基于双流网络与支持向量机融合的人体行为识别

童安炀[1,2] 唐 超[1,2] 王文剑[3]

摘 要 传统的双流卷积神经网络存在难以理解长动作信息的问题,并且当长时间流信息损失时,模型泛化能力降低. 针对此问题,文中提出基于双流网络与支持向量机融合的人体行为识别方法. 首先,提取视频中每帧 RGB 图像及其对应垂直方向的稠密光流序列图,得到视频中动作的空间信息和时间信息,分别输入空间域和时间域网络进行预训练,预训练完成后进行特征提取. 然后,针对双流网络提取的维度相同的特征向量执行并联融合策略,提高特征向量的表征能力. 最后,将融合后的特征向量输入线性支持向量机中进行训练及分类处理. 在 KTH、UCF sports 数据集上的实验表明文中方法具有较好的分类效果.

关键词 双流网络, 支持向量机, 特征融合, 光流

引用格式 童安炀,唐 超,王文剑. 基于双流网络与支持向量机融合的人体行为识别. 模式识别与人工智能, 2021, 34(9): 863-870.

DOI 10.16451/j.cnki.issn1003-6059.202109009 **中图法分类号** TP 391.41

Human Action Recognition Fusing Two-Stream Networks and SVM

TONG Anyang[1,2], TANG Chao[1,2], WANG Wenjian[3]

ABSTRACT It is difficult for the traditional two-stream convolutional neural network to understand the long-motion information, and when the long-time stream information is lost, the generalization ability of the model decreases. Therefore, a method for human action recognition fusing two-stream network and support vector machine is proposed. Firstly, RGB images of each frame in the video and their corresponding dense optical flow sequence diagrams in the vertical direction are extracted, and the spatial information and time information of actions in the video are obtained. The information is input into the spatial domain and time domain networks for pre-training, and feature extraction is carried out after pre-training. Secondly, the feature vectors with the same dimension extracted from the two-stream network are fused in parallel to improve the representation ability of feature vectors. Finally, the fused feature vectors are input into the linear support vector machine for training and classification. The experimental results based on the standard open database proves that the classification effect of the proposed method is good.

收稿日期:2021-04-28;录用日期:2021-06-24

Manuscript received April 28, 2021;
accepted June 24, 2021

安徽省自然科学基金项目(No. 2008085MF202)、安徽高校自然科学重点项目(No. KJ2020A0660)、多模态认知计算安徽省重点实验室(安徽大学)开放基金项目(No. MMC202003)资助

Supported by Natural Science Foundation of Anhui Province (No. 2008085MF202), University Natural Sciences Research Project of Anhui Province (No. KJ2020A0660), Open Project of Key Laboratory of Multimodal Cognitive Computation of Anhui University (No. MMC202003)

本文责任编委 陈松灿

Recommended by Associate Editor CHEN Songcan

1. 合肥学院 人工智能与大数据学院 合肥 230601
2. 安徽大学 多模态认知计算安徽省重点实验室 合肥 230601
3. 山西大学 计算机与信息技术学院 太原 030006

1. School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601
2. Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601
3. School of Computer and Information Technology, Shanxi University, Taiyuan 030006

Key Words Two-Stream Network, Support Vector Machine, Feature Fusion, Optical Flow

Citation TONG A Y, TANG C, WANG W J. Human Action Recognition Fusing Two-Stream Networks and SVM. Pattern Recognition and Artificial Intelligence, 2021, 34(9): 863-870.

人体行为识别的目的是通过系列性的观察,从不同的环境中识别不同的行为. 基于视觉的人体行为识别应用领域广泛,包括视频监控[1]、无人驾驶[2]、医疗保健[3]、人机交互[4]等. 然而,在现实生活中,人体动作的动态性、环境的复杂性及物体的尺度变化等,让准确有效的人体行为识别系统依然是计算机视觉中一个具有挑战性的研究领域.

目前,人体行为识别方法主要分为 2 类:基于传统的手工特征提取方法和深度学习方法. 手工特征提取方法通常包括 3 个连续的步骤:特征提取、局部描述子计算、分类[5]. Sullivan 等[6]将边缘信息与标记的关键姿态及部位进行匹配,再根据轮廓信息在连续帧之间进行跟踪. Oikonomopoulos 等[7]提出一种检测器,涉及计算特定时空位置周围圆柱形邻域的熵特征,通过突出运动特征以表示视频中的不同位置. Patrona 等[8]引入自动运动数据和动态运动数据加权,在动作参与的前提下调整人体数据的重要性,实现更有效的动作检测和识别.

