文章编号：2095-2163(2019)06-0083-06 中图分类号：TP391.41 文献标志码：A

基于关注度网络的行为识别

周 义，范楼苗，张 舟
（合肥工业大学 计算机与信息学院，合肥 230601）

摘 要： 行为识别是计算机视觉领域的一个重要研究课题，具有广泛的应用前景。针对现实中对视频整体序列结构建模会增加大量的冗余信息，提出了一种基于时空关注度长短期记忆网络（Spatial-Temporal Attention Long-Short Term Memory，STA-LSTM）的行为识别框架，提高了行为识别效率。利用 GoogLeNet 逐层卷积视频帧，自动聚合蕴含边、角和线等底层特征以生成具有显著结构性的高层语义特征。在 LSTM 中引入关注度网络来学习关注度权重，利用光流掩膜分割有效的运动前景区域，从而优化关注度权重，将其与卷积特征相结合作为 STA-LSTM 模型的输入特征，从而进行行为识别。在 UCF101 数据集上的实验结果表明，本文方法优于当前的一些先进方法。

关键词： 行为识别；长短期记忆网络；关注度；光流掩膜

Activity recognition based on attention network

ZHOU Yi, FAN Loumiao, ZHANG Zhou
(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

【Abstract】 Action recognition is an important research topic in the field of computer vision and has promising potential applications. On the basis of the fact that modeling the whole sequence of videos in reality will increase a large amount of noise, a framework of action recognition based on spatial-temporal attention Long-Short Term Memory (STA-LSTM) is proposed to improve the efficiency of action recognition. We use GoogLeNet to convolute the frames of videos layer-by-layer, so it can aggregate the low feature of sides, angles and lines automatically to generate high semantic feature which possesses salient structure. We introduce an attention network in LSTM to learn attention weight and use optical flow mask to segmente the foreground of effective motion for the optimization of attention weight. We combine attention weight matrix with convolutional features as the input of STA-LSTM for action recognition. We evaluate our method on UCF101 and experimental results demonstrate that our approach significantly outperforms state-of-the-art techniques.

【Key words】 action recognition; Long-Short Term Memory; attention; optical flow mask

0 引 言

识别视频中的行为动作是计算机视觉重要任务之一，其目的是从视频中提取、分析和表达行为动作信息。该技术正被广泛应用于视频监控、人机交互、医疗看护等领域[1]。随着深度学习技术在计算机视觉中越来越多的应用，也为研究行为识别开拓了新的方向。然而深度学习本身由于需要大数据量和网络参数数目过多等局限性，使得模型在计算方面付出了较大的代价。对此，本文重点研究如何挖掘视频中的有效信息，设计泛华能力强的深度神经网络，识别视频中的行为动作。

早期的一些研究主要是利用卷积神经网络来学习视频中行为的深度表达。Karpathy 等人[2]介绍了一种多规模 Sports-1M 视频数据集，来训练深度卷积神经网络。Simonyan 等人[3]提出一种双流卷积神经网络，通过分别处理 RGB 图像和光流图中的外观和运动信息达到了比较好的行为识别效果。然而，使用卷积神经网络仅能捕捉极少的时序信息。对此，循环神经网络能够较好地解决这个问题，尤其是 LSTM[4]，在视频序列建模方面效果显著。然而现实场景中，由于视频时长以及视频中动作所发生的区域不同，对视频整体序列结构建模会增加大量的冗余信息。对此，本文在循环神经网络中引入关注度机制，其能够模拟人类视觉注意力转移机制，将有限的认知资源聚集于场景中重要的刺激，而抑制那些不重要的信息。具体来说，利用 GoogLeNet[11] 逐层卷积视频帧，自动聚合蕴含边、角和线等底层特征，以生成具有显著结构性的高层语义特征。在 LSTM 模型中引入关注度机制，来学习关注度权重

