基于优化卷积神经网络结构的人体行为识别

孙月驰 平 伟* 徐明磊
(山东科技大学计算机科学与工程学院 山东 青岛 266590)

摘 要 为了提高卷积神经网络对非线性特征以及复杂图像隐含的抽象特征提取能力,提出优化卷积神经网络结构的人体行为识别方法。通过优化卷积神经网络模型,构建嵌套 Maxout 多层感知器层的网络结构,增强卷积神经网络的卷积层对前景目标特征提取能力。通过嵌套 Maxout 多层感知器层网络结构可以线性地组合特征图并选择最有效特征信息,获取的特征图经过矢量化处理,分类器 Softmax 完成人体行为识别。仿真实验结果表明,该方法对人体行为识别准确率取得较好结果。

关键词 深度学习 卷积神经网络 特征提取 行为识别

中图分类号 TP319 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.033

HUMAN BEHAVIOR RECOGNITION BASED ON OPTIMIZED CONVOLUTIONAL NEURAL NETWORK STRUCTURE

Sun Yuechi Ping Wei* Xu Minglei
(*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China*)

Abstract To improve the abstract feature extraction ability of convolutional neural networks on nonlinear features and complex images, a human behavior recognition method based on the structure of convolutional neural networks is proposed. By optimizing the convolutional neural network model, the method constructed the network structure of the nested Maxout multi-layer perceptron layer, enhanced the convolutional neural network's convolutional layer to extract the foreground target features, and nests the Maxout multi-layer perceptron layer network structure. The feature map could be linearly combined to select the most effective feature information, and the acquired feature map was subjected to vectorization processing, and the classifier softmax was used for classification and recognition of human behavior. The simulation results show that the method has a good result on the accuracy of human behavior recognition.

Keywords Deep learning Convolutional neural network Feature extraction Behavior recognition

0 引 言

近年来,人体行为识别研究逐渐成熟,在视频的智能分析、虚拟现实、人机交互、视频摘要、视频信息检索、运动分析方面都具有广阔的应用前景[1]。行为识别已经成为深度学习领域的研究热点之一,深度学习算法的研究推动了行为识别研究的进步。

深度学习的基本原理是通过构建具有提取非线性特征的卷积神经网络结构,完成学习、训练过程,提取数据集的本质特征。目前比较成熟的深度学习算法包括对抗神经网络(Generative Adversarial Networks, GAN)、深度置信网络(Deep Belief Network, DBN)、卷积神经网络(Convolutional Neural Network, CNN)等。深度学习在多个领域展现出强大的学习能力和适应能力,深入研究深度学习算法对推动人工智能及拓展其应用具有重要意义。随着运动目标识别技术的广泛应用,如何提高算法的泛化性能和非线性拟合能力,减少冗余特征信息的提取,提升算法对行为识别的准确率,将是未来研究的重点。很多学者采用深度学习算法获

收稿日期:2019-07-24。山东省高等学校学生教育与管理研究项目(2017A007);山东省研究生教育创新计划一般项目(SDYC16022)。**孙月驰**,硕士生,主研领域:智能信息处理。**平伟**,硕士生。**徐明磊**,讲师。

取深层次的特征信息,通过非监督学习方式来学习特征,训练模型进行目标和行为的识别。目前基于深度学习算法的行为识别研究可以分为如下四种类别:

1) 基于卷积神经网络的行为识别。卷积神经网络[2](Convolution neural network,CNN)是深度前馈神经网络的一个分支,在图像识别领域得到广泛的应用,并取得很大成功。CNN 由一维、二维以及三维卷积神经网络组成,分别应用于序列类的数据处理、图像类文本的识别、医学图像以及视频类数据识别。Ji 等[3]构建了一种新的 3D CNN 动作识别方法,通过 3D 卷积层卷积操作分别从空间、时间维度获得特征信息,从搭建的多信息通道中获得输入数据的运动信息,最终的特征表示组合来自所有通道的信息融合。Cheron 等[4]提出了一种新的基于姿势的卷积神经网络描述符(P-CNN)用于动作识别,描述符沿着人体部位的轨迹聚集运动和外观信息,通过研究时间聚合的不同方案,并且对自动估计和手动注释的人体姿势获得的 P-CNN 特征进行了实验,结果表明该模型在识别结果方面表现稳定。Yan 等[5]提出基于卷积神经网络对驾驶员行为识别的方法,首先利用高斯混合模型获取驾驶人员皮肤状区域的特征图像,提取的区域特征图输入深度卷积神经网络,即 R * CNN,以生成动作标签。皮肤状区域能够提供具有足够辨别能力的丰富语义信息。此外,R * CNN 能够从候选者中选择信息最丰富的区域以促进最终动作识别。

2) 基于自动编码器无监督行为识别。自动编码器[6](AutoEncoder)是一种无监督学习算法,该算法通过自动编码获取能够代表输入数据的主要成分,进行复现输入信息的处理。Le 等[7]通过对视频数据进行无监督特征学习,获取视频数据的学习特征,构建了独立子空间分析算法,从未标记的视频数据中学习不变的时空特征。Deng 等[8]提出了一种基于多层自动编码器和监督约束的深度学习方法,能够很好地使用有限的训练图像。

