基于时空深度神经网络的视频指纹算法

汪冬冬, 李岳楠

天津大学电气自动化与信息工程学院, 天津 300072

摘要 随着内容分享网络的发展,网络上的视频数据急剧增长,出现了大量的非法拷贝。为了减少版权侵犯纠纷,需要检测出网络上的非法拷贝。视频指纹是实现拷贝检测的关键技术,能够将视频感知内容表示为简短摘要。利用降噪自编码器(DAE)稳健性强的优点,通过逐层训练 DAE 构建独立提取各帧特征的深度网络,设计了一种基于时空神经网络的视频指纹算法。在此基础上,采用长短时记忆网络提取视频时序特征,并根据慢变特征分析理论设计了网络训练算法。实验结果表明:基于时空神经网络的视频指纹在视频拷贝检测中能够表现出较高的准确率,性能指标优于对比算法。

关键词 图像处理; 视频指纹; 神经网络; 拷贝检测; 特征提取

中图分类号 TP37 **文献标识码** A **doi**: 10.3788/LOP55.011006

Video Fingerprint Algorithm Based on Spatio-Temporal Deep Neural Network

Wang Dongdong, Li Yuenan

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract With the development of content-sharing networks, the on-line video data have grown dramatically and a large number of illegal copies have been appeared. To reduce any copyright infringement disputes, it is necessary to detect illegal copies on-line internet. Video fingerprint, which can express the video perceptual content as a compact description, is a key technology for copy detection. The video fingerprint algorithm based on spatio-temporal deep neural network is designed by the use of the excellent robustness of denoising auto-encoder (DAE) and building a deep neural network to extract features on frame level through greedily training DAE. Consequently, a long short-term memory network is adopted to extract each frame features of the deep network, and the training algorithm is designed on the basis of the theory of slow-feature analysis. Experimental results show that the proposed algorithm can reveal a high accuracy in video copy detection and outperform a number of the comparative algorithms.

Key words image processing; video fingerprinting; neural network; copy detection; feature extraction

OCIS codes 100.4998; 100.4996; 100.2000

1 引 言

随着视频分享网站和移动互联网的发展,网络上的视频数据急剧增加,由此带来了版权侵犯与非法内容传播等问题。由于数据量庞大,无法依靠人力完成对非法拷贝视频的搜索。为解决这一问题,近年来,陆续提出了一些视频拷贝的检测方法。视频拷贝检测是在已知源视频的前提下,从海量数据中搜索其拷贝版本。视频指纹算法是拷贝检测的关键技术,它将视频的主要内容描述为一个类似于人类指纹的简短的内容摘要。视频拷贝检测技术通过比较视频指纹辨别两段视频是否同源。要求视频指纹具备以下特征:

1) 稳健性:视频在传播过程中可能会经过几何变换、有损压缩、转码或编辑等,只要视频本质内容不发生改变,其指纹需要保持稳定,原视频和其拷贝之间的指纹应高度相似。

2) 区分性:内容完全不同的两段视频的指纹需

收稿日期: 2017-07-18; **收到修改稿日期**: 2017-08-22

基金项目: 深圳市互联网产业发展专项资金(ZDSY20120613125016389)