作者简介：周 义(1994-)，男，硕士研究生，主要研究方向：计算机视觉；范楼苗(1994-)，男，硕士研究生，主要研究方向：计算机视觉；张 舟(1995-)，男，硕士研究生，主要研究方向：计算机视觉。
通信作者：周 义 Email：18297927889@163.com

收稿日期：2019-04-27

系数矩阵。由于视频中的背景噪声和相机移动等因素的影响,利用卷积神经网络作用于 RGB 图像得到的特征不能准确地捕捉视频中的行为动作信息。针对这个问题,本文利用光流掩膜对视频中的运动前景区域进行分割,以此来校正网络所学习到的关注度权重。将关注度系数和卷积特征相结合,生成新的特征激活图序列。其中高值表示显著性区域,即得到 STA-LSTM 网络的显著性输入特征,然后对特征进行学习,从而识别视频中的行为。本文主要贡献是:

(1)提出了一种新颖的深度学习框架——STA-LSTM 用于视频中的行为识别,在端到端的处理过程中,本文方法可以准确地捕捉行为的外观信息和动作信息。

(2)提出的 STA-LSTM 模型能够有效地去除冗余信息,提取行为发生的有效区域,提高模型识别效率。

(3)将本文方法应用于 UCF101 数据集取得了良好的识别效果,与当前一些优秀的研究工作相比,在识别性能方面得到了显著地提升。

1 相关工作

行为识别的目的是从未知视频或图像序列中自动识别其中进行的行为动作,行为本身是相关联的一系列二维空间图像在时间方向上的连接。因此,行为本身具有空间和时间上的结构关联特性。行为特有的空间和时间结构特性,为许多研究者指明了行为识别的正确方向。

早期行为识别主要使用一些传统算法,Vemulapalli 等人[5]在 Lie 群组中用曲线表示每个动作并且使用 SVM 分类器来识别行为。Zanfir 等人[6]提出了一种移动姿态框架,结合修改后的 kNN 分类器进行低延迟行为识别。Carlsson 等人[7]通过从动作视频中提取到的关键帧以及保存的动作原型之间做模板来完成行为,其中,形状信息是用 Canny 边缘检测器得到的边缘数据来表示的。这种方法能够容忍图像和样本之间一定程度的形变,且能够准确识别不同人体姿态形成的相似的形状。Tang 等人[8]采用隐马尔科夫(HMM)模型建模行为的隐状态变化过程。Pei 等人[9]将行为分解为具有语义原子动作集合并定义原子为行为体与目标交互关系的集合,通过与或图学习原子动作的时序关系,能够有效剔除时序错误的与或图行为解释,提升了识别及预测行为的性能。

后来深度学习技术在计算机视觉中得到广泛应用,Heilbron 等人[10]使用序列编码器(即 LSTM),可以模拟随着时间推移的 C3D 特征的演变,使用定位模块生成整个输入视频中不同时间长度的候选提议的开始和结束时间,以进行行为提议。Simonyan 等人[12]通过在光流上训练一个神经网络来整合运动信息。利用外观和光流特性,动作识别的准确性显著提高。Lin 等人[13]尝试使用序列过程提取时空特征,即提取一维时间信息到二维空间信息。该端到端系统考虑长短运动模式,并实现良好的性能。Ng 等人[21]运用深度神经网络模型,结合帧序列分析视频的长期依赖信息用于行为识别 。Srivastava 等人[14]提出了一种基于兴趣点 LSTM 的无监督训练方法,使用编码器 LSTM 将输入序列映射成固定长度表示;然后使用单个或多个解码器 LSTM,对其进行解码以执行输入序列的重构或预测未来序列;最后对这个无监督的预训练 LSTM 进行微调,以适应人类行为识别任务。

