RGB-D Data-Based Action Recognition

Subjects: Computer Science, Information Systems
Contributor: Muhammad Bilal Shaikh

Early research on Human Action Recognition was dominated by the analysis of still images or videos, localizing the actor in a video spatio-temporally using bounding boxes, temporal extent, and a spatio-temporal cuboid which contains a particular action. Human Action Recognition has many applications, including the automated annotation of user videos, indexing and retrieving user videos, automated surveillance, monitoring elderly patients using specially adapted cameras, robot operations, and live blogging of actions. In recent times, the availability of massive amounts of video data has provided significance to the understanding of video data (through a sequence of images) with the possibility of solving problems such as scene identification, searching through video content, and interaction recognition through video scenes. RGB-D generally refers to Red, Green, Blue plus Depth data captured by RGB-D sensors. An RGB-D image provides a per-pixel depth information aligned with corresponding image pixels. An image formed through depth information is an image channel in which each pixel relates to a distance between the image plane and the corresponding object in the RGB image. The addition of depth information to conventional RGB image helps improve the accuracy and the denseness of the data.

Keywords: action recognition ; deep learning ; RGB-D ; assistive health

1. RGB-D

1.1. RGB-D Data Acquisition

Acquisition of depth information is mainly based on triangulation and Time-of-Flight (ToF) techniques. The former technique may be implemented passively using stereovision, which retrieves depth information by capturing the same scene from different point of views. Stereovision emulates a human vision principle where depth is computed as a disparity between images taken from different viewpoints. This may require knowledge of the geometry of cameras and calibration needs to be performed for each change in system configuration. An active approach relies on structured light, which uses an IR light pattern onto the scene to estimate disparity through varying object's depth. In addition to this, ToF and Light Detection and Ranging (LiDAR) scanners measure the time that light takes to hit an object's surface and return to the detector. LiDAR uses mechanical components to its surrounding. However, ToF performs distance computation using integrated circuits. Chen et al. [1] and others [2] have briefly surveyed depth data acquisition in RGB-D sensors.

1.2. RGB-D Sensors

Most of the consumer RGB-D sensors rely on structured light or ToF approaches. Such RGB-D sensors possess noise and data distortions, which are tackled by specifically designed algorithms. Nevertheless, ToF provides a better depth resolution than the others, which is about a few millimeters. Moreover, structured light systems are not beneficial in outdoor scenarios because solar light strongly affects IR cameras. HAR tasks that do not require very high depth resolution and precision have been easily implemented using both structured light sensors and ToF devices. Such devices represented a very good compromise between cost, performance, and usability, and allowed implementation of unobtrusive and privacy-preserving solutions. Some consumer-preferred RGB-D sensors are outlined in the following subsections.

1.2.1. Microsoft® Kinect™ Sensors

Microsoft released the Kinect RGB-D sensor, a low-cost but high-resolution tool that could be easily interfaced to a computer, and whose signals could be easily manipulated through common academic practices. The Kinect sensor V1 uses structured light, and Kinect V2 is based on ToF. The latter exhibits less software complexity but requires fast hardware, such as pulse width modulation (PWM) drivers. The Kinect technology pushed the development of depth-based algorithms and processing approaches. Kinect has been discontinued, but alternative sensors are available in the market. Azure Kinect is a recent spatial computing developer kit with computer vision and speech models, and a range of development interfaces that can be connected to Azure cognitive services. Azure Kinect is not available for consumers

and thus not a replacement of Kinect. Michal et al. [3] presented a comprehensive evaluation of Azure Kinect and its comparison with both versions of Kinect. Different versions of Kinect Sensor are shown in **Figure 1**a (from bottom to top—Kinect v1, Kinect v2, and Azure Kinect).

Figure 1. Various RGB-D sensors: (a) Microsoft Kinect [4][5][6], (b) Intel RealSense L515 [7], and (c) Orbbec Astra Pro [8].

The Kinect sensor makes the task of capturing RGB-D data easier by sensing the depth dimension of the subject and its environment. It also interprets the movement performed by a subject and transforms it into a format that practitioners can use for new experiments. Computer vision researchers have leveraged Kinect's vision technology for performing tasks such as aiding children to overcome autism [9] and for doctors in their operating rooms. Azure Kinect has been released for developers and industries which will potentially transform human–computer interaction in various industries including manufacturing, education [10], healthcare [11], retail [12], transportation [13], and beyond.

1.2.2. Intel® RealSense™ Depth Cameras

Intel RealSense depth cameras encompass a family of stereoscopic and portable RGB-D sensors which includes subpixel disparity accuracy, assisted illumination, and performs well even in outdoor settings. Keselman et al. [14] provided a brief overview of Intel RealSense cameras. The R400 family is successor to the R200 family that includes improvements in its stereoscopic matching algorithm and correlation cost function as well as an optimization in design, which enables the R400 family to consume lower power than R200 while operating on the same image resolutions. Intel has divided its RGB-D sensors into different categories which includes stereo depth, LiDAR, coded light, and tracking sensors. The Intel RealSense LiDAR Camera L515 [15] shown in **Figure 1**b is the smallest high-resolution LiDAR depth camera to date. The Intel D400 [16] series uses Active IR stereo technology. The Intel SR [17] series uses coded light technology; however, the recently introduced L series uses LiDAR technology for acquiring depth information. The L series has significantly reduced the size of the sensor, which can accelerate the use of RGB-D sensors in HAR.

