讲，单模态哈希方法并不能直接应用于跨模态的人脸检索任务. 为了使单模态方法可以适

用于跨模态检索任务，我们将人脸视频片段看成是各帧数据的无序集合，然后，将一张人脸图像与一个人脸视频片段的距离用该图像与视频中每一帧的距离的平均值来表示，从而实现跨图像与视频模态的检索任务.对于多模态的哈希方法，由于除 HER 以外的方法都只能处理表示为欧式空间中的向量的数据，因此利用式(12)先将人脸视频的灰度 RBF 核矩阵特征向量化，然后再执行这些方法.

为了充分评价本文方法在跨模态人脸检索任务上的性能，我们进行了两方面的实验：基于图像的人脸视频检索和基于视频的人脸图像检索.对于基于图像的人脸视频检索任务，表 1 和表 2 分别列出了本文方法与对比方法在 ICT-TV 数据集和 Celeb1K 数据集上的 mAP 比较；对于基于视频的图像检索任务，表 3 和表 4 分别列出了本文方法与对比方法在 ICT-TV 数据集和 Celeb1K 数据集上的 mAP 比较.图 2、图 3 和图 4 分别显示了这些方法在 ICT-TV 的 BBT 数据集、ICT-TV 的 PB 数据集以及 Celeb1K 数据集上两种跨模态检索任务的 PR、PN、RN 以及 mAP 曲线的对比.在这 3 个图中，左栏为“基于图像的人脸视频检索”任务的相应曲线，右栏为“基于视频的人脸图像检索”任务的相应曲线.

表 1 ICT-TV 数据集上本文方法与其它方法在基于图像的人脸视频检索任务上的 mAP 比较

数据集	类型	方法	8 位	16 位	32 位	64 位
BBT	单模态哈希方法	LSH	0.2448	0.2919	0.3162	0.3726
		RR	0.2916	0.3551	0.3571	0.3929
		ITQ	0.3546	0.3599	0.3898	0.3998
		SH	0.2881	0.3421	0.3145	0.3177
		SSH	0.3432	0.2761	0.2541	0.2447
		KSH	0.7825	0.7988	0.8638	0.8262
		SITQ	0.4703	0.3818	0.3837	0.3379
	多模态哈希方法	CCA	0.3664	0.3067	0.2763	0.2440
		CMSSH	0.5582	0.5985	0.5794	0.5489
		PDH	0.2252	0.2223	0.2817	0.2669
		MMNN	0.6322	0.7296	0.7860	0.7819
		SCM	0.7659	0.8461	0.8690	0.8605
		HER	0.8017	0.8121	0.8677	0.8673
		HHN	**0.8218**	**0.8541**	**0.9051**	**0.9112**
PB	单模态哈希方法	LSH	0.1764	0.1728	0.1868	0.1857
		RR	0.1689	0.1786	0.1872	0.1896
		ITQ	0.1746	0.1728	0.1794	0.1860
		SH	0.1685	0.1752	0.1903	0.2028
		SSH	0.1838	0.1840	0.1799	0.1823
		KSH	0.3420	0.4006	0.4260	0.4414
		SITQ	0.1893	0.1818	0.1824	0.1831
	多模态哈希方法	CCA	0.2193	0.2363	0.2198	0.1982
		CMSSH	0.2567	0.2397	0.2162	0.2078
		PDH	0.2316	0.2318	0.2131	0.2142
		MMNN	0.4076	0.5522	0.5328	0.7126
		SCM	0.4946	0.6165	0.6247	0.6927
		HER	0.4601	0.5666	0.6123	0.6268
		HHN	**0.7193**	**0.7608**	**0.7956**	**0.8162**

