Deep Learning for Video Classification: A Review

Atiq ur Rehman, *Member, IEEE*, and Samir Brahim Belhaouari, *Senior Member, IEEE*

***Abstract*— Video classification task has gained a 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 to the importance of 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 a number of existing reviews and survey papers related to video classification in the scientific literature. However, the existing review papers are either outdated, and therefore, do not include the recent state-of-art works or they have some limitations. In order 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 data sets 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, and 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.**

***Impact Statement*— The importance of accurate video classification task can be realized by the large amount of video data available online. Moreover, with an increase in the availability of large-scale video data, deep learning methods have demonstrated a high recognition accuracy for classification of videos. In recognition to the importance of automatic video classification using deep learning, this paper provides a comprehensive review on the topic. The paper addresses the limitations of existing reviews, provides an updated review on the state-of-art approaches, and offers some useful future research directions to solve the challenging video classification problem.**

***Index Terms*— Automatic Video Classification, Deep Learning, Handcrafted Features, Video Processing.**

I. Introduction

VIDEO classification task has gained a 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. The importance of 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 is being watched by different people on every single day. In recognition to the importance of video classification task, a combined effort is being made by the researchers for proposing an accurate video classification framework. Companies like Google AI are investing in different competitions to solve the challenging problem under constrained conditions. To further advance the progress of 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 of a powerful video classification model.

An Artificial Neural Network (ANN) is an algorithm based on the interconnected nodes to recognize the relationships in a set of data. Algorithms based on ANNs have shown a great success in modeling both the linear and the non-linear relationships in the underlying data. Due to a 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. There are a number of existing reviews and survey papers related to video classification in the scientific literature. However, those review papers are either old and therefore do not include the recent state-of-art works or they have some limitations. Some of the recent reviews on video classification with their limitations are discussed as follows: (i) Z. Wu [5] 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. (ii) Q. Ren [6] conducted a recent and simple review on video classification methods, however the techniques covered in this review are not well

[A]tiq ur Rehman is with the ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar (email: atrehman2@hbku.edu.qa; atiqjadoon@gmail.com).

Samir Brahim Belhaouari, is also with the ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar (email: sbelhaouari@hbku.edu.qa).

described and the review also lacks in the description of research gaps, benchmark datasets, limitations of existing methods, and performance metrics. (iii) A more recent review is done by A. Anusya [7], this review covers few methods for video classification, clustering, and tagging. However, the review provided is not comprehensive and lacks in concise information, coverage of topic, datasets, analysis of state-of-art approaches, and research limitations. (iv) Rani et al [8] also conducted a recent review on video classification methods, their review covers some recent video classification approaches and summary-based description of some recent works. This review has also some limitations including the missing analysis of recent state-of-art approaches, and short description of topics covered. (v) Y. Li et al [9] have 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. (vi) A recent review is also done by Md Islam et al [10], in this review they have included all the methods for video classification including deep learning. However, as the focus of review is not on deep learning approaches, therefore these methods are not completely covered in this review.

In contrast to the existing reviews on classification of videos, this paper provides a more comprehensive, concise and up-to-dated 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 breakthrough in 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 II reviews some existing Convolutional Neural Networks (CNNs) for images, Section III provides an in-depth review on Deep Learning models for Video classification, Section IV provides a summary for benchmark datasets, evaluation metrics, and comparison of existing state-of-art for video classification task, and Section V provides conclusion and future research directions.

II. Convolutional Neural Networks (CNNs) 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 [11], AlexNet [12], VGGNet [13], GoogleNet [14], ResNet [15], and DenseNet [16]. 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 Fig. 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 [17]. Although, deeper networks are better in terms of having more discriminatory powers, but the deeper networks require more data for training and more parameters to tune. Finding a professionally labeled huge dataset is still a big challenge faced by the research community and therefore it limits the development of more deeper neural networks.

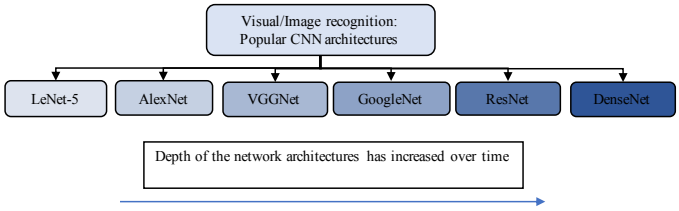

Fig. 1: State-of-art image recognition CNN networks. The trend is that the depth and discriminatory powers of network architectures increases from formerly proposed architectures towards the recently proposed architectures.

