

# 基于 3D 卷积的视频错帧筛选方法

缪宇杰<sup>1</sup>, 吴智钧<sup>1</sup>, 宫婧<sup>2</sup>

(1. 南京邮电大学 物联网学院, 江苏 南京 210003;

2. 南京邮电大学 理学院, 江苏 南京 210003)

**摘要:**为了提取更好的视频特征,提高训练精准度,提出了一个基于 CNN(convolutional neural network,卷积神经网络)的错帧筛选模型。所谓错帧,是指在时间上乱序的帧序列,相反,有序帧是指遵守时间顺序的帧序列。其目标是从若干组帧序列中,筛选出顺序错误的一组帧序列。采用无监督学习的方法来训练模型,因此不需要依赖有标签的数据集。基于这个模型的目标以及无标签的训练方式,采用了一个多分支的 CNN 结构,并且是端到端的。其输入的若干组帧序列从视频中采样获得,分别进行 3D 卷积编码后,能够提取出每组帧序列在时间和空间上的特征。为了找出帧顺序有误的一组序列,该模型对每组帧序列进行对比,找出它们之间的共同规则,从而筛选出违背此规则的那一组序列。在 UCF101 数据集上的实验结果证实了该方法的有效性,错帧筛选的准确率高。

**关键词:**无监督学习;卷积神经网络;错帧筛选;3D 卷积

**中图分类号:**TP301

**文献标识码:**A

**文章编号:**1673-629X(2018)05-0179-03

**doi:**10.3969/j.issn.1673-629X.2018.05.040

## A Wrong Temporal-order Frames Identification Method Based on 3D Convolution

MIAO Yu-jie<sup>1</sup>, WU Zhi-jun<sup>1</sup>, GONG Jing<sup>2</sup>

(1. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:**In order to extract better video features and improve training accuracy, we propose a model of wrong temporal-ordered frames based on CNN (convolutional neural network), whose task is identifying the sequence of wrong temporal-ordered frames from several sequences of frames. The sequence of wrong frames is wrong temporal-ordered while the right sequence is temporal-ordered. Unsupervised video representation learning is applied to train this model, therefore labeled data sets are unnecessary. Based on the task and no semantic labels, a multi-branched CNN structure is implemented which is learned end-to-end. As the model input, the sequences of frames are sampled from one video. Then, these sequences of frames are encoded with the method of 3D convolution to extract the temporal and spatial features of each sequence of frames. To find out the sequence of frames with wrong temporal-order, the model has to compare all the inputs, analyze the regularities among them, and identify the one with irregularities. The experiments on UCF101 dataset verify the effectiveness of the proposed method, and the accuracy of this model is high.

**Key words:**unsupervised learning; CNN; frame-sequence identification; 3D convolution

## 0 引言

近年来,随着深度学习的兴起,诸如 CNN 等深度学习框架的提出,很多机器学习的问题得到了解决,比如在真实场景下的目标识别、人体行为分析等等。但是,其识别结果的精准度还是不能令人满意,精准度的提高依然是深度学习领域一项具有挑战性的任务。

好的视频特征应该具备丰富的与识别内容相关的

信息。视频可以看作是一组连续帧,即静态图片。每张静态图片所提取的特征是独立的、互不相关的,并且只存在于空间维度上。为了更好地提取视频信息,有必要找到帧与帧之间的联系。文中采用 3D 卷积的方法,能够同时在时间和空间维度上提取视频特征<sup>[1]</sup>。

要正确提取视频的特征视频,首要条件是必须保证帧序列有序。假设帧序列是无序的,那么根据该序

收稿日期:2017-06-27

修回日期:2017-10-09

网络出版时间:2018-02-08

基金项目:国家自然科学基金(61373135);南京市六大高峰人才资助项目(C类)

作者简介:缪宇杰(1992-),男,硕士,研究方向为图像处理和模式识别;宫婧,博士,副教授,研究方向为信息网络、计算机网络及其应用。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20180207.1917.090.html>

