Human Activity Recognition in Video Benchmarks: A Survey

Tej Singh and Dinesh Kumar Vishwakarma

Abstract Vision-based Human activity recognition is becoming a trendy area of research due to its broad application such as security and surveillance, human—computer interactions, patients monitoring system, and robotics. For the recognition of human activity various approaches have been developed and to test the performance on these video datasets. Hence, the objective of this survey paper is to outline the different video datasets and highlights their merits and demerits under practical considerations. We have categorized these datasets into two part. The first part consists two-dimensional (2D-RGB) datasets and the second part has three-dimensional (3D-RGB) datasets. The most prominent challenges involved in these datasets are occlusions, illumination variation, view variation, annotation, and fusion of modalities. The key specification of these datasets are resolutions, frame rate, actions/actors, background, and application domain. All specifications, challenges involved, and the comparison made in tabular form. We have also presented the state-of-the-art algorithms that give the highest accuracy on these datasets.

Keywords Human activity recognition · Human–human interaction · RGB RGB-Depth (RGB-D) dataset

1 Introduction

In the present era, human activity recognition [1–5], in videos has become a prominent area of research in the field of computer vision. It has many daily living applications such as patient monitoring, object tracking, threat detection, and security and surveillance [6–9]. The motivation to work in this field is to recognize human

T. Singh (✉) · D. K. Vishwakarma
Department of Information Technology,
Delhi Technological University, New Delhi, Delhi, India
e-mail: tejsingh_2k16phdec05@dtu.ac.in; ttomar07@gmail.com

D. K. Vishwakarma
e-mail: dinesh@dtu.ac.in; dvishwakarma@gmail.com

B. S. Rawat et al. (eds.), *Advances in Signal Processing and Communication* ,
Lecture Notes in Electrical Engineering 526,
https://doi.org/10.1007/978-981-13-2553-3_24

gestures, actions and interactions in videos. The recognition of human activities in video involves various steps such as preprocessing, segmentation, feature extraction, dimension reduction and classification. We can save time if we have accurate knowledge of the publically available datasets [10, 11], so that there is no need to generate new dataset and a researcher's work will be easier to identify the datasets and a key focus will be on developing the new algorithm rather than gathering the information about datasets. With the advancement of labelling algorithm, it becomes an opportunity to label the dense dataset videos for activity recognition, object tracking, and scene reconstruction [12–14]. This work covers gesture recognition, daily living actions or activity, sports actions, human–human interactions and human–object interaction datasets. This paper consists of both RGB and RGB-D publically available datasets. This work provides datasets specifications such as year of publication, frame rates, spatial resolution, the total number of action and number of actors (subjects) performing in videos and state-of-the-art solutions on existing benchmarks. Tables 1 and 2 provides the details of RGB and RGB-D datasets, respectively. Before 2010, a large number of RGB video dataset was available to this community [15–17]. After the advancement of low-cost depth sensor, e.g. Microsoft Kinect, there has been a drastic increase in 3D, and multi-modal videos datasets. Due to low cost and lightweight sensors datasets are recorded with multiple modalities such as depth frames, accelerometer, IR sensors frames, acoustical data, and skeleton data information. The RGB-D datasets having multiple modalities reduce the chance of loss of information in videos as compared to traditional RGB datasets at the cost of increased complexities [18, 19].

2 Related Work

Chaquet et al. [20], focused on 28 publically available RGB datasets of human action and activity. The dataset characteristics are discussed such as ground truth, numbers of action/actors, views and area of applications. Their work does not cover RGB-depth dataset available at that time. Edwards et al. [3], focused on pose-based methods and presented a novel high-level activity dataset. Their work gives no information about state-of-the-art accuracies on existing dataset. Wang et al. [21], discussed specific novel techniques on RGB-D-based motion recognition. T. Hassner [22], focused on action recognition and accuracy of most of the RGB datasets. The very limitation of this work is the action in depth datasets and area of applications. M. Firman [23], analysed the depth dataset such as semantics, identification, face/pose recognition and object tracking. Borges et al. [24], discussed advantages and shortcomings of various methods for human action understanding. Zhang et al. [25], engrossed in action RGB-D benchmarks and lack of considered pose, human interaction activities. Besides, they intended to cover state-of-the-art accuracy and classification techniques on specific benchmarks. Compared with the existing surveys, the primary aim of this work will provide an accessible platform to the readers.

