Article

A Concise Overview of the Vision-based Human Action Recognition Field

Fernando Camarena * , Miguel Gonzalez-Mendoza * , Leonardo Chang , Ricardo J Cuevas-Ascencio

Posted Date: 3 February 2023

doi: 10.20944/preprints202302.0050.v1

Keywords: video-based human action recognition; Action Recognition; Deep Learning Methods; handcrafted Methods; Human Action; Overview

Article

A Concise Overview of the Vision-Based Human Action Recognition Field

Fernando Camarena*, Miguel Gonzalez-Mendoza*, Leonardo Chang and Ricardo J Cuevas-Ascencio

Tecnologico de Monterrey, School of Engineering and Science, Mexico
* fernando@camarenat.com; mgonza@tec.mx

Abstract: 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. Research in human video-based human action recognition is vast and ongoing, making it difficult to assess the full scope of available methods and current trends. This survey provides an in-depth exploration of the vision-based human action recognition field, comprehensively offering the available techniques and their evolution, highlighting the cutting-edge ideas driving its development. We also analyze the most used keywords in research papers over the past years to identify trends and predict possible future directions. Hence, this concise survey helps researchers understand the breadth of existing approaches, evaluate current research trends, and stay up-to-date on potential developments.

Keywords: video-based human action recognition; action recognition; deep learning methods; handcrafted methods; human action; overview

1. Introduction

Artificial intelligence (AI) redefines our understanding of the world by enabling high-impact applications such as intelligent video surveillance systems [1], self-driving vehicles [2], assisted living [3], and human-computer interface systems [4]. These applications frequently require a single core task: video-based human action recognition, an active research field with continuous improvements. Its principal purpose is to acquire insight into what a subject is doing in a video [5–7].

On the one hand, current research points out numerous directions, including effectively combining multi-modal information [8,9], learning without annotated labels [10], training with reduced data points [10,11], and exploring novel architectures [12,13].

On the other hand, recent surveys focused on providing an in-depth review of a specific contribution. For example, [14] categorized standard vision-based human action recognition datasets, whereas [15] analyzes the classification performance of standard action recognition algorithms. [16] was one of the first surveys to incorporate deep learning algorithms, providing a comprehensive overview of the datasets employed. [17] offers a comprehensive taxonomy centered on deep learning methodologies. While [18] and [19] concentrate on its applicability. Finally, [20] delved into future directions of the field.

Given the enormous knowledge and many directions, it can take time to introduce oneself to the subject thoroughly. As a result, in this work, we focus on offering an intuitive picture of the vision-based human action recognition field and describing how the approaches evolved and the intuitive notions that underpin them. Also, we gathered a collection of abstracts utilizing the papers-with-code portal services [21] and analyzed them to determine the most prevalent keywords by year. Thus, we can evaluate the most-discussed themes, the nature of their evolution, and potential prospects for the near future.

We divide the rest of the document as follows: in Section 2.1, we define the concept of human action. Then, we speak about the vision-based human action recognition taxonomy in 2.2. In Section 3, we explore the common terms in the paper's abstract to find future directions. Finally, we jump to conclusions in 4.

2. Action Recognition Taxonomy

The aim of this section is two-fold. First, in Section 2.1, we explain what this work understands as action. Finally, we introduce the action recognition taxonomy in Section 2.2.

