Vision-Based Human Action Recognition Field

Subjects: Computer Science, Artificial Intelligence

Contributor: Fernando Camarena , Miguel Gonzalez-Mendoza , Leonardo Chang , Ricardo Cuevas-Ascencio

Artificial intelligence's rapid advancement has enabled various applications, including intelligent video surveillance systems, assisted living, and human–computer interaction. These applications often require one core task: video-based human action recognition.

video-based human action recognition action recognition human action

1. Introduction

Artificial intelligence (AI) redefines the understanding of the world by enabling high-impact applications such as intelligent video surveillance systems [1], self-driving vehicles [2], and assisted living [3]. In addition, AI is revolutionizing areas such as education [4], healthcare [5], abnormal activity recognition [6], sports [7], entertainment [4][8], and human–computer interface systems [9]. These applications frequently rely upon the core task of video-based human action recognition, an active research field to extract meaningful information by detecting and recognizing what a subject is doing in a video [10][11][12]. Since its critical role in computer vision applications, the action recognition study can lead to innovative solutions that can benefit society in various ways. Nevertheless, it can take time to introduce oneself to the subject thoroughly.

2. Understanding Video-Based Human Action Recognition

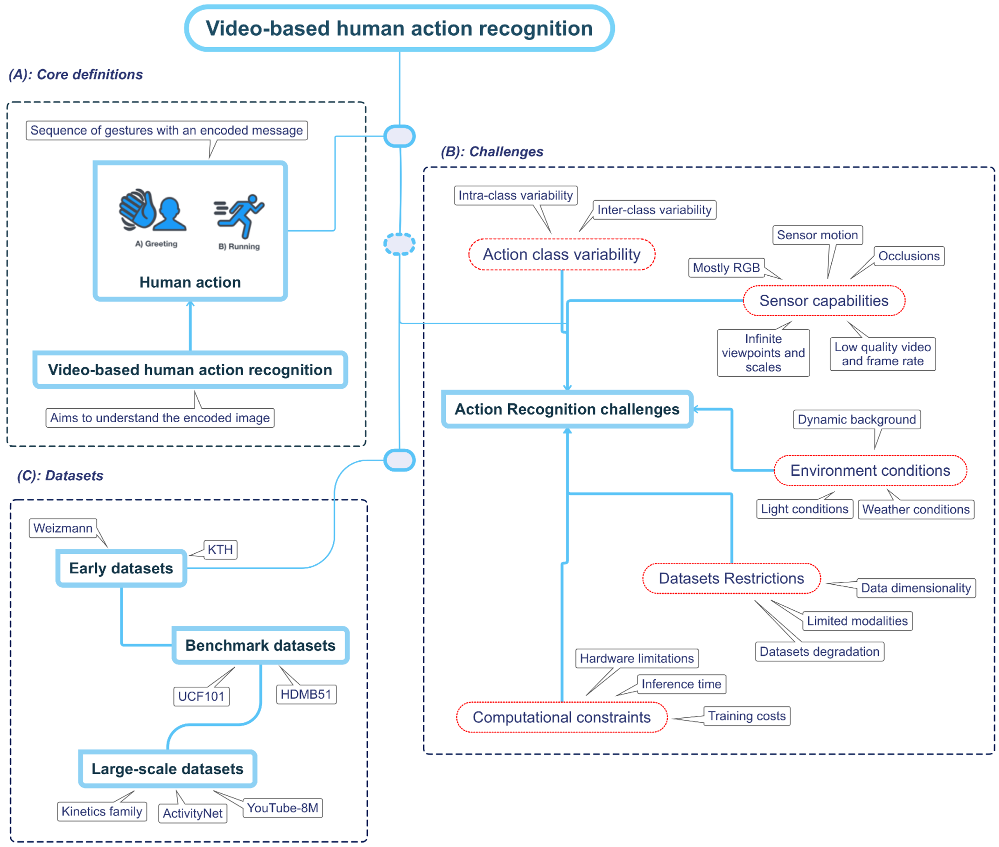

Figure 1. Video-based human action recognition overview. Part (**A**) represents human action; researchers instinctively associate a sequence of gestures with an action. For example, researchers might think of the typical hand wave when researchers think of the action greeting. On the contrary, imagining a person running will create a more dynamic scene with movement centered on the legs. Part (**B**) explains current challenges in the field, and Part (**C**) shows the relevant dataset used.