III. Video Classification

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

A. 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 Uni-modal video classification or Multi-modal video classification, as summarized in Fig. 2. The existing has literature utilized both these models for the video classification task and it is generally believed that models utilizing Multi-modal data perform better than the models based on Uni-modal data. Moreover, the visual

description of a video works better than the text and the audio description for the classification purpose of a video.

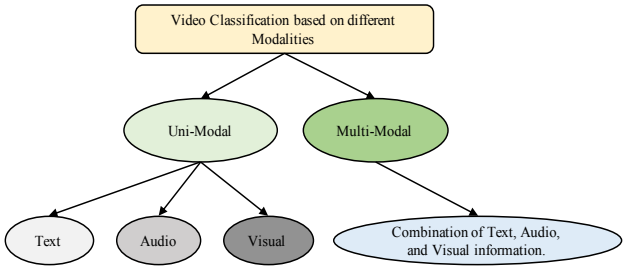

Fig. 2. : Different modalities used for classification of videos.

B. *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 literature are spatiotemporal interest points (STIPs) [18], improved Dense Trajectories (iDT) [19], SIFT-3D [20], HOG3D [21], Motion Boundary Histogram [22], Action- Bank [23], Cuboids [24], 3D SURF [25], and Dynamic-Poselets [26]. These hand-designed representations use different feature encoding schemes such as the ones based on pyramids and histograms. iDT is one of these hand-crafted representations, which is widely considered the state-of-the-art. Many recent competitive studies demonstrated that hand-crafted features [27]–[30], high-level [31], [32], and mid-level [33], [34] video representations have contributed towards the task of video classification with deep neural networks.

C. *Deep Learning frameworks*

Along with the development of more powerful deep learning architectures in the recent years, the trend for 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 [35]. 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. In order 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 multi-stream network model, where the temporal stream is processed using optical flow, is shown in Figure 4 [36]. 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 breakthrough 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.

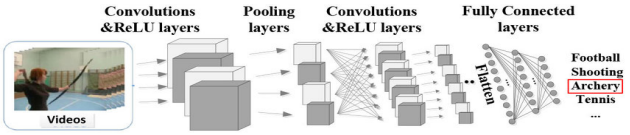

Fig. 3: An example of 3D-CNN architecture to classify videos [35].

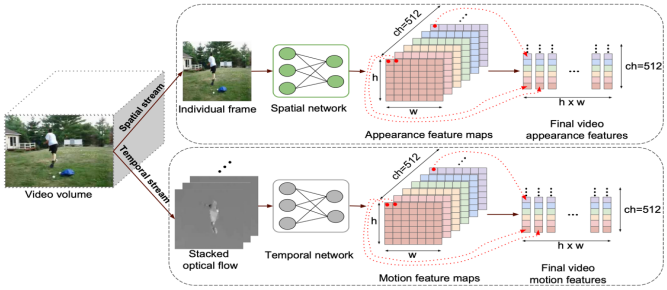

Fig. 4: An example of two stream architecture with optical flow [36].

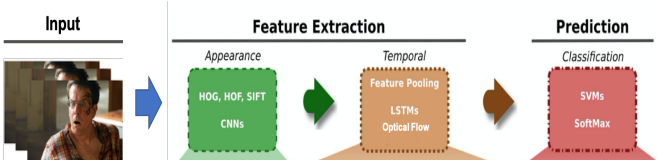

Figure 5: An overview of video classification process.

D. *Breakthrough*

The breakthrough in recognition of still-images originated with the introduction of deep learning model called AlexNet [37]. 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 [38], [39], or through the RNN architectures [40]–[43]. For un-supervised spatiotemporal feature learning in 3D convolutions, Restricted Boltzmann Machines [44] and stacked ISA [45] are also studied in parallel. The 3D CNNs using temporal convolutions to extract temporal features automatically were first proposed by Baccouche et al. [46] and by Ji et al. [47].

E. *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 comes 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 limit the understanding of video to only spatial domain. RNNs on the other hand have the ability to 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.

F. Developments In Video Classification Over Time

The existing approaches for video classification are categorized based on their working principle in Table I. 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.

Among the initially developed hand-crafted representations, improved Dense Trajectories (iDT) [19] is widely considered the state-of-the-art. Whereas, many recent competitive studies demonstrated that hand-crafted features [27]–[30], high-level [31], [32], and mid-level [33], [34] 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.