1.2.3. Orbbec® Depth Cameras

Orbbec Astra sensors incorporate processor which replaces traditional cable-based connection to sensor.The Orbecc camera-computer package is economical compared to Kinect or RealSense devices. Several SDKs are available including Astra SDK (developed by the manufacturers of the sensor) and OpenNI framework for 3D natural interaction sensors. The use of different sensors in the same problem could affect the accuracy of the process. Coroiu et al. [18] demonstrated safe exchange of Kinect sensor with the Orbbec sensor. According to the experiments, over 16 classifiers demonstrated that choice of sensor does not affect the accuracy. However, seven classifiers produced a drop-in accuracy. Furthermore, calibration algorithms using different RGB-D sensor are compared in [19]. In general, RGB-D sensors exhibit acceptable accuracy, but in some cases, calibration processes are critical to increase the sensor's accuracy and enable it to meet the requirements of such kinds of applications.

2. Classical Machine Learning-Based Techniques

Classical machine learning-based action recognition techniques use handcrafted features and can be classified on the basis of RGB data [20], depth data [21][22], skeleton sequences [23], and methods using a combination [24] of these data modalities (as illustrated in **Figure 2**). **Table 1** summarizes the best performing techniques which achieved benchmark accuracies for popular RGB-D datasets in action recognition research. The following subsections will discuss depth-, skeleton-, and RGB-D-based methods.

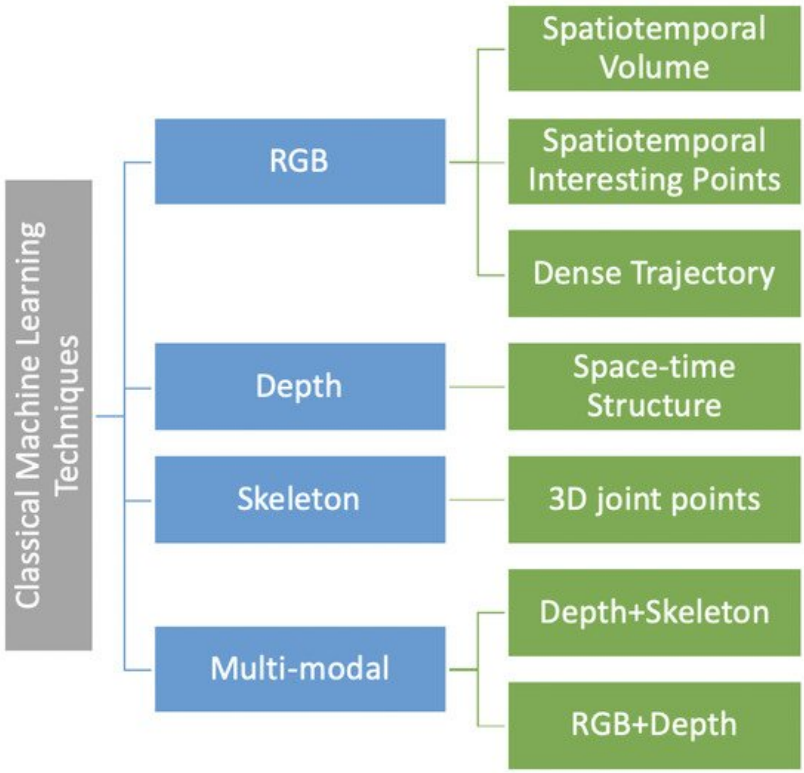

Figure 2. Hierarchy of action recognition techniques based on handcrafted features that use classical machine learning.

Table 1. Summary of popular action recognition datasets and methods that achieved the best recognition accuracy. Note that PDF stands for probability distribution function, i3D stands for inflated 3D, OF stands for Optical Flow, and GCN stands for Graph Convolutional Networks.

Year	Ref.	Methods (Modality)	Action Datasets	MSR Daily Activity 3D [25]	UT-Kinect [26]	EPIC Kitchen-55 [27]	NW-UCLA [28]	Toyota-SH [29]	HuDaAct [30]	UTD-MHAD [31]	Charades [32]	NTU RGB-D 120 [33]	miniSports [34]	Sports-1M [35]	IRD [36] [37]	HMDB-51 [38]	ICVL-4 [36] [37]	NTU RGB-D 60 [39]	MSR-Action3D [40]
2012	[41]	2D CNN (RGB-D)							89										
2015	[42]	DTQ-SVM (RGB-D)			100														90
2017	[43]	CNN (RGB-D)		98														75	
2018	[37]	i3D CNN + 2D CNN (RGB-D)								92								94	
2019	[44]	CNN (RGB + OF)									56								
2019	[34]	i3D CNN (RGB)											74						
2019	[45]	3D CNN (RGB)												75					
2019	[36]	GCN (Skeleton)													80		91		
2019	[46]	CNN (RGB)														82			
2019	[47]	TBN-Inception (RGB-Audio + OF)				35													

2.1. Depth Data-Based Techniques

Motion changes in the depth maps of the human body are used to represent action. Depth data can be observed as a space–time structure which is extracted from the appearance and motion information to describe human actions. Yang et al. [51] have proposed a supernormal vector feature through a depth map sequence for action representation. Oreifej et al.

