基于深度学习的视觉手势识别综述

邓智方 袁家政

（北京联合大学北京市信息服务工程重点实验室 北京 100101）

摘 要 深度学习可以让拥有多个处理层的计算模型学习具有多层次抽象数据的表示。这些方法给语音识别、视觉对象识别、对象检测等领域都带来了显著的改善，其优异表现使得作为人机交互重点方式之一的手势识别获得了一定的发展。文中就目前传统方法和基于深度学习的先进视觉手势识别方法进行归纳，并对视觉手势识别进行总结。

关键词 深度学习，视觉，手势识别，人机交互

中图法分类号 TN-91 **文献标识码** A

Survey of Visual Hand Gesture Recognition Based on Deep Learning

DENG Zhi-fang YUAN Jia-zheng

(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)

Abstract Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains. Its excellent performance makes hand gesture recognition, one of the key ways of human-computer interaction, obtain a certain development and progress. This paper summarized the traditional methods and advanced visual gesture recognition methods based on deep learning, and presented the prospect of visual gesture recognition in the end.

Keywords Deep learning, Vision-based, Hand gesture recognition, Human-computer interaction (HCI)

1 引言

手势为人机交互(HCI)提供了一种自然和直观的通信模式，我们可以开发有效的用户界面来使计算机实时地识别手势。然而，由于手势的复杂性以及人手所涉及的高自由度(DOF)而具有丰富的多样性，使得基于视觉的手部跟踪和手势识别成为了一个具有挑战性的问题。为了能够成功地实现它们的功能，人机交互中的手势必须满足实时性能、识别准确性以及针对变换和杂乱背景的鲁棒性的要求。为了满足这些要求，很多手势识别系统使用彩色标记或借助数据手套使得整个过程更为容易[1]。然而，使用标记和手套牺牲了用户的方便性，使得更多研究者关注没有任何标记和手套辅助的裸手势识别，也就是基于视觉的手势识别[2]。

视觉手势识别是指对视频采集设备拍摄到的包含手势的图像序列，通过计算机视觉技术进行处理，进而对手势加以识别，其基本流程如图1所示。

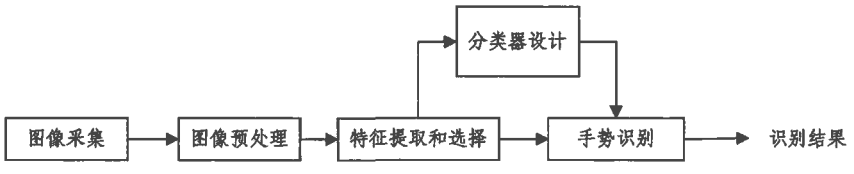

图1 视觉手势识别基本流程

本文受北京联合大学北京市属高校高水平教师队伍建设创新团队建设提升计划(IDHT20170511)资助。

邓智方(1992－)，男，硕士生，主要研究方向为数字图像处理、手势识别，E-mail：dengzfong@163.com；袁家政(1971－)，男，博士，教授，主要研究方向为图形图像处理、文物遗迹的数字化处理、数字博物馆、导航定位等，E-mail：jiazheng@buu.edu.cn(通信作者)。

本文基于视觉的手势识别技术可以进一步分为静态识别和动态手势识别[3]。静态手势被定义为一段时间内在空间中的定向和位置均没有任何移动的手，静态手势识别主要是基于单帧图像中对单只手在单个时间点上的形状、方向、轮廓进行识别。其优势是手处于静止不动状态，姿态、形状、位置等信息不随时间发生变化，容易实现且识别效率较高；但静态手势也有其自身的局限性，比如反映的信息量较少，不符合人手可以随意运动的特性，因此无法应用于复杂的场景中。动态手势识别则是在一段时间内识别手势姿态序列的变化，这个序列参数在空间中对应的是一条轨迹。连续的动态手势是由一帧一帧的静态手势所组成的，即静态手势就是动态手势的一种特殊状态。动态手势的优势是通过不同手势的灵活变化可以传递更多信息，在人机交互领域有广泛的应用。