列所提取的特征很有可能是不准确的,利用这样的特征来训练或者测试深度学习的模型,很可能会导致最终结果的误判。所以,验证帧序列是否有序是一项很重要的任务。

文中提出一种方法来验证视频帧序列的顺序。首先,提出错帧筛选模型,描述了其整体结构;其次,对该模型的主要技术关键点进行详细介绍;最后,通过实验对该方法进行验证。

## 1 相关研究

机器学习<sup>[2]</sup>分为有监督和无监督两个类,基本上可以从它们会不会得到一个特定的标签输出来区分。监督学习(supervised learning)是通过已有的训练样本(即已知数据及其对应的输出)来训练,从而得到一个最优模型,再利用这个模型将所有新的数据样本映射为相应的输出结果,对输出结果进行简单的判断从而实现分类的目的。那么这个最优模型也就具有了对未知数据进行分类的能力。而无监督学习(unsupervised learning)<sup>[3]</sup>事先没有任何训练数据样本,需要直接对数据进行建模。无监督学习在学习时并不知道其分类结果是否正确,亦即没有受到监督式增强(告诉它何种学习是正确的)。其特点是仅对此种网络提供输入范例,且自动从这些范例中找出其潜在类别规则。当学习完毕并经测试后,也可以将之应用到新的案例上。

现有的大多数深度学习模式识别方法通常由两个关键步骤组成:第一步是手工标注数据集的特征,第二步是在已标注的特征基础上学习分类器<sup>[4-7]</sup>。但是,手工标注作为有监督学习的特点之一正变得越来越不受欢迎,原因是耗费了大量的时间和精力,尤其在数据集更加复杂的情况下,手工标注的代价成倍增长。因此,文中采用无监督学习的方法来学习没有经过手工标注的视频特征。

文献[8-11]提出了一些无监督学习方法,这些文献阐明了在没有标签的情况下,视频或者图像的时间、空间结构也能够提供充分的信息。文献[12]提出了一种基于 CNN 的无监督学习方法,其主要目标是验证视频帧序列是否有序。该验证模型可看作是一个二元问题,它只是回答了一组帧是否是有序的。而文中提出的模型是一个多元的问题,能够从多组帧序列中找到错帧的一组序列。除此以外,还对帧序列进行了编码,以获取帧与帧之间在时间、空间上的信息。

## 2 错帧筛选模型

文中的目标是通过错帧筛选模型,从若干组帧序列中,将错误的一组帧序列筛选出来。从同一个视频中采样出若干组帧序列(详见 3.1 小节),假设有

$N + 1$  组帧序列,那么此模型的输入可表示为  $f = \{f_1, f_2, \dots, f_{N+1}\}$ ,其中,  $f_i$  为第  $i$  组帧序列。在这组输入中,有  $N$  组帧序列是有序的,只有一组帧序列是错序的,且这组错帧的位置随机。

将  $f$  的每一组帧分别输入错帧筛选模型的一个分支,如图 1 所示。首先对其进行编码(详见 2.2 小节),编码后每个分支会通过 5 个卷积层和 1 个全连接层,这一部分与 AlexNet<sup>[13]</sup>相同。每个分支网络的权值以及参数均相同。

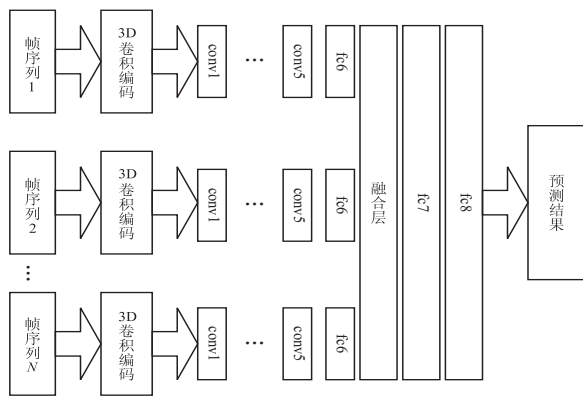


图 1 错帧筛选模型

错帧筛选的最终目的是从  $f$  中找到唯一的一组错帧,所以需要将每个分支融合到一起进行比较,从而筛选出错帧。因此,在每个分支的第一个全连接层之后,提出了新的一层—融合层,该层能够比较每个分支的特点并找到某种规则,通过这种规则来筛选错帧。设第  $i$  个分支的特征向量为  $v_i$ ,则融合层的输出为

