计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):42−48 [doi: 10.15888/j.cnki.csa.006873]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于视觉注意力的人体行为识别①

孔　言, 梁　鸿, 张　千

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)
通讯作者: 梁　鸿, E-mail: liangh6666@163.com

摘　要: 视频中人体行为识别是近年来计算机视觉中的一个重要研究领域, 但是现有的方法对于视频表示方式存在不足, 无法聚焦于图像内的显著区域. 提出了一种基于视觉注意力的深度卷积神经网络, 可以有效地为视频表示特征附加一个权重, 对特征中的有益区域进行注意, 实现更加准确的行为识别. 在自建的 Oilfield-7 油田数据集和 HMDB51 数据集上进行了实验, 以此来验证适用于油田现场人体行为所提出的网络模型的有效性. 实验结果表明, 所提的方法与已取得优异表现的双流架构相比具有一定的优越性.

关键词: 行为识别; 双流架构; 卷积神经网络 (CNN); 视频表示; 视觉注意力

引用格式: 孔言,梁鸿,张千.基于视觉注意力的人体行为识别.计算机系统应用,2019,28(5):42–48. http://www.c-s-a.org.cn/1003-3254/6873.html

Human Action Recognition Based on Visual Attention

KONG Yan, LIANG Hong, ZHANG Qian

(College of Computer &Communication Engineering,, China University of Petroleum, Qingdao 266580, China)

Abstract: Recognition of human actions in videos is an important research field in computer vision in recent years. However, existing methods have insufficient representation of video and cannot focus on significant areas within the image. We propose a deep convolutional neural network based on visual attention, which can effectively add a weight to the video representation features, pay attention to the beneficial regions in the features, and achieve more accurate behavior recognition. We conducted experiments on HMDB51 and our own Oilfield-7 dataset to verify the validity of the model proposed for human actions on the oilfield. The experimental results show that the proposed method has certain advantages compared with the two-stream architectures which have achieved excellent performance.

Key words: action recognition; two-stream architecture; Convolutional Neural Network (CNN); video representation; visual attention

1　引言

人类行为识别作为计算机视觉中的一个基本问题, 现在已经引起了业内的广泛关注. 视频中的行为识别是高级别的视频理解任务中的一个关键问题, 虽然深度卷积神经网络 (CNNs) 凭借其强大的性能已经在图像识别任务中取得了巨大的成功[1−4], 但是在行为识别任务中还没有取得类似于图像识别那样的进展. 在某种程度上, 视频中的行为识别与静态图像中的对象识别有着相似的问题, 这两项任务都必须处理显著的类内变化、背景杂乱和遮挡[5]. 但是, 两者之间存在的明显差异是视频比图像多了一项额外的 (也是重要的) 时间线索, 它可以帮助获得运动信息, 凭借着运动信息可以可靠的识别多种行为[6]. 最近用于动作识别的视频表示主要基于两种不同的 CNN 架构: (1) 3D 时空卷积[7,8];

① 基金项目: 国家科技部创新方法工作专项 (2015IM010300)
Foundation item: Special Fund for Innovation Method by Ministry of Science and Technology of the People's Republic of China (2015IM010300)
收稿时间: 2018-11-05; 修改时间: 2018-11-23; 采用时间: 2018-11-27; csa 在线出版时间: 2019-05-01

(2) 双流架构[6]. 虽然二者都在行为识别中取得了很好的表现, 但是双流结构凭借其容易利用新的超深体系结构和静态图像分类的预训练模型[9], 性能通常优于 3D 时空卷积.

然而, 行为识别中面临的主要挑战仍然是缺乏视频表示方式. 对于人类而言, 可以很容易的将目光聚焦于视频里图像的突出区域, 关注所感兴趣的部分. 但是, 现有的行为识别方法是对视频切分的每个短片段, 平均地汇集该片段所有的局部特征形成全局特征, 针对每个片段的全局特征进行行为类别分类. 采用平均汇聚的方式并不是一个适当的方式, 对于短片段中的每帧图像而言, 能够提供有益的特征的并不是其中的每一个像素或每一块区域, 对于某些区域应该重点关注 (例如, 人类运动, 人机交互), 另外的一些区域 (例如, 背景, 遮挡) 应当有意识的忽略.