Table 1 RGB (2D) video dataset

Dataset	Year	Modality	Application domain
KTH	2004	Grey	Human action recognition in real outdoor conditions
Weizmann	2005	RGB	Human action recognition
IXMAS	2006	RGB	Multi-view-invariant action recognitions
CASIA Action	2007	RGB	Human behaviour and human–human interaction
UCF Sports	2008	RGB	Sports actions recognition
Olympic Games	2008	RGB	Sports actions recognition
Hollywood	2008	RGB	Realistic actions recognition from movies
UT- Interaction	2009	RGB	Human–Human interaction activity recognition
BEHAVE	2009	RGB	Human Group behaviour activity analysis
HMDB51	2011	RGB	human–human interaction, human – object interaction
UCF50	2011	RGB	Human Sports activity recognition
BIT-Interaction	2012	RGB	Human–human interaction in realistic scenarios
UCF101	2013	RGB	Human Sports activity recognition
YouTube Sports 1 M	2013	RGB	Human Sports activity recognition
ActivityNet	2015	RGB	Human activity understanding
THUMOS'15	2015	RGB	Action recognition in wild video
ChaLearn: Action/Interaction	2015	RGB	Automatic learning of human action and interactions
FCVID	2015	RGB	Human activity understanding
YouTube 8 M	2016	RGB	Human activity recognition, human interaction
Okutama Action	2017	RGB	Concurrent human action recognition form aerial view

3 Challenges in HAR Dataset

In this section, we discuss challenges involved in RGB and RGB-D dataset. It can be noticed that dataset videos are facing limitations in at least one of aspects such as similarity of actions, cluttered background, viewpoints variations, illuminations variations and occlusions.

Table 2 RGB-D (3D) video dataset

Dataset	Year	Modality	Application Domain
MSR Action 3D	2010	Depth + skeleton	Sports Gesture recognition
CAD-60	2011	RGB, Depth, skeleton	Daily activity recognition
RGB-D HuDaAct	2011	RGB, Depth	Daily activity recognition
Berkeley MHAD	2013	RGB, depth, skeleton	Human behaviour Recognition
CAD-120	2013	Depth, skeleton	Action labelling, human and object tracking
Hollywood 3D	2013	RGB, Depth	Natural action recognition in movies
MSR Action Pairs	2013	Depth	Action pairs recognitions
UWA3D Multi-View	2014	RGB, Depth, skeleton	Similar and cross-view action recognition
Northwestern UCLA	2014	RGB, Depth, skeleton	Cross- view action recognition
LIRIS	2014	RGB, Depth, grey	Human activity recognition
UTD-MHAD	2015	RGB, Depth, skeleton	View- invariant human action recognition
M^2I	2015	RGB, Depth, skeleton	Human–human, human–object interaction
SYSU-3D HOI	2015	RGB, Depth, skeleton	human–object interaction
G3Di	2015	RGB, Depth, skeleton	Gaming interaction activity
NTU RGB + D	2016	RGB, Depth, skeleton, IR sequences	Human Action Recognition
PKU-MMD	2017	RGB(image and video), Depth, skeleton, IR sequences	Multi-modal action recognition

3.1 Background and Environmental Conditions

The background in videos may be different types such as slow/high dynamic, static, occluded, airy, rainy and dense populated. It can be observed that KTH dataset is more challenging due to changing the background as compared to Weizmann dataset. The UT-Interaction, BEHAVE, BIT Interactions datasets recorded in the larger outdoor area and changing natural background conditions. The various datasets such as UCF sports activity, UIUC, Olympic sports, hollywood1, HMDB51, THUMOS, ActivityNet and YouTube 8 M recorded from online sources YouTube, Google, and various movies, are challenging due to having both dynamic objects and backgrounds conditions.

