applied sciences

MDPI

Review

On the Use of Deep Learning for Video Classification

Atiq ur Rehman [1,2,*], Samir Brahim Belhaouari [3], Md Alamgir Kabir [1] and Adnan Khan [3]

[1] Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Technology, Mälardalen University, Högskoleplan 1, 722 20 Västerås, Sweden

[2] Department of Electrical and Computer Engineering, Pak Austria Fachhochschule, Institute of Applied Sciences and Technology, Haripur 22621, Pakistan

[3] Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar

* Correspondence: atiq.ur.rehman@mdu.se or atiqjadoon@gmail.com

Abstract: The video classification task has gained significant success in the recent years. Specifically, the topic has gained more attention after the emergence of deep learning models as a successful tool for automatically classifying videos. In recognition of the importance of the video classification task and to summarize the success of deep learning models for this task, this paper presents a very comprehensive and concise review on the topic. There are several existing reviews and survey papers related to video classification in the scientific literature. However, the existing review papers do not include the recent state-of-art works, and they also have some limitations. To provide an updated and concise review, this paper highlights the key findings based on the existing deep learning models. The key findings are also discussed in a way to provide future research directions. This review mainly focuses on the type of network architecture used, the evaluation criteria to measure the success, and the datasets used. To make the review self-contained, the emergence of deep learning methods towards automatic video classification and the state-of-art deep learning methods are well explained and summarized. Moreover, a clear insight of the newly developed deep learning architectures and the traditional approaches is provided. The critical challenges based on the benchmarks are highlighted for evaluating the technical progress of these methods. The paper also summarizes the benchmark datasets and the performance evaluation matrices for video classification. Based on the compact, complete, and concise review, the paper proposes new research directions to solve the challenging video classification problem.

Keywords: automatic video classification; deep learning; handcrafted features; video processing

Citation: Rehman, A.u.; Belhaouari, S.B.; Kabir, M.A.; Khan, A. On the Use of Deep Learning for Video Classification. *Appl. Sci.* **2023**, *13*, 2007. https://doi.org/10.3390/app13032007

Academic Editors: Zahid Mehmood Jehangiri, Mohsin Shahzad and Uzair Khan

Received: 2 December 2022
Revised: 21 January 2023
Accepted: 30 January 2023
Published: 3 February 2023

1. Introduction

The task of automatically classifying videos has become very successful recently. Particularly, the subject has drawn increased interest since deep learning models became an effective method for automatically classifying videos. The importance of the accurate video classification task can be realized by the large amount of video data available online. People around the world generate and consume a huge amount of video content. Currently, on YouTube only, over 1 billion hours of video are being watched by different people every single day. In recognition to the importance of the video classification task, a combined effort is being made by researchers for proposing an accurate video classification framework. Companies such as Google AI are investing in different competitions to solve the challenging problem under constrained conditions. To further advance the progress of the automatic video classification task, Google AI has released a public dataset called YouTube-8M with millions of video features and more than 3700 labels. All these efforts being made demonstrate the need for a powerful video classification model.

An artificial neural network (ANN) is an algorithm based on interconnected nodes to recognize the relationships in a set of data. Algorithms based on ANNs have shown a great success in modeling both the lineßar and the non-linear relationships in the underlying data. Due to the huge success rate of these algorithms, they are extensively being used for different real-time applications [1–4]. Moreover, with an increase in the availability of huge datasets, the deep learning models have specifically shown a significant improvement in the classification of videos. This paper reviews studies based on deep learning approaches for video classification.

Contribution

There are several existing reviews and survey papers related to video classification in the scientific literature. Some of the recent works are summarized here in Table 1. However, these review papers do not include the recent state-of-art works, and they have some limitations. In the following text, the limitations and highlights of these works are discussed.

Table 1. Summary of recent related works.

Reference	Year	Coverage	Highlights	Drawbacks
A. Anusya [5]	2020	2014–2019	Video classification, tagging, and clustering.	Not comprehensive and lacks concise information.
Rani et al. [6]	2020	2001–2016	Text, audio, and visual modalities for video classification.	Missing analysis of recent state-of-art approaches.
Y. Li et al. [7]	2020	2012–2019	Live sport video classification.	More specific to live sport video classification.
Md Islam et al. [8]	2021	2004–2020	Machine learning approaches for video classification.	Focus of review is not on deep learning approaches.
Ullah. H. et al. [9]	2021	2015–2020	Human activity recognition using deep learning.	Focus only on the human activity recognition.
This study	2022	2000–2022	Comprehensive deep learning review for video classification.	-

1. A more recent review was done by A. Anusya [5]; this review covers very few methods for video classification, clustering, and tagging. However, the review provided is not comprehensive and lacks concise information, coverage of topic, datasets, analysis of state-of-art approaches, and research limitations;
2. Rani et al. [6] also conducted a recent review on video classification methods, and their review covered some recent video classification approaches and summary-based description of some recent works. This review also had some limitations including the missing analysis of recent state-of-art approaches and a very limited description of topics covered;
3. Y. Li et al. [7] recently conducted a systematic and good review on live sport video classification. This review covers most of the recent works in live sport video classification, including the tools, video interaction features, and feature extraction methods. This is a comprehensive review, but the findings are not summarized in tables for research gaps and advantages and disadvantages of existing methods for a quick review. Moreover, this review is more specific to live sport video classification;
4. A recent review was also done by Md Islam et al. [8]; in this review, they included all the methods for video classification, including deep learning. However, as the focus of review is not on deep learning approaches, these methods are therefore not completely covered in this review;
5. Ullah. H. et al. [9] also conducted a recent systematic review; however, the focus of their review remained on human activity recognition;
6. Z. Wu. [10] presented a concise review on video classification specific to deep learning methods. This review provides a good description on deep learning models, feature extraction tools, benchmark dataset, and comparison of existing methods for

video classification. However, this review was conducted in the year 2016, and it does not cover the recent state-of-art deep learning methods;

7. Q. Ren [11] conducted a simple review on video classification methods; however, the techniques covered in this review are not well described, and the review also lacks in the description of research gaps, benchmark datasets, limitations of existing methods, and performance metrics.

In contrast to the existing reviews on classification of videos, this paper provides a more comprehensive, concise, and up-to-date review of deep learning approaches for video classification. In this current review, most of the recent state-of-art contributions related to the topic are analyzed and critically summarized. Deep learning is an emerging and vibrant field for the analysis of videos; therefore, we hope this review will help in stimulating future research along the line. The following are the key contributions to this review paper:

1. A summary of state-of-art, CNN-based deep learning models for image analysis;
2. An in-depth review of deep learning approaches for video classification highlighting the notable findings;
3. A summary of breakthroughs in the automatic video classification task;
4. Analysis of research trends from past towards future;
5. Description of benchmark datasets, evaluations metrics, and comparison of recent state-of-art deep learning approaches in terms of performance.

The rest of the paper is organized as follows: Section 2 reviews some existing CNNs for images; Section 3 provides an in-depth review on deep learning models for video classification; Section 4 provides a summary for benchmark datasets, evaluation metrics, and comparison of existing state-of-art methods for the video classification task; and Section 5 provides conclusion and future research directions.

2. Convolutional Neural Networks (CNN) for Image Analysis

Deep learning models, specifically convolutional neural networks (CNNs), are well known for understanding images. A number of CNN architectures are proposed and developed in the scientific literature for image analysis. Among these, the most popular architectures are LeNet-5 [12], AlexNet [13], VGGNet [14], GoogleNet [15], ResNet [16], and DenseNet [17]. The trend that follows from the formerly proposed architectures towards the recently proposed architectures is to deepen the network. A summary of these popular CNN architectures along with trend of deepening the network is shown in Figure 1, where the depth of network increases from left-most (LeNet-5) to right-most (DenseNet). Deep networks are believed to better approximate the target function and to generate better feature representation with more powerful discriminatory powers [18]. Although deeper networks are better in terms of having more discriminatory powers, the deeper networks require more data for training and more parameters to tune [19]. Finding a professionally labeled, huge dataset is still a big challenge faced by the research community, and therefore, it limits the development of deeper neural networks.

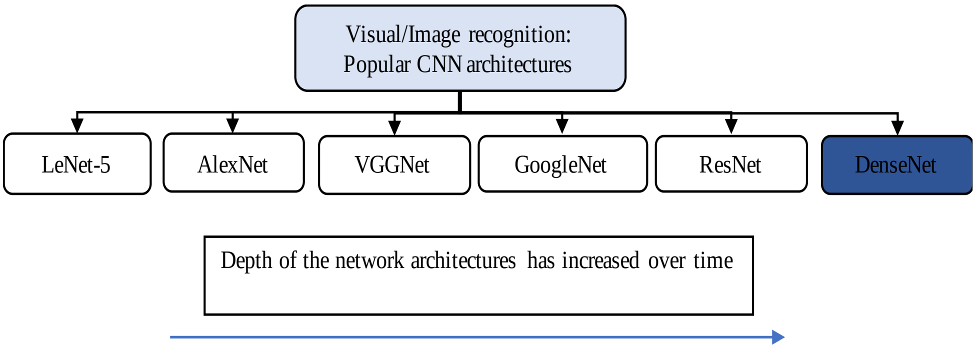

Visual/Image recognition:
Popular CNN architectures
LeNet-5
AlexNet
VGGNet
GoogleNet
ResNet
DenseNet
Depth of the network architectures has increased over time

Figure 1. State-of-art image recognition CNN networks. The trend is that the depth and discriminatory powers of network architectures increases from formerly proposed architectures towards the recently proposed architectures.

3. Video Classification

In this section, a very comprehensive and concise review for deep learning models employed in the video classification task is provided. This section covers a description on video data modalities, traditional handcrafted approaches, breakthroughs in video classification, and recent state-of-art deep learning models for video classification.

3.1. Video Data Modalities

As compared to images, videos are more challenging to understand and classify due to the complex nature of the temporal content. However, three different modalities, i.e., visual information, audio information, and text information, might be available to classify videos in contrast to image classification, where only a single visual modality can be utilized. Based on the availability of different modalities in videos, the task of classification can be categorized as a uni-modal video classification or a multi-modal video classification, as summarized in Figure 2. The existing literature has utilized both of these models for the video classification task, and it is generally believed that models utilizing multimodal data perform better than the models based on uni-modal data [20,21]. Moreover, the visual description [22] of a video works better than the text [23] and the audio [24,25] description for the classification purpose of a video.