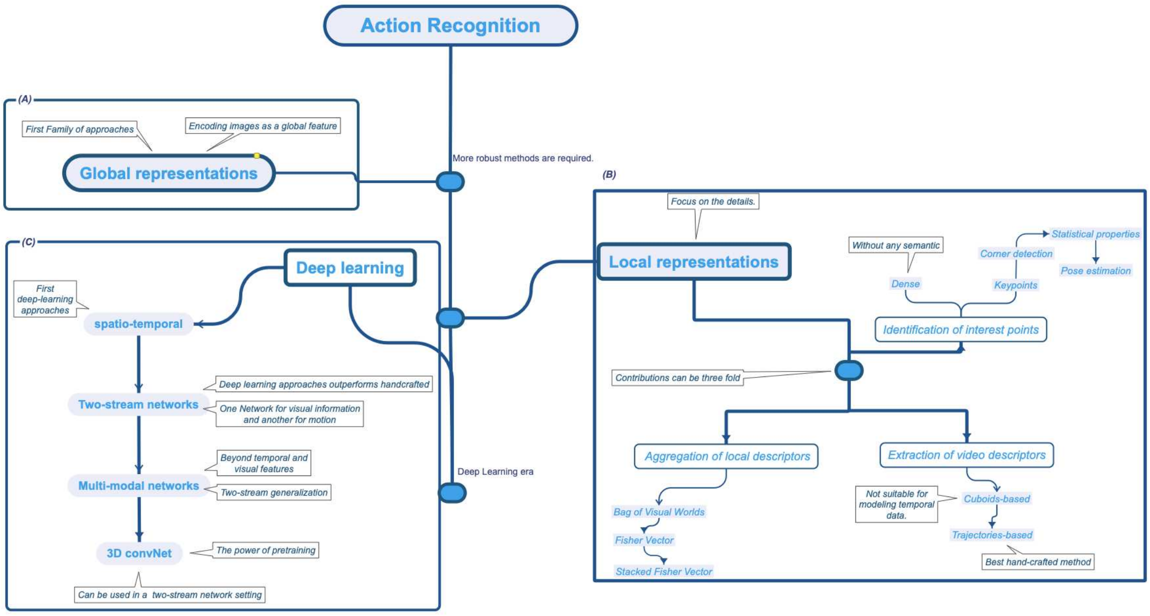
Action Recognition
(A)
First Family of approaches
Encoding images as a global feature
Global representations
More robust methods are required.
(B)
Focus on the details.
Local representations
Without any semantic
Dense
Corner detection
Keypoints
Statistical properties
Pose estimation
Identification of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

Figure 1. 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.

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 we create a picture of a man sprinting, we may build a more dynamic image by focusing on the person's legs, as shown in Figure 2. We unconsciously associate a particular message with a sequence of movements, which is what we call "an action" [22]. The human action recognition goal is to build approaches that can understand the encoded message in the sequence of gestures.

Although it is a natural talent for a person to recognize what others do, it is not an easy assignment for a computer since it faces numerous challenges [16], including the considerable variability in how a person can act. For example, an older person's running movement will differ from that of a younger person. In addition, it is challenging for an individual to perform the same activity in precisely the same way. Finally, actions can be influenced by cultural context, as in the case of greeting.

Similarly, there are restrictions on the camera's capabilities [16]: first, various camera placements achieve infinite action perspectives, which may generate object occlusions or limited viewpoints. Another limitation is the video resolution, which may not provide enough pixel information to identify the target. External factors, such as rain or wind, can also harm the camera's image quality.

doi:10.20944/preprints202302.0050.v1

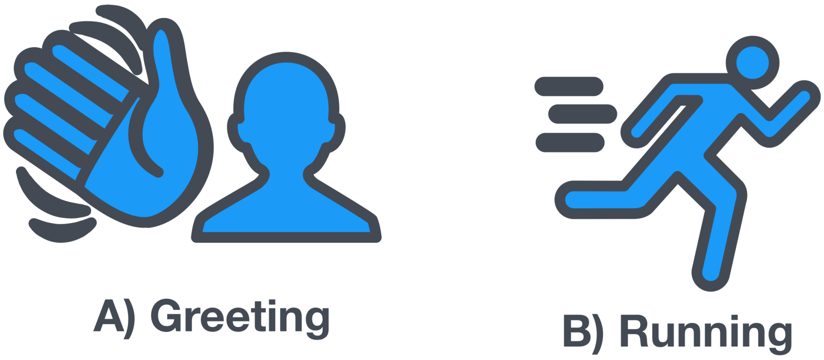

Figure 2. We instinctively associate a sequence of gestures with an action. For example, when we think of the action greeting, we might think of the typical hand wave. On the contrary, imagining a person running will create a more dynamic scene with movement centered on the legs. An action can be defined as a sequence of gestures that encode a message.