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. Besides these architectures, there are also hybrid approaches based on the integration of CNN and RNN architectures. A summary of these architectures in provided in Fig. 6.

TABLE I
DIFFERENT CATEGORIES OF APPROACHES FOR 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) [18], iDT [19], SIFT-3D [20], HOG3D [21], Motion Boundary Histogram [22], Cuboids [24], Action-Bank [23], 3D SURF [25], Dynamic-Poselets [26].
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.	[48]
3D-CNNs	2D image classification extension to 3D for video (For example the Inception 3D (I3D) architecture).	[49]
Spatiotemporal Convolutional Networks	To aggregate the temporal and the spatial information, these methods primarily depend on convolution and pooling.	[50], [47], [51]
Recurrent Spatial Networks	To represent temporal information in videos, recurrent neural networks such as LSTM or GRU are used.	[52], [40]
Two/multi Stream Networks	In addition to the context frame visuals, these methods use layered optical flow to identify movements.	[53], [54], [55], [43]
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.	[56], [57]
Hybrid Approaches	These are models based on integration of CNN and RNN architectures.	[58], [59], [60]

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 Fig. 7.

G. Summary of some notable deep learning frameworks

A summary of some deep learnings architectures for video classification is provided in Table II. These studies are summarized based on the architecture, the datasets, the evaluation metrics, the fusion strategy, and the notable findings. Some notable findings from these studies are as follows: (i) The architectures employing CNN/RNN for feature extraction have the ability to perform better than hand-crafted features provided that enough data is available for training. (ii) Tensor-Train Layer based RNN like LSTM and GRU perform better than the plain RNN architectures for video classification. (iii) It is sometimes necessary to use optical flow for datasets like UCF-101. (iv) It is not always helpful to use optical flow, especially for the case of videos taken from wild e-g Sports-1M. (v) It is important to use a sophisticated sequence processing architecture like LSTM to take advantage of optical flow. (vi) LSTMs when applied on both the optical flow and the image frames yield the highest performance measure for Sports-1M benchmark dataset. (vii) Augmenting optical flow and RGB input helps in improving the performance. (viii) Optical flow modality provides complementary information. (ix) A high computational requirement of optical flow limits its use in real-time systems. (x) Multi-Stream Multi-Class fusion can perform better than Average fusion, Weighted fusion, Kernel average fusion, MKL fusion, and Logistic regression fusion on datasets like UCF-101 and CCV. (xi) In 3D group convolutional networks, the volume of channel interactions play a vital role in achieving a high accuracy. (xii) 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. (xiii) 3D channel-separated convolutions results in a kind of regularization and prevents overfitting. (xiv) Popular frameworks of conventional semi-supervised algorithms (which were originally developed for 2D images) are unable to obtain good results for 3D video categorization. (xv) For semi-supervised learning, a calibrated employment of the object appearance cues keenly improves the accuracy of the 3D-CNN models.