融入注意力机制的循环网络模型可以提取行为发生的时空有效区域,有效剔除视频中的冗余信息。Yao 等人[15]介绍了一种时序注意力机制用于视频标题生成。Bazzani 等人[19]提出一种关注度模型学习视频中的重要区域,对每一帧使用高斯混合进行视觉关注度建模。Sharma 等人[18]使用三层 LSTM 网络,引入注意力机制,在网络中加入关注区域的移动、缩放机制,连续部分信息的序列化输入,学习视频的关键运动部位。受这些研究工作的启发,本文使用光流掩膜对视频中的运动前景区域进行分割,在不增加模型复杂度的情况下,还能利用重要的运动信息,能够有效提取场景中显著性区域,实验结果表明本文方法取得了良好的识别正确率。

2 模型框架

本文的模型架构如图 1 所示。首先利用 GoogLeNet 对视频帧序列进行卷积,提取最后一层卷积层特征;在 LSTM 中引入关注度机制[17],作用于卷积层特征的每一个区域;利用光流掩膜提取每一帧的运动前景区域作用于关注度网络,得到新的关注度权重矩阵,将之与卷积层特征相结合。作为 STA-LSTM 模型的输入特征,通过对特征的学习,进而对视频中的行为进行识别。

2.1 特征提取

本文使用在 ImageNet 数据集上预训练好的 GoogLeNet 模型,逐层卷积已重新调节大小为 224×

224 的视频帧序列，提取最后一层卷积层特征。此卷积层包含 1 024 个特征图，包含了输入视频帧的空间外观信息，其形状为 7×7×1 024 大小的特征立方体。因此，在每一个时间步长 t，提取的向量维度是 49×1 024。将这些特征立方体分解为特征片段：$G_t = [G_{t,1}, G_{t,2}, \ldots, G_{t,49}]$，这 49 个特征片段对应于输入视频帧的不同区域，本文的关注度模型就是选择性地关注这 49 个区域。

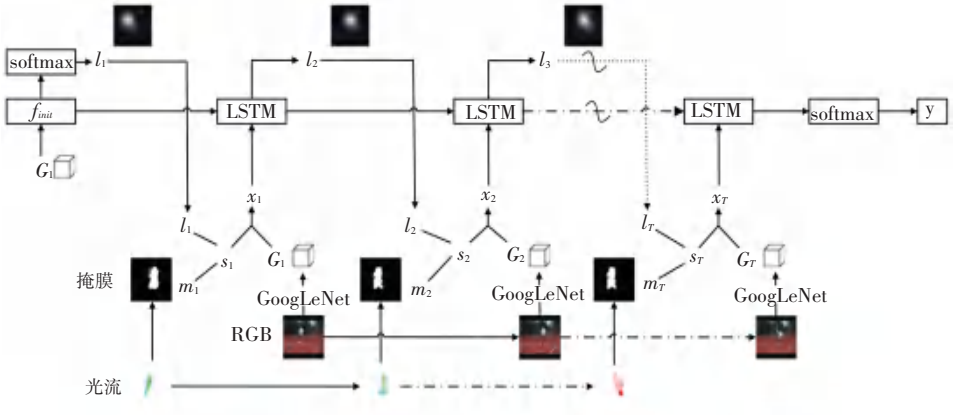

图 1　STA-LSTM 结构

Fig. 1　The structure of STA-LSTM

2.2　时空关注度的表达

使用 GoogLeNet 得到最后一层卷积层特征之后，在 LSTM 中引入关注度机制，作用于卷积层特征的每一个区域。同时，利用光流掩膜分割有效的运动前景，从而修正行为发生的有效区域，即本文提出的 STA-LSTM 模型，如图 2 所示。图中左侧蓝色框内为初始化记忆单元和隐单元。为了达到快速收敛的效果，使用两个三层感知器来初始化 STA-LSTM 模型的记忆单元和隐单元[20]，以此来计算初始的关注度得分公式如下：

$$c_0 = \frac{1}{49}\sum_{i=1}^{49} W_{c,i} G_{1,i},\ h_0 = \frac{1}{49}\sum_{i=1}^{49} W_{h,i} G_{1,i}, \qquad (1)$$

