跨层融合与多模型投票的动作识别

罗会兰* 卢 飞 严 源

(江西理工大学信息工程学院 赣州 341000)

摘 要：针对动作特征在卷积神经网络模型传输时的损失问题以及网络模型过拟合的问题，该文提出一种跨层融合模型和多个模型投票的动作识别方法。在预处理阶段，借助排序池化的方法聚集视频中的运动信息，生成近似动态图像。在全连接层前设置对特征信息进行水平翻转结构，构成无融合模型。在无融合模型的基础上添加第2层的输出特征与第5层的输出特征融合结构，构造成跨层融合模型。训练时，对无融合模型和跨层融合模型两种基本模型采用3种数据划分方式以及两种生成近似动态图像顺序进行训练，得到多个不同的分类器。测试时使用多个分类器进行预测，对它们得到的结果进行投票集成，作为最终分类结果。在UCF101数据集上，提出的无融合模型和跨层融合模型的识别方法与动态图像网络模型的方法相比，识别率有较大提高；多模型投票的识别方法能有效缓解模型的过拟合现象，增加算法的鲁棒性，得到更好的平均性能。

关键词：动作识别；跨层融合；多模型投票；近似动态图像；水平翻转

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2019)03-0649-07

DOI: 10.11999/JEIT180373

Action Recognition Based on Multi-model Voting with Cross Layer Fusion

LUO Huilan LU Fei YAN Yuan

(*School of Information Engineering*, *Jiangxi University of Science and Technology*, *Ganzhou* 341000, *China*)

Abstract: To solve the problem of the loss in the motion features during the transmission of deep convolution neural networks and the overfitting of the network model, a cross layer fusion model and a multi-model voting action recognition method are proposed. In the preprocessing stage, the motion information in a video is gathered by the rank pooling method to form approximate dynamic images. Two basic models are presented. One model with two horizontally flipping layers is called “non-fusion model”, and then a fusion structure of the second layer and the fifth layer is added to form a new model named “cross layer fusion model”. The two basic models of “non-fusion model” and “cross layer fusion model” are trained respectively on three different data partitions. The positive and negative sequences of each video are used to generate two approximate dynamic images. So many different classifiers can be obtained by training the two proposed models using different training approximate dynamic images. In testing, the final classification results can be obtained by averaging the results of all these classifiers. Compared with the dynamic image network model, the recognition rate of the non-fusion model and the cross layer fusion model is greatly improved on the UCF101 dataset. The multi-model voting method can effectively alleviate the overfitting of the model, increase the robustness of the algorithm and get better average performance.

Key words: Action recognition; Cross layer fusion; Multi-models voting; Approximate dynamic image; Horizontal flip

收稿日期：2018-04-24；改回日期：2018-11-02；网络出版：2018-11-12

*通信作者： 罗会兰 luohuilan@sina.com

基金项目：国家自然科学基金(61462035, 61862031)，江西省青年科学家培养项目(20153BCB23010)，江西省自然科学基金(20171BAB202014)

Foundation Items: The National Natural Science Foundation of China (61462035, 61862031), The Young Scientist Training Project of Jiangxi Province (20153BCB23010), The Natural Science Foundation of Jiangxi Province (20171BAB202014)

1 引言

视频中的动作识别是当今计算机视觉和人工智能中最活跃和最具挑战性的研究领域之一。它具有很多潜在的应用，比如视频监控与安全、人机交互、智能家居、视频检索、虚拟现实和医疗监控等。在过去的几年里，动作识别研究经历了传统的特征工程到深度网络工程的发展历程。传统的特征工程主要是提取与动作直接关联的特征进行动作识别，依靠领域专家为某一项目精心设计的特征提取方法，设计一个最适合当前任务数据的表示作为输入特征。而深度学习工程则是利用卷积神经网络自主提取动作特征进行分类识别。卷积神经网络的成功应用使得视频中的动作识别研究取得突破性进展，它模拟人的视觉系统逐层提取图像中的多级特征，具有较好的容错性和自适应性。

