融合空间-时间双网络流和视觉注意的人体行为识别

刘天亮*① 谯庆伟① 万俊伟① 戴修斌① 罗杰波②

①(南京邮电大学江苏省图像处理与图像通信重点实验室 南京 210003)

②(罗彻斯特大学计算机科学系 纽约州 罗彻斯特市 14627)

摘 要：该文受人脑视觉感知机理启发，在深度学习框架下提出融合时空双网络流和视觉注意的行为识别方法。首先，采用由粗到细Lucas-Kanade估计法逐帧提取视频中人体运动的光流特征。然后，利用预训练模型微调的GoogLeNet神经网络分别逐层卷积并聚合给定时间窗口视频中外观图像和相应光流特征。接着，利用长短时记忆多层递归网络交叉感知即得含高层显著结构的时空流语义特征序列；解码时间窗口内互相依赖的隐状态；输出空间流视觉特征描述和视频窗口中每帧标签概率分布。其次，利用相对熵计算时间维每帧注意力置信度，并融合空间网络流感知序列标签概率分布。最后，利用softmax分类视频中行为类别。实验结果表明，与其他现有方法相比，该文行为识别方法在分类准确度上具有显著优势。

关键词：人体行为识别；光流；双重时空网络流；视觉注意力；卷积神经网络；长短时记忆神经网络

中图分类号：TP391.41 **文献标识码**：A **文章编号**：1009-5896(2018)10-2395-07

DOI: 10.11999/JEIT171116

Human Action Recognition via Spatio-temporal Dual Network Flow and Visual Attention Fusion

LIU Tianliang① QIAO Qingwei① WAN Junwei① DAI Xiubin① LUO Jiebo②

①(*Jiangsu Provincial Key Laboratory of Image Processing and Image Communication*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)

②(*Department of Computer Science*, *University of Rochester*, *Rochester*, *NY* 14627, *USA*)

Abstract: Inspired by the mechanism of human brain visual perception, an action recognition approach integrating dual spatio-temporal network flow and visual attention is proposed in a deep learning framework. First, the optical flow features with body motion are extracted frame-by-frame from video with coarse-to-fine Lucas-Kanade flow estimation. Then, the GoogLeNet neural network with fine-tuned pre-trained model is applied to convoluting layer-by-layer and aggregate respectively appearance images and the related optical flow features in the selected time window. Next, the multi-layered Long Short-Term Memory (LSTM) neural networks are exploited to cross-recursively perceive the spatio-temporal semantic feature sequences with high level and significant structure. Meanwhile, the inter-dependent implicit states are decoded in the given time window, and the attention salient feature sequence is obtained from temporal stream with the visual feature descriptor in spatial stream and the label probability of each frame. Then, the temporal attention confidence for each frame with respect to human actions is calculated with the relative entropy measure and fused with the probability distributions with respect to the action categories from the given spatial perception network stream in the video sequence. Finally, the softmax classifier is exploited to identify the category of human action in the given video sequence. Experimental results show that this presented approach has significant advantages in classification accuracy compared with other methods.

收稿日期：2017-11-27；改回日期：2018-07-26；网络出版：2018-08-02

*通信作者： 刘天亮 liutl@njupt.edu.cn

基金项目：国家自然科学基金(61001152, 31200747, 61071091, 61071166, 61172118)，江苏省自然科学基金(BK2012437)，南京邮电大学校级科研基金(NY214037)，国家留学基金

Foundation Items: The National Natural Science Foundation of China (61001152, 31200747, 61071091, 61071166, 61172118), The Natural Science Foundation of Jiangsu Provice of China (BK2012437), The Natural Science Foundation of NJUPT (NY214037), China Scholarship Council

Key words: Human action recognition; Optical flow; Spatio-temporal dual network flow; Visual attention; Convolution Neural Network (CNN); Long Short-Term Memory (LSTM)

1 引言

人体行为识别是图像处理与视觉分析领域的主要任务和研究方向之一，其目标是从图像或视频中提取人体行为动作特征并分析识别动作类别，具有非常重要的现实意义，广泛应用于虚拟现实、智能视频监控、运动员动作分析和医疗辅助等方面。随着视觉计算和机器学习技术进步，尤其是深度学习飞速发展，为利用图像、视频等传感器设备识别人体行为动作提供更为便利和可能。行为动作识别可分为两类：一类依赖较优异的人工设计传统视觉特征，如密集轨迹特征[1]、多种特征融合的多实例学习[2]、视觉增强单词包法[3]等；后者是借助深度神经网络稳健感知视觉特征，如利用软注意力和长短时记忆LSTM(Long Short-Term Memory)学习法[4]。前者人工设计特征通常是一件非常费时费力的事，不但需要相关领域的专业知识，而且在很大程度上也要依靠经验和运气才能选好。后者需依赖大量标记数据学习但在动作表征能力上更强、更灵活且识别性能较高。深度神经网络从大量数据中监督式学习到表现力更强的高层特征，而这种学习方式符合人类感知世界的机理[5]。

