融合密集卷积与空间转换网络的手势识别方法

马 杰* 张绣丹 杨 楠 田亚蕾
(河北工业大学电子信息工程学院 天津 300401)

摘 要：手势识别作为人机交互的方式之一，在人工智能日益发展的今天备受瞩目。针对手势旋转、平移、缩放等形变导致识别率偏低的问题，该文基于密集卷积网络(Densenet)与空间转换网络(STN)提出了一种新型的网络结构Densenet_V2，先利用空间转换网络对输入的样本和特征图进行空间变换和对齐，再利用密集卷积网络自动提取手势的有效特征，最后通过线性分类器对手势进行分类。为防止网络模型对样本数据集过度拟合，对网络进行训练时在损失函数中加入 L2 正则项以实现权重衰减。在 Marcel 手势库上进行多次实验。实验结果表明，Densenet_V2 可以提高对静态形变手势的识别率。

关键词：手势识别；形变；密集卷积网络；空间转换网络；L2 正则项
中图分类号： TP391.41 **文献标识码**： A **文章编号**：1009-5896(2018)04-0951-06
DOI: 10.11999/JEIT170627

Gesture Recognition Method Combining Dense Convolutional with Spatial Transformer Networks

MA Jie ZHANG Xiudan YANG Nan TIAN Yalei
(*School of Electronic & Information Engineering*, *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: As an important milestone for the development of the artificial intelligence, gesture recognition enables the human-computer interaction and has received significantly growing research interest nowadays. However, the current technology for the gesture recognition has the low quality in the gesture rotation, translation and scaling. To solve the problem, a novel network structure named Densenet_V2 is proposed, and it is based on Dense Convolutional Networks (Densenet) and Spatial Transformer Networks (STN). Firstly, the input samples and feature maps are spatially transformed and aligned with the STN. Then the effective features of gestures are automatically extracted by using the Densenet. Finally, the linear classifier is adopted to classify the gestures. To prevent the network model from over-fitting the sample data set, the L2 regular term is involved into the loss function to achieve the weight decay when training the network. Experiments on the Marcel gesture database show that Densenet_V2 can improve the recognition rate of static deformation gestures.

Key words: Gesture recognition; Deformation; Dense convolutional networks; Spatial Transformer Networks (STN); L2 regular term

1 引言

手势识别作为人工智能的重要研究方向之一，近年来备受瞩目。手势识别系统分为两类：基于数据手套的手势识别[1]和基于计算机视觉的手势识别[2–10]。前者需要戴上数据手套，通过数据手套将用户的手势信息传递给计算机。这种方式虽定位准确，处理速度较快，但由于设备价格昂贵，用户体验差等缺点而难以推广。基于视觉的手势识别方式不需要用户穿戴任何设备，是目前手势识别研究的重点方向。

传统的基于计算机视觉的手势识别包括手势分割、特征提取和分类 3 步。手势分割是手势识别的前提，最常用的方式是利用肤色信息获取手势二值图像[2,3]。特征提取和分类是手势识别的关键，Liu 等人[4]利用 Hu 不变距提取手势区域特征，Hu 不变距具有旋转平移不变性，但对形变仍比较敏感；Dardas 等人[5]提取手势的 SIFT 特征并用支持向量机进行分类，完成对手势的识别；杨学文等人[6]综合手势主方向和类-Hausdorff 距离模板匹配法，较好地解决了手势发生形变时识别率较低的问题；刘淑

收稿日期：2017-06-29；改回日期：2017-11-28；网络出版：2018-01-23
*通信作者：马杰 jma@hebut.edu.cn
基金项目：国家自然科学基金(61203245)，河北省自然科学基金(F2012202027)
Foundation Items: The National Natural Science Foundation of China (61203245), The Natural Science Foundation of Hebei Province (F2012202027)

