计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190800101

行为关联网络:完整的变化行为建模

何 鑫[1] 许 娟[1,2] 金莹莹[1]
1 南京航空航天大学计算机科学与技术学院 南京 211100
2 东南大学计算机网络与信息集成教育部重点实验室 南京 210096
(hexin@nuaa.edu.cn)

摘 要 针对视频中的完整行为建模,目前常用的方法为时间分段网络(Temporal Segment Network,TSN),但TSN不能充分获取行为的变化信息。为了在时间维度上充分发掘行为的变化信息,文中提出了行为关联网络Action-Related Network(ARN),首先使用BN-Inception网络提取视频中行为的特征,然后将提取到的视频分段特征与Long Short-Term Memory(LSTM)模块输出的特征拼接,最后进行分类。通过以上方法,ARN可以兼顾行为的静态信息和动态信息。实验结果表明,在通用数据集HMDB-51上,ARN的识别准确率为73.33%,比TSN提高了7%;当增加行为信息时,ARN的识别准确率将比TSN提高10%以上。而在行为变化较多的数据集Something-Something V1上,ARN的识别准确率为28.12%,比TSN提高了51%。最后在HMDB-51数据集的一些行为类别上,文中进一步分析了ARN和TSN分别利用更完整的行为信息时识别准确率的变化情况,结果表明ARN的单个类别识别准确率高于TSN 10个百分点以上。由此可见,ARN通过关联行为变化,对完整行为信息进行了更充分的利用,从而有效地提高了变化行为的识别准确率。

关键词:行为识别;行为关联网络;深度学习;计算机视觉
中图法分类号 TP391

Action-related Network:Towards Modeling Complete Changeable Action

HE Xin[1],XU Juan[1,2] and JIN Ying-ying[1]
1 College of Computer Since and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211100,China
2 Key Laboratory of Computer Network and Information Integration,Ministry of Education,Southeast University,Nanjing 210096,China

Abstract When modeling the complete action in the video,the commonly used method is the temporal segment network (TSN),but TSN cannot fully obtain the action change information. In order to fully explore the change information of action in the time dimension,the Action-Related Network (ARN) is proposed. Firstly,the BN-Inception network is used to extract the features of the action in the video,and then the extracted video segmentation features are combined with the features output by the Long Short-Term Memory (LSTM),and finally classified. With the above approach,ARN can take into account both static and dynamic information about the action. Experiments show that on the general data set HMDB-51,the recognition accuracy of ARN is 73.33%,which is 7% higher than the accuracy of TSN. When the action information is increased,the recognition accuracy of ARN will be 10% higher than TSN. On the Something-Something V1 data set with more action changes,the recognition accuracy of ARN is 28.12%,which is 51% higher than the accuracy of TSN. Finally,in some action categories of HMDB-51 dataset,this paper further analyzes the changes of the recognition accuracy of ARN and TSN when using more complete action information respectively. The recognition accuracy of ARN is higher than TSN by 10 percentage points. It can be seen that ARN makes full use of the complete action information through the change of the associated action,thereby effectively improving the recognition accuracy of the change action.

Keywords Action recognition,Action-related network,Deep learning,Computer vision

1 引言

对视频内容的理解,尤其是对视频中人类行为的识别,是计算机视觉研究的核心课题之一。随着深度学习理论的发展,以及越来越多的行为识别视频数据集被公布[1-4],行为识别方法取得了很多突破[5-10]。但是,目前存在的大多数行为识别方法不能识别完整的行为信息,如果行为前后变化较大,则对部分行为信息的获取将导致错误的识别结果。例如助跑跳远,如果只识别前半部分的行为信息则很容易将其识别为跑步。Wang等在2016年提出的时间分段网络采用对视频

到稿日期:2019-08-22 返修日期:2019-11-23 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:许娟(juanxu@nuaa.edu.cn)