训练样本足够多，基于深度网络和视觉注意机制往往能感知抽象到高层语义特征，且更适合目标和行为的识别。通常，人类认知过程并非将注意平均分散至整个场景，而有意将目光聚集在不同感兴趣位置上准确获取目标信息[6]。于是，文献[7]引入基于视觉注意模型自动学习描述图像内容；可视化展示模型自动学习并修正对显著对象的注视，输出序列生成相应的单词。文献[8]提出允许模型自动或软搜索目标语句并预测目标词语之间相关关系，而不必硬性分割这些词语。文献[9]提出利用递归神经网络模型自适应地选择区域或位置的序列，并仅以图像或视频中高分辨率处理所选择的区域提取信息。针对图像分类和字幕生成的加速训练问题，文献[10]提出开关递归注意力模型，利用训练随机注意网络改进后验推论，以减少随机梯度变异。为了感知场景运动，文献[11]给出一种由粗到细策略的运动光流Lucas-Kanade特征提取方法。为了学习深度网络，文献[12]给出大规模分层图像数据库ImageNet。针对强化基本特征提取模块，文献[13]提出Inception概念下的GoogLeNet网络。文献[14]给出可视化并理解基于长短时记忆LSTM单元的递归神经网络建模序列。为了解决高效图像相似问题，文献[15]给出一种基于高斯分布之间KL散度逼近的测度方法。为了有效避免神经网络过拟合，文献[16]给出一种Dropout方法。针对大规模神经网络优化，文献[17]给出一种有效的随机优化方法Adam。

