5.算法实验结果对比

5.1 常用数据集介绍

一个视频内容理解算法的好坏与否需要与其他的同类算法在同一个环境中进行比较，于是产生了公开数据库。随着该领域大量的新方法、新模型的提出，越来越多的大规模的数据集也诞生了，这对研究视频内容理解方法起到了非常重要的推动作用。

现在根据视频理解下面的较小方向可将其分为4类数据集，分别是视频分类/动作识别、视频描述、时序行为检测数据集和其他视频理解数据集。表1-4列出了这四类数据集中的一些常用数据集的详细介绍，包括发表年份、动作数、简介、时长、目前最高识辨率等信息。比如在视频分类/动作识别中，主要有剪辑过的视频和未剪辑过的视频两种数据集。剪辑过的视频时间较短，是一段明确的动作，比如HMDB-51、UCF-101等，而未剪辑过的视频一般时间较长，而且其中包含很多冗余信息，比如Charades、ActivityNet (v1.3)等。

根据近些年数据集的发展趋势表示，数据集里包含越来越多的复杂的视频场景，让其更接近实际情况，这样在不受控制的自然情况下使视频内容理解更具有实用的价值。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据库名称 | 发布年份 | 动作数 | 简介 | 目前最高识辨率 |
| KTH[57] | 2004 | 6 | 包含2391个视频，背景单一、相机固定，是当时最大的人体动作数据库 | 98.83% |
| Olympic Sports[58] | 2010 | 16 | 源于YouTube，包含了远动员的各种运动视频，有16个运动类别 | 96.6% |
| HMDB51[59] | 2011 | 51 | 源于YouTube、谷歌视频，包括6766个剪辑视频，有人的面部、肢体动作和与物体交互的动作几类 | 80.7% (DeepMind I3D) |
| UCF101[60] | 2012 | 101 | 源于YouTube，包括13320个剪辑视频，包含五大类动作，分别是理发、体育运动、化妆刷牙、弹奏乐器、爬行 | 98%  (DeepMind I3D) |
| Sports-1M[61] | 2014 | 487 | 包含110万个视频，内容为各种体育运动 | 75.9% |
| ActivityNet v1.3[62] | 2015 | 200 | 包含19994个未剪辑视频，其中10024个用于训练，4926个用于验证，5044个用于测试。内容包括运动、饮食和家庭活动等 | 8.83% err (iBUG) |
| Charades[63] | 2016 | 157 | 每个视频大约有30秒，包含9848个未剪辑视频。视频中含有标记、每个动作的开始时间和结束时间 | 0.3441 mAP (DeepMind I3D) |
| YouTube-8M[64] | 2016 | 4716 | 包含800万个视频片段，是一个多标签分类的数据集 | 85% GAP (WILLOW) |
| Kinetics[65] | 2017 | 600 | 源于YouTube，包含50万个视频片段，每个视频大约10秒，包含多种多样的人类行为，是一个大规模的数据集 | Top-1:83.5%  Top-5: 96.8% |
| Moments in Time[66] | 2018 | 339 | 包含100万个视频片段，内容为人、动物、自然现象和物体等，时长3秒 | Top-1:38.1%,  Top-5:65.3% |
| HACS[67] | 2019 | 200 | 由麻省理工学院和Facebook联合创造，仅需8.8秒就能标记一个视频剪辑 | - |

表5-1 视频分类/动作识别常用数据集详细介绍

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据库名称 | 发布年份 | 视频片段数 | 简介 | 目前最高识辨率 |
| THUMOS 2014[78] | 2014 | 9.682 | 该数据集是在真实环境下的开源视频，包括吹干头发、刷牙等日常动作和驾驶、高尔夫挥杆等体育动作。判断检测到每个动作的开始和结束时间 | - |
| ActivityNet(v1.3)[62] | 2015 | ~20,000 | 包含200种不同的日常活动，比如遛狗、跳远等，约10000个视频用作训练集，验证集和测试集则各包含约5000个视频 | 0.344(SJTU &Columbia ) |
| MEXaction2[79] | 2015 | 1975 | 由YouTube、UCF101中的骑马视频以及INA视频三个部分组成，包含骑马和斗牛两类动作，其中未剪辑的视频长度很长 | - |
| Broad Video Highlights | 2017 | 18000 | 由百度发布，源于爱奇艺视频中的综艺节目，包括1500个长视频，提取视频中的精彩片段 | - |