[21] have proposed an orientation histogram feature of 4D normal vectors to represent the appearance information of a 3D spatio-temporal depth structure. Rehmani et al. [54] have proposed the idea of the main direction of a depth-curved surface where a perspective-independent feature and a principal component histogram are used to represent action. Yang et al. [53] have proposed the Depth Motion Map (DMM) to project spatio-temporal depth structure onto motion history maps. More recent motion history maps are represented by Histogram of Gradients (HoG) features in series to represent actions. Chen et al. [54] have used local binary features instead of HoG features; they [55] also investigated spatio-temporal depth structure from front, side, and upper directions. Miao et al. [56] have considered discrete cosine variation to compress the depth map and represent action through features using transform coefficients.

Year	Ref.	Methods (Modality)	Action Datasets	Daily Activity 3D [25]	UT-Kinect [29]	EPIC Kitchen 55 [27]	NW-UCLA [28]	Toyota-SH [24]	HuDaAct [30]	UTD-MHAD [31]	Charades [32]	YouTube RGB-D 120 [33]	miniSports [34]	Sports-1M [35]	IRD [36]	HMDB-51 [38]	ICVL-4 [37]	NTU RGB-D [39]	MSR-Action3D [40]
2020	[49]	3D CNN + GCN (RGB-D)	94	81	96														
2020	[50]	HAMLET (RGB-D) CNN (RGB-D)	98		95														
2020	[51]		94		92														
2020	[52]		96																
2020	[53]		94																
2020	[54]		96																
2020	[55]		94																
2020	[56]		96																

2.2. Skeleton Sequence-Based Techniques

Changes in position and appearance changes in human joint points between frames are used to describe action. Xia et al. [26] have modeled action through a discrete hidden Markov model. Action features have also been extracted through 3D Histograms of Oriented Displacements (HoD) [57], Accumulation of Motion Energy (AME) function aided with the Eigenjoint-based method [23], and through a longest common sequence algorithm [58] to select high-discriminative power features from the relative motion trajectories of the skeleton.

2.3. RGB-D Data-Based Techniques

The research results in [22][52][59] show that depth-based methods achieve better action recognition performance than RGB-based methods. Therefore, some researchers have also tried a fusion of different modalities. Chaaroui et al. [60] have investigated the fusion of skeleton and depth data to overcome problems caused by occlusion and perspective changes in skeleton features. In addition, a sparse regression learning-based method to fuse depth and skeleton features has been proposed by Li et al. [24]. A multi-kernel-based learning method for describing actions has been proposed by Althloothi et al. [61]; they calculated spherical harmonics through depth data and fused this with the spatial information of skeleton joints. Furthermore, RGB and depth data fusion has also been attempted by some researchers. For example, Liu et al. [62] used generic algorithms, Jalal et al. [63] merged spatio-temporal features, and Ni et al. [64] introduced the multi-level fusion of RGB and depth data features. However, to answer the missing modality problem, Kong et al. [65] have proposed a discriminative relational representation learning (DRRL) method. In the absence of a single modality in testing, this method transfers knowledge from training data to substitute the missing modality and achieves better recognition performance. The main concern with RGB-D data fusion is that it adds more computational complexity to the action recognition algorithm. Yu et al. [66] have proposed a binary representation for RGB-D data fusion with structure-preserving projections. This approach produced high efficiency and effectiveness on various action recognition benchmarks of RGB-D data.

3. Deep Learning

Computer vision researchers have directed considerable attention to the application of deep learning in action recognition. The classical machine learning-based methods are based on handcrafted features, which are not robust. Deep learning-based methods have been utilized due to their automated feature learning from images. Researchers have extracted action features from RGB data, depth data, and skeleton sequences using deep learning methods. The following subsections discuss the fundamental variants of neural networks, and later we present some modern deep learning-based approaches used in RGB-D data.

3.1. Neural Networks Variants

Recent successes in deep neural networks have boosted research in pattern recognition and computer vision. Some commonly used variants of neural networks are briefly outlined in following subsections.