2.1. What Is an Action?

To understand the idea behind an action, picture the image of a person greeting another. Probably, the mental image constructed involves the well-known waving hand movement. Likewise, if researchers create a picture of a man running, they may build a more dynamic image by focusing on his legs, as depicted in **Figure 1**. Researchers unconsciously associate a particular message with a sequence of movements, which they call “an action” [4][13]. In other words, human action is an observable entity that another entity, including a computer, can decode through different sensors. The human action recognition goal is to build approaches to understand the encoded message in the sequence of gestures.

2.2. Challenges Involved in Video-Based Human Action Recognition

While humans have a natural ability to perceive and comprehend actions, computers face various difficulties when recognizing such human actions [14]. Researchers categorize the challenges into five primary categories: action-class variability, sensor capabilities, environment conditions, dataset restrictions, and computational constraints. By understanding these challenges, researchers may build strategies to overcome them and, consequently, improve the model’s performance.

2.2.1. Action Class Variability

Both strong intra- and inter-class variations of an action class represent a challenge for video-based human action recognition [14]. The intra-class variations refer to differences within a particular action class [15]. These variations stem from various factors, such as age, body proportions, execution rate, and anthropometric features of the subject [16]. For example, the running action significantly differs between an older and a younger individual. Additionally, researchers have repeated some of the actions so many times that researchers already perform them naturally and unconsciously, making it difficult even for the same person to act precisely the same way twice [14][16]. Finally, cultural contexts can impact how humans act, such as in the case of the greeting action class [17]. Due to the variability, developing a single model that accurately represents all instances of the same class is challenging [14]. Therefore, mitigating intra-class variation is a crucial research area in computer vision to represent all instances of the same class accurately.

Conversely, inter-class variation refers to the dissimilarities between distinct action classes [14], representing a significant challenge because some actions could share major feature vectors [18]. For example, while standing up and sitting down may be perceived as distinct actions, they share the same structure and semantics, making it challenging to differentiate one from another if the model approach does not consider their temporal structure [19]. A similar case is the walking and running actions, which, despite being different, can be seen as variations of the same underlying action. Therefore, to make computer vision applications more accurate and valuable, it is essential to make models that can handle inter-class variations.

2.2.2. Sensor Capabilities

In computer vision, a sensor detects and measures environmental properties such as light, temperature, pressure, and motion to convert them into electrical signals for computer processing [20]. Due to the capture of rich visual information, the RGB camera is the most common sensor used in video-based human action recognition, which senses the light intensity of three color channels (red, green, and blue) [4][20].

Using an RGB camera entails some challenges, including a reduced perspective due to the limited field of view [14], which may cause the target to be partially or not present in the camera field; a partial temporal view of the target subject is known as occlusion [4][14][21] and can be caused either by an object, another subject, the same subject or even the light conditions. Dealing with missing information is difficult because the occlusion may hide the action’s representative features [14]. For example, if a player’s legs during a kick are not visible to the camera’s field of view throughout a soccer match, it can be challenging to establish if they made contact with the ball.

Furthermore, there is no semantic of how to place the camera sensor, which implies that the target subject can appear in infinite perspectives and scales [22]. On the one hand, some perspectives may not help recognize an action [22][23]; for instance, when a person is reading a book, they will usually hold it in front of them; if the camera viewpoint is the subject's back, it will not perceive the book, and therefore, it will not be able to recognize the action.

2.2.3. Environment Conditions

Environmental conditions can significantly impact the classification accuracy of a model to recognize human actions by affecting the significance of the captured data [4][14]. To illustrate, poor weather conditions such as rain, fog, or snow reduce the target subject's visibility and affect the appearance features extracted. Likewise, in "real" conditions, the target subject will find itself in a scene with multiple objects and entities, which will cause a dynamic, unpredictable, and non-semantic background [14]; the delineation and comprehension of the objective and background can become increasingly complex and challenging when additional factors or variables are presented, which obscure the distinction between the foreground and background. Additionally, environmental conditions can generate image noise that limits representative visual features' extraction and complicates the subject track over time [24].

