基于层次特征融合哈希的近似重复视频检索方法

聂秀山¹, 林培光¹, 杨明哲², 尹义龙^{2*}

1. 山东财经大学计算机科学与技术学院, 济南 250100

2. 山东大学软件学院, 济南 250101

* 通信作者. E-mail: ylyin@sdu.edu.cn

收稿日期: 2018-08-16; 接受日期: 2018-10-09; 网络出版日期: 2018-12-12

国家自然科学基金项目 (批准号: 61671274, 61876098, 61573219)、山东省重点研究与开发项目 (批准号: 2017CXGC1504) 和山东省高校优势学科人才团队培育计划资助项目

摘要 近年来, 由于互联网上视频数量的爆炸式增长, 基于视频哈希 (Hash) 的近似重复视频检索已经吸引了越来越多的研究者关注. 在现有方法中, 视频的视觉特征, 包括单一特征以及多特征融合的方法被广泛应用于近似重复视频检索算法中. 而低层视觉特征在表达高层语义方面存在不足, 使得近似重复视频检索的性能变低. 针对这一问题, 本文提出了一种基于层次特征融合的视频哈希方法, 用于近似重复视频检索. 该方法首先从视频中提取低层人工定义特征, 然后利用卷积神经网络提取中间层深度特征以及高层语义特征, 最后把不同层级的特征融合起来, 利用层次特征和样本之间的全局结构关系以及各特征之间的互补性, 学习得到视频哈希, 进而进行近似重复视频检索. 该方法在 CC-WEB-VIDEO 数据库上进行了实验, 实验结果证明本文方法与现有的方法相比在性能上有较大提升.

关键词 近似重复视频检索, 视频哈希, 层次特征, 特征融合, 监督学习

1 引言

随着视频应用的普及以及各种视频网站上视频数量的不断增加, 网上的重复或近似重复视频的数量也呈爆炸式增长. 这些不合理视频的存在使得用户在搜索相关视频时会花费更多的时间, 严重影响用户体验. 同时, 这一问题也给现有的一些应用领域, 例如视频推荐、版权保护等领域带来了严峻的挑战. 考虑到这些问题, 近似重复视频检索 (near-duplicate video retrieval, NDVR) 的研究变得越来越重要.

在当前的大数据时代, 大规模的近似重复视频检索面临着两项严峻挑战: (1) 存储消耗. 尽管近些年计算机硬件以及信息技术已经取得了空前的发展与进步, 数据的爆炸式增长仍然带来了严峻的存储成本问题; (2) 检索速度. 与存储消耗相比, 从规模庞大的数据中检索特定的视频将会是更加具有挑战

引用格式: 聂秀山, 林培光, 杨明哲, 等. 基于层次特征融合哈希的近似重复视频检索方法. 中国科学: 信息科学, 2018, 48: 1697–1708, doi: 10.1360/N112018-00136
Nie X S, Lin P G, Yang M Z, et al. Hierarchical feature fusion hashing for near-duplicate video retrieval (in Chinese). Sci Sin Inform, 2018, 48: 1697–1708, doi: 10.1360/N112018-00136

图 1 (网络版彩图) 低层视觉特征与真实语义的差异性图示

Figure 1 (Color online) Illustration of difference between low-level visual feature and semantic meaning

性的任务. 因此, 无论从检索精度还是计算复杂度上来讲, 实际需求与现有的解决方法之间都存在着巨大的差距.

从现有研究来看, 视频哈希 (Hash) 方法^[1~3]是解决大规模近似重复视频检索的一个有效的方案. 视频哈希的主要目标是把视频表示为低维的紧凑二进制码, 进而可以直接进行位运算, 使得之前十分耗时的检索过程变得快速、高效. 通过二进制表示的手段, 原有检索系统中, 因视频数据较多导致的存储消耗问题能够被有效地解决, 同时近似重复视频检索的过程能够被大大加快. 由于这些优势的存在, 基于视频哈希的大规模近似重复视频检索的方法受到了越来越广泛的关注^[4~6].

近似重复视频大多经历了一些不影响视频内容的简单修改, 例如对原视频添加噪声、模糊化、仿射变换等都可能得到近似重复视频^[6], 而这些编辑和修改手段大多和低层 (相对高层语义而言) 视觉特征相关, 因此, 现有的方法中, 大多利用人工定义的视觉特征来构造视频哈希. 视觉特征是指用户可以从视频中提取到的最常用的特征, 例如直方图特征、颜色空间特征^[7]、尺度不变特征变换 (scale invariant feature transform, SIFT)^[8], 以及局部二值模式 (local binary pattern, LBP)^[9]等.