作者简介: 汪冬冬(1991—),男,硕士研究生,主要从事图像处理方面的研究。E-mail: wddtju@aliyun.com

导师简介: 李岳楠(1981—),男,博士,副教授,主要从事多媒体信号处理方面的研究。E-mail: ynli@tju.edu.cn

要有较大差别。

3) 高效性：视频指纹应该简短、易于计算和搜索。

1999 年，Indyk 等[1]提出了视频指纹的概念。近年来，对视频拷贝检测技术的需求推动了视频指纹算法的研究，而特征提取是设计视频指纹算法的关键问题，直接决定了指纹的稳健性和区分性。按特征提取方式划分，视频指纹算法主要可分为帧级别的空间指纹级联和时空联合指纹等两类。第一类方法独立计算视频每一帧的指纹，然后级联得到整段视频的指纹。Roover 等[2]提出的径向哈希算法(Rash)是这类方法的代表，将沿每帧径向方向计算像素值的统计量作为指纹。由于梯度方向的重心具有稳健性的特点，因此 Lee 等[3]提出以梯度方向的重心(CGO)为特征计算指纹的算法。此外，一些算法以感兴趣区域[4]和稀疏特征[5]为基础设计视频指纹算法。第二类方法结合视频数据的时空关系计算视频指纹。Li 等[6]提出的结构图模型(SGM)通过时域和空间内的高效降维生成视频指纹。Nie 等[7]将视频表示为张量，提出了一种基于张量分解的视频指纹算法。Esmaeili 等[8]提出基于三维离散余弦变换(3D-DCT)的视频指纹算法，利用 DCT 变换的能量集中特性将视频的关键视觉信息抽象为简短指纹。

大部分传统视频指纹算法依赖于手工设计的特征提取方法，通常只能刻画视频某一方面的视觉特征。由于图像和视频数据所表达的信息十分复杂，因此手工设计的模型很难全面刻画这些信息，特别是一些抽象特征。神经网络能够自动提取这些特征，因此很多深度学习技术应用在计算机视觉领域，如行人检测[9]、目标跟踪[10]、图像超分辨率[11]等。研究人员将深度学习技术应用于视频拷贝检测。Jiang 等[12]将卷积神经网络应用于视频特征的提取，并且其性能高于传统方法的性能。Wang 等[13]将卷积神经网络和稀疏编码结合并应用于视频拷贝检测技术，与 Jiang 等提出的方法相比，该方法获得了更好的效果。Li 等[14]提出一种基于深度神经网络的视频(CRBM)算法，深度神经网络用于生成视频描述子，同样能学习到比传统方法稳健性更强的特征。本文利用神经网络来提取视频时空特征并将其压缩成稳健的视频指纹，通过构造空间特征提取网络(SFEN)和长短时记忆网络(LSTM)联合学习视频的时空特征，进一步提高视频指纹的稳健性；将慢变原则引入网络训练的目标函数中，并通过约束 LSTM 的隐层状态变化速度来学习时序稳健的特征。

2 空间特征提取

为了将不同时长和分辨率的视频映射为固定长度的指纹，首先需要对输入视频在时间和空间方向上做归一化处理。类似于文献[8]的方法，本算法将视频在时域上取平均得到的 20 个帧作为代表帧，再将每一帧在空间上降采样为 32×32 的相同尺寸，并将其变换成长度为 1024 的向量 $\boldsymbol{I}$ 作为输入。由于视频每帧数据量大并包含大量冗余信息，以帧为单位的空间特征提取是设计视频指纹算法中的关键问题。以降噪自编码器(DAE)为基本单元[15]，SFEN 结构如图 1 所示。

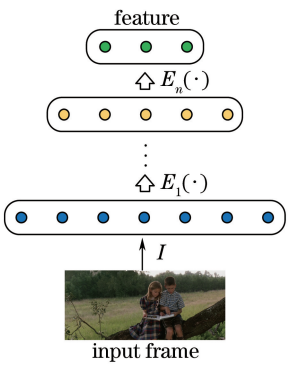

图 1 SFEN 结构图

Fig. 1 Architecture of SFEN

DAE 由 Pascal 等提出，通过训练神经网络完成降噪和压缩来实现稳健特征提取。为了使特征在各帧经过失真时仍然具有稳定性，本文算法以失真图像作为输入，训练网络将其重构成原始无失真图像。DAE 网络结构如图 2 所示，包括编码器与解码器两部分。

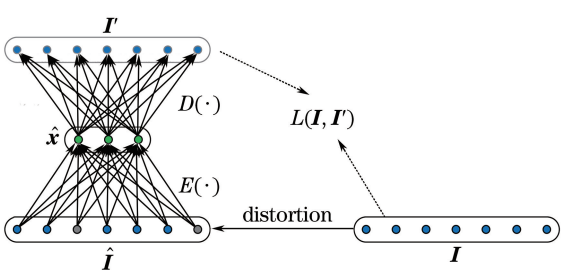

图 2 DAE 结构图

Fig. 2 Diagram of DAE

令 $\boldsymbol{I}$ 和 $\hat{\boldsymbol{I}}$ 分别表示输入的一帧图像和其失真后的图像，编码器将 $\hat{\boldsymbol{I}}$ 编码为特征 $\hat{\boldsymbol{x}}$：

$$\hat{\boldsymbol{x}}=E(\hat{\boldsymbol{I}})=\sigma(\boldsymbol{W}_E^1\hat{\boldsymbol{I}}+\boldsymbol{b}_E^1), \tag{1}$$