3) 基于受限制玻尔兹曼机及其扩展模型的行为识别。受限玻尔兹曼机[9](Restricted Boltzmann Machine,RBM)是一种基于能量函数能够描述变量之间的相互作用的建模方法,拥有比较健全的数学知识理论基础。Wu 等[10]提出了基于受限玻尔兹曼机器(RBM)及其变体构建的面部形状先验模型,构建一个基于深度信念网络的模型,以捕捉由于近前视图的面部表情变化而导致的脸部形状变化。为了处理姿势变化,将正面形状先验模型结合到三向 RBM 模型中,该模型可以捕获正面形状和非正面形状之间的关系。Feng 等[11]通过用模糊数替换所有实值参数,从限制玻尔兹曼机扩展模糊受限玻尔兹曼机,提出了基于模糊数的脆弱可能均值的新型学习算法,该算法利用模糊数的清晰可能平均值对模糊自由能函数进行去模糊化。

4) 基于递归神经网络及其扩展模型的行为识别。递归神经网络[12](Recursive neural network,RNN)可以分为两类:时间递归神经网络和结构递归神经网络。递归神经网络与传统算法相比,在处理声音、文本、视频等信息表征时,能够反映序列前后关联信息,学习到信息的逻辑顺序。Ng 等[13]构建含有特征池的递归神经网络,特征池网络使用 CNN 独立处理每个帧,然后使用各种池层组合帧级信息。与特征池一样,LSTM 网络在帧级 CNN 激活上运行,并且可以学习如何随时间整合信息。Yu 等[14]提出了一种适用于非常长期跟踪(例如一个月)的多摄像机监视场景的多人跟踪算法,跟踪算法利用身份信息在流形学习框架中用作稀疏标签信息。Du 等[15]提出了一种基于骨架的动作识别的端到端分层递归神经网络模型,根据人体骨骼划分为五部分作为该模型的输入层,分别进入五个子网络提取特征信息,并在高层进行特征的融合,最后由感知器输出结果。

上述基于卷神经网络扩展模型对人体行为识别的研究均需要人工完成特征标记,其计算量、模型的泛化能力,以及特征获取能力需要进一步提高。针对上述问题,本文提出了基于优化的卷积神经网络结构的人体行为识别方法。首先通过优化卷积神经网络模型,构建嵌套 Maxout 多层感知器层(Multy-Layer Perception,MLP)网络结构,卷积层对前景目标进行特征提取,通过嵌套 MaxoutMLP 网络结构可以线性地组合特征图并选择最有效特征信息,对获取的特征图进行向量化处理,利用分类器 Softmax 进行人体行为分类识别。实验结果表明,该方法的人体行为识别准确率取得较好结果。

1　相关理论

卷积神经网络是拥有深层网络结构的前馈神经网络,具有较强的容错、自学习及并行计算能力。近年来卷积神经网络广泛应用于处理分类和识别问题,特别是人脸识别、辅助医疗诊断、自动驾驶系统等领域,极大程度上促进了深度学习快速发展和推广应用。

1.1　卷积神经网络结构特点

卷积神经网络结构中的多层感知器、卷积核、池化

层、局部连接和权值共享等网络结构的应用，不但使神经网络的时间复杂度和空间复杂度得到降低，而且使网络结构的权值参数也大量地缩减，更利于神经网络的训练。

1.1.1 局部连接

局部连接也叫稀疏连接，受生物学中视觉神经结构的启发，视觉皮层的神经元进行局部信息的接收（即这些神经元只响应某些特定区域的刺激）。图像像素的空间联系与距离近的像素相关性强，反之相关性就弱。因此，神经元只接收自己负责的局部感受域而不需要对所有像素进行感知，感知的局部信息再由下一层进行局部信息融合，整合成全局感知。采用局部连接能够很大程度上减少卷积神经网络层与层之间的权值数量，进行特征降维处理并筛选有效的特征，进行神经网络的学习和训练提高模型的学习效率[16]。

1.1.2 权值共享

权值共享实现原理：使用同一个卷积核处理输入的整幅图像，局部提取的特征与其他部分提取的特征是相同的，其他位置都能使用同样的学习特征[17]。卷积神经网络权值共享降低了特征维度和参数数量，神经网络的时间复杂度和空间复杂度也得到降低。

1.1.3 多层卷积核

卷积神经网络的第一层卷积层进行卷积操作之后，卷积层得到的特征图是图像的一些浅层特征，如边缘信息、线条轮廓等信息。对于图像的识别，需要的是深层特征，浅层特征不能够充分表达图像的语义信息。一种卷积核只能获得同一种特征图，要获得更深层次特征，需要进行多层卷积核进行特征信息的提取，形成多种信息的特征图[18]。