一般来说, 使用单一特征来表示视频内容是不够充分的, 并且单一特征表示的方法在进行近似重复视频检索时往往缺乏鲁棒性. 因此, 研究者们提出基于多特征融合的思想用来解决这些问题. 多特征融合哈希 (multiple feature hashing, MFH)^[1]和随机多视角哈希 (stochastic multiview hashing, SMVH)^[2]这两种最新的基于哈希的近似重复视频检索方法, 它们都利用了多特征融合的思想.

然而, 在某些情况下, 无论是单一特征的方法, 还是多个视觉特征融合的方法都有其不足之处, 这主要是由于图像存在着“语义鸿沟”问题. 如图 1 所示的情况, 由于某些原因, 这两张图片的灰度直方

图(低层视觉特征)表示非常相似,但是,在语义层面上,这又是两张完全不同的图片,这就是所谓的“语义鸿沟”问题.通过从原始数据中提取一组语义特征的方式^[10,11],可以有效缓解这种局限性.因此,为了解决此类的问题,结合高层语义特征来构造视频哈希非常重要.

基于此研究动机,在本文研究中,把低层人工定义特征、深度网络中间层特征和高层语义特征定义为不同层级的特征,然后设计了一种基于层次特征融合的哈希方法来充分发掘不同层次特征所携带的互补性信息.在这个模型中,HSV (hue, saturation, value) 颜色直方图特征和局部二值模式特征被提取出来作为低层视觉特征,其他低层次视觉特征,例如尺度不变特征变换、光流场等特征同样也可以使用.另外两种特征,即中间层语义特征以及高层语义特征,分别来自于由 ImageNet 数据库预训练得到的卷积神经网络的隐藏层和输出层.然后,来自这3个层次的不同类型的特征被联合起来协作学习得到视频最后的哈希码.在哈希学习过程中,不同层次特征之间的互补性能被有效地利用起来.

总体来说,本文工作的主要贡献可以分为以下3个方面:

(1) 结合不同层级特征之间的关系,提出了一种基于层次特征融合的哈希学习方法用于近似重复视频检索,这种方法能够有效地利用低层视觉特征与高层语义特征之间的互补性,进而在一定程度上解决视频检索中遇到的语义鸿沟的问题.

(2) 在哈希学习过程中,本文方法把每个样本的每个特征作为个体同等对待,通过最小化具有相同标记个体之间的距离和最大化具有不同标记的个体之间的距离,充分利用了所有样本的多层次特征的全局性非成对 (non-pairwise) 关系,与传统使用成对关系的方法相比,提高了方法的可扩展性和检索效率.另外,方法利用了广义特征值分解高效地解决了具体优化的问题.

(3) 通过在公开视频数据库 CC-WEB-VIDEO 上的实验证明,本文提出的方法能够在使用较短的哈希码的前提下得到比目前已有方法更高的检索精度.

2 相关工作

根据国际视频检索任务评测组织 (TRECVID) 对近似重复视频的定义,近似重复视频一般经历了某些不影响语义内容的修改,例如对原视频进行添加噪声、模糊化、仿射变换等,都可能形成近似重复视频^[5,6],这些修改导致视频在视觉相似性和视觉特征上发生变化,因此,现有很多方法是基于视觉特征,例如直方图特征^[5]、颜色空间^[7]、尺度不变特征变换^[8]、快速稳健特征点^[12],以及局部二值模式^[9]等.但是,单一特征表示的方法在进行近似重复视频检测时缺乏鲁棒性.因此,基于多特征融合的方法被提出用来解决这一问题^[1,2].多特征哈希^[1]以及随机多角度哈希^[2]这两种最新的基于哈希的近重复视频检索方法都是基于多特征融合的策略.此外,Jiang等^[13]利用时域金字塔匹配结构融合多特征,构建视频拷贝检测系统;Nie等^[14,15]提出了利用时空关系图和球环结构融合视频特征进行近似重复视频检测的方法.

这些方法虽然都充分利用了视频的多特征信息,但这些特征大多是基于人工定义的视觉特征,正如本文在第1节所说,利用人工定义视觉特征会产生例如图1所示的“语义鸿沟”问题,因此需要寻找语义特征进行融合来解决此类问题.本文正是基于这一动机,提出利用低层视觉特征、中间层和高层语义特征多层次特征融合的思想来构建视频哈希用于近似重复视频检测.另一方面,多层次的特征可以看作对视频不同角度的认知,而多角度认知信息的融合,显然对于提升近似重复视频检测的精度有较好的促进作用.

图 2 (网络版彩图) 本文算法框架示意图
Figure 2 (Color online) Framework of the proposed method

3 层次特征融合哈希方法

本节将对所提出的方法做详细的介绍, 方法具体流程可以分为两个阶段: 特征提取阶段和基于不同层次特征融合的哈希学习阶段. 具体方法流程图如图 2 所示.

3.1 多层次特征提取

在本文方法中, 共用到了 3 种不同层次的特征: 低层人工定义视觉特征、中间层语义特征和高层语义特征.

