

基于改进多尺度深度卷积网络的手势识别算法



景雨 祁瑞华 刘建鑫 刘朝霞

大连外国语学院软件学院 辽宁 大连 116044

摘要 基于传统的浅层学习网络由于过度依赖于人工选择手势特征,因此不能实时适应复杂多变的自然场景。在卷积神经网络架构的基础上,提出了一种改进的多尺度深度网络手势识别模型,该模型能够利用卷积层自动学习手势特征,进而除去人工提取特征的弊端。该方法引入自适应多尺度特性来实现同一卷积层不同尺寸卷积核生成不同尺度特征,并通过级联浅层和深层的特征来达到不同抽象程度的特征图融合。同时,为了增强模型的泛化能力,提出了基于正则化约束的损失函数。实验结果表明,所提网络模型的识别精度高于普通单尺度卷积神经网络结构的识别精度,弥补了提取特征不够精细、全面及稳定性欠佳等缺点,同时网络训练所需的时间并没有大幅度增加。

关键词 手势识别;深度学习;多尺度;卷积特征;正则化;损失函数

中图分类号 TP391

Gesture Recognition Algorithm Based on Improved Multiscale Deep Convolutional Neural Network

JING Yu, QI Rui-hua, LIU Jian-xin and LIU Zhao-xia

School of Software, Dalian University of Foreign Languages, Dalian, Liaoning 116044, China

Abstract Since the traditional shallow learning networks rely too much on manual selection of gesture features, they cannot adapt to complex and varied natural scenes in real time. Based on the convolutional neural network architecture, this paper proposes an improved multi-scale deep network gesture recognition model, which makes it possible to overcome the drawbacks of manual extraction features by using the convolutional layer to automatically learn gesture features. In this method, the adaptive multi-scale features are introduced to realize that convolution kernels with different sizes at the same convolutional layer to generate different scale features, and achieves feature map fusion with different levels by cascading shallow and deep features. In addition, in order to enhance the generalization ability of the model, this paper proposes a loss function based on regularization constraints. The experimental results show that the recognition accuracy of the proposed network model is higher than that of the ordinary single-scale convolutional neural network, and the shortcomings of imprecise and incomprehensive extraction as well as poor stability are overcome, and the time required for network training is not greatly increased.

Keywords Gesture recognition, Deep learning, Multi-scale, Convolution feature, Regularization, Loss function

1 引言

随着硬件技术、信号处理技术的飞速发展,人机交互已经在日常生活中被频繁应用^[1]。现有的智能硬件已经可以利用可穿戴数据套件直接感知姿势与动作。然而,这类应用需要辅助套件进行信息采集,限制了交互信息的自然表达^[2]。通过视觉对手势进行识别,可以实现各种复杂信息的交互,该技术不需要复杂的数据采集设备,人机交互也就更加自然^[3]。

经过多年的发展,传统的手势识别算法因为各种原因而识别精度不高^[4-7]。Zhang等^[8]提出了一种基于特征包智能学习的手势识别方法,通过机器学习算法对复杂图像进行智能分析,找到潜在的手势区域,进而完成相应的高精度分类。

可以看出,通过直接对分割区域进行特征提取,再采用智能算法进行手势判断,虽然能够在一定程度上提高模型分类的准确率,但特征的表征能力直接影响系统的识别精度。相比支持向量机(Support Vector Machine, SVM)、决策树、神经网络等浅层学习模型的检测算法^[9],以卷积神经网络为代表的深度学习强调了模型结构的深度,更突出了手势特征学习的重要性,利用大数据来学习特征,更能表征数据丰富的特征信息。Wang等首次将卷积神经网络运用到手势识别领域,通过设计一种更深层次的网络结构,并采用Dropout技术来增强模型的泛化能力,其检测精度超过传统的浅层学习模型。深度模型本质上是一种深层非线性网络结构,通过网络中多层次节点将低层特征组合成更加抽象的高层特征,实现复杂

