ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 406－413
DOI:10. 16337/j. 1004-9037. 2019. 03. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于注意力机制的群组行为识别方法

王传旭　龚玉婷

(青岛科技大学信息科学技术学院,青岛,266100)

摘　要: 在基于视频图像的群组行为识别方法中,传统的深度学习方法大多使用标准(最大/平均)池化操作对卷积特征进行处理,并且未考虑群组行为中的关键人物对群组行为分类的重要性。针对以上问题,本文提出一种基于注意力机制的模型来检测群组行为视频中的行为,重点关注活动中的关键人物,根据注意力权重的不同分配动态地对卷积特征进行池化,最终正确识别视频图像中的群组行为。此模型在群组行为数据集CAD(Collective activity dataset)和CAE(Collective activity extended dataset)上的识别准确率优于许多使用标准池化结构的现有模型。

关键词: 群组行为;图像处理;注意力机制;行为识别

中图分类号: TN391　　**文献标志码:** A

Group Activity Recognition Method Based on Attention Mechanism

Wang Chuanxu, Gong Yuting

(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, 266100, China)

Abstract: In the video image based group activity recognition method, the traditional deep learning methods generally use the conventional(maximum / average)pooling to process the convolutional feature. However, these methods do not consider the importance of the key characters in the group activity which influence the classified result of group behavior. Therefore, we propose an attention based model to detect behavior in group activity videos. In order to identify the group behavior correctly in the video image, this model focuses on the key people in the activity and pools convolutional features dynamically according to the weight of the attention. We conduct extensive experiments on two group behavior datasets, CAD (Collective activity dataset) and CAE (Collective activity extended dataset). The recognition accuracy of our model is better than many existing models using conventional pooling structure.

Key words: group activity; image processing; attention mechanism; behavior recognition

引　　言

人体行为识别发展至今已经成为当今计算机视觉领域研究的一个热点,在智能监控、虚拟现实和视频检索等方面有着广泛的应用前景和巨大的经济价值。简单的行为识别即单人动作的分类[1],对于给定的一段视频,只需将其中的每个人的动作准确地分类到已知的动作类别中,然而这种局限于单

基金项目: 国家自然科学基金(61472196,61672305)资助项目。
收稿日期: 2018-10-26;**修订日期:** 2019-04-16

人活动的行为识别不足以满足真实的社会需要。较复杂的人体行为识别是给定的视频中不止包含一个动作分类,而是具有多个目标(人)多个动作类别,多个目标可能在同时做相同的动作,或者多个目标正在共同完成同一个复杂行为,将这些由多人完成的相同动作或者共同完成的行为称为“群组行为”[2]。

群组行为识别近年来吸引了许多研究者的目光。传统的经典方法有方向梯度直方图(Histogram of oriented gradient, HOG)结合使用支持向量机(Support vector machine, SVM)(HOG+SVM)、尺度不变特征转换(Scale invariant feature tvansform, SIFT)结合词袋模型(Bog of words, BOW)(SIFT+BOW)等方法进行行为识别。Lan等[3]和Ramanathan等[4]分别在全监督和弱监督的框架下,探讨社会角色的概念,以及单人在群组背景下的预测行为,建立了结构化模型来表示单人在空间以及时间区域的信息,属于浅层学习,对特征的刻画能力有限。并且这些模型都是基于手工制作特征的概率或者判别模型来识别群组行为,有很大的局限性,需要人力持续调整模型参数,不断迭代才能达到比较好的效果,开发周期长。针对以上问题,本文模型则使用深度学习网络来提取图像特征,使用时空特征更具代表性,模型更具泛化性,并且在使用深度学习网络的同类方法[5-7]中性能表现更佳。