2.2. Human Action Recognition Taxonomy

In this section, we describe the approaches to recognizing human actions. First, in Figure 1, we provide a formal description of the human action recognition taxonomy and summarize the evolution of the methods and some keypointsld. We categorized the methods into two main branches. On the one hand, we have the family of procedures whose feature engineering is performed manually, knowns as "handcrafted methods." On the other hand, there are methods whose representation is learned, as in the case of deep learning techniques.

2.2.1. handcrafted Approaches

As illustrated in Figure 1, handcrafted approaches entail manually engineering features, i.e., we must develop characteristics that support a computer to understand the human action concept, analogous to creating a translator app for those who do not speak the same language.

Global representations [16], shown in Figure 1 (A), are the first attempt to recognize actions whose intuition is to represent the content as a global feature. For example, as individuals, we can discern a person's actions by looking at their 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. Hence, a novel approach was needed. Among the most relevant methods are Motion Energy Image (MEI) [23], Motion History Image (MHI) [24], silhouettes [25], and Space-Time Volume (STV) [26].

The world is full of little details that are difficult to capture using the "big picture." Intuitively, as humans, to discover those little secrets, we need to explore, focus on the details, and zoom in on the regions of interest, which is the idea behind local representations [17,27], shown in Figure 1 (B). Local representations seek to extract descriptors from multiple regions of the video to get 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) [28] and hessian detector [29].

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. [30] propose to use of statistical properties as a pruning method.

Camarena et al. [31,32] 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 [33], which consists of placing points without semantics. Dense sampling increases the classification accuracy, but it is computationally expensive [31]. Besides, noise injected by other motion sources can affect the classifier's performance [31,32].

Once we have determined which regions to analyze, we must extract the corresponding region description. Visual and motion data are essential for accurately characterizing an action [31]. In this regard, the typical approach combines several descriptors to have a complete perspective of the target action. Regarding the visual information, we have histogram of oriented gradients 3D (HOG3D) [34], speed-up robust features (SURF) [35], 3D SURF [29], and pixel pattern methods [36–38]. On the other hand, descriptors that focus on motion information include histogram of Oriented Flow (HOF) [39], Motion Boundaries Histogram (MBH) [33], and MPEG flow [40].

Capturing motion information is a complex task; videos are composed of images in which the target person moves or changes location over time [33]. The naive method uses cuboids, which utilize static neighborhood patterns throughout time. However, cuboids are not suitable for modeling the temporal information of an object. its natural evolution were trajectory-based approaches [33,41,42] that rapidly becoming one of the most used methods [17,32].

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 [17]. Although there are several efficient optical flow algorithms, their application at different points of interest can be computationally expensive [32]. To reduce the computational time 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, we can use homographies [17] for reducing the motion's camera; on the other hand, pose estimation [32] 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) [43], Fisher Vectors [44], stacked fisher vector (SFV) [45], vector quantization (VQ) [46], vector of locally aggregated descriptors (VLAD) [47], super vector encoding (SVC) [48]. Among the handcrafted approaches, it is popularly referred to that FV and SFV, along with dense trajectories, achieve the best classification performance [16].

2.2.2. Deep Learning Approaches

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

Andrej et al. [49] 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 [20]. 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 [20]. 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, we use different parts of our brain to process the appearance and motion signals we perceive[50,51]. This understanding can be used for human action recognition, as suggested by [50]. 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 [50]. By combining spatial and temporal information, the network can effectively process visual information.

Due to the comparable performance of two-stream networks to trajectory-based methods, [20], 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 [52], 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 [52].

Because of the temporal nature of videos, researchers investigated the use of recurrent neural networks (RNN) [53] and long-term short-term memory (LSTM) [54,55] as the temporal stream for two-stream approaches. As proven by Ma et al. [56], pre-segmented data is necessary to explode 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 [20].

A generalization of two-stream networks is multi-stream networks [20], which describe actions using additional modalities like pose estimation [57], object information [58], audio signals [59], text transcriptions [60], and depth information [61].

