DOI: 10.3969/j.issn.1009-6868.2017.04.002
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20170707.1618.004.html

基于3D CNN的大规模视频手势识别研究

Large-Scale Video-Based Gesture Recognition Using 3D CNN Model

苗启广/MIAO Qiguang
李宇楠/LI Yunan
徐昕/XU Xin
（西安电子科技大学，陕西 西安 710071）
(Xidian University, Xi'an 710071, China)

中图分类号：TN929.5 文献标志码：A 文章编号：1009-6868 (2017) 04-0009-005

摘要： 提出了一种基于三维卷积神经网络（CNN）的大规模视频手势识别算法。首先，为了获得统一尺度的输入数据，在时域上对所有输入视频进行了归一化处理得到32帧的输入视频；然后，为了从不同的角度描述手势特征，通过真彩（RGB）视频数据生成了光流视频，并将RGB视频和光流视频分别通过C3D模型（一个3D CNN模型）提取特征，并通过特征连接的方式加以融合输入到支持向量机（SVM）分类器来提高识别性能。该方法在Chalearn LAP独立手势数据集（IsoGD）的验证集上达到了46.70%的准确率。

关键词： 手势识别；三维卷积神经网络；光流；SVM

Abstract: In this paper, an effective 3D convolutional neural network(CNN)-based method for large-scale gesture recognition is proposed. To obtain compact and uniform data for training and feature extracting, the inputs are unified into 32-frame videos. To describe features of gesture in different aspects, the optical flow data from red, green, blue (RGB) videos are generated. After that, the spatiotemporal features of RGB and optical flow data are extracted with the C3D model (a 3D CNN model) respectively and blended together in the next stage to boost the performance. Finally, the classes are predicted with a linear support vector machine (SVM) classifier. Our proposed method achieves 46.70% accuracy on the validation set of ChalearnLAP Isolated Gesture Dataset (IsoGD).

Key words: gesture recognition; 3D CNN; optical flow; SVM

手势是一种交流的形式，它指的是利用人的肢体动作来说明其意图或态度的行为。由于在视频监测控制、标志语言理解、虚拟现实和人机交互等领域有着巨大的应用前景，越来越多的研究人员开始研究手势识别算法，以实现将人类手势解释给机器的目标。

手势识别最早期的研究是从20世纪80年代开始，是一种典型的涉及到各方面知识的研究。为了对人体动作的含义加以识别，研究人员先后使用了大量不同种类的方法。早期的大部分方法都是基于复杂的人工制作特征。Stamer和Pentl[1]首先利用隐马尔科夫模型（HMM）进行美国手语的相关研究；Elmezain[2]等利用HMM模型对手势的动态轨迹进行了识别；Sgouropoulos等人[3]将神经网络方法和HMM方法结合使用，可提高动态手势的识别效果，并且该方法具有光照鲁棒性。Wan等人将尺度不变特征变换（SIFT）特征加以扩充，得到3D增强动作SIFT（EMoSIFT）[4]和3D稀疏动作SIFT（SMoSIFT）[5]，并通过稀疏关键点混合特征（MFSK）[6]来进行手势识别。随着近来深度学习技术的蓬勃发展，利用深度神经网络实现手势识别受到了研究者的广泛关注，且相对于传统手工特征方法，在识别率上也取得了重大突破。Karpathy等人[7]利用卷积神经网络（CNN）来实现手势行为识别；Simonyan和Zisserman[8]利用一个双流CNN网络同时提取手势视频中的时域和空域的特征；Tran等人[9]提出了一个3D CNN模型——C3D模型，解决了基于视频的手势识别需要同时处理时域和空域的特征这一问题。

在文章中，我们提出了一种基于同源数据融合的大规模的手势识别方法。首先，我们通过对数据分布特征的分析，将所有帧数不一的视频进行预处理，获得统一的帧数为32帧的视频；随后，我们由真彩（RGB）视频生成了光流视频，以进一步提取动