由于传统的人工特征提取方法设计复杂,普适性较低,难以提取深层特征,因此,深度学习方法开始替代传统的手工特征提取方法. Karpathy 等[9]将视频分解为连续 RGB 帧,输入卷积神经网络(Convolution Neural Networks, CNN). 然而,单个神经网络难以充分提取视频中的动作信息,Simonyan 等[10]提出双流卷积网络(Two-Stream Convolutional Networks, TCN),有效利用时间流信息,提供多特征提取及融合方法,进一步提高识别的准确率. Ji 等[11]提出 3D 卷积神经网络(3D-CNN),在卷积层中对空间维度和时间维度进行 3D 卷积,得到时空特征,提高性能. 为了让神经网络具有长期记忆,Hochreiter 等[12]提出长短期记忆网络(Long Short-Term Memory, LSTM),使用非线性机制,增强隐藏状态,使用简单的学习门控制函数实现状态传播,对于连续数据建模具有较优效果. Wang 等[13]提出双流 3D-CNN,通过多个特征的互补信息识别任意大小和长度视频中的人类行为. Li 等[14]提出视频长短时记忆网络(VideoLSTM),利用视频中的空间相关性形成注意力映射,通过动作标签定位动作的时空位置,对动作识别及定位具有较优效果.

在传统双流网络中,大都获取视频中目标运动的长时间流信息,而当视频中长时间流信息出现损失时,往往会对识别结果产生较大影响. 为了研究 RGB 图像与短时间流信息在双流网络下特征结合后的动作表征能力,以及进一步提高人体行为识别的准确率,本文提出基于双流网络与支持向量机(Support Vector Machine, SVM)融合的人体行为识别方法. 相比 CNN,在相同数据集上,SVM 训练时间更短,鲁棒性更好. 首先,提取视频中的每帧 RGB 图像及其对应垂直方向的稠密光流序列图,得到视频中动作的空间信息和时间信息,分别输入空间域和时间域网络进行预训练,预训练完成后进行特征提取. 然后,对双流网络提取的维度相同的特征向量执行并联融合策略,提高特征向量的表征能力. 最后,将融合后的特征向量输入线性支持向量机中进行训练及分类处理. 在标准公开数据集 KTH 和 UCF sports 上的实验表明本文方法具有较优的分类效果.

1 基于双流网络与支持向量机融合的人体行为识别

本文提出的基于双流网络与 SVM 融合的人体行为识别方法总体框架如图 1 所示.

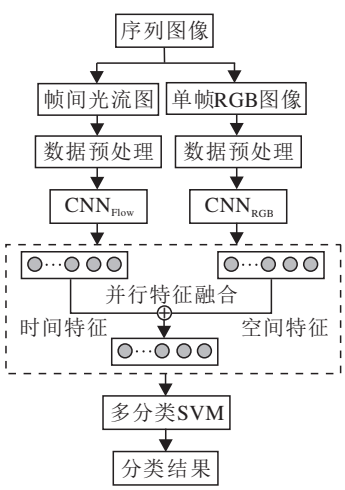


图 1 本文方法框图

Fig. 1 Framework of the proposed method

本文方法主要分为 3 部分:时空特征提取、时空特征融合、支持向量机分类. 首先读取视频中得到的图像,经过预处理后输入 2 个网络,提取时空特征.

然后对 CNN 提取的特征向量进行并联融合. 最后使用融合后的特征向量对线性分类器 SVM 进行训练及分类,设置分类准确率的平均值作为最后识别结果,完成人体行为识别.

1.1　时空特征提取

视频中的动作表征对识别系统的准确性具有重要意义. 本文方法的双流网络结构是以双流模型[10]为基础,按照实验需要进行改进并提取特征. 网络分为空间域和时间域,空间域上输入单帧 RGB 图像,经过灰度化和规范化处理后进行卷积、池化等操作. 训练完成后可提取视频帧中的目标物体在空间上的表征. 在时间域上,将相邻帧间垂直方向的光流图进行同样的数据预处理后输入网络,训练完成后得到动作的短时信息. 两条网络通过得分融合策略获取双流网络的分类结果. 双流卷积神经网络模型结构如图 2 所示.