到稿日期:2020-02-05 返修日期:2020-04-05 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61501082);大连外国语学院科研基金项目(2018XJYB27);辽宁省教育厅科学研究一般项目(2019JYT01, 2019JYT07);辽宁省自然科学基金项目(20180550018);辽宁省博士科研启动基金项目(2019BS061)

This work was supported by the National Natural Science Foundation of China(61501082), Scientific Research Fund Project of Dalian University of Foreign Languages (2018XJYB27), Scientific Research Project of Liaoning Province Educational Department (2019JYT01, 2019JYT07), Natural Science Foundation Project of Liaoning Province(20180550018) and Doctoral Research Start Fund Project of Liaoning Province (2019BS061).

通信作者:景雨(jingyu0814@126.com)

函数逼近,具有强本质特征的学习能力^[10-11]。

随着深度学习的迅速发展与广泛应用,现有的 Alex-Net^[12]、GoogleNet^[13]、残差网络(ResNet)^[14]等深度学习算法已经在图像分类领域取得了很好的成果,具有良好的应用前景,但也存在网络计算量大、模型复杂和实时性不高等缺点。同时,现有的深度学习网络结构只利用高层次特征进行图像的分类识别,其难以区分需要精细特征才能识别的目标,如手势类别、车辆型号等。Wu 等^[15]在 LeNet-5 网络的基础上设计了双通道 CNN 的静态手势识别方法。此方法不用人工提取特征,通过训练网络能自动学习特征,但提取不够精细,精度也不够高。在静态手势识别中采用常见的单一特征卷积神经网络很难取得很好的识别效果。因此,现有的深度学习模型不能直接用于手势识别。

针对自然状态下的手势特征,本文在多尺度深度模型的基础上提出了一种改进的多尺度深度学习网络,通过提取到的不同尺度的特征来更加准确地表征图像,使得卷积神经网络的识别率得到提升。该方法引入自适应多尺度特性,实现了同一卷积层不同尺寸卷积核生成不同尺度特征,通过级联浅层和深层的特征,达到不同抽象程度的特征图融合。同时,为增强模型的泛化能力,本文提出了基于正则化约束的损失函数。

2 改进的手势识别算法

2.1 多尺度卷积神经网络

在深度学习网络中,不同尺度的卷积层可获得不同抽象程度的特征图。尺度越小,所表征的目标细节越明显;而尺度越大,网络越深,获得的特征就越抽象^[16-17]。在手势识别领域中,肢体与镜头之间的距离和角度使得手势呈现出不同的外观特征,只有自适应地调节网络的深度与尺度才能适应不同复杂环境的手势识别。为了提取不同尺度下的目标特征,本文采用的自适应多尺度特性主要利用同一卷积层不同尺寸卷积核生成不同尺度特征,通过级联浅层和深层的特征,来达到不同抽象程度的特征图融合的目的。

本文设计的多尺度手势识别网络将 ReLU 激活层输出的特征图分两路输出:一路沿着正常的传播方向输出;另一路直接输出,经过均值池化后接入全连接层,最后对各个全连接层输出的特征向量进行特征融合,输入 Softmax 层进行分类识别。通过多层的特征提取,多尺度卷积神经网络能够利用低层、中层和高层图像特征进行图像分类识别,使得图像的分类识别能够更加精细化,并大大减小了网络优化的计算量。

2.2 尺度特征的选取

手势尺度的不确定性会直接影响模型的分类准确度,因此选择合适的尺度特征是卷积网络设计的关键。如果将每一个激活层输出的特征都叠加到一起,则很容易出现过拟合现象,并且由于层数的增加,会占用很大的运行内存。如果选取的尺度特征过少,则不能达到预期的实验效果,因此选取合适的尺度特征非常必要。受文献^[18]启发,本文尺度特征的选择采用贪心算法进行优化。同时,为了避免过拟合,对多尺度深度网络的损失函数进行正则化约束,通过引入一个额外的正则化项来获得不同尺度下的最优解。引入的正则化因子如式(1)所示:

$$C = C_0 + \frac{\lambda}{2} \sum_{\omega} \omega^2 \quad (1)$$

其中, C 代表新代价函数, C_0 表示原代价函数, λ 为正则参数, ω 为相应的权重。通过对新代价函数的权值 ω 求偏导可以得到如下偏导函数:

$$\frac{\partial C}{\partial \omega} = \frac{\partial C_0}{\partial \omega} + \lambda \omega \quad (2)$$