Chio和Savarese [8]同时跟踪多个人,并在一个联合框架中识别出单人行为、交互行为和群组行为。在文献[9]中,使用随机森林结构从输入视频中提取时空区域的特征,之后用于三维马尔科夫随机场,以定位场景中的群组行为。然而,上述方法他们并没有考虑群组行为中关键人物对群组行为识别的重要性。确定参与活动的关键人物、排除其他不相关人物是极其重要的,同时也是区分单人视频和多人视频中行为识别的关键之处。在视觉认知文献[10]中已经指出,人类不会把注意力集中在视觉范围内的整个场景上,相反,他们依次关注场景的不同部分以提取相关重要信息。而大多数传统的群组行为识别算法不采用注意力机制,对图像或者视频的重要部分无法给予关注。随着近年来深度神经网络的兴起,基于注意力的模型已经被证明在几个具有挑战性的任务上取得了良好的结果,包括图像识别、字幕生成[11]以及机器翻译[12]等。因此,本文提出一种基于注意力机制的群组行为识别模型,动态地对卷积特征进行池化,使用“循环注意力”在活动的不同阶段辨认出关键人物,为场景中的人物分配不同的注意力权重,最终识别其中的单人行为和群组行为类别。实验结果表明,此模型会更倾向于识别视频帧中的重要元素,为使用注意力机制进行群组行为识别提供了更令人信服的结果。

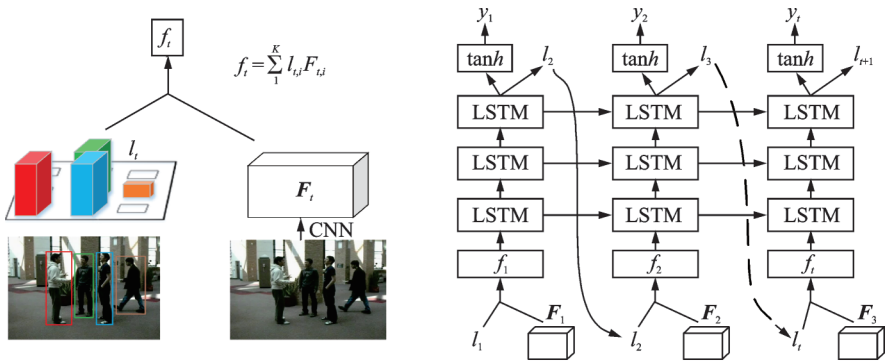
1 基于注意力机制的群组行为识别模型

1.1 模型结构

本文提出基于注意力机制的神经网络处理模型,如图1所示,其中软注意力机制即图1(a)所示,卷积神经网络(Convolution neural network,CNN)将视频帧作为输入并得到一个特征立方体,用符号表示为F_t。然后根据特征立方体F_t与l_t计算得到f_t。其中l_t是图1(b)循环模型中对特征立方体进行计算的输出层函数(Location softmax)的输出。图1(b)中,在每个时间步t,循环网络将图1(a)中生成的f_t作为输入,之后通过3层长短期记忆网络(Long short-term memory, LSTM)来预测群组行为类别标签y_t和下一个Location probability l_{t+1}(即$t+1$时刻Location softmax在特征区域的得分)。LSTM网络的具体细节会在1.3节进一步介绍。

1.2 特征提取

使用在ImageNet数据集上训练的CNN,将视频帧作为输入,得到大小为$K\times N\times D$的特征立方体,用符号表示为F_t,因此在每一时间步t,可以提取K个$N\times D$的特征向量(K代表图像中的人数,$N\times D$是每个人的特征维度),将这些特征向量称为特征立方体的特征片,即有$F_t=[F_{t,1},F_{t,2},\cdots,F_{t,i},\cdots,F_{t,k}]$。



(a) 软注意力机制 (a) Soft attention mechanism　　(b) 循环模型 (b) Recurrent model

图1　基于注意力机制的模型结构图

Fig.1　Model structure based on attention mechanism

其中，$F_{t,i}$表示第t帧第i个人的特征向量片。每个特征片映射到输入图像中，即每个人的图像区域(Location/patch)，模型选择将注意力集中在这K个区域上。如图1(a)所示，该图像经过CNN之后，会得到4个人的特征片组成的特征立方体，最终群组行为类别的判断模型则会选择将注意力集中在这4个人所在区域上。