受到上述启发, 借助于注意力机制来突出显示视频中的显著区域. 为此, 本文提出了一种基于视觉注意力的深度卷积神经网络, 它将注意力机制融入到双流卷积神经网络中. 注意力机制的特性使得我们可以在没有监督的情况下对每帧图像进行动作的区域定位, 对每个区域空间赋予权重, 然后根据加权求和将局部空间特征聚合起来. 这种不平凡的联合方式简单而有效, 可以容易地解决视频表示不突出的问题. 为了验证这一说法, 在油田现场行为数据集上进行了一系列的基于视频的行为识别实验, 所展现出来的结果表明, 基于视觉注意力深度卷积网络模型是行之有效的.

本文的其余内容如下: 第 2 节讨论了相关的工作, 第 3 节描述了视觉注意力深度卷积网络, 第 4 节给出了实验的细节, 第 5 节总结了本文的工作.

2 相关工作

行为识别作为视觉应用中的一项热门话题, 它的研究进展很大程度上是由于在图像识别方法的进步所推动的. 行为识别的目的是识别每个视频中的单个或多个动作, 通常被描述为一个简单的分类问题. 在 CNNs 还未取得如此巨大成功之前, Laptev I 等人提出利用时空特征将空间金字塔推广到时空域的方法, 检测稀疏时空感兴趣点并使用局部时空特征来进行描述 (包括 HOG 和 HOF), 将其编码入特征包 (BoF) 并结合 SVM 进行动作分类[10]. 随后的工作中, Wang H 等人拓展了四种特征描述符 (HOG3D、HOG/HOF、Cuboids、ESURF) 进行局部时空特征的描述[11], 实验表明局部特征的密集采样的方式比稀疏兴趣点检测表现出更优秀的性能. 随后, Wang H 等人又提出了一种密集轨迹算法进行行为识别, 通过从图像中采样密集点并根据密集光流场的位移信息进行跟踪, 这样密集的轨迹可以覆盖视频中的运动信息[12]. 基于改进的稠密轨迹算法[13]通过消除背景轨迹和扭曲光流获得了更加突出的表现.

随着深度学习的兴起, 具有强大性能的卷积神经网络已经在行为识别领域进行了广泛应用. Karpathy A 等人在 Sports-1M 数据集上使用深层卷积神经网络进行训练, 并对大规模行为分类进行了实证评估[14]. Simonyan K 和 Zisserman A 提出了双流架构, 输入一段视频分别获得视频图像信息和密集光流信息, 为两个流各自训练一个 CNNs 进行动作类别判断, 最后对两分支动作分类得分进行融合[6]. Feichtenhofer C 等人在双流架构的基础上改进了融合的方式, 他们将原本在 Softmax 层的融合提前到卷积层, 在最后一个卷积层对空间网和时态网 (spatial and temporal) 进行融合进一步提高了性能[15]. Ng 等人同样是对双流架构的融合方式上进行了研究, 他们利用 LSTM 对于时序信息具有强大的记忆功能这一特性, 将时态网进行了改进[16]. Wang LM 等人针对视频的特性, 提出了一种基于长范围时间结构建模的网络, 结合了稀疏时间采样策略和视频级别监督方式对整个视频段进行学习时的有效和高效[17]. Tran D 等人提出的 3D 卷积神经网络 (C3D) 是并列于双流架构的另一种处理视频级别的动作分类的主流方法, 由于 2D 卷积不能很好的捕获视频中的时序信息, 因此提出的 3D 卷积将原来的卷积层和池化层扩展为 3D 卷积和 3D 池化. 通过 3D 卷积可以直接处理视频[7]. Carreira J 等人提出了一种新的结合 3D 卷积和双流网络的模型, 称之为 I3D, 该模型可以从视频中学习无缝空间-时态特征提取器. 而且 I3D 模型是一种通用的视频表示的学习方法[18].

