doi:10.3772/j.issn.1002-0470.2019.12.005

基于卷积神经网络和频率域特征的视频拷贝检测方法①

石慧杰②

(北京工商大学计算机与信息工程学院 北京 100048)

摘 要 为了解决视频特征鲁棒性差、计算复杂度高等问题,提出一种新的视频拷贝检测方法。该算法将深度卷积网络特征和传统手工特征相结合,提升特征检测的维数,提升检测准确度。方法首先使用密集连接卷积网络(DenseNet)提取关键帧的深度特征,并对关键帧进行离散余弦变换(DCT)提取系数特征,然后使用基于典型相关分析(CCA)的特征融合算法将 2 种特征进行有效融合,最后使用融合特征进行特征匹配。在标准数据集上的实验表明,本文提出的算法检测效果较好,在常见的拷贝变化下可以得到更高的检测精度。该算法可以作为一种有效的数字视频版权保护技术应用于数字视频的监管领域。

关键词 视频拷贝检测;特征表示;卷积神经网络(CNN);典型相关分析(CCA);离散余弦变换(DCT);密集连接卷积网络(DenseNet)

0 引 言

随着科技的发展和互联网的广泛普及,数字视频盗版问题日益突出,根据 Digital TV Research 的统计,我国 2016 年因流媒体盗版造成的经济损失已经高达 42.36 亿美元,给整个数字文化创意产业带来了不可估量的损害,版权保护迫在眉睫。视频拷贝检测技术是一种通过比较待检测视频与原始视频的相似程度,判断检测视频是否构成侵权的方法[1],在视频版权保护和管理中发挥着重要作用。

在早期的视频拷贝检测算法研究中,研究人员主要使用各类传统手工特征进行检测,并取得了良好的实验结果。传统手工特征主要有基于全局特征和基于局部特征 2 种方法。有代表性的全局特征包括颜色直方图、GIST 特征、DCT 系数[2]等,该方法可以检测视频图像颜色等整体变化,适用于需求简单的大规模视频检索,但对于局部的图像偏移、剪切、旋转等拷贝变化检测效果不佳;代表性的局部特征包括 SIFT[3]、SURF[4]、BRIEF[5]、FREAK[6]等,局部特征对多种拷贝攻击具有较高的分辨能力,检测准确率高,但是该特征描述符的维度高,单个关键帧可能产生成百上千个局部特征,计算开销大,提取速度和匹配速度慢。

近年来,随着深度学习方法的引入,在计算机视觉领域引入了大量基于深度学习的方法,并取得了很大的成功[7]。卷积神经网络(convolutional neural network, CNN)在提取图像特征方面表现出强大的能力。卷积神经网络使用多层的神经网络,通过卷积运算提取输入信号的特征,通过池化层在提取的特征基础上进行池化抽象,得到较高层次的特征。Jiang 等人[8]最初尝试使用卷积神经网络特征进行拷贝检测,证明深度特征比现有传统方法更有优势,之后有许多新的深度网络被提出,较著名的有 VGGNet、GoogleNet、密集连接卷积网络(dense convolutional network, DenseNet)[9]等,这些网络具有越来越高的辨别能力。

在视频拷贝检测这一任务上,单一的特征往往

① 国家重点研发计划(2017YFB1401000)资助项目。

② 男,1991 年生,硕士生;研究方向:基于内容的视频检索,机器学习等;联系人,E-mail: shjzly@sina.cn

(收稿日期:2018-12-28)

具有自身的局限性，不能够准确地描述视频内容[10,11]，尤其面对复杂的拷贝变换时效果不理想，因此有关学者提出融合不同的特征来描述视频内容的算法。特征的融合方法可以分为特征级融合、分数级融合、决策级融合 3 类[12]。特征级融合算法对拷贝检测任务性能的提升是很明显的。特征级融合使用最原始的信息，通过对图像内容抽取不同特征并进行优化组合，不仅保留了参与融合的多特征的有效识别信息，而且在一定程度上消除了主观和客观因素造成的冗余信息。特征级融合算法可以分为串行融合、并行融合和基于统计分析的融合算法。串行融合是将 2 个特征直接首尾拼接连接成一个新的特征向量，并在高维空间中进行特征匹配；并行融合方法将 2 个特征向量组合成复矢量，并在复矢量空间中进行匹配。这 2 种融合算法的缺点是简单级联不同特征，导致组成的新特征维度很高，很容易引起“维度灾难”问题。孙权森等人[13]利用典型相关分析（canonical correlation analysis，CCA）进行特征级融合，降低特征的维度并获得了优于串行融合和并行融合的结果。