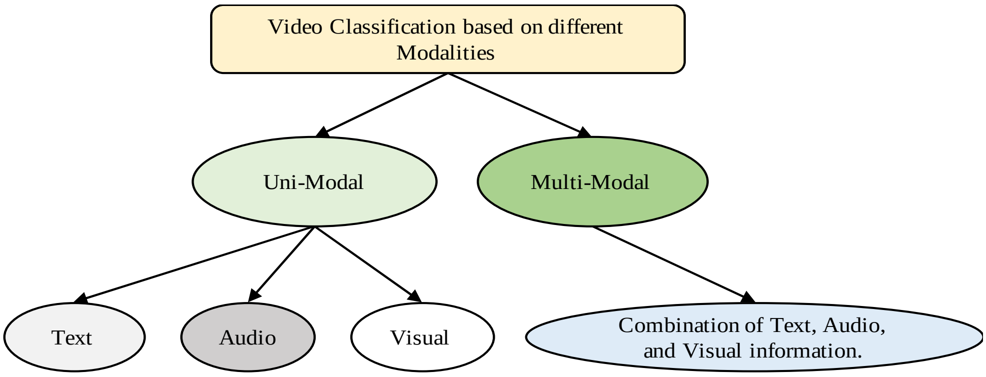

Video Classification based on different Modalities
Uni-Modal
Multi-Modal
Text
Audio
Visual
Combination of Text, Audio, and Visual information.

Figure 2. Different modalities used for classification of videos.

3.2. Traditional Handcrafted Features

During the earlier developments of the video classification task, the traditional handcrafted features were combined with state-of-art machine learning algorithms to classify the videos. Some of the most popular handcrafted feature representation techniques used in the literature are spatiotemporal interest points (STIPs) [26], improved dense trajectories (iDT) [27], SIFT-3D [28], HOG3D [29], motion boundary histogram [30], action-bank [31], cuboids [32], 3D SURF [33], and dynamic-poselets [34]. These hand-designed representations use different feature encoding schemes such as the ones based on pyramids and histograms. iDT is one of these handcrafted representations that is widely considered the state-of-the-art. Many recent competitive studies demonstrated that handcrafted features [35–38] and high-level [39,40] and mid-level [41,42] video representations have contributed towards the task of video classification with deep neural networks.

3.3. Deep Learning Frameworks

Along with the development of more powerful deep learning architectures in the recent years, the trend for the video classification task has followed a shift from traditional handcrafted approaches to the fully automated deep learning approaches. Among the very common deep learning architectures used for video classification is a 3D-CNN model. An example of 3D-CNN architecture used for video classification is given in Figure 3 [43]. In this architecture, 3D blocks are utilized to capture the video information necessary to classify the video content. One more very common architecture is a multi-stream architecture, where the spatial and temporal information is separately processed, and the features extracted from different streams are then fused to make a decision. To process the temporal information, different methods are used, and the two most common methods are based on (i) RNN (mainly LSTM) and (ii) optical flow. An example of a multi-stream network model [44], where the temporal stream is processed using optical flow, is shown in Figure 4. A high-level overview of the video classification process is shown in Figure 5, where the stages of feature extraction and prediction are shown with the most common type of strategies used in the literature. In the upcoming sections, the breakthroughs in video classification and studies related to classification of videos, specifically using deep learning frameworks, are summarized, describing the success rate of utilizing deep learning architectures and the associated limitations.

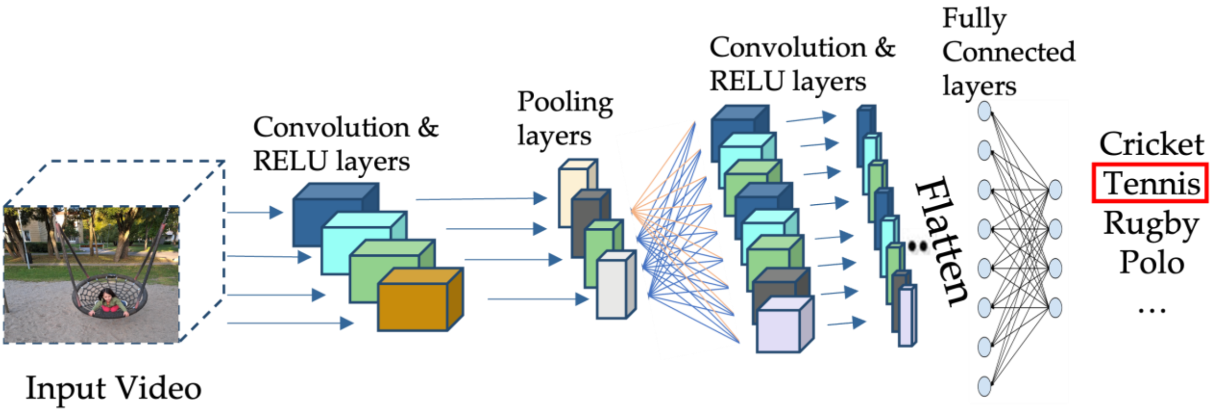


Figure 3. An example of 3D-CNN architecture to classify videos.

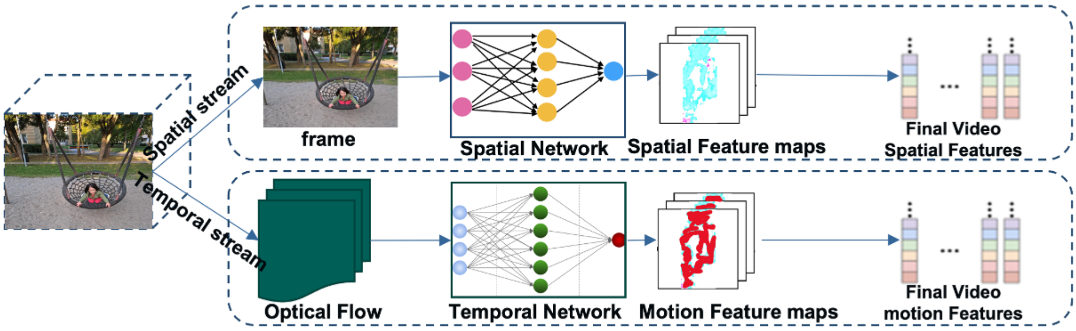


Figure 4. An example of two-stream architecture with optical flow.

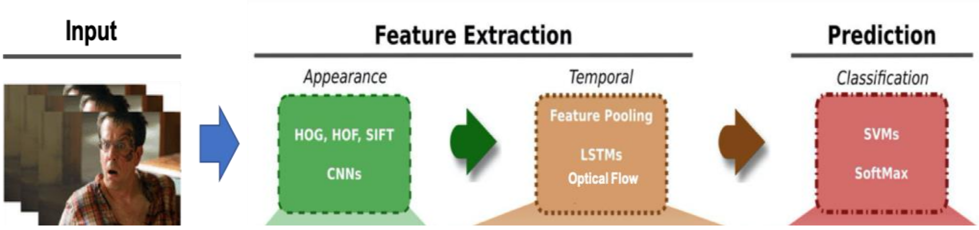


Figure 5. An overview of video classification process.

3.4. Breakthroughs

The breakthroughs in recognition of still-images originated with the introduction of a deep learning model called AlexNet [13]. The same concept of still-image recognition using deep learning is also extended for videos, where individual video frames are collectively processed as images by a deep learning model to predict the contents of a video. The features from individual video frames are extracted, and then, temporal integration of such features into a fixed-size descriptor using pooling is performed. The task is either done using high-dimensional feature encoding [45,46] or through the RNN architectures [47–50]. For un-supervised spatiotemporal feature learning in 3D convolutions, restricted Boltzmann machines [51] and stacked ISA [52] are also studied in parallel. The 3D-CNNs using temporal convolutions to extract temporal features automatically were first proposed by Baccouche et al. [53] and by Ji et al. [54].

3.5. Basic Deep Learning Architectures for Video Classification

The two most widely used deep learning architectures for video classification are convolutional neural network (CNN) and recurrent neural network (RNN). CNNs are mostly used to learn the spatial information from videos, whereas RNNs are used to learn the temporal information from videos, as the main difference between these two architectures is the ability to process temporal information or data that come in sequences. Therefore, both these network architectures are used for completely different purposes in general. However, the nature of video data with the presence of both the spatial and the temporal information demands the use of both these network architectures to accurately process the two-stream information. The architecture of a CNN applies different filters in the convolutional layers to transform the data. RNNs, on the other hand, reuse the activation functions to generate the next output in a series from the other data points in the sequence. However, the use of only 2D-CNNs alone limits the understanding of video to only spatial domain. RNNs, on the other hand, can understand the temporal content of a sequence. Both these basic architectures and their enhanced versions are applied in several studies for the task of video classification.

3.6. Developments in Video Classification over Time

The existing approaches for video classification are categorized based on their working principle in Table 2. The trend observed for the classification of videos from the existing literature is that the recently developed state-of-art deep learning models are outperforming the earlier handcrafted classical approaches. This is mainly due to the availability of large-scale video data for learning deep architectures of neural networks. Besides an improvement in classification performance the recently developed models are mostly self-learned and does not require any manual feature engineering. This added advantage makes them more feasible for use in real applications. However, the better performing recently developed architectures are deeper as compared to the previously developed architectures which brings a compromise on the computational complexity of the deep architectures.

Table 2. Different categories of approaches of video classification.