收稿日期：2017-05-21
网络出版日期：2017-07-07
基金项目：国家自然科学基金（61472302、U1404620、61672409）；模式识别国家重点实验室开放课题基金资助（201600031）；中央高校基本科研业务费专项资金（JB150317）；陕西省自然科学基金（2010JM8027）；航空科学基金（2015ZC31005）

作信息，同时避免表演者服饰、肤色等因素的干扰；我们再利用上文提到的C3D模型，对RGB数据和光流数据同时提取空域和时域上的特征信息，并将这些特征加以融合，最终通过支持向量机（SVM）分类器来获得分类结果，整个流程如图1所示。

1 基于C3D模型的视频手势识别

1.1 预处理

一般的CNN由于其中全连接层的限制，都要求输入数据具有相同的大小。因此我们首先需要对数据进行归一化处理，即统一的帧数，各帧相同的宽和高。为了尽可能地获取代表手势含义的特征，我们采取一种数据驱动的策略来实现这一过程，即通过对数据分布情况的分析来确定归一化的方式。首先，我们分析了实验数据集——Chalearn LAP IsoGD Database（简称IsoGD数据集），该数据集由Wan等人[10]建立，它源自于Guyon等人[11]建立的ChaLearn手势数据集（CGD）。IsoGD数据集包含了47 933个独立的视频，每个视频包含一个手势，这些手势被分为249类，它被用于2016年首届Chalearn LAP大规模手势识别竞赛，其详细信息如表1所示。由于IsoGD数据集中，每个视频的宽度和高度都是一致的，因此需要归一化处理的主要是时域，即帧数信息。

▼表1 Chalearn LAP IsoGD数据集详细信息

子集	类别数	手势数量	表演者数量
训练集	249	35 878	17
校验集	249	5 784	2
测试集	249	6 271	2

如图2所示，由于在数据集中，一些类别的手势看起来非常相似，因此在处理视频成统一帧数时，就需要在保留动作的运动路径信息和降低视频的空间占用之间进行折衷。在分析了35 878个训练集的视频的帧数后，我们发现：尽管视频的帧数从1~405帧各不相同，但是大部分视频的帧数在29~39之间，其中33帧的视频数量最多，达1 202个。为了便于处理，我们选择32作为视频的基准帧数，将所有视频统一至32帧。帧数大于32的视频需进行采样，而帧数小于32的视频则通过复制按一定比例选出的帧进行插值。通过这样的预处理方式，超过98%的视频至少每3帧进行了1次采样，大部分的运动路径信息得以保留。

1.2 光流特征提取

光流，是视觉场中的一种目标、表面和边缘的表征运动的模型，它是由观察者和场景之间的相对运动产生的。在文章中，我们通过RGB视频提取光流特征，一方面用于提取动作路径信息，另一方面也去除了背景、表演者肤色等与动作无关的信息。我们通过Brox等人[12]提出的基于亮度恒常性、梯度恒常性和时空平滑约束假设的能量方程来计算光流特征，该能量方程可表述为：

$$E(u,v)=E_{Data}+\alpha E_{Smooth} \quad (1)$$

其中，$\alpha>0$是一个正则化参数，E_{Data}可表述为：

$$E_{Data}(u,v)=\int_{\Omega}\psi(|\Delta I(x)|^2)+\gamma|\Delta G(x)^2|dx \quad (2)$$

