Journal of System Simulation

Volume 33 | Issue 5 Article 3

6-9-2021

Multi-view Human Action Recognition Based on Deep Neural Network

Zhao Ying
1. Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing 100081, China; ;2. Teachers College, Beijing Union University, Beijing 100011, China; ;

Lu Yao
1. Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing 100081, China; ;

Zhang Jian
3. School of Computer Science and Engineering, Central South University, Changsha 410083, China;

Qidi Liang
3. School of Computer Science and Engineering, Central South University, Changsha 410083, China;

See next page for additional authors

Multi-view Human Action Recognition Based on Deep Neural Network

Abstract

Abstract: *A novel deep neural network named CNN+CA*(*Convolutional Neural Network plus Context Attention*) *model is constructed and a new recognition algorithm based on sequence matching is presented* to improve the recognition accuracy of MVHAR (Multi-view Human Action Recognition). A CNN(Convolutional Neural Network) is designed to automatically learn multi-view fusion features; the CA (Context Attention) module is introduced to selectively focus on the parts of the features that are relevant for the recognition task; the proposed recognition algorithm based on sequence matching is used to realize MVHAR. The experimental results on the IXMAS dataset and the i3DPost dataset demonstrate that the recognition accuracy of the proposed method is higher than those of the state-of-the-art MVHAR methods.

Keywords

multi-view, human action recognition, convolutional neural network, context attention, sequence matching

Authors

Zhao Ying, Lu Yao, Zhang Jian, Qidi Liang, and Long Wei

Recommended Citation

Zhao Ying, Lu Yao, Zhang Jian, Liang Qidi, Long Wei. Multi-view Human Action Recognition Based on Deep Neural Network[J]. Journal of System Simulation, 2021, 33(5): 1019-1030.

第 33 卷第 5 期 系统仿真学报© Vol. 33 No. 5
2021 年 5 月 Journal of System Simulation May 2021

基于深度神经网络的多视角人体动作识别

赵瑛[1,2]，陆耀[1]，张健[3]，梁启弟[3]，龙炜[1]
(1. 北京理工大学 智能信息技术北京市重点实验室，北京 100081；
2. 北京联合大学 师范学院，北京 100011；3. 中南大学 计算机学院，长沙 410083)

摘要： *为提高多视角人体动作识别的精度，提出了一种新的深度神经网络模型——CNN+CA(Convolutional Neural Network plus Context Attention)模型和一种基于序列匹配的识别方法。*利用卷积神经网络自动学习出多视角融合特征；引入上下文注意力模块自动突出特征中有利于识别的区域，进一步提高特征的判别力；通过基于序列匹配的方法实现人体动作识别。在 IXMAS 数据集和 i3DPost 数据集上的实验结果表明，所提方法在识别精度上超过了其他同类方法。

关键词： 多视角；人体动作识别；卷积神经网络；上下文注意力；序列匹配
中图分类号：TP391 文献标志码：A 文章编号：1004-731X (2021) 05-1019-12
DOI: 10.16182/j.issn1004731x.joss.19-0448

Multi-view Human Action Recognition Based on Deep Neural Network

Zhao Ying[1,2], *Lu Yao*[1], *Zhang Jian*[3], *Liang Qidi*[3], *Long Wei*[1]
(1. Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing 100081, China; 2. Teachers College, Beijing Union University, Beijing 100011, China; 3. School of Computer Science and Engineering, Central South University, Changsha 410083, China)

Abstract: *A novel deep neural network named CNN+CA(Convolutional Neural Network plus Context Attention) model is constructed and a new recognition algorithm based on sequence matching is presented* to improve the recognition accuracy of MVHAR (Multi-view Human Action Recognition). A CNN(Convolutional Neural Network) is designed to automatically learn multi-view fusion features; the CA (Context Attention) module is introduced to selectively focus on the parts of the features that are relevant for the recognition task; the proposed recognition algorithm based on sequence matching is used to realize MVHAR. The experimental results on the IXMAS dataset and the i3DPost dataset demonstrate that the recognition accuracy of the proposed method is higher than those of the state-of-the-art MVHAR methods.

Keywords: multi-view; human action recognition; convolutional neural network; context attention; sequence matching

引言

人体动作识别是计算机视觉的研究热点之一，在智能视频监控(Smart Video Surveillance)、环境辅助生活(Ambient Assisted Living)、高级人机交互(Advanced Human-Computer Interaction)等领域有着重要的应用价值。许多研究者已经提出了各种有效的人体动作识别方法，但这些方法很多只是在单视角环境下开展研究，属于单视角人体动作识别。单视角人体动作识别要求测试阶段要具有和训练阶段相同的拍摄视角，否则识别精度就会受到影响。然而，在实际应用中，拍摄视角经常会有变化，人体朝向也会发生变化，因此，单视角人体动作识别

收稿日期：2019-08-26 修回日期：2019-10-08
基金项目：国家自然科学基金(61273273)；国家重点研发计划(2017YFC0112001)
第一作者：赵瑛(1977-)，女，博士，副教授，研究方向为人体行为分析、机器学习、智慧教育。E-mail：sftzhaoying@buu.edu.cn
通讯作者：陆耀(1958-)，男，博士，教授，研究方向为神经网络、图像和信号处理、模式识别。E-mail：vis_yl@bit.edu.cn

很难满足实际应用中对视角变化的鲁棒性要求。为了解决这个问题，近年来很多研究者开始把目光投向多视角人体动作识别(Multi-view Human Action Recognition，MVHAR)，即利用多个摄像机分别从不同的视角拍摄同一场景中的人体动作以获得多视角视频，然后利用多视角视频来开展人体动作识别。很显然，由于从多视角视频中可以抽取到更加丰富的特征，所以 MVHAR 有望取得比单视角人体动作识别更高的识别精度；而且，MVHAR 还有望通过不同视角之间的信息互补，来解决单视角人体动作识别中难以解决的环境遮挡和人体自遮挡问题。

然而，MVHAR 却也面临着诸多挑战。例如，在 MVHAR 中，由于人体的自遮挡，所以即使是同一个动作实例，从不同的视角进行拍摄，所获得的人体表观差异是很大的。为了充分利用各个视角上的信息来提高识别精度，MVHAR 需要进行多视角融合。

近年来，深度神经网络(Deep Neural Network)被广泛应用于图像分类、语音识别、自然语言理解、目标识别等领域，而且取得了很大成功。在人体动作识别领域也出现了利用深度神经网络进行人体动作识别的研究。