表 2 Celeb1K 数据集上本文方法与其它方法在基于图像的人脸视频检索任务上的 mAP 比较

数据集	类型	方法	8 位	16 位	32 位	64 位
Celeb1K	单模态哈希方法	LSH	0.0912	0.1130	0.1545	0.1915
		RR	0.1156	0.1447	0.1904	0.2094
		ITQ	0.1169	0.1686	0.2054	0.2317
		SH	0.1276	0.1648	0.1935	0.2063
		SSH	0.1342	0.1765	0.1823	0.1568
		KSH	0.2796	0.3468	0.3841	0.4383
		SITQ	0.1541	0.1937	0.2280	0.2341
	多模态哈希方法	CCA	0.1646	0.1950	0.1814	0.1470
		CMSSH	0.1738	0.1652	0.1500	0.1402
		PDH	0.1256	0.1264	0.1191	0.1229
		MMNN	0.1735	0.3656	0.3918	0.3672
		SCM	0.2654	0.3370	0.4157	0.4809
		HER	0.7324	0.7740	0.7659	0.7970
		HHN	**0.7705**	**0.7831**	**0.8028**	**0.8254**

表 3 ICT-TV 数据集上本文方法与其它方法在基于视频的人脸图像检索任务上的 mAP 比较

数据集	类型	方法	8 位	16 位	32 位	64 位
BBT	单模态哈希方法	LSH	0.2300	0.2969	0.3410	0.3426
		RR	0.3013	0.3537	0.3723	0.3865
		ITQ	0.3213	0.3524	0.3779	0.3986
		SH	0.2728	0.3336	0.3120	0.3098
		SSH	0.3266	0.2809	0.2571	0.2510
		KSH	0.7296	0.7852	0.8459	0.8023
		SITQ	0.4451	0.3787	0.3895	0.3600
	多模态哈希方法	CCA	0.3688	0.3057	0.2739	0.2423
		CMSSH	0.5905	0.6109	0.6096	0.5834
		PDH	0.2298	0.2278	0.2271	0.2517
		MMNN	0.6694	0.7552	0.8193	0.8256
		SCM	0.8079	0.8814	0.8966	0.8907
		HER	0.6567	0.7214	0.7354	0.7888
		HHN	**0.9080**	**0.9274**	**0.9364**	**0.9457**
PB	单模态哈希方法	LSH	0.1641	0.1817	0.1899	0.1850
		RR	0.1776	0.1728	0.1876	0.1934
		ITQ	0.1824	0.1745	0.1779	0.1869
		SH	0.1699	0.1762	0.1898	0.2014
		SSH	0.1900	0.1859	0.1832	0.1853
		KSH	0.3337	0.3920	0.4246	0.4309
		SITQ	0.1951	0.1847	0.1810	0.1847
	多模态哈希方法	CCA	0.2184	0.2375	0.2203	0.1977
		CMSSH	0.2084	0.1946	0.1817	0.1693
		PDH	0.2290	0.2292	0.2083	0.2076
		MMNN	0.3980	0.5522	0.5500	0.7067
		SCM	0.5195	0.6400	0.6458	0.7110
		HER	0.4551	0.5773	0.6580	0.6947
		HHN	**0.7265**	**0.7787**	**0.8040**	**0.8160**

表 4 Celeb1K 数据集上本文方法与其它方法在基于视频的人脸图像检索任务上的 mAP 比较

数据集	类型	方法	8 位	16 位	32 位	64 位
Celeb1K	单模态哈希方法	LSH	0.0768	0.1096	0.1486	0.1946
		RR	0.0989	0.1559	0.1813	0.2160
		ITQ	0.1130	0.1632	0.2028	0.2302
		SH	0.1193	0.1645	0.1965	0.2093
		SSH	0.1196	0.1761	0.1897	0.1638
		KSH	0.2496	0.3366	0.3773	0.4484
		SITQ	0.1270	0.1827	0.2225	0.2318
	多模态哈希方法	CCA	0.1634	0.1917	0.1801	0.1458
		CMSSH	0.1725	0.1601	0.1395	0.1269
		PDH	0.1193	0.1213	0.1150	0.1159
		MMNN	0.1851	0.3818	0.4018	0.3878
		SCM	0.2666	0.3351	0.4094	0.4733
		HER	0.5909	0.7260	0.7658	0.8177
		HHN	**0.7612**	**0.7786**	**0.7766**	**0.8323**