因此,权值的学习因子可以通过式(2)变换求解,其结果如下:

$$\omega' = \omega - \eta \frac{\partial C_0}{\partial \omega} - \eta \lambda \omega = (1 - \eta \lambda) \omega - \eta \frac{\partial C_0}{\partial \omega} \quad (3)$$

可以看出,新的权值更新规则可以设置为 $1 - \eta \lambda$,其中 η 是学习率, $\eta \lambda$ 称为权值衰减率。通过调节 λ 的大小,就可以改变整个网络对应的权值大小。当 λ 较大时,训练好的模型的权值就较小。由于网络中较小的权值对训练数据中的噪声不敏感,因此能够减少过拟合现象的出现。

2.3 网络训练

卷积神经网络的训练过程主要分为两个部分:前向传播和反向传播。为了对本文提出的多尺度卷积网络模型进行训练,文中采用的所有前向传播卷积公式如式(4)所示:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} k_{i,j}^l + b_j^l\right) \quad (4)$$

其中, l 表示卷积网络中的第 l 层; j 表示卷积层的第 j 个核; M 表示卷积核所在的区域; k 与 b 分别表示卷积核与偏置; x 表示特征图对应位置的值; f 表示激活函数。在卷积神经网络中,常用的池化(下采样)运算有最大池化、均值池化和高斯池化。在卷积神经网络的设计过程中,池化层的设计只需要定义池化窗口的大小、池化方法和步长。

众所周知,反向传播算法通常与梯度下降法相结合来训练人工神经网络。通过计算神经网络中代价函数对所有参数的梯度来更新参数值,使得代价函数不断减小,从而实现对神经网络的训练。为了对全连接网络部分的参数进行优化更新,只需要对全连接层的反向传播计算过程中输出层的残差与隐含层的残差进行计算,其等式如下:

$$\delta_j = (d_{q,h} - x_{out,j}) g'(x_j) \quad (5)$$

$$\delta_j^l = \left(\sum_{h=1}^{n^{l+1}} \delta_h^{l+1} \omega_{h,j}^{l+1}\right) g'(x_j^l) \quad (6)$$

式(5)为输出层残差,式(6)为隐含层残差。 $d_{q,h}$ 表示相应的期望输出, $x_{out,j}$ 表示实际输出, $g'(x_j)$ 表示激活函数的导数, x_j 表示上一层的输出, h, j 分别表示第 h 个神经元和第 j 个输入。因此,根据反向传播算法,全连接网络层的权重和偏置更新公式如下:

$$\Delta W^l = -\eta x^{l-1} (\delta^l)^T \quad (7)$$

$$\Delta b^l = \delta^l \quad (8)$$

其中, ΔW^l 表示第 l 层的权值, η 表示学习率, δ^l 表示 l 层的残差, x^{l-1} 表示 $l-1$ 层的输出, Δb^l 表示第 l 层的偏置。

3 实验仿真及分析

3.1 实验数据集

目前,针对手势识别所建立的数据集有 ChaLearn 挑战赛建立的 CGD 数据集^[19],以及约克大学建立并维护的 Hands 数据集。然而,这些数据集的样本量较少,且背景单一。为了获取更真实的手势数据实验效果,本文采用了东南大学自建的手势识别数据集(EShands)进行样本扩充。本文采用的整个样本集定义了复杂背景和简单背景下的6个手势,扩充后的手势集训练样本数量达到了10000,测试样本集1500个,