鉴于深度特征的辨别能力和鲁棒性优势以及离散余弦变换（discrete cosine transform，DCT）系数特征在效率上的优势，本文提出一种基于密集连接卷积网络（DenseNet）和 DCT 系数的视频拷贝检测算法，对查询视频进行关键帧提取并采用 DenseNet 网络进行深度特征提取，再利用 DCT 变换进行 DCT 系数特征提取，通过基于典型相关分析的特征融合算法将 2 种特征进行有效融合，最后使用融合特征进行特征匹配得到最终拷贝检测结果。

1　算法原理

基于关键帧的视频拷贝检测方法主要包含 4 个步骤：关键帧提取、特征提取、特征匹配、时间对齐。图 1 为本文提出的视频拷贝检测方法系统框架。框架中对视频库的建模为离线步骤（图 1 中的离线线路），对查询视频的拷贝检测需要进行在线步骤（图 1中的在线线路）。首先，对视频进行关键帧提取，传统算法通过检测镜头边界的方式提取关键帧，从而减少关键帧数量，避免沉重的内存和运算负担，但是这种稀疏采样的方法丢弃了大量的视频信息，对拷贝检测结果造成极大的影响。本文从准确性角度考虑，使用稠密提取关键帧的方式提取关键帧，然后对提取的视频关键帧分别进行深度特征提取和 DCT 系数特征提取，使用典型相关分析算法将 2 种特征进行特征级融合，使用融合特征进行特征匹配，最后使用时间信息将拷贝帧整合成拷贝视频片段。

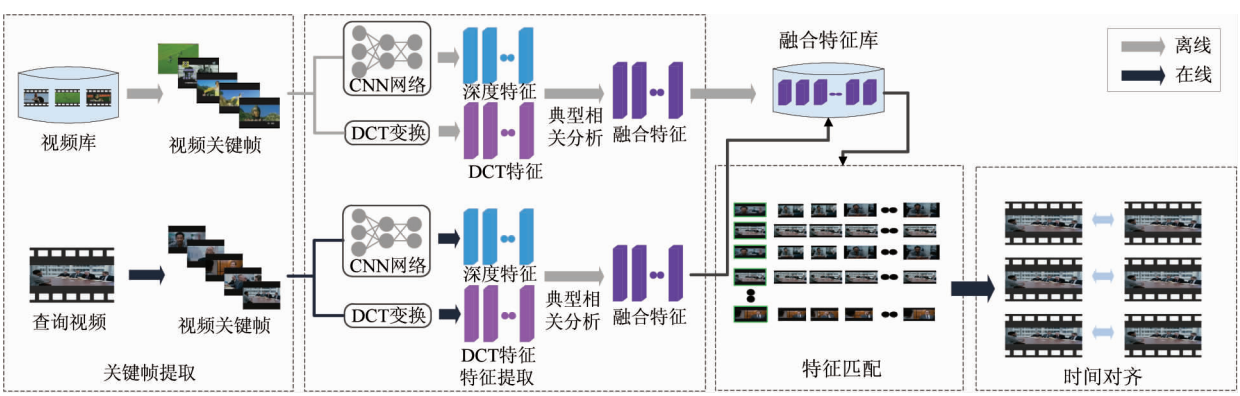

图 1　本文视频拷贝检测方法系统框架

1.1　视频关键帧深度特征提取

卷积神经网络（CNN）是一种前馈神经网络，可用于处理具有类似于网格结构的数据的网络，例如图像数据。卷积神经网络由生物学上的接受域提出。接受域主要指听觉系统，本体感受系统和视觉系统中神经元的某些特性。例如，在视觉神经系统中，神经元的接受域指的是视网膜上的特定区域，并且只有该区域中的刺激才能激活神经元。卷积神经网络有 3 个结构上的特性，即局部连接、权重共享以及空间或时间上的次采样，这些特性使得卷积神经

网络具有一定程度上的平移、缩放和扭曲不变性[14]。

一个前馈神经网络可以被认为是一系列函数的组合,定义如下:

$$f(x) = f_l(\cdots f_2(f_1(x;w_1);w_2)\cdots;w_l) \tag{1}$$

每一个函数f_l都有一个基准输入x_l和一个参数w_l,其输出为x_{l+1},参数$w = (w_1, w_2, \cdots, w_l)$是从目标问题中学习得来的。卷积神经网络中的数据输入x_1, x_2, $\cdots$, x_n通常为图像等网格数据。每个$\boldsymbol{x}_i$都是$M \times N \times K$的矩阵,像素为$M \times N$,通道数为K。设向量$\boldsymbol{y}$为卷积输出,过滤器是3维的,特征图(feature map)是K通道的,若有K'个过滤器,则经过卷积运算产生K'维的特征图$\boldsymbol{y}$:

$$\boldsymbol{y}_{i'j'k'} = \sum_{ijk} w_{ijk} x_{i+i',\, j+j',\, k+k'} \tag{2}$$

