Spatial and Temporal Human Action Recognition Analysis

Subjects: Robotics

Contributor: Stavros N. Moutsis , Konstantinos A. Tsintotas , Ioannis Kansizoglou , Antonios Gasteratos

Human action recognition in computer vision is the task that identifies how a person or a group acts on a video sequence. Early methods that rely on representation-based solutions, like the histogram of oriented gradients (HOG), local binary patterns (LBP), and motion analysis, have been used to address this problem over the years. Later works are based on machine and deep-learning techniques, such as support vector machines (SVM), two- or three-dimensional convolutional neural networks (2D-CNNs, 3D-CNNs), recurrent neural networks (RNNs), and vision transformers (ViT), aiming to enhance the performance and reduce bias.

human action recognition mobile-CNNs spatial analysis

1. Introduction

Human action (or activity) recognition attempts to determine what action is being performed by an individual or a group in a video sequence [1]. Even if this can be considered a simple task, it has puzzled computer vision scholars for several decades [2][3][4]. Throughout this period, human action recognition has been widely adopted by various scientific fields, such as human–machine interaction [5][6], medical assistive technologies [7][8][9][10], surveillance systems [11][12], sports analysis [13], and human–robot interaction [14][15]. Similarly, it assists in path planning for tasks like social collision avoidance and route optimization [16] in autonomous navigation [17]. However, the main reasons why human action recognition constitutes such a challenging task are the following. The first concerns the environment where the action occurs: entirely different surroundings may present the same act. Additionally, the direction might vary from one video to another (e.g., a person walks from right to left and vice versa). The second challenge is related to the sensor's position, which affects the recorded visual information. More specifically, the closer the sensor is to the scene, the more detailed information it provides, yet the more negligible the action it covers. Moreover, the video streams recorded by a steady camera should be perfectly stabilized; otherwise, the motion adds extra noise to the incoming data. To this end, the large amount of data needed is a common barrier to efficient solutions. Last is the curse of dimensionality: video sequences used for human action recognition use more than 500 image-frames with similar information, enlarging the size of the datasets.

Nevertheless, several methods have been proposed to address this problem based on different data types and techniques, which are mainly distinguished into two categories [18]. The first regards pipelines that apply representation-based solutions, such as global [18][19][20], local [18][19][21][22], and depth [23][24] ones. On the other hand, techniques of the second category concern frameworks implemented with deep-network-based pipelines, such as convolutional neural networks (CNNs) [25]. Despite the promising results of the former systems, their need to adapt to changes (e.g., environmental, frame background, or camera motion) between videos containing the same action presents a disadvantage. In contrast, the latter can adjust to the above challenges, showing remarkable outcomes in different computer vision and robotics tasks [26] (e.g., image recognition [27], object detection [28][29], visual-based navigation [30][31], place recognition [32][33][34], loop closure detection [35][36], and video description [37]). In particular, these approaches use two-dimensional CNNs (2D-CNNs) that receive a grid of values as input (i.e., an image) and subsequently perform spatial analysis via 2D convolutional filters. This way, they keep most of the image's information while reducing its dimensionality [38]. To this end, many CNN architectures have been proposed in the previous years, reaching improved performances (viz., LeNet [39], AlexNet [40], GoogleNet or InceptionNet [41], BN-Inception [42], VGGNet [27], and ResNet [43]). Still, many trainable parameters are retained in these models, rendering their training process time consuming (i.e., more than a week when employed on ImageNet [44][45]), even if modern graphics processing units (GPUs) are used. Because of this, mobile-CNNs (viz., YOLO [28], MobileNet-v1 [46], MobileNet-v2 [47], ShuffleNet-v1 [48], SuffleNet-v2 [49], NASNetMobile [50], FBNet [51], EfficientNet-b0 [52], MobileNet-v3 [53], and GhostNet [54]) were designed with fewer trainable parameters, simultaneously maintaining high outputs. In most of the cases, when 2D-CNNs are employed for human action recognition, these are large architectures [55], such as CNN-M-2048 [56][57][58], which is similar to ClarifaiNet [59], AlexNet [60], VGGNet [61][62][63], ResNet [64][65], GoogleNet [57], and BN-Inception [61][66]. When mobile versions are used, these include MobileNet-v2 [55][67] and EfficientNet-b0 [68]. Vision transformers (ViT) [69], which were recently introduced for different computer vision tasks (such as image classification [70], object detection [71], image segmentation [72], and action recognition [73][74][75], were inspired by their success in natural language processing [76]. However, while they show dominance over CNNs [77][78], they are characterized by high-complexity models in accordance with their demands for large-scale datasets [69][79]. On the contrary, lightweight models, such as Swin-T [80], MobileViT [81], and EVA-02-Ti [82], have also been applied, showing high performances.