人类的视觉注意一直以来都是计算机视觉界所关注的问题. Hou XD 等人基于连续的特征抽样提出了一种注意力模型, 对于显著性特征使用能量概念进行解释. 另外, 该模型可实现在静态场景和动态场景之间注意力的选择性[19]. Mathe S 等人针对视频中的动作识别进行了一系列的研究实验, 主要是人类视觉注意力和计算机视觉中的关联关系[20]. 与上述工作相比, 本文提

出的基于视觉注意力的深度卷积网络将人类的注意力机制融入到现存的深层 CNNs 中, 通过对模型结构轻量级的修改, 使得处理后的视频表示特征具有了局部显著性.

3 视觉注意力深度卷积网络

在本节中, 将介绍视觉注意力深度卷积网络, 即 AttConv-net. 最近的从视频中的行为识别的一种标准做法是使用多种信息流 (RGB 和光流) 的融合, 这样的方式可以获得显著的性能[6,15,21]. 在此之前, 先介绍网络的基础架构——时态段网络[17]. 最后, 对注意力机制进行了描述.

3.1 时态段网络

在时态段网络 (Temporal Segment Networks, TSN) 提出之前的双流卷积网络无法对远距离时间结构的视频数据进行建模, 只能处理空间网络中的单个帧或是时态网络单个栈, 不能有效地获取时序中上下内容的联系. 时态段网络通过一种稀疏采样的方式, 从整个视频中获取一系列短片段, 这样可以整合整个视频的视觉信息进行视频级别的分类. 每个片段都将给出其本身对于行为类别的初步预测, 从这些片段的"共识"来得到视频级别的预测结果[17].

具体来说, 给定一段视频V, 将其按相等间隔分为K段$\{S_1,S_2,\cdots,S_k\}$. 然后, 时态段网络按照如下方式对一系列短片段进行建模:

$$TSN(T_1,T_2,\cdots,T_k)= H(G(F(T_1;W),F(T_2;W),\cdots,F(T_k;W))) \tag{1}$$

其中, $(T_1,T_2,\cdots,T_k)$表示片段序列, 每个片段T_k从它对应的段S_k中随机采样得到, $F(T_k;W)$ 函数代表采用W作为参数的卷积网络作用于短片段T_k, $G()$代表段共识函数, 结合多个短片段的类别得分输出以获得它们之间关于类别判断的共识, 函数$H()$会根据这个共识预测整段视频属于每个行为类别的概率. 另外, 关于共识的损失函数G的形式为:

$$L(y,G)=-\sum_{i=1}^{C}y_i(G_i-\log\sum_{j=1}^{C}\exp G_j) \tag{2}$$

其中, C是行为类别的数量, y_i是关于i类的真实标签.

3.2 模型架构

AttConv-net 分别对双流中的空间网和时态网所提取的特征分配较大的权重, 使其容易地定位到感兴趣地区域, 从而可以更准确进行分类. 该结构如图 1 所示, 采用双流模型基础架构, 分为空间流网络和时态流网络. 本文的 AttConv-net 是在 TSN 的基础上进行了修改, 将注意力模型分别连接到空间网和时态网的最后一个卷积层所提取出的特征上, 之后将分配了权重的特征送入全连接层以及 Softmax 进行双流网络各自的类别概率的预测, 并且在评判最终视频所属类别之前会将空间流和时态流的网络结果进行合并. 给定一个完整视频 V, 将其处理成一系列的片段$S_i(i=1,2,\cdots k)$, k是一整个视频均等分的数量, 每个片段包含一帧 RGB 图和两帧光流图. 卷积神经网络 CNNs 分别提取 RGB 图的全局视觉特征$F_{\text{RGB}}=(F_1,F_2,F_3,\cdots,F_L)$ 和光流图的全局视觉特征$F_{\text{OF}}=(F_1,F_2,F_3,\cdots,F_L)$,$L$表示每张图像划分为了 L 块区域, 每个区域都是一个m维的向量. 融入注意力机制处理后得到特征$F_{att\text{RGB}}$和$F_{att\text{OF}}$, 之后便会得到每个片段S_i的双流网络中的所属类别得分C_{Si}和C_{Ti}, 经过共识函数$G()$后将双流结果送入 Softmax 函数算概率, 进而得到一个完整视频的分类结果W. 其中的工作流程可以概括为下列共识:

$$F_{att\text{RGB}}=f(F_{\text{RGB}}) \tag{3}$$

$$F_{att\text{OF}}=f(F_{\text{OF}}) \tag{4}$$

$$g_S=G\left(\sum_{i=1}^{k}C_{Si}\right) \tag{5}$$

$$g_T=G\left(\sum_{i=1}^{k}C_{Ti}\right) \tag{6}$$

$$W=\text{Softmax}(g_S,g_T) \tag{7}$$

式 (3) 和式 (4) 分别是用注意力模型对特征F_{RGB}和F_{OF}进行区域空间权重分配所得到的注意力特征$F_{att\text{RGB}}$和$F_{att\text{OF}}$, 式 (5) 和式 (6) 分别是用共识函数分别对空间流和时态流中所有片段的属于同一类别的得分做个求和均值得到g_S和g_T, 式 (7) 是融合双流网络的得分所获得的整个视频的分类结果W.

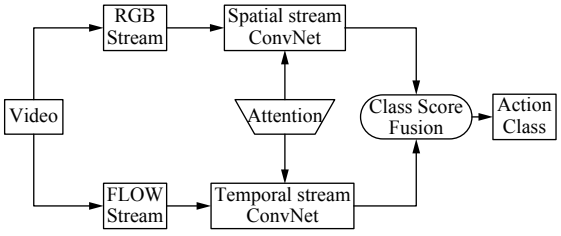

图 1 AttConv-net 模型结构图

3.3 视觉注意力模型

AttConv-net 中的注意力模型将从最后一个卷积层输出的特征向量附加一个介于 0 和 1 之间的权重, 以此聚焦于图像中的显著区域, 该模型结构如图 2 所示, 将视频片段输入到网络中, 空间流和时态流分别进行各自的卷积, 图中的虚框部分表示空间流和时态流进行相同的 Attention 处理, 输出的分数是两流网络的单独得分. 卷积神经网络提取的空间流特征F^t_{RGB}和时态流特征F^t_{OF}都是一个$L\times m$维的向量, 即图像有L个区域, 每个区域用m维的特征向量表示:

$$F^t_{\mathrm{RGB/OF}}=\left\{F^t_1,F^t_2,F^t_3,\cdots,F^t_L\right\},F_i\in R^m,t=(1,2,\cdots,k) \tag{8}$$

其中, R^m表示m维的特征表示, F_i表示第i个图像区域, F^t表示以时刻t为中心所代表的视频段的特征表示. 对于每个图像区域, 注意力函数O_{att}根据特征向量F^t_{RGB}和F^t_{OF}生成对应视频采样片段t的注意力权重α^t_i:

$$\alpha^t_i=O_{att}\left(F^t_{\mathrm{RGB/OF}}\right) \tag{9}$$

归一化处理:

$$\alpha^t_n=\frac{\exp\left(\alpha^t_i\right)}{\sum_{n=1}^{L}\exp\left(\alpha^t_n\right)} \tag{10}$$

其中, α^t_n表示注意力模型中第n个图像区域的在视频段t的权重.

经过注意力模型处理后的特征$F_{att\mathrm{RGB/OF}}$:

$$F_{att\mathrm{RGB/OF}}=\sum_{n=1}^{L}\alpha^t_nF_{\mathrm{RGB/OF}} \tag{11}$$

AttConv-net 之后将$F_{att\mathrm{RGB/OF}}$送入全连接层. 融入注意力机制的网络仍然是可以通过标准的反向传播来优化学习.