Categories	Working Principle	References
Hand-crafted approaches	These representations are handcrafted and employ various feature encoding techniques, such as histograms and pyramids.	Spatiotemporal Interest Points (STIPs) [26], iDT [27], SIFT-3D [28], HOG3D [29], Motion Boundary Histogram [30], Cuboids [32], Action- Bank [31], 3D SURF [33], Dynamic-Poselets [34].
2D- CNNs	These are image based models where frame level feature extraction is performed using CNN architecture and classification is performed using state-of-art classification models, for example SVM.	[55]
3D-CNNs	2D image classification extension to 3D for video (For example the Inception 3D (I3D) architecture).	[56]
Spatiotemporal Convolutional Networks	To aggregate the temporal and the spatial information, these methods primarily depend on convolution and pooling.	[54,57,58]
Recurrent Spatial Networks	To represent temporal information in videos, recurrent neural networks such as LSTM or GRU are used.	[47,53,59,60]
Two/multi Stream Networks	In addition to the context frame visuals, these methods use layered optical flow to identify movements.	[50,61–63]
Mixed convolutional models	Models built with the ResNet architecture in mind. They are particularly interested in models that utilize 3D convolution in the bottom or top layers but 2D in the remainder; these are referred to as "mixed convolutional" models. Or the methods based on mixed temporal convolution with different kernel sizes.	[64,65]
Hybrid Approaches	These are models based on integration of CNN and RNN architectures.	[66–68]

Among the initially developed hand-crafted representations, improved Dense Trajectories (iDT) [27] is widely considered the state-of-the-art. Whereas, many recent competitive studies demonstrated that hand-crafted features [35–38], high-level [39,40], and mid-level [41,42] video representations have contributed towards the task of video classification with deep neural networks. The hand-crafted models were among the very early developments of video classification problem. Later, 2D-CNNs were proposed for video classification, where image-based CNN models are used to extract frame level features and based on the frame level CNN features, some state-of-art classification models (for example SVM) are learned to classify videos. These 2D-CNN models do not require any manual feature extraction and these models performed better than the competing hand-crafted approaches. After successful development of 2D-CNN models where features are extracted from frame level, the same concept was extended to propose 3D-CNNs to extract features from videos. The proposed 3D-CNNs are computationally more expensive as compared to the 2D-CNN models. However, these models consider the time variations in feature extraction therefore these 3D-CNN models are believed to perform better as compared to 2D-CNN models for video classification [54,58,69].

The development of 3D-CNN models paved the way for fully automatic video classification models using different deep learning architectures. Among the developments using deep learning architectures, spatiotemporal convolutional networks are approaches based on integration of temporal and spatial information using convolutional networks to perform video classification. To collect temporal and spatial information, these methods primarily rely on convolution and pooling layers. Stack optical flow is used in two/multi-stream networks methods to identify movements in addition to context frame visuals. Recurrent spatial networks use recurrent neural networks (RNN) to model temporal information in videos, such as LSTM or GRU. The ResNet architecture is used to build mixed convolutional models. They are particularly interested in models that utilize 3D convolution in the bottom or top layers but 2D in the remainder; these are referred to as "mixed convolutional" models. These also include methods based on mixed temporal convolution with different kernel sizes. Advanced architectures based on DenseNet have also shown promising results for the video classification task. Some of these notable architectures based on DenseNet include region-based CNN (R-CNN) [70,71], faster R-CNN [72,73], and YOLO [74]. Besides these architectures, there are also hybrid approaches based on the integration of CNN and RNN architectures. A summary of these architectures is provided in Figure 6.

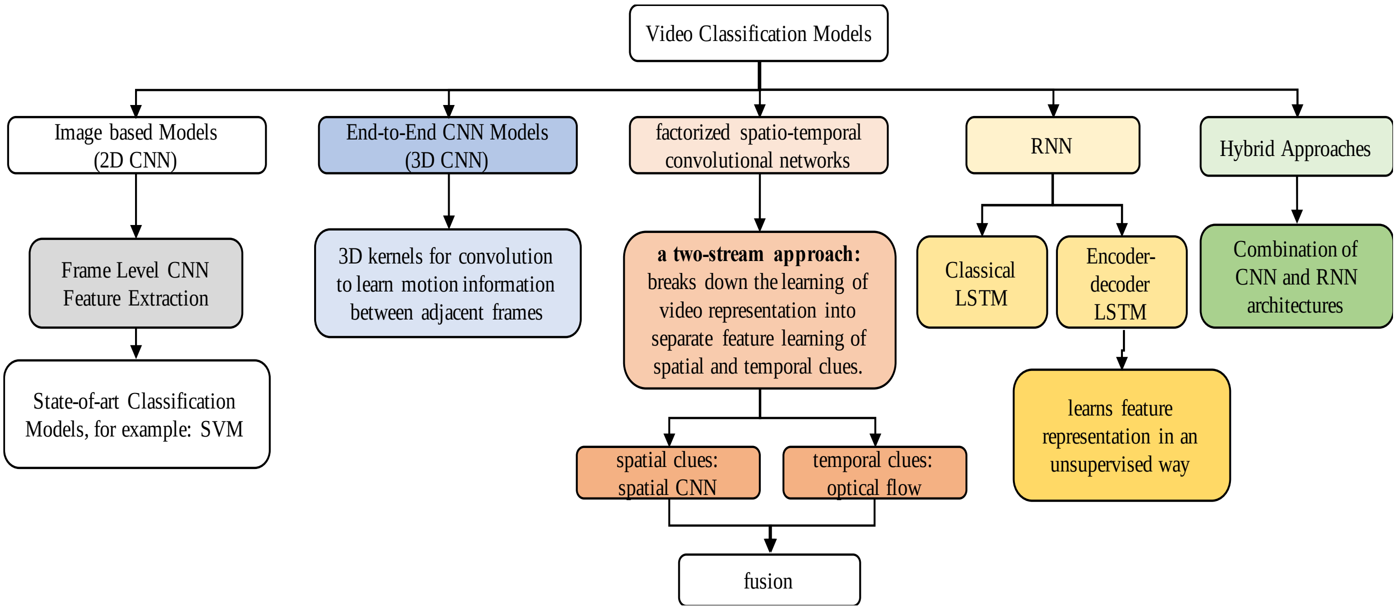

Video Classification Models
Image based Models (2D CNN)
End-to-End CNN Models (3D CNN)
factorized spatio-temporal convolutional networks
RNN
Hybrid Approaches
Frame Level CNN Feature Extraction
State-of-art Classification Models, for example: SVM
3D kernels for convolution to learn motion information between adjacent frames
a two-stream approach: breaks down the learning of video representation into separate feature learning of spatial and temporal clues.
spatial clues: spatial CNN
temporal clues: optical flow
fusion
Classical LSTM
Encoder-decoder LSTM
learns feature representation in an unsupervised way
Combination of CNN and RNN architectures

Figure 6. Summary of video classification approaches.

The different deep learning architectures described above employ different fusion strategies. These fusion strategies are either for the fusion of different features extracted from the video or for the fusion of different models used in the architecture. The fusion strategies mainly used for the extracted features are (i) concatenation, (ii) product, (iii) summation, (iv) maximum, and (v) weighted, where the concatenation approach simply combines all the features together, and all the features are used for classification. The product/summation approach performs the product/summation between the features extracted using different strategies and uses the result of product/summation to perform classification. The maximum approach takes the maximum value of the features extracted using different strategies and uses that for classification. The weighted approach gives different weights to different features and performs the classification using the weighted features. Different fusion methods are summarized in Figure 7.

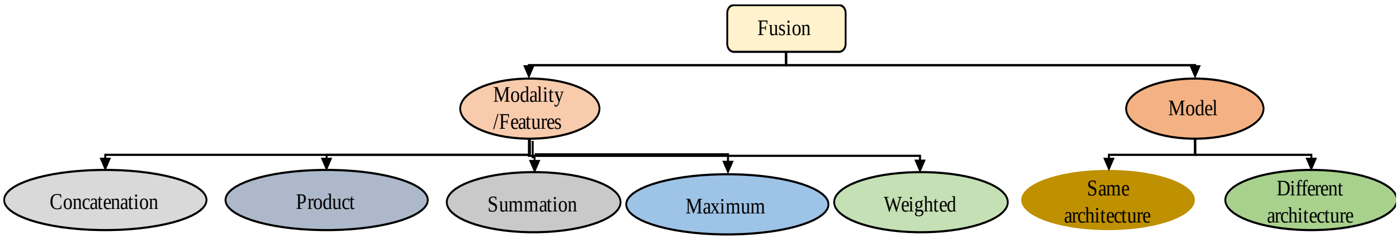


Figure 7. Different Fusion Types.

3.7. Summary of Some Notable Deep Learning Frameworks Developments

A summary of some deep learnings architectures for video classification is provided in Table 3. These studies are summarized based on the architecture, the datasets, the evaluation metrics, the fusion strategy, and the notable findings. The most common architectures for video classification are fundamentally based on the RNN and CNN architectures; classification accuracy is one of the most common evaluation metrics; UCF-101 and Sports-1M datasets are the choice for validation in most cases, multi-class classification problem is considered in almost all cases, SMART blocks outperform 3D convolutions in terms of spatiotemporal feature learning, and average fusion, kernel average fusion, weighted fusion, logistic regression fusion, and MKL fusion are all proven to be inferior compared to the multi-stream multi-class fusion technique. Moreover, a more applied form of classification in videos is to identify/recommend tags or thumbnails in videos, and this specific task is successfully caried out in [75–79].

3.8. Few-Shot Video Classification

FEW-SHOT learning (FSL) has received a great deal of interest in recent years. FSL tries to identify new classes with one or a few labeled samples [80–83]. However, due to most recent work in few-shot learning being centered on image classification, FSL in the video domain is still hardly being explored [84,85]. Some of the notable works done in this domain are discussed below.

A multi-saliency embedding technique was developed by Zhu et al. [85] to encode a variable-length video stream into a fixed-size matrix. Graph neural networks (GNN) were developed by Hu et al. [86] to enhance the video classification model's capacity for discrimination. The local–global link in a distributed representation space was still disregarded nevertheless. To categorize a previously unseen video, Cao et al. [87] introduced a temporal alignment module (TAM) that explicitly took advantage of the temporal ordering information in video data through temporal alignment. To combine the two-stream aspects of videos more effectively, Fu et al. [88] developed a depth-guided adaptive instance-normalization module (DGAdaIN). A C3D encoder was created by Zhang et al. [89] to record close-range action patterns for spatiotemporal video blocks. Few-shot video categorization was addressed by Qi et al. [90] by learning a collection of SlowFast networks enhanced with memory units. To comprehend realistic films of the target classes, Fu et al. [91] presented embodied agent-based one-shot learning, which made use of synthetic videos created in a virtual environment. For the issues of few-shot and zeroshot action recognition, Bishay et al. [92] presented the temporal attentive relation network (TARN), which was trained to compare representations of varying temporal length. By examining local–global linkages and preserving the specifics of properties, Y. Feng et al. [93] recently presented a dual-routing capsule graph neural network (DR-CapsGNN) to address the issue of severely constrained samples in few-shot learning.