2　深度学习简介

深度学习的概念起源于人工神经网络的研究，有多个隐层的多层感知器是深度学习模型的一个很好的范例。对神经网络而言，深度指的是网络学习得到的函数中非线性运算组合水平的数量。当前神经网络的学习算法多是针对较低水平的网络结构，将这种网络称为浅结构神经网络，如一个输入层、一个隐层和一个输出层的神经网络；与此相反，将非线性运算组合水平较高的网络称为深度结构神经网络，如一个输入层、两个隐层和一个输出层的神经网络，如图 2 所示。

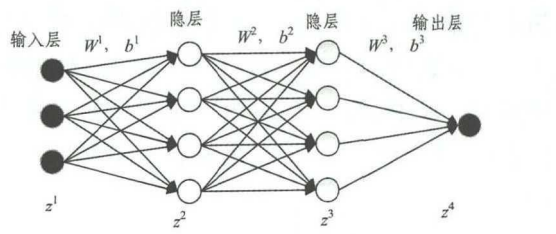

图 2　深度学习的简单网络架构

深度学习是一种特征学习方法，它通过一些简单的但是非线性的模型把原始数据转变成更高层次的、更加抽象的表达。通过足够多的转换的组合，非常复杂的函数也可以被学习[6]。对于分类任务，高层次的表达能够在强化输入数据的区分能力的同时削弱不相关因素。比如，一幅图像的原始格式是一个像素数组，那么在第一层上的学习特征表达通常指的是在图像的特定位置和方向上有没有边的存在；第二层通常会根据边的某些排放来检测图案，这时会忽略一些边上的小的干扰；第三层可能会把那些图案进行组合，从而使其对应于熟悉目标的某部分；随后的一些层会将这些部分再组合，从而构成待检测目标。深度学习的核心是，上述各层特征都不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的。

2.1　卷积神经网络(CNN)

卷积神经网络被设计的初衷是用于处理多维数组数据，比如一个由 3 个包含了像素值 2-D 图像组合成的具有 3 个颜色通道的彩色图像。很多数据形态都是这种多维数组：1D 用来表示信号和序列包括语言，2D 用来表示图像或者声音，3D 用来表示视频或者有声音的图像。卷积神经网络使用 4 个关键方法来利用自然信号的属性：局部连接、权值共享、池化以及多网络层的使用。CNN 内部结构如图 3 所示。

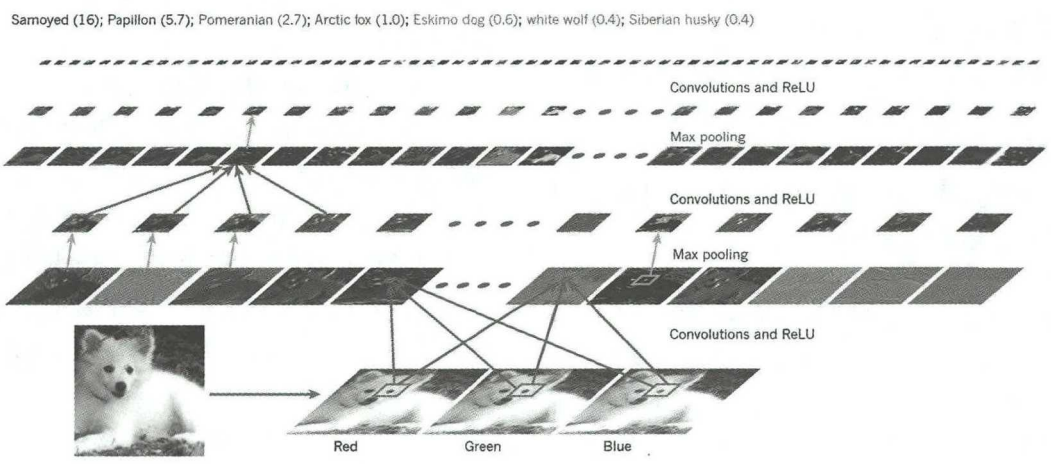

图 3　CNN 内部结构

2.2　递归神经网络(RNN)

首次引入反向传播算法时，最值得一提的是便是使用递归神经网络(Recurrent Neural Networks，RNNs)进行训练。对于涉及到序列输入的任务，比

如语音和语言，利用 RNNs 能获得更好的效果。RNNs 一次处理一个输入序列元素，同时维护网络中隐式单元中隐式的包含过去时刻序列元素的历史信息的“状态向量”。如果是深度多层网络不同神经元的输出，我们会考虑这种在不同离散时间步长的隐式单元的输出，这会使如何利用反向传播来训练 RNNs 的思路更加清晰。RNN 内部结构如图 4 所示。