综上问题，本文提出一种融合空间-时间双网络流感知和视觉注意模型的行为识别法。

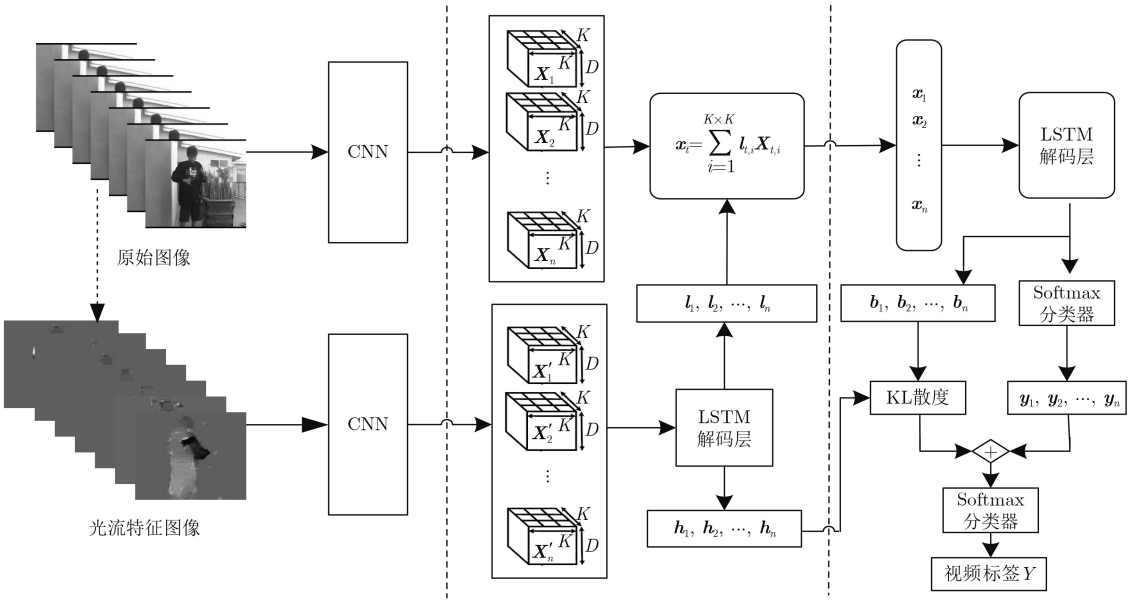
2 人体行为动作识别整体框架

本文提取每一视频帧的显著视觉特征表示人体行为动作，并利用空间-时间双网络流和视觉注意力机制解决行为动作识别，如图1所示。首先，采用GoogLeNet卷积神经网络CNN提取n幅连续原始图像及其相应光流特征图的中层视觉特征立方体描述$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$和$\boldsymbol{X}'_1,\boldsymbol{X}'_2,\cdots,\boldsymbol{X}'_n$。然后，根据空间维LSTM网络解码相应光流的$\boldsymbol{X}'_1,\boldsymbol{X}'_2,\cdots,\boldsymbol{X}'_n$视觉特征立方体，得到空间维注意力强度向量$(\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_n)$并在$K\times K$空间感知邻域上点乘原始图像序列的$\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$视觉特征立方体，以输出原始序列时间维的相应特征向量$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$。$\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$表示空间维LSTM解码输出的隐状态作为空间网络流高层语义描述序列。其次，将输出的特征向量$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$输入时间维LSTM网络，编码时间窗口内隐状态之间相互依赖，输出作为时间网络流高层语义描述的时间维LSTM解码的隐状态序列$\boldsymbol{B}=(\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_n)$和视频窗口中每帧标签概率分布$(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)$。再次，根据得到的$\boldsymbol{B}$和$\boldsymbol{H}$序列，采用KL散度测定相对熵计算时间维每帧注意力关注置信度，并与空间网络流感知序列标签概率分布矩阵中相应帧的列向量数乘，得到每帧相对视频序列帧动作类别的缩放概率分布，以约束原始图像流中关键帧。最后，利用softmax最大化似然函数分类识别视频中人体行为动作类别。

3 空-时双网络流高层特征感知

3.1 Lucas-Kanade运动光流提取

光流不仅刻画目标运动且蕴含丰富3维结构[11]，采用光流特征确定视觉注意的运动选择标准和反映其它丰富视觉信息。设$\boldsymbol{I}_0(\boldsymbol{x})$运动到$\boldsymbol{I}_1(\boldsymbol{x}+\boldsymbol{u}(\boldsymbol{x}))$期间亮度不变。采用类似文献[11]的由粗到细多尺度光流估计法和Munsell颜色转换逐帧提取视频中人体运动的光流图像：



(a) 视觉特征提取　　(b) 空间维注意力选择　　(c) 时间维注意力选择

图 1 本文行为识别流程图

$$\min_{\boldsymbol{u}}\left\{\int_{\Omega}\left(|\nabla u_1|^2+|\nabla u_2|^2\right)\mathrm{d}\Omega +\eta\int_{\Omega}\left(\boldsymbol{I}_1\left(\boldsymbol{x}+\boldsymbol{u}(\boldsymbol{x})\right)-\boldsymbol{I}_0(\boldsymbol{x})\right)^2\mathrm{d}\Omega\right\} \tag{1}$$

其中，$\boldsymbol{I}_0$和$\boldsymbol{I}_1$是视频中前后两帧的图像对，$\boldsymbol{u}(u_1(\boldsymbol{x}), u_2(\boldsymbol{x}))$是2维位移场，$\nabla u_1$和$\nabla u_2$分别为水平和垂直方向上邻域位移变化，$\eta$是平衡参数。第1项正则化惩罚$\boldsymbol{u}$中邻域位移变化的突变，以获得平滑的位移场；第2项称为光流约束的数据项。式(1)的目标是找到使基于邻近帧图像之间的误差标准与正则化惩罚均最小化的位移场视差图$\boldsymbol{u}$。

3.2 GoogLeNet空间维视觉特征

鉴于层数较深的神经网络能学习到较优越的语义特征，但是层数过深易导致欠拟合以及计算资源浪费，受GoogLeNet[13]启发，本文采用基于Inception 结构感知空间视觉特征，以既保持网络结构的稀疏性，又能利用密集矩阵的高性能计算。利用ImageNet数据集[12]预训练并微调得到的GoogLeNet深度卷积神经网络模型，分别逐层卷积给定时间窗口视频中外观图像和相应光流特征作为中间层特征，并自动聚合蕴含边、角和线等底层特征以生成有显著结构的时间流和空间流高层语义特征。类似文献[4]，每个时刻t，最后的卷积层具有D个卷积图，大小形状为$K\times K\times D$(本文实验设为7×7×1024)的特征立方体：

$$\boldsymbol{X}_t=\left[\boldsymbol{X}_{t,1},\boldsymbol{X}_{t,2},\cdots,\boldsymbol{X}_{t,K^2}\right],\quad \boldsymbol{X}_{t,i}\in\boldsymbol{R}_D \tag{2}$$

提取K^2个D维向量，称特征立方体中的特征切片。K^2大小垂直特征切片每个映射到输入空间中的不同重叠区域，根据空间维视觉注意在K^2垂直特征上选择注意力的集中区域。