分段的策略，使得 TSN 可以利用完整的视频信息。然而 TSN 对每个分段的识别结果采用平均得分的融合方式，这种融合方式在识别一些前后变化较大的行为时，由于每个分段的识别结果不一致，会误导模型产生错误的识别结果。

为了更有效地利用完整视频信息，本文提出了一种行为关联网络模型，这种模型采用稀疏采样策略进行端到端训练，并在时间维度上充分关联行为的变化信息。实验结果表明：1)在 UCF-101，HMDB-51 和 Something Something V1 数据集上，ARN 的识别准确率均高于 TSN；2)当可利用的行为信息增加时，ARN 的识别准确率将进一步提高；3)行为的前后差异越大，变化行为越多，ARN 识别精确度的优势就越明显。总之，ARN 通过关联行为变化信息的方式，有效地利用了完整行为信息，从而显著提高了变化行为的识别准确率。

2 相关工作

近年来，基于深度学习的行为识别技术已经被广泛研究。研究人员提出了许多优秀的模型来自动识视频中的人类行为。与本文相关的工作主要有以下两点：对视频的时空信息建模和对完整的视频建模。

(1)对视频的时空信息建模。随着深度学习的发展，卷积神经网络在图像分类任务上取得了非常好的表现[5]，研究人员开始尝试将卷积网络应用到视频分类任务中。由于视频比图片多了一个时间维度，探索视频的时空信息的建模方法就很重要。以下 3 种对视频的时空信息建模的方法被人们广泛研究：1)双流网络[6]，双流网络使用并行的网络结构，其中一个网络使用 RGB 帧对空间信息进行建模，另一个网络使用堆叠的光流对时间信息建模，最后将两个网络的识别结果进行融合；2)3D 卷积网络[7]，该类方法将 2D 卷积网络扩充为 3D 卷积网络，然后利用连续的 RGB 帧同时获得视频的时空信息；3)2D 卷积网络与 LSTM 结合[8]，这类方法将视频看作由有序的 RGB 图片构成的序列，先利用 2D 卷积网络提取特征，然后利用 LSTM 对时间信息进行建模。Carreira 等结合双流网络模型和 3D 网络模型[9]，将双流网络的特征提取网络由 2D 扩充到 3D，同时使用超大规模数据集 Kinects 对模型进行预训练，最终在公开数据集 UCF-101 和 HMDB-51 上取得了当时最高的识别准确率。但是 3D 卷积网络具有大量的参数，运算量极大，需要使用大规模数据集预训练才能缓解其在小规模数据集上过拟合的问题。针对 3D 卷积网络运算量大的问题，文献[10-12]分别将 3D 卷积核拆分成两个串联的空间卷积核和时间卷积核，该方法大大减少了模型的参数。由于光流提取过于耗时，Crasto 等使用 3D 卷积网络在仅使用 RGB 特征的情况下模拟光流特征[13]，该方法在保持较高识别准确率的情况下大大提高了行为识别的时间效率。Sun 等通过数学推导的方式得到了与光流正交的特征[14]，从而在不使用光流特征的情况下提升了行为识别的准确率。