其中，$W_{c,i}$ 和 $W_{h,i}$ 是产生初始记忆单元和隐单元的三层感知器的权重。

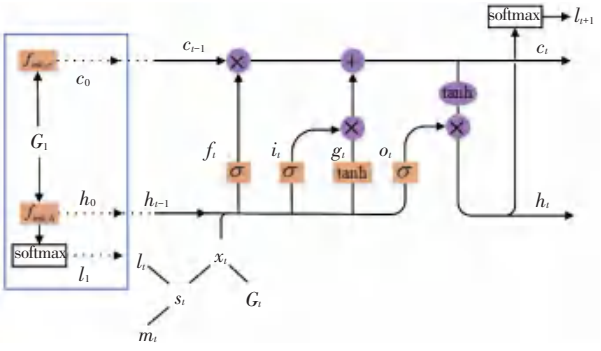

图 2　单个 STA-LSTM 单元

Fig. 2　Single unit of STA-LSTM

行为识别中，视频帧中仅有一部分区域和行为发生相关。显然，为视频帧中不同的区域分配不同的关注度权重，只需要关注这些行为发生的区域。如图 1 所示，针对打网球这一行为而言，主要关注点为手臂、球拍和网球本身。由于视频帧本身是连续的，相邻帧之间存在强烈的时序依赖关系，所以可以利用 $t-1$ 时刻的编码特征来预测 t 时刻的关注度权重，然后用此权重来精炼模型的输入特征，t 时刻单个 STA-LSTM 单元结构如图 2 所示。使用关注度模型作用于视频帧中的 7×7 个区域来预测 49 个区域的关注度权重，其得分 $l_{t,i}$ 可以表示为：

$$l_{t,i} = \frac{\exp(W_{l,i}^T h_{t-1})}{\sum_{j=1}^{49} \exp(W_{l,j}^T h_{t-1})}, \qquad (2)$$

其中，$W_{l,i}$ 表示 softmax 函数对应于第 i 个位置的权重，$i = 1,2,\ldots,49$，$t = 1,2,\ldots,T$；T 为序列化帧数的长度；$l_{t,i}$ 表示第 t 帧的第 i 个区域的关注度权重。

由于场景中存在背景噪声的干扰，而且同种行为可以发生在不同的场景中，因此，人们利用光流掩膜对运动前景和后景进行分割，对行为的发生区域进行初始划分，表示为 $m_{t,i}$，当分割后的第 i 个区域为运动前景时，$m_{t,i}$ 为 1；当分割后的第 i 个区域为背景噪声时，$m_{t,i}$ 为 0。对视频帧的前景和后景进行分割可以对关注度模型扫描区域加以有效地限制。提取出前景区域后，对前景区域中的关注度得分进行统计求和。此处，设置和的阈值为 Th，定义新的时空关注度得分 $s_{t,i}$，如下所示：

$$s_{t,i}=\begin{cases}\dfrac{m_{t,i}l_{t,i}}{\sum_j m_{t,j}l_{t,j}}, & \sum_j m_{t,j}l_{t,j}\geq Th\\ l_{t,i}, & other.\end{cases} \tag{3}$$

其中，$j=1,2,\ldots,49$。

2.3 STA-LSTM 模型

使用光流掩膜对行为前景和后景进行分割，有效地限制了关注度模型的关注范围，而不是利用光流特征和外观特征分别计算关注度得分。在利用外观和动作特征的同时还降低了网络复杂度，减少了计算量。得到上述关注度得分后，如图 2 所示，STA-LSTM 模型的输入可以表示为：