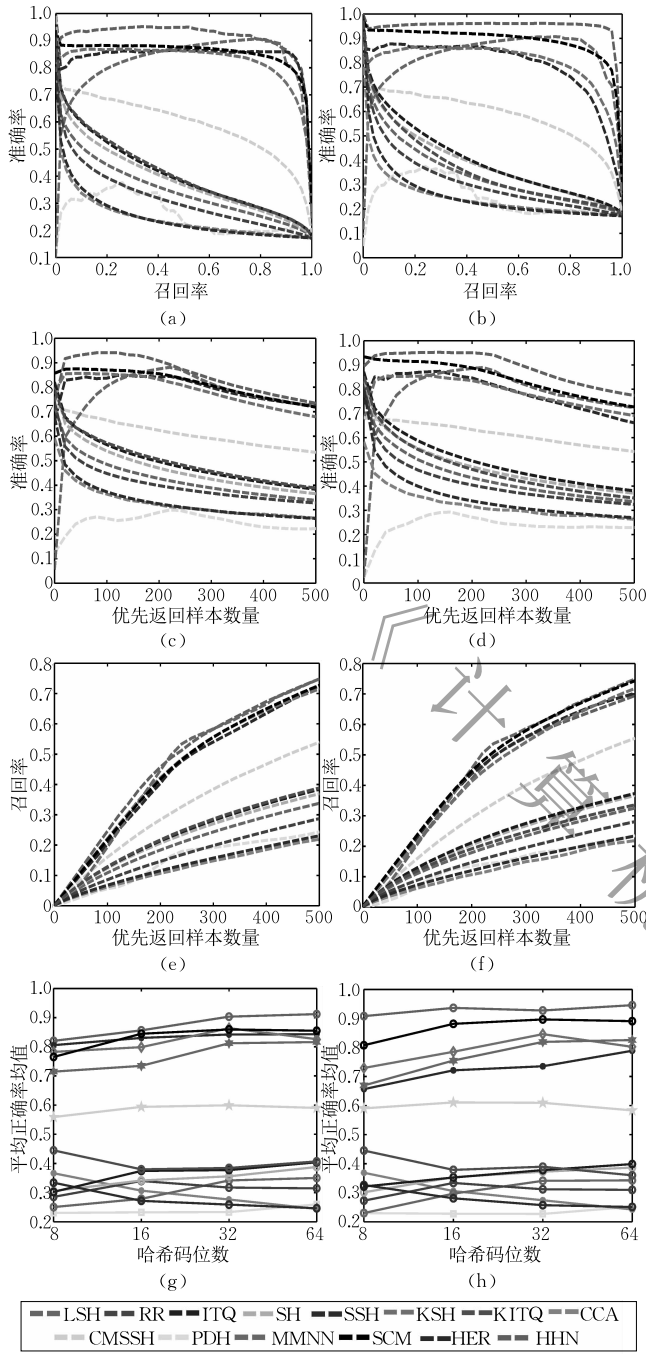

图 2　BBT 数据集上本方法与其它方法在两类跨模态人脸检索任务上的曲线对比图(左栏为基于图像的人脸视频检索的相应曲线,右栏为基于视频的人脸图像检索的相应曲线)