在图像识别领域，输入图像的特征层次结构是与生俱来的。如图 1 所示，从原始输入的像素开始，到由像素构成的简单的线条和纹理，再到由线条与纹理构成了图案，最终是由各个图案构成图像中的物体。整个过程通过原始输入找到浅层特征，再对浅层特征进一步挖掘找到中层特征，最后一步获得深层特征。要从原始输入直接找到深层特征无疑是行不通的，简而言之，单层的卷积获取到的往往是浅层的特征，增加卷积的层数才有可能获取到更深层的特征。

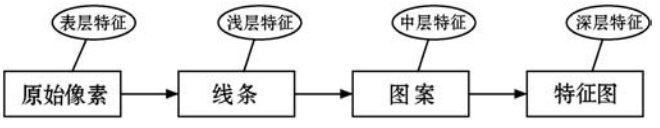

图 1 特征提取过程示意图

1.1.4 卷积过程的原理

卷积神经网络的卷积层对图像进行卷积操作，获取的特征图含有原始图像的结构性特征，深层次的特征更能够表达出图像本质含义信息。函数卷积的定义如下：对于 **R** 上可积的两个连续函数 $f(x)$、$g(x)$，它们的卷积 $h(x)$ 为：

$$h(x) = \int_{-\infty}^{\infty} f(\tau) g(x-\tau) \mathrm{d}\tau \tag{1}$$

式中：$f(x)$ 在 $g(x)$ 上的卷积记作 $f(x) * g(x)$，表示在定义域中 $f(x)$ 与 $g(\alpha - x)$ 乘积的积分；α 代表卷积函数 $h(x)$ 的自变量，即卷积的位置。

卷积计算过程即把图片转换成数据矩阵，游走的窗口为卷积核矩阵，一个 $N \times N$ 的图像经过 $M \times M$ 的卷积核卷积处理之后，将得到 $(N-M+1) \times (N-M+1)$ 的特征图。

1.2 Softmax 分类器

Logistic 回归模型的推广应用形成 Softmax 分类器用来解决多分类问题，本文优化的卷积神经网络模型使用 Softmax 分类器对行为进行分类处理。假设将异常行为分为 k 个，并对 k 个行为进行分类，样本数据视频序列有 m 个，视频序列的样本维度为 n。假设卷积神经网络训练数据集为 T：

$$T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(m)}, y^{(m)})\} \tag{2}$$

式中：$\boldsymbol{x}^{(i)}$ 为第 i 个输入样本；$y^{(i)}$ 为第 i 个样本的行为标签，$y^{(i)} \in \{1, 2, \cdots, k\}$。

对于每个输入 $\boldsymbol{x}^{(i)}$，Softmax 分类器会计算对应每个类的概率，计算公式如下：

$$P(y = j \mid \boldsymbol{x}) \quad y = 1, 2, \cdots, k \tag{3}$$

从向量角度来看，计算函数的公式如下：

$$f(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}) = \begin{bmatrix} p(y^{(i)} = 1 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \\ p(y^{(i)} = 2 \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \\ \vdots \\ p(y^{(i)} = k \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} \mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{x}^{(i)}}} \begin{bmatrix} \mathrm{e}^{\boldsymbol{\theta}_1^{\mathrm{T}} \boldsymbol{x}^{(i)}} \\ \mathrm{e}^{\boldsymbol{\theta}_2^{\mathrm{T}} \boldsymbol{x}^{(i)}} \\ \vdots \\ \mathrm{e}^{\boldsymbol{\theta}_k^{\mathrm{T}} \boldsymbol{x}^{(i)}} \end{bmatrix} \tag{4}$$

式中：$\boldsymbol{\theta}$ 表示神经网络参数。可见，行为有 k 个，每个行为对应一个概率值，概率的取值范围在 $[0,1]$ 之间，k 个异常行为的概率和为 1。神经网络的输出对应行为的概率以及该概率对应行为的标签。

神经网络训练过程中利用 Softmax 进行行为分类，损失函数计算公式如下：

$$L(\boldsymbol{\theta}) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{x}^{(i)}}}{\sum_{l=1}^{k} \mathrm{e}^{\boldsymbol{\theta}_l^{\mathrm{T}} \boldsymbol{x}^{(i)}}} \tag{5}$$

式中：$1\{y^{(i)} = j\}$ 表示指示函数，当 $y^{(i)}$ 与 j 相等时，输出为 1，反之，输出为 0，其输出为异常行为的标签矩阵。

通常情况下,利用梯度下降算法计算反向传播过程中损失函数,计算公式如下:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{1}{m}\boldsymbol{x}^{(i)}[1\{y^{(i)}=j\} - p(y^{(i)}=j \mid \boldsymbol{x}^{(i)};\boldsymbol{\theta})] \tag{6}$$

利用式(6)得到损失函数对权值参数的梯度,利用该梯度指导神经网络模型参数调整,直至神经网络训练结束并得到最佳的权值参数。

2　卷积神经网络结构构建

传统 CNN 在卷积层使用单层线性卷积,对非线性特征的提取和复杂图像隐含的抽象特征提取表现不突出。激活函数具有强大的拟合能力,在神经元数量足够的情况下,能够拟合所有特征模式,因此采用嵌套 MaxoutMLP 层与激活函数组合来提升算法的拟合能力,提高模型的识别准确率。