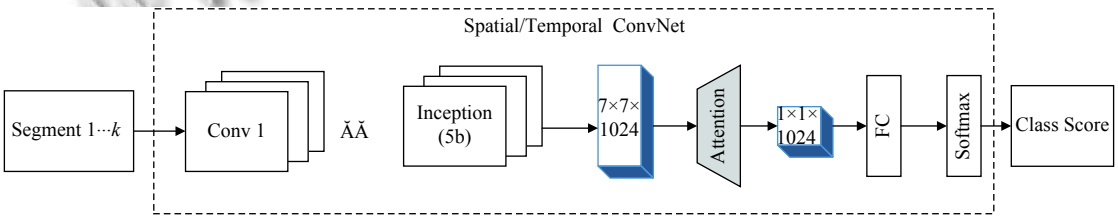

图 2 AttConv-net 网络结构图

4 实验

在本节中, 首先介绍了自建的油田视频数据集, 接下来评估 AttConv-net 在自建数据集以及 HMDB51 上的性能, 在此之外本文还对行为识别任务中所取得良好表现的时态段网络进行有效性实验, 即 AttConv-net 改进前的基础网络 TSN. 最后, 实验结果展现出与一些方法进行了比较, 并且可视化了注意力聚焦图.

4.1 油田人员行为数据集

油田人员行为数据集 (Oilfield-7) 包括 7 个动作类别: Pump Clean, Rig Walk, Room Operate, Site Walk, Room Stand, Tank Construction, Tank Walk. Oilfield-7 数据集包含剪辑好的视频 200 个, 每个视频平均 300 帧. 我们按照数据集的标准评估准则[22], 将数据划分为三种不同的训练和测试数据, 分类表现是按照三种划分数据多产生的平均识别精度. 另外, 数据集在进行剪辑时, 确保一个视频中只出现一种行为, 对于出现多种行为的视频中予以舍弃.

4.2 实验细节

时态段网络[17]是最近用于从视频中进行行为识别的表现最为突出的一个双流模型, 它在两个大型的行为数据集 HMDB51[23]和 UCF101[24]上分别获得 69.4% 和 93.5% 的准确率. 在本文实验中, 选用时态段网络在 Oilfield-7 数据集上进行训练去提取视频段的特征. 对于空间网和时态网, 选用的深层卷积网络结构是 BN-Inception[25], 这与时态段网络中的设置一致[17]. 由于用于行为是别的数据集数量较少的原因, 存在着过拟合的风险, 为此进行了数据增强. 调整输入的 RBG 图和光流图为 256×340, 并且从{256, 224, 192, 168}集合中进行宽度和高度的随机选择裁剪, 之后统一调整为224×224作为网络的输入. 其中的光流图是采用 TV-L1 光流算法得到[26]. 根据先前的工作[6,21], 测试时从每个视频中选择 25 帧 RGB 图或光流栈, 对于每个帧/栈, 通过裁剪四个角和一个中心的方式进行数据增强. 本文的网络参数的学习使用小批量随机梯

度下降算法进行, 用于特征提取的深层卷积神经网络是在 ImageNet[27]上事先预训练的. 网络训练过程中的学习率为10^{-3}, 辍学率为 0.8, 在视频类别最终得分进入融合时, 空间流的权重设置为 1, 时态流的权重设置为 0.5.

4.3 结果与分析

实验中通过与 AttConv-net 的两种变体来进行比较, 一个是基线深层卷积神经网络 BN-Inception, 不包含注意力机制和 TSN 结构; 另一个是不包含注意力机制的 TSN_BN-Inception. 如表 1 所示, 本文方法表现出了较好的优越性. 引入注意力去处理远距离时间结构的视频时, 对于片段图像的局部区域可以获得更高的关注度, 因此视频片段的特征表示更能反映出视频的所属类别.