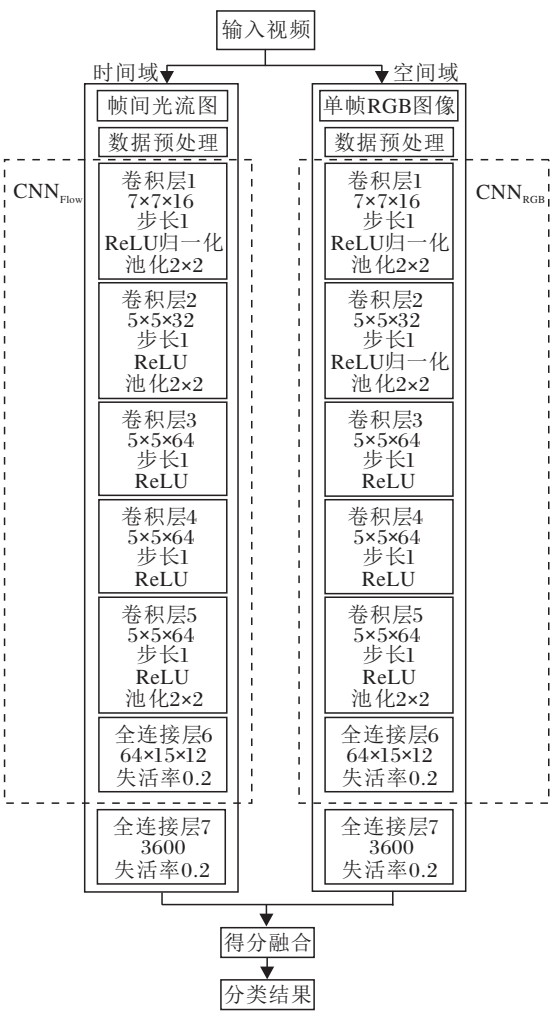


图 2　双流卷积神经网络模型结构

Fig. 2　Two-stream convolutional neural network

1.1.1　空间域

空间域包含数据预处理、5 个卷积层、5 个激活层、3 个池化层、2 个全连接层、2 个失活层、2 次数据归一化处理,采用自适应矩估计(Adaptive Moment Estimation, Adam)优化算法进行网络参数优化,交叉熵损失(Cross Entropy Loss)函数进行损失值的计算. 具体操作如下.

1)利用 CV 库上 COLOR_BGR2GRAY 对输入的 RGB 图像进行灰度化处理,并将图像尺寸规格化为 120 × 160 后输出.

2)卷积层通过多个卷积核对上一层的输入进行卷积运算,并以 0 填充的方式保证输入、输出特征图维度的一致及边缘信息的保留. $x_{i,j}^{l}$ 表示第 l 层 (i,j) 点的输入值,$Y_{i,j}^{l}$ 表示第 l 层输出的第 (i,j) 点的值. 卷积计算过程如下:

$$Y_{i,j}^{l}=f\left(\sum_{i\in m_h}\sum_{j\in n_w}x_{i,j}^{l}G_{i,j}^{l}+b^{l}\right),$$

其中,G 表示卷积核,b^{l} 表示偏置项,$m_h\times n_w$ 表示第 l 层中局部感受野的窗口大小.

3)激活层选择 ReLU 函数对网络进行稀疏化处理,加快训练速度及防止梯度消失,即

$$f(x)=\max(0,x).$$

4)池化层以降低特征面的分辨率为目的,获取空间不变性的特征,提高模型的容错率. 本文选取最大池化法对特征进行二次提取.

5)全连接层可整合卷积层及池化层中具有类别区分性的局部信息[15],解决非线性问题,起到分类的作用.

6)失活层的目的是让神经元以一定的概率失活,防止网络出现过拟合现象. 通过多组实验对比,当设置概率 $P=0.2$ 时,网络在防止过拟合的情况下达到最佳的分类效果.

7)两次数据归一化操作是为了加快模型的收敛速度,提高模型的非线性表达能力.

8)模型选取交叉熵损失函数,衡量真实概率分布与预测概率分布之间的差异性,进行损失值的计算:

$$Loss=-\sum_{i=1}^{n}p(x_i)\ln(q(x_i)),$$

其中,n 为样本类别数,$p(x_i)$、$q(x_i)$ 分别为变量 x_i 对应的真实概率分布和预测概率分布.

9)预训练完成后的空间域网络在第一层全连接层后输出空间特征向量为

$$\boldsymbol{f}_{\text{space}}=[x_1,x_2,\cdots,x_{3600}].$$

1.1.2 时间域

时间域的输入是相邻帧间对应的垂直方向的光流图,选用 Farneback 光流法[16] 进行光流提取.

