International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May - 2021 ISSN: 2582-3930

A review on Human Action Recognition in videos using Deep Learning

Varsha Devaraj[1], Dr. Nachappa MN[2]

[1]Student, Department of MSc- IT, Jain (Deemed-to-be-University)
[2]Professor & Head,Department of CS& IT, Jain(Deemed-to-be-University)

Abstract -Human Action Recognition (HAR) in video plays a vital role in today's world. The aim of HARis to automatically identify and analyse human activities using acquired information from video data. Some of the applications include security and surveillance, smart homes and assisted living, health monitoring, robotics, human–computer interaction, intelligent driving, video-retrieval, gaming and entertainment etc. This paper explores the impact of Deep Learning techniques on action recognition. We also explore how spatiotemporal features are aggregated through various deep architectures, the role of optical flow as an input, the impacts on real-time capabilities, and the compactness & interpretability of the learned features. Although several papers have already been published in the general HAR topics, the growing technologies in the field as well as the multi-disciplinary nature of HAR prompt the need for constant updates in the field. In this respect, this paper attempts to review and summarize the progress of recent advances, and also collectively generalizes the approaches for HAR and compares them in order to present the current state-of-the-art technique for HAR. The main challenges in HAR are also highlighted, along with discussing the benchmark datasets, and propose future directions.

Key Words:HAR, Human action recognition, video, surveillance, templates, journals

1.INTRODUCTION

It is hard to imagine a world without videos. Videos are a simple, popular, accessible, memorable and accurate way of capturing information. It eliminates the need for a narrative description and makes it easier to convey accurate information. Clearly, videos and video cameras have become an integral part of our lives. Video cameras are used almost everywhere e.g. for security, surveillance, for recording recreational or educational activities etc. They are used in the cities, workplaces, homes, schools, hospitals, banks, shops, indoors, outdoors, in the air and even under water. With the growing number of recorded videos and their widespread availability, the need for their computational understanding has become critical. Manual analysis of videos is time consuming and require more workforce. Therefore, there is a need for developing automatic techniques for video understanding and analysis. One important area in video analysis is automatic action recognition which is the focus in this paper.

1.1 Human Action Recognition

Human Action Recognition (HAR), has caught the interests of many researchers since the 1980s due to its broad applicability on different areas and has been extensively studied over the years. Some of the applications include smart homes, assisted living, health monitoring, robotics, human–computer interaction, intelligent driving, security and surveillance, gaming and entertainment etc. HAR mainly involves automatic detection, localization, recognition, and analysis of human actions from the data obtained from different types of sensors, including RGB camera, depth sensor, range sensor, etc. Action detection involves determining the presence of the action of interest in a continuous data stream, whereas action localization estimates when and where an action of interest appears. The goal of action recognition is to determine which action appears where in the data.

1.2 HAR Applications

Action recognition has been broadly studied by many researchers during the last decade with a significant evolution on the number of publications. This section highlights state-of-the-art applications that consider human action recognition methodologies to assist humans. Different applications of the current action recognition approaches are include:

1. Surveillance

Vision-based technologies are introduced in different security applications such as the surveillance system to recognize human behaviors such as fighting. Applications to identify vandalism events that may occur in a public places using one or several camera views. Multiple camera views used to detect and predict suspicious and aggressive behaviors in real time and in a crowded environment etc. are some of the applications researched on.

2. Assisted Living

Different modern technologies have provided a wide range of improvements in the performance of independent assisted living systems. Action recognition techniques can be used to monitor and assist the occupants. For example, human behavior monitoring and support (HBMS) i.e., a smart home system can analyze the activity patterns of an occupants to introduce automation based on the identified patterns to assist individuals in a smart way.

3. Healthcare Monitoring

The development of medical research and technology remarkably improved the quality of patients' life. However, higher demands of medical personnel made researchers try different technologies to improve healthcare monitoring methods that may be essential in emergency situations. Basically, one or more factors can be involved in the design of healthcare monitoring systems. This can include fall detection, human tracking, security alarm and cognitive assistance components.

