# 基于混合卷积神经网络的静态手势识别

石雨鑫 邓洪敏 郭伟林
(四川大学电子信息学院 成都 610065)

**摘 要** 静态手势识别在人机交互方面具有重要的应用价值,但手势背景的复杂性和手势形态的多样性给识别的准确性带来了一定的影响。为了提高手势识别的准确率,文中提出了一种基于卷积神经网络(Convolution Nenral Network,CNN)与随机森林(Random Forest,RF)的识别方法。该方法首先对静态手势的图片进行手势分割,然后利用卷积网络的特征提取功能提取特征向量,最后使用随机森林分类器对这些特征向量进行分类。一方面,卷积神经网络具有分层学习的能力,能够收集图片上更具代表性的信息;另一方面,随机森林对样本和特征选择具有随机性,并且对每个决策树结果进行了平均,不易出现过拟合问题。在静态手势数据集上进行验证,实验结果显示:所提方法能有效地对静态手势进行识别,平均识别率能够达到94.56%。文中进一步将所提方法与几种经典的特征提取方法(主成分分析(PCA)和局部二进制(LBP))进行对比,实验结果显示:相比于PCA和LBP特征提取方法,由CNN提取的特征向量进行分类识别的效果更好,该方法的识别率比PCA-RF方法高2.44%,比LBP-RF方法高1.74%。最后,在经典的MNIST数据集上进行验证,所提方法的识别率达到了97.9%,高于其他两种传统的特征提取方法。

**关键词** 卷积神经网络,随机森林,静态手势,识别
**中图法分类号** TP183 **文献标识码** A

## Static Gesture Recognition Based on Hybrid Convolution Neural Network

SHI Yu-xin DENG Hong-min GUO Wei-lin
(College of Electronics and Information Engineering,Sichuan University,Chengdu 610065,China)

**Abstract** Static gesture recognition has caught special attention for its great application value in man-machine interaction. At the same time,the accuracy of gesture recognition is affected by the complexity of gesture background and the diversity of gesture morphology in a certain extent. In order to improve the accuracy of gesture recognition,a method was proposed,which is based on convolutional neural network(CNN) and random forest(RF). Firstly,the image of the static gesture is segmented,then the feature extraction function of convolution network is used to extract feature vectors,and finally the random forest classifier is used to classify these feature vectors. On the one hand,the CNN has the ability of layered learning and is able to collect more representative information on the picture. On the other hand,random forest shows randomness for samples and feature selection,meanwhile,it can be avoided easily that the results of each decision tree is averaged over fitting problem. This paper verified by using the static gesture data set,and the experimental results show that the proposed method can effectively identify the static gestures and achieve an average recognition rate of 94.56%. The method proposed in this paper was further compared with principal component analysis (PCA) and partial binary(LBP). The experimental results show that the classification and recognition effect with feature extraction by CNN is better than PCA and LBP. The recognition rate is 2.44% higher than that of PCA-RF method and 1.74% higher than that of LBP-RF method. Finally,the recognition rate of the proposed method reaches 97.9%, which is higher than the other two traditional feature extraction methods.

**Keywords** Convolutional neural network,Random forest,Static gesture,Recognition

## 1 引言

随着计算机视觉技术的发展,手势识别成为了模式识别领域的研究热点,并且得到更为普遍的应用,如体感游戏、手语识别、辅助汽车控制等。手势的多样性和相似性,以及照明条件的影响,加大了识别难度。众多学者对手势识别进行了研究,如使用隐式马尔可夫模型(HMM)[1]、模板匹配[2]、基于几何特征识别[3]、神经网络[4]等方法。

卷积神经网络最初是受到生物视觉系统的神经机制启发,由LeCun等[5]提出的一种神经网络模型,用来处理时间序列和图像等类似网络结构的数据。卷积神经网络具有局部连接、权值共享的特点,并且可通过局部感受野、权值共享和