低层特征. 本文方法所涉及的低层特征主要是人工定义的视觉特征. 根据美国国家标准技术研究所等机构组织的视频检索评测 (TRECVID) 任务的定义, 近似重复视频一般指的是某视频片段经历了旋转、平移、放缩等几何变换, 或者是经历了随机噪声、字幕添加、画中画等操作后得到的视频, 显然这些操作是在保持视频语义内容不变的情况下, 针对视频视觉感知的一些修改, 因此低层的视觉特征在近似重复视频检测中具有重要的作用, 这也是近年来的关于近似重复视频检索比较注重低层视觉特征的原因. 在低层特征中, 考虑到视频的局部视觉特征信息以及全局视觉特征信息, 本文提取 HSV 颜色直方图特征和局部二值模式特征作为低层特征. 具体实施方面, 对于每一个视频, 本文从视频关键帧中分别提取 HSV 和 LBP 特征来描述视频的内容. HSV 特征是一种具有 162 维的全局性特征, 而 LBP 特征是一种具有 256 维的局部性特征. 为了减小计算复杂度, 本文采用 average pool 的思想把一个视频片段的所有关键帧特征整合为一个低层视觉特征. 需要说明的是, 本文只是使用了 HSV, LBP 等特征作为算法的一种实现, 视频的其他特征同样适用.

中间层语义特征. “语义鸿沟”是图像和视频理解中的一个经典问题, 本文利用中间层语义特征在一定程度上缓解这一问题, 所谓中间层特征, 主要是一种由低层视觉特征到视频类别标签的过渡信息, 在中间层特征中, 即隐含了低层视觉特征的信息, 也在一定程度上隐含了视觉特征和视频类别标签之间的映射关系.

为更好地体现中间层语义特征, 本文利用卷积神经网络 (convolutional neural network, CNN) 模型

来提取中间层语义特征. CNN 模型作为深度学习的主要模型之一,近年来在语音识别、图像理解、动作识别等计算机视觉领域显示了较强的能力. 根据计算机视觉方向的相关研究成果^[16,17],深度 CNN 特征往往包含更多关于目标物体的语义信息,而这些语义信息对于近似重复视频检索任务解决“语义鸿沟”问题十分重要. 在一个特定的网络结构中,中间隐藏层的激活神经元比其他任何层的神经元都要携带更加丰富和复杂的信息. 因此,这些神经元的特征在表示输入对象时表现更好.

在深度模型的诸多应用中,数据库 ImageNet 上的图像分类任务是最为经典的应用,因此,本文方法的中间层语义特征来自于在 ImageNet 上预训练 CNN 模型的隐藏层,本文采用在 ImageNet 上预训练模型的主要原因有两个: (1) 使用预训练的模型可以充分地利用模型在大型图像数据库上训练所获取的对样本的强大区分能力; (2) 成熟模型的隐藏层能够体现低层特征到高层特征的过渡信息. 在具体实施中,本文方法取深度模型的全连接层输出作为中间层语义特性,它在一定程度上也代表了低层和高层语义信息之间的桥梁.

高层语义特征. 在本文方法中,高层语义特征指的是深度卷积神经网络模型最终的输出向量,是针对模型输入得到属于某一标记分类的预测概率. 本文方法采用的深度学习模型是一个由大量网络图片训练得到的语义分类器. 所以,从语义的角度来看,可以利用深度卷积神经网络模型的输出来表示对其训练数据语义信息的区分,并且利用语义分类器来减轻低层特征的限制. 例如在图 1 中的图像可以很容易地通过语义标签来区分.

由以上描述可知,本文方法中的中间层和高层语义特征均来自于深度学习模型,众所周知,近年来,相关研究者提出了很多深度卷积神经网络模型,如 VGGNet^[18], AlexNet^[19] 和 GoogLeNet^[20] 等,这些深度卷积神经网络模型在计算机视觉领域都取得很好的应用. 本文方法采用 VGG 模型提取深层特征,其他的深度学习模型也可以应用于本文方法. 在本文方法中, VGG 模型的输入是视频关键帧. 为了得到一个视频片段的完整特征表示,本文方法采用了一个简单的平均合并思想,把单个视频所有关键帧的特征描述整合为一个特征来代表整个视频. 这种处理方式虽然在一定程度上损失了视频帧间的关系,但是提升了方法的运行效率. 另外,对于近似重复视频检测任务来说,丢帧或增加帧也是一类常见的编辑手段,因此,采用关键帧平均的思想也在一定程度上也增加了算法对此类编辑手段的鲁棒性.