引入激活函数来增加网络的非线性,即增加整个网络的表达能力,否则若干线性操作层的堆叠仍然只能起到线性映射的作用,无法形成复杂的函数。最常用的激活函数是Relu(rectified linear unit)函数:

$$\boldsymbol{y}_{ijk} = \max(0, x_{ijk}) \tag{3}$$

同时,在网络中引入池化层(pooling layer),池化操作使用特定位置的相邻输出的整体统计特征来替换该位置处的网络的输出。减少卷积特征尺寸,防止数据过度拟合,同时保持特征的一些不变性(旋转,平移,缩放等)。常用的池化操作有最大池化(max pooling)和平均池化(average pooling):

$$\boldsymbol{y}_{ijk} = \max\{\boldsymbol{y}_{i'j'k'}: i \leqslant i' < i + p, j \leqslant j' < j + p\} \tag{4}$$

$$\boldsymbol{y}_{i'j'k'} = \frac{1}{p \times p} \sum_{ijk} w_{ijk} x_{i+i',\, j+j',\, k+k'} \tag{5}$$

式中,p为池化操作时邻域的大小。

Huang等人[9]提出DenseNet网络并成为国际计算机视觉与模式识别会议2017年上最佳论文。DenseNet是一个将网络层进行密集连接的卷积神经网络,在该网络中,网络的每一层的输入是所有先前层的输出的并集,并且由该层学习的特征映射作为输入直接传输到所有后续层,图2是DenseNet的网络结构示意图。

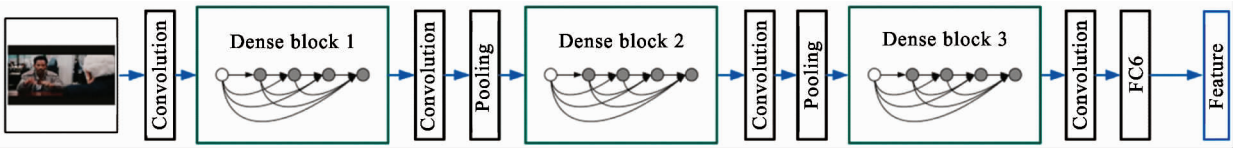

图2 DenseNet网络结构图

与其他卷积神经网络相比,DenseNet的每一层只需学习很少的特征,其参数量和计算量显著减少。并且综合利用浅层复杂度低的特征,得到一个光滑的具有更好泛化性能的决策函数。参照文献[8]的设置,本文选择了DenseNet-169网络的FC6层的输出作为视频关键帧的深度特征并进行了实验。使用文献[9]与Caffe预训练的模型对采样得到的视频关键帧进行深度特征的提取,得到每个图像的特征是1 000维。

1.2 视频关键帧DCT特征提取

离散余弦变换(DCT)将图像数据从空间域表示转换到频率域表示,通过DCT变换,图像的能量可以得到集中,图像的频域图像划分为低频、中频、高频,低频分量集中在矩阵的左上角,它是能量信息最集中的区域,反映图像慢变化,即图像整体部分;高频分量主要则集中在矩阵的右下角,代表图像跳变的位置,即图像细节如轮廓边缘等。根据心理视觉冗余原理,人的视觉系统对图像的高频成分没有低频成分敏感,可以使用DCT系数的低频分量信息来提取图像特征。文献[2]提出顺序测度算法证明DCT系数特征对缩放、直方图量化、模糊、高斯噪声等信号攻击有较好的鲁棒性,同时该算法还有计算复杂度低、特征提取速度快的优点。

DCT特征提取主要有以下几步:(1)将视频帧缩放到64×64像素,然后将图像转换到YUV空间,并只保留Y通道。(2)将得到的图像均分成64块(编号为0~63),每一块为8×8像素,并在每一块上进行2-D DCT变换,得到一个8×8的系数矩阵。