(2)对完整的视频信息进行建模。以上方法存在一个共同的问题就是无法对完整的视频信息进行建模。这些方法只利用了视频的一小段信息，得到的识别结果比较片面。Xiong 等于 2016 年提出了时间分段网络（Temporal Segment Network，TSN）[15]，可以用于完整的视频信息。TSN 在双流网络的基础上对视频进行分段，每个分段内仍采用之前的双流网络结构，最后对分段得分进行融合从而得到视频级的预测结果。TSN 在训练阶段使用数据增强以及 ImageNet 数据集预训练等训练策略，最后在 UCF-101 和 HMDB-51 数据集上分别取得了 94.0%和 68.5%的准确率。但是 TSN 对各个分段的融合采用了平均池化的融合方法，这种静态的融合策略无法发掘行为在视频分段间的联系，视频的各分段之间相互孤立，随着视频分段的增加，TSN 的识别准确率并没有获得提升，该结果说明 TSN 并不能很好地利用完整的视频信息。之后研究人员对 TSN 的分段融合策略做出了改进。Zhou 等于 2018 年提出利用简单的全连接层来关联视频分段中的行为信息[16]，该方法过于简单，无法发掘较长时间的行为联系。同年，Zolfaghari 等提出利用 3D 卷积网络融合视频分段[17]。3D 卷积网络的参数极多，并且需要使用大规模数据集 Kinetics 对 3D 卷积网络进行预训练，才可以在小规模数据集上取得较好的识别效果。同年，Chih-Yao 等基于 TSN 模型的分段思想，将每个分段内得到的时间特征和空间特征拼接，然后将拼接的时空特征输入到 LSTM 中对分段进行关联[18]。该模型由于对视频进行了分段，抽样得到的视频帧之间的差异较大。相比 CNN+LSTM 方法[8]，该模型的准确率得到了提升，但是提升空间非常小。在 UCF-101 数据集上，其准确率相对于 TSN 由 94.0%提升为 94.1%，在 HMDB-51 数据集上由 68.5%提升为 69.0%。

相比上述方法，本文提出的 ARN 主要有 3 点不同：1)类似于 TSN，通过稀疏采样策略得到横跨整个视频的帧，并且不受视频时长的影响；2)由于 LSTM 在许多时序任务上[19-20]取得了很好的表现，该方法使用 LSTM 来发掘运动模式的演化信息；3)与之前的 CNN+LSTM 方法不同，该方法将卷积特征与 LSTM 的输出征拼接之后再输入全连接层进行分类。该方法由于可以兼顾行为的外观信息和演化信息，因此可以显著弥补一些类别在 CNN+LSTM 系列模型上识别准确率下降的缺陷。实验结果表明，LSTM 在执行行为识别任务时是有效的，但是需要设计合适的方法才可以充分发挥它的作用。

3 行为关联网络

本文结合以下方法设计 ARN，其结构如图 1 所示。首先，可以利用 RGB 帧包含的外观信息来准确识别行为，与其相邻的帧包含大量的冗余信息，因此本文仅采样时间邻域内的单个 RGB 帧，利用特征提取模块来提取行为特征。然后，对于一些前后差异较大的复杂行为，需要关联这类行为随时间变化的完整信息才能使其被准确识别。LSTM 已经被证明可以有效地处理时序问题[19-20]，因此将第一步得到的特征表示输入到主要由 LSTM 构成的行为关联模块，从而对视频分段中包含的行为信息进行关联，最终得到视频级的预测结果。