Apart from this, contrastive learning has also proved successful in recognizing human actions. Some of the interesting works done in this regard are multi-granularity anchor-contrastive representation learning [94] and X-invariant contrastive augmentation and representation learning [95].

3.9. Geometric Deep Learning

Shape descriptors play a significant role in the description of manifolds for 3D shapes. In general, a global feature descriptor is created by aggregating local descriptors to describe the geometric properties of the entire shape, for example, using the bag-of-features paradigm. A local feature descriptor assigns a vector to each point on the shape in a multi-dimensional descriptor space, representing the local structure of the shape around that point. Most deep learning techniques that deal with 3D shapes essentially use the CNN paradigm. Volumetric 2D multi-view shape representations are applied directly using standard (Euclidean) CNN architectures in neural networks via methods such as

[96,97]. These techniques are unsuited for dealing with deformable shapes because the shape descriptors they use are dependent on extrinsic structures that are invariant under Euclidean transformations, as demonstrated in Figure 8a [98], while some other approaches [99–103] create a new framework by adopting the CNN feature extraction pattern to investigate the inherent CNN versions that would enable handling shape deformations by using intrinsic filter structure, as shown in Figure 8b [98]. Geometric deep learning deals with non-Euclidean graph and manifold data. This type of data (irregularly arranged/distributed randomly) is usually used to describe geometric shapes. The purpose of geometric deep learning is to find the underlying patterns in geometric data where the traditional Euclidean distance-based deep learning approaches are not suitable. There are basically two methods available in the literature to apply deep learning on geometric data: (i) extrinsic methods and (ii) intrinsic methods. The filters in extrinsic methods are applied on the 3D surfaces such that it effects the structural deformity due to the extrinsic filter structure. The key weakness of extrinsic approaches [96,97] is that they continue to consider geometric data as Euclidean information. When an object's position or shape changes, the extrinsic data representation fails. Additionally, for these methods to support the challenging-in-practice task of attaining the invariance of shape deformation, complicated models and extensive training are required. The filters in intrinsic approaches are applied on the 3D surfaces without being affected by the structural deformity. Rather than Euclidean realization, intrinsic methods work on the manifold and are isometry-invariant by construction. Some of the works based on intrinsic deep learning include (i) geodesic CNN [99], (ii) anisotropic CNN [100], (iii) mixture model network [101], (iv) structured prediction model [102], (v) localized spectral CNN [103], (vi) PointNet [104], (vii) PointNet++ [105], and (viii) RGA-MLP [106]. The application of geometric deep learning (mostly intrinsic methods) in analyzing videos can help in better understanding from the machine perspective, but it is still an open research problem and needs further investigation. For further details on geometric deep learning, readers are referred to [98,107].

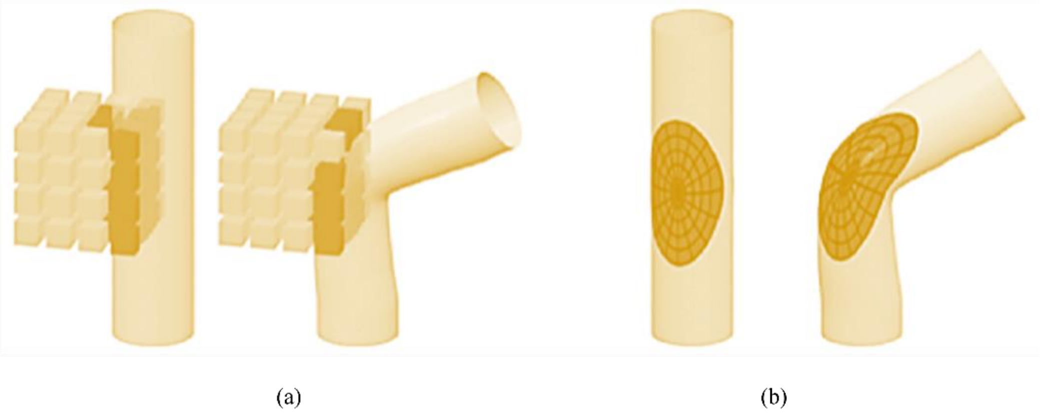

(a) (b)

Figure 8. Illustration of deep learning approaches on geometric data. (**a**) Extrinsic method and (**b**) intrinsic method.

Table 3. Summary and findings of studies based on deep learning models.

Study	Features	Model	Evaluation	Dataset	Problem	Fusion	Findings
[57]	Automatic spatio-temporal features/self-learning. Temporal features captured both locally and globally.	Multiresolution CNN architecture.	By the fraction of test samples that contained at least one of the ground truth labels in the top k predictions.	Sports-1M, UCF-101.	Multi-class	Single frame, Early Fusion, Late Fusion, Slow Fusion.	When compared to a multilayer neural network with rectified linear units followed by a Softmax classifier built using histogram features, the Softmax classifier performed better (both local features such as texton, HOG, cuboids, etc., and global features such as color moments, and hue–saturation).
[108]	Visual (dense trajectory descriptors): A 30-d trajectory shape descriptor, a 96-d HOG descriptor, a 108-d HOF descriptor, and a 108-d MBH descriptor (local visual descriptors). Audio Features: MFCCs and Spectrogram SIFT.	Deep neural network (DNN).	Mean average precision (mAP).	Hollywood2, Columbia Consumer Videos (CCV), and CCV+.	Multi-class	Regularized fusion of multiple features.	Found better than dense trajectory features and classification utilizing the basic early fusion technique.
[109]	Tensor-Train Factorization (global representation for the whole sequence).	Recurrent neural network (RNN).	Classification accuracy.	UCF11, Hollywood2, YouTube Celebrities Face Data.	Multi-class	-	Tensor-Train layer-based RNN such as LSTM and GRU perform better than the plain RNN architectures for video classification.
[110]	Improved Fisher vector (iFV) and explicit feature maps to represent features of conv and fc layers. Long-term temporal information.	A multilayer and multimodal fusion framework of deep neural networks based on fully connected (FC)-RNN.	Classification accuracy.	UCF101, HMDB51.	Multi-class	Multilayer and multi-modal fusion framework.	When compared to enhanced dense trajectories, which require a number of handcrafted procedures such as dense point tracking, camera motion estimation, person detection, and so on, the proposed FC-RNN obtained competitive results.
[50]	Convolutional temporal feature pooling architectures (conv pooling, late pooling, slow pooling, local pooling). Global video-level descriptors.	Two CNN architectures (AlexNet and GoogleNet) and LSTM.	By the fraction of test samples that contained at least one of the ground truth labels in the top k predictions.	UCF101, Sports 1 million.	Multi-class	Late fusion	(i) UCF-101 necessitates the utilization of optical flow. (ii) Optical flow is not always beneficial, especially when the videos are captured in the wild, such as Sports-1M. (iii) To make use of optical flow, a more advanced sequence processing architecture such as LSTM is required. (iv) The maximum documented performance is achieved by using LSTMs on both image frames and optical flow for the Sports-1M benchmark.
[111]	Spatiotemporal feature learning: a SMART block and ARTNet for short-term spatiotemporal feature learning with	ARTNet by integrating the SMART block into the C3D-ResNet18 architecture, where	Top-1 and Top-5 accuracy.	Kinetics, UCF101, and HMDB51.	Multi-class	Concatenation and	(i) In terms of spatiotemporal feature learning, SMART blocks outperform 3D convolutions (3D-CNN). (ii) In the case of ARTNet, supplementing RGB input with

	a possibility to explore long-term learning.	SMART block architecture is composed of appearance branch and relationship branch.				reduction operation.	optical flow improves performance. (iii) The optical flow modality can give additional information. (iv) Optical flow's high computing cost prevents it from being used in real-world systems.
[112]	Spatial, short-term motion and audio clues using CNN. Long-term temporal dynamics. (Multimodal features).	CNNs-LSTM model with multi-stream multi-class fusion process to adaptively determine the optimal fusion weights for generating the final scores of each class.	Classification accuracy.	UCF-101, Columbia Consumer Videos.	Multi-class	Multi-Stream Multi-Class Fusion.	Average fusion, kernel average fusion, weighted fusion, logistic regression fusion, and MKL fusion are all proven to be inferior to the proposed multi-stream multi-class fusion technique.
[113]	Two distinct layers: 1 × 1 × 1 conventional convolutions for channel interaction (but no local interaction) and k × k × k depth-wise convolutions for local spatiotemporal interactions (but not channel interaction). Global spatiotemporal average pooling layer.	Channel-separated convolutional network (CSN). Two models: interaction-preserved channel-separated network (ip-CSN) and interaction-reduced channel-separated network (ir-CSN).	Classification accuracy.	Sports1M and Kinetics.	Multi-class	-	(i) In 3D group convolutional networks, the number of channel interactions has a significant impact on accuracy. (ii) Separating channel interactions from spatiotemporal interactions in 3D convolutions improves accuracy and reduces computing cost. (iii) Three-dimensional channel-separated convolutions offer regularization and avoid overfitting.
[114]	The 3D network is optimized with three loss functions: (i) cross-entropy (CE) loss, (ii) pseudo-CE loss, and (iii) soft CE loss. 2D Image and 3D video model capture short and long visual descriptors.	Semi-supervised learning (VideoSSL) with 3D ResNet-18.	Top-1	UCF101, HMDB51, and Kinetics.	Multi-class	-	(i) For 3D video classification, a direct application of current semi-supervised algorithms (which were initially designed for 2D imagery) cannot yield adequate results. (ii) The accuracy of 3D-CNN models is much improved by a calibrated use of object appearance indicators for semi-supervised learning.
[115]	Modal- and channel-wise attentions.	Expansion-squeeze excitation fusion network	Accuracy, confusion matrix	ETRI-ACTIVITY3D, NUT RGB+D	Multi-class	Multi-modal	(i) Modal-fusion nets (M-Nets) and channel-fusion nets (C-Nets) are capable of capturing the modal and channel-wise dependencies between features in order to improve the discriminative power of features via modal and channel-wise ESEs. (ii) By adding the penalty of the difference between the minimum prediction losses on the single modalities and the prediction loss on the fused modality, multi-modal loss (ML) can further enforce the consistency between the single-modal features and the fused multi-modal features.