4 融合时-空双网络流和视觉注意机制的人体行为识别

4.1 长短时记忆LSTM编码行为

鉴于增加网络层数的递归神经网络通常易导致后面层节点对前面层时间节点的感知能力下降的问题，本文采用长短时记忆LSTM网络[14]建模行为序列。采用存储器单元存储、修改和访问内部状态，能更好地发现较长时间之间的依赖关系：

$$\left.\begin{aligned}
\boldsymbol{i}_t&=\sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+b_i\right)\\
\boldsymbol{f}_t&=\sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+b_f\right)\\
\boldsymbol{o}_t&=\sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+b_o\right)\\
\boldsymbol{g}_t&=\sigma\left(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+b_c\right)\\
\boldsymbol{c}_t&=\boldsymbol{f}_t\otimes\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\otimes\boldsymbol{g}_t\\
\boldsymbol{h}_t&=\boldsymbol{o}_t\otimes\tanh(\boldsymbol{c}_t)
\end{aligned}\right\} \tag{3}$$

式中，$\sigma(\cdot)$和$\tanh(\cdot)$分别是sigmoid激活函数和双曲正切函数。$\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{c}_t$, $\boldsymbol{o}_t$和$\boldsymbol{h}_t$分别是LSTM单元的输入门、遗忘门、存储单元、输出门以及隐状态。此处$\boldsymbol{x}_t$(参见式(5))表示时间步t处的LSTM网络输入，符号$\otimes$表示逐元素依次相乘。为了进一步学习长时间动态变化关系，本文采用多层LSTM网络建模隐状态，如图2所示，其中R为原始视频高层语义特征，F为光流特征高层语义特征，立方体代表LSTM, O代表网络输出；每一LSTM层中输出隐状

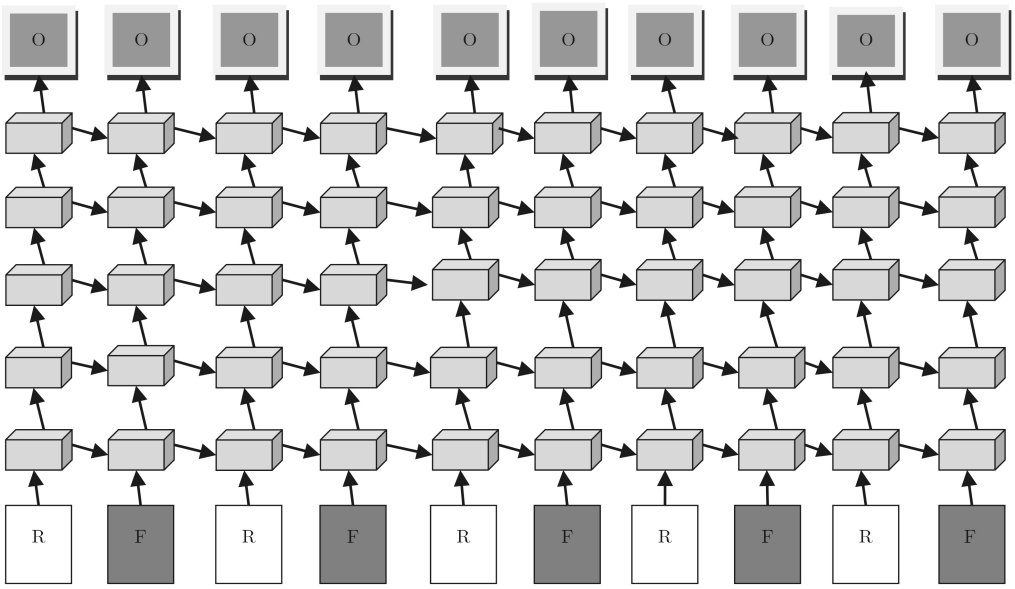


图 2 空间-时间双网络流多层LSTM网络示意图

态作为下一LSTM层的输入，将LSTM层层叠加构成多层LSTM。

4.2 空间维确定性软注意力机制

本文采用视觉注意力模型建模学习视频中空间关系上重点关注区域的概率分布。每个时刻t，光流特征感知网络流预测$K\times K$空间感知邻域位置上$\boldsymbol{l}_{t+1}$在标记类上的softmax概率分布(参见图1)，公式定义为

$$\begin{aligned}\boldsymbol{l}_{t+1,i}&=p(L_{t+1}=i|\boldsymbol{h}_{t,i})\\&=\frac{\exp\left(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{h}_{t,i}\right)}{\sum\limits_{j=1}^{K\times K}\exp\left(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{h}_{t,j}\right)},\quad t\in 1,2,\cdots,n\end{aligned}\tag{4}$$