2. Human Action Recognition through 3D-CNNs

As 3D convolutional filters can be applied to a sequence of consecutive images, performing spatial and temporal analysis simultaneously, 3D-CNNs have been proposed for human action recognition. Using a set of such models and capturing the video's visual appearance and motion dynamics, the authors in [83] propose a framework where the final prediction results from all the used models. Their pipeline is based on several architectures, giving better outcomes for different datasets. Nevertheless, they mention drawbacks, such as the high complexity and computational cost. Aiming to tackle this weakness, Sun et al. [84] simulate the 3D-CNN function utilizing 2D kernels at the first layers for spatial analysis, with 1D kernels following for the temporal analysis. Combined with improved dense trajectories (iDT) [21] and a linear support vector machine (SVM) classifier, C3D [85], a model employing 3 × 3 × 3 convolutional kernels to every layer, can achieve better results if it is previously trained on I380K [85]. Aiming to gain the high performance of 3D-CNNs in parallel with the low complexity of 2D-CNNs, Lin et al. [86] suggest a temporal shift module (TSM) that can be applied to 2D-CNNs without increasing the latency of the system. The basic concept lies in shifting parts of the channels through the time dimension, aiming at the information division between adjacent frames. More specifically, two strategies are introduced. The first concerns offline approaches, wherein all frames are processed, while the second regards the online systems, indicating that only the last frames are available during real-time activity recognition. In particular, the former utilizes a two-directional (+1 or −1) displacement, while during live demonstrations, the relocation happens only in one direction (+1). Lastly, ResNet-50 is adopted as the backbone network.

3. Human Action Recognition through Multiple-Stream CNNs

Two-stream CNN models perform spatial and temporal analysis via different 2D-CNNs that are trained separately [57], and the final prediction comes after the networks' later fusion. Specifically, the first 2D-CNN, responsible for spatial analysis, receives static RGB images as input, while the second network accepts a stack of dense optical flows between several consecutive images. The latter improves the framework's outcome as it is invariant to visual appearance, even when temporal coherence is not retained [87]. Therefore, temporal analysis is provided by the optical flow data. In a later work, Feichtenhofer et al. evaluate different fusion types, such as in a convolutional layer instead of the softmax layer [88]. Three different architectures, ClarifaiNet, GoogleNet, and VGGNet-16, are tested in [58], showing that the latter model is outperformed on spatial stream. At the same time, the performance of each network is improved on the temporal streams when they have previously trained on ImageNet. Since 3D-CNNs can learn spatiotemporal features only from RGB images, improved performance is attained when they are utilized with optical flows. The authors in [89] propose a two-stream 3D-ConvNet, wherein a 3D-CNN, called I3D, is used in each stream instead of a 2D-CNN. Previously trained on kinetics-400, I3D showed improved results [90].