表 1 在 Oilfield-7 数据集 (Split1) 上三种模型性能比较

模型 (AUC)	空间网	时态网	双流网
BN-Inception	0.896	0.536	0.873
BN-Inception_TSN	0.916	0.571	0.896
AttConv-net(我们)	**0.923**	**0.582**	**0.908**

针对 3 种划分数据分别进行实验, 每一部分所展示出的是融合了两流的准确率. 之后, 最终的比较结果是将 3 个部分取得平均准确率, 所展示的结果见表 2. AttConv-net 与另外两个变体方法比较, 表现出了最优的性能. 与 BN-Inception 相比, 平均准确率提高了 2.3%, 与 BN-Inception_TSN 相比, 提高了 1.4%.

表 2 在 Oilfield-7 数据集上三种模型性能比较

模型 (AUC)	Split1	Split2	Split3	Average
BN-Inception	0.873	0.876	0.871	0.873
BN-Inception_TSN	0.896	0.895	0.896	0.896
AttConv-net(我们)	**0.908**	**0.917**	**0.905**	**0.910**

进一步的为了验证本文方法的优越性, 用数据集 HMDB51[23]来测试其性能. HMDB51 数据集共有 6849 个视频段包含 51 类人体行为类别, 每个类别含有 101 个视频段且都经过人为标注. 同样的也是切分为 3 种不同的训练和测试数据, 分类表现是按照 3 种划分数据多产生的平均识别精度. 此次实验中所设置的相关参数和前文实验细节中的一样, 所展示结果见表 3. AttConv-net 与另外两个变体方法比较, 表现出了最优的性能. 与 BN-Inception 相比, 平均准确率提高了 1.3%, 与 BN-Inception_TSN 相比, 提高了 0.4%.

表 3 在 HMDB5 数据集 (Split1) 上三种模型性能比较

模型 (AUC)	空间网	时态网	双流网
BN-Inception	0.607	0.646	0.685
BN-Inception_TSN[17]	0.618	0.651	0.694
AttConv-net(我们)	**0.623**	**0.658**	**0.698**

在 Oilfield-7 数据集中, 使用 3 个模型进行了测试并使用 mAP 评价指标, 结果如表 4 所示. AttConv-net 与另外两种相比, 在 mAP 中取得了最好的表现. 但是, 在“Room Operate”和“Tank Construction”行为类别中, BN-Inception_TSN 展现出的结果优于 AttConv-net, 因为这两个类别中的人类行为表现的不为明显, 注意力更多地聚集在了背景当中, 而丢失了对动作的关注, 所带来的负面效果使得准确率降低. 为了更好地理解网络在学习过程中对图像局部区域的显著性, 本文可视化了部分注意力图, 如图 3 所示. 图中第 1 列代表的是从视频中提取的原始图像, 第 2 列是经过注意力关注后所得到最精准的效果, 第 3 列代表了注意力关注时的最宽泛的效果. 例如, 对于 Tank Walk(第 4 行), 可以关注并将焦点缩小到场地中行走的人, 但是对于 Pump Clean(第 2 行), 由于数据集数量的问题, 进而产生的过拟合的结果, 从而导致图像无法精准聚焦而产生偏离.

5 结论与展望

本文提出了基于视觉注意力的深度卷积的人体行为识别方法, 称之为 AttConv-net. 该方法利用注意力机制在图像中对于全局信息有了显著性理解, 聚焦于局部区域获取信息, 更加准确而高效的实现视频分类. 在自建的 Oilfield-7 数据集上进行的实验表明, AttConv-net 相较于基础深层卷积网络 BN-inception 和时态段网络 TSN 获得了更高的行为识别精度, 证明了注意力的有效性. 为了进一步证实本文方法性能的优势, 使用数据集 HMDB51 来验证, AttConv-net 也取得了较好的性能. 但是其中存在些许不足, 在时态流的网络训练过程中该方法所得到的精度不高, 这由于 Oilfield-7 数据集中的人类的动作幅度小, 提取的光流图中的信息丢失了大部分运动信息, 从而造成了较低的识别精度. AttConv-net 中两流卷积网络在进行特征融合时是采用共识函数去完成的, 视频中的些许片段存在噪声标签, 从而影响视频分类. 接下来的工作将探索一种片段特征聚合的方式来替代共识方式, 进一步的研究其对行为识别任务的影响.