The environment light is also critical in identifying human actions [14], primarily if the model approach only relies on visual data for feature representation. Lighting conditions can cause subjects to be covered by shadows, resulting in occlusions or areas of high/low contrast, making taking clear, accurate, and visual-consistent pictures of the target subject complex. These circumstances may also result in images differing from those used during model training, confounding the recognition process even further.

2.2.4. Dataset Restrictions

The effectiveness of a machine learning model for recognizing human actions heavily depends on the dataset's quality used in its training phase [25]. The dataset's features, such as the number of samples, diversity, and complexity, are crucial in determining the model's performance. However, using a suitable dataset to boost the model's accuracy takes time and effort [26].

The first approach is constructing the dataset from scratch, ensuring the action samples fit the application's requirements. However, this process can be resource-intensive [26] because most effective machine learning models work under a supervised methodology, and consequently, a labeling process is required [27]. Data labeling [27] involves defining labeling guidelines, class categories, and storage pipelines to further annotate each action sample individually, either manually or by outsourcing to an annotation service to ensure consistent and high-quality data samples.

For some application domains, data acquisition can be challenging due to various factors [28], such as the unique nature of the application, concerns regarding data privacy, or ethical considerations surrounding the use of certain types of data [29]. Consequently, data acquisition can be scarce, insufficient, and unbalanced in the action classes, presenting significant obstacles to developing effective models or conducting meaningful analyses [28].

The second approach involves utilizing well-known datasets with a predefined evaluation protocol, enabling researchers to benchmark their methodology against state-of-the-art techniques. Nevertheless, there are some limitations, including the availability of labeled data; for example, the UCF101 [30] and HMDB51 [31] are one of the most used benchmark datasets [32]. Still, their data dimensionality is insufficient to boost the deep-learning model [33]. Furthermore, current datasets for action recognition face the challenge of adequately representing and labeling every variation of a target action [14], which is nearly impossible due to the immense variability in human movements and environmental factors. This limitation can impact the accuracy and generalizability of action recognition models if the dataset does not represent the same data distribution of the target application [14].

Another main problem with publicly available datasets is their degradation over time [14]; for example, a researcher that aims to use the kinetics dataset [33] must download each video sample from the Internet. However, some download links may no longer work, and specific videos may have been removed or blocked. As a result, accessing the same dataset used in prior research is impossible, leading to inconsistent results [14].

Most of the datasets provide the video along with a textual label tag [34]. Although this is enough to train a model to recognize human action, they have two main limitations. On the one hand, there is no clear intuition that text label tags are the optimal label space for human action recognition [35], particularly in cases where a more nuanced or fine-grained approach to labeling is required or in an application scenario where multi-modal information is available [34]. On the other hand, the exclusive use of RGB information in current datasets overlooks the potential benefits of other input sensors [36], such as depth or infrared sensors, which may provide more detailed and complementary representations of human actions in specific application scenarios.

3. The Evolution of Video-Based Human Action Recognition Approaches

3.1. Handcrafted Approaches

As described in **Figure 2**, handcrafted approaches established the foundation for video-based human action recognition, which entails a manual feature engineering process, where human experts manually design features that support a computer to understand.