Farneback 光流法基于多项式展开的连续两帧空间图像稠密光流计算算法,计算图像上所有像素点的运动以提取稠密光流[17]. 主要步骤为多项式展开,使用二次多项式近似每个像素的某个邻域,将 2 幅图像的多项式展开结果用于位移估计. 采用迭代的多尺度位移估计方法获得较优性能. 对于大尺度运动产生的光流,具有较好的提取效果. 将二次多项式 $\boldsymbol{V}_t(\boldsymbol{x})$ 近似为某帧图像的空间信息,在理想状态下进行位移估计:

$$\boldsymbol{V}_t(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}_1\boldsymbol{x} + \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x} + c_1,$$

其中,$\boldsymbol{x}$ 表示二维坐标点$(x,y)^{\mathrm{T}}$,$\boldsymbol{A}$ 表示2×2的对称矩阵,$\boldsymbol{b}$ 表示向量,c 表示标量. $\boldsymbol{V}_t$ 帧图像在经过理想位移后得到下一帧图像:

$$\boldsymbol{V}_{t+1}(\boldsymbol{x}) = \boldsymbol{V}_t(\boldsymbol{x}-\boldsymbol{d}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}_2\boldsymbol{x} + \boldsymbol{b}_2^{\mathrm{T}}\boldsymbol{x} + c_2.$$

在恒定像素值假设的前提下,使 $\boldsymbol{V}_t(\boldsymbol{x})$ 和 $\boldsymbol{V}_{t+1}(\boldsymbol{x})$ 的对应项系数相等,可得

$$\boldsymbol{A}_2 = \boldsymbol{A}_1,$$

$$\boldsymbol{b}_2 = \boldsymbol{b}_1 - 2\boldsymbol{A}_1\boldsymbol{d},$$

$$c_2 = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{A}_1\boldsymbol{d} - \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{d} + c_1.$$

解出每个像素点位移量:

$$\boldsymbol{d} = -\frac{1}{2}\boldsymbol{A}_1^{-1}(\boldsymbol{b}_2 - \boldsymbol{b}_1).$$

计算得到每个像素点位移 $\boldsymbol{d}$ 后,可获得整个图像的光流场 $\boldsymbol{D}(x,y)$.

利用 Farneback 光流法进行稠密光流的提取,将垂直方向的光流信息 $\boldsymbol{D}_{y,i,j}$ 进行灰度化处理,得到 $\boldsymbol{D}_t$. 例如,图 3 为 UCF sports 数据集上相邻视频帧对应的光流图.

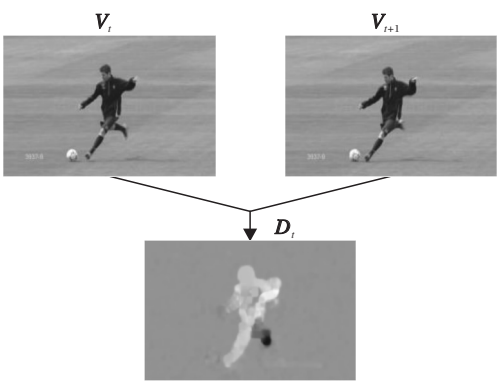


图 3 Farneback 光流法效果示意图

Fig. 3 Illustration of Farneback optical flow method

预处理后的光流图输入卷积层中,利用卷积核组提取动作的多个局部特征图,经过池化等操作,进一步提高特征图的表征能力,最后在全连接层进行分类. 相比空间域,在时间域的处理过程中,数据输入后在卷积层 2 上不进行数据归一化操作. 预训练完成后的时间域网络在第一层全连接层后输出时间特征向量:

$$\boldsymbol{f}_{\mathrm{time}} = [y_1, y_2, \cdots, y_{3600}].$$

1.2 时空特征融合

双流网络经过预训练后分别提取时间特征向量、空间特征向量,而单一特征向量的表征能力不足,训练的模型泛化能力不高. 大量文献[18-19] 表明,多种特征的融合可提升模型泛化能力,表现在局部特征的丢失信息可使用全局特征补偿,底层特征用于补充高级特征. 目前主流的特征融合方法分 3 种:像素级特征融合方法、特征级特征融合方法、决策级特征融合方法[20-21]. 本文通过实验对比,选取特征级特征融合方法中的并行特征融合方法,取代最大值融合方法和拼接融合方法,加强特征向量的表征能力,提高模型的泛化能力.