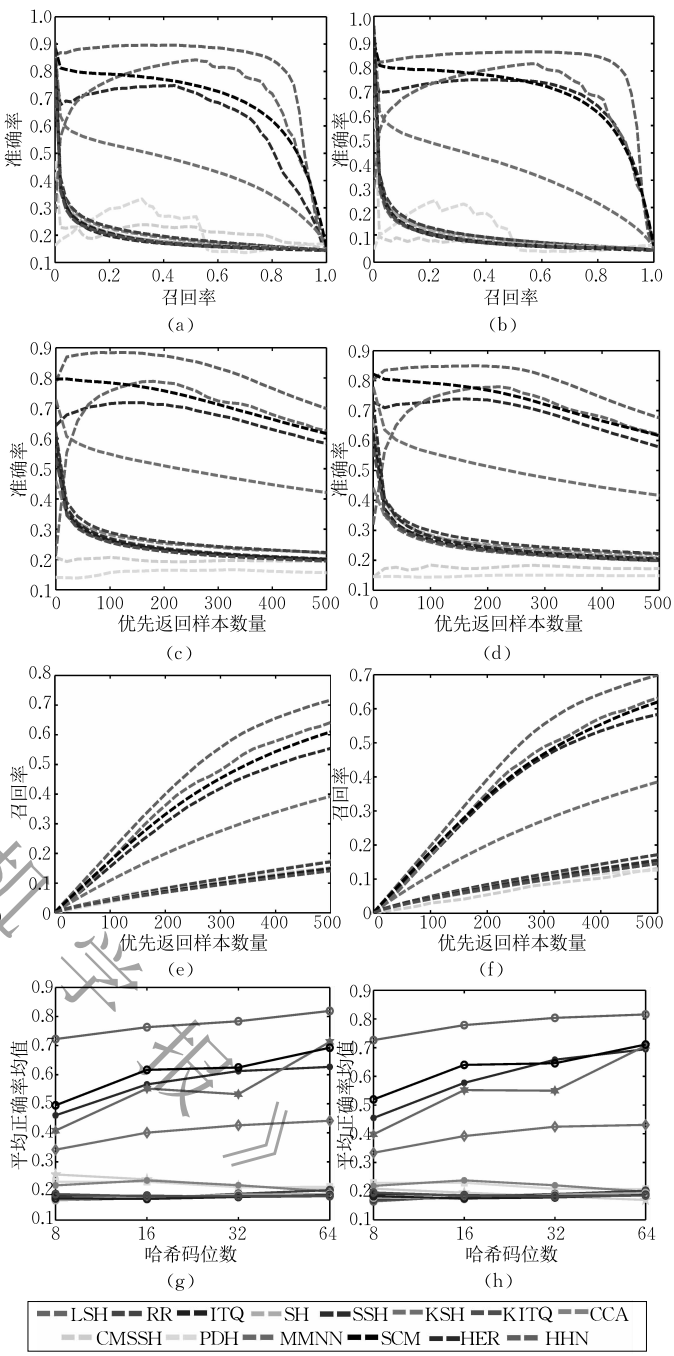

图 3　PB 数据集上本方法与其它方法在两类跨模态人脸检索任务上的曲线对比图(左栏为基于图像的人脸视频检索的相应曲线,右栏为基于视频的人脸图像检索的相应曲线)

可以看出,本文提出的跨模态人脸检索方法在多数情况下性能都显著优于其它的多模态哈希方法,原因有以下几方面:异构哈希网络优化过程中同时考虑了模态内与模态间的判别性;由于同时使用了 Fisher 损失、softmax 损失和三元排序三种损失,无论是公共空间的特征表示还是最终学得的二值哈希表示都具备较强的判别力;采用了两阶段的训练方式,这使得网络可以首先将数据从异构的空间中投影到具有较强判别力的公共空间,然后在公共空间上学习哈希函数,使得各阶段目的明确,便于提升特征表示的判别力;本文提出的异构哈希网络对原始数据及从原空间到二值空间的映射没有做任何先验性的假设,全部是基于数据学习而来,而 HER 方法的核函数可能在某些数据上并不适用.