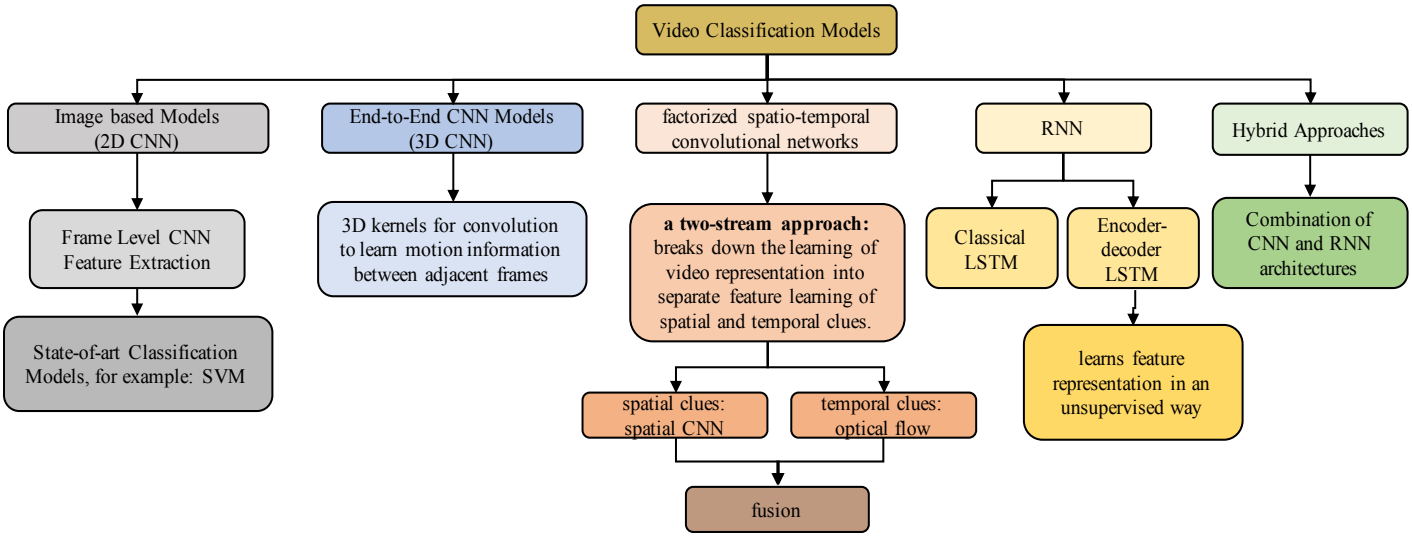

Figure. 6: Summary of video classification approaches.

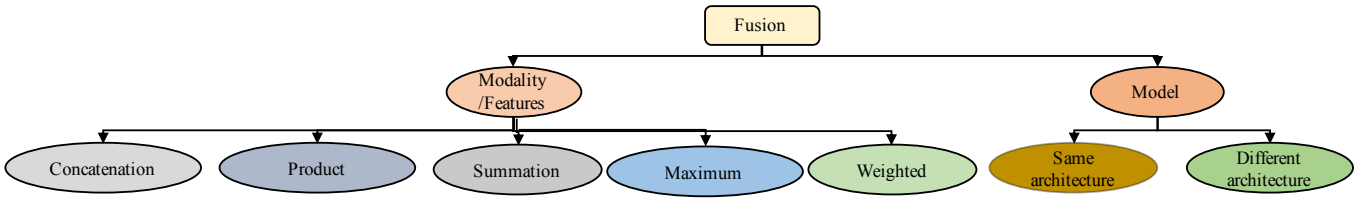

Figure. 7: Different Fusion Types.

TABLE II
SUMMARY AND FINDINGS OF STUDIES BASED ON DEEP LEARNING MODELS

Study	features	Model	Evaluation	Dataset	Problem	Fusion	Findings
[91]	Automatic Spatio-temporal features/ Self learning.	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 like Texton, HOG, Cuboids etc and global features like color moments , and Hue- Saturation).
[92]	**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. **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.
[93]	Tensor-Train Factorization	Recurrent neural network (RNN)	Classification accuracy.	UCF11, Hollywood2, Youtube Celebrities Face Data.	Multi-class	-	Tensor-Train Layer based RNN like LSTM and GRU perform better than the plain RNN architectures for video classification.
[87]	improved Fisher vector (iFV) and explicit feature maps to represent features of conv and fc layers.	A multilayer and multimodal fusion framework of deep neural networks based on Fully connected (FC)-RNN	Classification accuracy.	UCF101, HMDB51.	Multi-class	Multilayer and multimodal fusion framework.	When compared to enhanced dense trajectories, which require a number of hand-crafted procedures such as dense point tracking, camera motion estimation, person detection, and so on, the proposed FC-RNN obtained competitive results.
[43]	Convolutional temporal feature pooling architectures (Conv Pooling, Late Pooling, Slow Pooling, Local Pooling).	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 isn't always beneficial, especially when the videos are captured in the wild, such as Sports-1M. (iii) To take 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.
[78]	Spatiotemporal feature learning: a SMART block to learn spatiotemporal features.	ARTNet by integrating the SMART block into the C3D-ResNet18 architecture. Where, SMART block architecture is composed of appearance branch and relation branch.	Top-1 and Top-5 accuracy.	Kinetics, UCF101, and HMDB51.	Multi-class	Concatenation and reduction operation.	(i) In terms of spatiotemporal feature learning, SMART blocks outperform 3D convolutions (3D CNN). (ii) In the case of ARTNet, supplementing RGB input with 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.
[89]	Spatial, short-term motion and audio clues using CNN. (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.
[94]	Two distinct layers: 1×1×1 conventional convolutions for channel interaction (but no local interaction) and k×k×k depthwise convolutions for local spatiotemporal interactions (but not channel interaction).	Channel-Separated Convolutional Network (CSN). Two models: interaction-preserved channel-separated network (ip-CSN). 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.
[79]	The 3D network is optimized with three loss functions: (i) cross-entropy (CE) loss, (ii) Pseudo CE Loss, and (iii) Soft CE Loss.	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.