式中 $E(\cdot)$为编码器，$\sigma(x)=(1+\mathrm{e}^{-x})^{-1}$为 sigmoid 激活函数，$\boldsymbol{W}_E^1$ 为编码器权重矩阵，$\boldsymbol{b}_E^1$ 为编码器偏置向量。解码器将特征 $\hat{\boldsymbol{x}}$ 还原为图像 $\boldsymbol{I}'$：

$$\boldsymbol{I}'=D(\hat{\boldsymbol{x}})=f(\boldsymbol{W}_D^1\hat{\boldsymbol{x}}+\boldsymbol{b}_D^1), \tag{2}$$

式中 $D(\cdot)$为解码器，$\boldsymbol{W}_D^1$ 为解码器权重矩阵，$\boldsymbol{b}_D^1$ 为解码器偏置向量。网络训练的目标是使解码结果 $\boldsymbol{I}'$尽量接近原始无失真图像 $\boldsymbol{I}$，即最小化代价函数：

$$J_{\mathrm{DAE}}^1=\|\boldsymbol{I}-\boldsymbol{I}'\|_2^2+\alpha(\|\boldsymbol{W}_E^1\|_2^2+\|\boldsymbol{W}_D^1\|_2^2), \tag{3}$$

式中第一项为重构误差，第二项为正则项，α 为权重。SFEN 采用逐层贪婪训练方法，每个 DAE 网络采用上述训练方法进行独立训练，将得到的编码器 $E_i(\cdot)$作为 SFEN 的下一层；将解码器 $D_i(\cdot)$丢弃，然后在已训练好的编码器基础上继续训练下一层编码器。以第 $i+1$ 个 DAE 为例，将原始帧 $\boldsymbol{I}$ 和失真帧 $\hat{\boldsymbol{I}}$ 输入由之前已训练好的 i 个编码器级联所构成的网络 $\{E_1(\cdot),E_2(\cdot),\cdots,E_i(\cdot)\}$，分别得到编码结果 $\boldsymbol{x}^i=E_i\{\cdots E_2[E_1(\boldsymbol{I})]\}$，$\hat{\boldsymbol{x}}^i=E_i\{\cdots E_2[E_1(\hat{\boldsymbol{I}})]\}$，之后用 $\boldsymbol{x}^i$ 和 $\hat{\boldsymbol{x}}^i$ 训练编码器 $E_{i+1}(\cdot)$和解码器 $D_{i+1}(\cdot)$，如图 3 所示。

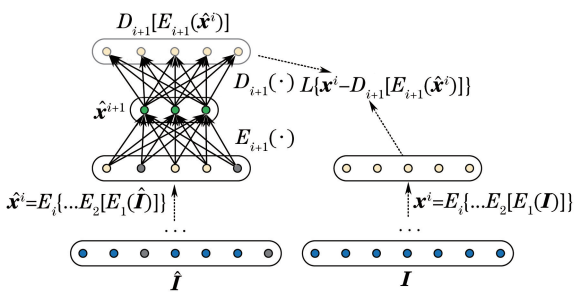

图 3 第 $i+1$ 个 DAE 训练

Fig. 3 Training of the $i+1^{\mathrm{th}}$ DAE

与(3)式类似，训练 $E_{i+1}(\cdot)$和 $D_{i+1}(\cdot)$的目标函数为

$$J_{\mathrm{DAE}}^{i+1}=\|\boldsymbol{x}^i-D_{i+1}[E_{i+1}(\hat{\boldsymbol{x}}^i)]\|_2^2+\alpha(\|\boldsymbol{W}_E^{i+1}\|_2^2+\|\boldsymbol{W}_D^{i+1}\|_2^2)。 \tag{4}$$