3.2 Similarity and Dissimilarity of Actions

The similarity between the actions classes in the datasets provides a fundamental challenge to the researcher. There are many actions which seem to be similar in videos such as jogging, running, walking, etc. The accuracy of classification is affected by the same type of actions. The same actions performed by different actors increase the complexity of the dataset such as YouTube Sports 1 M dataset having thousands of videos of same action class.

3.3 Occlusion

Occlusion is a thing where another object hides the object of interest. For the human action and activity recognition, occlusion can be categorized as self-occlusion and occlusion of another object/partial occlusion. The depth sensor is severely affected by internal noise data and self-occlusion by performing users such as in CAD-60, 50 salad, Berkeley MHAD, UWA3D activity, LIRIS, MSR Action pair, UTD-MHAD, M2I, SYSU-3D HOI, NTU RGB + D and PKU-MMD datasets.

3.4 View Variations

The viewpoint of any activity recorded inside the video dataset is a key attribute in the human activity recognition system. The multiple views have more robust information than single view and independent of captured view angle inside the dataset. However, multiple views increase the complexity such as more training as well as test data is required for classification analysis. Here, KTH, Weizmann, Hollywood, UCF Sports, MSR Action 3D, and Hollywood 3D, are single view datasets. The multi-view datasets are CAD-60, CAD-120, UWA3D, Northwestern-UCLA, LIRIS, UTD- MHAD, NTU RGB-D, IXMAS, CASIA Action, UT-Interaction, BEHAVE, BIT-Interaction, Breakfast Action.

4 Approaches for Human Action Recognitions

Based on the methodologies used in recent years to recognize human action and activities we can categorize the existing solutions to two major categories such as handcrafted features descriptor and deep learning approaches.

4.1 Local and Global Approaches

The initial work of human action recognition is limited to pose somewhat or gesture recognition. The first step to recognize the human action in videos was introduced by Bobik and Davis [26]. They simplified human action using Motion History Images (MHI) and Motion Energy Images (MEI). The global MHI template is given by

$$(x, y, t) = \sum_{\tau=0}^{i-1} B(x, y, t-i), \tag{1}$$

where E_τ is obtained MEI at particular time instant τ, while $B(x, y, t-i)$ is binary image sequences represents detected objects pixels.

The local representation STIPs for action recognition introduced by Laptev et al. [27]. A local 3D Harris operator [23] show a good performance to recognized 3D data objects with less number of interest points and widely used in computer vision applications. It is based on local autocorrelation function and defined as

$$e(x, y) = \sum_{x_i\, y_i} W\left(x_{i,} y_i\right)[I(x_i + \Delta x + y_i + \Delta y) - I(x_i, y_i)]^2, \tag{2}$$

where, $I(\cdot,\cdot)$ is defined as the image function and x_i, y_i are the points in the Gaussian function W centred on (x, y), which defines the neighborhood area in analysis.

4.2 Deep Learning Approaches

After 2012, these architecture received initial successes with supervised approaches which overcome vanishing gradient problem by using ReLU, GPUs (reduced time complexities). Deep learning technique is data driven it lacks when training samples are less, so in the case of small activity dataset local and global feature extractors are good and efficient for classification purpose.

Li et al. [28] showed that 3D convolutional networks outperform the 2D frame based counterparts with a noticeable margin. The 3D convolution value at position (x, y, z) on the j^{th} feature map in the i^{th} layer is defined as,

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m} \sum_{P=0}^{P_{i-1}} \sum_{Q=0}^{Q_{i-1}} \sum_{R=0}^{R_{i-1}} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \tag{3}$$

where, R_i is the size of the 3D kernel along the temporal dimension while w_{ijm}^{pqr} is the $(p, q, r)^{\text{th}}$ value of the kernel connected to the m^{th} feature map in the previous layer. Karpathy et al. [29] proposed the concept of slow fusion to increase the temporal awareness of a convolutional network. Donahue et al. [30] addressed the problem of

action recognition through the cascaded CNN and a class of recurrent neural network (RCNN) which is also known as Long Short Term Memory (LSTM) networks is given as

$$h^t = \sigma\left(w_x x^t + w_h h^{(t-1)}\right) \tag{4}$$

$$z^t = \sigma\left(w_z h^{(t)}\right) \tag{5}$$

$$w_x \in \mathbb{R}^{r\times d},\ w_h \in \mathbb{R}^{r\times r},\ w_z \in \mathbb{R}^{m\times r} \tag{6}$$

Here, $x^{(t)} \in \mathbb{R}^d$ (external signal), $z^{(t)} \in \mathbb{R}^m$ (output signal), and $h^{(t)} \in \mathbb{R}^r$ (hidden state). The recurrent neural network is found to be best model for video activity analysis.