本文采用的 VGG 框架共包含 5 个卷积层 (卷积 1~5) 和 3 个全连接层 (全连接 6~8),每个卷积层的详细描述可以参考文献 [19]. 在训练过程中,关键帧的原始颜色通道 (RGB) 信息被重新设置为 224×224 像素大小的格式输入到 VGG 网络结构中,并且在此网络中依次前向传播经过 5 个卷积层和 3 个全连接层. 在该模型中,全连接层和网络的预测输出层分别获得了 4096 维的特征向量和 1000 维的特征向量. 本文方法把这两层得到的特征分别称为中间层语义特征和高层语义特征. 加上另外两种低层人工定义特征,即 HSV 和 LBP 特征,一共获得了 3 层中的 4 种不同类型的特征. 显然,这些不同层次的特征能够从不同的角度更加合理地表示每个视频. 本文方法通过把这些不同层次的特征融合起来,利用层次融合之间的互补性,学习得到更可靠的视频哈希表示.

3.2 视频哈希学习

显然,不同层次的特征可以看作是从不同角度表示原始视频内容. 受到多视角判别性分析方法^[21]思想的启发,本文从全局性的角度利用不同层次的特征来学习得到视频对应的哈希码.

在现有的多视角哈希方案中,例如,在研究工作 [1,2] 中,涉及到两种被广泛应用的映射策略. 一种是最小化内部视角之间的差异,另一种是对任意两个视角 (特征) 之间的成对相关性的建模. 简单来说,第 1 种类型没有利用视角之间的差异,而第 2 种思路又缺乏对不同特征之间非成对关系的利用.

图 3 (网络版彩图) 类内散度和类间散度的示意图

Figure 3 (Color online) Illustration of within-class and between-class variation

因此, 在本文提出的方法中, 把每个视频的每一个特征向量看作一个个体, 也就是说, 假设数据集中 n 个视频样本, 每个样本有 M 个特征表示, 则可以得到 $N = M \times n$ 个个体. 然后从全局性考虑使用所有个体多个维度的相关性 (非成对的) 来学习得到哈希码.

在本文方法中, C 表示类别标记的数量, M 表示特征数量, N_k^m 表示第 k ($1 \leq k \leq C$) 个类别下第 m ($1 \leq m \leq M$) 个特征的样本个数, 本文方法中 $M = 3$ (把 HSV 和 LBP 当作一个低层视觉特征). 在 Hamming 空间中定义两种距离散度, 分别称为类间散度 B 和类内散度 S , 表示如下:

$$B = \sum_{k=1}^C (\bar{y}_k - \hat{y}_k)(\bar{y}_k - \hat{y}_k)^T, \quad (1)$$

$$S = \sum_{k=1}^C \sum_{m=1}^M \sum_{i=1}^{N_k^m} (y_{ki}^m - \hat{y}_{ki}^m)(y_{ki}^m - \hat{y}_{ki}^m)^T, \quad (2)$$

其中, $\bar{y}_k = \frac{1}{N_k} \sum_{m=1}^M \sum_{i=1}^{N_k^m} y_{ki}^m$ 表示标记为 k 的样本个体的平均哈希, $\hat{y}_k = \frac{C \times \bar{y} - \bar{y}_k}{C-1}$, 表示除了标记为 k 的样本个体以外的样本个体平均哈希, y_{ki}^m 表示标记为 k 且在第 m 个特征表示下的第 i 个样本个体哈希, $\hat{y}_{ki}^m$ 表示除 y_{ki}^m 以外的标记为 k 的样本个体的平均哈希. 图 3 展示了类间散度 B 和类内散度 S , 一些相关的符号也在图 3 中给出.

显然, 本文提出的哈希学习的目标是最大化类间散度矩阵 B 的同时最小化类内散度矩阵 S . 通常来说, 矩阵的行列式是解决此问题的一个简单方法, 一般来说, 行列式特征值的计算结果是行列式主方向上差异的结果^[22]. 因此, 本文利用行列式来进行散度量, 哈希学习的优化问题可以定义如下:

$$\max_{\mathbf{Y}} \frac{\det(\mathbf{B})}{\det(\mathbf{S})} \quad \text{s.t. } \mathbf{Y} \in \{0, 1\}^L, \quad (3)$$

其中 $\mathbf{Y}$ 代表对应视频的哈希码, L 为哈希码长度.

在哈希学习中, 因为哈希码是二值离散的, 因此, 优化问题 (3) 变成一个 NP-hard 问题. 为了放松离散性的约束, 本文定义了一个哈希函数进行放松操作, 通过放松操作可以把当前的优化问题 (3) 通过广义特征值分解的方式来解决, 对于得到连续最优解, 通过阈值再获取对应的哈希码.