其中静态手势集定义了6个手势,每个手势包含125个复杂背景样本和125个简单背景样本。为了更好地使用卷积神经网络进行手势识别,将每一种图片的大小统一化。此例中,静态手势识别的图片统一采用 66×76 像素的规格;为了减少计算量并缩短运行时间,将图片统一进行灰度化处理。

3.2 网络结构及参数设置

本文提出了一种改进的卷积神经网络手势识别算法,其软硬件平台如下:CPU为Intel(R) Core(TM) i7-6700 CPU @ 3.20GHz;GPU为NVIDIA GeForce GTX 1070Ti;操作系统为ubuntu 16.04;深度学习框架为伯克利视觉和学习中心(BVLC)开发的Caffe框架。

本文设计的网络包含5个卷积层、6个池化层和3个全连接层。第1个卷积层的卷积核大小为 11×11 ,步长为4,包含96个卷积核;第2个卷积层的卷积核大小为 5×5 ,步长为1,包含256个卷积核;第3个卷积层的卷积核大小为 3×3 ,步长为1,包含384个卷积核;第4个卷积层的卷积核大小为 3×3 ,步长为1,包含256个卷积核;第5个卷积层的卷积核与第4个卷积层相同。第1-3个池化层的池化窗口大小为 3×3 ,步长为2,采用最大池化方式;第4个池化层的池化窗口大小为 3×3 ,步长为1,采用最大池化方式;第5-6个池化层的池化窗口大小为 3×3 ,步长为2,采用平均池化方式。选取经激活函数ReLU输出的特征图作为多尺度特征,本文网络选取第2个、第4个和第5个卷积层输出的特征图,分别进行池化处理,通过一个全连接层后进行特征融合,最后输入Softmax层进行分类识别。

3.3 定量结果及分析

为了进一步验证多尺度卷积神经网络在性能上优于单尺度卷积神经网络,本文选择了当前比较流行的几种单尺度特征的深度卷积神经网络,如CaffeNet^[12],VGG_CNN^[20],DCNN^[18]和ResCNN^[14],参考这些网络结构,设计实现了这些网络结构的多尺度深度卷积神经网络模型,并在训练数据与测试数据统一的情况下进行了对比实验。单尺度网络的性能对比如表1和图1所示;多尺度网络的性能对比如表2和图2所示。

表1 单尺度网络的性能对比

Table 1 Performance comparison of single scale networks

Model	Accuracy	Time/s	Memory/G
CaffeNet	0.837	0.43	2.64
VGG_CNN	0.819	0.40	2.41
DCNN	0.797	0.59	2.45
ResCNN	0.753	0.68	2.72

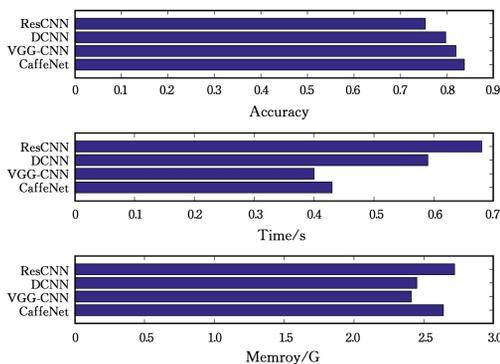


图1 各算法在单尺度网络中的性能对比

Fig. 1 Performance comparison of various algorithms in a single-scale network

表2 多尺度网络的性能对比

Table 2 Performance comparison of multi-scale networks

Model	Accuracy	Time/s	Memory/G
M_CaffeNet	0.903	0.60	3.52
M_VGG_CNN	0.871	0.51	3.29
M_DCNN	0.865	0.65	3.50
M_ResCNN	0.853	0.70	3.88