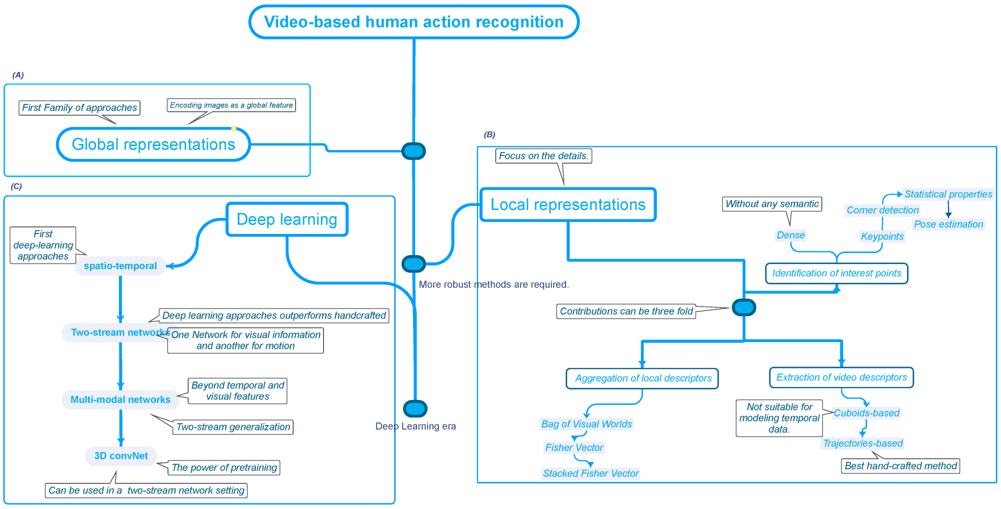
Video-based human action recognition
(A)
First Family of approaches
Encoding images as a global feature
Global representations
(B)
Focus on the details.
Local representations
Without any semantic
Dense
Corner detection
Statistical properties
Pose estimation
Keypoints
Identification of interest points
More robust methods are required.
Contributions can be three fold
Aggregation of local descriptors
Extraction of video descriptors
Bag of Visual Worlds
Fisher Vector
Stacked Fisher Vector
Not suitable for modeling temporal data.
Cuboids-based
Trajectories-based
Best hand-crafted method
(C)
Deep learning
First deep-learning approaches
spatio-temporal
Deep learning approaches outperforms handcrafted
Two-stream networks
One Network for visual information and another for motion
Beyond temporal and visual features
Multi-modal networks
Two-stream generalization
3D convNet
The power of pretraining
Can be used in a two-stream network setting
Deep Learning era

Figure 2. The Evolution of Action Recognition Approaches. The initial attempt at vision-based human action recognition relied on global representations (**A**), which were inferior to local representations (**B**). Lastly, deep learning approaches (**C**) became the most popular, with 3D convolutional neural networks becoming the most advanced because they can learn multiple levels of representations.

Two main components usually form handcrafted approaches. Firstly, feature extraction [4] transforms the input video into a video representation of the action. Secondly, the Action Classification [4] component maps the video representation onto a label tag.

3.1.1. Feature Extraction

Global representations [37] are the first attempt to recognize actions whose intuition is to capture the video input into one global feature. A simple intuition of the effects of this type of method is our natural ability to recognize human actions only by looking at the subject’s silhouette. However, this approach proved inadequate in addressing the numerous challenges posed by videos or images, such as different viewpoints and occlusions. Consequently, global representations could not fully capture the variability of an action. Among the most relevant methods are Motion Energy Image (MEI) [38], Motion History Image (MHI) [39], silhouettes [40], and Spacetime Volume (STV) [41].

The world is full of little details that are difficult to capture using the “big picture”. Intuitively, as humans, to discover those little secrets, researchers need to explore, focus on the details, and zoom in on the regions of interest, which is the idea behind local representations [32][42], as shown in **Figure 2**B. Local representations seek to extract descriptors from multiple regions of the video to obtain insights into the details. Local approaches break down into a sequence of steps: (a) detection of points of interest, (b) extraction of video descriptors, and (c) aggregations of local descriptors. As a consequence, the researcher’s contributions can be three-fold.

As the name suggests, the first step is to detect which regions of the video to analyze. Nevertheless, determining the significance of a region can be a relatively tricky undertaking. Applying edge detection algorithms is one method, such as Space-Time Interest Points (STIPs) [43] and hessian detector [44]. However, its application could lead to noise and lousy performance due to the extraction of edges that belong to something other than the target subject. To assess the regions' relevance and eliminate noisy information, Liu et al. [45] propose using statistical properties as a pruning method.

Camarena et al. [46][47] suggest that pose estimation can be used as the regions of interest, resulting in a method that has a fixed and low number of processing areas, which ensures a consistent frame processing rate. However, the approach is dependent on the subject body's visibility.

Another solution is to apply dense sampling [48], which consists of placing points without semantics. Dense sampling increases the classification accuracy, but it is computationally expensive [46]. In addition, noise injected by other motion sources can affect the classifier's performance [46][47].