H. Geometric Deep Learning

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 literature to apply deep learning on geometric data: (i) extrinsic methods, and (ii) intrinsic methods. An example of these two methods is illustrated in Fig. 8 [61]. 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 [62], (ii) Anisotropic CNN [63], (iii) Mixture model network [64], (iv) Structured Prediction Model [65], (v) Localized Spectral CNN [66], (vi) PointNet [67], (vii) PointNet++ [68], and (viii) RGA-MLP [69]. 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.

Fig. 8: Illustration of deep learning approaches on geometric data. (a) extrinsic method and (b) intrinsic method.

IV. Benchmark Datasets, Evaluation Metrics, and Comparison of existing state-of-art For Video Classification

A. Benchmark Datasets For Video Classification

There are several benchmark datasets being utilized for classification of videos, some of these notable datasets are summarized in Table III. 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.

B. Performance evaluation metrics for Video Classification

The evaluation of video classification models is done using different performance measures. The most common measures utilized to evaluate the models are Accuracy, Precision, Recall, F1 score, Micro F1, and K-fold [10]. Some of the recent studies using these measures are listed in Table IV.

TABLE III
COMMONLY USED EVALUATION METRICS FOR VIDEO CLASSIFICATION

Evaluation Metric	Year of Publication	Reference
Accuracy	2020-2021	[70]–[74]
Precision	2020-2021	[70], [72], [73]
Recall	2020-2021	[70], [72], [73]
F1 Score	2020-2021	[70], [72], [73]
Micro F1	2020	[75], [76]
K-Fold	2019	[77]
Top-k	2018,2021	[78], [79]

TABLE IV
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	234414	15	2009	Dynamic
Olympic Sports	800	16	2010	Dynamic
HMDB51	6766	51	2011	Dynamic
CCV	9317	20	2011	Dynamic
UCF-101	13320	101	2012	Dynamic
THUMOS-2014	18394	101	2014	Dynamic
MED-2014 (Dev. set)	31000	20	2014	Dynamic
Sports-1M	1133158	487	2014	Dynamic
ActivityNet	27901	203	2015	Dynamic
EventNet	95321	500	2015	Dynamic
MPII Human Pose	20943	410	2014	Dynamic
FCVID	91223	239	2015	Dynamic
UCF11	1600	11	2009	Dynamic
Youtube Celebrities Face	1910	47	2008	Dynamic
Kinetics	300000	400	2017	Dynamic
Youtube-8M	6.1 M	3862	2018	Dynamic
JHMDB	928	21	2011	Dynamic
Something-something	110000	174	2017	Dynamic

C. 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 [80], the videos in the dataset are collected from the YouTube. Total videos in the dataset are 13320 with 101 action categories. The dataset is challenging because of the uncontrolled environment in the captured videos and it is widely being used by researches working on 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 V, where the methods are arranged in ascending order based on the performance. The results reported in Table V are taken from the existing studies in the literature.

TABLE V
COMPARISON OF VIDEO CLASSIFICATION METHODS ON UCF-101

Method	Accuracy
LRCN [41]	82.9
DT + MVSV [81]	83.5
LSTM - Composite [42]	84.3
FSTCN [82]	88.1
C3D [83]	85.2
iDT + HSV [84]	87.9
Two-Stream [53]	88.0
RNN-FV [85]	88.0
LSTM [43]	88.6
MultiSource CNN [86]	89.1
Image-Based [48]	89.6
TDD [27]	90.3
Multilayer and Multimodal Fusion [87]	91.6
Transformation CNN [88]	92.4
Multi-Stream [89]	92.6
Key Volume Mining [90]	92.7
Convolutional Two-Stream [54]	93.5
Temporal Segment Networks [31]	94.2

V. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

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. From the analysis of the existing literature, following conclusions 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 hand-crafted features provided that enough data is available for training. (iii) Tensor-Train Layer based RNN like LSTM and GRU perform better than the plain RNN architectures for video classification. (iv) It is sometimes necessary to use optical flow for datasets like UCF-101. (v) It is not always helpful to use optical flow, especially for the case of videos taken from wild e-g Sports-1M. (vi) It is important to use a sophisticated sequence processing architecture like 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 Sports-1M benchmark dataset. (viii) Augmenting optical flow and RGB input helps in improving the performance. (ix) Optical flow modality provides complementary information. (x) A 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 like UCF-101 and CCV. (xii) In 3D group convolutional networks, the volume of channel interactions play 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) 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.