2.1　嵌套层数的确定

嵌套 Maxout 层的神经网络中线性区域的数量随着 Maxout 层的数量增加而增加,此外激活函数 ReLU 和 Maxout 网络中的线性区域的数量随着层数呈指数增长[19]。Maxout 网络在没有模型正则化的情况下容易过度拟合训练数据集,归因于 Maxout 网络在训练过程中能够识别输入的最有价值信息,并且易于进行特征共同适应[20]。

使用不同数量的 Maxout 层片段在数据集上测试了本文方法,如图 2 所示。不同数量的 Maxout 片段与使用 Maxout 层与批量归一化(Batch Normaliztion,BN)层片段组合测试结果,当 Maxout 片段为 5 时嵌套模型已经达到饱和状态。

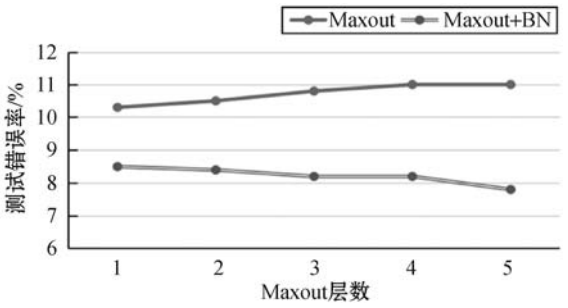

图 2　不同数量的 Maxout 层测试结果

2.2　池化层的选择

一般情况下,研究者会选择最大池化层进行下采样,最大池化层在提取特征方面更具代表性。在所有汇集层中使用平均池化汇聚有效特征,输入图像中的无关特征信息可以通过平均池化来抑制,并通过最大合并来丢弃。平均池是全局平均池的扩展,其中模型试图从每个本地补丁中提取信息以便于抽象到特征映射。嵌套结构能够从每个局部中获取抽象的代表性信息,使得更多可辨别的信息嵌入特征映射中,在每个池化层中使用空间平均池来聚合局部空间信息。在无数据扩充的 CIFAR-10 数据集上,最大、平均池化层测试错误率比较结果如表 1 所示。

表 1　最大、平均池化层测试错误率比较

采用的池化层	测试错误率/%
Max pooling	8.78
Avg pooling	7.85

2.3　构建嵌套层

嵌套多层 Maxout 网络的卷积层,即基于嵌套网络结构使用 MaxoutMLP 进行特征提取,构建的卷积神经网络模型使用批量标准化来降低饱和度并使用压差来防止过度拟合。此外,为了增加对象空间转换的稳健性,在所有池层中应用平均池以聚合 MaxoutMLP 获得的基本特征。

$$\boldsymbol{f}_{i,j,k} = \max_{m\in[1,n]}(\boldsymbol{w}_{k_m}^{\mathrm{T}}\boldsymbol{x}_{i,j} + \boldsymbol{b}_{k_m}) \tag{7}$$

式中:(i,j)是特征图中像素的位置;$\boldsymbol{x}_{i,j}$是以像素点(i,j)为中心的输入块;k_m是用于索引特征映射的通道$f_{i,j,k}$;n 则是 MLP 的层数。从另一个角度来看,Maxout 单位相当于卷积层上的跨通道最大池化层,跨通道最大池化层选择要输入下一层的最大输出。Maxout 单元有助于解决渐变消失的问题,因为渐变能够流过每个最大单元。

嵌套 Maxout MLP 层模块中的特征映射计算如下:

$$\boldsymbol{f}^{1}_{i,j,n_1} = BN((\boldsymbol{w}^{1}_{n_1})^{\mathrm{T}}\boldsymbol{x}_{i,j} + \boldsymbol{b}^{1}_{n_j}) \tag{8}$$

$$\boldsymbol{f}^{2}_{i,j,n_2} = \max_{m\in[1,k]}(BN((\boldsymbol{w}^{2}_{n_m})^{\mathrm{T}}\boldsymbol{f}^{1}_{i,j} + \boldsymbol{b}^{2}_{n_m})) \tag{9}$$

$$\boldsymbol{f}^{3}_{i,j,n_3} = \max_{m\in[1,k]}(BN((\boldsymbol{w}^{3}_{n_m})^{\mathrm{T}}\boldsymbol{f}^{2}_{i,j} + \boldsymbol{b}^{3}_{n_m})) \tag{10}$$

式中:$BN(\cdot)$表示批量归一化层;(i,j)是特征图中像素的位置;$x_{i,j}$是以像素点(i,j)为中心的输入块;k_n等是特征图中的各通道序号;n 则是嵌套 Maxout MLP 的层数。批量标准化层可以在激活函数之前应用,在这种情况下,非线性单元倾向于产生具有稳定分布的激活,降低饱和度。如图 3 所示,构建嵌套 Maxout 层的卷积层结构图。