3.1.1. Convolutional Neural Networks (CNN)

CNNs represent one of the most notable deep learning approaches, where they have been highly effective in a variety of computer vision applications. CNNs are good at recognizing patterns in Euclidean data, i.e., images, text, and videos. CNN works with a mathematical function called convolution, which is a special kind of linear operation. In convolution, input neurons are multiplied with a set of weights that are commonly known as filters or kernels. The filters act as a sliding window across the whole image and enable CNNs to learn features from neighboring cells. Within the same layer, the same filter will be used throughout the image, this is referred to as weight sharing. For example, using a CNN to classify images of dogs vs. non-dogs, the same filter could be used in the same layer to detect the nose and the ears of the cat. There are four basic ideas behind CNN that benefit from the characteristics of natural signals: local connections, shared

weights, pooling, and the use of many layers [67]. These four key ideas can be labeled as the Convolution layer, Rectified Linear Unit (ReLU) layer, Pooling, and Fully Connected (FC) Layer, respectively. An example of CNN architecture is presented in **Figure 3**a, originally shown in [68].

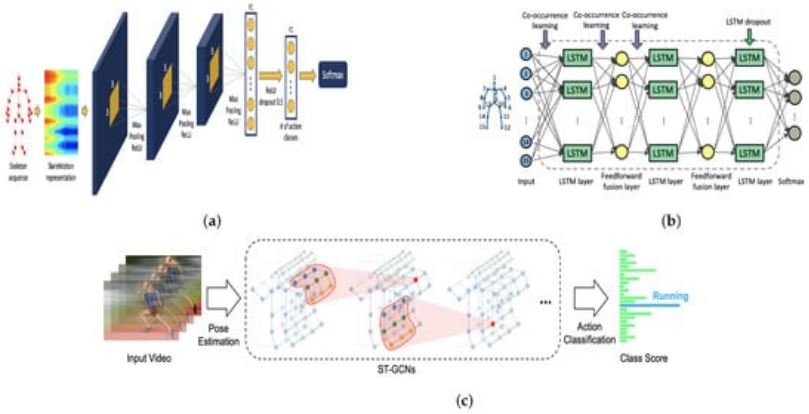

Figure 3. Illustration of deep learning techniques for processing RGB-D data. (**a**) Convolutional Neural Network (CNN). (**b**) Long Short-Term Memory (LSTM). (**c**) Graph Convolutional Network (GCN).

3.1.2. Recurrent Neural Networks (RNN)

The Recurrent Neural Network (RNN), Auto-Associative, or Feedback Network is a type of neural network that has variants including Gated Recurrent units. RNNs have been quite successful in conducting tasks like speech recognition, caption generation, machine translation, image/video classification, human dynamics, and action recognition, among other applications. The RNN function is an alternative to CNN; the RNN function is good at learning dependencies among spatially correlated data like image pixels [69]. RNN cannot store information for a longer duration. Long Short-Term Memory (LSTM) is a special kind of RNN capable of learning temporal relationships on a long-term scale. LSTM [70] uses a gates mechanism: write (input gate), read (output gate), and reset (forget gate), where this gates mechanism controls the behavior of its memory cells. The use of LSTM has produced effective results in speech recognition, especially in phoneme recognition. However, learning with LSTM is often challenging in real-time sequences [71]. An example of LSTM architecture is presented in **Figure 3**b, originally shown in [72].

3.1.3. Graph Convolutional Networks (GCN)

Earlier variants of neural networks are implemented using regular or Euclidean data. However, real-life data have a graph structure that is non-Euclidean. Therefore, the non-regularity of data structures has led to advancements in graph neural networks. Graph Convolutional Networks (GCN) are also considered as one of the basic Graph Neural Networks variants. Convolution in GCNs is the same operation as in CNNs. GCNs [73] are an efficient variant of CNNs on graphs. In GCNs, the model learns from the neighboring nodes by stacking layers of learned first-order spectral filters followed by a nonlinear activation function to learn graph representations. For simplicity, GCNs take a graph with some labeled nodes as input and generate label predictions for all graph nodes. GCNs could be divided into two approaches: Spatial GCNs and Spectral GCNs. An example of GCN is presented in **Figure 3**c, originally shown in [74].

3.2. Deep Learning-Based Techniques Using RGB-D Data

Deep learning can directly obtain hierarchical features from different data modalities and provides a more effective solution. Accordingly, appearance and optical sequences can be used as inputs to deep networks. Besides aspects of appearance and motion information, deep learning-based methods can also be applied using depth sequences and skeleton joint information. Wang et al. [75] have used convolution to learn action features from depth data. They [76] combined motion and structure information in a depth sequence by pairing structured dynamic images at the body, part, and joint levels through bidirectional rank pooling. Every pair is constructed from depth maps at each granularity level and serves as input to CNN. Song et al. [77] have proposed a model that uses different levels of attention in addition to an RNN with LSTM to learn discriminative skeleton joints. Ye et al. [78] have embedded temporal information with dense motion trajectories to learn actions.

Yan et al. [68] have modeled relationships between graphs and joints by using a graph-oriented CNN. Deep learning-based feature learning has been shown to provide better performance than handcrafted feature extraction methods; however, there are still challenges concerning RGB-D data fusion. Deep learning-based action recognition methods use different

standalone as well as hybrid neural network architectures, which can be classified as Single-Stream, Two-Stream, Long-term Recurrent Convolutional Network (LRCN), and Hybrid network-based architectures. The following subsections summarize these architectural styles.