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 we have, our network's performance will be higher. However, the datasets employed in vision-based human action recognition [62], [63], [64] do not have the scale that requires a deep learning model [65]. Do not dispose 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.[65] 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 [65] 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 [66]. Nevertheless, data notation is a time-intensive and costly process [66]. Pretrained is a frequent technique to reduce the required processing time and amount of labeled data [66]. Consequently, researchers explored the concept of 2D CNN inflation further [67,68], yielding innovative architectures like R(2+1)D [69].

Current research in vision-based human action recognition has several directions. First, novel architectures such as visual transformers have been ported to action recognition [69,70]. Second, there is a need for novel training methods like self-supervised learning (SSL) [71], which is a novel training technique that generates a supervisory signal from unlabeled data, thus eliminating the need for human-annotated labels. Third, few-shot learning action recognition is also being investigated [72].

Most of the architectures described are known as discriminative approaches [73], but there is another family of deep learning methods based on generative techniques [73]. Its core idea is based on the popular phrase "if I cannot create it, then I do not understand it" [74]. auto-encoders [75], variational autoencoders [76], and adversarial networks (GAN) [77] are examples of this approach.

3. Human Action Recognition Trends: Metadata Analysis

To gain a deeper understanding of the vision-based human action recognition discipline, we scanned the paper's abstracts to identify the most frequently used terms. Identifying the commonly used keyword allowed us to evaluate yearly trends regarding topics, techniques, and potential research directions. The extraction of the paper's abstracts could be accomplished in two ways: through the use of web scraping techniques [78] and web services such as programmatic interfaces (APIs) [79].

Web scraping techniques attempt to replicate human behavior by reading and parsing websites [78]. The significance of this type of technique [80] is that it permits the extraction of data even when there is no programmatic interface, such as in the case of Google Scholar [81]. However, a disadvantage

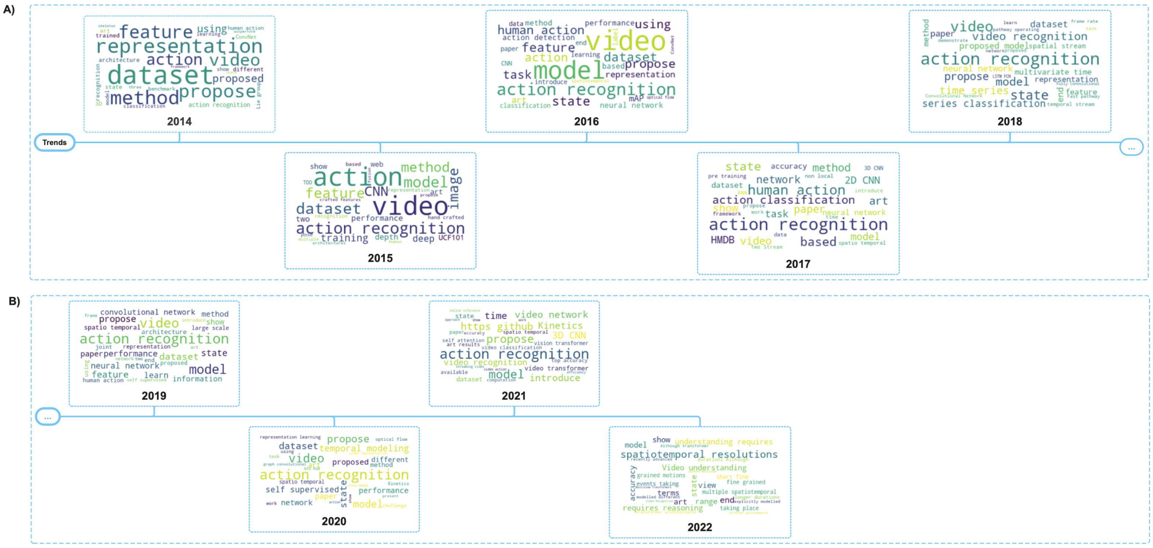

Figure 3. To gain a deeper understanding of the vision-based human action recognition field, we analyzed abstracts from papers published in 2014-2022. Popular keywords included CNN, performance, ConvNet architecture, and handcrafted in 2014 and 2015. Then, 3D CNN, pretraining, and the two-stream technique became widely used. Subsequently, there was a shift towards few-shot and self-supervised learning approaches to reduce the requirement for labeled data. Finally, novel architectures such as transforming and working with fine-grained data are increasingly popular.

of web scraping [78] is that some websites apply security measures like low-number of requests or the use of captchas, making it challenging to collect large-dimensional information. Also, there is a site structure dependency, which means that if the design of the website changes, the developed program could be affected.