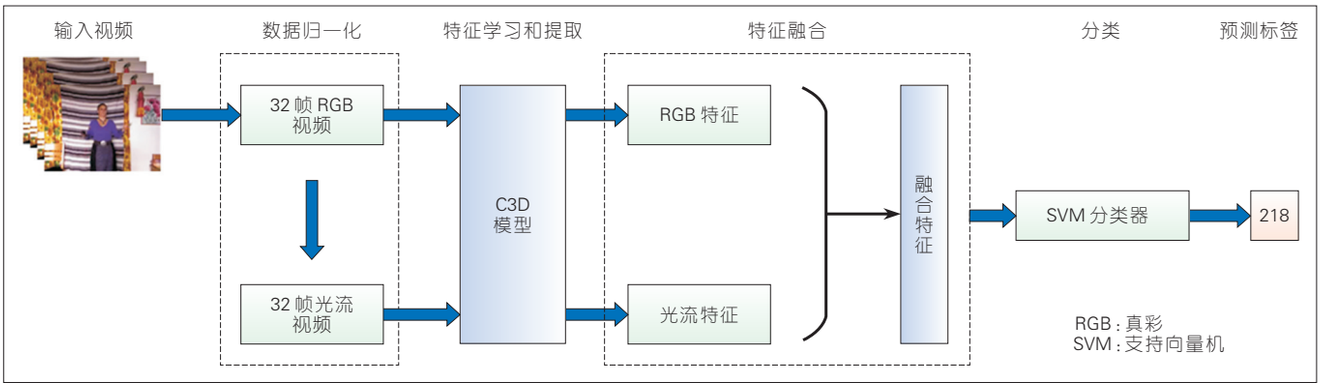
其中，γ是用于平衡两者的权重系数，ΔI和ΔG是视频两帧之间的灰度和梯度增量，$\psi(s^2)$用来增强能量方程的鲁棒性，Ω为积分区间，即整个视频，E_{Smooth}可表述为：

$$E_{Smooth}(u,v)=\int_{\Omega}\psi(|\nabla_3 u|^2+|\nabla_3 v|^2)dx \quad (3)$$

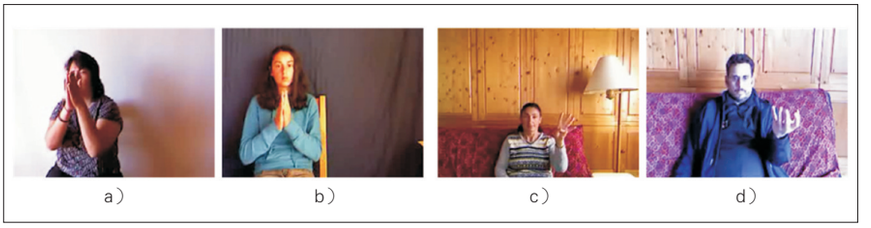
∇_3表示时空平滑约束假设中的时空梯度。通过利用拉格朗日方程和数值近似来最小化该能量函数，获得最终的光流结果。显然，光流数据更加关注运动信息，能够把运动无关的信息全部去除。

1.3 特征提取模型

如前所述，基于深度神经网络的特征提取由于能够更好地体现数据本身的信息，且不像手工特征那样需要研究者具备大量领域相关信息，因



▲图1 基于C3D模型和同源数据融合的大规模手势识别算法流程



▲图2 相似手势举例

而受到了研究者的青睐。文中所述需要提取的特征关注的手势是在视频中，所以解决手势识别任务更多的是依靠提取到的时序特征。因此，我们通过一种三维CNN——C3D模型来实现视频手势特征的自动提取。与二维的CNN相比，三维的CNN针对视频帧序列图像集合，并不仅仅是把视频划分成为帧集合，再用多通道输出到多个图像，而是将卷积核应用到时域，时空域的卷积核特性相互结合，更好地获取视频的特征。

如图3所示，C3D模型包括8个卷积层、5个池化层、2个全连接层来学习特征，和1个softmax层来提供预测的类别。8个卷积层的卷积核个数分别是64、128、256、256、512、512、512和512，卷积核的最佳大小是3×3×3。通过对视频的时空卷积，可以获得在不同尺度上的特征图。在1次或2次卷积操作之后，通过1次池化操作，来对特征进行降采样，以获得更具全局性的特征。在文中，第2~5层的池化层的卷积核大小是2×2×2，而第1个池化层的卷积核大小是1×2×2，以保证在网络中时域信息能够得到最大程度上的保留。在经过多次卷积和池化操作之后，特征图被抽象成一个4 096维的特征向量，用来标记样本的分类信息。