重复上述过程，将每次训练得到的编码器堆叠在网络的顶端得到空间特征提取网络(如图 1 所示)，网络的输出作为各帧的空间特征。本文采用了随机梯度下降的方法训练网络参数。

3 时序特征提取

由 SFEN 提取的帧特征本身虽然具有很强的稳健性，但每帧特征独立提取，无法描述视频的时序特性。为了提取这些特征，需要模型能够学习到各帧信息之间的关联。采用 LSTM 从 SFEN 输出的特征序列$(x_1,x_2,\cdots,x_T)$中学习视频的时序特征，整个模型的结构如图 4 所示。LSTM 网络利用循环结构将时序信息通过隐层传递到下一时刻，每个时刻 t 的隐层状态 $\boldsymbol{h}_t$ 由当前输入 $\boldsymbol{x}_t$ 与上一时刻的隐层状态 $\boldsymbol{h}_{t-1}$共同决定，从而具备学习序列时序特征的能力。

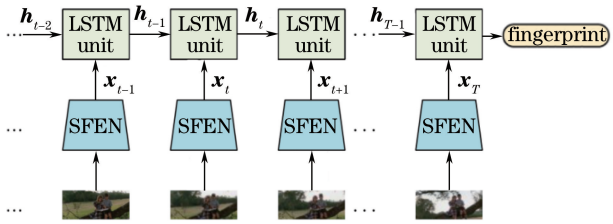

图 4 模型结构图

Fig. 4 Diagram of the whole network

为了防止网络训练陷入局部最优，采用“预训练＋微调”的方式训练 LSTM 网络。先用相对简单的训练目标预训练 LSTM，使 LSTM 从视频中提取包含慢变信息的特征；再以预训练得到的网络参数为初始值，以增强稳健性为目标微调 LSTM 网络。

3.1 预训练

采用的 LSTM 网络来自于文献[10]，其基本单元结构如图 5 所示。

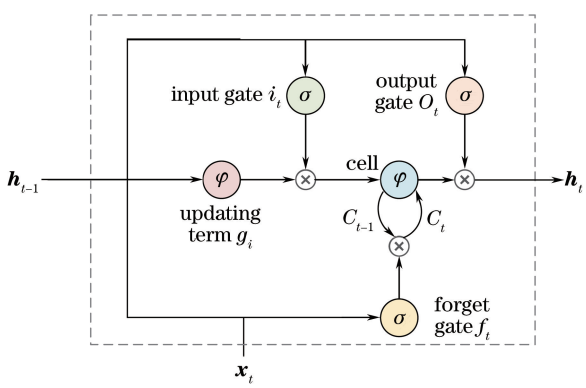

图 5 LSTM 基本单元结构图

Fig. 5 Structure of a basic LSTM unit

其中，记忆单元 $\boldsymbol{c}_t$ 保存着来自于上一个记忆单元 $\boldsymbol{c}_{t-1}$和当前的输入 $\boldsymbol{x}_t$ 的信息。每个 LSTM 单元包含 3 个调节门，即输入门、遗忘门、输出门；当门的值为 1 时，表示接受全部信息；门的值为 0 时，表示拒绝任何信息。将 sigmoid 函数 $\sigma(\cdot)$作为 3 个门的激活函数，其他激活函数选取双曲正切函数 $\phi(\boldsymbol{x})=\tanh\boldsymbol{x}$。当给定输入 $\boldsymbol{x}_t$、$\boldsymbol{h}_{t-1}$ 和 $\boldsymbol{c}_{t-1}$ 时，LSTM 单元在第 t 个时刻步骤的更新如下：

$$\begin{cases}\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t+\boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i)\\ \boldsymbol{f}_t=\sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t+\boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t+\boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o)\\ \boldsymbol{g}_t=\phi(\boldsymbol{W}_{gx}\boldsymbol{x}_t+\boldsymbol{W}_{gh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_g)\\ \boldsymbol{c}_t=\boldsymbol{f}_t\otimes\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\otimes\boldsymbol{g}_t\\ \boldsymbol{h}_t=\boldsymbol{o}_t\otimes\phi(\boldsymbol{c}_t)\end{cases}, \tag{5}$$