Application programming interfaces (APIs) [79] are another method for obtaining information because they enable integration and give capabilities for extracting pre-defined attributes. Another advantage over web scraping is that the quality of structure and cleanliness is believed to be higher and allows for greater modularity and scalability. There are several web services to extract research data, including arXiv [82], papers-with-code [21], Scopus API [83], and Science Direct API [84]. We found low-bias information while using the papers-with-code service, hence using it in this research. However, in the future, we might combine more than one service to make a more robust information base.

To extract the paper-related information, we request the query "action-recognition-in-video" through the task retrieval module of the service, which returned 1600 paper instances. Then, we split the abstracts by year and selected the 30 most frequently used words. The visualization is shown in Figure 3. The analysis helps us identify the key terms used to gather information about the topics that researchers face.

As expected, "CNN", "performance", "ConvNet", "architecture", and "handcrafted" were the most often used in 2014 and 2015. The keywords suggest that finding the optimal architecture for human action recognition was a substantial issue. We previously mentioned that the first attempts to recognize human actions using neural networks did not perform as well as their handcrafted counterpart due to the high dependence on temporal information. Therefore, finding the architecture that would take advantage of the neural network's ability to extract representations without manual feature engineering makes sense as the main issue. This intuition is aligned with the keywords presented in 2016 and 2017, where "optical flow", "3D CNN", "pre training", and "two-stream" are techniques that enable deep learning methods to outperform handcrafted approaches.

This tendency continued in 2018 and 2019, with the keywords strongly aligned to model the temporal information. Also, some relevant insights are that "cross-modal" and "self-supervised learning" are gaining popularity, which is one of the most promising research paths in our days [85]. Then, in 2020 and 2021, the terms "self-supervised learning" and "few-shot learning" increased in popularity, indicating a desire to reduce the amount of labeled data required. Finally, in 2022, some keywords appear that may show future directions, like 'graph convolutional neural networks, "understanding," and "reasoning". The keywords may indicate that current research not only focuses on the learning process but also ensures that the model truly understands what it is learning, opening the door to a more robust generalization and capabilities to learn new concepts. Furthermore, the granularity of actions and long-term behavior becomes a focused path for research. Finally, we see the application of vision transformers to vision-based human action recognition.

4. Conclusions

In this work, we thoroughly examine the field of study of human action recognition, delving into its history and evolution, and providing an in-depth overview of the method, from traditional handcrafted techniques to modern deep learning architectures. We present this information comprehensively, easy-to-understand manner, emphasizing how ideas in the field have evolved.

Although neural networks enabled a breakthrough in recognition task performance and removed the need for manual feature engineering, the process still has some limitations. Neural networks work under a supervised methodology, which requires a vast number of high-quality labels. Assuming unlabeled data is available. Also, the process of annotating data is a time-consuming and expensive operation.

To understand the current directions of the vision-based human action recognition field, we examined the scientific publication abstracts in which we found frequent terms among them. The keywords help to understand the yearly issues and provide clues to future directions.

To reduce the dependency on human-annotated labels, one of the research lines to pay attention to is self-supervised learning, which aims to generate a natural supervisory signal derived directly from the unlabeled data. In the same direction, few-shot learning techniques aim to generalize novel object classes in addition to new instances, reducing the amount of data required.

Reasoning instead of only learning is a novel trend in human action recognition, which can be helpful in a few-shot environment and when facing fine-grained activities. Creating models that perform well with fine-grained actions is a challenging problem because datasets cannot represent every possible variation of an action. Nevertheless, due to the increasing demand for specialized action comprehension in real-world applications, interest in this method has been growing.