Similarly, C3D [85] is employed in the two-stream CNN by the authors in [91], where two different types of fusion are tested. The first adopts a late fusion on the results provided by softmax average scores. The second uses an early fusion. Specifically, the vectors generated by the first fully connected (FC) layers are concatenated, creating a singular vector, which is subsequently loaded to an SVM. During training, a random frame is chosen from each video as input for the spatial stream. At the same time, an optical flow set of five or ten consecutive images is selected for the temporal stream [57]. Zhu et al. increase the accuracy by applying an end-to-end training protocol to both the spatial and temporal streams [66]. Their method is based on samples of 25 RGB images and optical flow stacks from each video. Aiming to achieve this result, at each epoch, the BN-Inception [42] model's last convolutional layer outputs end up in an FC layer via temporal pyramid pooling (TPP). The final prediction comes from the fusion of the spatial and temporal streams. Additionally, Feichtenhofer et al. [92] present a framework based on a two-stream 3D-CNN [93], wherein each model runs on different frame rates. Specifically, the one concerning the slow stream utilizes a low frame rate and is responsible for the spatial analysis. On the contrary, more frames are applied during rapid frames, aiming to handle the temporal analysis. The backbone networks ResNet-50 and ResNet-101 are also evaluated. Finally, apart from the video classification, the task of action detection is also tackled with high performance.

Because of the optical flow's high computational complexity, motion vectors are also proposed for lightweight live action recognition [94]. Similarly, Kim and Won [95] propose a stacked gray-scale three-channel image (SG3I) to replace the optical flow data and achieve faster implementation. Zong et al. use two additional streams, a spatial-saliency and a temporal-saliency stream, which capture the salient object and salient motion information, respectively, by taking as input sampled saliency maps. These two spatial and temporal streams create a four-stream feature extractor [65]. Huo et al. [55] propose a temporal trilinear pooling pipeline for lightweight action recognition, where three modalities are generated from compressed videos (viz., I-frames, motion vectors, and residuals serve as inputs to the framework). The CNN used as the backbone is MobileNet-v2, and three versions are employed, each dedicated to one of the three modalities. Finally, a processing speed of 40 frames per second is achieved on a mobile device. In [68], a multi-head attention mechanism [76] is applied after EfficientNet-bo, used as the backbone network to address the action recognition task based on [57]. The use of an attention module has been shown to improve performance in both spatial and temporal streams across all the examined networks serving as backbones, namely ResNet-18, ResNet-34 [43], ResNet-50, and the proposed EfficientNet-b0.

4. Human Action Recognition through Temporal Segment Networks

In temporal segment networks, each video is divided into three equal segments, wherein a snippet of each segment is randomly selected as an input of the two-stream CNN. Next, this input is applied three times in each video [96]. Finally, a segmental consensus function, such as max, average, or weighted average, is used for the spatial and temporal prediction before the fusion. This sampling strategy provides relevant information from the entire video, and learning is performed regardless of size. At the same time, the execution time and the system’s complexity remain constant. Instead of taking the segmental consensus function’s result as the final prediction for spatial and temporal streams, one more training step is employed in [61]. BN-Inception or VGGNet-16 is used as a feature extractor for the video’s three segments, aiming to train two separate SVMs for spatial and temporal analysis. This way, false matches between videos and labels are limited since the SVM’s final prediction comes from the feature extractor referring to the video’s three segments.

5. Human Action Recognition through CNN + RNN

Recurrent networks can be a valuable tool in human action recognition as a video is a temporal sequence of images. With that in mind, several techniques have been developed that use CNNs as feature extractors for feeding into a recurrent network, such as RNN, long short-term memory (LSTM) [97], or gated recurrent unit (GRU) [98]. AlexNet and GoogleNet, previously trained on ImageNet, are tested on two frameworks for activity identification [60]. The first method was explored using several types of pooling architectures, such as convolution pooling, late pooling, slow pooling, local pooling, and time-domain convolution, on the output of convolutional layers aiming to make the final prediction. Their findings show that convolution pooling outperformed the other architectures. The second pipeline connected an LSTM to the last convolutional layer outputs, intending to synthesize the temporal dynamics of the input stream. Five stacks of LSTM layers are used, followed by a softmax classifier to predict each frame. However, adding optical flow improved the performance in the LSTM approach but not in the convolutional pooling.