1.4 融合方案

Tran等人结合3个不同的网络提取出的特征来提高准确率。这给了我们灵感：使用特征融合的方法可以提高识别能力。通过实验我们发现：由于RGB视频和光流视频在描述特征的方式并不相同，因此直接将两个视频简单融合，反而不利于正确率的提升。相反，因为特征是视频的抽象，对于C3D提取出的特征向量，可以很好地阐述手势的特点。因此，我们选择了特征级融合。这样做的另一个优势是特征都是相同维度的，统一的格式有助于正确、有效地融合。为了保证两种数据的信息能够同时保留，我们选择通过将两种特征拼接得到高维特征的方式来实现融合。

2 手势识别实验结果和分析

由于目前IsoGD数据集的测试集部分的标签尚未公开，所以文中提到的所有实验和比较都是在该数据集的验证集上进行的。

2.1 实验环境

文中所提到的神经网络训练和特征提取在配有Intel Core i7-6700 CPU @ 3.40 GHz、16 GB内存和Nvidia Geforce GTX TITAN X图形处理器（GPU）的PC上实现，C3D模型依托Linux Ubuntu 16.04长期支持版本（LTS）系统和caffe框架实现32帧视频的生成，特征融合和SVM分类则在Windows 7系统（64 bit）上通过Matlab R2012b软件实现。

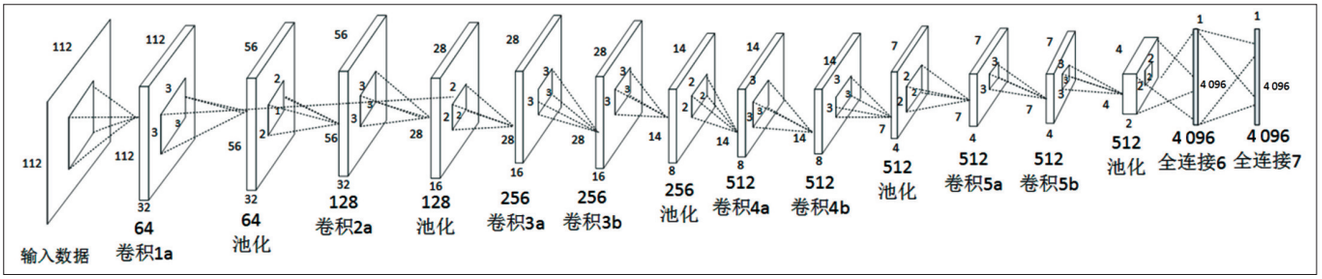
2.2 训练过程

考虑到训练一个较深的网络是很耗时的，特别是在IsoGD这样的大型数据集上，因此我们首先通过Sport-1M（最大的视频分类的标准数据集，包含110万个运动视频，共487类）预训练模型，使其能够适应视频动作分类的应用场景，随后再在实验所需的IsoGD数据集上调参。我们通过随机梯度下降法（SGD）来训练网络：首先将数据打乱，以减少数据排列信息对训练的干扰，在每一次训练的迭代过程中，有10个视频输入网络，网络初始学习率设为0.0001，并且在每5 000次迭代后以10%的比例下降，在10万次迭代后训练停止。