把视频 $\mathbf{X}$ 的第 m 个特征矩阵记作 $\mathbf{X}^m$, 定义一个哈希函数 $H(\mathbf{X}^m) = \mathbf{W}_m^T \mathbf{X}^m$, 即每层特征对应一种映射方法. 利用连续解来放松离散性约束, 然后整个优化问题就通过广义特征值分解的方式来解

决. 本文方法与常见线性判别分析 (LDA) 方法不同的是, 本文提出的方法能够为不同层次的特征同时实现映射变换. 将哈希函数 $H(\mathbf{X}^m)$ 代入式 (1), 可以得到关于 $\mathbf{W}$ 的类间散度 $\mathbf{B}$ 表达式如下所示:

$$\begin{aligned}
 \mathbf{B} &= \sum_{k=1}^C (\bar{\mathbf{y}}_k - \hat{\mathbf{y}}_k) (\bar{\mathbf{y}}_k - \hat{\mathbf{y}}_k)^T \\
 &= \sum_{k=1}^C (\bar{\mathbf{y}}_k \bar{\mathbf{y}}_k^T - \bar{\mathbf{y}}_k \hat{\mathbf{y}}_k^T - \hat{\mathbf{y}}_k \bar{\mathbf{y}}_k^T + \hat{\mathbf{y}}_k \hat{\mathbf{y}}_k^T) \\
 &= \sum_{k=1}^C \left(1 + \frac{2}{C-1} + \frac{1}{C-1^2}\right) \bar{\mathbf{y}}_k \bar{\mathbf{y}}_k^T - \sum_{k=1}^C \left(\frac{C}{C-1} + \frac{C}{(C-1)^2}\right) \bar{\mathbf{y}}_k \bar{\mathbf{y}}^T \\
 &\quad - \sum_{k=1}^C \left(\frac{C}{C-1} + \frac{C}{(C-1)^2}\right) \bar{\mathbf{y}} \bar{\mathbf{y}}_k^T + \sum_{k=1}^C \left(\frac{C}{C-1}\right)^2 \bar{\mathbf{y}} \bar{\mathbf{y}}^T \\
 &= \sum_{k=1}^C \left(\frac{1}{C-1} + 1\right)^2 \bar{\mathbf{y}}_k \bar{\mathbf{y}}_k^T - \left(\frac{2C^2}{C-1} + \frac{2C^2 - C^3}{(C-1)^2}\right) \bar{\mathbf{y}} \bar{\mathbf{y}}^T \\
 &= \left(\frac{1}{C-1} + 1\right)^2 \sum_{k=1}^C \left(\frac{1}{N_k} \sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)}\right) \left(\frac{1}{N_k} \sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)}\right)^T \\
 &\quad - \left(\frac{2C^2}{C-1} + \frac{2C^2 - C^3}{(C-1)^2}\right) \left(\frac{1}{N} \sum_{m=1}^M \sum_{k=1}^C \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)}\right) \left(\frac{1}{N} \sum_{m=1}^M \sum_{k=1}^C \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)}\right)^T \\
 &= \sum_{m=1}^M \sum_{\hat{m}=1}^M \mathbf{W}_m^T \mathbf{B}_{m\hat{m}} \mathbf{W}_{\hat{m}} \\
 &= [\mathbf{W}_1^T, \mathbf{W}_2^T, \mathbf{W}_3^T] \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \mathbf{B}_{23} \\ \mathbf{B}_{31} & \mathbf{B}_{32} & \mathbf{B}_{33} \end{pmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \mathbf{W}_3 \end{bmatrix}, \quad (4)
 \end{aligned}$$

其中, $\mathbf{B}_{m\hat{m}}$ 是一个分块矩阵. 具体定义如下:

$$\mathbf{B}_{m\hat{m}} = \left(\frac{1}{C-1} + 1\right)^2 \sum_{k=1}^C \frac{N_k^m N_k^{\hat{m}}}{(N_k)^2} \bar{\mathbf{x}}_k^m \bar{\mathbf{x}}_k^{\hat{m}T} - \frac{C^3}{(C-1)^2 N^2} \left(\sum_{k=1}^C N_k^m \bar{\mathbf{x}}_k^m\right) \left(\sum_{k=1}^C N_k^{\hat{m}} \bar{\mathbf{x}}_k^{\hat{m}T}\right)^T, \quad (5)$$

其中, $\bar{x}_k^m$ 表示标签为 k 的第 m 个特征表示下的所有样本的均值, N_k 表示第 k 个类别下的个体总数. 同样的, 类内散度矩阵 $\mathbf{S}$ 可以表示为

$$\begin{aligned}
 \mathbf{S} &= \sum_{k=1}^C \sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} (\mathbf{y}_{ki}^m - \hat{\mathbf{y}}_{ki}) (\mathbf{y}_{ki}^m - \hat{\mathbf{y}}_{ki})^T \\
 &= \sum_{k=1}^C \sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} (\mathbf{y}_{ki}^m (\mathbf{y}_{ki}^m)^T - \mathbf{y}_{ki}^m \hat{\mathbf{y}}_{ki}^T - \hat{\mathbf{y}}_{ki} (\mathbf{y}_{ki}^m)^T + \hat{\mathbf{y}}_{ki} \hat{\mathbf{y}}_{ki}^T) \\
 &= \sum_{k=1}^C \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \left(\frac{1}{N_k - 1} + 1\right)^2 \mathbf{y}_{ki}^{(m)} \mathbf{y}_{ki}^{(m)T}\right)
 \end{aligned}$$