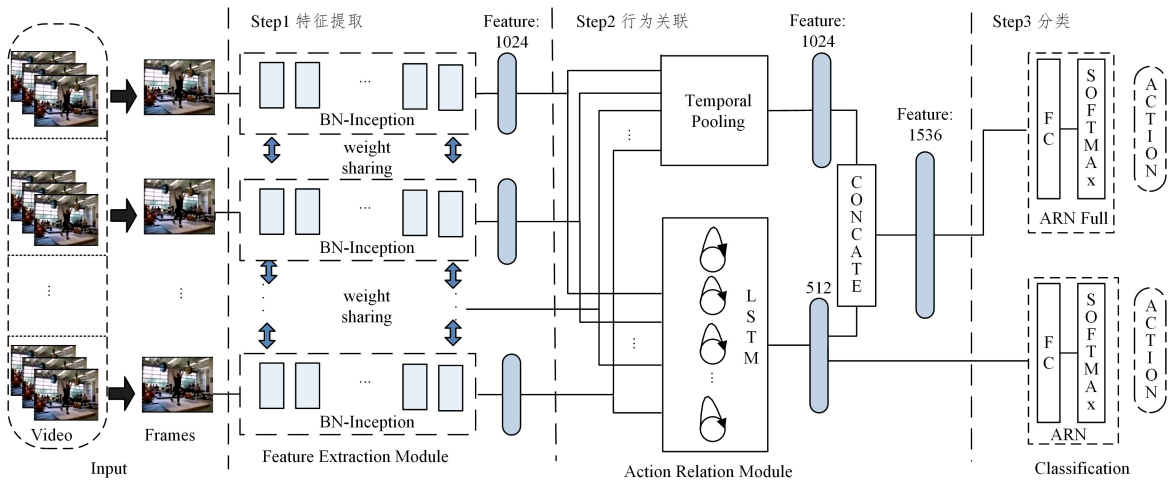

图 1　ARN 和 ARN Full 网络

Fig. 1　Architecture overview of ARN and ARN Full

3.1　ARN 和 ARN Full

ARN 中的行为关联模块主要用来学习行为在视频分段之间的联系,但是对于一些较简单的可以通过外观信息准确识别的行为,可以直接使用 RGB 特征进行识别。受文献[17]的启发,本文基于 ARN 提出了扩充版的 ARN Full。ARN Full 与 ARN 的主要区别在于关联模块,如图 2 所示。ARN Full 将特征提取模块提取到的特征以并行的方式分别输入时间池化模块(对每个分段的特征进行平均)和 ARN 的关联模块,然后将两者的输出结果进行拼接,最后进行分类;而 ARN 则将特征提取模块提取到的特征直接输入行为关联模块,然后直接对行为关联模块的输出进行分类。在 HMDB-51 数据集上的识别结果表明,ARN 属于 CNN+LSTM 类型的方法,因此 ARN 的识别准确率提升不高。ARN Full 由于能够兼顾行为的外观信息和行为的前后变化信息,因此可以很好地解决上述部分行为类别识别准确率下降的问题,并且可以进一步提升这些行为的识别准确率。

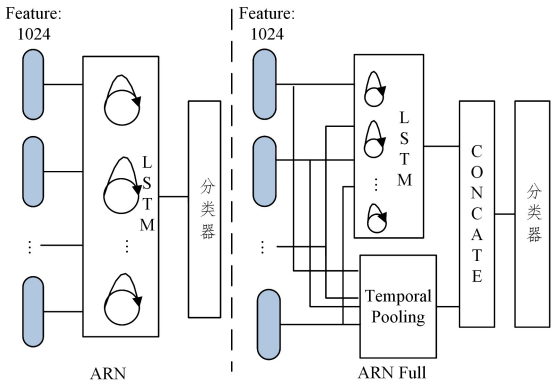

图 2　ARN 和 ARNFull 的结构对比

Fig. 2　Architecture comparison of ARN and ARN Full

3.2　详细的网络结构

(1)特征提取模块。将视频平均分为 N 个片段,并从每个分段中随机采样一张图片。这种采样策略可以确保采样结果覆盖整个视频的时间跨度,并且不受视频时间长度的影响。得益于深度学习技术的发展,尤其是卷积神经网络的发展,该模块可以选用任意的卷积神经网络作为特征提取网络[21-24],本文选用 BN-Inception[23]网络,该网络可以在识别准确率和效率之间取得很好的平衡。BN-Inception 网络主要由卷积层、池化层、Bath Normalization(BN)层和全连接层组成。文中移除 BN-Inception 网络最后的全连接层,选择 BN-Inception 全连接层的前一层——全局池化层的输出作为行为关联模块的输入,该层输出的特征维度为 1 024。