4. Entertainment and Games

In the recent years, gaming industries have developed a new generation of games based on the full body of a gamer such as

dance and sports games, which has made gaming more interesting and exciting. RGB-D sensors are used in this kind of games to improve the perception of human actions.

5. Human–Robot Interaction

Human–robot interaction is considerably adapted in home and industry environments. An interaction is achieved to perform a specific task such as "Passing a cup" or "locating an object". A vision-based method is one of the effective communication ways between human and robots.

6. Video Retrieval

Most search engines use the associated information to manage video data. Text data such as tag, description, title and keywords is one piece of information that can be used for such purposes. However, one piece of information can be incorrect, which results in unsuccessful video retrieval. An alternative approach is video retrieval by analyzing human actions in videos.

7. Autonomous Driving Vehicles

An automated driving system is aimed to ensure safety, security, and comfort. One of the most important components of this system is action prediction and recognition algorithms. These methods can analyze human action and motion information in a short period of time that helps to avoid critical issues such as collision.

The interest in the development of these human activity-based applications can be justified by the fact that they provide very valuable and useful means of communication. However, the progress of the research in this field is also affected by the considerable changes in the technology trend and overall ecosystems.

1.3 Challenges in HAR

Action recognition task involves the identification of different actions from video clips where the action may or may not be performed throughout the entire duration of the video. Despite the stratospheric success of deep learning architectures, progress in architectures for video classification and representation learning has been slower. Some of the Popular Challenges in Action Recognition are as follows:

1. Huge Computational Cost

A simple convolution 2D net for classifying 101 classes has just ~5M parameters whereas the same architecture when inflated to a 3D structure results in ~33M parameters. It takes 3 to 4 days to train a 3DConvNet on UCF101 and about two months on Sports-1M, which makes extensive architecture search difficult and over fitting.

2. Capturing long context

Action recognition involves capturing spatiotemporal context across frames. Additionally, the spatial information captured has to be compensated for camera movement. Even having strong spatial object detection doesn't suffice as the motion information also carries finer details. There are local as well as global context w.r.t. motion information which needs to be captured for robust predictions.

3. Occlusion

An action required to be recognized should be clearly visible in the video sequences. This is not true in the real case, especially in a normal surveillance video. Occlusion can be presented by the person itself or by any other objects in the field. This can make body parts performing an action invisible which can cause a big issue for the research community.

4. Cluttered Background

Cluttered background is a case that formed a distraction introducing ambiguous information in the video of an action. Different methods are affected by this issue with unwanted background motion (due to cluttered background) along with the required motion. In addition, this issue has a great influence on color-based and region-based segmentation approaches as these methods require uniform background to achieve high quality segmentation.

5. Variation in Viewpoint

In a real-world situations (unlike the experimental setups), the location and posture of the person vary considerably based on the viewpoint where the action is captured from. A variation in motion patterns might also appear in each different view which makes action recognition more difficult.

6. Designing classification architectures

Designing architectures that can capture spatiotemporal information involve multiple options which are non-trivial and expensive to evaluate. For example, some possible strategies could be

- One network for capturing spatiotemporal information vs. two separate ones for each spatial and temporal
- Fusing predictions across multiple clips
- End-to-end training vs. feature extraction and classifying separately

7. No standard benchmark

The most popular and benchmark datasets have been UCF101 and Sports1M for a long time. Searching for reasonable architecture on Sports1M can be extremely expensive. For UCF101, although the number of frames is comparable to ImageNet, the high spatial correlation among the videos makes the actual diversity in the training much lesser. Also, given the similar theme (sports) across both the datasets, generalization of benchmarked architectures to other tasks remained a problem. This has been solved lately with the introduction of Kinetics dataset.