比较流行的传统动作识别方法主要有流形学习法、轨迹法、堆叠法等等。利用流形学习进行动作识别的方法就是从提取的高维数据特征中寻找低维流形结构进行识别分类。文献[1]对提取到的特征向量使用等距映射方法降维，大大减少识别算法的处理时间。Qu等人[2]对等距映射算法进行改进，添加自适应距离因子，再结合最近邻分类器进行动作识别，获得了更优的识别效果。在光流特征基础上提出的轨迹类方法充分利用时间维度信息进行动作识别。Wang等人[3]借助动作的运动轨迹、方向梯度直方图(Histograms of Oriented Gradients, HOG)、光流直方图(Histograms of Optical Flow, HOF)和运动边界直方图(Motion Boundary Histograms, MBH)构造稠密轨迹特征描述子进行动作识别。Wang等人[4]进一步对轨迹特征的正则化方式和特征编码方式进行优化，极大地提高了识别效果。Ohnishi等人[5]在文献[4]的基础上借助空间流卷积特征与时间权重做点积的交叉流以及时间卷积特征与空间权重做点积的交叉流减弱背景运动轨迹对识别结果的影响。堆叠法[6,7]能更好地解决背景影响问题，如运动能量图像(Motion Energy Image, MEI)[6]通过聚集动作的空间位置变化进行动作识别，充分利用了动作变化的轮廓和能量的空间分布。在运动能量图的基础上生成运动历史图(Motion History Image, MHI)，将运动的变化以亮度的形式显示，但需要设置持续时间和衰退参数。Bilen等人[7]借助排序池化操作[8]将一段视频的运动信息表征到一张动态图像中进行识别分类。该方法编码视频时空演变的参数，有效地聚集了视频的运动信息。

近些年，采用卷积神经网络模型进行动作识别取得了较大的成功。经典的动作识别网络模型有双流模型、三流模型和多流模型。其中双流模型主要是对视频分别提取空间流和时间流特征，并对每个流进行训练，在分类识别输出前采用各种方法对流特征进行融合。如Simonyan等人[9]提出了一种双流模型，利用卷积神经网络分别提取基于时空流的运动信息和基于空间流的外观信息，在分类前将两个流的得分进行融合以提高动作识别率。Liu等人[10]受3维手势识别方面[11]的启发，在双流模型的基础上将2维结构扩充到3维，捕获3维信号的相关性同时提取时间和空间维度的特征，识别效果有所改进。此外，比较流行的双流模型还有隐藏双流卷积神经网络[12]、双流循环卷积神经网络[13]等等。部分研究者通过改进双流结构得到3流神经网络模型[14]，如Shi等人[14]分别提取视频的空间流，局部时间流和全局时间流构造3流神经网络模型进行动作识别。文献[14]在传统双流的基础上结合长时运动描述符进行动作识别，识别率较好。与双流、3流模型不同的是，多流模型则是对视频提取更多的运动信息，如运动流信息、外观流信息等等，Song等人[15]提出多模型多流深度学习框架进行动作识别，即同时使用多流卷积神经网络学习空间和时间特征，再利用长短时记忆框架模型从多个传感器中学习特征，最后通过两级池化融合技术来计算预测结果。文献[16]探索一种基于神经网络的多形式手势识别方法，使用递归方法融合长短期记忆模型与递归神经网络模型中的多流特征，进一步提高了动作识别的效果。Bilen等人[7]提出了多流动态图像网络模型，对每张动态图像分配一个卷积池化结构网络，分别提取多个状态的特征信息，在最后的卷积层借助排序池化操作聚集所有局部特征图，以便于后续识别分类。

一些研究者通过采用集成学习思想来提高动作识别的准确度，如朱丽等人[17]将随机森林和朴素贝叶斯两种分类器进行融合，实验结果表明可显著提高动作的识别率，稳定性大大加强。文献[18]借助集成的思想将卷积神经网络模型和自动编码器通过训练得到的权重进行融合，识别效果较单一动作模型提升明显，还避免了为获得单个模型最优的动作识别率而进行的多次调参处理步骤。

动作特征信息在深度网络模型[19]中传输时都存在一定的特征损失问题，特别是具有判别性的特征损失对最终的识别分类影响最大，因而如何处理深度网络模型中的传输损失是一个需要解决的问题。针对特征的传输损失问题以及模型的鲁棒性问题，本文旨在结合近似动态图像方法以及卷积神经网络