4. Benchmark Datasets, Evaluation Metrics, and Comparison of Existing State-of-the-Art for Video Classification

4.1. Benchmark Datasets for Video Classification

There are several benchmark datasets being utilized for classification of videos, AND some of these notable datasets are summarized in Table 4. The details related to these datasets, such as total number of videos contained in the dataset, number of classes present in the dataset, the year of publication of dataset, and the background of videos in the dataset, are included in the summary.

Table 4. Benchmark datasets.

Dataset	# of Videos	# of Classes	Year	Background
KTH	600	6	2004	Static
Weizmann	81	9	2005	Static
Kodak	1358	25	2007	Dynamic
Hollywood	430	8	2008	Dynamic
Hollywood2	1787	12	2009	Dynamic
MCG-WEBV	234,414	15	2009	Dynamic
Olympic Sports	800	16	2010	Dynamic
HMDB51	6766	51	2011	Dynamic
CCV	9317	20	2011	Dynamic
UCF-101	13,320	101	2012	Dynamic
THUMOS-2014	18,394	101	2014	Dynamic
MED-2014 (Dev. set)	31,000	20	2014	Dynamic
Sports-1M	1,133,158	487	2014	Dynamic
ActivityNet	27,901	203	2015	Dynamic
EventNet	95,321	500	2015	Dynamic
MPII Human Pose	20,943	410	2014	Dynamic
FCVID	91,223	239	2015	Dynamic
UCF11	1600	11	2009	Dynamic
YouTube Celebrities Face	1910	47	2008	Dynamic
Kinetics	300,000	400	2017	Dynamic
YouTube-8M	6.1 M	3862	2018	Dynamic
JHMDB	928	21	2011	Dynamic
Something-something	110,000	174	2017	Dynamic

4.2. Performance Evaluation Metrics for Video Classification

The evaluation of video classification models is performed using different performance measures. The most common measures utilized to evaluate the models are accuracy, precision, recall, F1 score, micro F1, and K-fold [8]. Some of the recent studies using these measures are listed in Table 5.

Table 5. Commonly used evaluation metrics for video classification.

Evaluation Metric	Year of Publication	Reference
Accuracy	2020–2021	[116–120]
Precision	2020–2021	[116,118,119]
Recall	2020–2021	[116,118,119]
F1 Score	2020–2021	[116,118,119]
Micro F1	2020	[121,122]
K-Fold	2019	[123]
Top-k	2018,2021	[111,114]

4.3. Comparison of Some Existing Approaches on UCF-101 Dataset

UCF-101 is a benchmark action recognition dataset published by the researchers of University of Central Florida in the year 2012 [124], and the videos in the dataset were collected from YouTube. The total videos in the dataset are 13,320, with 101 action categories. The dataset is challenging because of the uncontrolled environment in the captured videos, and it is widely being used by researchers working on the video classification problem. Therefore, it is easy to compare most of the existing literature based on this dataset. The existing works employing UCF-101 are compared in Table 6, where the methods are arranged in ascending order based on the performance. The results reported in Table 6 are taken from the existing studies in the literature.

Table 6. Comparison of video classification method on UCF-101.

Method	Accuracy
LRCN [48]	82.9
DT + MVSV [125]	83.5
LSTM–Composite [49]	84.3
FSTCN [126]	88.1
C3D [127]	85.2
iDT + HSV [128]	87.9
Two-Stream [61]	88.0
RNN-FV [129]	88.0
LSTM [50]	88.6
MultiSource CNN [130]	89.1
Image-Based [55]	89.6
TDD [35]	90.3
Multilayer and Multimodal Fusion [110]	91.6
Transformation CNN [131]	92.4
Multi-Stream [112]	92.6
Key Volume Mining [132]	92.7
Convolutional Two-Stream [62]	93.5
Temporal Segment Networks [39]	94.2

4.4. Comparison of Different Deep Learning Architectures

In Table 7, some important deep learning architectures are compared in terms of performance and computational requirement. These architectures are the basis of development of different deep learning models for video classification, and from this comparison, an estimation of the requirement of computational cost for each of these architectures can be drawn.

Table 7. Performance comparison of different deep architectures [127].

Architecture Name	Parameters	Error Rate	Depth	Category	Year
LeNet	0.060 M	[dist]MNIST: 0.8 MNIST: 0.95	5	Spatial exploitation	1998
AlexNet	60 M	ImageNet: 16.4	8	Spatial exploitation	2012
ZfNet	60 M	ImageNet: 11.7	8	Spatial exploitation	2014
VGG	138 M	ImageNet: 7.3	19	Spatial exploitation	2014
GoogLeNet	4 M	ImageNet: 6.7	22	Spatial exploitation	2015
Inception-V3	23.6 M	ImageNet: 3.5 multi-crop: 3.58 Single-Crop: 5.6	159	Depth + width	2015
Highway networks	2.3 M	CIFAR-10: 7.76	19	Depth + multi-path	2015

Inception-V4	35 M	ImageNet: 4.01	70	Depth +width	2016
Inception-ResNet	55.8 M	ImageNet: 3.52	572	Depth + width + multi-path	2016
ResNet	25.6 M	ImageNet: 3.6	152	Depth + multi-path	2016
	1.7 M	CIFAR-10: 6.43	110		
DelugeNet	20.2 M	CIFAR-10: 3.76	146	Multi-path	2016
		CIFAR-100: 19.02			
FractalNet	38.6 M	CIFAR-10: 7.27		Multi-path	2016
		CIFAR-10 +: 4.60			
		CIFAR-10 ++: 4.59	20		
		CIFAR-100: 28.20	40		
		CIFAR-100 +: 22.49			
		CIFAR100 ++: 21.49			
WideResNet	36.5 M	CIFAR-10: 3.89	28	Width	2016
		CIFAR-100: 18.85	–		
Xception	22.8 M	ImageNet: 0.055	126	Width	2017
Residual attention neural network	8.6 M	CIFAR-10: 3.90	452	Attention	2017
		CIFAR-100: 20.4			
		ImageNet: 4.8			
ResNeXt	68.1 M	CIFAR-10: 3.58	29	Width	2017
		CIFAR-100: 17.31	-		
		ImageNet: 4.4	101		
Squeeze and excitation networks	27.5 M	ImageNet: 2.3	152	Feature-map exploitation	2017
DenseNet	25.6 M	CIFAR-10 +: 3.46	190	Multi-path	2017
	25.6 M	CIFAR100 +:17.18	190		
	15.3 M	CIFAR-10: 5.19	250		
	15.3 M	CIFAR-100: 19.64	250		
PolyNet	92 M	ImageNet: Single:4.25	–	Width	2017
		Multi:3.45	–		
PyramidalNet	116.4 M	ImageNet: 4.7	200	Width	2017
	27.0 M	CIFAR-10: 3.48	164		
	27.0 M	CIFAR-100: 17.01	164		
Convolutional block attention Module (ResNeXt101 (32 × 4d) + CBAM)	48.96 M	ImageNet: 5.59	101	Attention	2018
Concurrent spatial and channel excitation mechanism	–	MALC: 0.12 Visceral: 0.09	–	Attention	2018
Channel boosted CNN	–	–	–	Channel boosting	2018
Competitive squeeze and excitation network CMPE-SE-WRN-28	36.92 M	CIFAR-10: 3.58	152	Feature-map exploitation	2018
	36.90 M	CIFAR-100: 18.47	152		