萍等人[7]提出了一种结合手指检测和方向梯度直方图的手势识别方法，成功地识别出 25 种手势。但在传统的手势识别中，基本都是采取人工提取特征的方式，这样有一定的主观性和局限性。而卷积神经网络(Convolutional Neural Network, CNN)可自动提取图像中的有用信息并学习，在一定程度上解决了这个问题。Lin 等人[8]利用肤色模型对手势进行分割，再根据手势主方向对手势姿势进行校准，最后输入到卷积神经网络(LeNet-5)进行训练，训练好的模型可识别出 7 种手势，平均识别率为 95.96%；杜坚等人[9]利用位运算代替滑动窗口完成目标的快速筛选，然后用卷积神经网络(LeNet-5)对目标区域进行 2 次判断和识别，该模型在 Marcel 手势库上的识别率达到 96.1%; Pyo 等人[10]利用卷积神经网络(AlexNet)对 NYU 手势库进行识别，平均识别率为 94.94%。

CNN 克服了人工提取特征的主观性和局限性，提高了识别率，但网络模型仍然对形变手势的鲁棒性不足。针对此问题，本文提出了一种新型网络结构 Densenet_V2，将最新的 CNN 结构—密集卷积网络与空间转换网络 STN 相合，STN 可以动态地对每个输入样本做相应的网格变换，无需通过手势主方向标记即可自适应地将数据进行空间变换和对齐。同时为防止网络模型对样本数据集过度拟合，对网络进行训练时在损失函数中加入 L2 正则项以实现权重衰减。实验结果表明，Densenet_V2 在静态手势识别上取得了很好的效果。

2 密集卷积网络

CNN 最早起源于 Lecun 等人[11]提出的 LeNet-5，它的本质是一种输入到输出的映射：

$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{x}) \tag{1}$$

其中，$f(\cdot)$ 是激活函数，$\boldsymbol{x}$ 为输入信息，$\boldsymbol{W}$ 为核函数，$\boldsymbol{h}$ 为特征映射。每一个通过卷积核得到的特征向量 $\boldsymbol{h}$ 都是一类特征映射，一个卷积层中包含若干个卷积核，用来提取图像中不同位置的信息。CNN 中的每一层卷积都会提取图像的有效特征，并将提取的特征输送到下一层卷积中。单层网络可以不断堆叠为深层网络，将底层特征逐步抽象为高阶特征，最后通过线性分类器完成分类。VGG[12]中提出越深的网络识别效果越好，因为深层网络可以将连续的特征信息结合构成高维特征，并使样本数据间的相关性得到充分表示。

密集卷积网络(Densenet)[13]是最新的CNN结构之一，网络结构如图 1 所示。Densenet 包括 3 个密集卷积块(每个密集卷积块中包含 4 个卷积层)，相邻的密集卷积块之间通过卷积层和平均池化层连接，网络中采用全局平均池化层代替传统的全连接层以减少参数量抑制过拟合，最后通过线性分类器完成分类。

在密集卷积块中，以每层之前所有层的输出作为输入。对于 l 层的传统卷积网络，连接数为 l，对于 Densenet，连接数则为 $l(l+1)/2$，这样可以充分利用之前层的所有信息，同时缩短了前层和后层之间的连接，有效地解决了随着网络的加深而产生的梯度消失问题。

3 空间转换网络

CNN 定义了一个非常强大的模型，但图像的旋转、平移等形变仍会导致模型的识别率偏低。在手势识别任务中，常采用基于手势主方向提取手势特征的方法，将特征数据在空间上对齐，从而保证手势的旋转、平移不变性[6,8]，提高识别率。DeepMind 提出的空间转换网络[14]无需手势主方向的标记，即可根据分类或者其他任务自适应地将数据进行空间变换和对齐。CNN 中的池化层也可以使网络具有一定的平移不变性[15]，减少几何变换对分类任务的影响。不同于池化层的可接受域是固定的，STN 是一种动态机制，可以通过为每个输入样本产生适当的变换来主动获得空间变换图像。