预训练双流网络后,空间域和时间域分别在第一层全连接层后输出提取的空间特征向量 $\boldsymbol{f}_{\mathrm{space}}$ 和时间特征向量 $\boldsymbol{f}_{\mathrm{time}}$. 并行特征融合后得到的特征向量如下:

$$\boldsymbol{f}_{\mathrm{fusion}} = [x_1 + y_1, x_2 + y_2, \cdots, x_{3600} + y_{3600}].$$

将融合后得到的 $\boldsymbol{f}_{\mathrm{fusion}}$ 放入 SVM 中进行训练和测试,提高识别的准确率.

1.3 支持向量机分类

SVM 是基于结构风险最小化理论的有监督学习算法[22],将低维数据通过核函数投影到高维空间,通过寻找最优分割超平面实现分类. 多分类 SVM 任务的实现可分为 3 类.

1)将多分类任务拆分成多个二分类 SVM 任务实现,如一对一(One-Versus-One, OVO)、一对其余(One-Versus-Rest, OVR) 等.

2)直接寻找最优化问题的解.

3)将 SVM 与其它可实现多分类的机器学习算法结合,如 K 近邻、决策树算法等.

本文选择 OVR 策略,以多个二分类 SVM 完成多分类 SVM 任务. 若数据集上有 N 种类别,分别将其中一个类别设置为正,其余类别设置为负,以此训练 N 种分类器. 若分类结果仅出现一个正,输出正类别;若出现多个分类器结果为正,选择置信度最大的正类别输出.

以 KTH 数据集分类为例,需要分别训练 6 个二分类 SVM,区分类别 m 和其余类别 n,即求解如下问题:

$$\min_{w^{mn},b^{mn},\varepsilon_i^{mn}} \frac{1}{2}\| \boldsymbol{w}^{mn} \|^2 + C\sum_{i=1}^{6}\varepsilon_i^{mn},$$

$$\text{s.t. } y_i[(w^{mn}\cdot\phi(\boldsymbol{f}_{\text{fusion}})) + b^{mn}] \geqslant 1 - \varepsilon_i^{mn},$$

$$\varepsilon_i^{mn} \geqslant 0,\ m = 1,2,\cdots,6,\ n = 1,2,\cdots,6.$$

其中:$\boldsymbol{w}$ 为法向量,决定超平面的方向;b 为位移项;$y_i \in \{m,n\}$, $i = 1,2,\cdots,289\ 116$;函数 ϕ 将输入的低维样本 $\boldsymbol{f}_{\text{fusion}}$ 映射到高位空间;ε_i 为松弛变量,C 为正则化系数.

当训练好 6 个分类器后,若只有一个分类器输出为正,标记该动作为分类结果;若出现多个分类器识别为正,选择置信区间最大的动作为分类结果.

2　实验及结果分析

2.1　实验环境

本文实验在 Windows10 操作系统下进行,选择 Pytorch 框架为深度学习环境. 双流网络设置学习率 $\eta = 0.001$. 一定范围内的批尺寸大小越大,收敛速度越快,训练过程的效率也越高,本文批尺寸大小依据实验硬件条件选择合理范围内的最优值,为 100. 多分类器 SVM 选取 linear 线性核函数,用于提高训练精度和效率. 核函数参数 $\gamma = 10$,惩罚系数 $C = 1$,通过对比实验可证实,相比 C 更大或更小时,SVM 会寻找到间隔最大的决策边界进行分类.

在常用的 2 个标准人体行为识别库——KTH 数据集[23] 和 UCF sports 数据集[24] 上进行有效性实验.

KTH 数据集是用固定摄像机以 25 帧/秒的帧率记录的一个背景简单、视图单一的数据集,共 599 个视频序列. 由 25 名受试者在 4 种场景下进行理想表演,分别为户外、户外规模变化、户外穿着不同衣服、室内. 运动状态分为 Walking, Jogging, Running, Boxing, Hand waving, Hand clapping.

UCF sports 数据集是从 ESPN 和 BBC 等电视频道中收集的 150 个体育视频. 这些动作被记录在真实的运动环境中,展示背景、光照条件和遮挡的变化,成为一个具有挑战性的数据集. 具体运动状态分为:Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skateboarding, Swing-Bench, Swing-Side, Walking.