In [37], an end-to-end trainable hybrid model, entitled long-term recurrent convolutional network (LRCN), consisting of a CNN and an LSTM, is proposed. More specifically, the used CNN is a minor variant of AlexNet, called CaffeNet [99], that has previously been trained on ILSVRC-2021 [45]. LRCN is trained to predict the action at each time step, and the final prediction comes from sixteen clip-frames. It is worth noting that using optical flow enhanced the performance compared to the framework using RGB images. In another work by Carreira and Zisserman [89], four human action recognition techniques are compared with the two-stream I3D. Among them, one uses InceptionNet-v1's last average pooling layer, trained on ImageNet, as a feature extractor, with an LSTM following. During training, the cross-entropy loss is applied to the output of all time steps. However, CNN + LSTM and 3D-CNN are the only ones that do not apply the frame's optical flow, showing the lowest performance. AlexNet, previously trained on ImageNet, is used as a feature extractor to accomplish the spatial analysis [100]. Moreover, consecutive frames, in step six, are chosen to feed into a bi-directional LSTM to avoid using redundant frames without losing the action sequence. Similarly, VGGNet-16 trained on ImageNet is employed for the images' feature extraction in [62]. Backpropagation is utilized only for the last eight layers of the network when the model is trained on the target dataset. Subsequently, thirty extracted vectors constitute the bi-directional LSTM input, from which the final prediction is derived. Ahme et al. [67] use MobileNet-v2 trained on ImageNet as a feature extractor. Their network's last layers are frozen when trained on the target dataset, and new layers are added for fine-tuning. After each video frame is transformed via the network, it is fed to a GRU aiming to make the final prediction. Zhang et al. [63] propose a method that combines three of the techniques above, wherein separated VGGNets are used for both spatial and temporal streams. In addition, the video is divided into *k* segment networks as proposed by [96], while the prediction is carried out by loading to a Bi-LSTM the fused features from the convolutional layers.

Finally, it is worth noting the significance of the features represented by CNNs, as they have a crucial role in the techniques mentioned earlier. These features are fed into the RNNs to perform the temporal analysis and predict the action within a video. Gao et al. [101] propose the Res2Net module as an alternative to the bottleneck block, achieving in this way the capture of more and richer information from the input image, enhancing the features. This improvement is done by dividing the input map into smaller segments. Each part, apart from the first one, is transformed by a 3 × 3 convolution, with the output of the previous convolution being added to the input of the next one. Finally, all the outputs are merged back into one map, and a last 1 × 1 convolution is applied, similar to the initial bottleneck block.

6. Skeleton Data Approaches

When only RGB data is used, skeleton data is often utilized for human action recognition as it provides a greater amount of information regarding the pose of an individual, which is directly related to the type of activity. Graph convolutional networks (GCN) have been used to tackle the action recognition task through skeleton data [102], where each joint is defined as a node, and the connections between the joints are considered as the edges. To address the high computational cost of GCNs, Cheng et al. [103] have proposed the shift GCN. In this approach, the features of the adjacent joints are loaded to its current 1 × 1 convolution node. Furthermore, in [104], the spatial-temporal GCN (ST-GCN) has been presented. Multiple skeleton data are fed into multiple ST-GCN layers to achieve action recognition according to the changes in the graphs over time. As ST-GCN requires complex pre-processing of the input, Peng et al. [105] proposed a framework that captures the features between sequential graphs but with a lower computational cost, achieved by transforming them into three dimensions. Finally, in [106], the joint-bone fusion GCN is presented, which combines two streams, one for the bones and one for the joints, aiming to analyze the relationship between these two dependencies. Additionally, a pose estimation transformer is applied for semi-supervised training.