通常，用深度神经网络学习得到的特征图是很难解释的。近年来，研究者们提出了一些注意力模型(Attention Model)，通过捕获“值得注意”的位置，为深度神经网络学习得到的特征图增加了一个可解释的维度。注意力模型在单幅静态图片的主题生成、机器翻译、行人检测、人体动作识别等领域显示出了很好的应用效果。在人体动作识别领域，注意力模型可以让算法关注每帧中“值得注意”的位置来推断正在发生的动作。例如，Sharma 等[1]用长短时记忆网络(Long Short Term Memory，LSTM)构建了一个用于人体动作识别的注意力模型，该模型能够检测出每一帧中“值得注意”的位置。Jaderberg 等[2]提出了一种叫做空间变换器(Spatial Transformer)的注意力模型，加在卷积神经网络(Convolutional Neural Network，CNN)的卷积层之间，并在 Street View House Numbers 数据集上达到了目前最好的识别效果。

本文用深度神经网络和注意力模型开展了 MVHAR 研究，提出了 CNN+CA(Convolutional Neural Network plus Context Attention)模型，实现了多视角融合特征的自动学习和自动突出。该模型由 CNN 和 CA 两个模块组成。CNN 模块能够从由同一时刻来自不同视角的感兴趣区域(Region of Interest，ROI)拼接而成的 ROI 拼接图中自动学习多视角融合特征得到卷积特征图；CA 模块能够自动突出卷积特征图中有利于动作识别的区域，进一步提高特征的判别力。

在此基础上，本文提出了一种新的基于序列匹配的动作识别方法。在 CNN+CA 模型的训练阶段，得到了训练集中每幅 ROI 拼接图的预测标签，进而将训练集中的每个多视角动作实例都用 ROI 拼接图的预测标签序列来表示；在测试阶段，利用已经训练好的 CNN+CA 模型，得到待识别多视角动作实例的每幅 ROI 拼接图的预测标签，然后将这个待识别的多视角动作实例也用 ROI 拼接图的预测标签序列来表示；最后通过序列匹配，从训练集中找出与之最近邻的预测标签序列。这个最近邻的预测标签序列所表示的多视角动作实例的动作类别就是最终的识别结果。

在 IXMAS 数据集和 i3DPost 数据集上的实验结果表明，本文所提方法能有效提高 MVHAR 的识别精度。

1 相关工作介绍

研究者们已经提出了一些 MVHAR 方法。根据特征抽取方式的不同，这些方法大体可以分为两类：人工设计特征(Handcrafted Features)的方法、自动学习特征(Automatic Learning Features)的方法。

人工设计特征的方法通常包括 3 个步骤：特征抽取、特征表示和分类。首先，先从视频中抽取人

DOI: 10.16182/j.issn1004731x.joss.19-0448

工设计的特征，例如：时空兴趣点(Spatio-Temporal Interest Points，STIP)、方向梯度直方图(Histogram of Oriented Gradients，HOG)、方向光流直方图(Histogram of Oriented Optical Flow，HOOF)、运动历史图(Motion History Image，MHI)、稠密轨迹和运动能量图等；然后利用傅里叶变换、K-means 聚类等方法对这些特征进行处理，构建出具有更强判别力的描述符；最后用分类器对这些描述符进行分类实现动作识别。

Aryanfar 等[3]先对每个视角的人体轮廓图进行边缘检测得到人体边缘；再计算每个人体边缘点到轮廓中心的距离组成每一帧的特征向量，这样每个视角上的人体动作就用一个特征矩阵来表示了，矩阵上的一行代表一帧的特征向量；然后把每个视角的特征矩阵合并在一起形成多视角特征矩阵；之后对这个多视角特征矩阵运用二维离散小波变化实现降维；最后利用支持向量机(Support Vector Machine, SVM)和贝叶斯分类器组合在一起形成的层次分类器实现了 MVHAR。

Liu 等[4]则是在每个视角上以人体形状为中心串联每一帧的 ROI 得到视频的子体积(Subvolume)；然后在子体积中进行时空立方体的采样得到低层描述符；再将采样位置相邻的时空立方体的低层描述符合并在一起，得到多视角低层描述符；之后利用多任务随机森林法学习得到多视角的中层描述符并以此表示动作；最后利用随机森林分类器实现了 MVHAR。

这些人工设计的特征是依据众多视觉理论设计出来的，融入了人类的先验知识，具有计算效率高、不依赖于大量样本训练的优点[5]。然而，这些人工设计的特征通常是根据数据的特点精心设计的，对特定数据集中的数据通常具有较强的表现力和判别力；而当数据集发生变化时，如果不进行重新设计和训练，可能就达不到很好的效果。所以人工设计特征泛化能力不强，在情况复杂多变的实际应用中难以达到很好的效果[6]。此外，在人工设计特征的方法中，特征抽取与分类训练被分割成两个独立的阶段依次进行，这导致每个训练样本的分类误差很难及时反向传播给特征抽取器，使得特征抽取过程缺乏分类结果的及时指导，不利于特征抽取质量的提高[6]。

近年来，基于深度神经网络的方法在语音识别、自然语言理解和单视角人体动作识别等领域都取得了很大的突破与成就。用深度神经网络开展人体动作识别属于自动学习特征的方法，因为深度神经网络的反向传播特性，能够使特征的构造过程自动化。与人工设计特征的方法不同，自动学习特征的方法将特征抽取器和动作分类器级联，形成一个前后联动的一体化工作链，这样分类器可以将分类误差及时反向传播给特征抽取器，用于指导特征抽取器的参数调节，从而使特征抽取器能自动抽取出更高质量的特征来提高分类器的识别精度。然而，在 MVHAR 领域，深度神经网络的应用仍然是少量的[6]。

Putra 等[6]提出了一种适用于 MVHAR 的深度神经网络模型，该模型由 3 部分组成：CNN、多堆叠长期短期记忆残差(Multiple Stacked Long Short-Term Memory Residual，MSLSTMRes)和稠密层(Dense Layers)。该模型把特征抽取器和动作分类器整合在一起，使得特征抽取器能够根据动作分类器的误分类结果及时调整特征抽取参数。所以该模型尽管以原始视频作为输入，但却取得了和依靠人工设计特征的传统方法相媲美的识别精度。

为了解决由于视角变化而导致识别精度下降的问题，Chuanxu 等[7]用深度学习和 K-SVD 稀疏算法实现了高精度的 MVHAR。首先用 CNN 和 LSTM 网络从不同的视角抽取特征图。这些特征图是具有语义信息的多视角高层特征。然后他们用 K-SVD 稀疏算法学习得到每个视角上的字典。这些字典对动作特征能非常好地进行稀疏表示。最后他们用 softmax 分类器实现了动作识别。

Kavi 等[8]把 CNN 的自动特征抽取能力和

LSTM 网络的时域记忆能力整合在一起，提出了适用于 MVHAR 的 CNN+LSTM 模型。该模型包括 3 个卷积层、两个稠密层、两个 LSTM 和 Softmax 分类器，并以 256×256 的灰度图作输入。算法在一个多摄像头收集到的驾驶员行为识别系统上进行评测，算法在使用视角融合的条件下分类精度取得了提升，并好于传统方法。然而，该算法并没有进行留一受试者(Leave-One-Actor-Out，LOAO)测验，所以算法是否支持跨受试者尚不确定。

