# Video Action Dataset Catalog, 2004-2021

Matthew S. Hutchinson (hutchinson@alum.mit.edu)

August 21, 2021

| | | | Action | | Actors | | Annotations | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Video Dataset | Year | Cited | Classes | Instances | H | N | C | T | S | Theme/Purpose |
| KTH [107] | 2004 | 4,246 | 6 | 2,391 | ✓ | | ✓ | | | B/W, static background |
| CAVIAR [31] | 2004 | 49 | 9 | >28 | ✓ | | ✓ | | ✓ | surveillance |
| Weizmann [7] | 2005 | 2,068 | 10 | 90 | ✓ | | ✓ | | | human motions |
| ViSOR [123, 122] | 2005 | 47 | n/a | n/a | ✓ | | ✓ | | | surveillance |
| ETISEO [119, 86] | 2005 | 183 | 15 | n/a | ✓ | | ✓ | | | human motions |
| IXMAS [129] | 2006 | 977 | 13 | 390 | ✓ | | ✓ | | | B/W, partial occlusion |
| UCF Aerial [32] | 2007 | n/a | 9 | n/a | ✓ | | ✓ | | ✓ | aerial-view |
| CASIA Action [127, 143] | 2007 | 242 | 15 | 1,446 | ✓ | | ✓ | | | multi-view, outdoors |
| Coffee & Cigarettes [61] | 2007 | 526 | 2 | 246 | ✓ | | ✓ | ✓ | ✓ | movies and TV |
| UIUC Action [120] | 2008 | 378 | 14 | 532 | ✓ | | ✓ | | | action repetition |
| UCF Sports [102] | 2008 | 1,269 | 10 | 150 | ✓ | | ✓ | | ✓ | sports |
| UCF ARG [33] | 2008 | n/a | 10 | 480 | ✓ | | ✓ | | ✓ | multi-view, aerial-view |
| Hollywood (HOHA) [60] | 2008 | 3,727 | 8 | n/a | ✓ | | ✓ | | | movies |
| Cambridge-Gesture [52] | 2008 | 298 | 9 | 900 | ✓ | | ✓ | | | gestures |
| BEHAVE [10] | 2009 | 134 | 10 | 163 | ✓ | | ✓ | | ✓ | human-human interaction |
| URADL [78] | 2009 | 574 | 10 | 150 | ✓ | | ✓ | | | daily activities |
| UCF11 [72] | 2009 | 1,183 | 11 | 3,040 | ✓ | | ✓ | | | web videos |
| MSR-I [140] | 2009 | 181 | 3 | n/a | ✓ | | ✓ | | ✓ | activities |
| i3DPost MuHAVi [35] | 2009 | 179 | 12 | >1,000 | ✓ | | ✓ | | | multi-view, studio |
| Hollywood2 [76] | 2009 | 1,488 | 12 | 3,669 | ✓ | | ✓ | | | movies |
| Collective Activity [132] | 2009 | 201 | 5 | 44 | ✓ | | ✓ | | ✓ | group activities |
| LabelMe [43] | 2009 | 203 | 70 | >300 | ✓ | ✓ | ✓ | ✓ | ✓ | actor-object interactions |
| Keck Gesture [147] | 2009 | 349 | 14 | 126 | ✓ | | ✓ | | | gestures |
| DLSPB [27] | 2009 | 307 | 2 | 89 | ✓ | | ✓ | ✓ | | movies and TV |
| Hollywood-Localization [53] | 2010 | 159 | 2 | 408 | ✓ | | ✓ | | ✓ | movies |
| VideoWeb [26] | 2010 | 55 | 51 | 368 | ✓ | | ✓ | | | multi-view, tasks |
| UT-Tower [16, 17] | 2010 | 61 | 9 | 648 | ✓ | | ✓ | | ✓ | aerial-view, human motions |
| UT-Interaction [104, 105] | 2010 | 593 | 6 | 60 | ✓ | | ✓ | | | human-human interaction |