式中，$\boldsymbol{W}_i$是映射到位置softmax的第i个元素的权重向量，L_{t+1}是区间[1, K^2]中随机变量。

受眼球运动机制[4]和光流特征蕴含3维空间和运动信息启发，利用注意力集中度机制选择原始视频和光流特征序列的中层卷积立方体特征。设$\boldsymbol{x}_t$为空间网络流LSTM解码层输出(即时间网络流LSTM输入$\boldsymbol{x}_t$)不是对特征立方体的所有切片取平均值，而是采样位置的切片。鉴于不可微分的硬注意力模型需采取式(4)抽样，本文采用确定性软注意力机制[8]计算关于前一时刻$\boldsymbol{h}_{t-1}$在不同感知区域的特征切片期望，作为下一刻t的时间网络流LSTM的输入$\boldsymbol{x}_t$:

$$\boldsymbol{x}_t=E_{P\left(L_t|\boldsymbol{h}_{t-1}\right)}\left[\boldsymbol{X}_t\right]=\sum_{i=1}^{K\times K}\boldsymbol{l}_{t,i}\boldsymbol{X}_{t,i}\tag{5}$$

式中，$\boldsymbol{X}_t$是原始视频帧特征立方体，$\boldsymbol{X}_{t,i}$是特征立方体$\boldsymbol{X}_t$在时间步t上的第i个切片，$\boldsymbol{l}_{t,i}$表示4.1节输出得到的光流特征学习到的空间位置softmax参数，K^2为立方体切片平面大小。LSTM模型的初始状态$\boldsymbol{c}_0$和隐含状态$\boldsymbol{h}_0$，初始化$\boldsymbol{x}_1$的第1个空间注意力初始参数$\boldsymbol{l}_1$:

$$\begin{aligned}\boldsymbol{c}_0&=\boldsymbol{f}_{\mathrm{init},c}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{K^2}\sum_{i=1}^{K\times K}\boldsymbol{X}_{t,i}\right)\right),\\\boldsymbol{h}_0&=\boldsymbol{f}_{\mathrm{init},h}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{K^2}\sum_{i=1}^{K\times K}\boldsymbol{X}_{t,i}\right)\right)\end{aligned}\tag{6}$$

式中，$\boldsymbol{f}_{\mathrm{init},c}$和$\boldsymbol{f}_{\mathrm{init},h}$是两个多层感知器，$n$是LSTM模型中的时间步数，即视频段帧数。

根据原始视频输入高层特征序列$\boldsymbol{x}_1$及其对应标签，利用LSTM递归神经网络解码为对应标签的概率分布，对原始视频序列的LSTM解码层设计与光流特征序列的设计是一致的。保存原始视频每帧图像LSTM解码后最后一层每个单元参数的隐状态序列$\boldsymbol{B}$。

4.3 时间维相关性注意力机制

鉴于视频常含易混淆帧并导致分类效果欠佳，本文提出时间维注意机制判读每帧相对视频段的相关性，利用光流特征序列基于空间LSTM模型解码到的隐状态参数$\boldsymbol{h}_t$，结合原始图像序列基于时间维LSTM模型解码到的隐状态参数$\boldsymbol{b}_t$，计算时间维注意力关注度权重值[15]:

$$\begin{aligned}\mathrm{KL}'_t=&\frac{1}{2}\sum_{k=1}^{q}\left(\boldsymbol{h}_{t,k}\lg\frac{\boldsymbol{h}_{t,k}}{\boldsymbol{b}_{t,k}}+\boldsymbol{b}_{t,k}\lg\frac{\boldsymbol{b}_{t,k}}{\boldsymbol{h}_{t,k}}\right),\\&t\in 1,2,\cdots,n\end{aligned}\tag{7}$$

式中，t表示时间帧，n为视频总长度；$\boldsymbol{h}$为隐状态层参数索引，q为其最大值；$\boldsymbol{b}_{k,t}$和$\boldsymbol{h}_{k,t}$分别表示原

始视频和光流序列的隐状态参数向量。鉴于时间维注意关注度权重系数可趋于正无穷而没有明确上界，本文采用sigmoid函数限制其幅值在[0, 1]区间更新：

$$\mathrm{KL}_t = \frac{\mathrm{sigmoid}\left(-\left|\mathrm{KL}_t'\right|\right)}{\sum_{t\subseteq T}\mathrm{sigmoid}\left(-\left|\mathrm{KL}_t'\right|\right)} \tag{8}$$