(2)行为关联模块。该模块由 LSTM 层、Bath Normalization(BN)层和 Dropout 层组成。本文将特征提取模块的输出输入到共有 512 个隐藏单元的 LSTM 层对行为变化信息进行关联。由于 LSTM 层的最后一个时间步学习到了更多的行为变化信息,文中选择 LSTM 层每个单元的最后一个时间步的输出结果作为行为模式的表达。文中尝试平均所有时间步的输出和采用较大的时间步的输出作为输出特征,其识别效果均低于使用最后一个时间步输出的识别准确率。然后,将 LSTM 层的输出输入 BN 层进行归一化处理。为了缓解过拟合的问题,本文将 BN 层的输出经过 Dropout 层之后再输入到全连接层,并根据采样帧的数目来动态调节丢弃率。与文献[18]的结论一致,本文在增加 LSTM 的层数以及隐藏单元的数目时,模型出现严重的过拟合。

4　实验和评估

首先在两个通用数据集(UCF-101[1]数据集和 HMDB-51[2]数据集)上进行实验,并使用这两个数据集的第一个分组来验证 ARN 的性能。ARN 在 UCF-101 数据集上取得了略高于 TSN 的准确率,在 HMDB-51 数据集上取得了远高于 TSN 的准确率,并取得了高于同类模型的识别准确率,如表 4 所列。为了进一步验证 ARN 对持续时间长的复杂行为的建模能力,本文在一个大规模且存在许多复杂行为(如 Pretending to put something behind something 和 Pretending to put something into something)的数据集 Something-Something V1[3]上进行了实验。这类行为具有很强的迷惑性,需要联系行为的上下文才可以做出准确的识别。

4.1 实验的训练和测试过程

在训练阶段，为了对比 ARN 和 TSN，本文保留了大部分 TSN 的训练策略，包括使用数据扩充技巧。本文将视频平均分为 N 个片段，每个片段内随机采样一帧图片。最后根据采样得到的图片数目(K)调节 dropout 层丢失率的大小。本文使用带有动量的小批量梯度下降算法来训练文中的网络，并设置模型最初的学习率为 0.001，batch 大小为 64。本文在连续训练 5 个 epochs 时，将学习率减小为原来的 1/10。

在测试阶段，与 TSN 模型对每个视频采样 25 帧不同的是，本文对每个视频仅采样与训练阶段相同的帧数(3～7 帧)，并且分别测试了是否使用数据增强(crop=1 或 crop=10)对模型的影响。结果证明，在测试阶段，ARN 在 crop=1 的情形下，其准确率已经高于 TSN，当增加 crop 与 TSN 一致(crop=10)时，ARN 的识别准确率会进一步得到提升。

4.2 TSN 实验结果

表 1 列出了 TSN 模型的空间网络和时间网络在 UCF-101 和 HMDB-51 数据集上的识别准确率。本文将 TSN 对空间网络和时间网络的识别结果加权平均，从而得到双流网络的识别结果。

表 1 TSN 实验结果

Table 1 Experiment results of TSN

(单位：%)

Modality	UCF-101	HMDB-51
RGB	86.0	53.5
Flow	88.1	66.5
RGB+Flow	94.0	69.0

4.3 ARN 在输入为 RGB 特征时的表现

本文使用 RGB 帧训练 ARN 模型，结果表明 ARN 可以从更完整的视频信息中获益，而 TSN 的空间网络在使用更多帧训练之后，准确率不再上升，甚至会略微下降。本文分别测试了使用 3,5,7,9 帧的情况下 TSN，ARN，ARN Full 的准确率变化情况。当帧数大于 7 帧时，它们的准确率变化较小，原因是 UCF-101 和 HMDB-51 数据集中的视频是已经裁剪好的视频片段，对其过多的分段会采样到冗余帧。相比 TSN 的空间网络，在 UCF-101 数据集上 ARN 的识别准确率由 86.0%提升到了 86.45%，在 HMDB 数据集上其由 53.53%提升到了 57.39%。由于对视频采样的帧数不同，行为识别的粗细粒度也不同，因此将采样不同帧数(3,5,7)的得分进行融合，得到 Multi-Scale ARN。实验结果表明，融合之后 ARN 在 HMDB-51 数据集上的识别准确率会得到进一步提升。表 2 详细对比了在 HMDB-51 数据集上，当使用不同数目输入帧时，TSN，ARN 以及 ARN Full 识别准确率的变化情况。