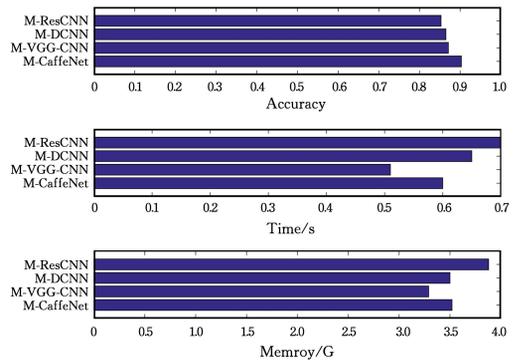


图2 各算法在多尺度网络中的性能对比

Fig. 2 Performance comparison of various algorithms in multi-scale networks

在单尺度卷积神经网络结构中,通常将全连接网络最后一层的输出作为特征;CaffeNet特征向量的维度为4096,VGG_CNN特征向量的维度为1000,DCNN特征向量的维度为1000,ResCNN的特征维度为1000。多尺度卷积神经网络的特征维度主要取决于两个方面:一是特征图的选择;二是特征图池化窗口大小的选择。本文选择第2个、第4个和第5个卷积层输出的特征图,特征维度变为9216;多尺度VGG_CNN选择第1层、第4层和第5层输出的特征,特征维度变为2000;多尺度DCNN选择第1个、第3个和第5个卷积层输出的特征图,特征维度变为2000;多尺度ResCNN选择第1个、第3个和第5个卷积层输出的特征图,特征维度变为2000。这些卷积神经网络引入多尺度特征进行实验,特征维度增加了2倍左右。从表1可以看出,多尺度卷积神经网络的特征维度增加,识别率也得到了很大的提升,说明引入多尺度特征能够提高卷积神经网络静态手势的识别率。但是,网络训练的时间并没有出现大幅度的增加,这是因为训练卷积神经网络卷积计算是耗时最大的操作,而在本文设计的多尺度卷积神经网络中并没有进行比原网络更多的卷积计算,因此网络训练所需的时间并没有大幅度的增加。从内存的使用情况来看,由于网络层数增多,网络需要保存的中间变量也随之增加,使得训练网络所需的内存增加较大。综上所述,多尺度卷积神经网络在性能上优于单尺度卷积神经网络。

为了便于对不同手势下的识别精度进行对比分析,本文对不同算法的识别精度进行了相应的定量统计,其结果如表3和图3所示。由于篇幅有限,本文没有提供识别图像,仅做了相应的文字描述。由于本文自建的数据集包含各种复杂的背景,尤其是在昏暗背景下,手势比较模糊,很难从个别图像中区分手指的现状。从表3可以看出,所提算法的准确率仅为0.64,但在所有对比算法中其结果是最好的,充分说明了本文提出的多尺度模型的有效性。

表3 不同数据集下的算法性能对比

Table 3 Comparison of algorithm performance on different data sets

Data Model	CaffeNet	VGG_CNN	DCNN	ResCNN	Proposed
CGD	0.69	0.69	0.74	0.72	0.75
Hands	0.74	0.76	0.69	0.66	0.80
EShands	0.59	0.61	0.57	0.53	0.64

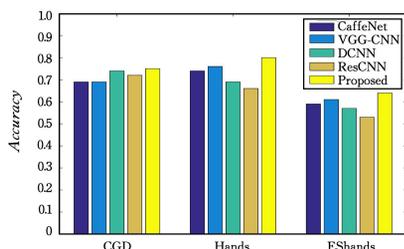


图3 各算法在不同数据集上的性能对比

Fig. 3 Performance comparison of various algorithms on different data sets

结束语 传统的浅层学习网络过度依赖于人工选择手势特征,因此无法适应复杂多变的自然场景。本文在卷积神经网络架构的基础上,提出了一种改进的多尺度深度网络手势识别模型,使得能够利用卷积层自动学习手势特征,去除了人工提取特征的弊端。该方法在多尺度模型下实现了尺度特征选择和网络模型训练。实验结果表明,所提网络模型的识别精度高于普通单尺度卷积神经网络结构的识别精度,弥补了提取特征不够精细、全面及稳定性欠佳等缺点,同时网络训练所需的时间并没有大幅度的增加。