模型方法的优势，提出了一种采用近似动态图像及多个卷积神经网络模型投票的识别算法。与文献[7]不同的是，为了避免近似排序池化过度压缩视频信息，本文视频预处理阶段对近似动态图像进行水平翻转操作，增加卷积神经网络训练的数据量；在全连接层前也添加水平翻转结构，使得输入全连接层的特征数据量翻倍。针对网络模型中的传输损失问题，受He等人[19]在残差网络模型中全等映射结构的启发，本文对动态图像网络模型[7]进行改进，添加跨式结构，使前层特征与后层特征进行融合。最后为进一步增强模型系统的鲁棒性，采用集成学习[20]的思想，设计方法生成不同训练数据集，并利用提出的2种深度网络模型在生成的不同训练数据集上，学习得到多个分类器，建立投票机制对测试结果进行集成，进一步增强模型系统的稳定性，提高动作识别率。

2　提出的方法

本文提出了一种基于近似动态图像及多个卷积神经网络模型投票集成的动作识别方法。首先将训练集视频按照部分重叠方法分成多个子视频帧序列，借助近似排序池化(Approximate Rank Pooling, ARP)函数聚集子视频帧序列运动信息，得到近似动态图像，然后对近似动态图像进行水平翻转，这样近似动态图像数据量增加1倍，使卷积神经网络模型得到充分训练。在卷积神经网络模型的构成方面，首先对动态图像网络模型进行改进，在将提取到的特征信息输入到全连接层前，设置特征信息水平翻转结构，得到无融合基础模型；为保持特征信息在传输过程的完整性，在无融合模型的基础上添加第2层的输出特征与第5层的输出特征融合结构，生成跨层融合模型。其次，将无融合基础模型和跨层融合模型分别在多种训练数据划分方式上，训练学习得到多个分类器，通过投票集成进行动作识别。为方便描述，本文提出的利用近似动态图像和多个卷积神经网络模型进行投票动作识别的算法简称为近似动态多模型投票识别算法(Voting with Approximate Dynamic and Multi-Model Recognition algorithm, VADMMR)。

2.1　近似排序池化与近似动态图像

文献[8,21]指出具有相似动态信息的视频都可用同一个排序函数来表示，这个排序函数可编码视频帧顺序模拟视频的时空演变，捕获视频的时空动态信息，实验验证了排序函数方法能有效地表征一个视频。为了更好地聚集视频中的运动信息，减少视频数据的冗余，提高网络模型参数训练的效率，本文采用排序函数的方法生成近似动态图像。

假设有一段共包含N帧的视频$I_1, I_2, \cdots, I_t, \cdots, I_N$，其中$I_t$表示视频$I$中的第$t$帧图像。用式(1)计算从第1帧到第$t$帧的平均特征向量$\boldsymbol{V}_t$，其中$\psi(I_t) \in \Re^D$表示第$t$帧图像的特征向量。

$$\boldsymbol{V}_t = \frac{1}{t}\sum_{\tau=1}^{t}\psi(I_\tau) \tag{1}$$

通过优化式(2)学习一个参数向量$\boldsymbol{d}^*$来表示视频信息，其中$S(t|\boldsymbol{d}) = \boldsymbol{d}^{\mathrm{T}}\cdot\boldsymbol{V}_t$用于计算视频第$t$帧的得分，即用向量$\boldsymbol{d}$与到$t$时刻为止的动作特征向量平均值$\boldsymbol{V}_t$的点积作为$I_t$的得分。

$$\left.\begin{aligned}\boldsymbol{d}^* &= \rho(I_1, I_2, \cdots, I_N; \psi) = \arg\min_{\boldsymbol{d}} E(\boldsymbol{d})\\ E(\boldsymbol{d}) &= \frac{\lambda}{2}||\boldsymbol{d}||^2 + \frac{2}{N(N-1)}\\ &\quad\times\sum_{q>t}\max\{0, 1 - S(q|\boldsymbol{d}) + S(t|\boldsymbol{d})\}\end{aligned}\right\} \tag{2}$$