在实验中,对 KTH、UCF sports 数据集上的视频进行分解,分别得到 289 116 幅、9 282 幅视频帧及对应的光流图. 为了提高方法的泛化能力,降低因数据量不足带来的误差,采用十折交叉验证的方法对数据集进行划分,轮流将十份中的一份作为测试集,剩余九份作为训练集,进行双流网络训练与测试. 同时,为了确保实验的真实可靠性,在 SVM 上进行分类任务时采用三折交叉验证方法,取 3 次测试结果的平均值作为最后结果.

评价标准采用识别准确率:

$$Accuracy = \frac{TP}{Total}\times 100\%,$$

其中,TP 表示分类正确的样本数,$Total$ 表示分类的样本总数. 识别准确率对模型整体的泛化能力有更直观的表达.

2.2　实验结果

实验选取如下对比方法:Two-stream Fusion[25]、CNN KNN+SVM[26]、Multi-stream CNN[27]、双流网络.

各方法在 2 个数据集上的识别准确率对比如表 1 所示. 由表可知,本文方法在 KTH 数据集上的分类效果最优. Two-stream Fusion 构建以牛津大学视觉几何组(Visual Geometry Group, VGG)开发的系列型网络 VGG-16 和 VGG-Medium(VGG-M)为基础的双流网络,设置批尺寸大小为 98,学习率 $\eta = 0.001$,精度达到饱和后 η 降低 90%. 通过强调学习空间、时间网络的卷积层之间的特征关系进行预测,但未避免方法对长时间流信息理解不充分的问题. CNN KNN+SVM 构建 AlexNet 网络,在迁移学习的背景下,使用预训练好的网络结合混合 K 近邻和 SVM 分类的方法,取得较优的识别效果,但未能利用时间信息对动作特征进行表征,在真实场景下的动作预测效果并不显著. Multi-stream CNN 构建以 VGG-Fast(VGG-f)、3D-CNN 等网络为基础的 3 个双流网络,基于整体检测的人体,构建外观和运动流,进行动作分类,但依赖于局部运动部位的变化,未理解整体运动信息,会出现检测到错误目标的情况.

表 1　各方法在 2 个数据集上的识别准确率

Table 1　Accuracies of different methods on 2 datasets

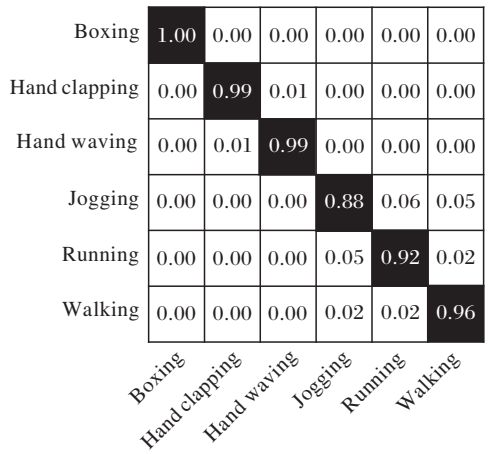
%

方法	KTH	UCF sports
Two-stream Fusion	94.3	–
CNN KNN+SVM	98.15	91.47
Multi-stream CNN	–	97.5
双流网络	95.8	98.2
本文方法	98.4	99.9

KTH、UCF sports 数据集上 2 种方法的行为识别

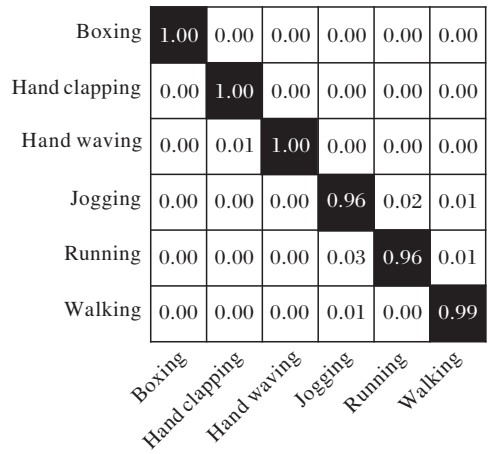
混淆矩阵如图 4、图 5 所示. 混淆矩阵直观表示每个动作的准确性,以及水平方向的真实动作与垂直方向的预测动作之间的对应关系.

由图 4 和图 5 可知,本文方法在 2 个数据集上各类别的准确率均得到显著提升. 在 KTH 数据集的 Boxing、Hand clapping、Hand waving 动作上实现 100% 的分类准确率,其余动作因相似度较高且存在空白帧而出现损失. 在 UCF sports 数据集的全部类别上实现近 100% 的准确率,充分利用空间信息、时间信息对真实场景下的动作进行准确分类.