表5-2 时序行为检测常用数据集详细介绍

|  |  |  |  |
| --- | --- | --- | --- |
| 数据库名称 | 发布年份 | 背景 | 简介 |
| MPII-MD[81] | 2015 | 电影 | 包含来自68HD电影的超过68K句子和视频片段的并行语料库 |
| MSR-VTT[82] | 2016 | 20个类别 | 从商业视频搜索引擎中收集了257个热门查询，其中每个查询包含118个视频，提供了总计41.2小时的10K网络视频剪辑和200K剪辑句对 |

表5-3 视频描述常用数据集详细介绍

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据库名称 | 发布年份 | 方向 | 简介 | 目前最高识辨率 |
| Jester[83] | 2001 | 视频中的手势 | 包含了150个笑话，使用-10到10的连续的等级，在量级别上有最高评分密度 | 95.34%(Ke Yang, NUDT\_PDL) |
| MovieQA[84] | 2016 | 视频问答 | 包含视频片段、脚本、字幕、DVS和情节等多种信息源，是一个大规模的关于电影的问答集 | - |
| AVA[85] | 2017 | 时空局部  原子视觉动作 | 由Google发布，注解了80种时空局部化原子动作，比如走、握手、踢等，包含5.76万个视频片段、21万个动作标签和9.6万个标记动作执行人 | - |

表5-4 视频内容理解其他方向常用数据集详细介绍

5.2 算法比较与分析

5.2.1视频分裂/动作识别方向的算法性能比较

该部分主要在HMDB51和UCF101数据集上对视频分类/动作识别方向的算法进行了性能比较，根据上文中介绍的视频内容理解分类方法，表1对基于手工提取特征的iDT算法，和基于深度学习的几种算法的性能进行了汇总，如表1所示。由表中的结果来看，虽然一开始传统方法的识别准确率比基于深度学习的要高，但是随着近些年深度学习的发展，UCF101的数据集准确率达到了98%，HMDB51数据集的准确率达到了80.9%。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参考文献 | 方法 | 预训练数据集 | UCF-101 | HMDB51 |
| [6] | iDT | — | 85.9 | 57.2 |
| [7] | CNN+LSTM | Sports-1M | 63.3 | — |
| [8] | Two-Stream+LSTM | ImageNet | 88.6 | — |
| [10] | AdaScan | ImageNet | 89.4 | 54.9 |
| AdaScan+iDT | 91.3 | 61.0 |
| AdaScan+iDT+C3D | 93.2 | 66.0 |
| [35] | C3D+LSTM | — | 92.9 | 70.1 |
| [16] | FstCN | ImageNet | 88.1 | 59.1 |
| [17] | Two-Stream(VGG-M) | ImageNet | 88.0 | 59.4 |
| [1] | TSN | ImageNet | 94.2 | 69.4 |
| [2] | Two-Stream fusion  （VGG-16） | ImageNet | 92.5 | 65.4 |
| [3] | ST-ResNet\*+IDT | ImageNet | 94.6 | 70.3 |
| [4] | TDD+iDT | ImageNet | 91.5 | 65.9 |

表5-5 视频分裂/动作识别方向的算法在UCF-101和HMDB51上的性能比较

[6] H. Wang and C. Schmid. Action recognition with improved trajectories. ICCV'13.

[7] A. Karpathy, et al. Large-scale video classification with convolutional neural networks. CVPR'14.

[8] J. Y.-H. Ng, et al. Beyond short snippets: Deep networks for video classification. CVPR'15.