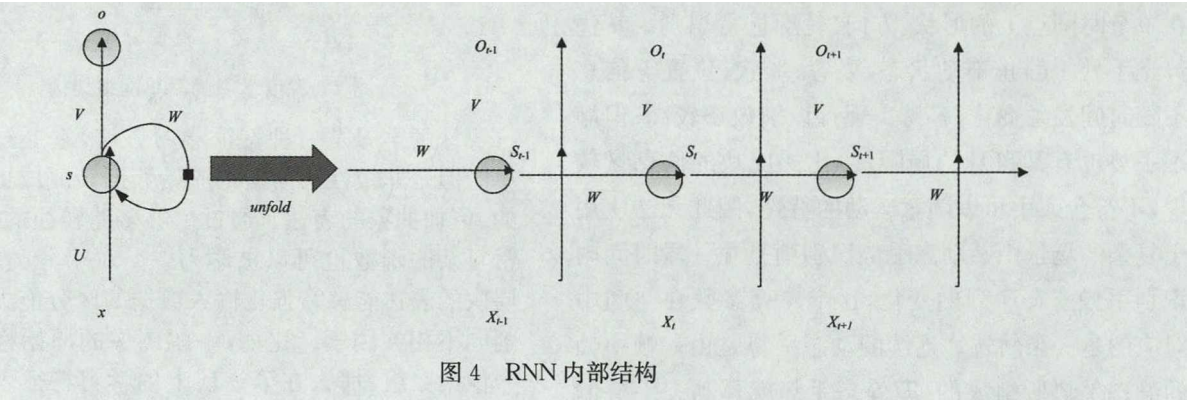

图 4 RNN 内部结构

3 基于传统方法的视觉手势识别

作为基于视觉的手势识别系统的构建模块，大多数完整的手动交互机制包括 3 个基本阶段：检测、跟踪和识别。

手势识别系统的主要步骤是检测手和相应图像区域的分割。这种分割是至关重要的，因为它将任务相关数据与图像背景隔离，然后将其传递到随后的跟踪和识别阶段。文献中已经提出了几种类型的视觉特征以及多数情况下它们的组合方法。这些特征包括手的肤色、形状、运动和解剖模型。文献[4-5]给出了一些手部分割技术的对比与讨论。

检测方法如果在图像采集帧速率下运行得足够快，则也可以用于跟踪。然而，手部跟踪是非常困难的，因为它们可以快速移动，并且它们的外观在几帧内的变化很大。跟踪可以被定义为分割的手部区域或特征的帧到帧对应关系，以理解观察到的手部移动[3]。

手势识别的总体目标是手（位置、姿势或姿态）传达的语义的解释。为了检测静态手势（即姿势），可以使用一般分类器或模板匹配器。然而，动态手势具有时间维度，需要使用处理该维度的方法，例如隐马尔可夫模型（HMM），通过手势表示来建模时间维度（例如基于运动的模型）。

以文献[2]为例，其提出了一种实时手势识别系统，用于通过手势与应用程序或视频游戏交互。

该系统在杂乱的背景下在面部减法[8]之后，再使用皮肤检测[7]和手姿势轮廓比较算法检测和跟踪裸手，通过特征袋和多分类支持向量机识别手势（SVM）并生成手势命令以控制应用程序的语法。训练阶段，在使用尺度不变性特征变换（SIFT）[9]提取每个训练图像的关键点之后，向量量化技术将在 K 均值聚类之后将来自每个训练图像的关键点映射到统一的维度直方图向量（词袋）[10]。此直方图被视为多类 SVM[20]的输入向量，以构建训练分类器。在测试阶段，对于从网络摄像头捕获的每一帧，使用本文算法检测手，然后针对仅包含检测到的手势的每个小图像提取关键点，并且将其发送到聚类模型中以将其映射到词袋向量，它们最终被发送到多类 SVM 训练分类器中以识别手势。该方法虽然最终实现了实时手势识别，但是其由于手势类别在一定程度上有些匮乏，中间处理过程的复杂性较高，且要求硬件及聚类处理的高度精确性，相对来说，并不是一个十分值得考虑的优秀方法。