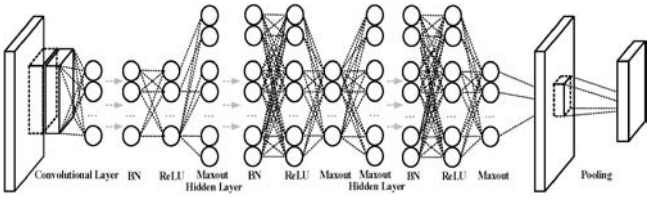

图 3　构建嵌套 Maxout 层的卷积层结构图

2.4 嵌套 Maxout 层的卷积神经网络模型

通过叠加四个嵌套 Maxout 层的卷积层模块形成本文嵌套 MaxoutMLP 层的卷积神经网络整体结构,如图 4 所示。

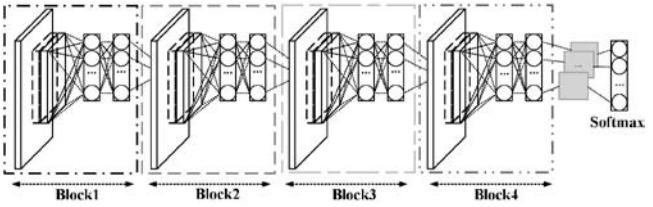

图 4 嵌套 Maxout 层的卷积神经网络整体结构

嵌套 MaxoutMLP 层的网络结构相当于级联的跨通道参数池和卷积层上的跨通道最大池,嵌套结构可以线性地组合特征图并选择最有效信息的组合输出到下一层。嵌套结构通过应用批量归一化来降低饱和度,并且可以对路径或 Maxout 碎片的激活模式中的信息进行编码,增强卷积神经网络深层架构的辨别能力。

3 神经网络训练

神经网络模型的训练过程采用误差反向传播算法,训练过程分为正向传播阶段和反向传播阶段。神经网络训练的正向传播为神经网络的各隐含层收到上一层的输出,利用激活函数激活计算出该层的输出;神经网络训练的反向传播阶段为利用损失函数计算神经网络的输出误差,并逐层向前传播计算神经网络各隐藏层的误差,各隐含层的误差作为前一隐含层权值参数的更新依据。神经网络算法的训练步骤,如图 5 所示。

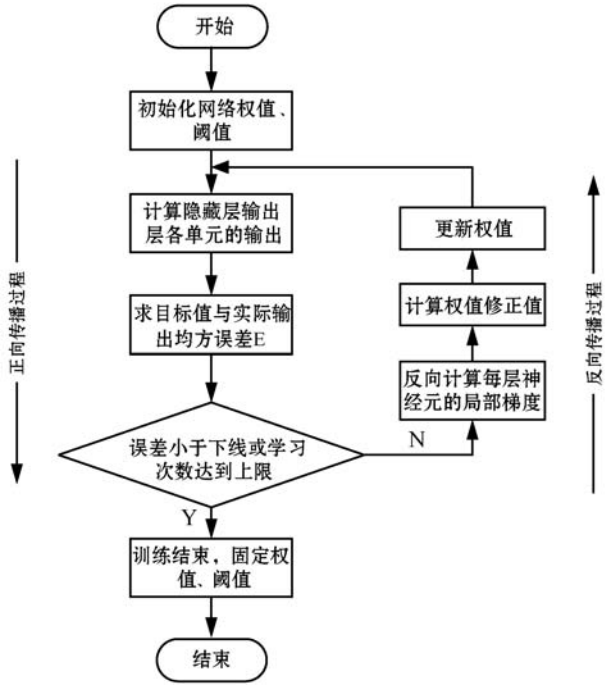

图 5 神经网络算法的训练步骤

步骤 1 随机初始化网络中的所有权值和阈值,取值范围为(-1,1)。

步骤 2 对训练样本$(\boldsymbol{x}_i,\boldsymbol{y}_i)$,计算网络的实际输出计算公式如下:

$$\hat{\boldsymbol{y}}_j^i = f(\boldsymbol{\beta}_j - \boldsymbol{\theta}_j) \tag{11}$$

式中:$f(\cdot)$表示激活函数 Sigmoid 函数;$\boldsymbol{\theta}_j$表示神经网络输出层第 j 个神经元的阈值;$\boldsymbol{\beta}_j$表示神经网络输出层第 j 个神经元的输入。

$$\boldsymbol{\beta}_j = \sum_{i=1}^{n} w_{ij}\boldsymbol{x}_i \tag{12}$$

式中:$w_{i,j}$表示神经网络隐含层第 i 个神经元与神经网络输出层第 j 个神经元之间的权重。

步骤 3 对卷积神经网络在$(\boldsymbol{x}_i,\boldsymbol{y}_i)$上的均方误差进行计算,计算公式如下:

$$E = \frac{1}{2}\sum_{j=1}^{m}(\hat{\boldsymbol{y}}_j^i - \boldsymbol{y}_j^i)^2 \tag{13}$$