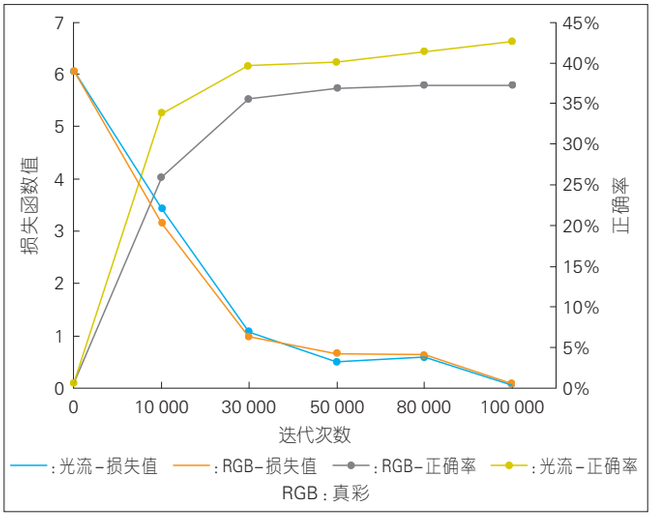
2.3 迭代的影响

作为一个基于学习的方法，迭代次数对分类结果有很大的影响。因此，在这个部分，我们分别在RGB和光流数据的输入上测试不同迭代次数的影响。识别率和损失函数值的变化情况如图4所示。

在训练过程的早期，网络的学习能力较强，损失函数值下降很快，在经过约3万次迭代后，RGB和光流数据的损失函数值都趋于稳定。最后，



▲图3 C3D网络结构



◀图4 RGB与光流数据损失函数值与正确率随迭代次数变化关系

在10万次迭代之后，损失函数值非常接近于0，这时网络已经从训练数据中学到了足够多的东西。另一方面，识别率显示出类似的趋势：正确率在训练早期上升很快。同时，不同数据的特征间的关系也在这个阶段展现出来。在1万次迭代之后，光流数据的优势开始显现出来，而且直到最后一次迭代，光流数据的准确率一直比RGB的高5%左右。

2.4 预处理效果

在本节中，我们验证预处理，即32帧归一化策略的效果。我们对比了输入视频分别为16帧和32帧的结果，这两种输入都是10万次迭代后的结果。

如图5所示，通过我们对输入数据的分析，32帧的归一化策略取得显著的效果，无论是在RGB还是光流数据上，相比16帧的输入，两种数据的32帧的输入都提高了约4%的识别率。这证明更多关于运动路径的信息有助于分辨不同的手势，从而在很大程度上提高了识别率。

2.5 融合结果

我们通过融合RGB和光流特征来提高性能。融合结果和单独RGB、光流数据的比较如图6所示。

相比于任何一种单一特征，融合策略对识别性能的提升都是显著的。融合特征相比单独RGB特征将正确率提升了将近10%，相比单独的光流特征，融合特征也有近5%的提升。这证明了特征融合的策略是行之有效的。

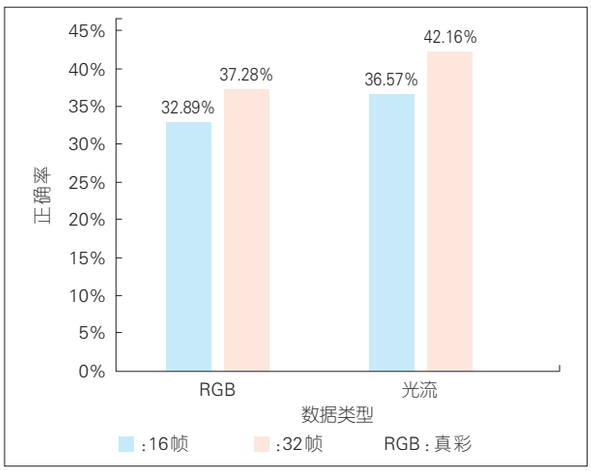


图5▶ 32帧归一化预处理效果

2.6 和传统方法的比较

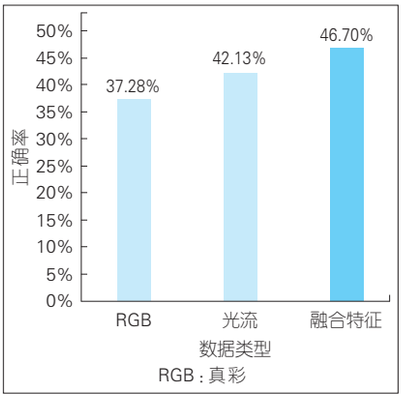
在本节中，我们将我们的方法与Wan等人提出的基于手工制作特征的方法进行对比，从图7中可以看出：CNN在对图片或视频的特征提取方面展示出了极大的优势，我们的方法大概将识别率提高了30%。