2 CNN+CA 模型

本文致力于用深度神经网络自动学习多视角融合特征并自动对特征中有利于动作识别的区域进行突出，为此，本文构建了一个新的深度神经网络——CNN+CA 模型。该模型由 CNN 和 CA 两个模块组成，如图 1 所示。该模型有 L 个输入单元，分别代表 L 个视角；有 M 个输出单元，分别代表 M 个动作类别。多视角动作实例各个视角的视频经过预处理，变成统一尺寸的 ROI 序列；各个视角的 ROI 序列是该模型的输入；模型先将各个视角对应时刻的 ROI 拼接起来，形成该时刻的 ROI 拼接图，再输入 CNN 模块；CNN 模块从 ROI 拼接图中自动学习多视角融合特征得到卷积特征图；然后将卷积特征图输入 CA 模块，自动突出卷积特征图中有利于动作识别的区域。

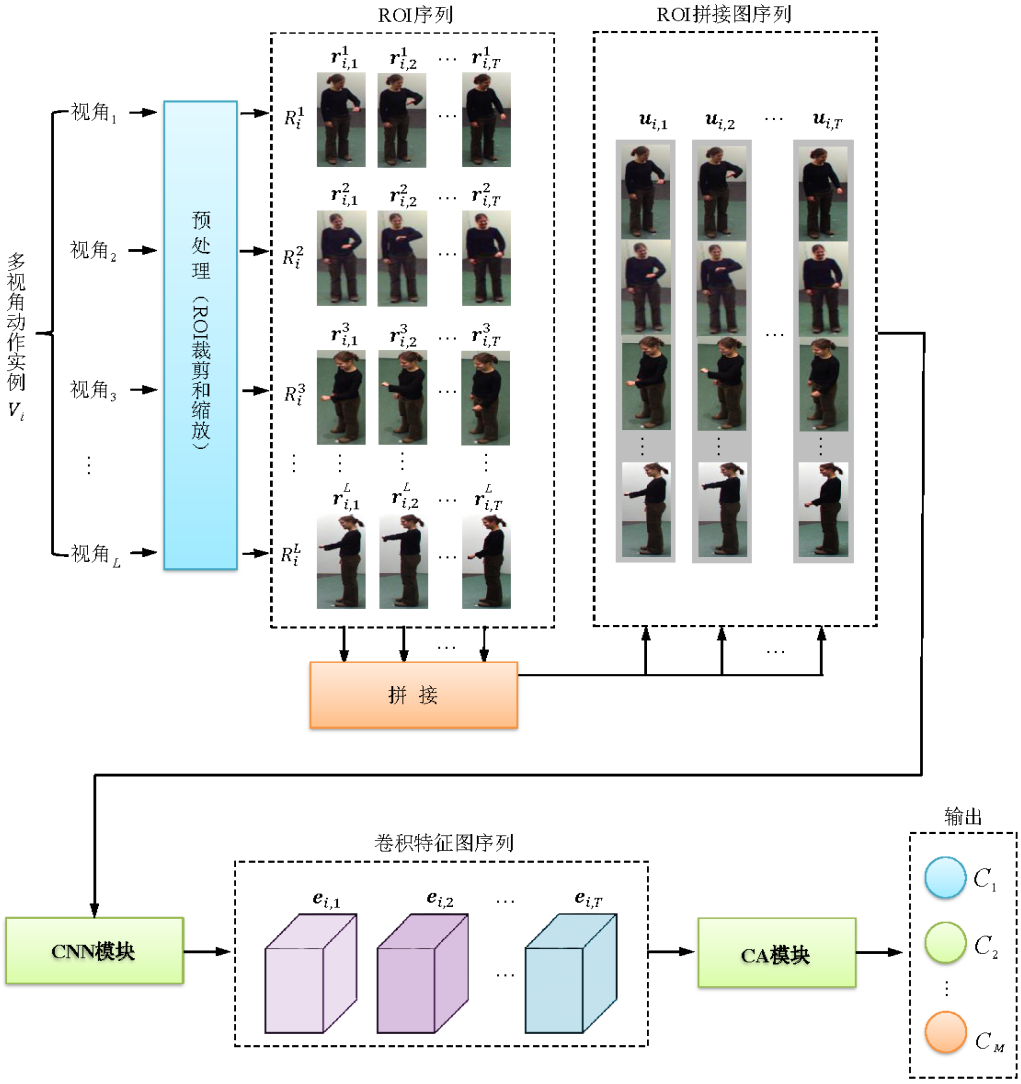

图 1 CNN+CA 模型的结构
Fig. 1 Structure of CNN+CA model

https://dc-china-simulation.researchcommons.org/journal/vol33/iss5/3
DOI: 10.16182/j.issn1004731x.joss.19-0448

ROI 拼接图中包含了同一行为主体在不同视角上的表观信息；ROI 拼接图输入 CNN 模块后得到的卷积特征图反映了每幅 ROI 拼接图的空间特征；将卷积特征图序列输入 CA 模块，利用时序的上下文信息，可以使行为主体执行动作的身体部位得到关注。图 2 给出了 CA 模块关注后的可视化效果图。图中，单数列表示原始 ROI 拼接图；偶数列表示经过CA模块关注后的效果图。在效果图中，白色区域就是被关注的区域，该区域是人体执行“看手表”这个动作时的身体部位(即左手)，白色区域亮度越高说明受关注的程度越高。

图 2　CA 模块关注的可视化效果图
Fig. 2　Visualization effect images after being concerned by CA module

给定用于训练的多视角动作实例集 $\{\boldsymbol{V}_i, i\in(1,N)\}$，其中，$\boldsymbol{V}_i=\{v_i^j, j\in(1,L)\}$ 表示第 i 个多视角动作实例，它有 L 个视角，v_i^j 表示第 i 个多视角动作实例在 j 个视角的视频，N 是用于训练的多视角动作实例总个数，模型训练的步骤如下：

step 1：预处理。将训练集中每个多视角动作实例的每个视角的视频用统一尺寸的 ROI 序列表示。给定一个视频 v_i^j，使用统一尺寸的边框对该视频的每一帧进行裁剪，得到它们的 ROI；然后再对训练集中所有 ROI 进行缩放操作，使其都变成 n 行、W 列。经过预处理，训练集中每个视角的视频都用一个 ROI 序列 $\boldsymbol{R}_i^j=(\boldsymbol{r}_{i,1}^j,\boldsymbol{r}_{i,2}^j,...,\boldsymbol{r}_{i,T}^j)$ 来表示，其中，T 表示视频 v_i^j 的总帧数，$\boldsymbol{r}_{i,t}^j, t\in(1,T)$ 表示视频 v_i^j 第 t 帧的 ROI。在本文中，n=128, W=128。