式中:$\hat{\boldsymbol{y}}_j^i$表示卷积神经网络的实际输出;$\boldsymbol{y}_j^i$表示卷积神经网络的期望输出。

步骤 4 判断是否达到介绍条件,即误差是否小于学习误差允许的最小值或者学习次数达到设置的最低次数。若未满足条件,则进行卷积神经网络的权值更新,神经网络的权值和阈值等参数依据目标的梯度方向调整。设神经网络训练过程的学习率为 η,神经网络的权值更新计算公式如下:

$$\Delta w_{ij} = -\eta\frac{\partial E}{\partial w_{ij}} = \eta\hat{\boldsymbol{y}}_j^i(1-\hat{\boldsymbol{y}}_j^i)(\boldsymbol{y}_j^i - \hat{\boldsymbol{y}}_j^i)\boldsymbol{x}_i \tag{14}$$

步骤 5 重复步骤 2 - 步骤 4,直到满足结束条件为止,神经网络训练过程结束,即神经网络训练完成,固定神经网络的权值、阈值。

4 实 验

4.1 实验环境

在 Intel(R) Core(TM) i5-2450M,3.0 GHz CPU、64 位 Windows 7 操作系统,采用 Open CV、Python 2.7 作为开发工具在 UCF-YouTube、KTH 两个数据库上进行本文算法实验验证。

4.2 KTH 数据集上的实验

KTH 数据集包含 25 个表演者在四个不同的场景下的 6 类动作包括:走、慢跑、跑、拳击、挥手、拍掌等,涉及的四个场景分别为:室外场景、室外且包含尺度变化、室外且有着装变化以及室内场景,如图 6 所示。选取每个场景不同动作的 25 个动作对象进行研究,20 个动作对象进行训练,5 个动作对象来测试。

图 6　KTH 数据集中人体行为示例图

本文方法与文献[21 – 23]中方法在 KTH 数据集上对每个行为类别识别率的对比结果分析如图 7 所示。本文方法在“跑”“慢跑”“拍掌”人体动作识别准确率上都有提高。

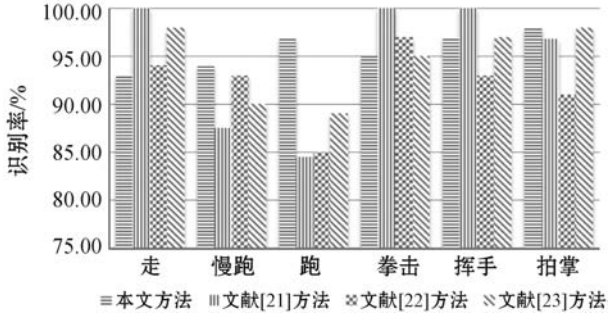

图 7　本文方法与其他方法在 KTH 数据上的识别率对比

结果显示,本文方法在 KTH 数据集上的准确识别率达到 95.6%,实验结果的混淆矩阵如图 8 所示。KTH 数据集上不同方法的平均准确率的比较如表 2 所示。本文方法比文献[24]提出改进的 3D CNN 的方法在识别准确率上提高了 1.8 个百分点;比文献[25]提出的 3D CNN 方法在识别准确率上提高了 5.4 个百分点;比文献[26]提出的最大边缘 HCRF 方法在识别准确率上提高了 3.1 个百分点;比文献[27]提出的局部三元模式 SVM 方法在识别准确率上提高了 5.5 个百分点;比文献[28]提出的空时词组核 SVM 方法在识别准确率上高出 1.0 个百分点。

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	0.93	0.05	0.02	0.00	0.00	0.00
Jogging	0.03	0.94	0.01	0.00	0.00	0.00
Running	0.01	0.02	0.97	0.00	0.00	0.00
Boxing	0.00	0.00	0.00	0.95	0.02	0.03
Waving	0.00	0.00	0.00	0.01	0.97	0.03
Clapping	0.00	0.00	0.00	0.01	0.01	0.98

图 8　本文算法在 KTH 数据集上的混淆矩阵

表 2　KTH 数据集上不同方法的平均准确率的比较

文献	使用方法	平均识别率/%
文献[24]	改进的 3D CNN	93.8
文献[25]	3D CNN	90.2
文献[26]	最大边缘 HCRF	92.5
文献[27]	局部三元模式 SVM	90.1
文献[28]	空时词组核 SVM	94.6
本文	CNN + GMM	95.6

4.3　数据集 UCF-YouTube 上的实验

UCF-YouTubeAction dataset 是一个人类动作视频数据集,包括 11 个动作类:篮球投篮、自行车、潜水、高尔夫挥杆、骑马、足球杂耍、荡秋千、网球、蹦床上跳来跳去、排球扣球和遛狗,如图 9 所示。视频被分为 25 组,其中有超过 4 个动作片段。同一组中的视频片段具有相同的特征,如相同的演员、相似的背景、相似的视角等。