3.2.1. Single Stream

A single-stream model is similar to the AlexNet [79] type of image classification network. Single-stream architecture can take advantage of regularization through local filters, parameter sharing at convolution layers, and local invariance building neurons (max pooling). Such neural network architecture shifts the engineering focus from feature design strategies to network structure and hyperparameter tuning strategies. Architectural details from AlexNet [79] can be used with different hyperparameter configurations. A single-stream architecture fuses information from all the frames in the softmax layer connected to the last fully connected layers with dense connections. Given an entire action video, the video-level prediction can be produced by forward propagation of each frame individually through the network and then averaging individual frame predictions over the duration of the video. However, single-stream architecture has been a foundation for other extended architectures. Some possible extensions to single-stream architecture have been explored by Baccouche et al. [80], Ji et al. [81], and Karpathy et al. [35].

3.2.2. Two Stream

The two-stream model uses two disjointed CNNs containing spatial and temporal information, which are later fused together. The spatial network performs action recognition from single video frames, while the temporal network learns to recognize action from motion, i.e., dense optical flow. The idea behind this two-stream model relates to the fact that the human visual cortex contains two pathways for object and motion recognition, i.e., the ventral stream performs object recognition and the dorsal stream recognizes motion. Spatial-stream CNN is modelled similar to the single-frame model discussed earlier. Given an action video, each frame is individually passed through the spatial network where an action label is assigned to each frame. The temporal-stream CNN is not the same as motion-aware CNN models (which use stacked single video frames as input). It takes stacked optical flow displacement fields between several consecutive frames as input to explicitly learn a temporal feature.

In two-stream models, the pioneering work of Simonyan and Zisserman [82] uses a single image and multi-optical flow sequence stack as input to the 2D CNN. Zhang et al. [83] have extended Simonyan and Zisserman's [82] work by using a motion vector instead of optical flow as an input to improve performance and comprehend real-time action recognition. Feichtenhoer et al. [84] have proposed an innovative approach involving moving the classification layer to the middle of the network for spatio-temporal information fusion, and this was shown to improve the accuracy. Wang et al. [85] have contributed to the input and training strategy of convolution and proposed Temporal Segment Network (TSN), improving the two-stream CNN. The notion of TSN was based on long-range temporal structural modeling. Later, Lan [86] and Zhou [87] enhanced the TSN. Carreira et al. [88] adopted the structure of Inception-v1 and inflated two-stream CNN to 3D CNN for action recognition. Zhu et al. [89] have expanded two-stream CNN to a 3D structure by drawing out the pooling operation.

3.2.3. Long-Term Recurrent Convolutional Network (LRCN)

LRCN uses CNN in co-ordination with an LSTM-based network. In LSTM-based deep learning methods, actions can be represented as feature changes between frames in the video, LSTM is widely used to improve action recognition techniques. Ng et al. [90] have presented a linear RNN for recognizing human actions that connects the output of a CNN with an LSTM cell. A new architecture—P3D ResNet—has been proposed by Qiu et al. [91], which uniquely places all the variants of blocks in a different placement of ResNet. In skeletal data, to deal with noise, Liu et al. [92] extended the idea to analyze spatio-temporal domains simultaneously by introducing an effective tree structure-based traversal framework. This framework uses a cross-modal feature fusion strategy within LSTM unit and a gating mechanism to learn the reliability of sequential data in long-term context representation. For mapping video frames with variable length inputs to variable length outputs, Donahue et al. [93] have proposed an LRCN. Unlike those methods that learn CNN filters based on a stack of a fixed number of input frames, LRCN [93] is not constrained to fixed-length input frames and thus can learn to recognize more complex action video. As illustrated in **Figure 4** (RGB-D input visuals taken from the work in [33]), individual video frames are first passed through CNN models with shared parameters and are then connected to a single-layer LSTM network. More precisely, the LRCN model combines a deep hierarchical visual-feature extractor, i.e., a CNN feature extractor, with an LSTM that can learn to recognize temporal variations in an end-to-end fashion.