表 2 TSN，ARN 和 ARN Full 的识别准确率对比

Table 2 Comparison of accuracy of TSN, ARN and ARN Full

(单位：%)

Frames	TSN	ARN	ARN Full
3	52.35	52.81	55.10
5	53.53	54.71	55.23
7	53.14	54.77	57.39
Multi-Scale	**54.05**	**57.25**	**59.47**

注：Multi-Scale 表示对不同帧数(3,5,7)的识别结果按照 1∶2∶3 的进行比例融合

4.4 融合光流特征后 ARN 的表现

本文使用稀疏采样策略对视频帧进行采样，得到的 RGB 帧的跨度较远，ARN 无法发掘非常短时间内潜在的行为模式变化，而光流特征能很好地弥补这个缺陷。因此，文中将 ARN 的得分与光流得分进行融合，其后在 HMDB-51 数据集上的识别准确率高于其他模型，如表 3 所列。注意，此时本文仅使用了 TSN 模型的时间网络在光流特征上训练的识别结果，并未将行为关联模块插入 TSN 的时间网络。

表 3 ARN 在输入为 RGB 和光流特征时的表现

Table 3 ARN performance with RGB and optical flow

Modality	UCF-101	HMDB-51
Multi-Scale/%	**94.5**	**73.33**

4.5 ARN 和其他行为识别方法的对比

表 4 列出了在 UCF-101 数据集和 HMDB-51 数据集上 ARN 和现有的行为识别方法的准确率的详细对比。ARN 在 UCF-101 数据集上取得了与其他方法相似的结果，因为 UCF-101 数据集中的运动信息变化较小，关联行为上下文信息对于提升行为的识别准确率作用不明显。由于运动信息对于 HMDB-51 数据集尤为重要，ARN 在 HMDB-51 数据集上取得了高于其他模型的识别准确率。注意，I3D[9] 和 ECO[17] 使用 Kinects[4] 数据集对模型进行预训练，它们在 UCF-101 数据集上的识别准确率高于 ARN。Kinects 是一个超大规模数据集，使用 Kinects 预训练可以有效地减轻模型在小规模数据集[1-2]上训练的过拟合问题，从而提升模型的识别准确率，但是其在训练阶段需要消耗非常多的硬件资源，因此本文仅使用了 ImageNet 预训练的模型。同时，在 ECO 使用 Kinects 数据集预训练的前提下，ARN 在 HMDB-51 数据集上的表现仍然好于 ECO，从而进一步证明了 ARN 对完整行为的建模能力。

表 4 在 UCF-101 数据集和 HMDB-51 数据集上行为识别方法的对比

Table 4 Action recognition comparison on UCF101[1] and HMDB51[2] datasets

(单位：%)

Method	UCF-101	HMDB-51
Two-Stream[6]	88.0	59.40
LSTM[8]	88.6	—
TSN[17]	**94.0**	**68.50**
TS-LSTM[18]	**94.1**	**69.00**
I3D(pre-training on Kinects[9]	**95.6**	**74.80**
I3D(pre-training on ImageNet)[9]	84.5	49.80
ECO(pre-training on Kinects)[17]	94.8	72.40
ARN(pre-training on ImageNet)	**94.5**	**73.33**

4.6 ARN 在 Something-Something V1 数据集上的表现

为了进一步分析 ARN 对于复杂行为的建模能力，本文对比了 ARN 模型与 TSN 模型在 something-something V1 数据集上的表现。该数据集包含大量需要联系行为上下文才可以做出准确识别的行为。表 5 列出了在 Something-Something V1 数据集上 TSN 和 ARN 的识别准确率随着输入帧数目变化的情况。结果表明，ARN 在这类包含复杂行为的数据集上的表现好于 TSN 10 个百分点左右，从而证明了关联行