本文受国家自然科学基金(61174025)资助。
石雨鑫(1994-),女,硕士生,主要研究方向为神经网络、模式识别,E-mail:shiyuxin2655209@vip.qq.com;邓洪敏(1969-),女,博士,副教授,主要研究方向为非线性动力学、模糊控制、神经网络,E-mail:denghongming@aliyun.com(通信作者);郭伟林(1992-),男,硕士生,主要研究方向为神经网络、人工智能。

下采样方法获得较好的平移、缩放和扭曲不变性，也可以缓解模型的过拟合问题。卷积神经网络在图像数据的识别过程中可以避免复杂的预处理工作，直接挖掘图像数据的局部特征，提取全局训练特征进行分类。如今，卷神经网络在图像分类、语音识别[6]、物体检测[7]、人脸识别[8]等多方面取得进展。

Lecun 等将卷积神经网络应用于手写字符识别，得到了高于传统方法的精度。近年来有学者对此问题展开进一步研究，Niu 等[9]提出了将 CNN 与 SVM 结合的方法，用 SVM 来生成预测；史鹤欢等[10]提出 CNN 与 PCA 相结合的方法，用 PCA 对 CNN 权值进行预处理，提升了 CNN 的分类性能。本文提出将 CNN 与随机森林相结合，随机森林[11]是基于多个决策树的随机集成学习方法。决策树结构简单，将其集成学习实现多分类，在运算量没有显著增加的前提下提升了分类性能。本文使用 CNN 的可训练的特性来代替传统的特征提取方法，从而收集更具代表性和相关性的信息，再通过随机森林进行分类。实验证明，其达到了比较好的效果。本文第 2 节介绍了静态手势分割；第 3 节介绍了 CNN、随机森林和本文所用模型的原理；第 4 节展示了实验过程与结果，并将本文方法与其他方法进行对比；最后总结全文并对未来研究进行展望。

## 2 静态手势分割

对静态手势分割的目的是将手势从复杂的背景中提取出来，因为背景较为复杂，直接采取灰度图像二值化无法取得较好的效果。为了消除背景对手势识别的影响，首先对静态手势进行分割处理。手势分割的方法有很多，如基于肤色的分割[12]、基于像素值的分割[13]、基于二维形状的分割[14]等。本文使用基于肤色的分割方法对静态手势进行分割。由于肤色的色度差异大于亮度差异，通过比较 RGB，HSV，YCbCr 空间，YCbCr 颜色空间在进行肤色分割方面有较好的聚类性和稳定性，在亮度和色度分离方面具有更好的性能。因此，我们采用 YCbCr 空间进行手势肤色的建模与分割。

首先，在 YCbCr 空间下利用高斯分布对肤色进行建模。高斯模型通过计算像素的概率值完成肤色的确认，并分割出手势区域。二维高斯型函数对肤色的建模公式如下：

$$P(Cr,Cb)=\exp[-0.5(x-m)^{\mathrm{T}}C^{-1}(x-m)] \tag{1}$$

其中，$x$ 为样本像素在 YCbCr 空间的矩阵，$x=(Cb,Cr)^{\mathrm{T}}$；$m$ 为肤色在 YCbCr 空间的样本均值，$m=E(x)$；$C$ 为肤色相似度模型的协方差矩阵，$C=E\{(x-m)(x-m)^{\mathrm{T}}\}$。

在计算出每个像素值的肤色概率值后，对肤色概率矩阵进行自适应阈值二值化处理。静态手势的预处理如图 1 所示。

(a)原始图像 (b)肤色似然图像 (c)二值图像 (d)填充后的二值图像

图 1 手势预处理

## 3 卷积神经网络与随机森林

### 3.1 卷积神经网络结构

标准的卷积神经网络一般由输入层、卷积层、下采样层(也称池化层)、全连接层和输出层组成。卷积神经网络的输入层为通常为一个矩阵，如一幅图像(一般使用原始图像)。本文卷积网络结构主要参照 LeNet-5[5]，其是由 Lecun 在 1998 年为手写数字识别设计的经典卷积网络结构。其基本结构如图 2 所示，卷积层以 $C$ 标识，下采样层以 $S$ 标识，全连接层以 $F$ 标识。