Once researchers have determined which regions to analyze, they must extract the corresponding region description. Visual and motion data are essential for accurately characterizing an action [46]. In this regard, the typical approach combines several descriptors to have a complete perspective of the target action. Regarding the visual information, researchers have a Histogram Of Oriented Gradients 3D (HOG3D) [49], Speed-Up Robust Features (SURF) [50], 3D SURF [44], and pixel pattern methods [51][52][53]. On the other hand, descriptors that focus on motion information include Histogram of Oriented Flow (HOF) [54], Motion Boundaries Histogram (MBH) [48], and MPEG flow [55].

Capturing motion information is a complex task; videos are composed of images in which the target person moves or changes location over time [48]. The naive method uses cuboids, which utilize static neighborhood patterns throughout time. However, cuboids are unsuitable for modeling an object's temporal information. Its natural evolution was trajectory-based approaches [48][56][57] that rapidly became one of the most used methods [32][47].

Trajectory-based methods use optical flow algorithms to determine the position of the object of interest in the next frame, which helps to improve the classification performance [32]. Although several efficient optical flow algorithms exist, their application at different points of interest can be computationally expensive [47]. To reduce the computational time, it is essential to know that there are several motion sources besides the subject of interest, including secondary objects, camera motions, and ambient variables. Focusing on the target motion may reduce the amount of computation required. On the one hand, researchers can use homographies [32] for reducing the motion's camera; on the other hand, pose estimation [47] can be used to remove the optical flow process thoroughly.

Descriptor aggregation is the final stage in which the video descriptor is constructed using the region descriptors acquired from the preceding processes. There are several methods, including Bag-of-Visual-Words (BoVW) [58], Fisher Vectors [59], Stacked Fisher Vector (SFV) [60], Vector Quantization (VQ) [61], Vector of Locally Aggregated Descriptors (VLAD) [62], Super Vector Encoding (SVC) [63]. Among the handcrafted approaches, it is popularly referred to as FV and SFV, along with dense trajectories achieving the best classification performance [37].

3.1.2. Action Classification

Action classification aims to learn a mapping function to convert a feature vector to a label tag. The literature exposes different approaches, including template-based [38][64][65], generative [66][67], and discriminative models [37][48].

Template-based models are the naive method that compares the feature vector to a set of predefined templates to assign the label tag of the closest instance given a similarity measure. The generative models [66][67] are based on probability and statistics techniques; some representative works include Bayesian Networks and Markov chains.

Discriminative models are one of the most common techniques, including most machine learning methods [37][48]. Due to its performance, handcrafted approaches commonly rely on Support Vector Machines (SVM).

Researchers rely on dimensionality reduction techniques [68] to lower the model's complexity and extract meaningful feature vectors that boost the performance in high-dimensional datasets. Standard techniques include Principal Component Analysis (PCA) [69] and Linear Discriminant Analysis (LDA) [70]. On the one hand, PCA assists in identifying the most representative features, while LDA aids in finding a linear combination of feature vectors that distinguish different action classes.

3.2. Deep Learning Approaches

Due to their strong performance in various computer vision tasks [1][2][3], Convolutional Neural Networks (CNNs) have become increasingly popular. Hence, its application to vision-based human action recognition appeared inevitable.

Andrej et al. [71] developed one of the first approaches, which involved applying a 2D CNN to each frame and then determining the temporal coherence between the frames. However, unlike other computer vision problems, using a CNN does not outperform handcrafted approaches [18]. The main reason was that human actions are defined by spatial and temporal information, and using a standalone CNN does not fully capture the temporal features [18]. Therefore, subsequent deep learning research for human action recognition has focused on combining temporal and spatial features.

As a common practice, biological processes inspire computer vision and machine learning approaches. For example, as individuals, researchers use different parts of our brain to process the appearance and motion signals researchers perceive [72][73]. This understanding can be used for human action recognition, as suggested by [72]. The concept is straightforward. On the one hand, a network extracts spatial characteristics from RGB images. On the other hand, a parallel network extracts motion information from the optical flow output [72]. The network can effectively process visual information by combining spatial and temporal information.

Due to the comparable performance of two-stream networks to trajectory-based methods [18], interest in these approaches grows, leading to novel research challenges such as how to merge the output of motion and appearance features. The most straightforward process, referred to as late fusion [74], is a weighted average of the stream's predictions. More sophisticated solutions considered that interactions between streams should occur as soon as possible and proposed the method of early fusion [74].