1.3　LSTM和特征提取

本文中使用的LSTM网络[13]的原理为

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{M} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3}$$

式中：i_t是输入门，f_t是遗忘门，o_t是输出门，g_t的计算如式(1)所示，c_t是细胞单元，h_t是隐藏状态，x_t表示LSTM网络中每个时间步的输入。$M:R^a \to R^b$是一个由可训练参数组成的仿射变换，其中$a=d+D$，$b=4d$，d代表i_t，f_t，o_t，g_t，c_t和h_t的维度。

针对图1(b) Recurrent model，其注意力机制的作用过程描述如下：

首先，经过特征提取得到的K个$N\times D$的特征向量，每个特征向量映射到输入帧中，即K个Location/patch，之后将特征向量送入Location softmax为K个Location打分。Location softmax的定义为

$$l_{t,i} = p(L_t = i|h_{t-1}) = \frac{\exp(\boldsymbol{W}_i^{\mathrm{T}} h_{t-1})}{\sum_{j=1}^{K} \exp(\boldsymbol{W}_j^{\mathrm{T}} h_{t-1})} \quad i \in 1, \cdots, K \tag{4}$$

式中：W_i是映射到Location softmax的第i个人的权重，L_t是取值范围为$(1,\cdots,K)$的随机变量，这里的$l_{t,i}$可以看做是模型认为第i个人的相应映射区域(Location/patch)对输入帧的重要程度(Attention probability)，即Location得分的高低代表着Attention对该位置人物关注的强弱；h_{t-1}为上一刻的隐含状态。

其次，学习到权重之后，对不同的Location进行计算打分(Score)，本文的操作是对特征和Score求期望，即Soft attention mechanism[14]通过对不同区域的特征向量进行求期望来计算下一时间步的输入期望值，即有

$$f_t = E_p(L_t|h_{t-1})[F_t] = \sum_{1}^{K} l_{t,i} F_{t,i} \tag{5}$$

式中:F_t是特征立方体,$F_{t,i}$是第t帧图像特征立方体中的第i个人的特征向量;$l_{t,i}$即式(4)中的Location得分(得分的高低代表着Attention对该位置人物关注的强弱);f_t则由每一个Location,共有K个,对应位置的特征向量$F_{t,i}$和Location score$l_{t,i}$相乘然后求和得到。例如图1(a)中,视频帧的群组行为标签为“Talking”,4个人的特征向量F_t分别与其对应的Location得分l_t相乘求和得到f_t,其中红绿蓝黄4种颜色的长方体分别对应帧中的4个人的Location得分,长方体体积大小表示得分的高低,可以看出图中黄色对应的人物并没有参与到“Talking”中,所以最终模型对其的打分最低,Attention对该位置人物关注的最弱,反映到图中即黄色长方体体积最小。

然后,将期望f_t作为输入送入3层的LSTM,如图1(b),之后经过tan$h$激活函数附加隐含层的输出,作为最终的群组行为标签y_t(即在Label类别上的Softmax得分);同时将不经过tanh激活函数的输出,作为下一时刻的l_{t+1}(即下一时刻在k个Location上的Softmax得分),之后l_{t+1}与$[F_{t+1}]$相乘作为下一时刻LSTM的输入f_{t+1},如此循环传递,形成Recurrent model使用注意力机制选择群组行为中重要人物的过程。

本文中第1个时刻的Cell stateC_0和Hidden state H_0使用以下的初始化策略[15],以加速收敛,有

$$c_0 = f_{\text{init},c}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{K}\sum_{i=1}^{K}F_{t,i}\right)\right) \tag{6}$$

$$h_0 = f_{\text{init},h}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{K}\sum_{i=1}^{K}F_{t,i}\right)\right) \tag{7}$$