5 Discussion

In this section, we briefly discuss the advantages and disadvantages of both types of 2D and 3D datasets.

5.1 Advantages of RGB and RGB-D Dataset

It can observe that from Tables 3 traditional human activity datasets are recorded with a small number of actions recognition from segmented videos under somewhat controlled conditions. Some benchmarks downloaded from online media such as YouTube, movies and social videos sharing sites represent a realistic action scene which is more practical for real-life applications. UCF 101 dataset is the largest dataset in the context of some classes, video clips than UCF 11, UCF 50, Olympic sports and HMDB51 datasets. ActivityNet is large-scale RGB video dataset captured with complete annotated labels and bounding box. The 3D datasets have advantages over visual 2D dataset as they are less sensitive to illuminations because they are captured with multiple sensors system such as visual, acoustical, and inertial sensors systems. It can be observed that from Table 4, that the fusion of information using different sensors increases the recognition accuracy on depth dataset at the cost of increased complexities. The 3D Online RGB-D action dataset was recorded in a living room environment used for cross-action environment and real online action recognition. The NTU RGB + D dataset is having a large number of actions/actors among existing datasets and was captured with multiple modalities and different camera views. PKU-MMD is large scale benchmark focused on continuous multi-modalities 3D complex human activities with complete annotation information, and it is suitable for deep learning methods.

Table 3 Technical specification RGB and RGB-D dataset

Dataset	Resolution	FPS	Actions	Actors	Videos
KTH	160×120	25	6	25	600
Weizmann	180×144	50	10	9	90
IXMAS	390×291	23	13	11	1650
CASIA Action	320×240	25	8	24	1446
UCF Sports	720×480	10	10	–	150
Olympic Games	–	–	16	–	783
Hollywood	400×300 300×200	24	8	–	233
UT- Interaction	720×480	30	6	–	160
BEHAVE	640×480	25	6	5	163
HMDB51	320×240	30	51/-	–	6766
UCF50	320×240	25	50	–	6681
BIT-Interaction	320×240	30	8	–	400
UCF101	320×240	25	101	–	13320
YouTube Sports 1 M	–	–	487	–	1133158
ActivityNet	1280×720	30	203	–	27801
THUMOS'15	–	–	101	–	5600
ChaLearn: Action/Interaction	480×360	15	235	14	235
FCVID	–	–	239	–	91223
YouTube 8 M	–	–	4716	–	~800, 000
Okutama Action	3840×2160	30	12	9	44
MSR Action 3D	640×480	15	20	7	567
RGB-D HuDaAct	640×480	30	12	30	1189
CAD-60	640×480	25	12	4	60
Berkeley MHAD	640×480	30	11	12	
CAD-120	640×480	25	10	4	120
Hollywood 3D	1920×1080	24	14	*	650
MSR Action Pairs	320×240	30	10	6	180
UWA3D Multi-view	640×480	30	30	10	900
Northwestern-UCLA	640×480	30	10	10	1475
LIRIS	640×480, 720×576	25	828	21	*
UTD-MHAD	512×424	30	27	8	861
M^2I	320×240	30	22	22	1784
SYSU- 3D HOI	640×480	30	40	12	~ 480
G3Di	640×480	30	12	15	574
NTU RGB + D	512×424, 1920×1080	30	60	40	56880
PKU-MMD	512×424, 1920×1080	30	66/60	51/40	1076