[10] A. Kar, et al. AdaScan: Adaptive scan pooling in deep convolutional  neural networks for human action recognition in videos. CVPR'17.

[16] L. Sun, et al. Human action recognition using factorized spatio-temporal convolutional networks. ICCV'15.

[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. NIPS'14.

[35] ARIF S, WANG J, HASSAN U T, et al. 3D-CNN-based fused feature maps with LSTM applied to action recognition[J]. Future Internet, 2019, 11(2): 42.

1. Wang, Limin, et al. "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition.." abs/1608.00859 (2016).
2. Feichtenhofer, Christoph, et al. "Convolutional Two-Stream Network Fusion for Video Action Recognition.." abs/1604.06573 (2016).
3. Feichtenhofer, Christoph, et al. "Spatiotemporal Residual Networks for Video Action Recognition.." abs/1611.02155 (2016).
4. Wang, Limin, et al. "Action recognition with trajectory-pooled deep-convolutional descriptors." abs/1505.04868 (2015).

5.2.2时序行为检测算法性能比较

在IoU值确定的情况下，mAP的值是时序行为检测的常用评价指标，mAP值越大，算法准确率越高。其中动作检测中loU(tIoU)是对时间(维度是1)的IOU。MAP为所有类别的平均精度求和与所有类别的视频数目(数据集大小)的比。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| tIoU | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| Karaman et al.[36] | - | - | 0.5 | 0.3 | 0.2 |
| Wang et al.[37] | 18.2 | 17 | 14 | 11.7 | 8.3 |
| Heilbron et al.[38] | - | - | - | - | 13.5 |
| Escorcia et al.[39] | - | - | - | - | 13.9 |
| Oneata et al.[40] | 36.6 | 33.6 | 27 | 20.8 | 14.4 |
| Richard and Gall[41] | 39.7 | 35.7 | 30 | 23.2 | 15.2 |
| Yeung et al.[42] | 48.9 | 44 | 36 | 26.4 | 17.1 |
| Yuan et al.[43] | 51.4 | 42.6 | 33.6 | 26.1 | 18.8 |
| S-CNN[22] | 47.7 | 43.5 | 36.3 | 28.7 | 19 |
| CDC[23] | - | - | 40.1 | 29.4 | 23.3 |
| TAG[24] | 64.1 | 57.7 | 48.7 | 39.8 | 28.2 |
| TURN[25] | 54 | 50.9 | 44.1 | 34.9 | 25.6 |

表5-6为不同时序检测方法在THUMOS2014数据集上不同tIOU时，mAR值的比较。

5.2.3视频描述算法性能比较

在 VTT (Video to Text) 任务中，需要衡量模型输出的视频描述语句的准确性，因为衡量的对象是一个个的自然语言文本，所以通常会选择自然语言处理领域的相关评价指标。这些指标原先都是用来度量机器翻译结果质量的，并且被证明可以很好的反映待评测语句的准确性，主要包含4种： METEOR[44]，BLEU[45]，ROUGE[46]，CIDEr[47]。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | METEOR | BULE@4 | ROUGE-L | CIDEr |
| Mean-Pooling[48] | 23.7 | 30.4 | 52 | 35 |
| Soft-Attention[49] | 25 | 28.5 | 53.3 | 37.1 |
| S2VT[50] | 25.7 | 31.4 | 55.9 | 35.2 |
| ruc-uva[51] | 27.5 | 39.4 | 60 | 48 |
| VideoLAB[52] | 27.7 | 39.5 | 61 | 44.2 |
| AaIto[53] | 27.7 | 41.1 | 59.6 | 46.4 |
| v2t\_navigator[54] | 29 | 43.7 | 61.4 | 45.7 |
| Multi-Task [55] | 28.8 | 40.8 | 60.2 | 47.1 |
| MIMLL [56] | 29.4 | 44.2 | 62.6 | 50.5 |

表5-7 MRS-VTT为训练集，视频描述的算法比较