| UCF50 [101] | 2010 | 537 | 50 | >5,000 | ✓ | | ✓ | | | web videos, expand UCF11 |
| TV-Human Interaction [92] | 2010 | 176 | 4 | 300 | ✓ | | ✓ | | ✓ | TV, human-human interaction |
| Olympic Sports [144] | 2010 | 745 | 16 | 800 | ✓ | | ✓ | | | sports |
| MSR-II [12] | 2010 | 232 | 3 | n/a | ✓ | | ✓ | | ✓ | activities |
| MSR-Action3D [64] | 2010 | 1,285 | 20 | 4,020 | ✓ | | ✓ | | | RGB-D, gestures and motions |
| CMU MoCap [1] | 2010 | n/a | n/a | 2,605 | ✓ | | ✓ | | | RGB-D, human motions |
| VIRAT [90] | 2011 | 634 | 23 | ∼10,000 | ✓ | | ✓ | ✓ | ✓ | surveillance, aerial-view |
| HMDB51 [57] | 2011 | 2,428 | 51 | ∼7,000 | ✓ | | ✓ | | | human motions |
| CAD-60 [118] | 2011 | 549 | 12 | 60 | ✓ | | ✓ | | ✓ | RGB-D, daily activities |
| GTEA [30, 67] | 2011 | 492 | 71 | 526 | ✓ | | ✓ | | | egocentric, kitchen |
| CCV [49] | 2011 | 288 | *20 | 9,317 | ✓ | | ✓ | | | web videos |
| ChaLearn [2] | 2011 | n/a | 86 | 50,000 | ✓ | | ✓ | | | RGB-D, gestures and motions |
| RGBD-HuDaAct [88] | 2011 | 393 | 12 | 1,189 | ✓ | | ✓ | | | RGB-D, daily activities |
| NATOPS [113] | 2011 | 111 | 24 | 400 | ✓ | | ✓ | | | aircraft hand signaling |
| GTEA Gaze [29] | 2012 | 331 | 40 | 331 | ✓ | | ✓ | ✓ | | egocentric, kitchen |
| GTEA Gaze+ [29, 67] | 2012 | 165 | 44 | 1,958 | ✓ | | ✓ | ✓ | | egocentric, kitchen |
| BIT-Interaction [55] | 2012 | 109 | 8 | 400 | ✓ | | ✓ | | | human-human interaction |
| LIRIS [131] | 2012 | 60 | 10 | n/a | ✓ | | ✓ | | ✓ | RGB-D, office environment |
| MSR-DailyActivity3D [124] | 2012 | 1,339 | 16 | 320 | ✓ | | ✓ | | | RGB-D, gestures |
| UCF101 [114] | 2012 | 3,183 | 101 | 13,320 | ✓ | | ✓ | | | web videos, expand UCF50 |
| UTKinect-A [134] | 2012 | 1,216 | 10 | 200 | ✓ | | ✓ | | | RGB-D, indoors |
| MSR-Gesture3D [59] | 2012 | 317 | 12 | n/a | ✓ | | ✓ | | | RGB-D, gestures |
| ASLAN [54] | 2012 | 106 | 432 | 3,631 | ✓ | | ✓ | | | web videos, action similarity |
| ADL [95] | 2012 | 619 | 18 | ∼1,200 | ✓ | | ✓ | | ✓ | egocentric, daily activities |
| ACT4$^2$ [18] | 2012 | 122 | 14 | 6,844 | ✓ | | ✓ | | | RGB-D, multi-view |
| SBU-Kinect-Interaction [141] | 2012 | 339 | 8 | ∼170 | ✓ | | ✓ | ✓ | | RGB-D, human-human inter. |
| MPII-Cooking [103] | 2012 | 436 | 65 | 5,609 | ✓ | | ✓ | ✓ | | kitchen, fine-grained actions |
| Osaka Kinect [75] | 2012 | 31 | 10 | 80 | ✓ | | ✓ | | | RGB-D, gestures |

*Only a portion of classes are actions. Some are objects or visual tags.