(3)按照 Zig-zag 顺序,计算每个子块的前 4 个子带的系数,将每个子带的 DCT 系数相加作为该子带的系数。(4)计算相邻图像子块对应子带系数大小,形成一个 256 维的 DCT 特征 D_{256},计算公式如下:

$$d_{i,j}=\begin{cases}1, & e_{i,j}\geqslant e_{i,(j+1)\%64},\ 0\leqslant i\leqslant 3, 0\leqslant j\leqslant 63\\ 0, & \text{其他}\end{cases} \tag{6}$$

$$\boldsymbol{D}_{256}=<d_{0,0},\cdots,d_{0,63},\cdots,d_{3,0},\cdots,d_{3,63}> \tag{7}$$

式中,$e_{i,j}$ 表示第 j 个图像子块的第 i 个子带系数。

1.3　视频关键帧特征融合

典型相关分析(CCA)是用于处理 2 组特征向量之间相互依赖关系的统计方法,该方法将相关性研究转化为几对不相关变量之间的相关性研究。给定2 个零均值的特征向量 $\boldsymbol{x}\in R^{p\times 1}$ 和 $\boldsymbol{y}\in R^{q\times 1}$,CCA 将找到 1 对投影方向 u_1 和 v_1 使得 $x_1^*=u_1^{\mathrm{T}}\boldsymbol{x}$ 与 $y_1^*=v_1^{\mathrm{T}}\boldsymbol{y}$ 之间具有最大相关性,并称为典型相关。x_1^*,y_1^* 称为第 1 对典型相关,同理可找到第 2 对 x_2^*,y_2^*,并使其与 x_1^*,y_1^* 不相关,类推使 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的相关性提取完毕。这样,只需要分析前 m 对典型变量的关系就可以得到 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的相关性。对一批特征向量样本集 $\{(x_i,y_i)\}_{i=1}^{n}\in R^{p\times 1}\times R^{q\times 1}$,记作 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in R^{p\times n}$,$\boldsymbol{Y}=[y_1,y_2,\cdots,y_n]\in R^{q\times n}$,使用 CCA 求得 2 组基向量 $\boldsymbol{u}$,$\boldsymbol{v}$,使 $\boldsymbol{x}^*=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}$ 和 $\boldsymbol{y}^*=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{y}$ 之间的关系最大,问题转化为求解相关系数的最大值问题:

$$\max\frac{E[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{v}]}{\sqrt{E[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}]\cdot E[\boldsymbol{v}^{\mathrm{T}}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{v}]}}=\max\frac{\boldsymbol{u}^{\mathrm{T}}S_{xy}\boldsymbol{v}}{\sqrt{\boldsymbol{u}^{\mathrm{T}}S_{xx}\boldsymbol{u}\cdot\boldsymbol{v}^{\mathrm{T}}S_{yy}\boldsymbol{v}}} \tag{8}$$

式中,E 为期望,S_{xx} 为 $\boldsymbol{X}$ 的协方差,S_{yy} 为 $\boldsymbol{Y}$ 的协方差,S_{xy} 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差。

1.4　视频关键帧时间对齐

根据视频的时间信息,将符合时间递增并且相邻视频帧差值不大于阈值的视频帧看做图的节点,使用有向边连接节点以构成路径,所有的路径起始于源节点,终止于尾节点。定义边的权重为其目的节点与查询帧的相似性评分,路径的权值就是其所经过的边的权值的累加,并获得最大累加权值的路径,从而确定拷贝视频片段。将时间对齐问题转化为网络流问题,借助网络流问题[15]解决时间对齐问题。

2　实验及结果分析

测试视频数据集使用 VCDB[8],VCDB 数据集共有 100 528 个视频,其中 528 个视频是核心视频,包含 9 236 对视频副本,其余 10 万个视频为干扰视频。数据集完全在网络上获取得到,不需要通过模拟拷贝攻击操作进行额外处理。VCDB 数据集中的拷贝方案很复杂,而且某些复制段的长度是源视频长度的一小部分,因此检测非常困难。本文实验所使用的计算机配置为 Intel ® Xeon(R) CPU E5-2630 v3 @ 2.40 GHz x32、62.6 G 内存、Tesla K40m。

为了评估检测的准确性,在实验中使用准确率(precision)和召回率(recall)2 个指标,并构造 PR 曲线与文献[8]的结果对比。

图 3 是本文算法在 VCDB 核心数据集上进行拷贝检测得到的 PR 曲线,Baseline(Temporal network)和Baseline(Hough voting)代表的曲线是基于局部特征 SIFT 的视频拷贝检测算法[8]的结果,Standard CNNofAlexNet是基于标准卷积神经网络的检测结果,Fusion of AlexNet and SCNN 是双胞胎网络特征和标准卷积神经网络特征相结合的检测结果。通过对比可以发现,使用本算法进行拷贝检测结果有更高的查全率和查准率。