式中:$f_{\text{init},c}$和$f_{\text{init},h}$是2个多层感知器,T是模型中的时间步数。这些值用于计算初始输入f_1的第1个Location softmax l_1。在实验中,使用多层的LSTM,如图1(b)所示。

1.4 训　　练

本文模型中,卷积神经网络的主要作用是学习图像中每个人物的特征表示。注意力机制模块利用LSTM的属性,并将其作为注意力机制的基础单位进行注意力的选择,使模型分配不同的注意力权重给图像中不同的人物区域。因此,注意力模块的权重、整个网络的输入数据以及网络的中间输出数据相互作用构成整个模型。训练模型时,需要联合训练卷积神经网络以及注意力机制模块,从而获得最高的识别精确度。

本文使用交叉熵损失以及双重随机惩罚[11]对模型进行训练,有

$$L = -\sum_{t=1}^{T}\sum_{i=1}^{C} y_{t,i}\log\hat{y}_{t,i} + \lambda\sum_{i=1}^{K}\left(1-\sum_{t=1}^{T} l_{t,i}\right)^2 + \gamma\sum_{i}\sum_{j}\theta_{i,j}^2 \tag{8}$$

式中:$y_{t,i}$代表真实的标签;$\hat{y}_{t,i}$代表模型预测的群组行为类的概率;T是时间步总数;C是输出类数;λ是注意力惩罚系数;γ是权重衰减系数;θ代表所有的模型参数。第1项是带有时间维的交叉熵损失函数。第2项是注意力正则化项,着重于Location softmax中的Score的注意力惩罚,对Location softmax施加一个额外的约束,$\sum_{t=1}^{T} l_{t,i} \approx 1$,即在总时间步$T$内,图像中所有Location的Score之和为1,也就是Attention对图片中所有人物区域的关注权重概率之和为1,使得模型在某时间点查看帧中每个人的区域,鼓励模型同等重视图像中的每个人;对于第2项的系数λ,实验抽取了不同取值的λ,分析其对最终识别率的影响,结果表明Attention更能关注图像中的关键人物区域。第3项是模型参数的正则项,防止整个模型过拟合。

2 实验结果与分析

为了验证本文算法的识别精度,选用CAD和CAE作为测试数据集进行实验。首先,简单介绍这两个群组行为数据集;然后阐述训练过程中参数的设置以及分析;最后将本文算法以及其他方法的实验结果进行比较和数据分析。

2.1 数据集

实验中使用的第一个数据集是群体活动数据集(Collective activity dataset,CAD),该数据集是使用低分辨率的手持摄像机获取的44个视频片段,此数据集有5种行为标签:Crossing,Queuing,Walking,Talking和Waiting;8种姿势标签(本文中并未用到);以及5种群组行为标签即每帧活动中k个人共同完成的场景标签:Crossing,Queuing,Walking,Talking和Waiting。每个人都有1个行为标签,每帧图像都有1个群组行为标签(场景标签)。

实验中使用的另一个数据集是群体活动扩展数据集(Collective activity extended dataset ,CAE),该数据集共有6个行为标签,分别是Crossing,Queuing,Dancing,Talking,Waiting和Jogging,以及6种群组行为标签,即每帧活动中k个人共同完成的场景标签:Crossing,Queuing,Dancing,Talking,Waiting和Jogging,同样每个人都有1个行为标签,每帧图像都有一个场景标签。

本文使用数据集中的所有视频帧,随机抽取其中的60%用于训练,20%用于验证,20%用于测试。将视频帧输入到ImageNet数据集上训练的VGG-16中,并对其进行了微调,每张图像得到的$k\times1\times3\,000$输出,被用作模型的输入。

2.2 参数设置与分析

实验中,使用交叉验证训练模型以及其他超参数。对于所有数据集,分别试验了LSTM层数为1,2,3,4,5层时的模型,3层LSTM时识别效果最佳,随着LSTM层数递增没有观察到模型性能的显著改进。其中LSTM网络隐藏层的维度设置为512。对于注意力惩罚系数,用0,1,10进行了实验;模型的权重衰减系数设置为10^{-5},并且在所有非循环连接中使用0.5的Droupout[15],使用Adam优化算法[16]进行15个Epoch训练。