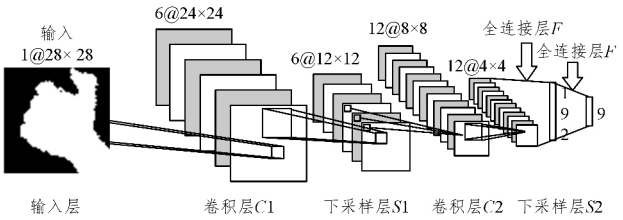


图 2 卷积神经网络的结构

如图 2 所示，此网络结构是由 2 个卷积层和 2 个子采样层交替组成。卷积层，又叫特征提取层，采用 $5\times5$ 大小的卷积核，对输入的图像进行卷积操作。输入的原始图像用 $X$ 表示，$H_i$ 为卷积网络第 $i$ 层的特征图($H_0=X$)。卷积的产生过程可以表示为：

$$H_i=f(H_{i-1}\otimes W_i+b_i) \tag{2}$$

其中，$W_i$ 表示第 $i$ 层卷积核的权值矩阵，"$\otimes$"表示卷积运算，$b_i$ 为偏置矩阵。每个特征图中所有单元共享权值，从而大大减少了权值参数的数量，加速了收敛。$f(\cdot)$表示为神经元激励函数，常用的神经元激活函数有 Sigmoid 函数、Tanh 函数、Relu 的函数等。由于前两种激活函数使得网络训练速度慢，容易陷入过拟合，因此本文采用 Relu 函数作为激活函数来提高训练速度，同时也解决了梯度消失的问题[15]。

下采样层又称为池化层，池化层可以有效地缩小矩阵的尺寸，从而减少全连接层的参数。下采样层主要对特征进行模糊，从而获得平移、尺度等不变性，依据一定的采样规则对特征图进行采样。采样过程可以表示为：

$$H_i=f(down(H_{i-1})) \tag{3}$$

其中，$down(\cdot)$表示采样函数。常用的采样函数有最大值采样函数和均值采样函数，本文采用均值采样函数，即计算区域元素的算术平均值作为函数输出，从而提取特征平面局部相应的均值。采样的大小取为 $2\times2$，取样过程与卷积相似。

全连接层的作用即对特征样本进行分类。全连接层中每个单元都与下采样层的每个单元相连接，每个神经元的输出如下：

$$H_i=f(W^{\mathrm{T}}H_{i-1}+b) \tag{4}$$

其中，$H_{i-1}$为全连接层的输入，即上层采样层的输出；$H_i$ 为全连接层的输出；$f(\cdot)$表示激活函数；$b$ 为偏置。全连接层将输出节点拉成一列向量，采用 Softmax 分类器对测试标签进行预测，得到不同种类的概率分布情况。

### 3.2 卷积神经网络的训练

卷积神经网络的训练与一般的神经网络相同，使用误差反向传播算法进行训练。

1)计算平方误差代价函数：

$$E^N=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}(t_k^n-y_k^n)^2 \tag{5}$$

其中，$N$ 是样本个数，$c$ 是标签的维度，$t_k^n$ 表示第 $n$ 个样本实际输出标签的第 $k$ 维，$y_k^n$ 是第 $n$ 个样本理论输出标签的第 $k$ 维。

2）按极小误差的方法调整权值矩阵 $W$ 以及偏置值 $b$。更新过程运用 BP 神经网络中的梯度下降法：

$$W_2 = W_1 - \eta \frac{\partial E}{\partial W_1} \tag{6}$$

$$b_2 = b_1 - \eta \frac{\partial E}{\partial b_1} \tag{7}$$

其中，$\eta$ 为梯度下降的学习率，$W_2$ 和 $b_2$ 为更新后的值。反向传播训练所做的就是更新网络的权值，使得网络输出 $y_k^n$ 与实际值 $t_k^n$ 的误差最小化。