式中 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为权重矩阵和偏置向量，不同下标表示输出和起作用的输入项，如 $\boldsymbol{W}_{ix}$ 为输入 $\boldsymbol{x}$ 对输出 i 的权重矩阵，$\boldsymbol{b}_i$ 为输出 i 的偏置向量。$\boldsymbol{i}_t$ 表示输入门，用来控制是否从当前更新项 $\boldsymbol{g}_t$ 中获取信息。$\boldsymbol{f}_t$ 为遗忘门，用来控制是否遗忘 LSTM 之前的记忆信息 $\boldsymbol{c}_{t-1}$。$\boldsymbol{o}_t$ 为输出门，用来控制当前记忆 $\boldsymbol{c}_t$ 转换为隐层状态 $\boldsymbol{h}_t$。3 个调节门和更新项的值均由当前时刻 t 的输入 $\boldsymbol{x}_t$ 和前一时刻 $t-1$ 的隐含层状态 $\boldsymbol{h}_{t-1}$ 计算得到。

利用 LSTM 构造的时序特征提取网络模型，如图 6 所示。

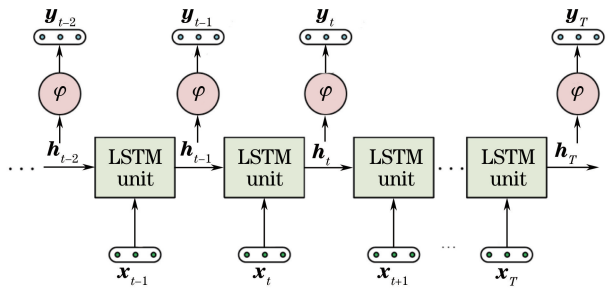

图 6 LSTM 网络

Fig. 6 LSTM network

图 6 中输出 $\boldsymbol{y}_t$ 由 LSTM 的隐层状态 $\boldsymbol{h}_t$ 计算得到：$\boldsymbol{y}_t=\varphi(\boldsymbol{W}_y\boldsymbol{h}_t+\boldsymbol{b}_y)$，使输出 $\boldsymbol{y}_t$ 接近空间特征 $\boldsymbol{x}_t$，强迫 $\boldsymbol{h}_t$ 保留输入 $\boldsymbol{x}_t$ 的信息。为了提高时序特征的稳健性，将慢变特征分析引入网络的训练过程。慢变特征分析是一种从快速变化的输入信号中提取稳定特征的无监督算法[17-18]，其基本思想来源于视觉系统的慢变原则：当视频中的场景中的人物或物体等目标发生变化时，相对于快速变化的像素值，其中的重要特征的变化相对缓慢。为模拟这一特征，使 $\boldsymbol{h}_t$ 在能重构出输入信号 $\boldsymbol{x}_t$ 的前提下，尽可能与前一时刻特征 $\boldsymbol{h}_{t-1}$ 相近，由此定义代价函数：

$$J_{\mathrm{LSTM1}}=\sum_{t=1}^{T}\left(\left\|\boldsymbol{y}_t-\boldsymbol{x}_t\right\|_2^2+\lambda\left\|\boldsymbol{h}_t-\boldsymbol{h}_{t-1}\right\|_2^2\right),\tag{6}$$