为了分析注意力惩罚系数λ(式(7))对最终识别率的影响,抽取实验中λ取值为0,1,10时对比比较明显的结果列入表1中。$\lambda=0$时,模型倾向于减少过多的注意力,将注意力集中于对识别结果有重要影响的区域;$\lambda=1$时会鼓励模型进一步探索更多的不同的注意位置;在$\lambda=10$时,类似于平均池化的情况,将注意力放在了整个图像场景,而不是把注意力有选择地放在图像的关键区域。

表1 不同模型结构在数据集上的平均识别准确率比较

Tab. 1 Comparison of average recognition accuracy of different model structures on datasets %

Model	CAD	CAE
Softmax regression(Full CNN feature cube)	83.54	83.96
Avg pooled LSTM	83.84	84.06
Max pooled LSTM	82.48	83.67
Soft attention model($\lambda=0$)	89.00	91.01
Soft attention model($\lambda=1$)	87.93	88.49
Soft attention model($\lambda=10$)	82.33	83.07

同时,表1的结果表明,本文提出的注意力模型比使用平均和最大池化结构的LSTM表现更好。究其原因,平均池化是对所有的特征信息做了一个均衡的处理,更多的是对背景信息的保留;最大池化是取特征信息的最大值,保留更多的是纹理信息;而本文用到的注意力机制则综合考虑了图像的所有特征信息,从中选择出对当前群组行为识别任务更重要的人物行为特征信息,并将注意力权重更多的分配给关键人物区域,最终提高了识别结果的平均准确率,表1的实验结果也验证了这一点。

2.3 模型评估

2.3.1 本文模型与其他方法在CAD上的结果比较

表2给出了人体行为识别经典模型HOG+SVM,Bag of words以及近几年的群组行为识别方法与本文模型在群体活动数据集CAD上的实验结果,CAD数据集包括Walking,Crossing,Waiting,Queuing和Talking五类群组行为,根据文献[17],Walking和Crossing的定义不明确,因为这两类行为唯一的区别是人与街道之间的关系,且两类行为更像是一个人的行为而不是群组行为。因此,本文将数据集中Walking和Crossing合并为Moving进行训练学习以及最终的测试。表2中包括4类群组行为的平均识别率(Mean per class accuracy,MPCA)以及每个行为的识别率。

表2 本文模型与其他方法在CAD数据集上的识别准确率对比

Tab. 2 Comparison of recongnition accuray of our method and other method %

Model	HOG+SVM	Bag of words	Ref.[18]	Ref.[19]	Ref.[5]	Ref.[7]	Ref.[6]	Ours
Moving	62.2	58.32	84.0	92	87.0	90.77	94.40	82.43
Waiting	53.0	61.10	84.0	69	75.0	81.37	63.60	75.90
Queuing	64.0	66.00	86.0	76	93.0	99.16	100.00	99.35
Talking	40.0	59.80	75.0	99	99.0	84.62	99.50	100.00
MPCA	54.8	61.30	82.8	84	88.3	88.98	89.37	89.42