| Video Dataset | Year | Cited | Action | | Actors | | Annotations | | | Theme/Purpose |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Classes | Instances | H | N | C | T | S | |
| DHA [70] | 2012 | 66 | 23 | 483 | ✓ | | ✓ | | | RGB-D, gestures and motions |
| Falling Event [142] | 2012 | 138 | 5 | 200 | ✓ | | ✓ | | | RGB-D, daily activities |
| G3D [8, 9] | 2012 | 207 | 20 | 659 | ✓ | | ✓ | ✓ | | RGB-D, gaming gestures |
| MSR-3DActionPairs [91] | 2013 | 866 | 12 | 360 | ✓ | | ✓ | | | RGB-D, gestures |
| Multiview 3D Event [128] | 2013 | 119 | 8 | 3,815 | ✓ | | ✓ | | | RGB-D, multi-view |
| RGBD-SAR [137] | 2013 | 29 | 12 | 810 | ✓ | | ✓ | | | RGB-D, monitoring seniors |
| CAD-120 [56] | 2013 | 587 | 10 | 120 | ✓ | | ✓ | | ✓ | RGB-D, daily activities |
| JPL Interaction [106] | 2013 | 253 | 7 | ∼85 | ✓ | | ✓ | ✓ | | egocentric, human-human inter. |
| MHAD [89] | 2013 | 336 | 11 | ∼650 | ✓ | | ✓ | | | RGB-D, multi-view, gestures |
| Florence3D [79, 108] | 2013 | 189 | 9 | 213 | ✓ | | ✓ | | | RGB-D , gestures |
| THUMOS'13 [47, 42, 114] | 2013 | 191 | *101 | 13,320 | ✓ | | ✓ | | ✓ | web videos, extend UCF101 |
| J-HMDB-21 [44] | 2013 | 567 | 51 | 928 | ✓ | | ✓ | | ✓ | re-annotate HMDB51 subset |
| Mivia [13] | 2013 | 21 | 7 | 490 | ✓ | | ✓ | | | RGB-D, daily activities |
| IAS-lab [82, 83] | 2013 | 31 | 15 | 540 | ✓ | | ✓ | | | RGB-D, human motions |
| WorkoutSU-10 [84] | 2013 | 66 | 10 | 1,200 | ✓ | | ✓ | | | RGB-D, group activities |
| 50Salads [115] | 2013 | 177 | 17 | 966 | ✓ | | ✓ | ✓ | | RGB-D, kitchen |
| UWA3D-I [96] | 2014 | 141 | 30 | ∼900 | ✓ | | ✓ | | | RGB-D, multi-view |
| MANIAC [4] | 2014 | 43 | 8 | 120 | ✓ | | ✓ | ✓ | | RGB-D, egocentric, manipulations |
| Breakfast Action [58] | 2014 | 203 | 48 | 11,267 | ✓ | | ✓ | ✓ | | kitchen |
| Northwester-UCLA [125] | 2014 | 222 | 10 | 1,475 | ✓ | | ✓ | | | RGB-D, multi-view |
| Sports-1M [50] | 2014 | 5,667 | 487 | 1,000,000 | ✓ | | ✓ | | | multi-label, sports |
| ORGBD (3D Online) [139] | 2014 | 136 | 7 | 336 | ✓ | | ✓ | | | RGB-D, human-object inter. |
| THUMOS'14 [48, 42] | 2014 | 146 | **101 | 15,904 | ✓ | | ✓ | ✓ | | extends THUMOS'13 |
| Office Activity [126] | 2014 | 94 | 20 | 1,180 | ✓ | | ✓ | | | RGB-D, office environment |
| Composable [69] | 2014 | 81 | 16 | 693 | ✓ | | ✓ | | | RGB-D, gestures and motions |
| CMU-MAD [40] | 2014 | 80 | 20 | 1,400 | ✓ | | ✓ | ✓ | | RGB-D, gestures and motions |
| FFPA [149] | 2015 | 48 | 5 | 591 | ✓ | | ✓ | | | egocentric, daily activities |
| TJU [71] | 2015 | 69 | 15 | 1,200 | ✓ | | ✓ | | | RGB-D, static background |
| $M^2I$ [135] | 2015 | 23 | 22 | 1,760 | ✓ | | ✓ | | | RGB-D, multi-view |
| FCVID [46] | 2015 | 219 | ***239 | 91,223 | ✓ | | ✓ | | | web videos, diverse categories |
| ActivityNet100 (v1.2) [11] | 2015 | 797 | 100 | 10,733 | ✓ | | ✓ | ✓ | | untrimmed web videos |
| THUMOS'15 [36, 42] | 2015 | 146 | **101 | 21,037 | ✓ | | ✓ | ✓ | | extends THUMOS'14 |
| MEXaction [117, 116] | 2015 | 16/3 | 2 | 1,108 | ✓ | | ✓ | ✓ | | culturally relevant actions |
| MEXaction2 [21] | 2015 | n/a | 2 | 1,975 | ✓ | | ✓ | ✓ | | culturally relevant actions |