### 3.3　随机森林

随机森林（Random Forest，RF）是 Breiman 提出的一个分类器融合算法，可以很好地解决多分类问题。随机森林具有分类速度快、泛化能力强的特点，在文本分类、人体行为识别、对象跟踪等领域被广泛应用。它包含大量的决策树，每棵决策树为一个分类器[11]。对于一个新的输入样本，随机森林中的所有决策树被用来决策属性分类，最后将每个决策树的结果汇总，获得票数最多的分类结果为最终生成结果。随机森林算法与 Bagging 算法[16]类似，均是采用 Boostrap 方法抽样[17]，产生多个训练集。不同的是，随机森林算法在构建决策树时，采用了随机分裂属性集的方法。

#### 3.3.1　Boostrap 抽样

Boostrap 抽样是从给定训练集中有放回地均匀抽样，即每当选中一个样本，它等可能地被再次选中并被再次添加到训练集中。设定集合 $S$ 中含有 $n$ 个不同样本 $\{x_1, x_2, x_3, \cdots, x_n\}$，每次有放回地从集合 $S$ 中抽取一个样本，一共抽取 $n$ 次，形成新的集合 $S^*$，则集合 $S^*$ 中不包含某个样本 $x_i$（$i=1, 2, \cdots, n$）的概率为：

$$p = (1 - \frac{1}{n})^n \tag{8}$$

当 $n \to \infty$ 时，有：

$$\lim_{n\to\infty} p = \lim_{n\to\infty} (1 - \frac{1}{n})^n = e^{-1} \approx 0.368 \tag{9}$$

因此，集合 $S^*$ 中可能包含了重复的样本，除去重复样本，集合 $S$ 中约有 $1 - 0.368 = 63.2\%$ 的样本出现在集合 $S^*$ 中。

#### 3.3.2　随机森林分类器构造

在训练过程中，随机森林中的每棵决策树的训练样本都是从总样本集中随机、有放回地选取一个子集，决策树在每个节点都选取当前分类效果最好的弱分类器。所有决策树分类器构成一个随机森林分类器。

1）利用 Boostrap 方法抽样，随机产生 $T$ 个训练集 $S_1$，$S_2$，…，$S_T$，每个训练集生成对应的决策树 $C_1$，$C_2$，…，$C_T$；

2）设样本特征数为 $K$，从 $K$ 个特征中随机抽取 $k$ 个特征作为当前节点的分裂特征集（$0 < k < K$），并以这 $k$ 个特征中最好的分裂方式对该节点进行分裂，此处选择吉尼指数来对决策树进行分裂[18]；