本文以及文献[5-7]都是使用深度学习网络对群组行为进行识别的方法,从表2中可以明显看出,此类方法表现优异,比HOG+SVM和Bag of words两种经典行为识别模型的平均识别率高了20%左右,究其原因,是这两种传统模型的较多特征是在背景区域提取得到,背景信息干扰较大。本文模型对于群组特征比较明显的Talking,Queuing行为类表现出了优异的识别性能,与文献[18]的对比则更加明显,文献[18]在群组特征较强的Talking,Queuing行为类的识别率明显低于Moving,Waiting两类群组特征较弱的行为,原因是该方法使用手工设计的特征,对人类行为的刻画能力有限,模型缺乏泛化性,当数据来源发生变化时,需要重新设计特征描述符,自适应效果差,不利于群组行为特征的识别。本文模型在同样使用深度学习网络的同类方法[5-7]中性能表现最佳,平均识别率最高,达到了89.42%,并且对Talking类达到了完全正确的识别水平。同时,使用3层LSTM的循环注意力机制关注活动中的关键人物,合理分配了注意力权重,特征数据处理的速度以及识别效果明显要高于使用最大池化结构的其他方法,文献[6]中仅单人行为特征向量维数是本文模型的1.5倍,文献[19]使用AC描述符(Action context descriptor)构建图形化模型的群组行为识别方法,在模型推理阶段花费的周期约是本文的3倍,文献[7]则是将深度光流、场景和个人行为等多种特征信息进行融合,模型复杂且参数量大,对于排球比赛等运动竞技类的行为识别可能更有优势。

表3 本文模型与其他方法在CAE数据集上的平均识别准确率对比

Tab. 3 Performance comparison of our method and other method %

Model	CAE
Bag of words-Person	65.45
Bag of words-Group	57.67
VGG16-Person [20]	83.96
VGG16-Group [20]	77.63
LRCN-Person [21]	74.12
LRCN-Group [21]	69.79
Ours-Person	86.48
Ours-Group	91.23

2.3.2 本文模型与其他方法在CAE上的结果比较

表3是在CAE数据集上本文模型与其他方法得到的单人行为(Person)和群组行为(Group)平均识别准确率的列表,根据文献[17]去除了数据集中群组特征弱的Crossing类别进行实验,避免了误判的情况。本文模型分别在单人行为及群

组行为识别上较Bag of words,文献[20-21]方法识别准确率优势明显,而且与使用相同卷积神经网络的VGG16[20]相比,使用了软注意力循环结构的本文模型在单人行为和群组行为的识别上效果显著,分别达到了86.48%和91.23%的准确率。表中所有方法都是在单人行为识别的基础上对群组行为进行识别,Bag of words,文献[20-21]这3种方法的群组行为识别率皆低于单人行为识别的准确率,而本文模型情况则相反,群组行为的识别率较单人行为明显提高了,证明了本文模型优异的群组行为识别性能。实验效果如图2所示。

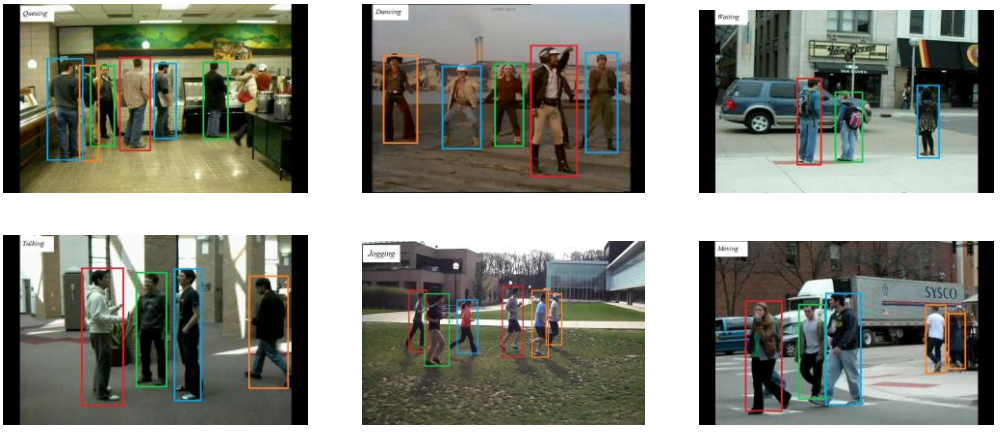


图2 本文模型识别效果图

Fig.2 Examples of recognition with the proposed model

本文提出的方法使用神经网络从人的身上提取外观特征,利用软注意力机制对场景中所有人的行为信息进行池化操作,并重点关注群组行为中的关键人物,合理分配注意力权重以对群组行为做出识别。相对表中的其他方法,本文模型没有繁琐的预处理操作和复杂的建模过程,综合考虑到了活动场景中的所有人,比表中直接排除背景人物的Bag of words模型更具理论说服力,同时识别效果优于使用了长期时间递归卷积网络LRCN的方法[21],证明了本文加入注意力机制之后的群组行为识别模型的有效性。