学习到的最优参数向量$\boldsymbol{d}^*$包含了可对视频帧进行排序的信息，同时也聚集了视频帧中所有的运动信息，因此$\boldsymbol{d}^*$可看作是视频的描述符。向量$\boldsymbol{d}^*$具有与每帧特征向量相同的维度，可以看成是一个特征图像，称为视频的动态图像。虽然借助现代高性能的计算机可进行精确的动态图像运算，但会导致程序的运行时间长且内存消耗大。Bilen等人[7]在文献[8]的基础上对排序池化操作进行优化，提出近似排序池化方法，该方法可有效地加快视频动态图像的生成。近似排序池化操作就是对式(2)进行一次梯度下降迭代求解，作为最优值$\boldsymbol{d}^*$的近似。

2.2　跨层融合网络模型

文献[7]对动态图像采用卷积神经网络模型进行特征提取，为使网络模型参数得到充分训练，本文提出在视频的预处理阶段对生成的近似动态图像进行水平翻转，使训练和验证的数据量翻倍。即在将近似动态图像输入到卷积神经网络前进行水平翻转，将翻转前后的近似动态图像都输入网络进行训练。添加近似动态图像翻转结构的网络模型框架图如图1所示，其中ARP表示近似排序池化层，CONV表示卷积层，TEMP-POOL表示最大池化聚集层，用于聚集属于同一个视频的所有子视频获得的卷积特征，FC表示全连接层，FH(Flip Horizontal)表示水平翻转结构。本文还在全连接层前设置水平翻转结构，使得进入全连接层的特征矢量翻倍。为方便后续描述，本文将提出的如图1所示的模型结构简称为无融合模型，方便与后续有跨层融合的模型区分开，它增加了两个水平翻转层。该方