$$x_t=\sum_{i=1}^{49}s_{t,i}G_{t,i}, \tag{4}$$

STA-LSTM 公式如下：

$$f_t=\sigma(W_f^x x_t+W_f^h h_{t-1}+b_f), \tag{5}$$

$$i_t=\sigma(W_i^x x_t+W_i^h h_{t-1}+b_i), \tag{6}$$

$$g_t=\tanh(W_g^x x_t+W_g^h h_{t-1}+b_g), \tag{7}$$

$$o_t=\sigma(W_o^x x_t+W_o^h h_{t-1}+b_o), \tag{8}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t, \tag{9}$$

$$h_t=o_t\odot\tanh(c_t), \tag{10}$$

其中，W 和 b 表示 LSTM 参数。公式(5)、(6)和(8)中的 f_t 是忘记门，i_t 是输入门，o_t 是输出门。g_t 如公式(7)计算所得，表示 t 时刻候选记忆单元状态。公式(9)和(10)中的 c_t 和 h_t 表示 t 时刻记忆单元状态和隐单元状态，x_t 代表 t 时刻的输入特征。$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 表示 sigmoid 和 tanh 激活函数，$\odot$ 表示哈达马积。

STA-LSTM 模型的核心就是忘记门和输入门，忘记门根据当前的输入 x_t、上一时刻状态 c_{t-1} 和上一时刻输出 h_{t-1}，共同决定哪一部分记忆需要被遗忘。输入门根据 x_t，c_{t-1} 和 h_{t-1} 决定哪些部分将进入当前时刻的状态 c_t。STA-LSTM 结构在计算得到新的状态 c_t 后，通过输出门根据最新的状态 c_t、上一时刻的输出 h_{t-1} 和当前的输入 x_t 来决定该时刻的输出 h_t。

最后，使用 softmax 函数作用于最后一个隐单元得到最终结果：

$$y_d=\mathrm{softmax}(W_s h_T+b_s). \tag{11}$$

其中，y_d 代表模型预测值；d 表示子序列的样本编号；W_s 和 b_s 为 softmax 函数的参数。

2.4 损失函数

本文的样本损失函数如下：

$$L=-\sum_{d=1}^{frame-\mathrm{T}+1}\sum_{n=1}^{C}\hat{y}_{d,n}\log y_{d,n}+\lambda\|\theta\|_2 \tag{12}$$

其中，第一项表示交叉熵损失函数[20]，第二项表示模型其它参数的正则化约束。

式中，$frame$ 为视频的帧数；T 为序列化帧数的长度；$\hat{y}_{d,n}$ 为类别的真实标签；$y_{d,n}$ 为模型预测值；C 为类别种类数；n 为类别编号；λ 为权值衰减系数；θ 表示模型所有参数的集合。

3 实 验

3.1 数据集

本文方法所用的数据集为 UCF101[16]，其中包含 13 320 个视频，分为 101 种行为类别，选取每个类别视频总数的三分之二作为训练集，剩下的作为测试集。所有视频均采集于现实场景，在相机移动、物体外观、人物姿态等方面变化多样，因此广泛应用于各种行为分析的研究。

3.2 实验细节及评价标准

将所有视频分解为视频帧序列，并将分辨率重新调整为 224×224 大小，将视频帧序列输入在 ImageNet 数据集预训练好的 GoogLeNet 模型中。本实验取其最后一层卷积层特征作为 STA-LSTM 模型的输入，STA-LSTM 结构隐单元的数量为 1 024，权值衰减系数 λ 设为 10^{-5}，优化算法使用 Adadelta[23]，深度学习框架为 Theano[22]。模型在训练和测试时序列化输入帧的数量均为 $T(T=16)$ 帧，将视频帧按照步长为 1 分成多个 T 帧的子序列。在测试阶段，针对每个视频预测其所有子序列的所属类别，并和标签值相比较统计正确的类别数，作为该视频的识别正确率，最后对所有视频的正确率求均值作为最终的识别正确率。