5. Key Findings

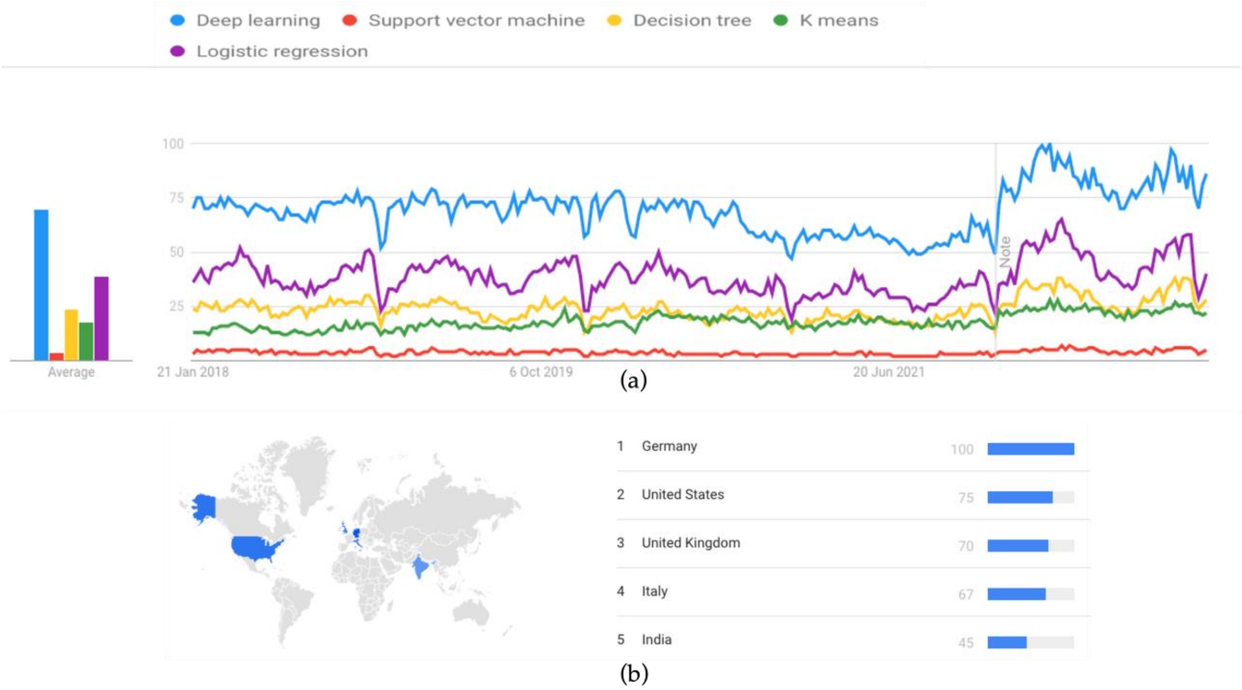
From the analysis of the existing literature, the following key findings are drawn for video classification task: (i) The visual description works better than the text and the audio description, and the combination of all modalities can contribute to better performance with an increase in computational cost. (ii) The architectures employing CNN/RNN for feature extraction have the ability to perform better than handcrafted features provided that enough data are available for training. (iii) Tensor-Train layer-based RNN such as LSTM and GRU perform better than the plain RNN architectures for video classification. (iv) It is sometimes necessary to use optical flow for datasets such as UCF-101. (v) It is not always helpful to use optical flow, especially for the case of videos taken from the wild, e.g., Sports-1 M. (vi) It is important to use a sophisticated sequence processing architecture such as LSTM to take advantage of optical flow. (vii) LSTMs, when applied on both the optical flow and the image frames, yield the highest performance measure for the Sports-1M benchmark dataset. (viii) Augmenting optical flow and RGB input helps in improving the performance. (ix) Optical flow modality provides complementary information. (x) The high computational requirement of optical flow limits its use in real-time systems. (xi) Multi-stream multi-class fusion can perform better than average fusion, weighted fusion, kernel average fusion, MKL fusion, and logistic regression fusion on datasets such as UCF-101 and CCV. (xii) In 3D group convolutional networks, the volume of channel interactions plays a vital role in achieving a high accuracy. (xiii) The factorization of 3D convolutions by separating spatiotemporal interactions and channel interactions can lead to an improvement in accuracy and a decrease in the computational cost. (xiv) Further, 3D channel-separated convolutions results in a kind of regularization and prevents overfitting. (xv) Popular frameworks of conventional semi-supervised algorithms (which were originally developed for 2D images) are unable to obtain good results for 3D video categorization. (xvi) For semi-supervised learning, a calibrated employment of the object appearance cues keenly improves the accuracy of the 3D-CNN models.

6. Conclusions

This article reviews deep learning approaches for the task of video classification. Some of the notable studies are summarized in detail, and the key findings in these studies are highlighted. The key findings are reported as an effort to help the research community in developing new deep learning models for video classification.

The latest developments in deep learning models have demonstrated the potential of these approaches for the video classification task. However, most of the existing deep learning architectures for video classification are basically adopted from the favored deep learning architectures in image/speech domain. Therefore, most of the existing architectures remain insufficient to deal with the more complicated nature of video data that contain rich information in the form of spatial, temporal, and acoustic clues. This calls for attention towards the need for a tailored network capable of effectively modeling the spatial, temporal, and acoustic information. Moreover, training CNN/RNN models requires labeled datasets, and acquiring those datasets is usually time-consuming and expensive, and hence, a promising research direction is to utilize the considerable amount of unlabeled video data to derive better video representations.

Furthermore, the deep learning approaches are outperforming other state-of-the-art approaches for video classification. The deep learning Google trend is still growing, and it is still above the trend for some other very well-known machine learning algorithms, as shown in Figure 9a. However, the recent developments in deep learning approaches are still under-evaluated and require further investigations for the video classification task. One such example is geometric deep learning approaches, and the worldwide research interest in this specific topic is shown in Figure 9b, which describes that this topic is still confined to some states of U.S., Europe, and India. Therefore, it has yet to be developed and investigated further. The use of geometric deep learning in extracting rich spatial information from videos can also be a new research direction as a future work for better accuracy in the video classification task.



(a)

1	Germany	100
2	United States	75
3	United Kingdom	70
4	Italy	67
5	India	45

(b)

Figure 9. (**a**) Google trend on deep learning vs. some other state-of-the-art methods. (**b**) Worldwide research interest in geometric deep learning.

Author Contributions: Conceptualization, A.u.R. and S.B.B.; methodology, M.A.K.; validation, A.K.; formal analysis, S.B.B.; data curation, A.u.R.; writing—original draft preparation, A.u.R.; writing—review and editing, M.A.K.; visualization, A.u.R.; supervision, S.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to acknowledge Mälardalen University for supporting the publication charges.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Muller, K.-R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* **2021**, *109*, 247–278. https://doi.org/10.1109/JPROC.2021.3060483.
2. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. https://doi.org/10.1016/j.ymssp.2020.107398.
3. Minallah, N.; Tariq, M.; Aziz, N.; Khan, W.; Rehman, A.; Belhaouari, S.B. On the performance of fusion based planet-scope and Sentinel-2 data for crop classification using inception inspired deep convolutional neural network. *PLoS ONE* **2020**, *15*, e0239746. https://doi.org/10.1371/journal.pone.0239746.
4. Rehman, A.; Bermak, A. Averaging Neural Network Ensembles Model for Quantification of Volatile Organic Compound. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 848–852. https://doi.org/10.1109/IWCMC.2019.8766776.
5. Anushya, A. Video Tagging Using Deep Learning: A Survey. *Int. J. Comput. Sci. Mob. Comput.* **2020**, *9*, 49–55.
6. Rani, P.; Kaur, J.; Kaswan, S. Automatic Video Classification: A Review. *EAI Endorsed Trans. Creat. Technol.* **2020**, *7*, 163996. https://doi.org/10.4108/eai.13-7-2018.163996.
7. Li, Y.; Wang, C.; Liu, J. A Systematic Review of Literature on User Behavior in Video Game Live Streaming. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3328. https://doi.org/10.3390/ijerph17093328.

8. Islam, M.S.; Sultana, M.S.; Roy, U.K.; al Mahmud, J. A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work. *J. Ilm. Tek. Elektro Komput. Dan Inform.* **2021**, *6*, 47. https://doi.org/10.26555/jiteki.v6i2.18978.
9. Ullah, H.A.; Letchmunan, S.; Zia, M.S.; Butt, U.M.; Hassan, F.H. Analysis of Deep Neural Networks for Human Activity Recognition in Videos—A Systematic Literature Review. *IEEE Access* **2021**, *9*, 126366–126387. https://doi.org/10.1109/ACCESS.2021.3110610.
10. Wu, Z.; Yao, T.; Fu, Y.; Jiang, Y.-G. Deep learning for video classification and captioning. In *Frontiers of Multimedia Research*; ACM: New York, NY, USA, 2017; pp. 3–29. https://doi.org/10.1145/3122865.3122867.
11. Ren, Q.; Bai, L.; Wang, H.; Deng, Z.; Zhu, X.; Li, H.; Luo, C. A Survey on Video Classification Methods Based on Deep Learning. *DEStech Trans. Comput. Sci. Eng.* **2019**. https://doi.org/10.12783/dtcse/cisnrc2019/33301.
12. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based Learning Applied to Document Recognition. *Intell. Signal Process.* **2001**, 306–351. https://doi.org/10.1109/9780470544976.ch9.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *2*, 1097–1105. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001284.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
15. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
17. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
18. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **2020**, *53*, 5455–5516. https://doi.org/10.1007/s10462-020-09825-6.
19. Ian, G.; Yoshua, B.; Aaron, C. *Deep Learning (Adaptive Computation and Machine Learning Series)*; The MIT Press: Cambridge, MA,USA, 2016.
20. Shah, A.M.; Yan, X.; Shah, S.A.A.; Mamirkulova, G. Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. *J. Ambient Intell. Humaniz Comput.* **2020**, *11*, 2925–2942. https://doi.org/10.1007/s12652-019-01434-8.
21. De Jong, R.J.; de Wit, J.J.M.; Uysal, F. Classification of human activity using radar and video multimodal learning. *IET Radar Sonar Navig.* **2021**, *15*, 902–914. https://doi.org/10.1049/rsn2.12064.
22. Truong, B.T.; Venkatesh, S.; Dorai, C. Automatic genre identification for content-based video categorization. In Proceedings of the International Conference on Pattern Recognition 2000, Barcelona, Spain, 3–7 September 2000; Volume 15, pp. 230–233. https://doi.org/10.1109/icpr.2000.902901.
23. Huang, C.; Fu, T.; Chen, H. Text-based video content classification for online video-sharing sites. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 891–906. https://doi.org/10.1002/asi.21291.
24. Lee, K.; Ellis, D.P.W. Audio-based semantic concept classification for consumer video. *IEEE Trans. Audio Speech Lang Process.* **2010**, *18*, 1406–1416. https://doi.org/10.1109/TASL.2009.2034776.
25. Liu, Z.; Huang, J.; Wang, Y. Classification TV programs based on audio information using hidden Markov model. In Proceedings of the 1998 IEEE 2nd Workshop on Multimedia Signal Processing, Redondo Beach, CA, USA, 7–9 December 1998; pp. 27–32. https://doi.org/10.1109/MMSP.1998.738908.
26. Laptev, I.; Lindeberg, T. Space-time interest points. In Proceedings of the IEEE International Conference on Computer Vision, 2003, Nice, France, 13–16 October 2003; Volume 1, pp. 432–439. https://doi.org/10.1109/iccv.2003.1238378.
27. 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. https://doi.org/10.1109/ICCV.2013.441.
28. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the ACM International Multimedia Conference and Exhibition, Augsburg, Germany, 25–29 September 2007; pp. 357–360. https://doi.org/10.1145/1291233.1291311.
29. Kläser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the BMVC 2008—British Machine Vision Conference 2008, Leeds, UK, September 2008. https://doi.org/10.5244/C.22.99.
30. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3952 LNCS, pp. 428–441. https://doi.org/10.1007/11744047_33.
31. Sadanand, S.; Corso, J.J. Action bank: A high-level representation of activity in video. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1234–1241. https://doi.org/10.1109/CVPR.2012.6247806.

32. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; Volume 2005, pp. 65–72. https://doi.org/10.1109/VSPETS.2005.1570899.
33. Willems, G.; Tuytelaars, T.; Van Gool, L. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5303 LNCS, pp. 650–663. https://doi.org/10.1007/978-3-540-88688-4_48.
34. Wang, L.; Qiao, Y.; Tang, X. Video action detection with relational dynamic-poselets. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693 LNCS, pp. 565–580. https://doi.org/10.1007/978-3-319-10602-1_37.
35. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314. https://doi.org/10.1109/CVPR.2015.7299059.
36. Kar, A.; Rai, N.; Sikka, K.; Sharma, G. AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5699–5708. https://doi.org/10.1109/CVPR.2017.604.
37. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal multiplier networks for video action recognition. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 7445–7454. https://doi.org/10.1109/CVPR.2017.787.
38. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3D residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
39. Wang, L. et al. Temporal segment networks: Towards good practices for deep action recognition. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9912 LNCS, pp. 20–36. https://doi.org/10.1007/978-3-319-46484-8_2.
40. Wang, Y.; Long, M.; Wang, J.; Yu, P.S. Spatiotemporal pyramid network for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
41. Lan, Z.; Zhu, Y.; Hauptmann, A.G.; Newsam, S. Deep Local Video Feature for Action Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017; pp. 1219–1225. https://doi.org/10.1109/CVPRW.2017.161.
42. Duta, I.C.; Ionescu, B.; Aizawa, K.; Sebe, N. Spatio-temporal vector of locally max pooled features for action recognition in videos. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3205–3214. https://doi.org/10.1109/CVPR.2017.341.
43. Shen, J.; Huang, Y.; Wen, M.; Zhang, C. Toward an Efficient Deep Pipelined Template-Based Architecture for Accelerating the Entire 2-D and 3-D CNNs on FPGA. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2020**, *39*, 1442–1455. https://doi.org/10.1109/TCAD.2019.2912894.
44. Duta, I.C.; Nguyen, T.A.; Aizawa, K.; Ionescu, B.; Sebe, N. Boosting VLAD with double assignment using deep features for action recognition in videos. In Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2016; pp. 2210–2215. https://doi.org/10.1109/ICPR.2016.7899964.
45. Xu, Z.; Yang, Y.; Hauptmann, A.G. A discriminative CNN video representation for event detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1798–1807. https://doi.org/10.1109/CVPR.2015.7298789.
46. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 971–980.
47. Ballas, N.; Yao, L.; Pal, C.; Courville, A. Delving deeper into convolutional networks for learning video representations. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016—Conference Track Proceedings, San Juan, PR, USA, 2–4 May 2016.
48. Donahue, J. *et al.* Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878.
49. Srivastava, N.; Mansimov, E.; Salakhutdinov, R. Unsupervised learning of video representations using LSTMs. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015; Volume 1, pp. 843–852.
50. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101.
51. Taylor, G.W.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional learning of spatio-temporal features. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6316 LNCS, pp. 140–153. https://doi.org/10.1007/978-3-642-15567-3_11.
52. Le, Q.V.; Zou, W.Y.; Yeung, S.Y.; Ng, A.Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3361–3368. https://doi.org/10.1109/CVPR.2011.5995496.

53. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 7065 LNCS, pp. 29–39. https://doi.org/10.1007/978-3-642-25446-8_4.
54. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. https://doi.org/10.1109/TPAMI.2012.59.
55. Zha, S.; Luisier, F.; Andrews, W.; Srivastava, N.; Salakhutdinov, R. Exploiting Image-trained CNN Architectures for Unconstrained Video Classification. In Proceedings of the BMVC, Swansen, UK, 7–10 September 2015; pp. 60.1–60.13. https://doi.org/10.5244/c.29.60.
56. Carreira, J.; Zisserman, A. Quo Vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733. https://doi.org/10.1109/CVPR.2017.502.
57. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.F. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732. https://doi.org/10.1109/CVPR.2014.223.
58. 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; Volume 2015, pp. 4489–4497. https://doi.org/10.1109/ICCV.2015.510.
59. Shu, X.; Tang, J.; Qi, G.-J.; Liu, W.; Yang, J. Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1110–1118. https://doi.org/10.1109/TPAMI.2019.2942030.
60. Shu, X.; Zhang, L.; Qi, G.-J.; Liu, W.; Tang, J. Spatiotemporal Co-Attention Recurrent Neural Networks for Human-Skeleton Motion Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3300–3315. https://doi.org/10.1109/TPAMI.2021.3050918.
61. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* 2**014**, *1*, 568–576.
62. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016; pp. 1933–1941. https://doi.org/10.1109/CVPR.2016.213.
63. Wu, Z.; Jiang, Y.-G.; Wang, X.; Ye, H.; Xue, X.; Wang, J. Fusing Multi-Stream Deep Networks for Video Classification. *arXiv* **2015**, arXiv:1509.06086.
64. 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.
65. Shan, K.; Wang, Y.; Tang, Z.; Chen, Y.; Li, Y. MixTConv: Mixed Temporal Convolutional Kernels for Efficient Action Recognition. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1751–1756. https://doi.org/10.1109/icpr48806.2021.9412586.
66. Wu, Z.; Wang, X.; Jiang, Y.G.; Ye, H.; Xue, X. Modeling spatial-Temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the MM 2015—2015 ACM Multimedia Conference, Brisbane, Australia, 26–30 October 2015; pp. 461–470. https://doi.org/10.1145/2733373.2806222.
67. Tanberk, S.; Kilimci, Z.H.; Tukel, D.B.; Uysal, M.; Akyokus, S. A Hybrid Deep Model Using Deep Learning and Dense Optical Flow Approaches for Human Activity Recognition. *IEEE Access* **2020**, *8*, 19799–19809. https://doi.org/10.1109/ACCESS.2020.2968529.
68. Alhersh, T.; Stuckenschmidt, H.; Rehman, A.U.; Belhaouari, S.B. Learning Human Activity From Visual Data Using Deep Learning. *IEEE Access* **2021**, *9*, 106245–106253. https://doi.org/10.1109/access.2021.3099567.
69. Kopuklu, O.; Kose, N.; Gunduz, A.; Rigoll, G. Resource efficient 3D convolutional neural networks. In Proceedings of the 2019 International Conference on Computer Vision Workshop, ICCVW 2019, Seoul, Korea, 27–28 October 2019; pp. 1910–1919. https://doi.org/10.1109/ICCVW.2019.00240.
70. Liu, H.; Bhanu, B. Pose-guided R-CNN for jersey number recognition in sports. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 2457–2466. https://doi.org/10.1109/CVPRW.2019.00301.
71. Huang, G.; Bors, A.G. Region-based non-local operation for video classification. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2020; pp. 10010–10017. https://doi.org/10.1109/ICPR48806.2021.9411997.
72. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
73. Biswas, A.; Jana, A.P.; Mohana; Tejas, S.S. Classification of objects in video records using neural network framework. In Proceedings of the International Conference on Smart Systems and Inventive Technology, ICSSIT 2018, Tirunelveli, India, 13–14 December 2018; pp. 564–569. https://doi.org/10.1109/ICSSIT.2018.8748560.
74. Jana, A.P.; Biswas, A.; Mohana. YOLO based detection and classification of objects in video records. In Proceedings of the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2018, Bangalore, India, 18–19 May 2018, pp. 2448–2452. https://doi.org/10.1109/RTEICT42901.2018.9012375.
75. Zhou, R.; Xia, D.; Wan, J.; Zhang, S. An intelligent video tag recommendation method for improving video popularity in mobile computing environment. *IEEE Access* **2020**, *8*, 6954–6967. https://doi.org/10.1109/ACCESS.2019.2961392.