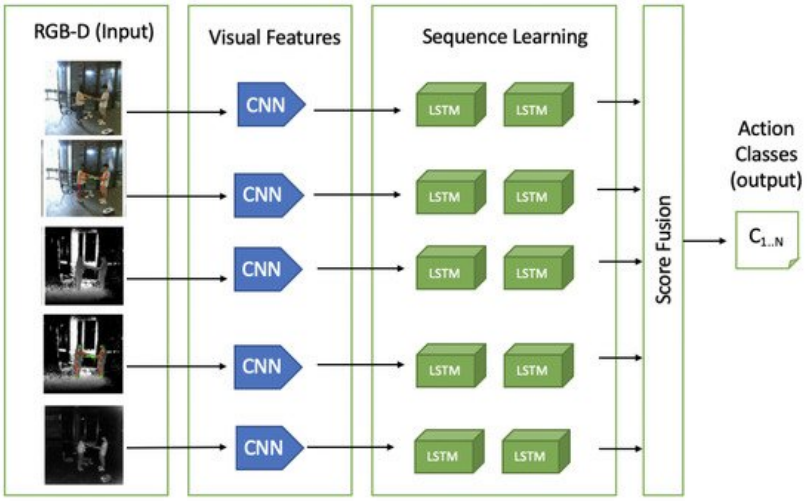

Figure 4. A possible architecture of LRCN with RGB-D input. Input from each modality, i.e., RGB, RGB + Skeleton joints, Depth, Depth + Skeleton joints, and IR are passed through a CNN layer for extracting visual features and an LSTM layer for sequence learning. Scores from each model are then fused and mapped to the number of classes for predictions. Visuals of RGB-D input are taken from NTU RGB-D 60 dataset [39].

References

1. Chen, C.; Jafari, R.; Kehtarnavaz, N. A Survey of Depth and Inertial Sensor Fusion for Human Action Recognition. Multimed. Tools Appl. 2017, 76, 4405–4425.
2. Rosin, P.L.; Lai, Y.K.; Shao, L.; Liu, Y. RGB-D Image Analysis and Processing; Springer: Berlin/Heidelberg, Germany, 2019.
3. Tölgyessy, M.; Dekan, M.; Chovanec, L.; Hubinskỳ, P. Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2. Sensors 2021, 21, 413.
4. Microsoft. Buy the Azure Kinect Developer kit–Microsoft. 2019. Available online: (accessed on 14 June 2021).
5. EB Games. Kinect for Xbox One (Preowned)-Xbox One-EB Games Australia. 2015. Available online: (accessed on 14 June 2021).
6. EB Games. Kinect for Xbox 360 without AC Adapter (Preowned)-Xbox 360-EB Games Australia. 2013. Available online: (accessed on 14 June 2021).
7. Intel Corporation. LiDAR Camera L515 – Intel® RealSense™ Depth and Tracking Cameras. 2019. Available online: (accessed on 14 June 2021).
8. Orbbec 3D. Astra Series-Orbbec. 2021. Available online: (accessed on 14 June 2021).
9. Lee, I.J. Kinect-for-windows with augmented reality in an interactive roleplay system for children with an autism spectrum disorder. Interact. Learn. Environ. 2020, 1–17.
10. Yukselturk, E.; Altıok, S.; Başer, Z. Using game-based learning with kinect technology in foreign language education course. J. Educ. Technol. Soc. 2018, 21, 159–173.
11. Pal, M.; Saha, S.; Konar, A. Distance matching based gesture recognition for healthcare using Microsoft's Kinect sensor. In Proceedings of the International Conference on Microelectronics, Computing and Communications (MicroCom), Durga, India, 23–25 January 2016; pp. 1–6.
12. Ketoma, V.K.; Schäfer, P.; Meixner, G. Development and evaluation of a virtual reality grocery shopping application using a multi-Kinect walking-in-place approach. In Proceedings of the International Conference on Intelligent Human Systems Integration, Dubai, UAE, 7–9 January 2018; pp. 368–374.
13. Zhang, Y.; Chen, C.; Wu, Q.; Lu, Q.; Zhang, S.; Zhang, G.; Yang, Y. A Kinect-based approach for 3D pavement surface reconstruction and cracking recognition. IEEE Trans. Intell. Transp. Syst. 2018, 19, 3935–3946.
14. Keselman, L.; Woodfill, J.I.; Grunnet-Jepsen, A.; Bhowmik, A. Intel(R) RealSense(TM) Stereoscopic Depth Cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1267–1276.
15. Drouin, M.A.; Seoud, L. Consumer-Grade RGB-D Cameras. In 3D Imaging, Analysis and Applications; Springer: Berlin/Heidelberg, Germany, 2020; pp. 215–264.