step 2：拼接。将多视角动作实例各个视角对应时刻的 ROI 拼接起来，形成该时刻的 ROI 拼接图，然后将该实例用 ROI 拼接图序列来表示。给定一个用 ROI 序列表示的多视角动作实例 $\boldsymbol{V}_i^{'}=\{\boldsymbol{R}_i^j, j\in(1,L)\}$，将该实例每个视角对应时刻的 ROI 拼接起来形成一幅该时刻的 ROI 拼接图。设 $\boldsymbol{u}_{i,t}$ 表示该实例在 t 时刻的 ROI 拼接图，则 $\boldsymbol{u}_{i,t}$ 是由 $\{\boldsymbol{r}_{i,t}^j, j\in(1,L)\}$ 拼接而成的，且 $\boldsymbol{u}_{i,t}$ 的尺寸是 H 行、W 列，其中 H=n×L。这样，多视角动作实例 $V_i^{'}$ 就可以用一个 ROI 拼接图序列 $\boldsymbol{U}_i=(\boldsymbol{u}_{i,1},\boldsymbol{u}_{i,2},...,\boldsymbol{u}_{i,T})$ 来表示了。将训练集中的所有 ROI 拼接图序列组织在一起，形成 ROI 拼接图序列集 $\{\boldsymbol{U}_i, i\in(1,N)\}$。

step 3：自动学习特征。用 CNN 模块从 ROI 拼接图中自动学习多视角融合特征，得到相应的卷积特征图。将 ROI 拼接图序列集 $\{\boldsymbol{U}_i, i\in(1,N)\}$ 输入 CNN 模块，通过卷积、池化等操作得到卷积特征图序列集 $\{\boldsymbol{E}_i, i\in(1,N)\}$，其中 $\boldsymbol{E}_i=(\boldsymbol{e}_{i,1},\boldsymbol{e}_{i,2},...,\boldsymbol{e}_{i,T})$ 是 ROI 拼接图序列 $\boldsymbol{U}_i=(\boldsymbol{u}_{i,1},\boldsymbol{u}_{i,2},...,\boldsymbol{u}_{i,T})$ 的卷积特征图序列，$\boldsymbol{e}_{i,t}, t\in(1,T)$ 是ROI拼接图 $\boldsymbol{u}_{i,t}, t\in(1,T)$ 的卷积特征图。每幅卷积特征图有 D 个通道，每个通道有 $K\times L$ 行、K 列。在本文中，D=512, K=4。CNN 模块的网络结构如图 3 所示，采用交叉熵损失函数，参数如表 1 所示，每一层的卷积核大小都是 3×3，步长都是 1。

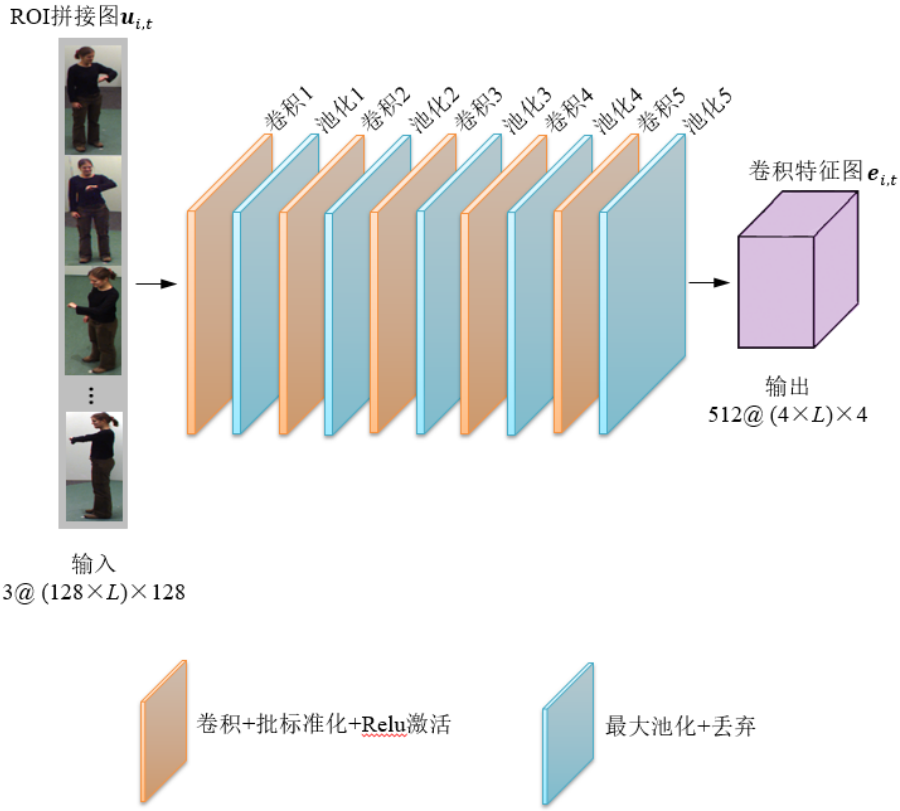

图 3 CNN 模块的网络结构
Fig. 3 Network Structure of CNN module

表 1 CNN 模块的参数
Tab. 1 Parameters of CNN model

层	输入	输出
卷积 1	3@(128×L)×128	32@(128×L)×128
池化 1	32@(128×L)×128	32@(64×L)×64
卷积 2	32@(64×L)×64	64@(64×L)×64
池化 2	64@(64×L)×64	64@(32×L)×32
卷积 3	64@(32×L)×32	128@(32×L)×32
池化 3	128@(32×L)×32	128@(16×L)×16
卷积 4	128@(16×L)×16	256@(16×L)×16
池化 4	256@(16×L)×16	256@(8×L)×8
卷积 5	256@(8×L)×8	512@(8×L)×8
池化 5	512@(8×L)×8	512@(4×L)×4

step 4：自动突出特征中有利于识别的区域：将 CNN 模块输出的卷积特征图序列集输入 CA 模块，自动突出特征图中有利于动作识别的区域。给定一幅卷积特征图 $\boldsymbol{e}_t$，将该特征图各通道上对应位置的元素组织在一起形成一个 D 行、1 列的特征切片 $\boldsymbol{x}_{t,k}$，得到该特征图的特征切片表示：

$$\boldsymbol{e}_t^{'} = \{\boldsymbol{x}_{t,k}, k \in (1, K \times L \times K)\} \tag{1}$$

式中：每个特征切片都分别代表特征图中一个相应的位置区域。CA 模块将为卷积特征图 $\boldsymbol{e}_t$ 学习出一幅注意力地图 A_t，用来突出卷积特征图 $\boldsymbol{e}_t$ 中有利于动作识别的区域。受文献[1]的启发，本文设计了如图 4 所示的 CA 模块。和文献[1]一样，该模块也由 3 个 LSTM 单元组成，但是每个 LSTM 单元所采用的神经元个数和文献[1]不一样，从下往上，3 个 LSTM 单元所用的神经元个数分别是 128，256 和 100。每个 LSTM 单元，本文采用文献[1]所提的方式来实现。

$$\begin{bmatrix} l_t \\ f_t \\ o_t \\ g_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \mathbb{M} \begin{bmatrix} h_{t-1} \\ X_t \end{bmatrix} \tag{2}$$

$$c_t = f_t \odot c_{t-1} + l_t \odot g_t \tag{3}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4}$$