Table 4 RGB and RGB-D dataset with state-of-the-art accuracy and techniques

Dataset	Classification technique	Max avg. accuracy (%)	Evaluation protocol	Reference year
Weizmann	Hybrid (SDGs + AESIs)	100	LOOCV	2016
KTH	Interest points (IP) with differential motion information	98.20	3-fold cross-validation	2016
IXMAS	HC-MTL + L/S Reg	94.7	Cross-View	2017
CASIA Action	Hierarchical Spatio-Temporal model (HSTM)	95.24	–	2017
Olympic Games	Motion Part Regularization	92.3	leave-one-group-out cross-validation	2015
Hollywood	Joint max margin semantic features, DCNN	48.58	Cross-View	2016
UT- Interaction	Hierarchical Spatio-Temporal Model (HSTM)	94.17	leave-one-out cross-validation (LOOCV)	2017
UCF-YouTube	Interest points (IP) with differential motion information	91.30	3-fold cross-validation	2016
BEHAVE	Group interaction zone(GIZ), (ARF + GCT + AF)	93.74	3-folds-cross-validation	2014
HMDB51	Multi-Stream Deep Network	67.8	–	2017
UCF50	HC-MTL + L/S Reg	80.63	LOGO (Cross-View)	2017
BIT-Interaction	4-level, Pachinko Allocation Model	93	10-fold cross-validation	2016
UCF101	Multi-Stream Deep Network	93.3	–	2017
YouTube Sports 1 M	HC-MTL + L/S Reg	89.7	LOGO (Cross-View	2017
ActivityNet	Spatial CNN + Motion features	53.8	–	2017

(continued)

Table 4 (continued)

Dataset	Classification technique	Max avg. accuracy (%)	Evaluation protocol	Reference year
THUMOS'15	Pyramid of Score Distribution Feature (PSDF)	40.9(0.1)	–	2016
ChaLearn: Action/Interaction	Fisher vector + iDT features	53.85	cross-validation	2015
FCVID	rDNN	76.0	–	2017
YouTube 8 M	NetVLAD + CG after pooling and MoE	83.0	–	2017
Okutama Action	SSD(RGB)	18.80	Cross-validation	2017
MSR Action 3D	ConvNets	100	cross-subject	2015
RGB-D HuDaAct	BoW with χ^2 kernel SVM	82.9	Cross-subject validation	2014
CAD-60	Decision-level fusion	96.4	cross-subjects	2015
Berkeley MHAD	Hierarchy of LDSs, HBRNN-L	100	–	2013
CAD-120	QQSTR with feature selection	95.2	4-fold cross-validation	2015
Hollywood 3D	Bag of features (BoF) with Disparity Pyramids	36.09	cross-validation	2014
MSR Action Pairs	HON4D + Ddisc	96	cross-validation	2013
UWA3D Multi-view	MSO-SVM	91.79 (0 degree)	Cross-view	2015
Northwestern –UCLA	CNN + Synthesized + Pre-trained	92.3	Cross-view	2017
LIRIS	Pose + Appearance + context	74 (recall)	–	2014
UTD-MHAD	Depth plus RGB using product rule	91.2	Cross-subject	2016
M^2I	FV/BoVW	92.33	Cross –view	2017
SYSU- 3D HOI	Joint heterogeneous features learning (JOULE)model	84.89 ± 2.29 (S2)	Cross-subject	2016

(continued)

Table 4 (continued)

Dataset	Classification technique	Max avg. accuracy (%)	Evaluation protocol	Reference year
G3Di	Hierarchical Transfer Segments (HiTS)	the average latency time 2 frames (66 ms)	Cross-subject	2016
NTU RGB + D	CNN + Synthesized + Pre-trained	87.21	Cross-view	2017
PKU-MMD	Joint Classification Regression RNN	64.20	Cross-view	2017

5.2 Disadvantages of RGB and RGB-D Dataset

Currently, there are many video datasets, despite this, there are limitations in automatically recognize and classify the human activities. The main reasons of such limitations in at least one of the form are the number of samples for each action, the length of clips, capturing environmental conditions, background clutter and viewpoints changes and some activities. The 2D datasets were recorded with a small number of actions to complex actions with a broad range of applications. The 2D datasets are faced more challenges like view variations, intra-class variations, cluttered background, partial occlusions, and camera movements than depth datasets. The RGB-D dataset is facing limitations of low resolutions, less training samples, the number of camera view, different actions, various subjects and less precision. Initial RGB-D datasets captured single actions videos frames under controlled indoor or lab environments. MSR Action 3D is restricted to gaming actions depth frames only. Northwestern-UCLA dataset was recorded with more than one Kinect sensors at the same time to collect multi-view representations. It becomes a challenge to handle and synchronize all sensors data information simultaneously.