Although, the latest developments in deep learning models have demonstrated the potential of these approaches for 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 a rich information in the form of spatial, temporal, and acoustic clues. This calls an attention towards the need for a tailored network capable of effectively modeling the spatial, temporal, and acoustic information. Moreover, training CNN/RNN models require labeled datasets and acquiring those datasets are 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 Fig. 9 (a). However, the recent developments in deep learning approaches are still under evaluated and require further investigations for video classification task. One such example is geometric deep learning approaches, the worldwide research interest in this specific topic is shown in Figure 9 (b). Which describes that this topic is still confined to some states of US and has yet to be developed and investigated further. The use of geometric deep learning in extracting rich spatial information from the videos can also be a new research direction for better accuracy in video classification task.

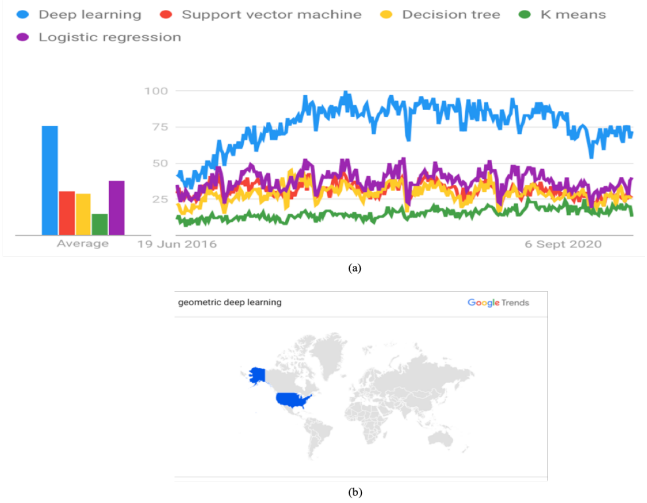

Figure 9: (a) Google trend on deep learning Vs some other state-of-the-art methods. (b) Worldwide research interest in geometric deep learning.

REFERENCES

[1] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[2] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, 2021.

[3] N. Minallah, M. Tariq, N. Aziz, W. Khan, A. ur Rehman, and S. B. Belhaouari, "On the performance of fusion based planet-scope and Sentinel-2 data for crop classification using inception inspired deep convolutional neural network," *PLoS One*, vol. 15, no. 9 September, 2020.

[4] A. U. Rehman and A. Bermak, "Averaging neural network ensembles model for quantification of volatile organic compound," *2019 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2019*, pp. 848–852, 2019.

[5] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," *Front. Multimed. Res.*, pp. 3–29, 2017.

[6] Q. REN *et al.*, "A Survey on Video Classification Methods Based on Deep Learning," *DEStech Trans. Comput. Sci. Eng.*, no. cisnrc, 2019.

[7] A. Anushya, "Video Tagging Using Deep Learning : a Survey," *Int. J. Comput. Sci. Mob. Comput.*, vol. 9, no. 2, pp. 49–55, 2020.

[8] P. Rani, J. Kaur, and S. Kaswan, "Automatic Video Classification: A Review," *EAI Endorsed Trans. Creat. Technol.*, vol. 7, no. 24, p. 163996, 2020.

[9] Y. Li, C. Wang, and J. Liu, "A systematic review of literature on user behavior in video game live streaming," *Int. J. Environ. Res. Public Health*, vol. 17, no. 9, 2020.

[10] M. S. Islam, M. S. Sultana, U. K. Roy, and J. Al Mahmud, "A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 6, no. 2, p. 47, 2021.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Intell. Signal Process.*, pp. 306–351, 2001.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, vol. 2, pp. 1097–1105.

[13] Simonyan.Karen and Zisserman.Andrew, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[14] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 1–9.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 770–778.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 2261–2269.

[17] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, 2020.

[18] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, vol. 1, pp. 432–439.

[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3551–3558, 2013.

[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, 2007, pp. 357–360.

[21] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," *BMVC 2008 - Proc. Br. Mach. Vis. Conf. 2008*, 2008.