STN包括3部分：定位网络，网格生成器和采样器，如图2所示。其中 U 是输入特征图，V 是输出特征图。定位网络通过一个子网络生成空间变换参数 θ ,$\theta = f_{\text{loc}}(U)$。网格生成器根据 V 中坐标 $\left(x_i^t, y_i^t\right)$ 找到它在 U 中对应坐标 $\left(x_i^s, y_i^s\right)$，即找出输入特征图 U 和输出特征图 V 的关系 T_θ：

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G) = \boldsymbol{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{2}$$

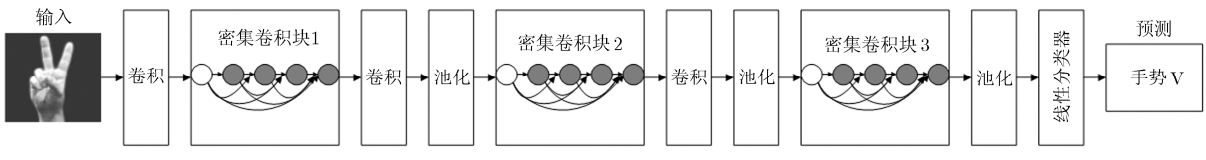

图 1 Densenet 网络结构图

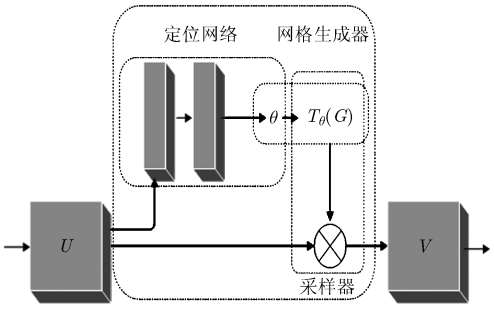

图 2　STN 结构图

其中，$\boldsymbol{A}_\theta$ 是仿射变换矩阵；采样器根据 T_θ 通过双线性插值的方法在 U 中采样出真实的像素值放入 V 中对应坐标 $\left(x_i^t, y_i^t\right)$ 中，

$$V_i^C = \sum_n^H \sum_m^W U_{nm}^C \max\left(0,1-\left|x_i^s - m\right|\right) \cdot \max\left(0,1-\left|y_i^s - n\right|\right) \tag{3}$$

其中，U_{nm}^C 是输入特征图在位置 (n,m) 处的像素值，H 表示图像的高，W 表示图像的宽，V_i^C 是输出通道对应位置的像素值，C 是通道数。在实验中要求对输入的每个通道进行相同的采样，以保证通道之间的空间一致性。

4　基于 Densenet_V2 的手势识别

为了进一步提升 CNN 对形变手势的鲁棒性，本文基于密集卷积网络(Densenet)和空间转换网络(STN)提出了一个新型的网络结构 Densenet_V2。在实验中通过对不同的网络结构进行多次测试，确定网络的最佳参数如下：卷积层的总数为 15(其中每个密集卷积块中包含 4 个卷积层)，每个卷积层中卷积核的个数均设为 12，密集卷积块内的卷积核大小均设为 3×3，其余卷积核大小为 1×1；前两个池化操作选用 2×2 的平均池化，最后一个池化操作选用 2×2 的全局平均池化。网络结构如图 3 所示。