2.7 最终结果对比

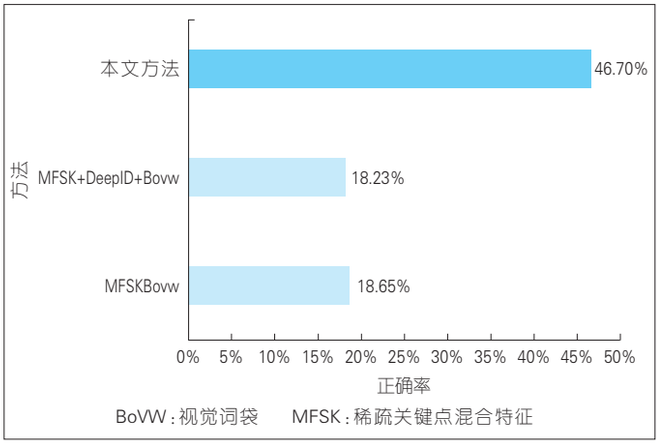
我们的方法和大赛中前3名的方法，以及大赛的基准方法在校验集上的结果对比如图8所示。结果显示：相对于使用手工制作特征的基准方法，我们基于深度学习的方法在特征提取上具有更好的性能。此外，使用光流数据进行去背景处理，使我们的识别率更进一步提升，相对于大赛中的方法，加入光流数据使得准确率提升了4%。

3 结束语

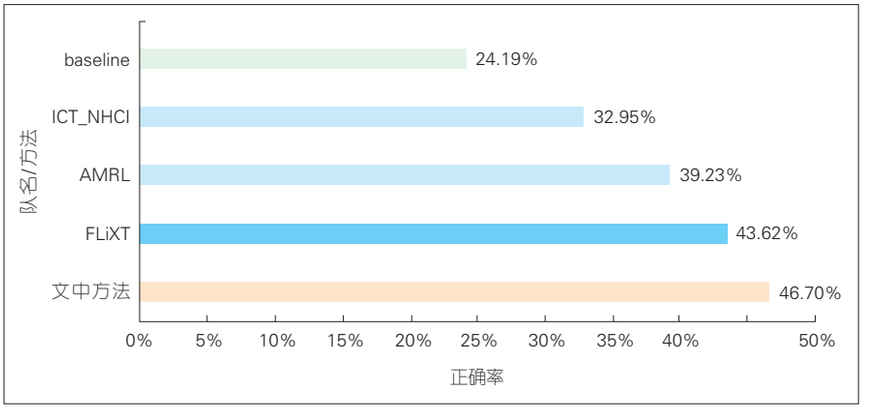
我们提出了一种基于RGB和光流数据及三维CNN的手势识别方法。输入的数据首先被统一成32帧的视频，以便更好地保存动作路径信息；然后，我们通过RGB数据生成了光流数据来去除视频中与手势无关的因素；接着，RGB和光流视频的特征被C3D模型分别提取并加以融合来提高识别性能；之后，我们使用



▲图6 RGB与光流数据融合结果



◀图7 基于C3D的手势识别算法与传统手工特征方法结果比较



▲图8 文中方法与Chalearn LAP大规模手势识别竞赛方法结果比较

SVM分类器进行最终分类。实验证明：我们的策略是有效的，而且我们的方法优于其他现有的技术。相较于Chalearn LAP大规模手势识别竞赛中的方法及传统手工特征方法，我们方法的识别正确率有了较大提升。