(a)双流网络

(a)Two-stream network

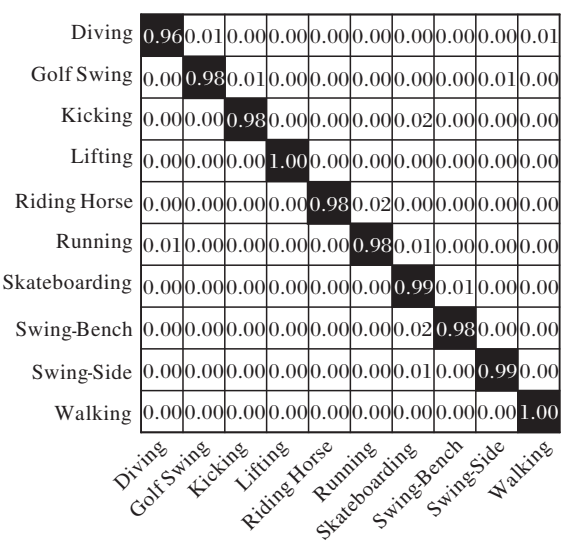


(b)本文方法

(b)The proposed method

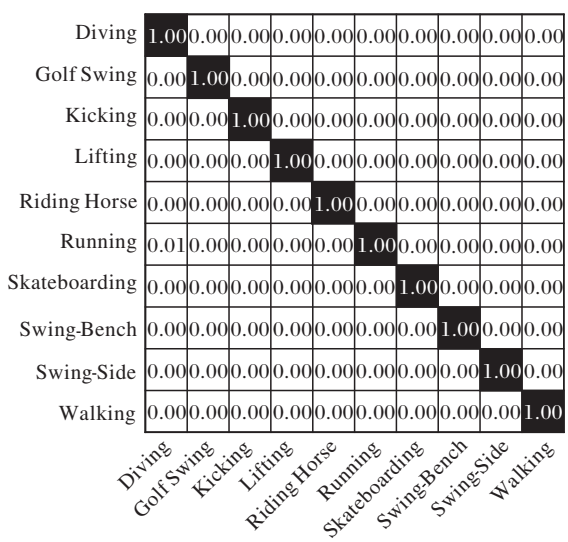
图 4 各方法在 KTH 数据集上的混淆矩阵

Fig. 4 Confusion matrix of different methods on KTH dataset



(a)双流网络

(a)Two-stream network



(b)本文方法

(b)The proposed method

图 5 各方法在 UCF sports 数据集上的混淆矩阵

Fig. 5 Confusion matrix of different methods on UCF sports dataset

3　结 束 语

本文提出基于双流网络与 SVM 融合的人体行为识别方法. 在双流卷积神经网络的基础上进行改进,将提取的空间特征向量、时间特征向量进行融合,采用支持向量机分类. 在小规模数据集 KTH 和 UCF sports 上的分类效果显著,验证本文方法的有效性. 今后考虑将本文方法扩展到更复杂的数据集,如 IXMAS、UCF-50、UCF-101、HMDB-51 数据集. 本文方法在处理视频中空白帧较多和相机剧烈抖动的数据集时,无法提取表征能力较强的特征向量,降低模型的泛化能力,这也是今后需要深入研究的地方.

参 考 文 献

[1] BRÉMOND F, THONNAT M, ZÚÑIGA M. Video-Understanding Framework for Automatic Behavior Recognition. Behavior Research Methods, 2006, 38(3): 416-426.

[2] RAMEZANI M, YAGHMAEE F. A Review on Human Action Analysis in Videos for Retrieval Applications. Artificial Intelligence Review, 2016, 46(4): 485-514.

[3] AZKUNE G, NÚÑEZ-MARCOS A, ARGANDA-CARRERAS I. Vision-Based Fall Detection with Convolutional Neural Networks. Wireless Communications and Mobile Computing, 2017. DOI: 10.1155/2017/9474806.

[4] KOPPULA H S, SAXENA A. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 14-29.

[5] AL-FARIS M, CHIVERTON J, NDZI D, *et al*. A Review on Computer Vision-Based Methods for Human Action Recognition. Journal of Imaging, 2020, 6(6). DOI: 10.3390/jimaging6060046.

[6] SULLIVAN J, CARLSSON S. Recognizing and Tracking Human Action // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2002: 629-644.

[7] OIKONOMOPOULOS A, PATRAS I, PANTIC M. Spatiotemporal Salient Points for Visual Recognition of Human Actions. IEEE Transactions on Systems, Man, and Cybernetics (Cybernetics), 2006, 36(3): 710-719.