时间流LSTM网络解码后的标签概率分布$P(y_t=c)$，与每一帧对应得分系数内积后，利用softmax分类器分类判别得视频窗口对应类别概率分布：

$$P(y'=c)=\frac{\exp\left(\sum_{t=1}^{n}P(y_t=c)\mathrm{KL}_t\right)}{\sum_{c\in C}\exp\left(\sum_{t=1}^{n}P(y_t=c)\mathrm{KL}_t\right)},\quad t\in 1,2,\cdots,n \tag{9}$$

式中，t代表时间帧，c表示动作类别，$P(y'=c)$为最大概率值对应标签为人体行为动作类别。

将视频分成包含固定帧数n的若干小剪辑，每个剪辑根据时空维注意机制判断对应帧在时间和空间上对剪辑片段的重要程度，得若干剪辑判读序列标签值，利用众数原理选择最大可能标签作为序列最终类别。

4.4 注意力惩罚和损失函数优化

受文献[4]和文献[7]启发，本文采用正则化交叉熵损失和时空视觉注意力网络惩罚的优化目标函数，在空间位置和时间关注softmax处施加额外附加约束，使得$\sum_{t=1}^{n}\boldsymbol{l}_{t,i}\approx 1$，以使模型在某个时刻点关注帧内的每个相应空间区域：

$$\begin{aligned}L=&-\sum_{t=1}^{n}\sum_{c=1}^{C}y_{t,c}\lg\hat{y}_{t,c}+\lambda_1\sum_{t=1}^{n}\|KL_t\|_1\\&+\lambda_2\sum_{i=1}^{K\times K}\left\|1-\sum_{t=1}^{n}\boldsymbol{l}_{t,i}\right\|_2+\lambda_3\sum_{i}\sum_{j}\|\boldsymbol{\theta}_{i,j}\|_2\end{aligned} \tag{10}$$

其中，$\boldsymbol{y}_t=(y_{t,1},\ldots,y_{t,C})^{\mathrm{T}}$表示时刻$t$相应类别标签出现概率的向量，$n$和$C$分别为时间步和预定类别的总数。如果属于第$i$类，则对于$j$不等于$i$，则$y_{t,c}=1$；否则，$y_{t,c}=0$。$\hat{y}_{t,c}$表示$t$时刻被预测为第$c$类的概率，$\boldsymbol{\theta}_{i,j}$表示设计网络中所有需要学习的参数。$\lambda_1$，$\lambda_2$和$\lambda_3$分别平衡时间注意力惩罚、空间注意力惩罚以及权重衰减的正则贡献。L1范数为正则控制时间维注意力强度，而非无限增大。第1个L2范数正则鼓励空间维注意力动态分布在时序列中更多的空间位置。第2个L2范数正则抑制网络过度配置。鉴于整体优化相互影响的各网络参数有相当困难，本文采用联合训练策略随机梯度下降法有效训练时空注意力网络和相关LSTM网络。单独预训练空间和时间注意模块，以确保网络融合和参数的收敛。

5　实验结果与分析

5.1 实验方案与参数设置

采用UCF-11数据集定量分析验证Theano深度学习框架下实现的本文方法。数据集由11个类1600个视频组成：basketball, biking, diving, g_swing, ho_riding, s_juggling, swing, t_swing, t_jumping, v_spiking, walking。采用975个剪辑用于训练和625个测试，剪辑的帧率为29.97 fps，每个视频剪辑只关联一个动作，以30 fps帧率分割保存为224×224像素图片集，模型的训练与测试是在NVIDIA TITAN BLACK GPU上，每128批处理一次。LSTM解码器训练5层LSTM模型，LSTM隐藏状态的维数、门的状态、隐藏层大小均设置为512。注意力惩罚系数$\lambda_1=\lambda_3=1$，权重衰减处罚$\lambda_2=10^{-5}$，dropout[16]值设为0.5。采用GoogLeNet模型提取的最后卷积层7×7×1024维数据作为中间层输入特征。

5.2 实验结果与分析

(1) 空间维与时间维视觉注意机制的有效性验证：表1给出了本文方法在不同子模块组合下平均准确度和模型测试耗时情况的综合性能比较。由表1可知，仅利用光流特征和传统深度学习CNN+LSTM的“2+3”子模块组合方法的平均准确度最低(仅为72.9%)，其原因是所提取光流特征保存为图像的过程除了保留运动以及空间信息外，丢失了很多视频中固有的纹理和色彩等其他有效的视觉判别特征。可见，蕴含运动信息的光流特征只能作为有效的补充手段或者判别依据。相较于不含时间维注意模型的“1+2+3+4”方法和“1+2+3”方法，“1+2+3+5”子模块组合方法的平均准确度更高；由此可知，时间维度注意力选择机制的贡献大于空间维注意力选择机制，引入时间维关键帧选择约束能减少因易混淆动作图像影响视频分类。而本文方法(“1+2+3+4+5”)相较于其他各子模块组合方法