$$\sum_{j>i} [\frac{1}{j} \sum_{q=1}^j v_q - \frac{1}{i} \sum_{p=1}^i v_p]。$$

将上述计算结果输入最后两个全连接层和一个线性分类器,这个分类器能够对  $N + 1$  个输入进行分析对比,进而预测出错帧的一组视频序列。

## 3 关键技术

### 3.1 视频帧采样

视频帧的采样也是一项非常重要的工作,具有良好特性的帧序列有助于预测结果精准度的提升。如果采样的帧序列之间的变化很小,那么很难判断出这组帧序列是顺序还是乱序。图 2 所示为一组有序的视频帧,帧 a 和 b、帧 e 和 f 之间的动作变化很小,而帧 b、c、d、e 之间差别很明显。由图 2 易知,很难判断  $\{a, b, c\}$ 、 $\{b, a, c\}$  哪组是乱序,但  $\{c, d, e\}$ 、 $\{e, c, d\}$  很容易判断,因此需要选取帧间差异较大的帧作为输入。使用粗帧级光流<sup>[14]</sup>来测量帧与帧之间的变化程度,把每个帧的平均流量大小作为该帧的权重,并用它来偏置采样较大变化帧的窗口。此方法保证了采样的帧序列不会出现难以分辨是否为错帧的情况。

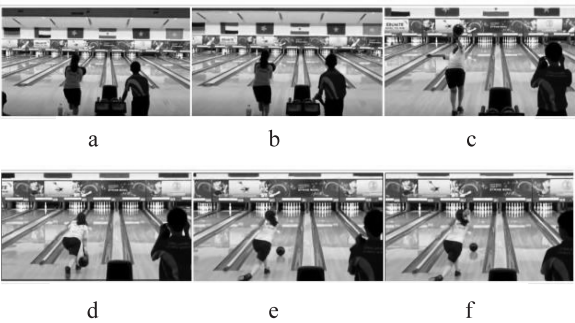


图2 帧间差异示例

设采样结果为 $I = I_1, I_2, \dots, I_n$  这样一组包含  $n$  帧的视频序列,且有序,即  $I_1 < I_2 < \dots < I_n$ 。在上述的采样结果  $I$  中,还需要再进行一次采样作为错帧筛选模型的输入。在这个步骤中,采用了随机采样的方法。在  $I$  中随机采样出  $X$  帧图像  $N$  次,则产生了  $N$  组有序的帧序列,每组有  $X$  帧。乱序的帧序列的采样也是随机采样  $X$  帧,并保证乱序。

3.2 视频帧编码

在完成视频帧采样之后,需要对每一组帧序列进行编码,编码的目的是提取帧序列的结构信息。完成编码后,可以将多帧图片合并为一帧。这样做的好处是在训练错帧筛选模型时,不需要限定每组输入的帧数,因为不论每组输入的帧数是多少,通过编码都可以提取为一帧的信息。

文中采用3D 卷积<sup>[15]</sup>的方法进行编码。在2D CNN 中,2D 卷积在卷积层具有提取局部邻域上层特征映射的功能。坐标为  $(x,y)$  的单元在第  $i$  层的第  $j$  个特征映射的值为:

$$v_{ij}^{xy} = \tanh(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}) \tag{1}$$

其中,  $\tanh(\cdot)$  是双曲正切函数;  $b_{ij}$  是该特征映射的偏置;  $m$  是  $(i-1)$  层上的特征映射集合的索引,该层连接了当前特征映射;  $w_{ijm}^{pq}$  是坐标  $(p,q)$  的核值,该核连接了第  $m$  个特征映射;  $P_i$ 、 $Q_i$  分别是核的高度与宽度。

2D CNN 中的特征映射也是2D 的,只反映了图像空间上的信息,没有考虑时间上的信息。文中采用3D CNN 中的3D 卷积方法,在卷积层的特征映射连接了上一层多帧连续图像,这样既包含了空间信息,也包含了时间信息,从而可以获取一组帧序列的信息。则式1 可以改写为:

$$v_{ij}^{xyz} = \tanh(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}) \tag{2}$$