3）对于每个测试样本 $X$，利用每棵决策树进行测试，得到对应类别 $C_1(X)$，$C_2(X)$，…，$C_T(X)$；

4）最后，采用投票的方式将 $T$ 个决策数中输出最多的类别作为测试集样本 $X$ 的所属类别。

## 4　混合卷积神经网络及实验结果分析

### 4.1　混合卷积神经网络

混合卷积神经网络是卷积神经网络与随机森林结合的模型。本文对静态手势进行识别，识别过程如图 3 所示。

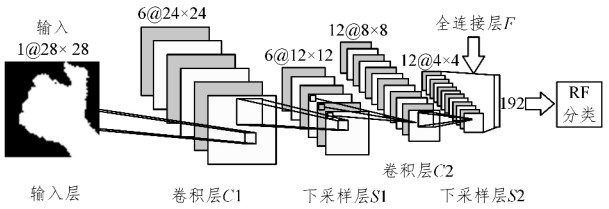


图 3　卷积神经网络与随机森林的识别过程

模型的输入层为 $28 \times 28$ 大小的静态手势图片，第一层卷积层 $C1$ 的滤波器尺寸为 $5 \times 5$，深度为 6，步长为 1，故输出的尺寸为 $24 \times 24$，$C1$ 层参数个数为 $6 \times (5 \times 5 + 1) = 156$，$C1$ 层每个像素都与前一个输入层像素连接，共 $156 \times 28 \times 28 = 122\,304$ 个连接。第二层下采样层 $S2$ 的输入为 $C1$ 层输出 $24 \times 24 \times 6$ 的矩阵，采用 $2 \times 2$ 均值滤波器，步长为 2。第三层为卷积层 $C3$，输入为上层输出 $12 \times 12 \times 6$ 大小的矩阵，使用过滤器的大小为 $5 \times 5$，深度为 12，其连接方式如表 1 所列。该层的输出矩阵大小为 $8 \times 8 \times 12$，共有 $6 \times (3 \times 5 \times 5 + 1) + 3 \times (4 \times 5 \times 5 + 1) + 3 \times (4 \times 5 \times 5 + 1) = 1\,062$ 个参数，因此有 $1\,062 \times 8 \times 8 = 67\,968$ 个连接。第四层为下采样层 $S4$，输入为 $C3$ 的输出，与 $S2$ 层一样，采用 $2 \times 2$ 的均值滤波器，步长为 2，输出矩阵为 $4 \times 4 \times 12$。最后一层为全连接层，全连接层将 $C4$ 输出的特征矩阵拉成一个向量，向量维度为 $4 \times 4 \times 12 = 192$。该层的输出可视为静态手势图片的特征向量，直接送入随机森林分类器进行分类识别。

表 1　$C3$ 层的连接方式

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | X | | | | X | X | X | | | X | | X |
| 2 | X | X | | | | X | X | X | | X | X | |
| 3 | X | X | X | | | | X | X | X | | X | X |
| 4 | | X | X | X | | | X | X | X | X | | X |
| 5 | | | X | X | X | | | X | X | X | X | |
| 6 | | | | X | X | X | | | X | | X | X |

### 4.2　实验结果及分析

本文实验在室内场景采集的手势图像数据集下进行，并且在经典的 MNIST 手写体字符数据集上验证。手势图像数据集共有 9 类手势，每类手势有 1 600 幅形状略有差异的图片，一共有 14 400 张图片，从中随机抽取 10 000 张作为训练样本，其余 4 400 张作为测试样本。

为验证 CNN-RF 性能，本文做了以下实验。

1）分析了随机森林中决策树对分类性能的影响，实验结果如图 4 所示。

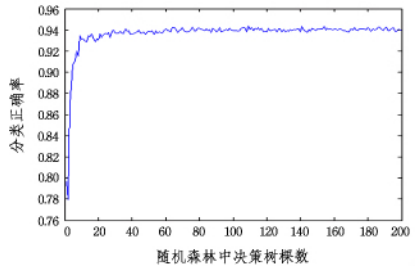


图 4　决策树对随机森林性能的影响

实验结果表明，针对静态手势数据集而言，综合考虑随机森林中包含的决策树与建模速度，决策树选择为

40～200 棵较为理想。

2)对训练样本进行了不同次数的迭代,迭代次数与所对应的识别准确率如图 5 所示。

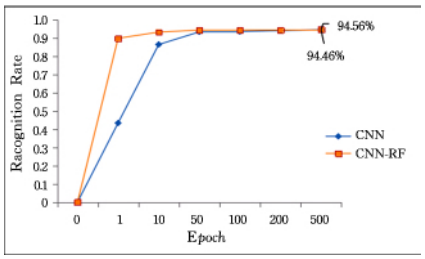


图 5 迭代次数与所对应的识别准确率

从图 5 可以看出,CNN-RF 能比 CNN 更快地拟合,且识别率更高,在迭代次数达到 500 时趋于饱和,此时 CNN-RF 的识别率为 94.56%,比 CNN 的识别率高 0.10%。

3)比较了两种经典方法 PCA 和 LBP 与 RF 结合和 CNN 与 RF 结合的识别精度,实验结果如表 2 所列。

表 2 3 种特征提取方式的对比

| 方法 | PCA-RF | LBP-RF | CNN-RF |
|---|---|---|---|
| 识别率/% | 92.12 | 92.82 | 94.56 |

实验结果表明,用 CNN 作为特征提取,效果优于传统的 PCA 和 LBP 提取方式。

4)在 MNIST 数据集下 CNN-RF 与两种经典特征提取方法的对比结果如表 3 所列。

表 3 3 种特征提取方式的对比

| 方法 | PCA-RF | LBP-RF | CNN-RF |
|---|---|---|---|
| 识别率/% | 95.57 | 94.87 | 97.96 |