Because of the temporal nature of videos, researchers investigated the use of Recurrent Neural Networks (RNN) [75] and Long-Term Short-Term Memory (LSTM) [76][77] as the temporal stream for two-stream approaches. As proven by Ma et al. [78], pre-segmented data are necessary to fully explore the performance of an LSTM in videos thoroughly, eventually leading to Temporal Segment Networks (TSN), which has become a popular configuration for two-stream networks [18].

A generalization of two-stream networks is multi-stream networks [18], which describe actions using additional modalities such as pose estimation [79], object information [80], audio signals [81], text transcriptions [82], and depth information [83].

One factor that impacts the performance of deep neural networks is the amount of data used to train the model. In principle, the more data researchers have, the higher our network performance. However, the datasets employed in vision-based human action recognition [30][84][85] do not have the scale that requires a deep learning model [33]. Not disposing of enough data has various implications, one of which is that it is difficult to determine which neural network architecture is optimal. Carreiera et al. [33] introduced the Kinetics dataset as the foundation for re-evaluated state-of-the-art architectures and proposed a novel architecture called Two-Stream Inflated 3D ConvNet (I3D) architecture, based on 2D ConvNet inflation. I3D [33] demonstrates that 3D convolutional networks can be pre-trained, which aids in pushing state-of-the-art action recognition further. Deep learning methods work under a supervised methodology implicating considerable high-quality labels [86]. Nevertheless, data notation is a time-intensive and costly process [86]. Pretraining is a frequent technique to reduce the required processing time and amount of labeled data [86]. Consequently, researchers explored the concept of 2D CNN inflation further [87][88], yielding innovative architectures such as R(2+1)D [89].

References

1. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. A combined multiple action recognition and summarization for surveillance video sequences. Appl. Intell. 2021, 51, 690–712.
2. Badue, C.; Guidolini, R.; Carneiro, R.V.; Azevedo, P.; Cardoso, V.B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T.M.; Mutz, F.; et al. Self-driving cars: A survey. Expert Syst. Appl. 2021, 165, 113816.
3. Martinez, M.; Rybok, L.; Stiefelhagen, R. Action recognition in bed using BAMs for assisted living and elderly care. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 329–332.

4. Pareek, P.; Thakkar, A. A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. Artif. Intell. Rev. 2021, 54, 2259–2322.
5. Nweke, H.F.; Teh, Y.W.; Mujtaba, G.; Al-Garadi, M.A. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Inf. Fusion 2019, 46, 147–170.
6. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis. Comput. 2021, 106, 104078.
7. Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Future Gener. Comput. Syst. 2019, 96, 386–397.
8. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. 2013, 117, 633–659.
9. Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Mavroudi, E.; Katsamanis, A.; Tsiami, A.; Maragos, P. Multimodal human action recognition in assistive human-robot interaction. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2702–2706.
10. Meng, Y.; Panda, R.; Lin, C.C.; Sattigeri, P.; Karlinsky, L.; Saenko, K.; Oliva, A.; Feris, R. AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition. arXiv 2021, arXiv:2102.05775.
11. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Trans. Pattern Anal. Mach. Intell. 2021, 44, 3316–3333.
12. Ullah, A.; Muhammad, K.; Hussain, T.; Baik, S.W. Conflux LSTMs network: A novel approach for multi-view action recognition. Neurocomputing 2021, 435, 321–329.
13. Borges, P.V.K.; Conci, N.; Cavallaro, A. Video-based human behavior understanding: A survey. IEEE Trans. Circuits Syst. Video Technol. 2013, 23, 1993–2008.
14. Jegham, I.; Khalifa, A.B.; Alouani, I.; Mahjoub, M.A. Vision-based human action recognition: An overview and real world challenges. Forensic Sci. Int. Digit. Investig. 2020, 32, 200901.
15. Cherla, S.; Kulkarni, K.; Kale, A.; Ramasubramanian, V. Towards fast, view-invariant human action recognition. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, Alaska, 23–28 June 2008; pp. 1–8.
16. Stergiou, N.; Decker, L.M. Human movement variability, nonlinear dynamics, and pathology: Is there a connection? Hum. Mov. Sci. 2011, 30, 869–888.