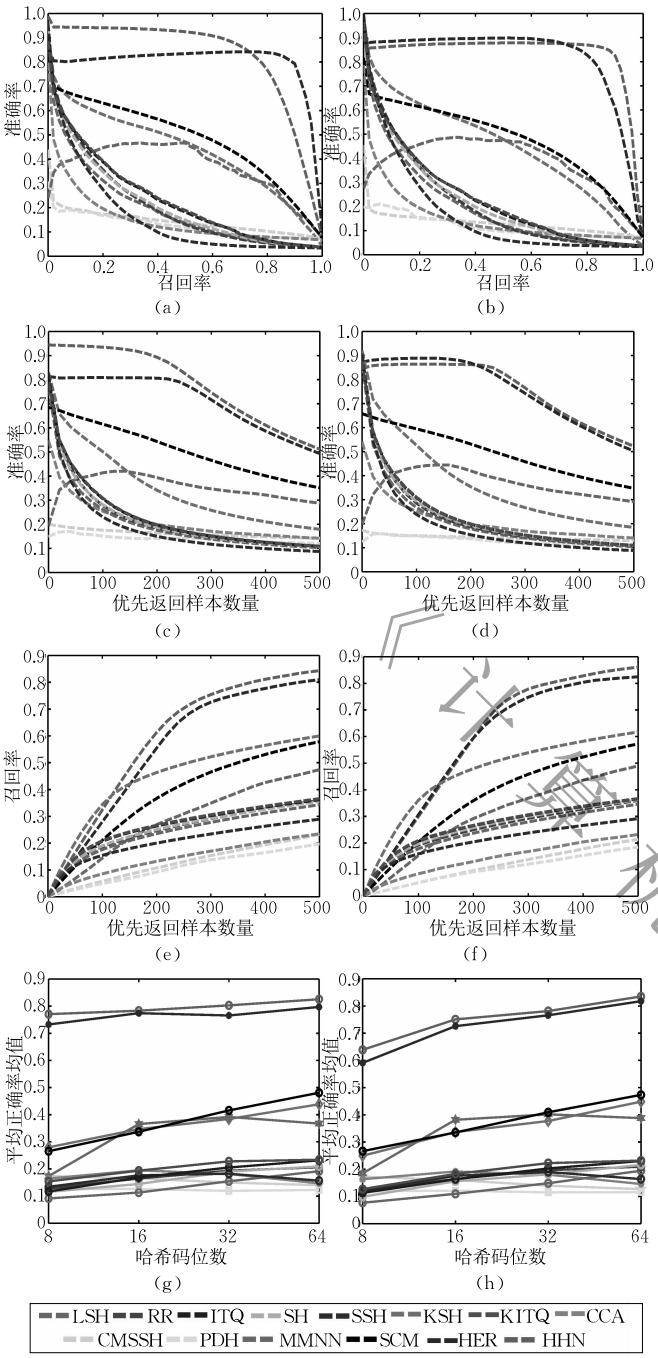

图 4 Celeb1K 数据集上本方法与其它方法在跨模态人脸检索任务上的曲线对比图(左栏为基于图像的人脸视频检索的相应曲线,右栏为基于视频的人脸图像检索的相应曲线)

从实验结果中还可以发现,多模态方法要好于单模态方法,这主要是因为多模态方法利用了灰度 RBF 核矩阵特征描述符来表示人脸视频,该描述符可以充分利用视频各帧之间的相关关系来建模人脸视频中包含的诸如光照、姿态等的复杂变化.此外,在单模态哈希方法中,LSH、RR、ITQ 和 SH 方法属于无监督的哈希方法,SSH、KSH、SITQ 属于半监督或有监督的哈希方法;在多模态哈希方法中,PDH 是无监督方法,CCA、CMSSH、MMNN、SCM、HER 和本文的 HHN 方法是有监督方法.可以发现,有监督和半监督哈希方法在大多数情况下会好于非监督的哈希方法,主要原因是有监督和半监督哈希方法在训练过程中充分利用了标签信息,而标签信息是对检索系统进行评价时唯一可依赖的信息.