为在时间维度上的信息对识别复杂行为的重要性。

表 5　TSN 和 ARN 在 Something-Something V1 数据集上识别准确率随输入视频帧数目的变化情况

Table 5　Chang of recognition accuracy of TRN and TSN as number of frames(fr.) varies on Something-Something V1

(单位:%)

Frames	TSN	ARN	ARN Full
3	16.30	20.25	21.24
5	16.88	24.44	24.88
7	17.34	25.87	26.42
Multi-Scale	**18.63**	**27.52**	**28.12**

4.7　TSN 和 ARN 对某些行为类别的识别准确率

当使用更多的视频帧训练 TSN 时,TSN 的准确率不再提升,如表 2 所列。该结果说明 TSN 的识别准确率并不能受益于更多的视频信息,原因是 TSN 对分段得分进行融合的策略过于简单,无法发掘行为在分段间的联系。随着视频分段的增多,复杂的背景信息容易使模型过拟合。本文通过分析 TSN 在 HMD-51 数据集的每个行为类别的识别准确率发现,随着视频分段的增加,有 14 个类别的分类准确率会下降,表 6 列出了部分类别识别准确率随采样帧数目的变化情况。然而,ARN 随着视频分段的增加,其识别效果会越来越好。同时,那些识别准确率下降的行为,在使用 ARN 对分段行为信息进行关联之后,它们的识别准确率会得到很大提升。表 7 列出了在使用 7 帧输入帧的情况下,ARN 和 TSN 对 HMDB-51 数据集一些行为类别的识别准确率。

表 6　随着输入帧数的增加,TSN 在 HMDB-51 数据集上部分行为类别识别准确率的变化情况

Table 6　With increase of number of input frames, variation of TSN's recognition accuracy of some behavior categories on HMDB-51 data set

(单位:%)

Label	3 Frames	7 Frames	变化量
chew	60	46	−14
fencing	50	40	−10
flic_flac	43	40	−3
jump	26	16	−10
turn	6	0	−6

表 7　在使用 7 帧 RGB 帧的情况下,TSN 和 ARN 对 HMDB-51 数据集的部分行为的识别准确率

Table 7　In the case of 7 RGB frames, recognition accuracy of some actions on HMDB-51 dataset by TSN and ARN

(单位:%)

Label	TSN	ARN	变化量
chew	46	66	+20
fencing	40	70	+30
flic_flac	40	66	+26
jump	16	43	+27
turn	0	26	+26

ARN 可以对复杂的行为进行建模,由表 2、表 3 可知,相比 TSN,ARN 的总体表现好于 TSN。但是对于一些行为类别,ARN 的识别准确率低于 TSN。这可能是 CNN+LSTM 类方法[8,20]在 UCF-101 和 HMDB-51 数据集上表现不好的原因。本文提出的 ARN Full 由于可以同时兼顾视频的外观信息和行为在视频分段间的关联信息,因此表现最好。最终 ARN Full 在 HMDB-51 数据集上的识别结果要远远好于之前提出的 CNN+LSTM 类方法,如表 4 所列。

4.8　实验结果及分析

本文使用 3 个通用行为识别数据集评估了 ARN。由于 UCF-101 数据集中 52%的行为在使用单帧图片的情况下就能以高于 90%的准确率被准确识别,行为关联模块对这类行为的识别准确率的提升较小。最终 ARN 在 UCF-101 上取得了 94.5%的准确率,该结果略高于 TSN(94.0%)。数据集 HMDB-51 和 Something-Something V1 中 90%以上的行为类别需要详细的运动信息才能被准确识别,通过行为关联的方式可以有效地提升这些行为的识别准确率。ARN 在 HMDB-51 和 Something-Something V1 数据集上分别取得了 73.33%和 28.12%的准确率,明显高于 TSN 的识别准确率(68.5%和 18.63%),同时 ARN 在 HMDB-51 数据集上的识别准确率好于其他方法。本文对比了模型利用更多行为信息时行为识别准确率的变化情况。随着模型可利用的行为信息增多,ARN 的识别准确率明显好于 TSN。最后本文分析了随着 ARN 和 TSN 利用更多的行为信息,两者对 HMDB-51 数据集中单个行为类别识别准确率的变化情况,实验结果表明,随着模型可利用的行为信息增多,采用静态融合方式的 TSN 表现较差,而采用行为关联模块的 ARN 的识别准确率会有较大的提升。以上实验结果充分说明,相比 TSN 的静态融合策略,ARN 通过关联行为信息的方式可以提升对行为建模的能力,从而可以更好地利用完整的行为信息。