式中 T 为时间长度，λ 为慢变项的权重。通过(6)式优化学习网络中的参数。

3.2 微 调

当视频经过有损编码或加噪声等失真时，失真视频的指纹应当与原始视频指纹高度相似。为了实现这一目标，以预训练所得到的网络参数为初始值，利用成对的原始和失真视频组合进一步微调 LSTM 网络参数。如图 4 所示，LSTM 网络的输入为 SFEN 计算得到的各帧特征序列 $(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$。在 LSTM 的循环结构中，最后时刻的隐层状态 $\boldsymbol{h}_T$ 包含了之前所有输入帧的信息，将 $\boldsymbol{h}_T$ 作为视频指纹。令 $(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$ 和 $(\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\cdots,\hat{\boldsymbol{x}}_T)$ 分别表示原始和失真视频的各帧特征序列，令 $\boldsymbol{h}_T=F_{\mathrm{LSTM}}(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$ 和 $\hat{h}_T=F_{\mathrm{LSTM}}(\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\cdots,\hat{\boldsymbol{x}}_T)$ 分别为二者的视频指纹，$F_{\mathrm{LSTM}}(\cdot)$ 为 LSTM 网络将特征序列映射为视频指纹的函数，则微调 LSTM 的目标是最小化指纹距离：

$$J_{\mathrm{LSTM2}}=\left\|\hat{\boldsymbol{h}}_T-\boldsymbol{h}_T\right\|_2^2。\tag{7}$$

(6)、(7)式均采用自适应估计时刻(ADAM)的优化方法训练网络参数，微调的过程如图 7 所示。完成微调后，将 LSTM 和前文所述的 SFEN 级联构成计算视频指纹的深度网络，如图 4 所示。

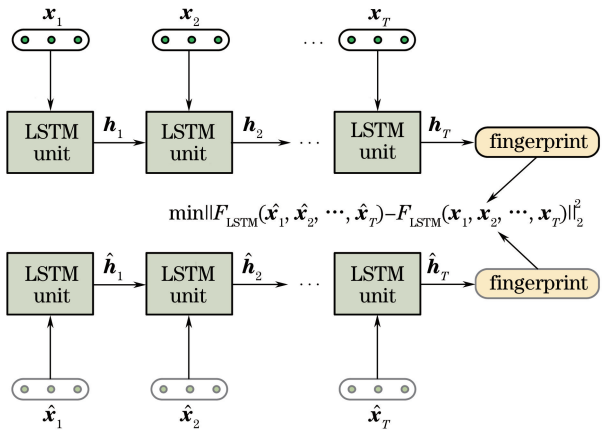

图 7 微调 LSTM 网络

Fig. 7 Fine-tuning of the LSTM network

4 实验结果及分析

本文的拷贝检测实验数据库取自公共标准测试数据集，其中包含了 600 个从 You Tube 上下载的视频序列和另外 201 个来自 TRECVID 的视频序列，任意两段视频是不同的。对每段视频施加 9 种常见的内容保持失真，主要包括信号处理失真、几何失真和时域失真等。具体失真类型和参数设置如表 1 所示。每个原始视频经过这些失真处理后产生 17 个拷贝版本，测试库中的原始视频和拷贝的总数目为 14418 个。为了降低视频指纹提取的复杂度，采用文献[8]方法先对视频数据进行空间和时序归一化预处理，处理后的视频尺寸固定为 32×32×20，每帧像素经过归一化后均值为 0，方差为 1。

表 1 失真类型与参数设置

Table 1 Distortion types and parameters setting

Distortion	Description
Compression	Encoder: XVid, frame rate: 25 frame/s, bit rate: 256 kb/s, fixed resolution: 480×320
Median filtering	Filter size ∈[10,20]
Gaussian noise	Zero mean, variance ∈[0.1,0.5,1]
Rotation+croping	Pixel ∈[2,5,10]
Histogram equalization	Number of gray levels ∈[16,32,64]
Frame dropping	Delete 25% frames then linearly interpolate
Frame resizing	Ratio ∈[0.2,4]
Joint spatio-temporal distortion 1	Combine median filtering (filter size is 10) and frame dropping (25%)
Joint spatio-temporal distortion 2	Combine rotation (pixel is 5) and frame dropping (25%)

视频指纹提取网络由 SFEN 和 LSTM 级联而成，如图 4 所示。其中，SFEN 的结构为(1024-800-400)，而 LSTM 的结构为(400-128)(分别为 $\boldsymbol{x}_i$、$\boldsymbol{h}_i$ 的节点数)，时间长度 $T=10$(即每 10 个归一化帧为单位计算指纹)。算法将 LSTM 的隐藏层输出作为指纹，因此归一化后的视频(20 帧)通过网络后被映射为长度为 256 的指纹。