[22] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3952 LNCS, pp. 428–441, 2006.

[23] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1234–1241, 2012.

[24] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings - 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS*, 2005, vol. 2005, pp. 65–72.

[25] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5303 LNCS, no. PART 2, pp. 650–663, 2008.

[26] L. Wang, Y. Qiao, and X. Tang, "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)*, 2014, vol. 8693 LNCS, no. PART 5, pp. 565–580.

[27] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 4305–4314.

[28] A. Kar, N. Rai, K. Sikka, and G. Sharma, "AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 5699–5708.

[29] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 7445–7454.

[30] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation

with pseudo-3D residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[31] L. Wang *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)*, 2016, vol. 9912 LNCS, pp. 20–36.

[32] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2097–2106.

[33] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, "Deep Local Video Feature for Action Recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017, vol. 2017-July, pp. 1219–1225.

[34] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe, "Spatio-temporal vector of locally max pooled features for action recognition in videos," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 3205–3214.

[35] J. Shen, Y. Huang, M. Wen, and C. Zhang, "Toward an Efficient Deep Pipelined Template-Based Architecture for Accelerating the Entire 2-D and 3-D CNNs on FPGA," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 39, no. 7, pp. 1442–1455, 2020.

[36] I. C. Duta, T. A. Nguyen, K. Aizawa, B. Ionescu, and N. Sebe, "Boosting VLAD with double assignment using deep features for action recognition in videos," *Proc. - Int. Conf. Pattern Recognit.*, vol. 0, pp. 2210–2215, 2016.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[38] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 1798–1807, 2015.

[39] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 971–980.

[40] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, 2016.

[41] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 2625–2634, 2015.

[42] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 843–852, 2015.

[43] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4694–4702, 2015.

[44] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6316 LNCS, no. PART 6, pp. 140–153, 2010.

[45] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3361–3368, 2011.

[46] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7065 LNCS, pp. 29–39, 2011.

[47] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.

[48] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting Image-trained CNN Architectures for Unconstrained Video Classification," in *BMVC*, 2015, pp. 60.1-60.13.

[49] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4724–4733, 2017.

[50] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 4489–4497.

[52] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "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)*, 2011, vol. 7065 LNCS, pp. 29–39.

[53] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, vol. 1, no. January, pp. 568–576.

[54] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 1933–1941.

[55] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang, "Fusing Multi-Stream Deep Networks for Video Classification," in *CoRR*, 2015.

[56] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[57] K. Shan, Y. Wang, Z. Tang, Y. Chen, and Y. Li, "MixTConv: Mixed Temporal Convolutional Kernels for Efficient Action Recognition," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1751–1756.

[58] Z. Wu, X. Wang, Y. G. Jiang, H. Ye, and X. Xue, "Modeling spatial-Temporal clues in a hybrid deep learning framework for video classification," in *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, 2015, pp. 461–470.

[59] S. Tanberk, Z. H. Kilimci, D. B. Tukel, M. Uysal, and S. Akyokus, "A Hybrid Deep Model Using Deep Learning and Dense Optical Flow Approaches for Human Activity Recognition," *IEEE Access*, vol. 8, pp. 19799–19809, 2020.

[60] T. Alhersh, H. Stuckenschmidt, A. U. Rehman, and S. B. Belhaouari, "Learning Human Activity From Visual Data Using Deep Learning," *IEEE Access*, vol. 9, pp. 106245–106253, 2021.

[61] W. Cao, Z. Yan, Z. He, and Z. He, "A Comprehensive Survey on Geometric Deep Learning," *IEEE Access*, vol. 8, pp. 35929–35949, 2020.

[62] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst, "Geodesic Convolutional Neural Networks on Riemannian Manifolds," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015-Febru, pp. 832–840, 2015.

[63] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 3197–3205, 2016.

[64] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using

mixture model CNNs," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5425–5434, 2017.

[65] O. Litany, T. Remez, E. Rodola, A. Bronstein, and M. Bronstein, "Deep Functional Maps: Structured Prediction for Dense Shape Correspondence," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 5660–5668, 2017.

[66] D. Boscaini, J. Masci, S. Melzi, M. M. Bronstein, U. Castellani, and P. Vandergheynst, "Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks," *Eurographics Symp. Geom. Process.*, vol. 34, no. 5, pp. 13–23, 2015.

[67] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 77–85, 2017.

[68] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 5100–5109, 2017.