图 9　UCF-YouTube 数据集中人体行为示例图

本文方法在 UCF-YouTube 数据集上达到 88.5% 识别精度,实验结果的混淆矩阵如图 10 所示,在 UCF-YouTube 数据集上不同方法的平均准确率比较如表 3 所示。本文方法比文献[28]提出的基于词组核的 SVM 方法在识别准确率上高出 15.6 个百分点;比文献[29]提出的扩散图内嵌方法在识别准确率上高出 12.4 个百分点;比文献[30]提出的基于 BOW 的 SVM 方法在识别准确率上高出 3.1 个百分点。

	Basketball shooting	Swinging	Biking	Diving	Walking with a dog	Soccer juggling	Volleyball spiking	Trampoline jumping	Horse riding	Tennis swinging	Golf swinging
Basketball shooting	0.85	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.00
Swinging	0.00	0.92	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.04	0.00
Biking	0.00	0.04	0.84	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.00
Diving	0.04	0.00	0.00	0.86	0.05	0.00	0.00	0.00	0.00	0.05	0.00
Walking with a dog	0.04	0.00	0.04	0.00	0.83	0.00	0.00	0.00	0.05	0.00	0.04
Soccer juggling	0.00	0.04	0.00	0.04	0.00	0.88	0.00	0.00	0.04	0.00	0.00
Volleyball spiking	0.00	0.00	0.04	0.00	0.00	0.00	0.93	0.00	0.03	0.00	0.00
Trampoline jumping	0.00	0.03	0.00	0.00	0.02	0.00	0.00	0.95	0.00	0.00	0.00
Horse riding	0.04	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.92	0.00	0.00
Tennis swinging	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.90	0.00
Golf swinging	0.00	0.04	0.00	0.04	0.04	0.00	0.00	0.04	0.00	0.00	0.80

图 10　本文算法在 UCF-YouTube 数据集上的混淆矩阵

表 3 UCF-YouTube 数据集上不同方法的平均准确率的比较

文献	使用方法	平均识别率/%
文献[28]	基于词组核的 SVM	72.9
文献[29]	扩散图内嵌	76.1
文献[30]	基于 BOW 的 SVM	85.4
本文	CNN + GMM	88.5

5 结 语

本文提出一种基于优化卷积神经网络结构的人体行为识别算法,通过嵌套 Maxout MLP 层的网络结构,提高了神经网络对非线性特征以及复杂图像隐含的抽象特征提取能力。嵌套层中使用激活函数 ReLU,在神经元数量足够的情况下,能够提升神经网络拟合特征模式的性能,嵌套结构使用批量归一化来降低饱和度,并且可以对路径或 Maxout 碎片的激活模式中的信息进行编码,增强卷积神经网络深层架构的辨别能力。在 UCF-YouTube、KTH 数据库上进行实验验证,该模型在泛化性能和非线性拟合能力两方面都有所提高,与传统方法和传统 CNN 方法比较,取得了较高的识别率。

参 考 文 献

[1] 单言虎,张彰,黄凯奇. 人的视觉行为识别研究回顾、现状及展望[J]. 计算机研究与发展,2016,53(1):93 - 112.

[2] Sainath T N, Kingsbury B, Saon G, et al. Deep convolutional neural networks for large-scale speech tasks[J]. Neural Networks,2014,64:39 - 48.

[3] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1): 221 - 231.

[4] Cheron G, Laptev I, Schmid C. P-CNN: Pose-based CNN features for action recognition[C]//2015 IEEE International Conference on Computer Vision(ICCV). IEEE,2015:3218 - 3226.

[5] Yan S Y, Teng Y X, Smith J S, et al. Driver behavior recognition based on deep convolutional neural networks[C]// 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE,2016:636 - 641.

[6] Kampffmeyer M, Løkse S, Bianchi F M, et al. The deep kernelized autoencoder[J]. Applied Soft Computing,2018, 71: 816 - 825.

[7] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2011:3361 - 3368.

[8] Deng S, Du L, Li C, et al. SAR automatic target recognition based on Euclidean distance restricted autoencoder[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2017,10(7):3323 - 3333.

[9] 朱煜,赵江坤,王逸宁,等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报,2016,42(6):848 - 857.

[10] Wu Y, Wang Z G, Ji Q. Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machines[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2013:3452 - 3459.

[11] Feng S, Chen C L P. A fuzzy restricted Boltzmann machine: novel learning algorithms based on the crisp possibilistic mean value of fuzzy numbers[J]. IEEE Transactions on Fuzzy Systems,2018,26(1):117 - 130.

[12] Gregor K, Danihelka I, Graves A, et al. DRAW: a recurrent neural network for image generation[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. ACM,2015:1462 - 1471.

[13] Ng J Y, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2015:4694 - 4702.

[14] Yu S I, Yang Y, Li X C, et al. Long-term identity-aware multi-person tracking for surveillance video summarization [EB]. arXiv:1604.07468v2,2016.

[15] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2015:1110 - 1118.

[16] 钱坤,王天真,马斌,等. 一种新型的局部连接 BP 网络模型及应用[J]. 系统科学与数学,2014,34(7):792 - 804.