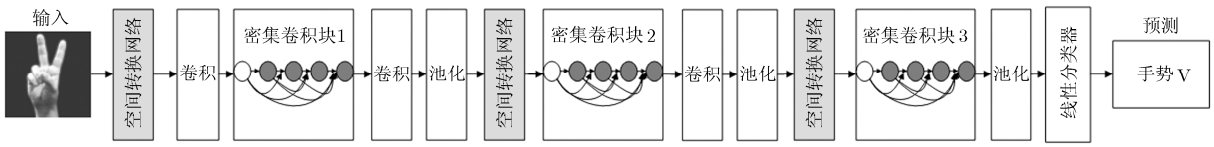

图 3　Densenet_V2 网络结构图

Densenet_V2 是在 Densenet 中加入 STN，使样本在进入 CNN 之前先通过一层 STN 进行空间对齐，并在密集卷积块 2、密集卷积块 3 前各加入一层STN，对网络中提取的特征图进行空间对齐。STN 可以动态地对每个输入样本做相应的网格变换，从而将输入图像或者学习到的特征在空间上对齐，如图 4 所示，其中 U 为 STN 的输入图，V 为输出图。

Densenet_V2 是一个深层网络模型，而实验中选用的 Marcel 手势库的训练样本仅有 4872 张图片，对于小样本数据集的训练，深度学习模型容易陷入过拟合，从而导致网络泛化能力不足。为抑制样本量偏少而带来的过拟合问题，在训练网络时引入 L2 正则化[16]。L2 正则化是在原损失函数中加入一个正则项：

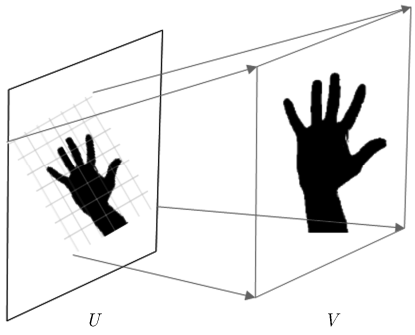

图 4　STN 效果图

$$L = -\sum y_i' \lg(y_i) + \frac{\lambda}{2n}\sum \|\boldsymbol{W}\|_2^2 \tag{4}$$

其中，L 为正则化后的损失函数，$-\sum y_i' \lg(y_i)$ 为交叉熵，记为 L_0，y_i' 是网络预测值，y_i 是样本实际对应的标签值，$\frac{\lambda}{2n}\sum \|\boldsymbol{W}\|_2^2$ 为L2正则项，λ 是正则化系数 $(0<\lambda<1)$，n 是训练集个数，$\boldsymbol{W}$ 是网络的参数。

式(4)对应的梯度为

$$\nabla_{\boldsymbol{W}} L = \nabla_{\boldsymbol{W}} L_0 + \lambda \boldsymbol{W} \tag{5}$$

使用单步梯度下降更新权重：

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \epsilon\left(\nabla_{\boldsymbol{W}} L_0 + \lambda \boldsymbol{W}\right) \tag{6}$$

即

$$\boldsymbol{W} \leftarrow (1-\epsilon\lambda)\boldsymbol{W} - \nabla_{\boldsymbol{W}} L_0 \tag{7}$$

其中，ϵ 为任意小的正数。通过式(7)可以看出引入L2正则化会修改学习准则，在执行每一步梯度下降算法之前都会对权重向量乘以一个常数因子 $(1-\epsilon\lambda)$ 以收缩权重，$\boldsymbol{W}$ 越小，网络复杂程度越低，过拟合产生的概率就越小。L2正则化通过在损失函数中加入约束项的方法，抑制网络模型对训练样本的过度拟合。

5 实验结果及分析

本文是在Linux环境下，基于TensorFlow平台，实现手势识别。硬件配置参数如下：4VCPU, Tesla K80 GPU, 64 G内存，12 G显存。实验中选取Marcel标准手势库进行手势识别实验，该手势库训练样本包含4872张图片共6种手势(V, A, B, C, point,five, 分别对应标签值0~5)，如图5所示。

在实验中将 Densenet_V2 的学习率设置为0.001, dropout 参数设置为0.8，最高迭代次数设置为 10000 次。同时采用交叉验证的方式确定正则化系数，当正则化系数为 0.001 时，网络的泛化能力最佳。