3 结束语

本文建立了循环的基于注意力机制的群组行为识别模型,重点关注活动中的关键人物,描述了如何动态地对卷积特征进行池化;实验表明使用本文模型进行群组行为识别的效果要优于使用最大池化和平均池化的其他模型;并且进一步证明此模型会更倾向于识别视频帧中的重要元素。实验还表明,本文模型比不使用任何注意机制的网络结构表现更好。未来计划探索混合软硬注意力的方法以降低模型的计算成本,从而扩展到更大的数据集,如排球数据集,其中注意力机制也可以选择集中在较早的卷积层上从而关注视频帧中的较低层特征。

参考文献:

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]// Conference and Workshop on Neural Information Processing Systems. Montreal, CAN: Curran Associates, 2014: 568-576.

[2] Ibrahim M S, Muralidharan S, Deng Zhiwei, et al. A hierarchical deep temporal model for group activity recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA:IEEE, 2016: 1971-1980.

[3] Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition[C]// Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. USA:IEEE, 2012: 1354-1361.
[4] Ramanathan V, Yao B, Li F F. Social role discovery in human events[C]// IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013: 2475-2482.
[5] Hajimirsadeghi H, Yan W, Vahdat A, et al. Visual recognition by counting instances: A multi-instance cardinality potential kernel[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston:IEEE, 2015: 2596-2605.
[6] Wang M, Ni B, Yang X. Recurrent modeling of interaction context for collective activity recognition[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 2-8.
[7] Li X, Chuah M C. Sbgar: Semantics based group activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 2876-2885.
[8] Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition[C]// European Conference on Computer Vision. Firenze, Italy:Springer-Verlag, 2012: 215-230.
[9] Choi W, Shahid K, Savarese S. Learning context for collective activity recognition[C]// Computer Vision and Pattern Recognition. Colorado Springs, USA:IEEE, 2011: 3273-3280.
[10] Rensink R A. The dynamic representation of scenes[J].Visual Cognition, 2000, 7(1/3): 17-42.
[11] Xu K, Ba J, Kiros R, Cho K, 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: IMLS, 2015: 2048-2057.
[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. Long Beach, USA: Curran Associates, 2017: 5998-6008.
[13] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[EB/OL]. https://arxiv.org/abs/1409.2329, 2014.
[14] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]// International Conference on Learning Representations. San Diego, USA: ICLR, 2015: 1-15.
[15] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J].JMLR, 2014, 15(1): 1929-1958.
[16] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]// International Conference on Learning Representations. San Diego, USA: ICLR, 2015: 1-13.
[17] Choi W, Shahid K, Savarese S. What are they doing? : Collective activity classification using spatio-temporal relationship among people[C]// IEEE International Conference on Computer Vision Workshops. Kyoto, Japan: IEEE, 2009: 1282-1289.
[18] Kaneko T, Shimosaka M, Odashima S, et al. A fully connected model for consistent collective activity recognition in videos[J]. Pattern Recognition Letters, 2014, 43(1): 109-118.
[19] Azar S M, Atigh M G, Nickabadi A. A multi stream convolutional neural network framework for group activity recognition [EB/OL]. https://arxiv.org/abs/1812.10328, 2018.
[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations. San Diego, USA: ICLR, 2015: 1-14.
[21] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677-691.

作者简介:

王传旭(1968-),男,博士,教授,研究方向:图像处理、模式识别、计算机视觉;E-mail:543908563@qq.com。

龚玉婷(1994-),女,硕士研究生,研究方向:图像处理、模式识别、计算机视觉。

(编辑:刘彦东)