| Watch-n-Patch [133] | 2015 | 119 | 21 | ∼2,000 | ✓ | | ✓ | ✓ | | RGB-D, daily activities |
| TVSeries [24] | 2016 | 109 | 30 | 6,231 | ✓ | | ✓ | ✓ | | TV |
| OAD [65] | 2016 | 109 | 10 | n/a | ✓ | | ✓ | ✓ | | RGB-D, daily activities |
| CONVERSE [28] | 2016 | 20 | 7 | n/a | ✓ | | ✓ | ✓ | | RGB-D, human-human inter. |
| OA [63] | 2016 | 11 | 48 | 480 | ✓ | | ✓ | | | action semantic hierarchy |
| Volleyball [41] | 2016 | 215 | 6 | 1,643 | ✓ | | ✓ | | | sports (volleyball motions) |
| UWA3D-II [97] | 2016 | 117 | 30 | 1,075 | ✓ | | ✓ | | | RGB-D, multi-view |
| ActivityNet200 (v2.3) [11] | 2016 | 1,118 | 200 | 23,064 | ✓ | | ✓ | ✓ | | untrimmed web videos |
| YouTube-8M [3] | 2016 | 607 | n/a | n/a | ✓ | | ✓ | | | multi-label |
| Charades [112] | 2016 | 343 | 157 | 66,500 | ✓ | | ✓ | ✓ | | crowd-sourced, daily activities |
| NTU RGB-D [109] | 2016 | 792 | 60 | 56,880 | ✓ | | ✓ | | | RGB-D, multi-view |
| Micro-Videos [87] | 2016 | 27 | n/a | n/a | ✓ | ✓ | ✓ | | | micro-videos (e.g., Vine, Tik-Tok) |
| JAAD [99, 100] | 2017 | 53 | n/a | 654 | ✓ | | ✓ | | ✓ | pedestrians |
| DAHLIA [121] | 2017 | 9 | 7 | 51 | ✓ | | ✓ | | | RGB-D, daily activities |
| PKU-MMD [20] | 2017 | 67 | 51 | 3,366 | ✓ | | ✓ | | | RGB-D, multi-view |
| SYSU 3DHOI [39] | 2017 | 302 | 12 | 480 | ✓ | | ✓ | | | RGB-D, human-object inter. |
| DALY [130] | 2017 | 26 | 10 | 3,600 | ✓ | | ✓ | | ✓ | daily activities |
| Okutama Action [6] | 2017 | 55 | 12 | 4,700 | ✓ | | ✓ | | ✓ | aerial view |
| Kinetics-400 [51] | 2017 | 1,380 | 400 | 306,245 | ✓ | | ✓ | | | diverse web videos |
| AVA [38] | 2017 | 404 | 80 | >392,416 | ✓ | | ✓ | | ✓ | atomic visual actions |
| Something-Something [37] | 2017 | 182 | 174 | 108,499 | ✓ | | ✓ | | | human-object interactions |
| SLAC [146] | 2017 | 19 | 200 | ∼1,750,000 | ✓ | | ✓ | ✓ | | sparse-labelled web videos |
| Moments in Time (MiT) [80] | 2017 | 212 | 339 | 836,144 | ✓ | ✓ | ✓ | | | intra-class variation, web videos |
| MultiTHUMOS [138] | 2017 | 305 | 65 | ∼16,000 | ✓ | | ✓ | ✓ | | multi-label, extends THUMOS |
| $VIENA^2$ [5] | 2018 | 7 | 25 | 15,000 | ✓ | ✓ | ✓ | ✓ | | pedestrians and vehicles |
| PRAXIS Gesture [85] | 2018 | 16 | 29 | ∼4,600 | ✓ | | ✓ | | | RGB-D, gestures |
| UAV-GESTURE [93] | 2018 | 10 | 13 | 119 | ✓ | | ✓ | | ✓ | aerial-view, gestures |
| Diving48 [68] | 2018 | 25 | 48 | 18,404 | ✓ | | ✓ | | | diving motions (sports) |
| EPIC-KITCHENS-55 [22] | 2018 | 209 | 125 | 39,594 | ✓ | | ✓ | ✓ | ✓ | egocentric, kitchen |
| YouCook2 [148] | 2018 | 96 | n/a | ∼15,400 | ✓ | | ✓ | ✓ | | web videos, kitchen |
| Kinetics-600 [14] | 2018 | 115 | 600 | 495,547 | ✓ | | ✓ | | | extends Kinetics-400 |
| VLOG [34] | 2018 | 41 | 30 | ∼122,000 | ✓ | | ✓ | ✓ | | web videos, human-object inter. |
| EGTEA Gaze+ [66] | 2018 | 94 | 106 | 10,325 | ✓ | | ✓ | ✓ | ✓ | egocentric, kitchen |
| Something-Something-v2 [74] | 2018 | 12 | 174 | 220,847 | ✓ | | ✓ | | | extends Something-Something |
| Charades-Ego [111] | 2018 | 39 | 157 | 68,536 | ✓ | | ✓ | ✓ | | egocentric, daily activities |