本文具体手势识别步骤如下：

(1)对原数据集进行数据增广，并对图像进行标准化处理；

(2)将处理后的数据集按照 4:1 的比例，分为训练集和验证集；

(3)将训练集和验证集作为 Densenet_V2 的输入，采用反向传播算法对网络进行逐层训练，用随机梯度下降算法进行权值更新，不断调整网络的迭代次数和学习率，当验证集的准确率不再提升时，停止迭代；

(4)训练完成后，将测试集作为网络的输入，对训练好的网络进行测试；

(5)对测试集中的手势进行形变处理(随机旋转、缩放、平移)后，观察实验结果。

用测试集对训练好的网络进行测试，同时为了验证本文提出的网络结构的性能，实验中将 5 种不同的手势识别方法(文献[7–10]模型及 Densenet)作为 Densenet_V2 的比较对象进行对比研究。文献[8–10]均利用 CNN 进行手势识别研究，网络参数设置参考对应文献。在文献[8,9]模型中的卷积层个数为 2，第 1 个卷积层有 6 个卷积核，第 2 个卷积层有 12 个卷积核，卷积核的大小定为5×5；文献[10]模型的卷积层个数为 5，每层中卷积核的个数均为12，大小均为5×5; Densenet 模型中卷积层个数为12，每层中卷积核的个数均为 12，大小为3×3。

观察表 1 和表 2 中的数据，可以发现：(1)在迭代次数为 10000 次时，测试集的识别率最高，对于未发生形变手势的平均识别率可以达到 98.8%，对于发生形变的识别率也可以达到 97.6%。Densenet_V2 对于形变手势和未形变手势都有很好的识别效果。(2)在损失函数没有加入 L2 正则化时网络呈现过拟合状态；加入 L2 正则化后该网络模型对测试集的识别率提高了 17%左右。L2 正则化有效地抑制了网络模型对训练样本的过度拟合。表 3 中数据显示，Densenet_V2 模型在卷积层个数为 15 时，识别效果效果最佳。该实验结果表明，当卷积层个数为 15 时，网络对手势图像特征的抽象能力最佳并达到饱和状态，此时增加卷积层个数，对网络识别率没有影响，而减少卷积层个数，则会减弱网络的抽象能力，使识别率降低。

通过图 6 和图 7 可以明显地看出，在手势未发生形变时，以下几种模型：文献[7]模型、文献[8]模

表 1 Densenet_V2 对未形变手势的平均识别率(%)

迭代次数		1000	2000	5000	8000	10000
训练集识别率		95.9	99.9	100.0	100.0	100.0
测试集识别率	有正则化	91.4	95.8	98.1	98.6	98.8
	无正则化	71.4	76.9	78.9	79.1	81.2

表 2 Densenet_V2 对形变手势的平均识别率(%)

迭代次数		1000	2000	5000	8000	10000
训练集识别率		95.9	99.9	100.0	100.0	100.0
测试集识别率	有正则化	79.3	80.2	92.0	96.2	97.6
	无正则化	60.1	63.7	70.9	74.7	79.9

表 3 卷积层个数对识别率的影响(迭代次数 10000，有正则化)(%)

卷积层个数		13	14	15	16	17
训练集识别率		100.0	100.0	100.0	100.0	100.0
测试集识别率	未形变手势	95.9	97.1	98.8	98.8	98.8
	形变手势	95.0	95.9	97.6	97.6	97.6