参 考 文 献

- [1] MITRA S, ACHARYA T. Gesture Recognition; A Survey[J]. IEEE Transactions on Systems, Man and Cybernetics, Part C, (Applications and Reviews), 2007, 37(3): 311-324.
- [2] WU Y, HUANG T S, MATHEWS N. Vision-Based Gesture Recognition: A Review[C]// International Gesture Workshop on Gesture-based Communication in Human-computer Interaction. Springer-Verlag, 1999.
- [3] WANG S B, QUATTONI A, MORENCY L P, et al. Hidden Conditional Random Fields for Gesture Recognition[C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006). New York: IEEE, 2006.
- [4] LEE H K, KIM J H. An HMM-Based Threshold Model Approach for Gesture Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1999, 21(10): 961-973.
- [5] BRETZNER L, LAPTEV I, LINDBERG T. Hand Gesture Recognition using Multi-Scale Colour Features. Hierarchical Models and Particle Filtering[C]// Face and Gesture FG'02. IEEE Computer Society, 2002.
- [6] SUN Y, LI C, LI G, et al. Gesture Recognition Based on Kinect and sEMG Signal Fusion[J]. Mobile Networks and Applications, 2018, 23(4): 797-805.
- [7] YANG L, HU G M, HUANG D F, et al. Static Gesture Recognition Algorithm of Combining Skin Color Segmentation with ELM[J]. Journal of Guangxi University (Nat Sci Ed), 2015, 40(2): 444-450.
- [8] ZHANG Q Y, WANG D D, ZHANG M Y, et al. Hand Gesture Recognition Based on Bag of Features and Support Vector Machine[J]. Journal of Computer Applications, 2012, 32(12): 3392-3396.
- [9] LATORRECARMONA P, PLA F. 3D human gesture recognition using integral imaging[J]. Spienewsroom, 2018, 12(1): 15-20.
- [10] CORNEANU C, NOROOZI F, KAMINSKA D, et al. Survey on Emotional Body Gesture Recognition [J/OL]. IEEE Transactions on Affective Computing. <https://ieeexploreieee.org/document/8493586>.
- [11] HAO Y, BAI Y P, ZHANG X F, et al. Application of Convolution Neural Network in SAR Target Recognition[J]. Journal of Chongqing University of Technology (Natural Science), 2018, 32(5): 204-209.
- [12] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks[C]// IEEE. The IEEE Conference on Computer Vision and Pattern Recognition (CPVR). Las Vegas: IEEE, 2017: 4700-4708.
- [13] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. 2016: 21-37.
- [14] XIE S, GIRSHICK R B, DOLLAR P, et al. Aggregated Residual Transformations for Deep Neural Networks[C]// Computer Vision and Pattern Recognition. 2017: 5987-5995.
- [15] WU, XIAO Y. A Hand Gesture Recognition Algorithm Based on DC-CNN[J]. Multimedia Tools and Applications, 2019: 1-13.
- [16] HE Y, LI G, LIAO Y, et al. Gesture Recognition Based on An Improved Local Sparse Representation Classification Algorithm [J]. Cluster Computing, 2019, 22(10): 935-946.
- [17] PI S Y, TANG H, XIAO N F. Fully Convolutional Deep Learning Model Based Graspable Object detection [J]. Journal of Chongqing University of Technology (Natural Science), 2018, 32(2): 166-173.
- [18] LATORRECARMONA P, PLA F. 3D Human Gesture Recognition Using Integral Imaging[J]. Spienewsroom, 2018, 12(1): 15-20.
- [19] GUYON I, ATHITSOS V, JANGYODSUK P, et al. ChaLearn Gesture Challenge: Design and First Results[C]// Computer Vision & Pattern Recognition Workshops. 2012.
- [20] XIN S, HEE-SEUNG K, KOMATSU S, et al. Spatial-temporal Human Gesture Recognition Under Degraded Conditions Using Three-dimensional Integral Imaging[J]. Optics Express, 2018, 26(11): 13938.



JING Yu, born in 1982, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include image processing, pattern recognition and computer vision.