式中：l_t 为输入门；f_t 为遗忘门；o_t 为输出门；g_t 为门控状态，用公式(2)进行计算；$\mathbb{M}$ 是由可训练

https://dc-china-simulation.researchcommons.org/journal/vol33/iss5/3
DOI: 10.16182/j.issn1004731x.joss.19-0448

参数 $m=d+D$ 和 $b=4d$ 组成的从 m 维空间到 d 维空间的仿射变换，其中 d 是 l_t, f_t, o_t, g_t, c_t 和 h_t 的维度；σ 为 sigmoid 激活；c_t 为 cell 状态，其初始化公式如公式(5)；h_t 为隐藏状态，其初始化公式如公式(6)。

$$c_0 = f_{\text{init},c}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{K\times L\times K}\sum_{k=1}^{K\times L\times K}\boldsymbol{x}_{t,k}\right)\right) \tag{5}$$

$$h_0 = f_{\text{init},h}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{K\times L\times K}\sum_{k=1}^{K\times L\times K}\boldsymbol{x}_{t,k}\right)\right) \tag{6}$$

式中：$f_{\text{init},c}$ 和 $f_{\text{init},h}$ 为 2 个多层感知器；T 为 CA 模块中的时刻总数，在本文中，它等于一个 ROI 拼接图序列的长度。

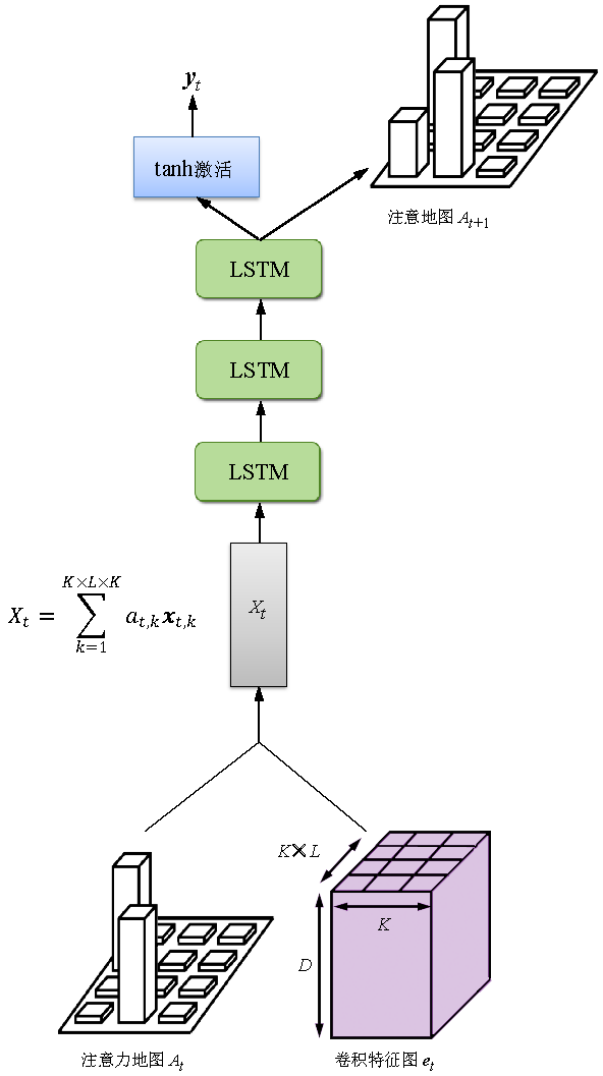

图 4　CA 模块的结构
Fig. 4　Structure of CA module

$A_t = \{a_{t,k}, k \in (1, K\times L\times K)\}$ 是 t 时刻的注意力地图，其中，$a_{t,k}$ 是对 t 时刻第 k 个位置的重要性的预测概率值，其计算公式如公式(7)所示。

$$a_{t,k} = \exp(W_k^T h_{t-1}) \Big/ \sum_{\beta=1}^{K\times L\times K}\exp(W_\beta^T h_{t-1}) \tag{7}$$

式中：W_k 为一个权重，对应 A 中的第 k 个元素。$a_{t,k}$ 值越大，说明这位置对动作识别越重要。

X_t 是从 t 时刻所有位置的特征切片中计算得出的对下一时刻输入的期望值，是 CA 模块在 t 时刻的输入，其计算公式如公式(8)所示。

$$X_t = \sum_{k=1}^{K\times L\times K} a_{t,k}\boldsymbol{x}_{t,k} \tag{8}$$

在每个时刻 t，CA 模块要通过 Softmax 分类器输出 2 个预测：A_{t+1} 和 $\boldsymbol{y}_t$。其中，$A_{t+1} = \{a_{t+1,k}, k \in (1, K\times L\times K)\}$ 是 $t+1$ 时刻的注意力地图；$\boldsymbol{y}_t$ 则是对 t 时刻的 ROI 拼接图在 M 个动作类别标签上的预测概率分布。

本文对每一幅原始视频帧进行裁剪得到 ROI，用 ROI 拼接图序列作为模型的输入，而不是直接将各个视角的原始视频帧拼接起来形成原始视频帧拼接图序列来作为模型的输入，是为了通过 ROI 裁剪过滤掉一些环境噪声，更好地把注意力放到“值得注意”的人体动作区域上，而不是背景环境区域上，以此提高识别效果。

3　基于序列匹配的动作识别

经过训练，CNN+CA 模型具有了自动抽取多视角融合特征并且突出特征中有利于动作识别区域的能力，而且还为训练集中的每幅 ROI 拼接图 $\boldsymbol{u}_{i,t}$ 都输出了一个在动作类别标签集上的预测概率分布 $\boldsymbol{y}_{i,t}$。本文利用最近邻法，为每幅 ROI 拼接图 $\boldsymbol{u}_{i,t}$ 都找出最近邻的动作类别标签，称之为预测标签，记为 $s_{i,t}, t \in (1,T)$，并用预测标签 $s_{i,t}$ 来表征 ROI 拼接图 $\boldsymbol{u}_{i,t}$。这样，就得到了训练集中每个多视角动作实例 V_i 的预测标签序列表示 $S_i = (s_{i,1}, s_{i,2}, ..., s_{i,T})$，其中 $s_{i,t}=C_\alpha$，C_α 是第 α 类动作的类别标签，$\alpha \in (1, M)$。