76. Khan, U.A.; Martinez-Del-Amor, M.A.; Altowaijri, S.M.; Ahmed, A.; Rahman, A.U.; Sama, N.U.; Haseeb, K.; Islam, N. Movie Tags Prediction and Segmentation Using Deep Learning. *IEEE Access* **2020**, *8*, 6071–6086. https://doi.org/10.1109/ACCESS.2019.2963535.
77. Apostolidis, E.; Adamantidou, E.; Mezaris, V.; Patras, I. Combining adversarial and reinforcement learning for video thumbnail selection. In Proceedings of the ICMR 2021—2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 1–9. https://doi.org/10.1145/3460426.3463630.
78. Carta, S.; Giuliani, A.; Piano, L.; Podda, A.S.; Recupero, D.R. VSTAR: Visual Semantic Thumbnails and tAgs Revitalization. *Expert Syst. Appl.* **2022**, *193*, 116375. https://doi.org/10.1016/j.eswa.2021.116375.
79. Yang, Z.; Lin, Z. Interpretable video tag recommendation with multimedia deep learning framework. *Internet Res.* **2022**, *32*, 518–535. https://doi.org/10.1108/INTR-08-2020-0471.
80. Wang, Y.; Yan, J.; Ye, X.; Jing, Q.; Wang, J.; Geng, Y. Few-Shot Transfer Learning With Attention Mechanism for High-Voltage Circuit Breaker Fault Diagnosis. *IEEE Trans. Ind. Appl.* **2022**, *58*, 3353–3360. https://doi.org/10.1109/TIA.2022.3159617.
81. Zhong, C.; Wang, J.; Feng, C.; Zhang, Y.; Sun, J.; Yokota, Y. PICA: Point-wise Instance and Centroid Alignment Based Few-shot Domain Adaptive Object Detection with Loose Annotations. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 398–407. https://doi.org/10.1109/WACV51458.2022.00047.
82. Zhang, A.; Liu, F.; Liu, J.; Tang, X.; Gao, F.; Li, D.; Xiao, L. Domain-Adaptive Few-Shot Learning for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**. https://doi.org/10.1109/LGRS.2022.3217502.
83. Zhao, A. *et al.* Domain-Adaptive Few-Shot Learning. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021; pp. 1389–1398. https://doi.org/10.1109/WACV48630.2021.00143.
84. Gao, J.; Xu, C. CI-GNN: Building a Category-Instance Graph for Zero-Shot Video Classification. *IEEE Trans. Multimedia* **2020**, *22*, 3088–3100. https://doi.org/10.1109/TMM.2020.2969787.
85. Zhu, L.; Yang, Y. Compound Memory Networks for Few-Shot Video Classification. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11211, pp. 782–797. https://doi.org/10.1007/978-3-030-01234-2_46.
86. Hu, Y.; Gao, J.; Xu, C. Learning Dual-Pooling Graph Neural Networks for Few-Shot Video Classification. *IEEE Trans. Multimedia* **2021**, *23*, 4285–4296. https://doi.org/10.1109/TMM.2020.3039329.
87. Cao, K.; Ji, J.; Cao, Z.; Chang, C.-Y.; Niebles, J.C. Few-Shot Video Classification via Temporal Alignment. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10615–10624. https://doi.org/10.1109/CVPR42600.2020.01063.
88. Fu, Y.; Zhang, L.; Wang, J.; Fu, Y.; Jiang, Y.-G. Depth Guided Adaptive Meta-Fusion Network for Few-shot Video Recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1142–1151. https://doi.org/10.1145/3394171.3413502.
89. Zhang, H.; Zhang, L.; Qi, X.; Li, H.; Torr, P.H.S.; Koniusz, P. Few-Shot Action Recognition with Permutation-Invariant Attention. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12350, pp. 525–542. https://doi.org/10.1007/978-3-030-58558-7_31.
90. Qi, M.; Qin, J.; Zhen, X.; Huang, D.; Yang, Y.; Luo, J. Few-Shot Ensemble Learning for Video Classification with SlowFast Memory Networks. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3007–3015. https://doi.org/10.1145/3394171.3416269.
91. Fu, Y.; Wang, C.; Fu, Y.; Wang, Y.X.; Bai, C.; Xue, X.; Jiang, Y.G. Embodied One-Shot Video Recognition. In Proceedings of the 27th ACM International Conference on Multimedia, Nice France, 21–25 October 2019; pp. 411–419. https://doi.org/10.1145/3343031.3351015.
92. Bishay, M.; Zoumpourlis, G.; Patras, I. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv* **2019**, arXiv:1907.09021.
93. Feng, Y.; Gao, J.; Xu, C. Learning Dual-Routing Capsule Graph Neural Network for Few-shot Video Classification. *IEEE Trans. Multimedia* **2022**, 1. https://doi.org/10.1109/TMM.2022.3156938.
94. Shu, X.; Xu, B.; Zhang, L.; Tang, J. Multi-Granularity Anchor-Contrastive Representation Learning for Semi-Supervised Skeleton-Based Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, 1–18. https://doi.org/10.1109/TPAMI.2022.3222871.
95. Xu, B.; Shu, X.; Song, Y. X-Invariant Contrastive Augmentation and Representation Learning for Semi-Supervised Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 3852–3867. https://doi.org/10.1109/TIP.2022.3175605.
96. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D ShapeNets: A deep representation for volumetric shapes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015 pp. 1912–1920. https://doi.org/10.1109/CVPR.2015.7298801.
97. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; pp. 945–953. https://doi.org/10.1109/ICCV.2015.114.
98. Cao, W.; Yan, Z.; He, Z.; He, Z. A Comprehensive Survey on Geometric Deep Learning. *IEEE Access* **2020**, *8*, 35929–35949. https://doi.org/10.1109/ACCESS.2020.2975067.

99. Masci, J.; Boscaini, D.; Bronstein, M.M.; Vandergheynst, P. Geodesic Convolutional Neural Networks on Riemannian Manifolds. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 832–840. https://doi.org/10.1109/ICCVW.2015.112.
100. Boscaini, D.; Masci, J.; Rodolà, E.; Bronstein, M. Learning shape correspondence with anisotropic convolutional neural networks. *Adv. Neural Inf. Process. Syst* **2016**, *29*, 3197–3205.
101. Monti, F.; Boscaini, D.; Masci, J.; Rodolà, E.; Svoboda, J.; Bronstein, M.M. Geometric deep learning on graphs and manifolds using mixture model CNNs. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5425–5434. https://doi.org/10.1109/CVPR.2017.576.
102. Litany, O.; Remez, T.; Rodola, E.; Bronstein, A.; Bronstein, M. Deep Functional Maps: Structured Prediction for Dense Shape Correspondence. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5660–5668. https://doi.org/10.1109/ICCV.2017.603.
103. Boscaini, D.; Masci, J.; Melzi, S.; Bronstein, M.M.; Castellani, U.; Vandergheynst, P. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. *Eurographics Symp. Geom. Process.* **2015**, *34*, 13–23. https://doi.org/10.1111/cgf.12693.
104. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 77–85. https://doi.org/10.1109/CVPR.2017.16.
105. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5100–5109.
106. Li, Y.; Cao, W. An Extended Multilayer Perceptron Model Using Reduced Geometric Algebra. *IEEE Access* **2019**, *7*, 129815–129823. https://doi.org/10.1109/ACCESS.2019.2940217.
107. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42. https://doi.org/10.1109/MSP.2017.2693418.
108. Wu, Z.; Jiang, Y.G.; Wang, J.; Pu, J.; Xue, X. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In Proceedings of the MM 2014—2014 ACM Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 167–176. https://doi.org/10.1145/2647868.2654931.
109. Yang, Y.; Krompass, D.; Tresp, V. Tensor-train recurrent neural networks for video classification. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; Volume 8, pp. 5929–5938.
110. Yang, X.; Molchanov, P.; Kautz, J. Multilayer and multimodal fusion of deep neural networks for video classification. In Proceedings of the MM 2016—2016 ACM Multimedia Conference, Amsterdam, The Netherlands, 15–29 October 2016; pp. 978–987. https://doi.org/10.1145/2964284.2964297.
111. Wang, L.; Li, W.; Li, W.; Van Gool, L. Appearance-and-relation networks for video classification. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1430–1439.
112. Wu, Z.; Jiang, Y.G.; Wang, X.; Ye, H.; Xue, X. Multi-stream multi-class fusion of deep networks for video classification. In Proceedings of the MM 2016—Proceedings of the 2016 ACM Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; pp. 791–800. https://doi.org/10.1145/2964284.2964328.
113. Tran, D.; Wang, H.; Torresani, L.; Feiszli, M. Video classification with channel-separated convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 Octobet–2 November 2019; pp. 5552–5561.
114. Jing, L.; Parag, T.; Wu, Z.; Tian, Y.; Wang, H. VideoSSL: Semi-Supervised Learning for Video Classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1110–1119.
115. Shu, X.; Yang, J.; Yan, R.; Song, Y. Expansion-Squeeze-Excitation Fusion Network for Elderly Activity Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5281–5292. https://doi.org/10.1109/TCSVT.2022.3142771.
116. Li, Z.; Li, R.; Jin, G. Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary. *IEEE Access* **2020**, *8*, 75073–75084. https://doi.org/10.1109/ACCESS.2020.2986582.
117. Zhen, M. et al. Learning Discriminative Feature with CRF for Unsupervised Video Object Segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12372 LNCS, pp. 445–462. https://doi.org/10.1007/978-3-030-58583-9_27.
118. Ruz, G.A.; Henríquez, P.A.; Mascareño, A. Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Gener. Comput. Syst.* **2020**, *106*, 92–104. https://doi.org/10.1016/j.future.2020.01.005.
119. Fantinel, R.; Cenedese, A.; Fadel, G. Hybrid Learning Driven by Dynamic Descriptors for Video Classification of Reflective Surfaces. *IEEE Trans. Industr. Inform.* **2021**, *17*, 8102–8111. https://doi.org/10.1109/TII.2021.3062619.
120. Costa, F.F.; Saito, P.T.M.; Bugatti, P.H. Video action classification through graph convolutional networks. In Proceedings of the VISIGRAPP 2021—16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Vienna, Austria, 8–10 February 2021; Volume 4, pp. 490–497. https://doi.org/10.5220/0010321304900497.
121. Xu, Q.; Zhu, L.; Dai, T.; Yan, C. Aspect-based sentiment classification with multi-attention network. *Neurocomputing* **2020**, *388*, 135–143. https://doi.org/10.1016/j.neucom.2020.01.024.
122. Bibi, M.; Aziz, W.; Almaraashi, M.; Khan, I.H.; Nadeem, M.S.A.; Habib, N. A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis. *IEEE Access* **2020**, *8*, 68580–68592. https://doi.org/10.1109/ACCESS.2020.2983859.

123. Sailunaz, K.; Alhajj, R. Emotion and sentiment analysis from Twitter text. *J. Comput. Sci.* **2019**, *36*, 101003. https://doi.org/10.1016/j.jocs.2019.05.009.
124. 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.
125. Cai, Z.; Wang, L.; Peng, X.; Qiao, Y. Multi-view super vector for action recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 596–603. https://doi.org/10.1109/CVPR.2014.83.
126. 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. https://doi.org/10.1109/ICCV.2015.522.
127. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. C3D: Generic Features for Video Analysis. 2015. Available online: https://vlg.cs.dartmouth.edu/c3d/ (accessed on 20 January 2023)
128. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Comput. Vis. Image Underst.* **2016**, *150*, 109–125. https://doi.org/10.1016/j.cviu.2016.03.013.
129. Lev, G.; Sadeh, G.; Klein, B.; Wolf, L. RNN fisher vectors for action recognition and image annotation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9910 LNCS, pp. 833–850. https://doi.org/10.1007/978-3-319-46466-4_50.
130. Park, E.; Han, X.; Berg, T.L.; Berg, A.C. Combining multiple sources of knowledge in deep CNNs for action recognition. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, 7–10 March 2016. https://doi.org/10.1109/WACV.2016.7477589.
131. Wang, X.; Farhadi, A.; Gupta, A. Actions ~ Transformations. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2658–2667. https://doi.org/10.1109/CVPR.2016.291.
132. Zhu, W.; Hu, J.; Sun, G.; Cao, X.; Qiao, Y. A Key Volume Mining Deep Framework for Action Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1991–1999. https://doi.org/10.1109/CVPR.2016.219.

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.