图 5　Marcel 手势样本示例

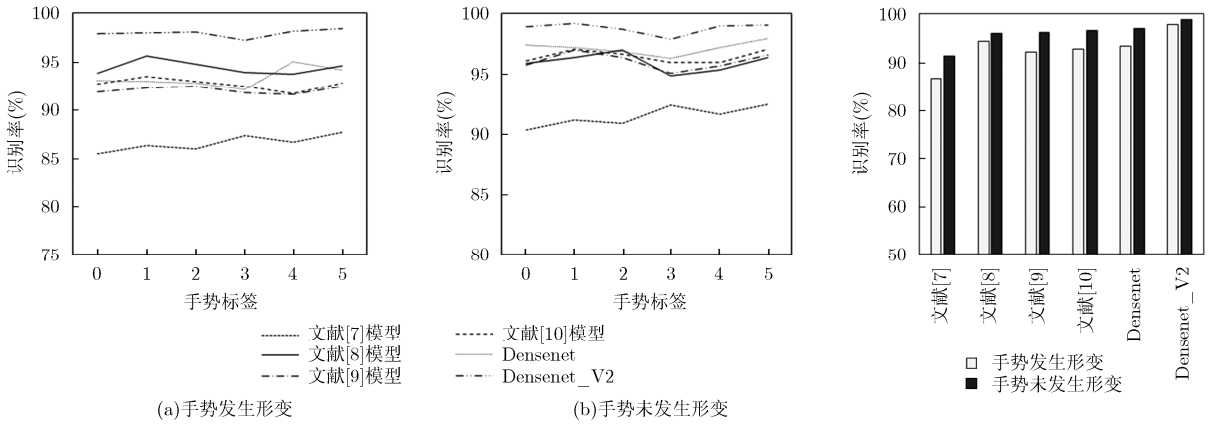

图 6 不同识别方法的识别率对比

图 7 不同情况下平均识别率对比

型、文献 [9] 模型、文献 [10] 模型、Densenet, Densenet_V2 的识别率逐步提升。可见相比传统手势识别方法，CNN 可有效提高对手势的识别率，且网络层数越深，对图像特征的抽象能力越强，识别率越高。当测试集手势发生形变时，几种识别方法的识别率有不同程度的下降：分别下降了 4.7%, 1.6%, 3.8%, 3.8%, 0.9%。可见相比其他方法，文献 [8]模型和 Densenet_V2 对形变手势的鲁棒性较高。文献[8]模型识别率下降较少，因为该模型根据手势主方向对手势姿势进行了校准，在一定程度上保证了手势的旋转平移不变性。本文提出的 Densenet_V2 模型识别率下降最少，该模型在 Densenet 的基础上加入 STN，无需手势主方向的标记，可动态地对每个输入样本做相应的网格变换，自适应地将输入图像和特征图进行空间变换和对齐，减少了手势形变对分类任务产生的影响，从而提高了 CNN 对形变手势的鲁棒性。

6　结束语

本文提出了一种新型网络模型 Densenet_V2，该模型融合了密集卷积网络和空间转换网络，让网络自动学习一个仿射变换矩阵，将输入图像或者学习到的特征在空间上对齐，减少了手势形变对分类任务产生的影响，从而更好地实现分类任务。同时在网络训练时在损失函数中加入 L2 正则化，有效地解决了数据集样本过少而导致的网络过拟合问题，提高了网络的泛化能力。但由于网络层数较深，虽然采用 GPU 进行加速，网络的训练时间依然较长，后期需继续对网络结构或训练方法进行进一步调整。

参 考 文 献

[1] PIYUSH K, SIDDHARTH S R, and Anupam A. Hand data glove: A new generation real-time mouse for human-computer interaction[C]. International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, Jharkand, India, 2012: 750–755. doi: 10.1109/RAIT.2012.6194548.

[2] WEI W and JING P. Hand segmentation using skin color and background information[C]. International Conference on Machine Learning and Cybernetics, Xi'an, China, 2012: 1487–1492. doi: 10.1109/ICMLC.2012.6359584.

[3] 阮晓钢, 林佳, 于乃功, 等. 基于多线索的运动手部分割方法[J]. 电子与信息学报, 2017, 39(5): 1088–1095. doi: 10.11999/JEIT160730.