4.5 三元排序损失函数的作用

为了验证三元排序损失函数的作用,我们对比了使用三元排序损失函数所得到的哈希编码(HHN)与直接量化得到的哈希编码(HHN-2)在跨模态人脸检索任务上的性能.表 5 和表 6 分别列出了三元损失函数在基于图像的人脸视频检索任务上和基于视频的人脸图像检索任务上对 mAP 的影响.可以看出,通过三元排序损失函数所优化得到的哈希编码比直接量化得到的哈希编码取得了更好的检索结果,说明了三元排序损失函数对于检索的重要作用.

表 5 三元损失函数在基于图像的人脸视频检索任务上对 mAP 的影响

数据集	方法	8 位	16 位	32 位	64 位
BBT	HHN-2	0.6438	0.6412	0.6394	0.6435
	HHN	**0.8218**	**0.8541**	**0.9051**	**0.9112**
PB	HHN-2	0.7022	0.6904	0.7001	0.6939
	HHN	**0.7193**	**0.7608**	**0.7956**	**0.8162**
Celeb1K	HHN-2	0.6909	0.7005	0.6971	0.6961
	HHN	**0.7705**	**0.7831**	**0.8028**	**0.8254**

表 6 三元损失函数在基于视频的人脸图像检索任务上对 mAP 的影响

数据集	方法	8 位	16 位	32 位	64 位
BBT	HHN-2	0.6674	0.7210	0.7486	0.7776
	HHN	**0.9080**	**0.9274**	**0.9364**	**0.9457**
PB	HHN-2	0.6910	0.7210	0.7486	0.7776
	HHN	**0.7265**	**0.7787**	**0.8040**	**0.8160**
Celeb1K	HHN-2	0.7021	0.7398	0.7545	0.7754
	HHN	**0.7612**	**0.7786**	**0.7766**	**0.8323**

4.6 检索效率

在本实验中,我们在 Celeb1K 数据集上对比了我们的方法与其它对比方法的检索效率.由于所有的方法都使用相同的特征,因此没有计算特征提取的时间,而只计算了检索的时间.表 7 列出了各个方

表 7 检索效率 (单位:s)

方法	检索时间	方法	检索时间
LSH	2.2863e-06	CCA	1.0901e-06
RR	4.0581e-06	CMSSH	2.1930e-06
ITQ	3.7152e-06	PDH	2.1930e-06
SH	1.0528e-04	MMNN	7.6754e-06
SSH	3.5437e-06	SCM	1.0965e-06
KSH	0.0154	HER	0.0433
SITQ	9.8881e-06	**HHN**	**5.9125e-06**

法的平均检索时间.可以看出,KSH,HER 以及 SH 所需的检索时间较长,而我们的方法与其它的方法具有相似的检索效率.

5 总　结

本文提出了一种使用异构哈希网络的跨模态人脸检索方法.异构哈希网络能够将位于异构空间的人脸图像和视频映射到一个公共且有判别力的二值空间上,获得相应的哈希表示,实现异构数据的特征表示同构化.网络的训练同时使用了 Fisher 损失,softmax 损失和三元排序损失,其中的 Fisher 损失使用了类内散度和类间散度的差分形式,并保持均值向量的可学习性,以方便使用基于批次的优化方法.Fisher 损失关注于公共空间的判别力,softmax 损失强调公共空间上特征表示的可分性,三元排序损失旨在提升最终的检索性能,三者相互补充,共同作用于网络训练.异构哈希网络为基于深度学习的跨模态哈希方法提供了一套通用的解决框架,不仅可以完成跨图像和视频模态的检索任务,还可以容易地扩展到其他跨模态的检索任务上.