*Only 24 classes have spatiotemporal annotations. This subset is also known as UCF101-24.

**Only 20 classes have temporal annotations.

***Only a portion of classes are actions. Some are objects or visual tags.

| Video Dataset | Year | Cited | Action | | Actors | | Annotations | | | Theme/Purpose |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Classes | Instances | H | N | C | T | S | |
| Youtube-8M Segments [3] | 2019 | 0 | *n/a | n/a | ✓ | | ✓ | ✓ | | multi-label, extends YouTube-8M |
| Jester [77] | 2019 | 37 | 27 | 148,092 | ✓ | | ✓ | | | crowd-sourced, gestures |
| LSVV-HRI [45] | 2019 | 4 | 83 | 25,600 | ✓ | | ✓ | | | RGB-D, human-robot interaction |
| PIE [98] | 2019 | 10 | 6 | ∼1,800 | ✓ | | ✓ | | ✓ | pedestrians |
| Kinetics-700 [15] | 2019 | 96 | 700 | ∼650,000 | ✓ | | ✓ | | | extends Kinetics-600 |
| Multi-MiT [81] | 2019 | 10 | 313 | ∼1,020,000 | ✓ | ✓ | ✓ | | | multi-label, extends MiT |
| HACS Clips [145] | 2019 | 64 | 200 | ∼1,500,000 | ✓ | | ✓ | | | trimmed web videos |
| HACS Segments [145] | 2019 | 64 | 200 | ∼139,000 | ✓ | | ✓ | ✓ | | extends and improves SLAC |
| NTU RGB-D 120 [73] | 2019 | 168 | 120 | 114,480 | ✓ | | ✓ | | | extends NTU RGB-D 60 |
| EPIC-KITCHENS-100 [23] | 2020 | 8 | 97 | ∼90,000 | ✓ | | ✓ | ✓ | ✓ | extends EPIC-KITCHENS-55 |
| AVA-Kinetics [62] | 2020 | 17 | 80 | >238,000 | ✓ | | ✓ | | ✓ | adds annotations, AVA+Kinetics |
| ARID [136] | 2020 | 0 | 11 | 3,784 | ✓ | | ✓ | | | dark (low-lighting) videos |
| AViD [94] | 2020 | 5 | 887 | ∼450,000 | ✓ | ✓ | ✓ | | | diverse peoples, anonymized faces |
| FineGym [110] | 2020 | 25 | 10 | 4,883 | ✓ | | ✓ | ✓ | | gymnastics w/ sub-actions |
| TinyVIRAT [25] | 2020 | 1 | 26 | 12,829 | ✓ | | ✓ | | | low-resolution videos |
| HAA500 [19] | 2020 | 2 | 500 | ∼10,000 | ✓ | | ✓ | ✓ | | course-grained atomic actions |