References

1. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. Sensors 2019, 19, 1005.
2. Arseneau, S.; Cooperstock, J.R. Real-time image segmentation for action recognition. In Proceedings of the 1999 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 1999), Conference Proceedings (Cat. No. 99CH36368), Victoria, BC, Canada, 22–24 August 1999; IEEE: Piscataway, NJ, USA, 1999; pp. 86–89.
3. Masoud, O.; Papanikolopoulos, N. A method for human action recognition. Image Vis. Comput. 2003, 21, 729–743.
4. Charalampous, K.; Gasteratos, A. A tensor-based deep learning framework. Image Vis. Comput. 2014, 32, 916–929.
5. Gammulle, H.; Ahmedt-Aristizabal, D.; Denman, S.; Tychsen-Smith, L.; Petersson, L.; Fookes, C. Continuous Human Action Recognition for Human-Machine Interaction: A Review. arXiv 2022, arXiv:2202.13096.
6. An, S.; Zhou, F.; Yang, M.; Zhu, H.; Fu, C.; Tsintotas, K.A. Real-time monocular human depth estimation and segmentation on embedded systems. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 55–62.
7. Yin, J.; Han, J.; Wang, C.; Zhang, B.; Zeng, X. A skeleton-based action recognition system for medical condition detection. In Proceedings of the 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), Nara, Japan, 17–19 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
8. Cóias, A.R.; Lee, M.H.; Bernardino, A. A low-cost virtual coach for 2D video-based compensation assessment of upper extremity rehabilitation exercises. J. Neuroeng. Rehabil. 2022, 19, 1–16.
9. Moutsis, S.N.; Tsintotas, K.A.; Gasteratos, A. PIPTO: Precise Inertial-Based Pipeline for Threshold-Based Fall Detection Using Three-Axis Accelerometers. Sensors 2023, 23, 7951.

10. Moutsis, S.N.; Tsintotas, K.A.; Kansizoglou, I.; An, S.; Aloimonos, Y.; Gasteratos, A. Fall detection paradigm for embedded devices based on YOLOv8. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques, Copenhagen, Denmark, 1 May–19 October 2023; pp. 1–6.
11. Hoang, V.D.; Hoang, D.H.; Hieu, C.L. Action recognition based on sequential 2D-CNN for surveillance systems. In Proceedings of the IECON 2018—44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3225–3230.
12. Tsintotas, K.A.; Bampis, L.; Taitzoglou, A.; Kansizoglou, I.; Kaparos, P.; Bliamis, C.; Yakinthos, K.; Gasteratos, A. The MPU RX-4 project: Design, electronics, and software development of a geofence protection system for a fixed-wing vtol uav. IEEE Trans. Instrum. Meas. 2022, 72, 7000113.
13. Wei, D.; An, S.; Zhang, X.; Tian, J.; Tsintotas, K.A.; Gasteratos, A.; Zhu, H. Dual Regression for Efficient Hand Pose Estimation. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6423–6429.
14. Carvalho, M.; Avelino, J.; Bernardino, A.; Ventura, R.; Moreno, P. Human-Robot greeting: Tracking human greeting mental states and acting accordingly. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1935–1941.
15. An, S.; Zhang, X.; Wei, D.; Zhu, H.; Yang, J.; Tsintotas, K.A. FastHand: Fast monocular hand pose estimation on embedded systems. J. Syst. Archit. 2022, 122, 102361.
16. Charalampous, K.; Kostavelis, I.; Gasteratos, A. Robot navigation in large-scale social maps: An action recognition approach. Expert Syst. Appl. 2016, 66, 261–273.
17. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Online Appearance-Based Place Recognition and Mapping: Their Role in Autonomous Navigation; Springer Nature: Berlin/Heidelberg, Germany, 2022; Volume 133.
18. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. Image Vis. Comput. 2017, 60, 4–21.
19. Poppe, R. A survey on vision-based human action recognition. Image Vis. Comput. 2010, 28, 976–990.
20. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 2001, 23, 257–267.
21. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–

3558.