结束语 本文提出了一种基于卷积神经网络与随机森林的静态手势识别方法,首先对静态手势图片进行预处理,分割出手势,接着利用卷积神经网络的可训练功能来代替传统的图像特征提取法方法,以收集更具代表性的特征,最后用随机森林分类器对图像特征进行识别分类。实验结果表明:本文的识别方法可以得到较好的分类性能,其识别率能够达到 94.56%,还具有更好的收敛性能,CNN-RF 识别率高于 PCA-RF 方法 2.44%,高于 LBP-RF 方法 1.74%。本文最后在 MNIST 数据集上验证了卷积神经网络与随机森林结合同样能达到较好的识别效果。本文仅对单一环境下的手势进行分割与分类,下一步将考虑在复杂背景下对手势进行分割,且要考虑如何优化卷积网络的权值,使其更快地收敛。

## 参 考 文 献

[1] ZAKI M M,SHAHEEN S I. Sign language recognition using a combination of new vision based features[J]. Pattern Recognition Letters,2011,32(4):572-577.

[2] ALKHATEEB J H,KHELIFI F,JIANG J,et al. A new approach for off-line handwritten Arabic word recognition using KNN classifier[C]// IEEE International Conference on Signal and Image Processing Applications. IEEE,2010:191-194.

[3] LIU Y,YIN Y,ZHANG S. Hand Gesture Recognition Based on HU Moments in Interaction of Virtual Reality[C]// International Conference on Intelligent Human-Machine Systems and Cybernetics. IEEE,2012:145-148.

[4] RONCANCIO C. Combined Gesture-Speech Recognition and Synthesis Using Neural Networks[J]. IFAC Proceedings Volumes,2008,41(2):2968-2973.

[5] LECUN Y,BENGIO Y. Convolutional networks for images, speech,and time series[M]// The handbook of brain theory and neural networks. MIT Press,1998.

[6] WAIBEL A,HANAZAWA T,HINTON G,et al. Phoneme recognition using time-delay neural networks[J]. Readings in Speech Recognition,1990,1(2):393-404.

[7] VAILLANT R,MONROCQ C,CUN Y L. An original approach for the localization of objects in images[C]// International Conference on Artificial Neural Networks. IET,1993:26-30.

[8] LAWRENCE S,GILES C L,TSOI A C,et al. Face recognition: a convolutional neural-network approach[J]. IEEE Transactions on Neural Networks,1997,8(1):98-113.

[9] NIU X X,SUEN C Y. A novel hybrid CNN-SVM classifier for recognizing handwritten digits[J]. Pattern Recognition,2012,45(4):1318-1325.

[10] 史鹤欢,许悦雷,马时平,等. PCA 预训练的卷积神经网络目标识别算法[J]. 西安电子科技大学学报(自然科学版),2016,43(3):161-166.

[11] BREIMAN L. Random forest[J]. Machine Learning,2001,45:5-32.

[12] STERGIOPOULOU E,PAPAMARKOS N. Hand gesture recognition using a neural network shape fitting technique[J]. Engineering Applications of Artificial Intelligence,2009,22(8):1141-1158.

[13] ESCALERA S,RADEVA P,DIMOV D,et al. Graph cuts optimization for multi-limb human segmentation in depth maps[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2012:726-732.

[14] BELONGIE S,MALIK J,PUZICHA J. Shape matching and object recognition using shape contexts[C]// IEEE International Conference on Computer Science and Information Technology. IEEE,2010:483-507.

[15] NAIR V,HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// International Conference on International Conference on Machine Learning. Omnipress,2010:807-814.

[16] QUINLAN J R. Bagging,boosting,and C4.5[C]// Proceedings of the National Conference on Artificial Intelligence. AMER ASSOC ARTFICIAL INTELL,1996:725-730.

[17] JOHNSON R W. An Introduction to the Bootstrap[J]. Teaching Statistics,2001,23(2):49-54.

[18] 王全才. 随机森林特征选择[D]. 大连:大连理工大学,2011.