16. Grunnet-Jepsen, A.; Sweetser, J.N.; Woodfill, J. Best Known Methods for Tuning Intel® RealSense™ Depth Cameras D415. 2018. Available online: (accessed on 28 January 2021).
17. Zabatani, A.; Surazhsky, V.; Sperling, E.; Moshe, S.B.; Menashe, O.; Silver, D.H.; Karni, T.; Bronstein, A.M.; Bronstein, M.M.; Kimmel, R. Intel® RealSense™ SR300 Coded light depth Camera. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2019, 2333–2345.
18. Coroiu, A.D.C.A.; Coroiu, A. Interchangeability of Kinect and Orbbec Sensors for Gesture Recognition. In Proceedings of the 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 6–8 September 2018; pp. 309–315.
19. Villena-Martínez, V.; Fuster-Guilló, A.; Azorín-López, J.; Saval-Calvo, M.; Mora-Pascual, J.; Garcia-Rodriguez, J.; Garcia-Garcia, A. A Quantitative Comparison of Calibration Methods for RGB-D Sensors Using Different Technologies. Sensors 2017, 17, 243.
20. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-based Action Recognition Datasets: A Survey. Pattern Recognit. 2016, 60, 86–105.
21. Oreifej, O.; Liu, Z. 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), Portland, OR, USA, 23–28 June 2013; pp. 716–723.
22. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time Human Action Recognition Based on Depth Motion Maps. J. Real Time Image Process. 2016, 12, 155–163.
23. Yang, X.; Tian, Y. Effective 3D Action Recognition using EigenJoints. J. Vis. Commun. Image Represent. 2014, 25, 2–11.
24. Li, M.; Leung, H.; Shum, H.P. Human Action Recognition via Skeletal and Depth based Feature Fusion. In Proceedings of the 9th International Conference on Motion in Games, Burlingame, CA, USA, 10–12 October 2016; pp. 123–132.
25. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 13–18 June 2010; pp. 9–14.
26. Xia, L.; Chen, C.; Aggarwal, J. View Invariant Human Action Recognition using Histograms of 3D Joints. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 20–27.
27. Damen, D.; Doughty, H.; Farinella, G.M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 720–736.
28. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv 2012, arXiv:cs.CV/1212.0402.
29. Das, S.; Dai, R.; Koperski, M.; Minciullo, L.; Garattoni, L.; Bremond, F.; Francesca, G. Toyota Smarthome: Real-World Activities of Daily Living. In Proceedings of the Internation Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
30. Ni, B.; Wang, G.; Moulin, P. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In Proceedings of the Internation Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1147–1153.
31. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing A Depth Camera and A Wearable Inertial Sensor. In Proceedings of the Int. Conf. on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.
32. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Proceedings of the European Conference Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 510–526.
33. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.Y.; Kot Chichung, A. NTU RGB + D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2019, 2684–2701.
34. Korbar, B.; Tran, D.; Torresani, L. SCSampler: Sampling Salient Clips from Video for Efficient Action Recognition. In Proceedings of the International Conference on Compututer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6231–6241.
35. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale Video Classification With Convolutional Neural Networks. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognit (CVPR), Columbus, OR, USA, 24–27 June 2014; pp. 1725–1732.

36. Kim, S.; Yun, K.; Park, J.; Choi, J. Skeleton-Based Action Recognition of People Handling Objects. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WCACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 61–70.

37. Zhu, J.; Zou, W.; Xu, L.; Hu, Y.; Zhu, Z.; Chang, M.; Huang, J.; Huang, G.; Du, D. Action Machine: Rethinking Action Recognition in Trimmed Videos. arXiv 2018, arXiv:cs.CV/1812.05770.

38. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.

39. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB + D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognition (CVPR), Los Alamitos, CA, USA, 27–30 June 2016; pp. 1010–1019.

40. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view Action Modeling, Learning and Recognition. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656.

41. Zhao, Y.; Liu, Z.; Yang, L.; Cheng, H. Combing RGB and Depth Map Features for human activity recognition. In Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Hollywood, CA, USA, 3–6 December 2012; pp. 1–4.

42. Ye, J.; Li, K.; Qi, G.J.; Hua, K.A. Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 99–106.

43. Shahroudy, A.; Ng, T.T.; Gong, Y.; Wang, G. Deep multimodal feature analysis for action recognition in RGB + D videos. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2017, 40, 1045–1058.

44. Ryoo, M.S.; Piergiovanni, A.; Tan, M.; Angelova, A. AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures. arXiv 2020, arXiv:cs.CV/1905.13209.

45. Tran, D.; Wang, H.; Torresani, L.; Feiszli, M. Video Classification with Channel-separated Convolutional Networks. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5552–5561.

46. Wang, L.; Koniusz, P.; Huynh, D.Q. Hallucinating iDT Descriptors and i3D Optical Flow Features for Action Recognition with CNNs. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8698–8708.

47. Kazakos, E.; Nagrani, A.; Zisserman, A.; Damen, D. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

48. Das, S.; Sharma, S.; Dai, R.; Brémond, F.; Thonnat, M. VPN: Learning Video-Pose Embedding for Activities of Daily Living. In ECCV 2020; Springer: Cham, Switzerland, 2020; pp. 72–90.

49. Islam, M.M.; Iqbal, T. HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. arXiv 2020, arXiv:cs.RO/2008.01148.

50. Davoodikakhki, M.; Yin, K. Hierarchical action classification with network pruning. In International Symposium on Visual Computing; Springer: Berlin/Heidelberg, Germany, 2020; pp. 291–305.

51. Yang, X.; Tian, Y. Super Normal Vector for Activity Recognition using Depth Sequences. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OR, USA, 23–28 June 2014; pp. 804–811.

52. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. Real Time Action Recognition using Histograms of Depth Gradients and Random Decision Forests. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, (WCACV), SteamBoats Springs, CO, USA,, 24–26 March 2014; pp. 626–633.