给定一个待识别的多视角动作实例 V^{test}，先采用和训练阶段一样的步骤，利用已经训练好的 CNN+CA 模型，为该实例的每幅 ROI 拼接图输出一个在动作类别标签上的预测概率分布 $\boldsymbol{y}_t^{\text{test}}$；然后用最近邻法，为该实例的每幅 ROI 拼接图找出最近邻的动作类别标签作为该 ROI 拼接图的预测标签，并用预测标签来表征该 ROI 拼接图；在此基础上，为该实例构建出它的预测标签序列表示 $S^{\text{test}}=(s_1^{\text{test}},s_2^{\text{test}},...,s_T^{\text{test}})$，其中 $s_t^{\text{test}},t\in(1,T)$ 代表该多视角动作实例 t 时刻 ROI 拼接图的预测标签，$s_t^{\text{test}}=C_\alpha$，$C_\alpha$是第$\alpha$类动作的类别标签，$\alpha\in(1,M)$。这样，多视角动作识别问题就转化成了预测标签序列的匹配问题。

由于不同的受试者执行动作的方式存在差异，即使是执行相同的动作，有的受试者执行得快，而有的则执行得慢。所以，相同的动作被不同的受试者执行，所产生的预测标签序列，其时间长度是不一致的。为了能够有效进行序列比较和匹配，采用了动态时间归整(Dynamic Time Warping，DTW)法，先将 2 个不同时间长度的序列进行对齐，然后再计算 2 个序列之间的距离。

给定 2 个预测标签序列 $S^{\text{train}}=(s_1^{\text{train}},s_2^{\text{train}},...,s_t^{\text{train}},...,s_T^{\text{train}})$ 和 $S^{\text{test}}=(s_1^{\text{test}},s_2^{\text{test}},...,s_\gamma^{\text{test}},...,s_\Gamma^{\text{test}})$，用式(9)计算它们之间的 DTW 距离 $S^{\text{train}}-S^{\text{test}}$：

$$S^{\text{train}}-S^{\text{test}}=\text{dtw}(T,\Gamma) \tag{9}$$

$$\text{dtw}(t,\gamma)=\min\begin{Bmatrix}\text{dtw}(t-1,\gamma)\\ \text{dtw}(t,\gamma-1)\\ \text{dtw}(t-1,\gamma-1)\end{Bmatrix}+\delta(s_t^{\text{train}},s_\gamma^{\text{test}}) \tag{10}$$

式中：

$$\delta\left(s_t^{\text{train}},s_\gamma^{\text{test}}\right)=\begin{cases}0 & \text{当}s_t^{\text{train}}\text{和}s_\gamma^{\text{test}}\text{相同时}\\ 1 & \text{其他}\end{cases}$$

这样，通过为 S^{test} 计算到训练集中每个预测标签序列 $S_i,i\in(1,N)$ 之间的 DTW 距离，就可以从训练集中找出与之最近邻的一个预测标签序列，该序列所代表的多视角动作实例的动作类别标签就是这个待识别的多视角动作实例的最终识别结果。

4 实验结果及其比较

为验证所提算法的有效性，本文在公开的多视角人体动作识别数据集 IXMAS[9]和 i3DPost[10]上进行了实验验证。在实验中，用每一帧的原始图像作输入。实验结果表明，本文所提的方法比其他方法识别精度更高。

4.1 数据集

IXMAS 数据集是用于验证 MVHAR 算法有效性的最流行的数据集之一。它包含 14 种动作，每种动作分别由 12 个受试者执行 3 次，每个动作实例都由 5 台摄像机从 5 个不同的视角(四周各 1 个和顶部 1 个)进行拍摄。该数据集所包含的 14 种动作分别是：看手表、抱胳膊、抓头、坐下、起来、转身、走、挥手、拳击、踢、指、捡、从下方扔和过头扔。为了公平比较，本文与文献[3,4,6, 11-13]一样，在实验中丢弃了“指”、“从下方扔”和“过头扔”这 3 种动作，而只保留了其余的 11 种动作(如图 5 所示)。在 IXMAS 数据集中，虽然 5 个视角上的摄像机都是固定的，但是由于受试者可以自由选择自己的位置和朝向，所以识别难度较大。

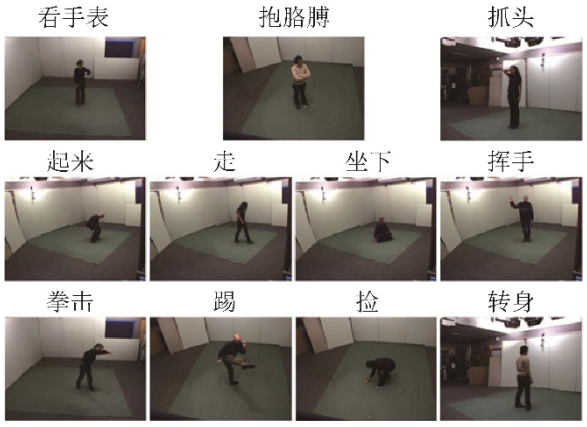

图 5 IXAMS 数据集各种动作的帧示例图

Fig. 5 Example frames of the human actions in the IXMAS dataset

i3DPost 数据集也是一个用于验证 MVHAR 算法有效性的常用数据集。它包括 8 种动作，如图 6 所示，分别是：走、跑、原地跳、向前跳、弯腰、

DOI: 10.16182/j.issn1004731x.joss.19-0448

跌倒、坐在椅子上和单手挥。每种动作由 8 个受试者执行，每个动作实例由 8 台摄像机从不同的角度进行拍摄。这 8 台摄像机布置在拍摄现场离地面 2 m 高的空中、呈环形排列。

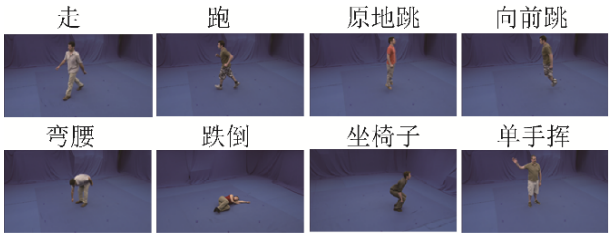

图 6　i3Dpost 数据集各种动作的帧示例图
Fig. 6　Example frames of the human actions in the i3DPost dataset

4.2 实验设置

验证过程中，采用了 LOAO 交叉验证方案，因为 LOAO 是 MVHAR 研究者们常用的验证方案(如文献[1-3,7,9,12-18])。实验中，本文随机把 1 位受试者的多视角视频留出来用作测试集，其余受试者的多视角视频用作训练集，进行训练和测试得出识别精度；这样周而复始，直到把所有受试者都留出来做过测试集为止。IXMAS 数据集共有 12 位受试者，本文进行了 12 次这样的训练和测试，最终的识别精度是这 12 次实验所得出的识别精度的平均值。同理，i3DPost 数据集共有 8 位受试者，共进行了 8 次训练和测试，最终的识别精度是这 8 次实验所得出的识别精度的平均值。