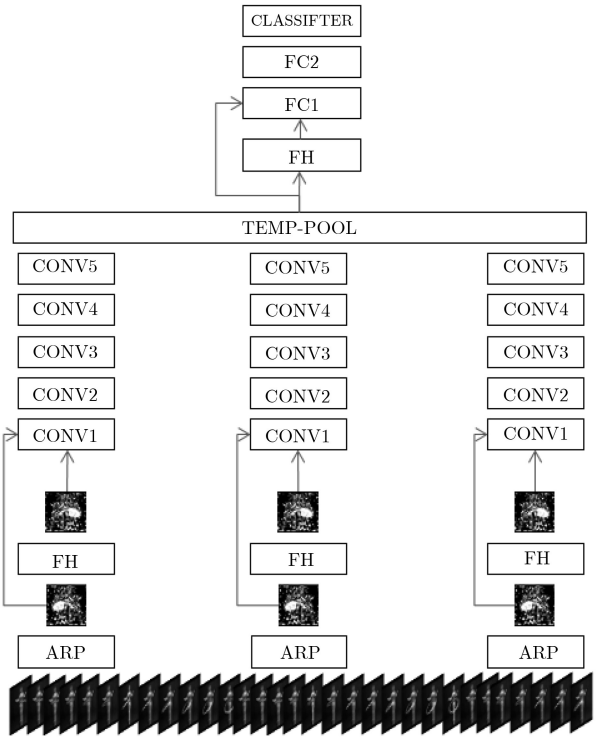

图 1 无融合模型

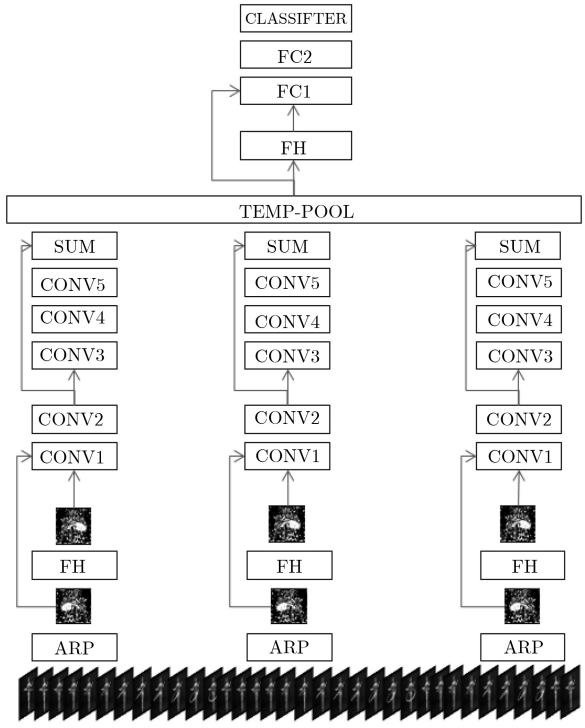

图 2 跨层融合模型

法可直接增加全连接层训练特征数据量的输入，且更充分表达了近似动态图像的判别性特征。

为了进一步提高模型的识别率，受He等人[19]残差网络模型结构的启发，本文对提出的无融合模型添加跨层融合结构，构成跨层融合模型。依据文献[22]中特征的可视化分析方法可知，网络模型的前两层卷积层可提取到的特征主要为颜色和边缘等最底层的特征信息，而第3个卷积层提取到的特征以纹理特征信息为主，第4个卷积层提取的特征开始比较有可区别性，第5层提取的特征最完整，而且是比较关键的判别性特征。依据网络模型各层提取到的特征信息的特点，为解决深度卷积神经网络模型传输过程中的特征损失问题，本文将第2个卷积层的输出特征与第5个卷积层的输出特征相融合，保证特征信息的完整性。鉴于相融合的两张特征图含有的特征信息不一样，本文实验部分对特征图融合的权重问题进行了实验分析和探讨。在无融合模型基础上提出的跨层融合卷积神经网络模型如图2所示，其中的SUM表示两层特征信息通过加权和的方式进行融合。

2.3 投票识别模型的建立

为增强模型的稳定性，本文借助集成学习的思想，训练多个分类器进行动作识别。本文提出一种融合多个结构相似的模型进行动作识别，以图1的无融合模型和图2中的跨层融合模型为基本模型，通过输入不同的训练数据进行训练，得到多个分类器。首先对训练数据集进行3种不同类型的划分，分别表示为split1, split2和split3，例如：某个动作类有15个训练视频样本，本文实验部分的split1的划分方式是前10段视频用作训练集，后5段用作验证集，同理split2是前5段和后5段视频用作训练集，中间5段用作验证集，而split3是使用前5个样本做验证集，其他作训练集。然后，用正序和反序，两种顺序将视频帧输入到近似排序池化中生成近似动态图像。虽然同一视频序列包含的视频帧相同，但输入的顺序不同，生成的近似动态图像也不同。这样利用3种训练数据划分方法，2种生成近似动态图像的顺序，对2种基础模型分别进行训练，就可以学习到12个分类器。在训练得到多个动作分类器后，将测试视频输入就可以得到多个分类结果，本文提出使用均值投票的方法进行集成，得到最终的分类结果。即每个动作类的最终输出结果是12个分类器中对应动作类的输出结果的均值。

3 实验结果与分析

本文使用公开数据集UCF101[22]进行实验。UCF-101数据集由101个人体动作类组成，共有13320段视频，其中训练集包含9537段视频，测试集包含3783段视频。每个视频都在无约束的环境下进行拍摄的，部分视频包含摄像机的抖动、光照条件变化、运动背景以及动作遮挡等因素。

3.1 实验设置

本文参照文献[7]中生成动态图像的参数设置方法，在将视频序列分割成子视频时，将帧数和步长分别设置为10和6。

为充分验证算法的有效性，本文采用交叉验证方法进行训练识别。首先对UCF101训练数据集进行3种不同的划分，分别表示为split1, split2和split3。每种数据划分方式将全部训练视频数据以7: 3的比例分为训练集与验证集，但采用不同训练与验证数据选择方法，从而得到不同的训练数据输入方式。具体为：每个动作类共有25组训练视频，其中split1将前面7组视频用作验证，剩下18组视频作为训练；split2将第8组到第14组视频用作验证，其余用作训练；split3则将第15组到第21组视频用作验证，剩下视频用作训练。