*Only a portion of classes are actions. Some are objects or visual tags.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

# References

[1] Cmu graphics lab motion capture database, 2010.

[2] The chalearn gesture dataset (cgd 2011), 2011.

[3] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016.

[4] Eren Erdal Aksoy, Minija Tamosiunaite, and Florentin Wörgötter. Model-free incremental learning of the semantics of manipulation actions. *Robotics and Autonomous Systems*, 71:118 – 133, 2015. Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence.

[5] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Viena2: A driving anticipation dataset. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 449–466, Cham, 2019. Springer International Publishing.

[6] Mohammadamin Barekatain, Miquel Marti, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402 Vol. 2, 2005.

[8] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12, 2012.

[9] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. Dynamic feature selection for online action recognition. In Albert Ali Salah, Hayley Hung, Oya Aran, and Hatice Gunes, editors, *Human Behavior Understanding*, pages 64–76, Cham, 2013. Springer International Publishing.

[10] Scott Blunsden and RB Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4, 2010.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, June 2015.

[12] Liangliang Cao, Zicheng Liu, and Thomas S Huang. Cross-dataset action detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1998–2005. IEEE, 2010.

[13] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, and Mario Vento. Recognition of human actions from rgb-d videos using a reject option. In Alfredo Petrosino, Lucia Maddalena, and Pietro Pala, editors, *New Trends in Image Analysis and Processing – ICIAP 2013*, pages 436–445, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[14] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[15] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[16] C. Chen and J. K. Aggarwal. Recognizing human action from a far field of view. In *2009 Workshop on Motion and Video Computing (WMVC)*, pages 1–7, 2009.

[17] Chia-Chih Chen, M. S. Ryoo, and J. K. Aggarwal. UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html, 2010.

[18] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 52–61, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[19] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. *arXiv preprint arXiv:2009.05224*, 2020.

[20] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.

[21] Michel Crucianu. Mexaction2: action detection and localization dataset, 07 2015.

[22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, September 2018.

[23] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

[24] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 269–284, Cham, 2016. Springer International Publishing.

[25] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: Low-resolution video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7387–7394, 2021.

[26] Giovanni Denina, Bir Bhanu, Hoang Thanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. *VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication*, pages 335–347. Springer London, London, 2011.

[27] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1491–1498, 2009.

[28] M. Edwards, J. Deng, and X. Xie. From pose to activity: Surveying datasets and introducing CONVERSE. *Computer Vision and Image Understanding*, 144:73 – 105, March 2016. Special Issue on Individual and Group Activities in Video Event Analysis.

[29] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[30] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[31] Robert B. Fisher. The pets04 surveillance ground-truth data set. *Proeedings of the Sixth IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS04)*, 11, 05 2004.

[32] UCF Center for Research in Computer Vision. Ucf aerial action dataset, 2007.

[33] UCF Center for Research in Computer Vision. Ucf arg, 2008.

[34] David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4991–5000, June 2018.