其中,  $R_i$  是3D 卷积核在时间维度上的大小。  
在视频帧编码中只进行了一次卷积运算,然后对一组帧序列的特征映射求均值,结果即为编码的结果。

4 实验结果与分析

使用UCF101 数据集进行实验。UCF101 数据集是由真实用户上传的具有复杂背景的视频,共有101 个动作类别,13 000 个视频片段,时长共27 小时。  
实验中,从同一个视频片段中采样7 组帧序列,其中6 组是有序的,1 组是无序的。每组帧序列有7 帧图像,大小为  $80 \times 60$ ,卷积核大小为  $7 \times 7 \times 3$  ( $7 \times 7$  表示空间维度,3 表示时间维度)。将帧序列输入网络,首先对帧序列进行编码,编码结果如图3 所示。

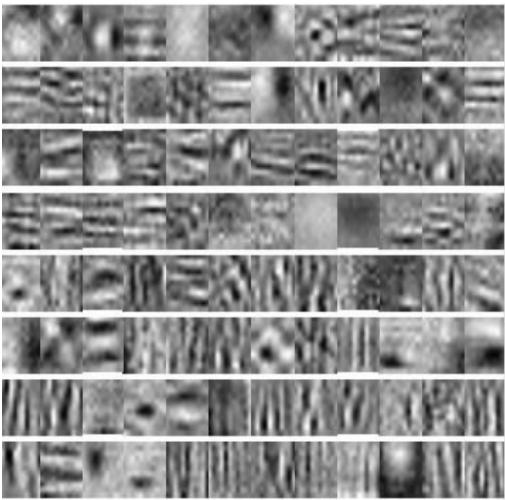


图3 帧序列编码结果

为了验证该方法的有效性,将实验获得的验证错帧结果的准确性与文献[12] 比较,结果如表1 所示。从表1 可见,文中方法提高了对错帧检验的准确性。

表1 不同帧顺序验证方法准确性对比

方法	准确性/%
元组验证方法 <sup>[12]</sup>	72.1
文中方法	75.6

5 结束语

文中提出了一种基于3D 卷积的视频帧顺序验证方法,能够对视频帧序列顺序与否进行验证。通过无监督学习视频特征的方法,避免了有监督学习中所需的手工标注标签的过程,很大程度上减少了时间与精力的耗费。3D 卷积对视频序列特征的提取,不仅获取了该序列空间上的信息,同时获取到了时间上的信息,提升了验证的准确性。

参考文献:

[1] 林海波,李 扬,张 毅,等. 基于时序分析的人体运动模式的识别及应用[J]. 计算机应用与软件,2014,31(12): 225-228.  
[2] 郭丽丽,丁世飞. 深度学习研究进展[J]. 计算机科学, 2015,42(5):28-33.



信息更多、更明显,而且加快了处理速度。

C++AMP 在并行处理上高速且高效,由于 C++AMP 的可扩展性使得 C++AMP 作为一种新的异构并行技术得以快速发展<sup>[19]</sup>。C++AMP 是一种新的具有集成优势的并行化技术,而基于 C++AMP 的图像并行化技术对于图像处理的速度以及效率都有很大的优势。C++AMP 在图像处理中应用广泛,也可以在其他需要并行的地方使用。随着 C++AMP 技术的不断成熟与改进,C++AMP 在并行计算领域会占据更重要的地位<sup>[20]</sup>。