17. Matsumoto, D. Cultural similarities and differences in display rules. Motiv. Emot. 1990, 14, 195–214.
18. Zhu, Y.; Li, X.; Liu, C.; Zolfaghari, M.; Xiong, Y.; Wu, C.; Zhang, Z.; Tighe, J.; Manmatha, R.; Li, M. A comprehensive study of deep video action recognition. arXiv 2020, arXiv:2012.06567.
19. Huang, D.A.; Ramanathan, V.; Mahajan, D.; Torresani, L.; Paluri, M.; Fei-Fei, L.; Niebles, J.C. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7366–7375.
20. Bradski, G.; Kaehler, A. Learning OpenCV: Computer Vision with the OpenCV Library; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2008.
21. Ramanathan, M.; Yau, W.Y.; Teoh, E.K. Human action recognition with video data: Research and evaluation challenges. IEEE Trans. -Hum.-Mach. Syst. 2014, 44, 650–663.
22. Yang, W.; Wang, Y.; Mori, G. Recognizing human actions from still images with latent poses. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2030–2037.
23. Piergiovanni, A.; Ryoo, M.S. Recognizing actions in videos from unseen viewpoints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4124–4132.
24. Kaur, A.; Rao, N.; Joon, T. Literature Review of Action Recognition in the Wild. arXiv 2019, arXiv:1911.12249.
25. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
26. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 4037–4058.
27. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. Technologies 2020, 9, 2.
28. Kumar Dwivedi, S.; Gupta, V.; Mitra, R.; Ahmed, S.; Jain, A. Protogan: Towards few shot learning for action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
29. Mittelstadt, B.D.; Floridi, L. The ethics of big data: Current and foreseeable issues in biomedical contexts. Sci. Eng. Ethics 2016, 22, 303–341.
30. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv 2012, arXiv:1212.0402.

31. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
32. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. Image Vis. Comput. 2017, 60, 4–21.
33. 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.
34. Alayrac, J.B.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; Zisserman, A. Self-Supervised MultiModal Versatile Networks. NeurIPS 2020, 2, 7.
35. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The something something video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850.
36. 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. 2022, 45, 3200–3225.
37. 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.
38. Bobick, A.; Davis, J. An appearance-based representation of action. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–19 August 1996; Volume 1, pp. 307–312.
39. Huang, C.P.; Hsieh, C.H.; Lai, K.T.; Huang, W.Y. Human action recognition using histogram of oriented gradient of motion history image. In Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Beijing, China, 21–23 October 2011; pp. 353–356.
40. 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; pp. 9–14.
41. Poppe, R. A survey on vision-based human action recognition. Image Vis. Comput. 2010, 28, 976–990.
42. Rodríguez-Moreno, I.; Martínez-Otzeta, J.M.; Sierra, B.; Rodriguez, I.; Jauregi, E. Video activity recognition: State-of-the-art. Sensors 2019, 19, 3160.
43. Laptev, I. On space-time interest points. Int. J. Comput. Vis. 2005, 64, 107–123.

44. Willems, G.; Tuytelaars, T.; Van Gool, L. An efficient dense and scale-invariant spatio-temporal interest point detector. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 650–663.
45. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos "in the wild". In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1996–2003.
46. Camarena, F.; Chang, L.; Gonzalez-Mendoza, M. Improving the Dense Trajectories Approach Towards Efficient Recognition of Simple Human Activities. In Proceedings of the 2019 7th International Workshop on Biometrics and Forensics (IWBF), Cancun, Mexico, 2–3 May 2019; pp. 1–6.
47. Camarena, F.; Chang, L.; Gonzalez-Mendoza, M.; Cuevas-Ascencio, R.J. Action recognition by key trajectories. Pattern Anal. Appl. 2022, 25, 409–423.
48. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176.
49. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3d-Gradients. Available online: https://class.inrialpes.fr/pub/klaser-bmvc08.pdf (accessed on 30 January 2023).
50. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 2008, 110, 346–359.
51. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 915–928.
52. Norouznezhad, E.; Harandi, M.T.; Bigdeli, A.; Baktash, M.; Postula, A.; Lovell, B.C. Directional space-time oriented gradients for 3d visual pattern analysis. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 736–749.
53. Tuzel, O.; Porikli, F.; Meer, P. Region covariance: A fast descriptor for detection and classification. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 589–600.
54. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 428–441.
55. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2593–2600.