3.3 实验结果及分析

首先，通过表 1 来验证本文的时空关注度对识别效果产生的影响。其次，通过设置前景区域中时空关注度得分和不同阈值 (Th)，观察模型在 UCF101 数据集上的识别效果，见表 2。最后将本文方法和当前一些优秀方法进行比较，比较结果见表 3。

表 1 时空关注度因素对实验结果的影响

Tab. 1 The influence of experiment result by spatial-temporal attention factor %

方法	UCF101
GoogLeNet+LSTM	81.5
GoogLeNet+STA-LSTM	88.7

由表 1 可明显看出，在引入时空关注度后，本文

所提出的新模型所取得的效果显著，从而证实了本文方法可以应用于行为识别。

表 2　不同 *Th* 值对实验结果的影响

Tab. 2　The influence of experiment result by different *Th* value　%

Th	0.5	0.6	0.7	0.8	0.9
UCF101	85.8	87.3	88.7	85.2	83.0

由表 2 可知，不同的 *Th* 值对实验结果有很大的影响。当 *Th* 较小时，不能提供有效的参考区域，当 *Th* 较大时，由于背景噪声、相机移动、光照条件等影响，造成前景分割的不准确，容易对关注度模型矫正过度。经实验验证，当 *Th* 值为 0.7 时，识别效果最佳。

表 3　不同方法识别性能对比

Tab. 3　Comparison of recognition performance of different methods　%

方法	UCF101
Soft Attention Model[18]	84.96
Composite LSTM Model[14]	84.3
Beyond Short Snippets Models[21]	88.6
RMDN[19]	82.8
本文方法	88.7

表 3 表明，与当前一些优秀方法相比，本文方法所达到的识别正确率更高。而且，相比于其它关注度方法而言，本文通过光流掩膜分割运动前景区域后，模型能够更有效地关注视频中显著区域，提高识别效率的同时并没有增加模型复杂度。如图 3 所示，图中(a)、(b)、(c)分别表示原始视频帧、本文方法所学习到的显著性区域、Soft Attention Model 学习到的显著性区域，可以看出本文方法能够更准确地学习显著性区域。为了进一步论证本文关注度网络的效果，如图 4 所示，在“颠球”这一行为中，本文方法可以准确地捕捉足球、膝盖和脚等显著性区域。

(a) 原始视频帧　(b) 本文关注度　(c) 软关注度

(a) raw frame　(b) our attention　(c) soft attention

图 3　本文方法与 Soft Attention Model 的关注度可视化比较

Fig. 3　Visual attention comparison between soft attention model and our method

图 4　关注度的效果

Fig. 4　The effect of attention

为了更详细地观察本文方法的细节效果，逐帧定位单个视频的具体识别情况。这里以该帧为首的子序列的识别正确率作为该帧的识别正确率。抽取一个行为类别为“扣篮(Basketball Dunk)”的视频，如图 5 所示，观察该视频全部帧的识别情况。为了便于观察，本图只选取识别正确率排名前三的类别，如图 6 所示，分别为“扣篮”、“投篮(Basketball Shooting)”和“扣球(Volleyball Spiking)”。显然，本文方法将该视频正确地识别为“扣篮”，因为“投篮”和“扣篮”的相同点就是这两种行为都需要篮球，“扣球”和“扣篮”相似之处在于“扣”这一动作特性，在不影响判别准确性的前提下，本文方法也将“扣篮”这一行为以微小的概率预测成“投篮”或者“扣球”这两种行为。