Author Contributions: Conceptualization, Leonardo Chang; Formal analysis, Fernando Camarena; Investigation, Fernando Camarena; Methodology, Fernando Camarena and Miguel Gonzalez-Mendoza; Project administration, Miguel Gonzalez-Mendoza; Resources, Miguel Gonzalez-Mendoza; Supervision, Miguel Gonzalez-Mendoza and Leonardo Chang; Validation, Miguel Gonzalez-Mendoza; Writing – original draft, Fernando Camarena and Ricardo J Cuevas-Ascencio; Writing – review & editing, Fernando Camarena and Ricardo J Cuevas-Ascencio.

Funding: F.Camarena gratefully acknowledges the scholarship no. 815917 from CONACyT to pursue his postgraduate studies. The scholarship had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest: he author declares that he has no conflict of interest.

References

1. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence* **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 Systems with Applications* **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). IEEE, 2015, pp. 329–332.
4. 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). IEEE, 2016, pp. 2702–2706.
5. 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 preprint arXiv:2102.05775* **2021**.
6. 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 Transactions on Pattern Analysis and Machine Intelligence* **2021**.
7. 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.
8. 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.
9. Valverde, F.R.; Hurtado, J.V.; Valada, A. There is More than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking with Sound by Distilling Multimodal Knowledge. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11612–11621.
10. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* **2021**.
11. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733* **2020**.
12. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169* **2021**.
13. Hafiz, A.M.; Parah, S.A.; Bhat, R.U.A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. *arXiv preprint arXiv:2106.07550* **2021**.
14. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* **2013**, *117*, 633–659.
15. Cheng, G.; Wan, Y.; Saudagar, A.N.; Namuduri, K.; Buckles, B.P. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964* **2015**.
16. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A review on human activity recognition using vision-based method. *Journal of healthcare engineering* **2017**, *2017*.
17. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image and vision computing* **2017**, *60*, 4–21.
18. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *International Journal of Computer Vision* **2022**, *130*, 1366–1401.
19. Lei, Q.; Du, J.X.; Zhang, H.B.; Ye, S.; Chen, D.S. A survey of vision-based human action evaluation methods. *Sensors* **2019**, *19*, 4129.
20. 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 preprint arXiv:2012.06567* **2020**.
21. rstojnic, Viktor Kerkez, J.B. API Client for paperswithcode.com. https://github.com/paperswithcode/paperswithcode-client, 2021.
22. Borges, P.V.K.; Conci, N.; Cavallaro, A. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology* **2013**, *23*, 1993–2008.
23. Bobick, A.; Davis, J. An appearance-based representation of action. In Proceedings of the Proceedings of 13th International Conference on Pattern Recognition. IEEE, 1996, Vol. 1, pp. 307–312.
24. 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. IEEE, 2011, pp. 353–356.
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-Workshops. IEEE, 2010, pp. 9–14.

26. Poppe, R. A survey on vision-based human action recognition. *Image and vision computing* **2010**, *28*, 976–990.
27. Rodríguez-Moreno, I.; Martínez-Otzeta, J.M.; Sierra, B.; Rodriguez, I.; Jauregi, E. Video activity recognition: State-of-the-art. *Sensors* **2019**, *19*, 3160.
28. Laptev, I. On space-time interest points. *International journal of computer vision* **2005**, *64*, 107–123.
29. 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. Springer, 2008, pp. 650–663.
30. 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. IEEE, 2009, pp. 1996–2003.
31. 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). IEEE, 2019, pp. 1–6.
32. Camarena, F.; Chang, L.; Gonzalez-Mendoza, M.; Cuevas-Ascencio, R.J. Action recognition by key trajectories. *Pattern Analysis and Applications* **2022**, *25*, 409–423.
33. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 3169–3176.
34. Klaser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3d-gradients. In Proceedings of the BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008, pp. 275–1.
35. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Computer vision and image understanding* **2008**, *110*, 346–359.
36. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* **2007**, *29*, 915–928.
37. 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. Springer, 2012, pp. 736–749.
38. Tuzel, O.; Porikli, F.; Meer, P. Region covariance: A fast descriptor for detection and classification. In Proceedings of the European conference on computer vision. Springer, 2006, pp. 589–600.
39. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In Proceedings of the European conference on computer vision. Springer, 2006, pp. 428–441.
40. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2593–2600.
41. 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. IEEE, 2009, pp. 104–111.
42. 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. IEEE, 2009, pp. 514–521.
43. 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*.
44. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European conference on computer vision. Springer, 2010, pp. 143–156.
45. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. In Proceedings of the European Conference on Computer Vision. Springer, 2014, pp. 581–595.
46. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the null. IEEE, 2003, p. 1470.
47. 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. IEEE Computer Society, 2010, pp. 3304–3311.
48. 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. Springer, 2010, pp. 141–154.
49. 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 Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

50. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* **2014**.
51. Goodale, M.A.; Milner, A.D. Separate visual pathways for perception and action. *Trends in neurosciences* **1992**, *15*, 20–25.
52. 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 Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 435–442.
53. 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.
54. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. *Applied soft computing* **2020**, *86*, 105820.
55. 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). IEEE, 2017, pp. 177–186.
56. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication* **2019**, *71*, 76–87.
57. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7024–7033.
58. 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. Springer, 2010, pp. 494–507.
59. 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 preprint arXiv:1806.10319* **2018**.
60. Hsiao, J.; Li, Y.; Ho, C. Language-guided Multi-Modal Fusion for Video Action Recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3158–3162.
61. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *Journal of real-time image processing* **2016**, *12*, 155–163.
62. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: a local SVM approach. In Proceedings of the Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, Vol. 3, pp. 32–36.
63. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* **2012**.
64. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4305–4314.
65. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
66. Tao, L.; Wang, X.; Yamasaki, T. Pretext-Contrastive Learning: Toward Good Practices in Self-supervised Video Representation Leaning. *arXiv* **2020**, [2010.15464].
67. 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. IOP Publishing, 2019, Vol. 569, p. 032035.
68. Liu, G.; Zhang, C.; Xu, Q.; Cheng, R.; Song, Y.; Yuan, X.; Sun, J. I3D-Shufflenet Based Human Action Recognition. *Algorithms* **2020**, *13*, 301.
69. 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 Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
70. Chen, J.; Ho, C.M. MM-ViT: Multi-modal video transformer for compressed video action recognition. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1910–1921.
71. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **2020**, *9*, 2.

72. Kumar Dwivedi, S.; Gupta, V.; Mitra, R.; Ahmed, S.; Jain, A. Protogan: Towards few shot learning for action recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
73. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications* **2020**, *79*, 30509–30555.
74. Gleick, J. *Genius: The life and science of Richard Feynman*; Vintage, 1993.
75. Xing, L.; Qin-kun, X. Human action recognition using auto-encode and pnn neural network. *Software Guide* **2018**, *1*, 1608–01529.
76. Mishra, A.; Pandey, A.; Murthy, H.A. Zero-shot learning for action recognition using synthesized features. *Neurocomputing* **2020**, *390*, 117–130.
77. Ahsan, U.; Sun, C.; Essa, I. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint arXiv:1801.07230* **2018**.
78. Zhao, B. Web scraping. *Encyclopedia of big data* **2017**, pp. 1–3.
79. de Souza, C.R.; Redmiles, D.; Cheng, L.T.; Millen, D.; Patterson, J. Sometimes you need to see through walls: a field study of application programming interfaces. In Proceedings of the Proceedings of the 2004 ACM conference on Computer supported cooperative work, 2004, pp. 63–71.
80. Glez-Peña, D.; Lourenço, A.; López-Fernández, H.; Reboiro-Jato, M.; Fdez-Riverola, F. Web scraping technologies in an API world. *Briefings in bioinformatics* **2014**, *15*, 788–797.
81. Google Scholar, Official website. https://scholar.google.com/intl/es/scholar/about.html.
82. ArXiv API. https://arxiv.org/help/api/user-manual.
83. Scopus APIs. https://dev.elsevier.com/sc_apis.html.
84. ScienceDirect APIs. https://www.elsevier.com/solutions/sciencedirect/librarian-resource-center/api.
85. Hendrycks, D.; Mazeika, M.; Kadavath, S.; Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems* **2019**, *32*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.