6 Conclusion

A review of the various state-of-the-art datasets on human action has been presented. Human action datasets have been categorized into two major categories: RGB and RGB-D datasets. The challenges involved and specifications of these datasets have been discussed. The conventional RGB dataset faces the problems of a cluttered background, illumination variations, camera motion, viewpoints change and occlusions. It is a challenge for feature descriptors in activity recognitions datasets that meets the changing real-world environments. It is required robust evaluation techniques for cross-dataset validation, which will be useful for realistic scenarios applications.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. **43**, 1–43 (2011)
2. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**, 983–1009 (2013)
3. Edwards, M., Deng, J., Xie, X.: From pose to activity: surveying datasets and introducing CONVERSE. Comput. Vis. Image Underst. **144**, 73–105 (2016)
4. Dawn, D.D., Shaikh, S.H.: A comprehensive survey of human action recognition with spatiotemporal interest point (STIP) detector. Vis. Comput. **32**, 289–306 (2016)
5. Bux, A., Angelov, P., Habib, Z.: Vision-based human activity recognition: a review. Adv. Comput. Intell. Syst. **513**, 341–371 (2016)
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision. Beijing (2005)
7. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision (2006)
8. Xu, W., Miao, Z., Zhang, X.P., Tian, Y.: A hierarchical spatio-temporal model for human activity recognition. IEEE Trans. Multimedia **99**, 1 (2017)
9. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: ActivityNet: a large-scale video benchmark for human activity understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston (2015)
10. Ryoo, M.S., Chen, C.C., Aggarwal, J., Chowdhury, A.R.: An overview of contest on semantic description of human activities. In: Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 (2010)
11. Vishwakarma, D. K., Singh, K.: Human activity recognition based on spatial distribution of gradients at sub-levels of average energy silhouette images. In: IEEE Transactions on Cognitive and Development Systems, vol. 9, no. 4, pp. 316–327. (2017)
12. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco (2010)
13. Kong, Y., Liang, W., Dong, Z., Jia, Y.: Recognizing human interaction from videos by a discriminative model. IET Comput. Vis. **8**, 277–286 (2014)
14. Ni, B., Moulin, P., Yang, X., Yan, S.: Motion part regularization: Improving action recognition via trajectory group selection. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston (2015)
15. Aggarwal, J., Xia, L.: Human activity recognition from 3D data- a review. In: Pattern Recognition Letters. vol. 48 (2013)
16. Lun, R., Zhao, W.: A survey of applications and human motion recognition with Microsoft Kinect. In: International Journal of Pattern Recognition and Artificial Intelligence, vol. 29 (2015)
17. Presti, L.L., Cascia, M.L.: 3D skeleton-based human action classification: a survey. Pattern Recogn. **53**, 130–147 (2016)
18. Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: RGB-D based action recognition datasets: a survey. Pattern Recogn. **60**, 86–105 (2016)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of the Advances in Neural Information Processing Systems. (2014)
20. Chaquet, J.M., Carmona, E.J., Caballero, A.F.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. **117**, 633–659 (2013)
21. Wang, P., Li, W., Ogunbona P.O., Escalera, S.: RGB-D-based motion recognition with deep learning: a survey. Int. J. Comput. Vis. (2017)
22. Hassner, T.: A critical review of action recognition benchmarks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland (2013)
23. Firman, M.: RGBD datasets: past, present and future. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016)

24. Borges, P.-V.K., Conci, N., Cavallaro, A.: Video-based human behavior understanding: a survey. IEEE Trans. Circuits Syst. Video Technol. **23**, 1993–2008 (2013)
25. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)
26. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**, 107–123 (2005)
27. Li, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 221–231 (2013)
28. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Columbus (2014)
29. Donahue, J., Hendricks, L., Guadarrama, S., Rohrbach, M.V., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
30. Sipiran, I., Bustos, B.: Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes. In: The Visual Computer, vol. 27 (2011)