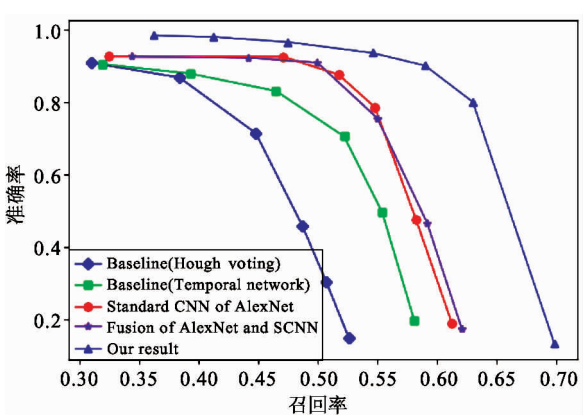

图 3　PR 曲线对比图

3 结 论

本文使用深度卷积神经网络 DenseNet 和 DCT 变换来提取视频帧的特征,并使用 CCA 特征融合算法提升算法性能,实验结果表明该算法对视频拷贝检测有较高的查全率和查准率,可以用于大型数据库上的视频拷贝检测。但是从实验结果可以发现,由于直接使用在 ImageNet 数据集上训练的 DenseNet-169 模型,目前的算法存在一定程度的漏检,下一步工作将会从重新训练或者微调神经网络模型等方面进一步提高拷贝检测的查全率和准确率。

参考文献

[1] 张三义, 张兴忠, 郝晓燕. 基于 ORB 和灰度序特征的视频拷贝检测[J]. 计算机应用研究, 2014,31(10): 3113-3116

[2] Kim C. Content-based image copy detection[J]. *Signal Processing Image Communication*, 2003, 18(3):169-184

[3] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2):91-110

[4] Herbert B, Tinne T, Luc V G. SURF: speed-up robust features[C]//European Conference on Computer Vision, Graz, Austria, 2006: 404-417

[5] Calonder M, Lepetit V, Strecha C, et al. BRIEF: binary robust independent elementary features[C]// European Conference on Computer Vision, Crete, Greece, 2010: 778-792

[6] Alahi A, Oritiz R, Vandergheynst P. Freak: fast retina keypoint[C]//IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 510-517

[7] Zhang X Y, Wang S, Yun X. Bidirectional active learning: a two-way exploration into unlabeled and labeled data set[J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2015, 26(12):3034-3044

[8] Jiang Y G, Wang J. Partial copy detection in videos: a benchmark and an evaluation of popular methods[J]. *IEEE Transactions on Big Data*, 2016, 2(1):32-42

[9] Huang G, Liu Z, Maaten L V D, et al. Densely connected convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2261-2269

[10] Zhang X Y, Xu C, Cheng J, et al. Effective annotation and search for video blogs with integration of context and content analysis[J]. *IEEE Transactions on Multimedia*, 2009, 11(2):272-285

[11] Zhang X Y, Xu C, Jian C, et al. Automatic semantic annotation for video blogs[C]//IEEE International Conference on Multimedia & Expo, Hannover, Germany,2008: 121-124

[12] Ross A, Jain A. Information fusion in biometrics[J]. *Pattern Recognition Letters*, 2003, 24(13): 2115-2125

[13] 孙权森, 曾生根, 王平安, 等. 典型相关分析的理论及其在特征融合中的应用[J]. 计算机学报, 2005, 28(9):1524-1533

[14] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11):2278-2324

[15] Ahuja R K, Magnanti T L, Orlin J B. Network flows: theory, algorithms, and applications[J]. *Journal of the Operational Research Society*, 1993, 45(11):791-796

Video copy detection based on deep learning features and DCT coefficients

Shi Huijie

(School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048)

Abstract

Many content-based video copy detection (CBCD) techniques have been proposed to identify the copies of a copyrighted video. Early algorithms using traditional features has the problems of weak robustness to viewing angle changes and high computational complexity. In recent years, deep learning has been used to get better detection results. Due to the poor performance from single visual features, an algorithm combining deep convolutional network features with traditional features is proposed. The deep features with dense convolutional network (DenseNet) and the ordinal measures of the coefficients of its discrete cosine transform (DCT) from the sampled video frames are first extracted, then the fusion features of these two features based on the canonical correlation analysis (CCA) algorithm are obtained. With the matching of fusion features, the copied videos are detected. Experiments on standard datasets show that the proposed method obtains a significant performance against common geometric attacks.

Key words: video copy detection, convolutional neural network (CNN), canonical correlation analysis (CCA), discrete cosine transform(DCT), dense convolutional network(DenseNet)