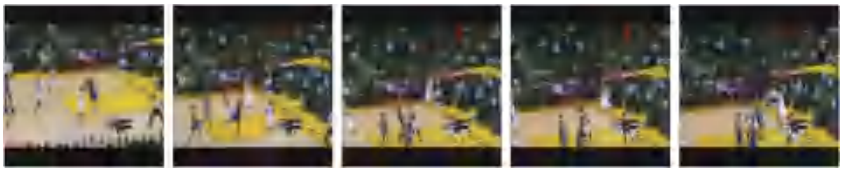

图 5　表示扣篮行为的五帧

Fig. 5　Five frames representing the activity of basketball dunk

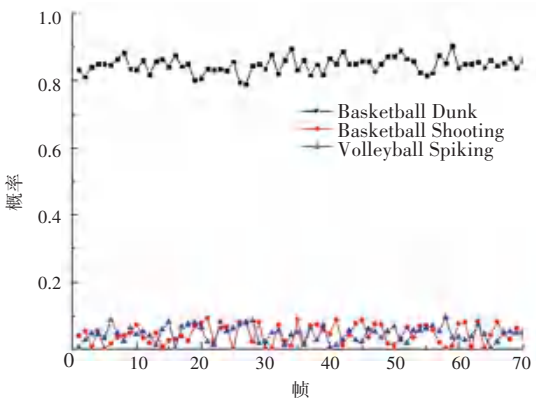

图6 每一帧属于所有类别的概率

Fig. 6 The probability of each frame belonging to all classes

4 结束语

本文提出一种循环时空关注度网络,用于视频中的行为识别。通过外观等特征学习视频中的显著性区域,同时利用光流掩膜分割运动前景区域对关注度网络学习到的显著性区域进行校准划分,使得模型能够更准确地关注视频中的显著性区域从而捕捉更重要的信息,提高行为识别效率。实验结果表明,与当前一些优秀方法相比,本文方法所达到的识别正确率更高。相对于 UCF101 的行为类别较为简单易理解。未来,希望本文的方法可以应用于更加复杂的视频场景中,如大型监控场景下的视频理解、异常检测等,将有助于维护公共安全等领域。

参考文献

[1] POPPE R. A survey on vision-based human action recognition[J]. Image and vision computing, 2010, 28(6): 976-990.

[2] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks [J]. ComputerVision and Pattern Recognition (CVPR), IEEE, 2014: 1725.

[3] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.

[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[5] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3d skeletons as points in a lie group [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 588-595.

[6] ZANFIR M, LEORDEANU M, SMINCHISESCU C. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection [C]//Proceedings of the IEEE international conference on computer vision. 2013: 2752-2759.

[7] CARLSSON S, SULLIVAN J. Action recognition by shape matching to key frames [C]//Workshop on models versus exemplars in computer vision. 2001, 1(18).

[8] TANG K, FEI-FEI L, KOLLER D. Learning latent temporal structure for complex event detection[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1250-1257.

[9] PEI M, JIA Y, ZHU S C. Parsing video events with goal inference and intent prediction [C]//2011 International Conference on Computer Vision. IEEE, 2011: 487-494.

[10] HEILBRON FC, BARRIOS W, ESCORCIA V, et al. Scc: Semantic context cascade for efficient action detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 3175-3184.

[11] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[12] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.

[13] SUN L, JIA K, YEUNG D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 4597-4605.

[14] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using lstms[C]// International conference on machine learning. 2015: 843-852.

[15] YAO L, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure [C]//Proceedings of the IEEE international conference on computer vision. 2015: 4507-4515.

[16] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[17] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[18] SHARMA S, KIROS R, SALAKHUTDINOV R. Action recognition using visual attention[J]. arXiv preprint arXiv:1511.04119, 2015.

[19] BAZZANI L, LAROCHELLE H, TORRESANI L. Recurrent mixture density network for spatiotemporal visual attention[J]. arXiv preprint arXiv:1603.08199, 2016.

[20] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International conference on machine learning. 2015: 2048-2057.

[21] YUR-HEI Ng J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.

[22] BASTIEN F, LAMBLIN P, PASCANU R, et al. Theano: new features and speed improvements[J]. arXiv preprint arXiv:1211.5590, 2012.

[23] ZEILER M D. ADADELTA: an adaptive learning rate method [J]. arXiv preprint arXiv:1212.5701, 2012.