4 基于深度学习的视觉手势识别的研究方法

深度学习随着自身的优良特性及其快速发展，在很多方面都带来了明显的改善，给手势识别领域也带来了新的研究热点和进步，对手势识别改善较大的方法有三维卷积神经网络（3DCNN）、递归三维神经网络（R3DCNN）等。

4.1 运用 3DCNN 进行手势识别

文献[11]介绍了采用 3D 卷积神经网络的深度和强度通道信息构建的手势识别系统。其在 Molchanov 等人[12]的帮助下，交织两个通道以建立标准化的时空体积，并训练两个单独的子网络。同时，为了减少潜在的过度拟合并改善手势分类器的泛化，提出了一种有效的时空数据增强方法来改变手势的输入体积，增强方法也包含现有的空间增强技术[13]。这项工作与 Molchanov 等人[12]的多传感器方法相似，但在使用两个单独的子网络和数据增加方面不同。其网络结构如图 5 所示，采用时空数据增强进行训练，其在 VIVA 挑战的数据集[19]上的最终结果优于单个 CNN 和基于特征的算法[14]。

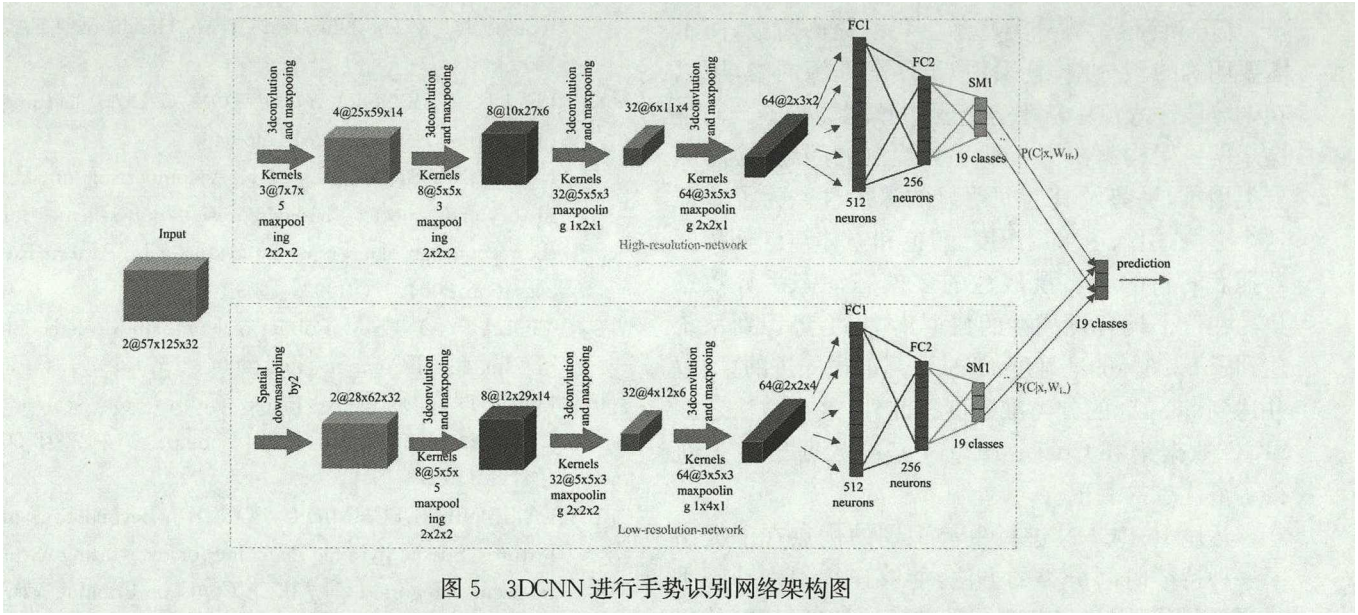

图 5　3DCNN 进行手势识别网络架构图

此方法的优点在于能够动态地识别手势，使用归一化深度和图像梯度值的融合动量，并且增加时空数据来避免过拟合，利用低分辨率和高分辨率子网络的组合显著提高了分类精度；缺点在于手势限于自动驾驶领域，对于序列化的信息处理不足，而且对于复杂环境的鲁棒性不足。