3.2 跨层融合权重设置实验分析

在提出的跨层融合模型中，参与融合的两个不同特征层包含的特征信息是不一样的，因而进行跨层融合时需要考虑两个特征图融合的权重。为了更简单地分析不同融合权重设置方法对后续识别结果的影响，本实验在去掉两层翻转结构的跨层融合模型基础上，分析比较了不同的加权融合方法对训练过程的影响，以及它们在验证集上的平均识别准确度，实验结果如表1所示。实验中分析比较了4种加权融合方法，即第2层的融合权重分别设置为0.50, 0.25, 0.20和0.10，对应的第5层特征图的权重为0.50, 0.75, 0.80和0.90，分别表示为“融合0.50”、“融合0.25”、“融合0.20”、“融合0.10”。在数据集划分split3的验证集上进行识别验证比较4种不同权重融合方法的平均识别率如表1所示。从表1可知，随着第2层的融合权重降低，验证的平均识别准确度逐渐升高，当第2层特征图的权重为0.10时，跨层融合的识别效果最优，因此本文后续的对比实验中，在跨层融合时，将第2层权重设置为0.10，第5层权重设置为0.90。

表 1　4种不同权重融合模型的平均识别准确度(%)

模型	融合0.50	融合0.25	融合0.20	融合0.10
平均准确度	53.89	63.12	63.94	64.82

3.3 跨层融合模型识别效果实验分析

由于UCF101数据集有101个动作类，主要包括5个方面：人体运动、人与物互动、人与人互动、乐器表演和体育竞赛运动。跨层融合模型在5种典型动作上的识别准确度如表2所示。由表2可知，学习到的6种基于跨层融合模型的分类器在同一动作类视频下表现各异。在同一种数据划分方式下，因视频帧正序和反序的输入不同训练得到的分类器效果差别不大。在不同的数据划分方式下，即使视频帧输入顺序相同，同一动作类的识别率也有较大差别，如呼啦圈动作类中split1正序的识别率要比split2正序的识别率要高近10%。不过，像军队前进类和弹吉他类，这两个动作类在不同数据划分和视频帧输入顺序下识别率变化不大。跨层融合模型在不同训练方法下得到的6个分类器的平均识别准确度为82.90%。

表 2　跨层融合模型动作识别准确度(%)

动作类	转呼啦圈	键盘打字	军队行进	弹吉他	掷铁饼	类平均
split1+正序	**87.14**	80.40	87.14	91.33	77.45	82.47
split1+反序	86.29	79.63	**87.90**	**91.65**	76.86	82.16
split2+正序	77.28	88.35	86.64	89.29	73.60	83.06
split2+反序	76.66	88.88	86.27	90.88	71.31	**83.87**
split3+正序	78.72	**89.25**	87.02	91.21	**78.20**	83.03
split3+反序	78.91	86.46	86.99	90.66	76.65	82.79

注：粗体数字代表动作类中识别率最高，带下划线数字代表动作类的识别率次高。

3.4 投票集成模型识别效果实验分析

本实验首先在5种经典的动作类上分析比较了本文提出的多模型投票集成动作识别方法(VADMMR)的识别准确度，实验结果如表3所示，在5种经典的动作类方面，掷铁饼类的动作识别准确度较低，为79.83%；弹吉他类的动作识别准确度最高，为91.58%；在这5类上的平均识别准确度为84.67%，略高于跨层融合模型。

表4所示是本文提出的多模型投票集成动作识别方法VADMMR与其它先进方法的比较结果，比较的是在UCF101数据集的测试集上对101个类测试的平均准确度结果。与2014年比较经典的双流模型相比，本文提出的多模型投票集成动作识别方法VADMMR的识别准确度比Spatial Stream ConvNet[9]网络高11.67%，比Temporal Stream ConvNet[9]约高1%。与2017年的Spatial-C3D方法[23]相

表 3 VADMMR在5类动作上的识别准确度(%)

动作类	转呼啦圈	键盘打字	军队行进	弹吉他	掷铁饼	类平均
VADMMR	83.77	87.43	88.83	91.58	79.83	84.67

表 4 本文提出的VADMMR与其它动作识别方法对比

文献	技术策略	年份	平均识别率(%)
文献[9]	Spatial Stream ConvNet	2014	73.0
文献[9]	Temporal Stream ConvNet	2014	83.7
文献[24]	Composite LSTM	2015	84.3
文献[7]	动态图像网络(MDI)	2016	70.9
文献[23]	Spatial-C3D	2017	83.6
本文方法	VADMMR	2018	84.67