利用 IXMAS 数据集提供的人体轮廓图，先计算出每一帧的人体轮廓边框；然后，对每一个视频，从该视频每一帧的人体轮廓边框中找出一个尺寸最大的边框；之后，将这个尺寸最大的边框与该视频每一帧的人体轮廓边框进行中心点对齐，再用这个最大尺寸的边框对该原始视频帧进行裁剪，得到该原始视频帧的 ROI。当所有视频的 ROI 都裁剪完成之后，再进行缩放操作，将所有视频每一帧的 ROI 都缩放成 n=128 行、W=128 列大小。对 i3DPost 数据集，利用该数据集提供的每个视角的背景图，先进行背景差分，得到每一帧的人体轮廓图，再用和 IXMAS 数据集一样的预处理方式，得到所有视频每一帧的 128×128 大小的 ROI。

实验在 TensorFlow 框架下用 Python 语言开发。运行平台是内存 160 G 的 GPU 服务器、显卡型号是 Titan Xp、操作系统是 Ubuntu。Softmax 加交叉熵损失作为网络优化的目标函数。优化算法使用 Adam，初始学习率设为 1e-5。整个网络的训练周期为 40 个 epoch。Dropout 层的丢失率为 0.95。整个训练过程中未使用数据增广的策略。

4.3 实验结果

实验在 IXMAS 数据集上达到了 93.6%的识别精度。图 7 给出算法在 IXMAS 数据集上的混淆矩阵。从混淆矩阵可以看出，一些由手或胳膊执行的动作，例如，“看手表”、“抱胳膊”、“抓头”、“挥手”等由于从外形上看有较大的相似性，所以误识别率较高。另外，“转身”基本上是沿着一个特定的方向行“走”，所以算法也产生了一些误识别。但是，由于算法在识别阶段采用了序列匹配，考虑到了 ROI 拼接图之间的时序关系，所以对“起来”和“坐下”这 2 种姿态变化刚好相反的动作达到了完美的识别精度。另外，算法对“踢”、“捡”和“走”等动作也达到了完美的识别精度。

	1	2	3	4	5	6	7	8	9	10	11
1	0.9	0	0	0	0	0.1	0	0	0	0	0
2	0.1	0.9	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0.1	0	0.9	0	0	0	0	0
7	0	0	0	0	0	0	0.8	0	0	0	0.2
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0.9	0.1	0
10	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0.1	0	0	0	0.9

1：看手表　2：抱胳膊　3：起来　4：踢　5：捡　6：拳击　7：抓头　8：坐下　9：转身　10：走　11：挥手

图 7　IXMAS 数据集的混淆矩阵
Fig. 7　Confusion matrix for the IXMAS dataset

表 2 给出了在 IXMAS 数据集上本文所提方法与其他方法在识别精度上的比较。从比较结果可以看出，本文所提方法在识别精度上超过了其他同类方法。表中的 Putra 等[6]也用了基于深度神经网络的方法，但是 Putra 等[6]是直接用原始视频帧作输入的，而且没有引入上下文注意力模块来突出特征中有利于动作识别的区域。本文用在原始视频帧上裁剪出 ROI 区域作为输入，初步过滤掉了一些背景噪声，让算法把注意力放到 ROI 区域中；然后又引入了 CA 模块，将用 CNN 网络学习得到的特征图中对动作识别有利的区域进行突出，进一步提高了特征的判别力，得到了比 Putra 等[6]更高的识别精度。

表 2 IXMAS 数据集识别精度的比较
Tab. 2 Comparison of recognition accuracy on the IXMAS dataset

方法	动作类别数量	视角数量	识别精度/%
Chuanxu 等，2019[7]	13	5	89.2
Putra 等，2018[6]	11	5	90.2
Liu 等，2018[4]	11	5	90.3
Sargano 等，2016[11]	11	5	89.8
Zhang 等，2016[12]	11	5	84.3
Aryanfar 等，2015[3]	11	5	88.1
Liu 等，2014[13]	11	5	88.0
本文所提方法	11	5	93.6

实验在 i3DPost 数据集上达到了 96.3%的识别精度。混淆矩阵如图 8 所示。

从混淆矩阵中可以看出，本文所提算法在绝大多数动作上都达到了完美的识别精度，识别误差较大的是“向前跳”这种动作。由于在裁剪 ROI 的时候，采用了轮廓中心对准的方式，这种方式丢失了一些人体在空间位置上的变化信息，从而导致在 ROI 序列中，“向前跳”这种动作与“原地跳”很相似。表 3 给出了在 i3DPost 数据集上本文所提方法与其他方法在识别精度上的比较。从比较结果可以看出，本文所提方法的识别精度最高。

图 8 i3DPost 数据集的混淆矩阵
Fig. 8 Confusion matrix for the i3DPost dataset

表 3 i3DPost 数据集识别精度的比较
Tab. 3 Comparison of recognition accuracy on the i3DPost dataset

方法	识别精度/%
Zhang 等，2018[14]	94.6
Chalearnnetkul 等，2018[15]	93.0
Iosifidis 等，2014[16]	95.9
Sadek S 等，2013[17]	81.7
Iosifidis 等，2012[18]	94.4
Holte 等，2011[19]	92.2
本文所提方法	96.3

为了验证CA模块对提高特征判别力进而提高识别精度的有效性，本文去除了 CNN+CA 模型中的 CA 模块，只留下 CNN 模块，仍然用 LOAO 方案在 IXMAS 数据集上开展了 MVHAR 实验，结果发现识别精度下降到了 87.3%。用 CNN+CA 模型和去除 CA 模块后只用 CNN 模块在每种动作上的识别精度比较如图 9 所示。从图 9 可以看出，CNN 模块加上 CA 模块之后，各种动作的识别精度都得到了提高。特别是“抱胳膊”、“挥手”这种动作幅度不是很大的动作，由于增加了 CA 模块，使特征图中有利于动作识别的区域得到了突出，因而识别精度得到了明显提高。

CA 模块的设计，本文参考文献[1]，也采用了 3 个 LSTM 单元。为了检验 LSTM 单元个数对识别

精度的影响，分别用由 1 个 LSTM 单元、2 个 LSTM 单元和 3 个 LSTM 单元构成 CA 模块，在 IXMAS 数据集上进行了实验验证。实验结果如图 10 所示。从图 10 可以看出，随着 LSTM 单元个数的增加，每种动作的识别精度都有了提升。当只用 1 个 LSTM 单元时，平均识别精度只有 89.2%；用 2 个 LSTM 单元时，平均识别精度是 92.3%。

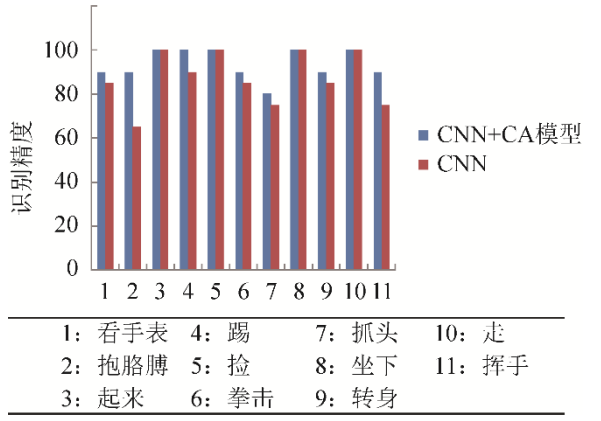

图 9　用 CNN+CA 模型和单用 CNN 模块在 IXMAS 数据集上识别精度的比较

Fig. 9　Comparison of recognition accuracy on the IXMAS dataset using CNN+CA model and CNN module only