53. Yang, X.; Zhang, C.; Tian, Y. Recognizing Actions using Depth Motion Maps-based Histograms of Oriented Gradients. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1057–1060.

54. Chen, C.; Jafari, R.; Kehtarnavaz, N. Action Recognition from Depth Sequences using Depth Motion Maps-based Local Binary Patterns. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WCACV), Waikola, HI, USA, 5–9 January 2015; pp. 1092–1099.

55. Chen, W.; Guo, G. TriViews: A General Framework to use 3D Depth Data Effectively for Action Recognition. J. Vis. Commun. Image Represent. 2015, 26, 182–191.

56. Miao, J.; Jia, X.; Mathew, R.; Xu, X.; Taubman, D.; Qing, C. Efficient Action Recognition from Compressed Depth Maps. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 16–20.

57. Gowayyed, M.A.; Torki, M.; Hussein, M.E.; El-Saban, M. Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.

58. Pazhoumand-Dar, H.; Lam, C.P.; Masek, M. Joint Movement Similarities for Robust 3D Action Recognition using Skeletal Data. J. Vis. Commun. Image Represent. 2015, 30, 10–21.

59. Papadopoulos, G.T.; Axenopoulos, A.; Daras, P. Real-time Skeleton-tracking-based Human Action Recognition using Kinect Data. In Proceedings of the International Conference on Multimedia Modeling, Dublin, Ireland, 6–10 January 2014; pp. 473–483.

60. Chaaraoui, A.; Padilla-Lopez, J.; Flórez-Revuelta, F. Fusion of Skeletal and Silhouette-based Features for Human Action Recognition with RGB-D Devices. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 91–97.

61. Althloothi, S.; Mahoor, M.H.; Zhang, X.; Voyles, R.M. Human Activity Recognition using Multi-features and Multiple Kernel Learning. Pattern Recognit. 2014, 47, 1800–1812.

62. Liu, L.; Shao, L. Learning Discriminative Representations from RGB-D Video Data. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 1493–1500.

63. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust Human Activity Recognition from Depth Video using Spatiotemporal Multi-fused Features. Pattern Recognit. 2017, 61, 295–308.

64. Ni, B.; Pei, Y.; Moulin, P.; Yan, S. Multilevel Depth and Image Fusion for Human Activity Detection. IEEE Trans. Syst. Man Cybern. 2013, 43, 1383–1394.

65. Kong, Y.; Fu, Y. Discriminative relational representation learning for RGB-D action recognition. IEEE Trans. Image Process. 2016, 25, 2856–2865.

66. Yu, M.; Liu, L.; Shao, L. Structure-preserving binary representations for RGB-D action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 38, 1651–1664.

67. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. Neurocomputing 2016, 187, 27–48.

68. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.

69. Miikkulainen, R.; Liang, J.; Meyerson, E.; Rawal, A.; Fink, D.; Francon, O.; Raju, B.; Shahrzad, H.; Navruzyan, A.; Duffy, N.; et al. Chapter 15-Evolving Deep Neural Networks. In Artificial Intelligence in the Age of Neural Networks and Brain Computing; Academic Press: Cambridge, MA, USA, 2019; pp. 293–312.

70. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.

71. Ronao, C.A.; Cho, S.B. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. 2016, 59, 235–244.

72. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.

73. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv 2017, arXiv:cs.LG/1609.02907.

74. Caetano, C.; Sena de Souza, J.; Santos, J.; Schwartz, W. SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.

75. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P. Deep Convolutional Neural Networks for Action Recognition Using Depth Map Sequences. arXiv 2015, arXiv:cs.CV/1501.04686.

76. Wang, P.; Wang, S.; Gao, Z.; Hou, Y.; Li, W. Structured Images for RGB-D Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 1005–1014.

77. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. IEEE Trans. Image Process. TIP 2018, 27, 3459–3471.

78. Ye, Y.; Tian, Y. Embedding Sequential Information into Spatiotemporal Features for Action Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1110–1118.
79. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc.: Lake Tahoe, CA, USA, 2012; pp. 1097–1105.
80. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential Deep Learning for Human Action Recognition. In International Workshop on Human Behavior Understanding; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
81. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2013, 35, 221–231.
82. Simonyan, K.; Zisserman, A. Two-stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
83. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-Time Action Recognition With Deeply Transferred Motion Vector CNNs. IEEE Trans. Image Process. TIP 2018, 27, 2326–2339.
84. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
85. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
86. Lan, Z.; Zhu, Y.; Hauptmann, A.G.; Newsam, S. Deep Local Video Feature for Action Recognition. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1219–1225.
87. Zhou, B.; Andonian, A.; Torralba, A. Temporal Relational Reasoning in Videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11205, pp. 831–846.
88. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? In A New Model and the Kinetics Dataset. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognit. (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
89. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden Two-Stream Convolutional Networks for Action Recognition. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 363–378.
90. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. In Proceedings of the IIEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
91. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5533–5541.
92. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2018, 40, 3007–3021.
93. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 2017, 39, 677–691.