比，本文的方法约提高1%。其中，Composite LSTM[24]使用了LSTM来学习视频特征表达，结合自动编解码网络和预测网络，得到了最好的识别率，但是它使用了相对复杂的模型来学习视频信息表达。文献[23]在双流模型的基础上引入了注意机制，通过引入从时域网络到空间网络的交叉连接层，指导空间流多注意前景人的动作信息，获得了不错的效果。文献[7]将视频表达成一张动态图像，这样就可以使用适合于图像处理的任何网络模型来学习动作分类，表4中文献[7]的结果是将动态图像输入简单的Alexnet网络进行训练的结果，相较其他方法，文献[7]的模型简单，训练速度快，但只用一张动态图像来表达整个视频信息，带来了视频信息的较大损失，所以效果比不上其他复杂多流模型，这有待研究者对动态图像模型进一步改进。本文提出的多模型均值投票法的平均识别率为84.67%，比基础动态图像网络模型[7]的平均识别率提高了约14%，表明本文方法较好地改善了动态图像模型信息丢失的问题。

4 结束语

本文在动态图像模型的基础上，提出对生成的近似动态图像进行水平翻转操作，使得训练图像的数据量翻倍，同时针对全连接层参数量过多而导致的网络模型过拟合问题，提出在卷积神经网络模型的卷积层和全连接层之间添加水平翻转结构，使得卷积层提取到具有动作区分性的特征信息增加1倍，实验结果证明该方法有效缓解了网络模型过拟合现象。针对网络模型中特征信息传输的损失问题，本文提出了跨层融合模型，将模型第2层的输出特征与模型第5层的输出特征进行跨层融合。在此基础上，本文通过将训练数据集划成不同的训练数据子集，并分别按正序和反序方法生成近似动态图像，从而产生6种不同的训练方法，对提出的两种模型分别进行训练，学习得到12个分类器。测试阶段采用均值投票集成的方法，集成12个分类器的识别结果，以期得到更稳定、更好的平均性能。实验结果表明，该方法不仅可整体提升每个动作类的识别效果，还可增强模型系统的鲁棒性。本文提出的多模型投票均值的识别方法依赖于近似动态图像和原始动作识别模型提取特征信息的能力，所以下一步的研究工作是研究更加高效的视频预处理方法以及更优的基础模型。

参 考 文 献

[1] BLACKBURN J and RIBEIRO E. Human Motion Recognition Using Isomap and Dynamic Time Warping[M]. Berlin Heidelberg: Springer, 2007: 285–298. doi: 10.1007/978-3-540-75703-0_20.

[2] QU Hang and CHENG Jian. Human action recognition based on adaptive distance generalization of isometric mapping[C]. Proceedings of the International Congress on Image and Signal Processing, Bangalore, India, 2013: 95–98. doi: 10.1109/cisp.2012.6469785.

[3] WANG Heng, KLÄSER A, SCHMID C, *et al*. Dense trajectories and motion boundary descriptors for action recognition[J]. *International Journal of Computer Vision*, 2013, 103(1): 60–79. doi: 10.1007/s11263-012-0594-8.

[4] WANG Heng and SCHMID C. Action recognition with improved trajectories[C]. Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 3551–3558. doi: 10.1109/iccv.2013.441.

[5] OHNISHI K, HIDAKA M, and HARADA T. Improved dense trajectory with cross streams[C]. ACM on Multimedia Conference, Amsterdam, Holland, 2016: 257–261. doi: 10.1145/2964284.2967222.

[6] AHAD M A R, TAN J, KIM H, *et al*. Action recognition by employing combined directional motion history and energy images[C]. IEEE Conference On Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 73–78. doi: 10.1109/CVPRW.2010.5543160.

[7] BILEN H, FERNANDO B, GAVVES E, *et al*. Dynamic image networks for action recognition[C]. Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 3034–3042. doi: 10.1109/cvpr.2016.331.

[8] CHERIAN A, FERNANDO B, HARANDI M, *et al*. Generalized rank pooling for activity recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 3222–3231. doi: 10.1109/cvpr.2017.172.