[8] PATRONA F, CHATZITOFIS A, ZARPALAS D, *et al*. Motion Analysis: Action Detection, Recognition and Evaluation Based on Motion Capture Data. Pattern Recognition, 2018, 76(11): 612-622.

[9] KARPATHY A, TODERICI G, SHETTY S, *et al*. Large-Scale Video Classification with Convolutional Neural Networks // Proc of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, USA: IEEE, 2014: 1725-1732.

[10] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos // GHAHRAMANI Z, WELLING M, CORTES C, *et al*., eds. Advances in Neural Information Processing Systems 27. Cambridge, USA: The MIT Press, 2014: 568-576.

[11] JI S W, XU W, YANG M, *et al*. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[12] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.

[13] WANG X H, GAO L L, WANG P, *et al*. Two-Stream 3D convNet Fusion for Action Recognition in Videos with Arbitrary Size and Length. IEEE Transactions on Multimedia, 2018, 20(3): 634-644.

[14] LI Z Y, GAVRILYUK K, GAVVES E, *et al*. VideoLSTM Convolves, Attends and Flows for Action Recognition. Computer Vision and Image Understanding, 2018, 166: 41-50.

[15] SAINATH T N, MOHAMED A, KINGSBURY B, *et al*. Deep Convolutional Neural Networks for LVCSR // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Washington, USA: IEEE, 2013: 8614-8618.

[16] FARNEBÄCK G. Two-Frame Motion Estimation Based on Polynomial Expansion // Proc of the Scandinavian Conference on Image Analysis. Berlin, Germany: Springer, 2003: 363-370.

[17] GUAN Q, HUA M, HU H G. A Modified Grabcut Approach for Image Segmentation Based on Local Prior Distribution // Proc of the International Conference on Wavelet Analysis and Pattern Recognition. Washington, USA: IEEE, 2017: 122-126.

[18] DOLLAR P, RABAUD V, COTTRELL G, *et al*. Behavior Recognition via Sparse Spatio-Temporal Features // Proc of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Washington, USA: IEEE, 2005: 65-72.

[19] WANG H, KLÄSER A, SCHMID C, *et al*. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. International Journal of Computer Vision, 2013, 103(1): 60-79.

[20] HALL D L, LLINAS J. An Introduction to Multisensor Data Fusion. Proceedings of the IEEE, 1997, 85(1): 6-23.

[21] YANG J, YANG J Y, ZHANG D, *et al*. Feature Fusion: Parallel Strategy vs. Serial Strategy. Pattern Recognition, 2003, 36(6): 1369-1381.

[22] TONG S, KOLLER D. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 2002, 2: 45-66.

[23] RODRIGUEZ M D, AHMED J, SHAH M. Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2008. DOI: 10.1109/CVPR.2008.4587727.

[24] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing Human Actions: A Local SVM Approach // Proc of the 17th International Conference on Pattern Recognition. Washington, USA: IEEE, 2004, III: 32-36.

[25] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional Two-Stream Network Fusion for Video Action Recognition // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2016: 1933-1941.

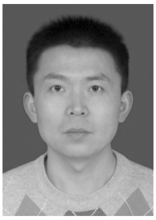
[26] SARGANO A B, WANG X F, ANGELOV P, *et al*. Human Action Recognition Using Transfer Learning with Deep Representations // Proc of the International Joint Conference on Neural Networks. Washington, USA: IEEE, 2017: 463-469.

[27] TU Z G, XIE W, QIN Q Q, *et al*. Multi-stream CNN: Learning Representations Based on Human-Related Regions for Action Recognition. Pattern Recognition, 2018, 79(2): 32-43.

作者简介

童安炀,硕士研究生,主要研究方向为深度学习、计算机视觉. E-mail:1350466625@qq. com.

(**TONG Anyang**, master student. His research interests include deep learning and computer vision.)



唐 超(通信作者),博士,副教授,主要研究方向为机器学习、计算机视觉. E-mail:tangchao77@sina. com.

(**TANG Chao**(Corresponding author), Ph.D., associate professor. His research interests include machine learning and computer vision.)

王文剑,博士,教授,主要研究方向为机器学习、计算智能. E-mail:wjwang@sxu. edu. cn.

(**WANG Wenjian**, Ph.D., professor. Her research interests include machine learning and computing intelligence.)

* *

(上接 862 页)