表 4 在 Oilfield-7 数据集上 3 种模型的 AP 评价指标比较 (第 1 行代表 7 种类别)

模型 (AP)	PC	RW	RO	SW	RS	TC	TW	mAP
BN-Inception	0.594	0.745	0.683	0.838	0.786	0.674	0.738	0.723
BN-Inception_TSN	0.656	0.779	**0.766**	0.885	0.828	**0.796**	0.835	0.792
AttConv-net(本文)	**0.696**	**0.825**	0.716	**0.936**	**0.866**	0.737	**0.868**	**0.806**

图 3 Oilfield-7 数据集部分行为注意力变化的可视化图像

参考文献

1 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. Las Vegas, NV, USA. 2016. 770–778.

2 He KM, Zhang XY, Ren SQ, *et al*. Identity mappings in deep residual networks. European Conference on Computer Vision. Springer. The Netherlands. 2016. 630–645.

3 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

4 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. arXiv:1602.07261, 2017.

5 Nguyen TV, Song Z, Yan SC. STAP: Spatial-temporal attention-aware pooling for action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(1): 77–86. [doi: 10.1109/TCSVT.2014.2333151]

6 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. arXiv: 1406.2199, 2014.

7 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4489–4497.

8 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]

9 Girdhar R, Ramanan D, Gupta A, *et al*. ActionVLAD: Learning spatio-temporal aggregation for action classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3165–3174.

10 Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies. 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. 2008. 1–8.

11 Wang H, Ullah MM, Klaser A, *et al*. Evaluation of local spatio-temporal features for action recognition. BMVC 2009-British Machine Vision Conference. London, UK. 2009. 124.1–124.11.

12 Wang H, Kläser A, Schmid C, *et al*. Action recognition by dense trajectories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2011). Providence, RI, USA. 2011. 3169–3176.

13 Wang H, Schmid C. Action recognition with improved trajectories. Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia. 2013. 3551–3558.

14 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1725–1732.

15 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition.

Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1933–1941.

16 Ng JYH, Hausknecht M, Vijayanarasimhan S, *et al*. Beyond short snippets: Deep networks for video classification. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4694–4702.

17 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. European Conference on Computer Vision. The Netherlands. 2016. 20–36.

18 Carreira J, Zisserman A. Quo Vadis, Action recognition? A new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 4724–4733.

19 Hou XD, Zhang LQ. Dynamic visual attention: Searching for coding length increments. Advances in Neural Information Processing Systems. 2009: 681–688.

20 Mathe S, Sminchisescu C. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In: Fitzgibbon A, Lazebnik S, Perona P, et al, eds. Computer Vision–ECCV 2012. Berlin, Heidelberg: Springer, 2012. 842–856.

21 Wang LM, Xiong YJ, Wang Z, *et al*. Towards good practices for very deep two-stream ConvNets. arXiv: 1507.02159, 2015.

22 Idrees H, Zamir AR, Jiang YG, *et al*. The THUMOS challenge on action recognition for videos “in the Wild”. Computer Vision and Image Understanding, 2017, 155: 1–23. [doi: 10.1016/j.cviu.2016.10.018]

23 Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: A large video database for human motion recognition. 2011 International Conference on Omputer Vision. Barcelona, Spain. 2011. 2556–2563.

24 Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012.

25 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv: 1502.03167, 2015.

26 Wedel A, Pock T, Zach C, *et al*. An Improved Algorithm for TV-L1 optical flow. In: Cremers D, Rosenhahn B, Yuille A L, et al, eds. Statistical and Geometrical Approaches to Visual Motion Analysis. Berlin, Heidelberg: Springer, 2009. 23–24.

27 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–25.