56. Messing, R.; Pal, C.; Kautz, H. Activity recognition using the velocity histories of tracked keypoints. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 104–111.
57. Matikainen, P.; Hebert, M.; Sukthankar, R. Trajectons: Action recognition through the motion analysis of tracked features. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 29 September–2 October 2009; pp. 514–521.
58. Chang, L.; Pérez-Suárez, A.; Hernández-Palancar, J.; Arias-Estrada, M.; Sucar, L.E. Improving visual vocabularies: A more discriminative, representative and compact bag of visual words. Informatica 2017, 41, 333–347.
59. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 143–156.
60. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 581–595.
61. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Computer Vision, IEEE International Conference, Nice, France, 13–16 October 2003; p. 1470.
62. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
63. Zhou, X.; Yu, K.; Zhang, T.; Huang, T.S. Image classification using super-vector coding of local image descriptors. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 141–154.
64. Rabiner, L.R.; Juang, B.-H. Fundamentals of Speech Recognition; PTR Prentice Hall: Englewood Cliffs, NJ, USA, 1993; Volume 14.
65. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 2001, 23, 257–267.
66. Natarajan, P.; Nevatia, R. Online, real-time tracking and recognition of human actions. In Proceedings of the 2008 IEEE Workshop on Motion and Video Computing, Copper Mountain, CO, USA, 8–9 January 2008.
67. Oliver, N.M.; Rosario, B.; Pentland, A.P. A Bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Mach. Intell. 2000, 22, 831–843.

68. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. Science 2006, 313, 504–507.

69. Pearson, K. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 1901, 2, 559–572.

70. Fisher, R.A. The use of multiple measurements in taxonomic problems. Ann. Eugen. 1936, 7, 179–188.

71. 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 Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.

72. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. arXiv 2014, arXiv:1406.2199.

73. Goodale, M.A.; Milner, A.D. Separate visual pathways for perception and action. Trends Neurosci. 1992, 15, 20–25.

74. Ye, H.; Wu, Z.; Zhao, R.W.; Wang, X.; Jiang, Y.G.; Xue, X. Evaluating two-stream CNN for video classification. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 435–442.

75. 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.

76. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. Appl. Soft Comput. 2020, 86, 105820.

77. Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Two stream lstm: A deep fusion framework for human action recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 177–186.

78. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. Signal Process. Image Commun. 2019, 71, 76–87.

79. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7024–7033.

80. Ikizler-Cinbis, N.; Sclaroff, S. Object, scene and actions: Combining multiple features for human action recognition. In Proceedings of the European Conference on Computer Vision, Heraklion,

Greece, 5–11 September 2010; pp. 494–507.

81. He, D.; Li, F.; Zhao, Q.; Long, X.; Fu, Y.; Wen, S. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. arXiv 2018, arXiv:1806.10319.

82. Hsiao, J.; Li, Y.; Ho, C. Language-guided Multi-Modal Fusion for Video Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3158–3162.

83. 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.

84. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Cambridge, UK, 23–26 August 2004; Volume 3, pp. 32–36.

85. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.

86. Tao, L.; Wang, X.; Yamasaki, T. Pretext-Contrastive Learning: Toward Good Practices in Self-supervised Video Representation Leaning. arXiv 2020, arXiv:2010.15464.

87. Wang, X.; Miao, Z.; Zhang, R.; Hao, S. I3d-lstm: A new model for human action recognition. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kazimierz Dolny, Poland, 21–23 November 2019; IOP Publishing: Bristol, UK, 2019; Volume 569, p. 032035.

88. Liu, G.; Zhang, C.; Xu, Q.; Cheng, R.; Song, Y.; Yuan, X.; Sun, J. I3D-Shufflenet Based Human Action Recognition. Algorithms 2020, 13, 301.

89. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.

Retrieved from https://encyclopedia.pub/entry/history/show/97946