[17] 刘威,刘尚,白润才,等. 互学习神经网络训练方法研究[J]. 计算机学报,2017,40(6):1291 - 1308.

[18] 常亮,邓小明,周明全,等. 图像理解中的卷积神经网络[J]. 自动化学报,2016,42(9):1300 - 1312.

[19] Montúfar G, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks[C]//2014 28th Annual Conference on Neural Information Processing Systems, 2014: 2924 - 2932.

[20] Liao Z B, Carneiro G. On the importance of normalisation layers in deep learning with piecewise linear activation units [C]//2016 IEEE Winter Conference on Applications of Computer Vision(WACV). IEEE,2016:1 - 8.

[21] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2011:3169 - 3176.

(下转第 269 页)

推荐算法[J]. 计算机应用与软件,2017,34(4):305-308.

[8] 吴宾,娄铮铮,叶阳东. 一种面向多源异构数据的协同过滤推荐算法[J]. 计算机研究与发展,2019,6(5):1034-1047.

[9] Hu Y,Shi W S,Li H,et al. Mitigating data sparsity using similarity reinforcement-enhanced collaborative filtering[J]. ACM Transactions on Internet Technology,2017,17(3):31.

[10] Zhang L M,Ma J F,Lu D,et al. Attribute clustering based collaborative filtering[C]//2014 International Conference on Advances in Materials Science and Information Technologies in Industry (AMSITI),2014:965-968.

[11] Huang Z,Chen H,Zeng D,et al. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information Systems,2004,22(1):116-142.

[12] Patra B K,Launonen R,Ollikainen V,et al. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data[J]. Knowledge Based Systems,2015,82(C):163-177.

[13] 黄震华,张佳雯,张波,等. 语义推荐算法研究综述[J]. 电子学报,2016,44(9):2262-2275.

[14] Gradgyenge L,Kiss A,Filzmoser P,et al. Graph embedding based recommendation techniques on the knowledge graph[C]//Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. ACM,2017:354-359.

[15] 陈鹤. 基于语义本体的社交网络服务推荐系统[D]. 长春:吉林大学,2014.

[16] Chen H,Zhang M F. Improve tagging recommender system based on tags semantic similarity[C]//2011 IEEE 3rd International Conference on Communication Software and Networks. IEEE,2011:94-98.

[17] Shambour Q,Lu J. A Hybrid multi-criteria semantic-enhanced collaborative filtering approach for personalized recommendations[C]//2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. IEEE,2011:71-78.

[18] 王俊红. 基于语义网的农业学习资源推荐系统研究[J]. 计算机应用与软件,2013,30(8):233-235.

[19] Wang H,Shi X J,Yeung D Y. Relational stacked denoising autoencoder for tag recommendation[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. ACM,2015:3052-3058.

[20] 常亮,张伟涛,古天龙,等. 知识图谱的推荐系统综述[J]. 智能系统学报,2019,14(2):207-216.

[21] 吴玺煜,陈启买,刘海,等. 基于知识图谱表示学习的协同过滤推荐算法[J]. 计算机工程,2018,44(2):226-232.

[22] Bordes A,Usunier N,Garcia-Duran A,et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. ACM,2013:2787-2795.

[23] 方阳,赵翔,谭真,等. 一种改进的基于翻译的知识图谱表示方法[J]. 计算机研究与发展,2018,55(1):139-150.

[24] Harper F M,Konstan J A. The movieLens datasets:History and context[J]. ACM Transactions on Interactive Intelligent Systems,2015,5(4):19.

[25] Ziegler C,McNee S M,Konstan J A,et al. Improving recommendation lists through topic diversification[C]//Proceedings of the 14th International Conference on World Wide Web. ACM,2005:22-32.

[26] Wang H W,Zhang F Z,Wang J L,et al. RippleNet:propagating user preferences on the knowledge graph for recommender systems[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM,2018:417-426.

(上接第 204 页)

[22] Minhas R, Mohammed A A, Wu Q M J. Incremental learning in human action recognition based on snippets[J]. IEEE Transactions on Circuits & Systems for Video Technology,2012,22(11):1529-1541.

[23] Wu X X, Xu D, Duan L X, et al. Action recognition using context and appearance distribution features[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2011:489-496.

[24] 范晓杰,宣士斌,唐凤. 基于 Dropout 卷积神经网络的行为识别[J]. 广西民族大学学报(自然科学版),2017,23(1):76-82.

[25] Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013,35(1):221-231.

[26] Wang Y, Mori G. Max-margin hidden conditional random fields for human action recognition[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2009:872-879.

[27] Yeffet L, Wolf L. Local trinary patterns for human action recognition[C]//2009 IEEE 12th International Conference on Computer Vision. IEEE,2009:492-497.

[28] Zhang Y, Liu X, Chang M C, et al. Spatio-temporal phrases for activity recognition[C]//The 2012 European Conference on Computer Vision. Springer,2012:707-721.

[29] Liu J, Yang Y, Shah M. Learning semantic visual vocabularies using diffusion distance[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2009:461-468.

[30] Wang H, Klaser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision,2013,103:60-79.