由于 SFEN 独立从各帧中提取特征，不涉及时序特征提取，因此 SFEN 的训练数据为静态图像，训练图像从 ImageNet 中随机选取，设置权重衰减系数 $\alpha=10^{-5}$，学习率为 0.014。LSTM 的训练数据为 400 个来自于好莱坞场景数据集视频，与测试视频无重叠。对于每个训练视频，将经过归一化后先由训练好的 SFEN 从各帧中提取的特征序列作为 LSTM 网络的训练数据。预训练 LSTM 时，选取权重衰减系数 $\lambda=0.5$。在微调过程中，对 400 个训练视频进行失真处理，并约束失真视频所提取到的指纹与已有的原始视频指纹相似。

选用 5 个具有代表性的视频指纹算法作为对比，分别为 SGM 算法[7]、3D-DCT 算法[8]、Rash 算法[2]和 CGO 算法[3]。另外，同样使用 CRBM[16]作为对比。对于14418个视频，当2个视频数据的视频指纹之间的欧氏距离小于阈值 β 时，则判定这 2 个视频来自于同一个原始视频的 2 个版本，否则判定为来自不同原始视频。当 β 取不同值时的判定结果与实际情况比较，计算出错误拒绝率(FRR)和错误接受率(FAR)，依此绘制出受试者工作特征(ROC)曲线(图 8)并计算 F_1 值

$$F_1=\max_T\left\{\frac{2\cdot[1-\mathrm{FAR}(\beta)]\cdot[1-\mathrm{FRR}(\beta)]}{[1-\mathrm{FAR}(\beta)]+[1-\mathrm{FRR}(\beta)]}\right\},\tag{8}$$

设所有测试视频中指纹欧氏距离最大值为 $d_{\max}$，则 β 的变化范围为$[0,d_{\max}]$。

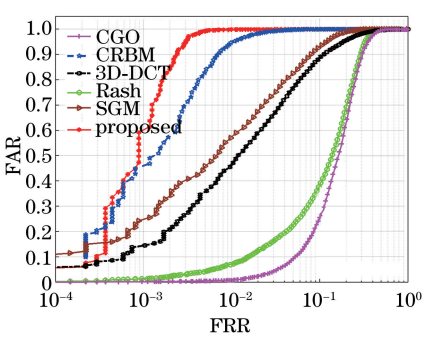

图 8 ROC 曲线对比

Fig. 8 Comparison of ROC curves

对于每种失真，分别计算本文算法和对比方法的 F_1 指标，如表 2 所示。

表 2 F_1 指标对比

Table 2 Comparison of the F_1 scores

Distortion	Proposed algorithm	CRBM algorithm	CGO algorithm	SGM algorithm	3D-DCT algorithm	Rash algorithm
Compression	0.966	0.907	0.801	0.929	0.889	0.845
Filtering	0.998	0.994	0.887	0.978	0.991	0.966
Noise	0.999	0.999	0.813	0.890	0.896	0.624
Rotation	0.999	0.988	0.534	0.939	0.838	0.950
Equalization	0.996	0.980	0.839	0.836	0.851	0.519
Dropping	0.999	0.999	0.996	0.994	0.999	0.999
Resizing	0.999	0.999	0.903	0.994	0.998	0.948
Joint distortion 1	0.998	0.996	0.941	0.985	0.997	0.981
Joint distortion 2	0.999	0.993	0.516	0.968	0.888	0.965
Overall	0.995	0.982	0.783	0.915	0.894	0.793

实验结果显示,本文算法的 ROC 曲线图和 F_1 指标在所有对比实验中均优于对比算法,说明由时序深度神经网络计算得到的视频指纹具有更好的稳健性和区分性,可以实现精确的视频拷贝检测。值得注意的是,在所有方法中性能较好的是基于视频时序特征的一类指纹算法(本文算法、CRBM 算法、SGM 算法和 3D-DCT 算法),因为这类算法可以抵抗有损压缩和丢帧等带来的非同步失真。本文算法具有高稳健性的另外一个原因是在训练深度神经网络时考虑到了网络的慢变特征学习能力,而慢变特征在视频内容发生失真时能够相对保持稳定。