表 1　本文方法在不同子模型组合下平均准确度和模型测试用时性能比较

子模型组合	1+3	2+3	1+2+3	1+2+3+4	1+2+3+5	1+2+3+4+5
平均准确度(%)	76.1	72.9	82.0	83.2	85.5	**87.5**
测试时间(s)	129	130	130	132	132	133

注：数字含义：1为原始图像，2为光流特征，3为CNN+LSTM模型，4为空间注意力，5为时间注意力

的平均准确度更高。由此可见，本文方法有效结合空间维和时间维注意力模型，既可选择视频帧中图片内重点关注的有关目标区域，亦可兼顾选择视频序列中与动作类别紧密关联且重点关注的相应视频帧。其中模型测试用时是将测试集625个视频预处理后于相同环境下在5个不同模型上测试所消耗的时间。

(2) 纵向实验比较和混淆矩阵：表2给出了UCF-11数据集上本文方法与传统方法在平均准确度角度的客观评价对比实验结果。由表2得知，与传统提取整体特征方法的MIL方法[1]和Dense trajectories方法[2]相比，本文方法的平均准确度分别高出了12.3%和3.3%。本文方法明显优越有如下两方面的原因。第一，相对于传统方法未借助其他先验知识提取视觉特征，本文方法利用GoogLeNet深度卷积感知网络从空间视野角度根据大量数据预训练自动学习的有用特征表示，从底层视觉特征逐步卷积聚合到高层视觉语义特征表示，提取鲁棒性和识别度高的特征；其二，本文方法也利用了类似于视点关注度移动的空间维和时间维注意力选择机制，以关注行为动作图像中最具判别性区域和有选择地摒弃无关动作特征，从而提高动作分类的准确度。

表 2 本文方法与其他传统方法在UCF-11数据集下实验结果比较

模型方法	数据模式	平均准确度(%)
MIL方法[1]	多种特征融合	75.2
Dense trajectories方法[2]	密度轨迹特征	84.2
Attention+LSTM方法[4]	外观图像	84.9
本文方法	外观图像+光流特征	**87.5**

与传统Attention+LSTM方法[4]相比较，本文方法平均准确度提高了2.6%。其原因是，文献[4]方法仅从原始视频基于注意力机制和LSTM建模，而本文方法不仅感知原始视频且感知学习其视频内部3维运动信息参数的相应光流特征，同时融合空间-时间网络流拟合注意力选择机制视点的移动；而文献[4]仅仅考虑了空间维注意力选择机制，未充分考虑时间维注意力选择机制的重要性，在最终分类过程中未排除易混淆图像信息，导致分类效果相对欠佳。

为了更好且细粒度地展示本文方法识别性能，图3给出了数据集UCF-11的11类行为动作混淆矩阵。11类人体行为动作中全部正确分类6类，部分错误分类5类。由图3可知，“basketball”，“t_swing”，“v_spiking”，“s_juggling”4种行为最容易判别错误，其主要原因是这4种行为动作均含有类似的手部挥动、复杂的室内室外场景、图像画质模糊等不利因素影响。另外，这4种行为的光流图像3维运动信息极为相似，容易导致注意力选择机制关注不明显；动作本身具有较高程度的相似性，因此存在较高的错误分类现象。

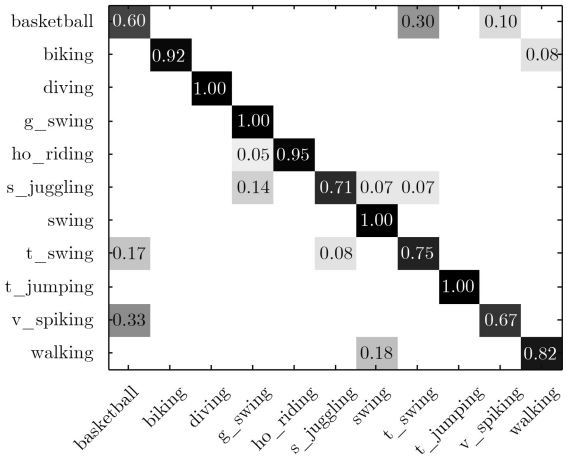


图 3 UCF-11数据集上11类行为动作分类识别实验的混淆矩阵

6 结论

针对利用深度学习识别人体行为通常仅考虑神经网络的设计和规模且未充分考虑类似眼球感知运动的注意力选择机制等问题，本文在结合光流特征捕获场景运动信息基础上，提出在空间维和时间维上利用注意力选择机制约束筛选视频序列中重点关注区域的对象，摒弃视频序列中包含无关信息帧或重点计算视频序列中动作类相关对象，减少运算参数，提高模型鲁棒性。实验结果表明，与其他传统方法相比，本文基于融合双网络流感知和视觉关注度模型框架是一种有效人体行为识别方法，对人体行为识别问题提供一种有价值的解决方案。