[35] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3dpost multi-view and 3d human action/interaction database. In *2009 Conference for Visual Media Production*, pages 159–168, 2009.

[36] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2015.

[37] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5842–5850, Oct 2017.

[38] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, June 2018.

[39] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[40] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. Sequential max-margin event detectors. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 410–424, Cham, 2014. Springer International Publishing.

[41] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[42] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[43] Jenny Yuen, B. Russell, Ce Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1451–1458, 2009.

[44] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199, December 2013.

[45] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition, 2019.

[46] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):352–364, 2018.

[47] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2013.

[48] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. 2014.

[49] Yu-Gang Jiang, Guangnan Ye, S. Chang, Daniel Ellis, and Alexander Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. page 29, 01 2011.

[50] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, June 2014.

[51] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[52] T. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1415–1428, 2009.

[53] Alexander Kläser, Marcin Marszałek, Cordelia Schmid, and Andrew Zisserman. Human focused action localization in video. In Kiriakos N. Kutulakos, editor, *Trends and Topics in Computer Vision*, pages 219–233, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[54] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):615–621, 2012.

[55] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 300–313, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[56] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[57] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, 2011.

[58] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[59] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1975–1979, 2012.

[60] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[61] I. Laptev and P. Perez. Retrieving actions in movies. In *2007 IEEE 11th International Conference on Computer Vision*, 2007. DOI: 10.1109/ICCV.2007.4409105.

[62] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

[63] W. Li and M. Fritz. Recognition of ongoing complex activities by sequence prediction over a hierarchical label space. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[64] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14, 2010.

[65] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 203–220, Cham, 2016. Springer International Publishing.

[66] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, September 2018.

[67] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[68] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[69] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[70] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, page 1053–1056, New York, NY, USA, 2012. Association for Computing Machinery.

[71] An-An Liu, Wei-Zhi Nie, Yu-Ting Su, Li Ma, Tong Hao, and Zhao-Xuan Yang. Coupled hidden conditional random fields for rgb-d human action recognition. *Signal Processing*, 112:74 – 82, 2015. Signal Processing and Learning Methods for 3D Semantic Analysis.

[72] J. Liu, Jiebo Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003, 2009.

[73] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Duan, and A. Kot Chichung. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2684–2701, 2019.

[74] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018.

[75] A. Mansur, Y. Makihara, and Y. Yagi. Inverse dynamics for action recognition. *IEEE Transactions on Cybernetics*, 43(4):1226–1236, 2013.

[76] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, 2009.

[77] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2874–2882, Oct 2019.

[78] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th International Conference on Computer Vision*, pages 104–111, 2009.

[79] Media Integration & Communication Center (MICC). Florence 3d actions dataset, 2013.

[80] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, 2020.

[81] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding, 2019.

[82] Matteo Munaro, Gioia Ballin, Stefano Michieletto, and Emanuele Menegatti. 3d flow estimation for human action recognition from colored point clouds. *Biologically Inspired Cognitive Architectures*, 5:42 – 51, 2013. Extended versions of selected papers from the Third Annual Meeting of the BICA Society (BICA 2012).

[83] Matteo Munaro, Stefano Michieletto, and Emanuele Menegatti. An evaluation of 3d motion flow and 3d pose estimation for human action recognition. In *RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras*, 2013.

[84] Farhood Negin, Fırat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In Mohamed Kamel and Aurélio Campilho, editors, *Image Analysis and Recognition*, pages 648–657, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[85] Farhood Negin, Pau Rodriguez, Michal Koperski, Adlen Kerboua, Jordi Gonzàlez, Jeremy Bourgeois, Emmanuelle Chapoulie, Philippe Robert, and Francois Bremond. Praxis: Towards automatic cognitive assessment using gesture recognition. *Expert Systems with Applications*, 2018.

[86] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 476–481, 2007.

[87] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. The open world of micro-videos, 2016.

[88] B. Ni, Gang Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1147–1153, 2011.

[89] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, 2013.

[90] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, 2011.

[91] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[92] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, volume 1, page 33. Citeseer, 2010.