参　考　文　献

[1] Sun Y, Wang X, Tang X. Sparsifying neural network connections for face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4856-4864

[2] Taigman Y, Yang M, Ranzato M A, Wolf L. Web-scale training for face identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2746-2754

[3] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[4] Yang J, Ren P, Chen D, et al. Neural aggregation network for video face recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5216-5225

[5] Li Y, Wang R, Huang Z, et al. Face video retrieval with imagequery via hashing across Euclidean space and Riemannian manifold//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4758-4767

[6] Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing//Proceedings of the IEEE Conference on Very Large Data Bases. Edinburgh, UK, 1999: 518-529

[7] Weiss Y, Torralba A, Fergus R. Spectral hashing//Proceedings of the Advances in neural information processing systems. Vancouver, Canada, 2009: 1753-1760

[8] Gong Y, Lazebnik S. Iterative quantization: A procrustean approach to learning binary codes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 817-824

[9] Wang J, Kumar S, Chang S-F. Semi-supervised hashing for scalable image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3424-3431

[10] Liu W, Wang J, Ji R, et al. Supervised hashing with kernels//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2074-2081

[11] Feng F, Wang X, Li R. Cross-modal retrieval with correspondence autoencoder//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 7-16

[12] 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

[13] Rastegari M, Choi J, Fakhraei S, et al. Predictable dual-view hashing//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 1328-1336

[14] Bronstein M M, Bronstein A M, Michel F, Paragios N. 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

[15] Zhen Y, Yeung D-Y. A probabilistic model for multimodal hash function learning//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 940-948

[16] Masci J, Bronstein M M, Bronstein A M, Schmidhuber J. Multimodal similarity-preserving hashing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36: 824-830

[17] Zhai D, Chang H, Zhen Y, et al. Parametric local multimodal hashing for cross-view similarity search//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2754-2760

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

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

[20] Liu X, He J, Deng C, Lang B. Collaborative hashing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2139-2146

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

[22] Cao Y, Long M, Wang J, et al. Deep visual-semantic hashing for cross-modal retrieval//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1445-1454

[23] Yang M, Zhang L, Feng X, Zhang D. Fisher discrimination dictionary learning for sparse representation//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 543-550

[24] Norouzi M, Fleet D J, Salakhutdinov R R. Hamming distance metric learning//Proceedings of the Advances in neural information processing systems. Lake Tahoe, USA, 2012: 1061-1069

[25] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding//Proceedings of the ACM International Conference on Multimedia. Orlando, USA, 2014: 675-678

[26] Li Y, Wang R, Shan S, Chen X. Hierarchical hybrid statistic based video binary code and its application to face retrieval in TV-series//Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Ljubljana, Slovenia, 2015: 1-8

[27] Liu L, Zhang L, Liu H, Yan S. Toward large-population face identification in unconstrained videos. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24: 1874-1884

[28] Tuzel O, Porikli F, Meer P. Pedestrian detection via classification on Riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30: 1713-1727

DONG Zhen, born in 1990, Ph. D. His research interests include computer vision, face recognition and retrieval.

PEI Ming-Tao, born in 1977, Ph. D., associate professor, Ph. D. supervisor. His research interests include computer vision and face recognition.

Background

Cross-modality face retrieval is to retrieve faces of a particular person in one modality given his/her face information in another modality, such as retrieving video shots containing particular person given one image of him/her (query-by-image video retrieval), or retrieving the face images of one person by using his/her video clip as query (query-by-video image retrieval). It is an important problem in computer vision.

The current methods all concentrate on face video retrieval, and only one previous work is about face retrieval across image and video domains as far as we know. Li et al. proposed a hashing method across the Euclidean space and the Riemannian manifold to measure the similarity of face images and videos for cross-domain face retrieval. Although they achieve superior performances than many traditional single-modality and multi-modality methods, the extracted features (DCT and covariance matrix) and the learned hash functions are incompatible due to the independent feature extraction and hash learning procedures, which might influence the effects of the generated hash codes. In contrast, our heterogeneous hashing network integrates the feature extraction and hash learning into a unified optimization framework where two procedures interact for the final retrieval task. Experiments on a large-scale face video dataset and two challenging TV-series datasets demonstrate the effectiveness of the proposed method.

This work was supported by the National Natural Science Foundation of China (No. 61472038).