22. Tsintotas, K.A.; Giannis, P.; Bampis, L.; Gasteratos, A. Appearance-based loop closure detection with scale-restrictive visual features. In Proceedings of the International Conference on Computer Vision Systems, Thessaloniki, Greece, 23–25 September 2019; Springer: Cham, Switzerland, 2019; pp. 75–87.
23. 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-Workshops, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 9–14.
24. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A review on human activity recognition using vision-based method. J. Healthc. Eng. 2017, 2017, 3090343.
25. Kansizoglou, I.; Bampis, L.; Gasteratos, A. Do neural network weights account for classes centers? IEEE Trans. Neural Netw. Learn. Syst. 2022, 34, 8815–8824.
26. Kansizoglou, I.; Bampis, L.; Gasteratos, A. Deep feature space: A geometrical perspective. IEEE Trans. Pattern Anal. Mach. Intell. 2021, 44, 6823–6838.
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
28. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
29. Tsintotas, K.A.; Bampis, L.; Taitzoglou, A.; Kansizoglou, I.; Gasteratos, A. Safe UAV landing: A low-complexity pipeline for surface conditions recognition. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Virtual, 24–26 August 2021; pp. 1–6.
30. An, S.; Zhu, H.; Wei, D.; Tsintotas, K.A.; Gasteratos, A. Fast and incremental loop closure detection with deep features and proximity graphs. J. Field Robot. 2022, 39, 473–493.
31. Tsintotas, K.A.; Sevetlidis, V.; Papapetros, I.T.; Balaska, V.; Psomoulis, A.; Gasteratos, A. BK tree indexing for active vision-based loop-closure detection in autonomous navigation. In Proceedings of the 2022 30th Mediterranean Conference on Control and Automation (MED), Athens, Greece, 28 June–1 July 2022; pp. 532–537.
32. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Probabilistic appearance-based place recognition through bag of tracked words. IEEE Robot. Autom. Lett. 2019, 4, 1737–1744.
33. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Tracking-DOSeqSLAM: A dynamic sequence-based visual place recognition paradigm. IET Comput. Vis. 2021, 15, 258–273.

34. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Visual Place Recognition for Simultaneous Localization and Mapping. In Autonomous Vehicles Volume 2: Smart Vehicles; Scrivener Publishing LLC: Beverly, MA, USA, 2022; pp. 47–79.

35. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Modest-vocabulary loop-closure detection with incremental bag of tracked words. Robot. Auton. Syst. 2021, 141, 103782.

36. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection. IEEE Trans. Intell. Transp. Syst. 2022, 23, 19929–19953.

37. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; 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, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.

38. Oikonomou, K.M.; Kansizoglou, I.; Gasteratos, A. A Framework for Active Vision-Based Robot Planning using Spiking Neural Networks. In Proceedings of the 2022 30th Mediterranean Conference on Control and Automation (MED), Vouliagmeni, Greece, 28 June–1 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 867–871.

39. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.

40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90.

41. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

42. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 448–456.

43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.

44. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.

45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, 115, 211–252.

46. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.

47. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.

48. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.

49. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.

50. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710.

51. Wu, B.; Dai, X.; Zhang, P.; Wang, Y.; Sun, F.; Wu, Y.; Tian, Y.; Vajda, P.; Jia, Y.; Keutzer, K. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10734–10742.

52. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.

53. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.

54. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.

55. Huo, Y.; Xu, X.; Lu, Y.; Niu, Y.; Lu, Z.; Wen, J.R. Mobile video action recognition. arXiv 2019, arXiv:1908.10155.

56. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. arXiv 2014, arXiv:1405.3531.

57. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. 2014, 27, 568–576.

58. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards good practices for very deep two-stream convnets. arXiv 2015, arXiv:1507.02159.

59. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 818–833.

60. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.

61. Lan, Z.; Zhu, Y.; Hauptmann, A.G.; Newsam, S. Deep local video feature for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1–7.

62. Chenarlogh, V.A.; Jond, H.B.; Platoš, J. A Robust Deep Model for Human Action Recognition in Restricted Video Sequences. In Proceedings of the 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), Milan, Italy, 7–9 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 541–544.

63. Zhang, Y.; Guo, Q.; Du, Z.; Wu, A. Human Action Recognition for Dynamic Scenes of Emergency Rescue Based on Spatial-Temporal Fusion Network. Electronics 2023, 12, 538.

64. Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Attention transfer from web images for video recognition. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1–9.

65. Zong, M.; Wang, R.; Ma, Y.; Ji, W. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. Appl. Soft Comput. 2023, 132, 109884.

66. Zhu, J.; Zhu, Z.; Zou, W. End-to-end video-level representation learning for action recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 645–650.