参 考 文 献

[1] IKIZLER-CINBIS N and SCLAROFF S, Object, scene and actions: Combining multiple features for human action recognition[C]. European Conference on Computer Vision, Heraklion, Crete, Greece, 2010, 6311: 494–507.

[2] WANG Heng, KLASER A, and SCHMID C. Action recognition by dense trajectories[C]. IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2011: 3169–3176. doi: 10.1109/CVPR.2011.5995407.

[3] 张良, 鲁梦梦, 姜华. 局部分布信息增强的视觉单词描述与动作识别[J]. 电子与信息学报, 2016, 38(3): 549–556. doi: 10.11999/JEIT150410.

ZHANG Liang, LU Mengmeng, and JIANG Hua. An improved scheme of visual words description and action recognition using local enhanced distribution information[J].

Journal of Electronics & Information Technology, 2016, 38(3): 549–556. doi: 10.11999/JEIT150410.

[4] SHARMA S, KIROS R and SALAKHUTDINOV R. Action recognition using visual attention[C]. International Conference on Neural Information Processing Systems Times Series Workshop, Montreal, Canada, 2015: 1–11.

[5] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. *Neural Networks*, 2015, 61: 85–1117. doi: 10.1016/j.neunet.2014.09.003.

[6] RENSINK R A. The dynamic representation of scenes[J]. *Visual Cognition*, 2000, 1(1/3): 17–42.

[7] XU Kelvin, BA Jimmy, KIROS R, *et al*. Show, attend and tell: Neural image caption generation with visual attention[C]. Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, 14: 77–81.

[8] BAHDANAU D, CHO K, and BENGIO Y. Neural machine translation by jointly learning to align and translate[C]. International Conference on Learning Representation, San Diego, USA, 2015: 1–15.

[9] MNIH V, HEESS N, GRAVES A, *et al*. Recurrent models of visual attention[C]. Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2204–2212.

[10] BA Jimmy Lei, GROSSE R, SALAKHUTDINOV R, *et al*. Learning wake-sleep recurrent attention models[C]. International Conference on Neural Information Processing Systems, Montreal, Canada, 2015: 2593–2601.

[11] AND J S P. Horn-Schunck optical flow with a multi-scale strategy[J]. *Image Processing on Line*, 2013, 20: 151–172. doi: 10.5201/ipol.2013.20.

[12] RUSSAKOVSKY O, DENG Jia, SU Hao, *et al*. ImageNet: Large scale visual recognition challenge[J]. *International Journal of Computer Vision*, 2015, 115(3): 211–252.

[13] SZEGEDY Christian, LIU Wei, JIA Yangqing, *et al*. Going deeper with convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9. doi: 10.1109/CVPR.2015.7298594.

[14] ANDREJ K, JUSTIN J, and LI Feifei. Visualizing and understanding recurrent networks[C]. International Conference on Learning Representation Workshop, Caribe Hilton, USA, 2016: 1–11.

[15] GOLDBERGER J, GORDON S, and GREENSPAN H. An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures[C]. IEEE International Conference on Computer Vision, Nice, France, 2003: 487–493.

[16] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, *et al*. Dropout: A simple way to prevent neural networks from overfitting[J]. *Journal of Machine Learning Research*, 2014, 15: 1929–1958.

[17] KINGMA D P and BA J. Adam: A method for stochastic optimization[C]. International Conference on Learning Representation, San Diego, USA, 2015: 1–15.

刘天亮：1980年生，男，博士，副教授，硕士生导师，研究方向为图像处理、计算机视觉.

谯庆伟：1989年生，男，硕士生，研究方向为图像处理与多媒体通信.

万俊伟：1993年生，男，硕士生，研究方向为图像处理与多媒体通信.

戴修斌：1980年生，男，博士，副教授，硕士生导师，研究方向为医学图像重建、图像处理和模式识别.

罗杰波：1968年生，博士，教授，博士生导师，研究方向为图像处理、计算机视觉、机器学习、数据挖掘和社交网络媒体等.