结束语　对于前后差异较大的行为,只识别完整行为中的一部分行为信息会得到前后不一致的识别结果,这类行为需要考虑完整的行为信息才能被准确识别。然而,对视频中行为信息的静态融合方式无法发掘行为的变化信息。本文提出了新的行为识别模型 Action-Relation Network(ARN),其可以利用 LSTM 关联视频中行为的起始和演变过程,从而更有效地利用完整的视频信息。在多个数据集上的实验结果证明了 ARN 可以准确识别前后差异较大的复杂行为。同时,本文提出的 ARN Full 由于可以兼顾视频的动态信息和静态的外观信息,显著提升了 CNN+LSTM 类行为识别方法的识别准确率。

本文的研究工作的不足之处在于未引入注意力机制。注意力机制在序列问题中已经被广泛采用,并已被证明可以有效提升算法的表现。下一步工作将结合注意力机制,着重发掘视频中行为变化较明显的部分,以进一步提升复杂行为的识别准确率。

参 考 文 献

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

[2] KUEHNE H,JHUANG H,GARROTE E,et al. HMDB:a large video database for human motion recognition[C]// 2011 Interna-

tional Conference on Computer Vision. IEEE, 2011: 2556-2563.
[3] GOYAL R, KAHOU S E, MICHALSKI V, et al. The Something Something Video Database for Learning and Evaluating Visual Common Sense[C]// ICCV. 2017, 1(2): 3.
[4] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv: 1705.06950, 2017.
[5] RUSSAKOVSKY O, DENG J, SUH, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[6] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Advances in neural information processing systems. 2014: 568-576.
[7] DU T, BOURDEV L D, FERGUS R, et al. C3d: generic features for video analysis[J]. Eprint Arxiv, 2014, 2(8).
[8] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.
[9] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
[10] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.
[11] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 305-321.
[12] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.
[13] CRASTO N, WEINZAEPFEL P, ALAHARI K, et al. MARS: Motion-Augmented RGB Stream for Action Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7882-7891.
[14] SUN S, KUANG Z, SHENGL, et al. Optical flow guided feature: A fast and robust motion representation for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1390-1399.
[15] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016: 20-36.
[16] ZHOU B L, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos[C]// Proceedings of the European Confe-rence on Computer Vision (ECCV). 2018: 803-818.
[17] ZOLFAGHARI M, SINGH K, BROX T. Eco: Efficient convolutional network for online video understanding[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 695-712.
[18] MA C Y, CHEN M H, KIRA Z, et al. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition[J]. Signal Processing: Image Communication, 2019, 71: 76-87.
[19] CHEN Q, ZHU X, LING Z, et al. Enhanced lstm for natural language inference[J]. arXiv: 1609.06038, 2016.
[20] GAO L, GUO Z, ZHANG H, et al. Video captioning with attention-based LSTM and semantic consistency[J]. IEEE Transactions on Multimedia, 2017, 19(9): 2045-2055.
[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.
[22] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv: 1502.03167, 2015.
[23] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[24] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

HE Xin, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include deep learning and action recognition.

XU Juan, born in 1981, associate professor, is a member of China Computer Federation. Her main interests include quantum computing and quantum information, cloud computing and deep learning.