4.2　运用 R3DCNN 进行手势识别

文献[15]通过对之前手势识别 CNN 网络架构的改进，提出了一种具有连续时间分类(CTC)[16]的 R3DCNN 网络架构来进行手势识别，如图 6 所示。它们可以从多模态[17-18]数据中同时检测和分类动态手势。

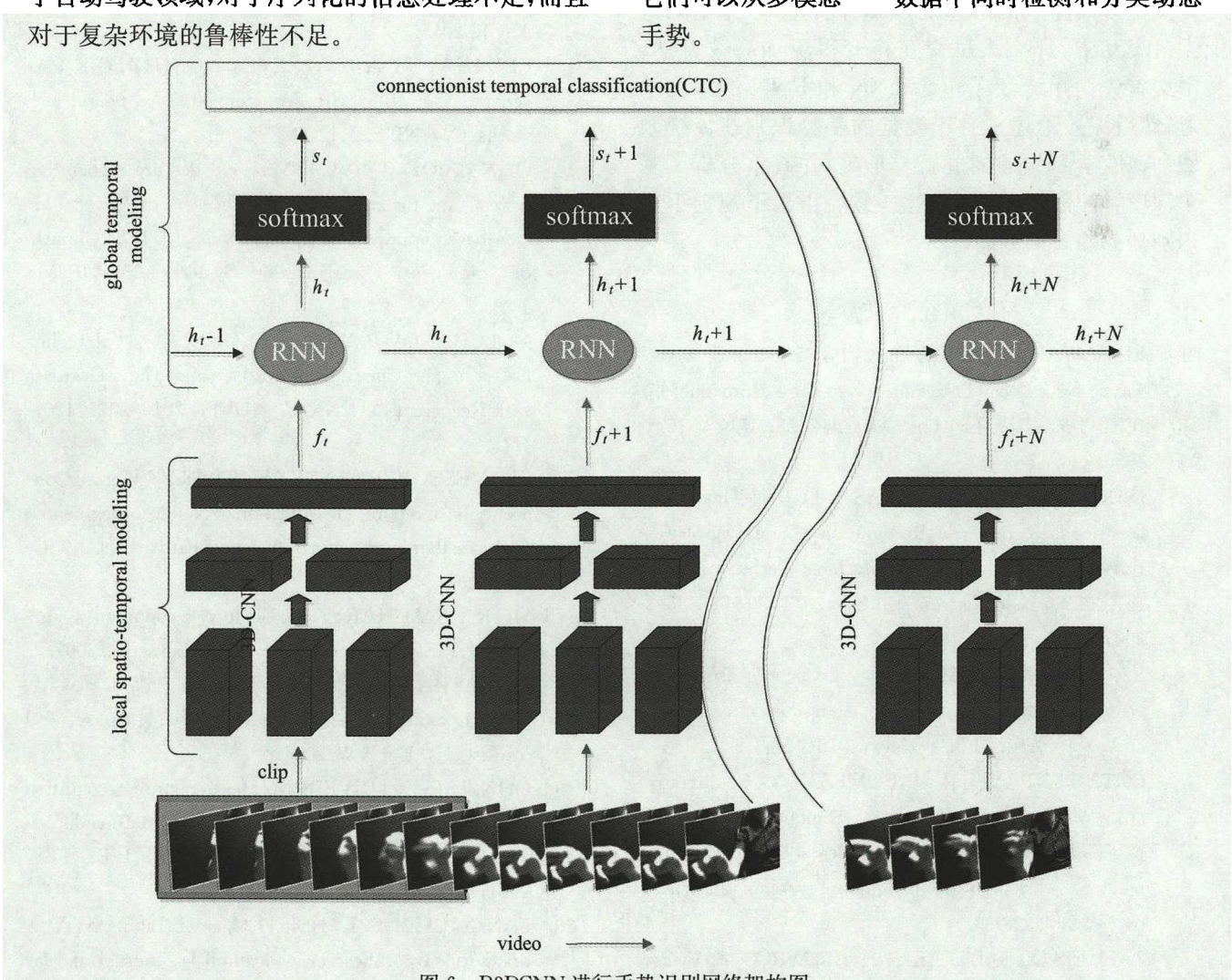

图 6　R3DCNN 进行手势识别网络架构图

CTC 使得手势分类基于手势的核心阶段，而不需要明确的预分割，且采用 CTC 来训练网络，可以在未分段输入流中从手势过程中预测类标签。而且这个网络架构解决了早期手势检测的问题：检测和分类困难，手势及其之间明显的滞后性。为了验证这个方法，引入了采用深度、颜色和立体声红外传感器捕获的新的具有挑战性的多模态动态手势数据集。在这个具有挑战性的数据集中，手势识别系统达到了 83.8%的准确度，超过了当前最先进的算法，并达到 88.4%的分类准确度。此外，这个方法还在 SKIG 数据集和 ChaLearn2014 基准数据集[18] 上获得了最优越的表现。