67. Ahmed, W.; Naeem, U.; Yousaf, M.H.; Velastin, S.A. Lightweight CNN and GRU Network for Real-Time Action Recognition. In Proceedings of the 2022 12th International Conference on Pattern Recognition Systems (ICPRS), Saint-Etienne, France, 7–10 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–7.

68. Zhou, A.; Ma, Y.; Ji, W.; Zong, M.; Yang, P.; Wu, M.; Liu, M. Multi-head attention-based two-stream EfficientNet for action recognition. Multimed. Syst. 2023, 29, 487–498.

69. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv 2020, arXiv:2010.11929.

70. Chen, C.F.R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
71. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring Plain Vision Transformer Backbones for Object Detection. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 280–296.
72. Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; Hong, Q. LViT: Language meets Vision Transformer in Medical Image Segmentation. IEEE Trans. Med. Imaging 2023, 1.
73. Ma, Y.; Wang, R. Relative-position embedding based spatially and temporally decoupled Transformer for action recognition. Pattern Recognit. 2024, 145, 109905.
74. Ulhaq, A.; Akhtar, N.; Pogrebna, G.; Mian, A. Vision Transformers for Action Recognition: A Survey. arXiv 2022, arXiv:cs.CV/2209.05700.
75. Yang, J.; Dong, X.; Liu, L.; Zhang, C.; Shen, J.; Yu, D. Recurring the Transformer for Video Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14063–14073.
76. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In Advances in Neural Information Processing Systems; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
77. Amjoud, A.B.; Amrouch, M. Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review. IEEE Access 2023, 11, 35479–35516.
78. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Appl. Sci. 2023, 13, 5521.
79. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. ACM Comput. Surv. 2022, 54, 1–41.
80. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv 2021, arXiv:abs/2103.14030.
81. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. arXiv 2021, arXiv:abs/2110.02178.
82. Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. EVA-02: A Visual Representation for Neon Genesis. arXiv 2023, arXiv:2303.11331.
83. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 35, 221–231.

84. Sun, L.; Jia, K.; Yeung, D.Y.; Shi, B.E. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.

85. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.

86. Lin, J.; Gan, C.; Han, S. TSM: Temporal Shift Module for Efficient Video Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.

87. Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M.J. On the integration of optical flow and action recognition. In Proceedings of the German Conference on Pattern Recognition, Stuttgart, Germany, 9–12 October 2018; Springer: Cham, Switzerland, 2018; pp. 281–297.

88. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1933–1941.

89. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.

90. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. arXiv 2017, arXiv:1705.06950.

91. Khong, V.M.; Tran, T.H. Improving human action recognition with two-stream 3D convolutional neural network. In Proceedings of the 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Ho Chi Minh City, Vietnam, 5–6 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.

92. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. arXiv 2019, arXiv:cs.CV/1812.03982.

93. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human Action Recognition from Various Data Modalities: A Review. IEEE Trans. Pattern Anal. Mach. Intell. 2023, 45, 3200–3225.

94. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with deeply transferred motion vector cnns. IEEE Trans. Image Process. 2018, 27, 2326–2339.

95. Kim, J.H.; Won, C.S. Action recognition in videos using pre-trained 2D convolutional neural networks. IEEE Access 2020, 8, 60179–60188.

96. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 20–36.
97. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.
98. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv 2014, arXiv:1409.1259.
99. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
100. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access 2017, 6, 1155–1166.
101. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. IEEE Trans. Pattern Anal. Mach. Intell. 2021, 43, 652–662.
102. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
103. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
104. 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.
105. Peng, W.; Shi, J.; Varanka, T.; Zhao, G. Rethinking the ST-GCNs for 3D skeleton-based human action recognition. Neurocomputing 2021, 454, 45–53.
106. Tu, Z.; Zhang, J.; Li, H.; Chen, Y.; Yuan, J. Joint-Bone Fusion Graph Convolutional Network for Semi-Supervised Skeleton Action Recognition. IEEE Trans. Multimed. 2023, 25, 1819–1831.

Retrieved from https://encyclopedia.pub/entry/history/show/119771