$$\begin{aligned}
 & - \left(\frac{2}{N_k - 1} + \frac{2 - N_k}{(N_k - 1)^2} \right) \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)} \right) \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)} \right)^T \\
 & = \sum_{k=1}^C \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \left(\frac{1}{N_k - 1} + 1 \right)^2 \mathbf{y}_{ki}^{(m)} \mathbf{y}_{ki}^{(m)T} \right. \\
 & \quad \left. - \left(\frac{2}{N_k - 1} + \frac{2 - N_k}{(N_k - 1)^2} \right) \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)} \right) \left(\sum_{m=1}^M \sum_{i=1}^{N_k^{(m)}} \mathbf{y}_{ki}^{(m)} \right)^T \right) \\
 & = \sum_{m=1}^M \mathbf{W}_m^T \left(\sum_{k=1}^C \sum_{i=1}^{N_k^{(m)}} \left(\frac{1}{N_k - 1} + 1 \right)^2 \mathbf{x}_{ki}^{(m)} \mathbf{x}_{ki}^{(m)T} \right) \mathbf{W}_m \\
 & \quad - \sum_{m=1}^M \sum_{\hat{m}=1}^M \mathbf{W}_m^T \left(\left(\frac{2}{N_k - 1} + \frac{2 - N_k}{(N_k - 1)^2} \right) \sum_{k=1}^C N_k^{(m)} N_k^{(\hat{m})} \overline{\mathbf{x}_k^{(m)} \mathbf{x}_k^{(\hat{m})}}^T \right) \mathbf{W}_m \\
 & = \mathbf{W}^T \{ \mathbf{S}_{m\hat{m}} \} \mathbf{W} \\
 & = [\mathbf{W}_1^T, \mathbf{W}_2^T, \mathbf{W}_3^T] \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{pmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \mathbf{W}_3 \end{bmatrix}, \tag{6}
 \end{aligned}$$

其中 $\{\mathbf{S}_{m\hat{m}}\}$ 是一个已知的分块矩阵.

$$\begin{aligned}
 \mathbf{S}_{mm} & = \sum_{k=1}^C \left(\sum_{i=1}^{N_k^m} \frac{N_k^2}{(N_k - 1)^2} \mathbf{x}_{ki}^m (\mathbf{x}_{ki}^m)^T - \frac{N_k (N_k^m)^2}{(N_k - 1)^2} \overline{\mathbf{x}_{ki}^m \mathbf{x}_{ki}^m}^T \right), \\
 \mathbf{S}_{m\hat{m}} & = - \sum_{k=1}^C \frac{N_k N_k^m N_k^{\hat{m}}}{(N_k - 1)^2} \overline{\mathbf{x}_{ki}^m \mathbf{x}_{ki}^{\hat{m}}}^T, \quad m \neq \hat{m}.
 \end{aligned} \tag{7}$$

然后, 优化问题 (3) 可以转换为问题 (8), 此问题可以通过广义特征值分解来解决

$$\arg \max_{\mathbf{W}} \frac{\det(\mathbf{W}^T \{ \mathbf{B}_{m\hat{m}} \} \mathbf{W})}{\det(\mathbf{W}^T \{ \mathbf{S}_{m\hat{m}} \} \mathbf{W})}. \tag{8}$$

最后, 通过 $H(\mathbf{X}) = \mathbf{W}^T \mathbf{X}$ 进行映射并比较 $H(\mathbf{X})$ 中每个维度的值与平均值的大小来确定二进制哈希码的值. 如果某个维度的值比平均值大, 将哈希码置为 1, 否则置为 0. 因此, 每个视频都被表示为一个长度为 L 的二值哈希码向量.

4 实验与结果

4.1 实验设置

本文在近似重复视频检索领域最常用的公开数据库 CC-WEB-VIDEO^[23] 上做了充分的实验, CC-WEB-VIDEO 数据库共包含 12790 个视频片段 (398015 个关键帧). 在实验中, 取 90% 的数据作为训练集, 10% 的数据作为测试集. 相关实验设备的具体参数如下: Intel(R) Core(TM) i7-4790 CPU at 3.60 GHz, 6 GB RAM, AMD Radeon R7 350 GPU.