[9] SIMONYAN K and ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. Proceedings of the International Conference on Neural

Information Processing Systems, Sarawak, Malaysia, 2014: 568–576. doi: 10.1109/iccvw.2017.368.

[10] LIU Hong, TU Juanhui, and LIU Mengyuan. Two-stream 3D convolutional neural network for skeleton-based action recognition[OL]. https://arxiv.org/abs/1705.08106, 2017.

[11] MOLCHANOV P, GUPTA S, KIM K, *et al*. Hand gesture recognition with 3D convolutional neural networks[C]. Proceedings of the Computer Vision and Pattern Recognition Workshops, Boston, USA, 2015: 1–7. doi: 10.1109/cvprw.2015.7301342.

[12] ZHU Yi, LAN Zhenzhong, NEWSAM S, *et al*. Hidden two-stream convolutional networks for action recognition[OL]. https://arxiv.org/abs/1704.00389, 2017.

[13] WEI Xiao, SONG Li, XIE Rong, *et al*. Two-stream recurrent convolutional neural networks for video saliency estimation[C]. Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Cagliari, Italy, 2017: 1–5. doi: 10.1109/bmsb.2017.7986223.

[14] SHI Yemin, TIAN Yonghong, WANG Yaowei, *et al*. Sequential deep trajectory descriptor for action recognition with three-stream CNN[J]. *IEEE Transactions on Multimedia*, 2017, 19(7): 1510–1520. doi: 10.1109/TMM.2017.2666540.

[15] SONG Sibo, CHANDRASEKHAR V, MANDAL B, *et al*. Multimodal multi-stream deep learning for egocentric activity recognition[C]. Proceedings of the Computer Vision and Pattern Recognition Workshops, Las Vegas, USA, 2016: 24–31. doi: 10.1109/cvprw.2016.54.

[16] NISHIDA N and NAKAYAMA H. Multimodal Gesture Recognition Using Multi-Stream Recurrent Neural Network[M]. New York, Springer-Verlag, Inc., 2015: 682–694.

[17] 朱丽, 吴雨川, 胡峰, 等. 老年人动作识别系统研究[J]. 计算机工程与应用, 2017, 53(14): 24–31. doi: 10.3778/j.issn.1002-8331.1703-0470.
ZHU Li, WU Yuchuan, HU Feng, *et al*. Study on action recognition system for the aged[J]. *Computer engineering and Application*, 2017, 53(14): 24–31. doi: 10.3778/j.issn.1002-8331.1703-0470.

[18] 寿质彬. 基于神经网络模型融合的图像识别研究[D]. [硕士论文], 华南理工大学, 2015.
SHOU Zhibin. Research on image recognition base on neural networks and model Combination[D]. [Master dissertation], South China University of Technology, 2015.

[19] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[20] DIETTERICH T G. Ensemble methods in machine learning[J]. *1st International Workshgp on Multiple Classifier Systems*, 2000, 1857(1): 1–15. doi: 10.1007/3-540-45014-9_1.

[21] FERNANDO B, GAVVES E, ORAMAS M J, *et al*. Modeling video evolution for action recognition[C]. Proceedings of the Computer Vision and Pattern Recognition, Boston, USA, 2015: 5378–5387. doi: 10.1109/cvpr.2015.7299176.

[22] SOOMRO K, ZAMIR A R, and SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[OL]. https://arxiv.org/abs/1212.0402, 2012.

[23] TRAN A and CHEONG L F. Two-stream flow-guided convolutional attention networks for action recognition[C]. Proceedings of the IEEE International Conference on Computer Vision Workshop, Venice, Italy, 2017: 3110–3119. doi: 10.1109/iccvw.2017.368.

[24] SRIVASTAVA N, MANSIMOV E, and SALAKHUTDINOV R. Unsupervised learning of video representations using LSTMs[C]. International Conference on Machine Learning, Lille, France, 2015: 843–852.

罗会兰：女，1974年生，博士，教授，研究方向为机器学习和模式识别等.

卢　飞：男，1994年生，硕士生，研究方向为视频中的动作识别、图像语义分割等.

严　源：男，1991年生，硕士生，研究方向为视频中的动作识别等.