[93] Asanka G. Perera, Yee Wei Law, and Javaan Chahl. Uav-gesture: A dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[94] AJ Piergiovanni and Michael S. Ryoo. Avid dataset: Anonymized videos from diverse countries. *arXiv preprint arXiv:2007.05515*, 2020.

[95] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2012.

[96] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 742–757, Cham, 2014. Springer International Publishing.

[97] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[98] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[99] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–213, 2017.

[100] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. It's not all about size: On the role of data properties in pedestrian detection. In *ECCVW*, 2018.

[101] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[102] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[103] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2012.

[104] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1593–1600, 2009.

[105] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[106] Michael S. Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[107] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36 Vol.3, 2004.

[108] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2013.

[109] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, June 2016.

[110] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, June 2020.

[111] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[112] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing.

[113] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *Face and Gesture 2011*, pages 500–506, 2011.

[114] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[115] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, page 729–738, New York, NY, USA, 2013. Association for Computing Machinery.

[116] A. Stoian, M. Ferecatu, J. Benois-Pineau, and M. Crucianu. Scalable action localization with kernel-space hashing. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 257–261, 2015.

[117] A. Stoian, M. Ferecatu, J. Benois-Pineau, and M. Crucianu. Fast action localization in large-scale video archives. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(10):1917–1930, 2016.

[118] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation*, pages 842–849. IEEE, 2012.

[119] A. t. NGHIEM, F. BREMOND, M. THONNAT, and R. MA. A new evaluation approach for video processing algorithms. In *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*, pages 15–15, 2007.

[120] Du Tran and Alexander Sorokin. Human activity recognition with metric learning. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 548–561, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[121] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. The daily home life activity dataset: A high semantic activity dataset for online recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 497–504, 2017.

[122] R. Vezzani and R. Cucchiara. Annotation collection and online performance evaluation for video surveillance: The visor project. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 227–234, 2008.

[123] R. Vezzani and R. Cucchiara. Visor: Video surveillance on-line repository for annotation retrieval. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1281–1284, 2008.

[124] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012.

[125] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[126] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 3d human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 97–106, New York, NY, USA, 2014. Association for Computing Machinery.

[127] Y. Wang, K. Huang, and T. Tan. Human activity recognition based on r transform. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[128] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[129] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249 – 257, 2006. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[130] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision, 2016.

[131] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14 – 30, 2014.

[132] Wongun Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1282–1289, 2009.

[133] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4362–4370, June 2015.

[134] L. Xia, C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, 2012.

[135] Ning Xu, Anan Liu, Weizhi Nie, Yongkang Wong, Fuwu Li, and Yuting Su. Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1195–1198, New York, NY, USA, 2015. Association for Computing Machinery.

[136] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark, 2020.

[137] Z. Yang, L. Zicheng, and C. Hong. Rgb-depth feature for 3d human activity recognition. *China Communications*, 10(7):93–103, 2013.

[138] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2):375–389, 2018.

[139] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 50–65, Cham, 2015. Springer International Publishing.

[140] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1728–1743, 2011.

[141] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012.

[142] Chenyang Zhang and Yingli Tian. Rgb-d camera-based daily living activity recognition. *Journal of computer vision and image processing*, 2(4):12, 2012.

[143] Z. Zhang, K. Huang, T. Tan, and L. Wang. Trajectory series analysis based event rule induction for visual surveillance. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[144] Zhang Zhang, Kaiqi Huang, and Tieniu Tan. Multi-thread parsing for recognizing complex events in videos. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 738–751, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[145] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8668–8678, October 2019.

[146] Hang Zhao, Zhicheng Yan, Heng Wang, Lorenzo Torresani, and Antonio Torralba. Slac: A sparsely labeled dataset for action classification and localization. 12 2017.

[147] Zhe Lin, Zhuolin Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *2009 IEEE 12th International Conference on Computer Vision*, pages 444–451, 2009.

[148] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[149] Yipin Zhou and Tamara L. Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.