表 1 各种方法的平均检索精度 (mAP) 结果对比

Table 1 mAP performance comparison

GF	SPH	MFH	SMVH	Proposed-L	Proposed-H	Proposed
0.892	0.826	0.934	0.957	0.962	0.950	0.990

图 4 (网络版彩图) 数据库中 24 类查询的平均检索精度

Figure 4 (Color online) Average precision (AP) of the 24 queries

在近似重复视频检索结果的度量中, 本文采用了常用的精度 – 召回率 (precision-recall, P-R) 曲线和平均准确率 (mean average precision, mAP) 来衡量检索结果的好坏. 精度 (precision) 是指检索出来的正确的视频数占检索出来视频总数的比例, 召回率 (recall) 指的是检索出来的正确的视频数占近似重复视频样本总数的比例.

为了证明所提出方法的性能, 本文将提出的方法分别与基于全局特征的方法 (global feature, GF)、谱哈希 (spectral hashing, SPH)^[22]、多特征融合哈希^[1], 以及随机多视角哈希^[2]进行了性能比较. 另外, 为了证明本文所提出的方法中不同层次特征之间的互助性, 本文同样分别利用单一层次特征进行了实验, 例如, 单独利用了低层特征 (LBP 和 HSV 特征) 进行实验, 方法名称简称为 Proposed-L, 同时单独利用了高层语义特征 (中间层特征和高层语义特征) 进行实验, 方法名称简称为 Proposed-H. 本提出的方法简称为 Proposed. 实验结果见表 1、图 4 和 5.

4.2 实验结果分析

表 1 展示了本文所提出的方法与现有的研究方法的对比, 从这个表中可以看出本文所提出的方法取得了最好的性能. 同时, 还可以看到, 单独使用低层的人工定义特征的性能高于使用高层次的深度特征, 这是和近似视频本身的特性吻合的, 正如 2.1 小节中所述, 近似重复视频之间的差别在于噪声、几何变换等低层的视觉感知的改变, 因此, 反映视觉感知信息的低层特征更加重要. 但是, 当遇到本文

图 5 (网络版彩图) 各种对比方法的 P-R 曲线示意图

Figure 5 (Color online) Precision and Recall performance

第 1 节介绍的一些情形的时候, 高层特征的作用就显现出来了, 因此融合多层次的特征对总体性能的提升是有帮助的. 从表 1 中也可以看出, 无论是单独利用低层特征, 还是单独利用高层语义特征 (包括中间层特征和高层语义特征) 进行实验, 结果都不如本文所提出的融合多层次特征的方法. 这意味着高级语义特征的使用弥补了低层人工定义特征在语义表达上的缺陷, 同时也表明了这些不同层次特征之间的互补性有效提升了近似重复视频检索的性能.

图 4 展示了在数据库 24 类样本中, 本文方法和部分对比方法在每一类上的平均检索精度, 可以看出本文提出的方法在绝大多数类别中的检索效果都是明显优于其他方法. 图 5 展示了各方法 P-R 曲线 (查准率 – 召回率曲线) 之间的对比, 而 P-R 曲线是最常用的评价指标之一, 从图 5 的曲线也可以看出, 本文所提出的方法明显优于其他对比方法.

哈希码的长度同样可以影响实验性能以及检索结果. 本文算法通过多次实验, 把哈希码长度设置为 275, 相关研究表明^[2,3], 哈希码的长度和最终的检索精度并不完全成正向关系, 本文所设定的哈希码长度是精度和效率均衡的结果.

5 结论

本文提出了一种基于多层次特征融合的哈希方法来解决近似重复视频检索问题, 该方法首先提取视频的低层人工定义特征, 然后利用深度卷积神经网络提取视频的中层和高层语义特征, 利用不同层级特征和样本之间的结构和关系, 建立一个目标优化函数, 通过广义特征值分解来得到最优解, 从而获取视频哈希. 与现有方法相比, 本文提出的方法能够充分利用不同层次特征之间的互补性, 弥补了低层人工定义特征在语义表达上存在的缺陷, 最终实现更高效的检索性能. 本文实验也证明了所提出的方法的有效性.

参考文献

- 1 Song J K, Yang Y, Huang Z, et al. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Trans Multimedia*, 2013, 15: 1997–2008