#### 参考文献:

- [1] 胡 琼,秦 磊,黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报,2013,36(12):2512-2524.
  - [2] 董荣胜.《九校联盟(C9)计算机基础教学发展战略联合声明》呼唤教育的转型[J]. 中国大学教学,2010(10):14-15.
  - [3] 陈冠诚. C++AMP 异构并行编程解析[J]. 程序员,2012(4):104-106.
  - [4] 肖 汉. 基于 CPU+GPU 的影像匹配高效能异构并行计算研究[D]. 武汉:武汉大学,2011.
  - [5] 丁 鹏,陈利学,龚 捷,等. GPU 通用计算研究[J]. 计算机与现代化,2010(1):12-15.
  - [6] 陈宏希. 基于边缘保持平滑滤波的 Canny 算子边缘检测[J]. 兰州交通大学学报,2006,25(1):86-90.
  - [7] 唐志文. 浅析数字图像处理技术的研究现状及其发展方向[J]. 硅谷,2010(5):30.
  - [8] 吴学明,李灿平. 边缘检测算法在不同分辨率图像中的性能研究[J]. 计算机测量与控制,2006,14(2):166-169.
  - [9] 王 兰,吴 谨. 一种改进的 Canny 边缘检测算法[J]. 微计算机信息,2010,26(2):198-199.
  - [10] LIANG L R, LOONEY C G. Competitive fuzzy edge detection[J]. Applied Soft Computing, 2003, 3(2):123-137.
  - [11] PELLEGRINO F A, VANZELLA W, TORRE V. Edge detection revisited[J]. IEEE Transactions on Systems, Man and Cybernetics, 2004, 34(3):1500-1518.
  - [12] 王 蓉,高立群,柴玉华,等. 综合 Canny 法与小波变换的边缘检测方法[J]. 东北大学学报:自然科学版,2005,26(12):1131-1133.
  - [13] 朱仲涛,张 钹,张再兴. 图像关于边缘提取算子的微分不变性[J]. 计算机学报,1999,22(9):903-910.
  - [14] 姚 平. CUDA 平台上的 CPU/GPU 异步计算模式[D]. 合肥:中国科学技术大学,2010.
  - [15] SANDERS J. GPU 高性能编程 CUDA 实战[M]. 聂雪军,译. 北京:机械工业出版社,2011.
  - [16] COOK S. A developer's guide to parallel computing with GPUs[M]. [s. l.]:Morgan Kaufmann,2012.
  - [17] 宗露艳,吴 陈. 一种改进的 Canny 算子边缘检测算法[J]. 现代电子技术,2011,34(4):104-106.
  - [18] 张焕龙,胡士强,杨国胜. 基于外观模型学习的视频目标跟踪方法综述[J]. 计算机研究与发展,2015,52(1):177-190.
  - [19] GOMEZ-LUNA J, GONZALEZ-LINARES J M, BENAVIDES J I, et al. An optimized approach to histogram computation on GPU[J]. Machine Vision and Applications, 2013, 24(5):899-908.
  - [20] DESTREMPES F, MIGNOTTE M. A statistical model for contours in image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(5):626-638.
- +++++
- (上接第 181 页)
- [3] 殷瑞刚,魏 帅,李 晗,等. 深度学习中的无监督学习方法综述[J]. 计算机系统应用,2016,25(8):1-7.
  - [4] 王满一,宋亚玲,李 玉,等. 结合区域光流特征的时序模板行为识别[J]. 系统仿真学报,2015,27(5):1146-1151.
  - [5] JHUANG H, SERRE T, WOLF L, et al. A biologically inspired system for action recognition[C]//International conference on computer vision. Rio de Janeiro, Brazil: IEEE, 2007:1-8.
  - [6] 杨祎玥,伏 潜,万定生. 基于深度循环神经网络的时间序列预测模型[J]. 计算机技术与发展,2017,27(3):35-38.
  - [7] 徐庆伶,汪西莉. 一种基于支持向量机的半监督分类方法[J]. 计算机技术与发展,2010,20(10):115-117.
  - [8] DOERSCH C, GUPTA A, EFROS A A. Unsupervised visual representation learning by context prediction[C]//International conference on computer vision. [s. l.]:IEEE,2015.
  - [9] 朱 陶,任海军,洪卫军. 一种基于前向无监督卷积神经网络的人脸表示学习方法[J]. 计算机科学,2016,43(6):303-307.
  - [10] PICKUP L C, PAN Z, WEI D, et al. Seeing the arrow of time[C]//IEEE conference on computer vision and pattern recognition. Columbus, OH, USA:IEEE,2014:2043-2050.
  - [11] JAYARAMAN D, GRAUMAN K. Learning image representations tied to ego-motion[C]//International conference on computer vision. Santiago, Chile:IEEE,2015:1413-1421.
  - [12] MISRA I, ZITNICK C L, HEBERT M. Shuffle and learn; unsupervised learning using temporal order verification[C]//European conference on computer vision. Berlin: Springer, 2016:527-544.
  - [13] JEFF D, JIA Yangqing, VINYALS O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition[C]//International conference on machine learning. Beijing, China:ACM,2014.
  - [14] FARNEBÄCK G. Two-frame motion estimation based on polynomial expansion[M]//Scandinavian conference on image analysis. Berlin: Springer,2003:363-370.
  - [15] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013,35(1):221-231.