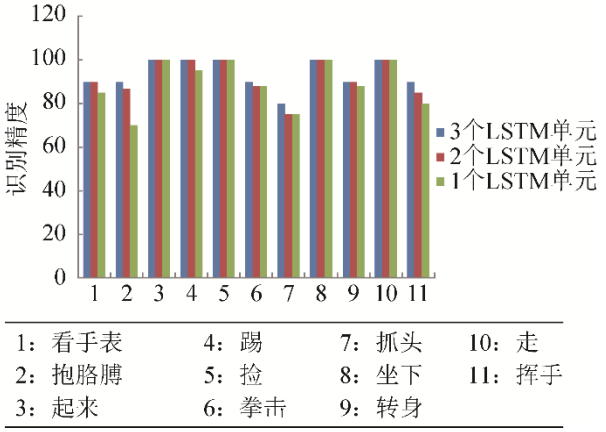

图 10　由不同个数 LSTM 单元构成的 CA 模块在 IXMAS 数据集上识别精度的比较

Fig. 10　Comparison of recognition accuracy on the IXMAS dataset using CA Module Composed of LSTM Units with Different Numbers

5　结论

MVHAR 由于可以利用来自多个视角的信息实现视野互补，因而可以克服单视角人体动作识别中难以克服的环境遮挡与人体自遮挡问题；而且可以利用来自多个视角的丰富信息提高识别精度。因此，有较高的研究和应用价值。

与人工设计特征的方法相比，用深度神经网络方法可以实现特征的自动学习，而且利用深度神经网络的反向传播特性，可以将分类结果误差及时反向传播回来用于指导特征学习过程中的参数调整，从而使网络能学到更具表现力和判别力的特征。

本文将深度学习方法应用到 MVHAR 中，构建了一个新的深度神经网络——CNN+CA 模型，实现了多视角融合特征的自动学习和自动突出。在此基础上，提出了一种基于序列匹配的动作识别方法。在 IXMAS 数据集和 i3DPost 数据集上的实验结果表明，本文所提方法是有效的。

参考文献：

[1] Sharma S, Kiros R, Salakhutdinov R. Action Recognition Using Visual Attention[C]// International Conference on Learning Representations Workshop. San Juan, Puerto Rico, USA: Springer, 2016: 1-11.

[2] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial Transformer Networks[C]// Advances in Neural Information Processing Systems. Montréal CA: Springer, 2015: 2017-2025.

[3] Aryanfar A, Yaakob R, Halin A A, et al. Multi-view Human Action Recognition Using Wavelet Data Reduction and Multi-Class Classification[J]. Procedia Computer Science (S1877-0509), 2015, 100(62): 585-592.

[4] Liu C, Li Z, Shi X, et al. Learning a Mid-level Representation for Multi-View Action Recognition[J]. Advances in Multimedia (S1687-5680), 2018: 1-10.

[5] Zhen X, Shao L, Maybank S, et al. Handcrafted vs. Learned Representations for Human Action Recognition[J]. Image and Vision Computing (S0262-8856), 2016, 55(2): 39-41.

[6] Putra P U, Shima K, Shimatani K. Markerless Human Activity Recognition Method Based on Deep Neural Network Model Using Multiple Cameras[C]// IEEE International Conference on Control, Decision and Information Technologies. Thessaloniki, Greece: IEEE, 2018: 13-18.

[7] Chuanxu W, Guofeng H, Yun L. Multi-views Action

Recognition on Deep Learning and K-SVD[J]. Journal of Physics Conference Series (S1742-6588), 2019, 1176(6): 062015.

[8] Kavi R, Kulathumani V, Rohit F, et al. Multiview Fusion for Activity Recognition Using Deep Neural Networks[J]. Journal of Electronic Imaging (S1017-9909), 2016, 25(4): 043010.

[9] Weinland D, Ronfard R, Boyer E. Free Viewpoint Action Recognition Using Motion History Volumes[J]. Computer Vision and Image Understanding (S1077-3142), 2006, 104(2/3): 249-257.

[10] Gkalelis N, Kim H, Hilton A, et al. The i3DPost Multi-View and 3D Human Action/Interaction Database[C]// IEEE Conference for Visual Media Production. London, UK: IEEE, 2009: 159-168.

[11] Sargano A, Angelov P, Habib Z. Human Action Recognition From Multiple Views Based on View-invariant Feature Descriptor Using Support Vector Machines[J]. Applied Sciences (S2076-3417), 2016, 6(10): 309-322.

[12] Zhang J, Zhang L, Shum H P H, et al. Arbitrary View Action Recognition Via Transfer Dictionary Learning on Synthetic Training Data[C]// IEEE International Conference on Robotics and Automation. Stockholm, Sweden: IEEE, 2016: 1678-1684.

[13] Liu C W, Pei M T, Wu X X, et al. Learning a Discriminative Mid-Level Feature for Action Recognition[J]. Science China-Information Sciences (S1862-2836), 2014, 57(5): 1-13.

[14] Zhang J, Shum H P, Han J, et al. Action Recognition From Arbitrary Views Using Transferable Dictionary Learning[J]. IEEE Transactions on Image Processing (S1057-7149), 2018, 27(10): 4709-4723.

[15] Chalearnnetkul P, Suvonvorn N. Multiview Layer Fusion Model for Action Recognition Using RGBD Images[J]. Computational Intelligence and Neuroscience (S1687-5265), 2018: 1-22.

[16] Iosifidis A, Tefas A, Pitas I. Regularized Extreme Learning Machine for Multi-view Semi-supervised Action Recognition[J]. Neurocomputing (S0925-2312), 2014, 145(5): 250-262.

[17] Sadek S, Al-Hamadi A, Krell G, et al. Affine-Invariant Feature Extraction for Activity Recognition[J]. ISRN Machine Vision (S2090-7796), 2013(1): 1-7.

[18] Iosifidis A, Tefas A, Nikolaidis N, et al. Multi-view Human Movement Recognition Based on fuzzy Distances and Linear Discriminant Analysis[J]. Computer Vision and Image Understanding (S1077-3142), 2012, 116(3): 347-360.

[19] Holte M B, Moeslund T B, Nikolaidis N, et al. 3D Human Action Recognition for Multi-view Camera Systems[C]// International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission. Hangzhou, China: IEEE, 2011: 342-349.

https://dc-china-simulation.researchcommons.org/journal/vol33/iss5/3
DOI: 10.16182/j.issn1004731x.joss.19-0448