- 2 Hao Y B, Mu T T, Hong R C, et al. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Trans Multimedia*, 2017, 19: 1–14
- 3 Liu H, Zhao Q J, Wang H, et al. An image-based near-duplicate video retrieval and localization using improved Edit distance. *Multimed Tools Appl*, 2017, 76: 24435–24456
- 4 Lv J, Wu B, Yang S, et al. Efficient large scale near-duplicate video detection base on spark. In: *Proceedings of IEEE International Conference on Big Data*, Washington, 2016. 957–962
- 5 Chou C L, Chen H T, Lee S Y. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Trans Multimedia*, 2015, 17: 382–395
- 6 Nie X S, Jing W Z, Ma L Y, et al. Two-layer video fingerprinting strategy for near-duplicate video detection. In: *Proceedings of IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Hong Kong, 2017. 555–560
- 7 Shen H T, Zhou X F, Huang Z, et al. UQLIPS: a real-time near-duplicate video clip detection system. In: *Proceedings of the 33rd International Conference on Very Large Data Bases*, Vienna, 2007. 1374–1377
- 8 Wei S K, Zhao Y, Zhu C, et al. Frame fusion for video copy detection. *IEEE Trans Circ Syst Video Technol*, 2011, 21: 15–28
- 9 Zhao G Y, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intel*, 2007, 29: 915–928
- 10 Kordopatis-Zilos G, Papadopoulos S, Patras I, et al. Near-duplicate video retrieval by aggregating intermediate cnn layers. In: *Proceedings of International Conference on Multimedia Modeling*, Cham, 2017. 251–263
- 11 Cai J J, Merler M, Pankanti S, et al. Heterogeneous semantic level features fusion for action recognition. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Shanghai, 2015. 307–314
- 12 Nie X S, Yin Y L, Sun J D, et al. Comprehensive feature-based robust video fingerprinting using tensor model. *IEEE Trans Multimedia*, 2017, 19: 785–796
- 13 Jiang M L, Tian Y H, Huang T J. Video copy detection using a soft cascade of multimodal features. In: *Proceedings of International Conference on Multimedia and Expo*, Melbourne, 2012. 374–379
- 14 Nie X S, Liu J, Sun J D, et al. Robust video hashing based on representative-dispersive frames. *Sci China Inf Sci*, 2013, 56: 068104
- 15 Nie X S, Chai Y N, Liu J, et al. Spherical torus-based video hashing for near-duplicate video detection. *Sci China Inf Sci*, 2016, 59: 059101
- 16 Lin G S, Shen C H, van den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 3194–3203
- 17 Liu Z W, Li X X, Luo P, et al. Semantic image segmentation via deep parsing network. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015. 1377–1385
- 18 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 19 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, 2012. 1097–1105
- 20 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015
- 21 Kan M N, Shan S G, Zhang H H, et al. Multi-view discriminant analysis. *IEEE Trans Pattern Anal Mach Intel*, 2016, 38: 188–194
- 22 Weiss Y, Torralba A, Fergus R. Spectral hashing. In: *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, 2008. 1753–1760
- 23 Wu X, Hauptmann A G, Ngo C W. Practical elimination of near-duplicates from web video search. In: *Proceedings of the 15th ACM International Conference on Multimedia*, New York, 2007. 218–227

Hierarchical feature fusion hashing for near-duplicate video retrieval

Xiushan NIE¹, Peiguang LIN¹, Mingzhe YANG² & Yilong YIN^{2*}

1. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250100, China;

2. School of Software, Shandong University, Jinan 250101, China

* Corresponding author. E-mail: ylyin@sdu.edu.cn

Abstract In recent years, owing to the rapid growth of the number of videos on the Internet, near-duplicate video retrieval (NDVR) by video hashing has attracted huge attention. In the existing methods, the visual features of videos, including single feature and multiple visual feature fusion, are widely used in the NDVR algorithms. However, low-level visual features have some disadvantages in expressing high-level semantics, which may lead to low performance in NDVR. In this paper, we propose a video hashing method for NDVR based on hierarchical feature fusion to address this issue. In the proposed method, low-level handcrafted features from videos were first extracted; then, the intermediate-level deep features and high-level semantic features extracted from the convolutional neural network are obtained. Finally, these semantic features are combined with low-level visual features, where the global structural relationships and complementarity discovered among the hierarchical features are utilized to learn the hash code for NDVR. Extensive experiments are performed on the CC-WEB-VIDEO dataset; the proposed framework proves to have a better retrieval performance compared with the state-of-the-art approaches.

Keywords near-duplicate video retrieval, video hashing, hierarchical feature, feature fusion, supervised learning

Xiushan NIE was born in 1981. He obtained his Ph.D. degree from Shandong University, Jinan, in 2011. Currently, he is a professor at Shandong University of Finance and Economics. His research interests include multimedia retrieval and indexing, machine learning, and computer vision.

Peiguang LIN was born in 1978. He obtained his Ph.D. degree in computer science from Beijing Institute of Technology, Beijing, in 2006. Currently, he is an associate professor at Shandong University of Finance and Economics. His research interests include information retrieval, NLP, and machine learning.

Mingzhe YANG was born in 1991. He obtained his master's degree from Shandong University, Jinan, in 2018. Currently, he is a researcher at Baidu, Inc. His research interests include multimedia retrieval and computer vision.

Yilong YIN was born in 1972. He obtained his Ph.D. degree from Jilin University, Changchun, in 2000. Currently, he is a professor at Shandong University. His research interests include machine learning, data mining, computational medicine, and biometrics.