运用 R3DCNN 进行手势识别的优点在于提出了一种持续时间分类的方法，能够连续识别动态手势，对视频等序列信息有一个良好的处理过程，使得手势的分类更为准确、实时；但是其缺点在于构造模型及预训练复杂，对于复杂环境的识别仍然不具有鲁棒性。

结束语 面对蒸蒸日上的人工智能，深度学习无疑是为其注入的加速剂，如果能将手势完全融入人机交互中，那么人机交互将变得更加高效自然。本文首先简单介绍了包含 CNN 和 RNN 的深度学习，然后深入论述了基于视觉的手势识别在传统领域和深度学习领域的成就与不足，并着重分析了基于深度学习的手势识别方法，最后指出了视觉手势识别任务的发展方向。

参 考 文 献

[1] EL-SAWAH A, GEORGANAS N, PETRIU E. A prototype for 3-D hand tracking and gesture estimation[J]. IEEE Trans. on Instrum. Meas., 2008, 57(8): 1627-1636.

[2] DARDAS N H, GEORGANAS N D. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques[J]. IEEE Trans. on Instrum. Meas., 2011, 60(11): 3592-3607.

[3] RAUTARAY S, AGRAWAL A. Vision based hand gesture recognition for human computer interaction: a survey[J]. Artif. Intell. Rev., 2015(43): 1-54.

[4] COTE M, PAYEUR P, COMEAU G. Comparative study of adaptive segmentation techniques for gesture analysis in unconstrained environments[C]// IEEE international workshop on imagining systems and techniques. 2006: 28-33.

[5] ZABULIS X, BALTZAKIS H, ARGYROS A. Vision-based Hand gesture recognition for human-computer interaction[C]// The Universal Access Handbook. LEA, 2009.

[6] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(28): 436-444.

[7] MCKENNA S, MORRISON K. A comparison of skin history and trajectory-based representation schemes for the recognition of userspecific gestures[J]. Pattern Recognition, 2004, 37(5): 999-1009.

[8] VIOLA P, JONES M. Robust real-time object detection [J]. Int. J. of Comput. Vis., 2004, 2(57): 137-154.

[9] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. Int. J. of Comput. Vis., 2004, 60(2): 91-110.

[10] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]// IEEE Conf. on Comput. Vis. and Pattern Recog.. 2006: 2169-2178.

[11] MOLCHANOV P, GUPTA S, KIM K, et al. Hand Gesture Recognition with 3D Convolutional Neural Networks[C]// CVPR 2015. 2015.

[12] MOLCHANOV P, GUPTA S, KIM K, et al. Multi-sensor System for Driver's Hand-gesture Recognition[C]// AFGR. 2015.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// NIPS. 2012: 1097-1105.

[14] OHN-BAR E, TRIVEDI M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations[J]. IEEE Trans. on Intelligent Transportation Systems, 2014, 15(6): 1-10.

[15] MOLCHANOV P, YANG X D, GUPTA S, et al. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D CNN[C]// CVPR 2016. 2016: 4207-4215.

[16] GRAVES A, FERNANDEZ S, GOMEZ F J, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// ICML. 2006: 369-376.

[17] MITRA S, ACHARYA T. Gesture recognition: a survey[C]// IEEE Systems, Man, and Cybernetics. 2007.

[18] ESCALERA S, BARO X, GONZALEZ J, et al. ChaLearn Looking at People Challenge 2014: dataset and results[C]// ECCVW. 2014.

[19] OHN-BAR E, TRIVEDI M. Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations[J]. IEEE ITS, 2014, 15(6): 1-10.

[20] FAN R, CHANG K, HSIEH C, et al. Liblinear: A library for large linear classification[J]. Journal of Machine Learning Research, 2008, 9(9): 1871-1874.