[69] Y. Li and W. Cao, "An Extended Multilayer Perceptron Model Using Reduced Geometric Algebra," *IEEE Access*, vol. 7, pp. 129815–129823, 2019.

[70] Z. Li, R. Li, and G. Jin, "Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary," *IEEE Access*, vol. 8, pp. 75073–75084, 2020.

[71] M. Zhen *et al.*, "Learning Discriminative Feature with CRF for Unsupervised Video Object Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12372 LNCS, pp. 445–462, 2020.

[72] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Futur. Gener. Comput. Syst.*, vol. 106, pp. 92–104, 2020.

[73] R. Fantinel, A. Cenedese, and G. Fadel, "Hybrid Learning Driven by Dynamic Descriptors for Video Classification of Reflective Surfaces," *IEEE Trans. Ind. Informatics*, 2021.

[74] F. F. Costa, P. T. M. Saito, and P. H. Bugatti, "Video action classification through graph convolutional networks," *VISIGRAPP 2021 - Proc. 16th Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl.*, vol. 4, pp. 490–497, 2021.

[75] Q. Xu, L. Zhu, T. Dai, and C. Yan, "Aspect-based sentiment classification with multi-attention network," *Neurocomputing*, vol. 388, pp. 135–143, 2020.

[76] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem, and N. Habib, "A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis," *IEEE Access*, vol. 8, pp. 68580–68592, 2020.

[77] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *J. Comput. Sci.*, vol. 36, 2019.

[78] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 1430–1439, 2018.

[79] L. Jing, T. Parag, Z. Wu, Y. Tian, and H. Wang, "VideoSSL: Semi-Supervised Learning for Video Classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1110–1119.

[80] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *CRCV-TR-12-01*, 2012.

[81] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 596–603.

[82] L. Sun, K. Jia, D. Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter. pp. 4597–4605, 2015.

[83] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D : Generic Features for Video Analysis C3D," in *ICCV*, 2015.

[84] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," in *Computer Vision and Image Understanding*, 2016, vol. 150, pp. 109–125.

[85] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "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)*, 2016, vol. 9910 LNCS, pp. 833–850.

[86] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep CNNs for action recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016.

[87] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," *MM 2016 - Proc. 2016 ACM Multimed. Conf.*, pp. 978–987, 2016.

[88] X. Wang, A. Farhadi, and A. Gupta, "Actions ~ Transformations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2658–2667.

[89] Z. Wu, Y. G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," *MM 2016 - Proc. 2016 ACM Multimed. Conf.*, pp. 791–800, 2016.

[90] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A Key Volume Mining Deep Framework for Action Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 1991–1999.

[91] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1725–1732, 2014.

[92] Z. Wu, Y. G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," *MM 2014 - Proc. 2014 ACM Conf. Multimed.*, pp. 167–176, 2014.

[93] Y. Yang, D. Krompass, and V. Tresp, "Tensor-train recurrent neural networks for video classification," *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 8, pp. 5929–5938, 2017.

[94] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

Atiq Ur Rehman (S'17–M'19) received the bachelor's degree in computer engineering (with distinction) from COMSATS University Islamabad, Pakistan, in 2010. He received the master's degree in computer engineering from National University of Sciences and Technology (NUST), Pakistan, in 2013 and PhD degree in Computer Science and Engineering from Hamad Bin Khalifa University, Qatar in 2019. He is currently working as a Post doc researcher with the College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His research interests include the development of evolutionary computation, pattern recognition and machine learning algorithms.

Samir Brahim Belhaouri (SM'19) received the master's degree in telecommunications from the National Polytechnic Institute (ENSEEIHT) of Toulouse, France, in 2000, and the Ph.D. degree in Applied Mathematics from the Federal Polytechnic School of Lausanne (EPFL), in 2006. He is currently an associate professor in the Division of Information and Computing Technology, College of Science and Engineering, HBKU. He also holds and leads several academic and administrator positions, Vice Dean for Academic & Student Affairs at College of Science and General Studies and University Preparatory Program at ALFAISAL university (KSA), University of Sharjah (UAE), Innopolis University (Russia), Petronas University (Malaysia), and EPFL Federal Swiss school (Switzerland). His main research interests include Stochastic Processes, Machine Learning, and Number Theory. He is now working actively on developing algorithms in machine learning applied to visual surveillance, sensing technologies and biomedical data, with the support of several international fund for research in Russia, Malaysia, and in GCC.