RUAN Xiaogang, LIN Jia, YU Naigong, *et al*. Moving hand segmentation based on multi-cues[J]. *Journal of Electronics & Information Technology*, 2017, 39(5): 1088–1095. doi: 10.11999/JEIT160730.

[4] LIU Y, YIN Y, and ZHANG S. Hand gesture recognition based on HU moments in interaction of virtual reality[C]. International Conference on Intelligent Human-Machine Systems and Cybernetics, Nanchang, China, 2012: 145–148.

doi: 10.1109/IHMSC.2012.42.

[5] DARDAS N H and GEORGANAS N D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques[J]. *IEEE Transactions on Instrumentation & Measurement*, 2011, 60(11): 3592–3607. doi: 10.1109/TIM.2011.2161140.

[6] 杨学文, 冯志全, 黄忠柱, 等. 结合手势主方向和类-Hausdorff距离的手势识别[J]. 计算机辅助设计与图形学学报, 2016, 28(1): 75–81. doi: 10.3969/j.issn.1003-9775.2016.01.010.
YANG Xuewen, FENG Zhiquan, HUANG Zhongzhu, *et al*. Gesture recognition based on combining main direction of gesture and Hausdorff-like distance[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2016, 28(1): 75–81. doi: 10.3969/j.issn.1003-9775.2016.01.010.

[7] 刘淑萍, 刘羽, 於俊, 等. 结合手指检测和 HOG 特征的分层静态手势识别[J]. 中国图象图形学报, 2015, 20(6): 781–788. doi: 10.11834/jig.20150607.
LIU Shuping, LIU Yu, YU Jun, *et al*. Hierarchical static hand gesture recognition by combining finger detection and HOG features[J]. *Journal of Image and Graphics*, 2015, 20(6): 781–788. doi: 10.11834/jig.20150607.

[8] LIN H I, HSU M H, and CHEN W K. Human hand gesture recognition using a convolution neural network[C]. IEEE International Conference on Automation Science and Engineering, Taipei, China, 2014: 1038–1043. doi: 10.1109/CoASE.2014.6899454.

[9] 杜堃, 谭台哲. 复杂环境下通用的手势识别方法[J]. 计算机应用, 2016, 36(7): 1965–1970. doi: 10.11772/j.issn.1001-9081.2016.07.1965.
DU Kun and TAN Taizhe. General method for gesture recognition in complex environment[J]. *Journal of Computer Applications*, 2016, 36(7): 1965–1970. doi: 10.11772/j.issn.1001-9081.2016.07.1965.

[10] PYO J, JI S, and YOU S. Depth-based hand gesture recognition using convolutional neural networks[C]. International Conference on Ubiquitous Robots and Ambient Intelligence, Xi'an, China, 2016: 225–227. doi: 10.1109/URAI.2016.7625742.

[11] LECUN Y, BOTTOU L, BENGIO Y, *et al*. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278–2324. doi: 10.1109/5.726791.

[12] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[OL]. http://arxiv.org/abs/1409.1556, 2014.

[13] HUANG G, LIU Z, WEINBERGER K Q, *et al*. Densely connected convolutional networks[OL]. http://arxiv.org/abs/1608.06993, 2016.

[14] JADERBERG M, SIMONYAN K, ZISSERMAN A, *et al*. Spatial transformer networks[OL]. https://arxiv.org/abs/1506.02025v3,2015.

[15] LECUN Y, BENGIO Y, and HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436–444. doi: 10.1038/nature14539.

[16] GOODFELLOW I, BENGIO Y, and COURVILLE A. Deep Learning[M]. Massachusetts, USA: MIT Press, 2016: 231–234.

马 杰： 男，1978 年生，教授，研究方向为图像处理与模式识别.

张绣丹： 女，1992 年生，硕士生，研究方向为图像处理与模式识别.

杨 楠： 女，1992 年生，硕士生，研究方向为图像处理.

田亚蕾： 女，1993 年生，硕士生，研究方向为图像处理与模式识别.