然而，仍然存在着很多因素影响着识别率。由于运动信息还不足以区分那些差别细微的类别，还需要学习更多复杂的特征来解决这些问题。同时，还有很多其他的深度学习网络结构，如深度置信网络，在目标识别方面展示出了很大的优势。这些网络结构在视频手势识别方面的使用还值得更多的研究。

致谢

本文的部分实验由西安电子科技大学计算机学院硕士研究生田宽和范莹莹完成，在此对他们表示衷心感谢！

参考文献

[1] STARNER T, PENTL A. Visual Recognition of American Sign Language Using Hidden Markov Models[J]. International Workshop on Automatic Face&Gesture Recognition, 1995(2):189–194
[2] ELMEZAIN M, HAMADI A, MICHAELIS B. Hand Trajectory–Based Gesture Spotting and Recognition using HMM[C]//The 16th IEEE International Conference on Image Processing (ICIP). USA:IEEE,2009:3577–3580. DOI: 10.1109/ICIP.2009.5414322
[3] SGOUROPOULOS K, STERGIOPOULOU E, PAPAMARKOS N.A Dynamic Gesture and Posture Recognition system[J]. Journal of Intelligent&Robotic Systems, 2013(1):1–14
[4] WAN J, RUAN Q, LI W, et al. One–Shot Learning Gesture Recognition from RGB–D Data Using Bag of Features[J]. Journal of Machine Learning Research, 2013, 14(1): 2549–2582
[5] WAN J, RUAN Q, LI W, et al. 3D SMoSIFT: Three–Dimensional Sparse Motion Scale Invariant Feature Transform for Activity Recognition from RGB–D Videos[J]. Journal of Electronic Imaging, 2014, 23(2): 023017–023017
[6] WAN J, RUAN Q, LI W, et al. One–Shot Learning Gesture Recognition from RGB–D Data Using Bag of Features[J]. Journal of Machine Learning Research, 2013, 14(1): 2549–2582
[7] KARPATHY A, TODERICI G, SHETTY S, et al. Large–Scale Video Classification with Convolutional Neural Networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. USA: IEEE, 2014: 1725–1732
[8] SIMONYAN K, ZISSERMAN A. Two–Stream Convolutional Networks for Action Recognition in Videos[C]//Advances in Neural Information Processing Systems. Canda: NIPS, 2014: 568–576
[9]TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3d Convolutional Networks[C]//Proceedings of the IEEE International Conference on Computer Vision. USA: IEEE, 2015: 4489–4497
[10] WAN J, ZHAO Y, ZHOU S, et al. Chalearn Looking at People RGB–D Isolated and Continuous Datasets for Gesture Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. USA:IEEE, 2016: 56–64
[11] GUYON I, ATHITSOS V, JANGYODSUK P, et al. The ChaLearn Gesture Dataset (CGD 2011)[J]. Machine Vision and Applications, 2014, 25(8): 1929–1951
[12] BROS T, BRUHN A, PAPENBERG N, et al. High Accuracy Optical Flow Estimation Based on a Theory for Warping[J]. Computer Vision–ECCV 2004, 2004(3024): 25–36. DOI: 10.1007/978–3–540–24673–2_3

作者简介

苗启广，西安电子科技大学计算机学院教授；主要从事计算机视觉、机器学习、大数据分析方面的研究；主持在研或完成国家核高基重大专项、国家重点研发计划、国家自然科学基金、省自然科学基金、国防预研、国防“863”和“新世纪优秀人才支持计划”项目20余项；在重要期刊上发表论文100余篇。

李宇楠，西安电子科技大学计算机学院博士研究生；主要从事计算机视觉、深度学习方面的研究。

徐昕，西安电子科技大学计算机学院硕士研究生；主要从事计算机视觉、深度学习方面的研究。