5 结 论

提出了一种新型的基于深度神经网络的视频指纹学习模型,使用降噪自编码网络提取帧特征,使用 LSTM 网络提取帧之间的时间相关信息。在此基础上通过约束 LSTM 隐层在连续时刻的相似性提取视频的慢变特征,并通过微调进一步增强视频指纹的稳健性。实验结果显示本方法在视频拷贝检测中具有较强的准确率,对于常见的内容失真,其 F_1 指标均在 0.96 以上。

参 考 文 献

[1] Indyk P, Iyengar G, Shivakumar N. Finding pirated video sequences on the internet [R]. Stanford: Stanford University, 1999.

[2] De Roover C, De Vleeschouwer C, Lefebvre F, *et al*. Robust video hashing based on radial projections of key frames [J]. IEEE Transactions on Signal processing, 2005, 53(10): 4020-4037.

[3] Lee S, Yoo C D. Robust video fingerprinting for content-based video identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(7): 983-988.

[4] Li J N, Guo X Q, Yu Y, *et al*. A robust and low-complexity video fingerprint for multimedia security [C]. 2014 International Symposium on Wireless Personal Multimedia Communications (WPMC), 2014: 97-102.

[5] Wu B, Krishnan S S, Zhang N, *et al*. Compact and robust video fingerprinting using sparse represented features[C]. 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016: 1-6.

[6] Li M, Monga V. Compact video fingerprinting via structural graphical models [J]. IEEE Transactions on Information Forensics and Security, 2013, 8(11): 1709-1721.

[7] Nie X S, Yin Y L, Sun J D, *et al*. Comprehensive feature-based robust video fingerprinting using tensor model[J]. IEEE Transactions on Multimedia, 2017, 19(4): 785-796.

[8] Esmaeili M M, Fatourechi M, Ward R K. A robust and fast video copy detection system using content-based fingerprinting[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(1): 213-226.

[9] Ye G L, Sun S Y, Gao K J, *et al*. Nighttime pedestrian detection based on faster region convolution neural network[J]. Laser & Optoelectronics Progress, 2017, 54(8): 081003.
叶国林, 孙韶媛, 高凯珺, 等. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8): 081003.

[10] Gao L, Wang J F, Fan Y, *et al*. Robust visual tracking based on convolutional neural networks and conformal predictor[J]. Acta Optica Sinica, 2017, 37(8): 0815003.
高琳, 王俊峰, 范勇, 等. 基于卷积神经网络与一致性预测器的稳健视觉跟踪[J]. 光学学报, 2017, 37(8): 0815003.

[11] Xiao J S, Liu E Y, Zhu L, *et al*. Improved image super-resolution algorithm based on convolutional neural network [J]. Acta Optica Sinica, 2017, 37(3): 0318011.
肖进胜, 刘恩雨, 朱力, 等. 改进的基于卷积神经网络的图像超分辨率算法[J]. 光学学报, 2017, 37(3): 0318011.

[12] Jiang Y G, Wang J 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.

[13] Wang L, Bao Y, Li H J, *et al*. Compact CNN based video representation for efficient video copydetection [C]. International Conference on Multimedia Modeling, 2017: 576-587.

[14] Li Y N, Chen X P. Robust and compact video descriptor learned by deep neural network[C]. 2017 IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: 2162-2166.

[15] Franc V, Laskov P, Müller K R. Stopping conditions for exact computation of leave-one-out error in support vector machines [C]// Proceedings of the 25th International Conference on Machine Learning, 2008: 328-335.

[16] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]. 2013 IEEE International conference on Acoustics, Speech and Signal Processing, Vancouver (ICASSP), 2013: 6645-6649.

[17] Wiskott L, Berkes P, Franzius M, *et al*. Slow feature analysis[J]. Scholarpedia, 2011, 6(4): 5282.

[18] Wiskott L, Sejnowski T J. Slow feature analysis: Unsupervised learning of invariances [J]. Neural Computation, 2002, 14(4): 715-770.