2. LITERATURE SURVEY

In 2017, 3D convolutional networks as feature extractors was introduced [1]. It uses 3D convolutions on video frames (where convolution is applied on a spatiotemporal cube). They trained the network on a large dataset of Sports 1M and then used the model as a feature extractor for other datasets. Their finding was a simple linear classifier like SVM on top of an ensemble of extracted features worked better than the state-of-the-art algorithms. The network focused on spatial appearance in the first few frames and tracked the motion in the subsequent frames. But the long-range temporal modelling was a problem, also training such huge networks is computationally a problem.

Heng Wang [2], in 2011 introduced an approach to model videos by combining dense sampling with feature tracking.

They introduce an efficient solution to remove camera motion by computing the motion boundaries descriptors along the dense trajectories. Local descriptors computed in a 3D video volume around interest points have become a popular way for video representation. To leverage the motion information in our dense trajectories, they compute descriptors within a space-time volume around the trajectory. Issue was that the trajectories tend to drift from their initial location during tracking.

Piotr Dollar, Vincent Rabaud, Garrison Cottrell, Serge Belongie [3] introduced a new spatiotemporal interest point detector and analyzes various cuboid descriptors. And concludes that cuboid prototyping (using K-means clustering) is a good behavior descriptor. Possible improvements could be using the spatiotemporal layout of the features, using features detected at multiple scales, and incorporating a dynamic model on top of their representation.

[4] Heng Wang and Cordelia Schmid, in 2013, improves dense trajectories by explicitly estimating camera motion. It demonstrates how the performance can be improved by removing background trajectories. It also uses state of the art human detectors to remove potentially inconsistent matches during camera motion estimation.

[5] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu in 2013, proposed to perform 3D convolutions to extract spatial and temporal features from the video. It discuss a 3D CNN model that uses the proposed 3D convolutions. The CNN architecture generates multiple channels of information from adjacent video frames and performs convolution and subsampling separately in each channel. And also proposes to regularize the 3D CNN models by augmenting the models with auxiliary outputs computed as high-level motion features. Issue was that it uses supervised training where labeling of data is a painful job. The number of labeled samples can be significantly reduced when such a model is pre-trained using unsupervised algorithms.

[6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei introduced Early fusion, Late fusion, and slow fusion connectivity for fusing time information in CNN models. They conclude that slow fusion consistently performs better. It also introduces a multi-resolution architecture for CNN to reduce the computation cost without affecting performance. It uses 2 separate streams of processing over 2 spatial resolutions. One of the streams is fed with down-sampled frames (context) and the other stream is fed with the center portion of the image(fovea). But it was computationally intensive and very little performance improvement is achieved. Scope for improvements were to consider broader categories in the dataset, investigate approaches that take camera motion into account, explore RNNs for learning global video-level predictions.

Quoc V. Le, Will Y. Zou, Serena Y. Yeung, Andrew Y. Ng [7] extended the Independent Subspace Analysis for learning features from Spatio-temporal data. It scales up the ISA algorithm to large receptive fields by convolution and stacking and learning hierarchical representations.

In 2014,[8] Karen Simonyan, Andrew Zisserman proposed a model that uses separate spatial and temporal recognition streams based on ConvNets. But it had issues like the spatial pooling in the network does not take the trajectories into account, and the camera motion wasn't handled properly and was compensated by mean displacement subtraction.

In 2015, Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici [9] explored the idea of incorporating information across longer video sequences. It introduced feature pooling method that processes each frame independently and uses max-pooling on local information to combine frame-level information. Also demonstrates the usage of an RNN that uses LSTM cells which are connected to the output of the underlying CNN. And validates the effectiveness of using Optical flow for motion information. Improvements such as an RCNN can be used to generate better features by utilizing its own activations in the last frame in conjunction with the image from the current frame.

[10] In 2016, Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell proposed the LRCN(Long term Recurrent Convolutional Networks) which combines convolutional layers with long-range temporal recursion. But it gives a single prediction for the entire video. If there are multiple actions in the clip, it takes the average of the probabilities of the softmax layer's output.

In 2011, Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, Atilla Baskurt [11] introduced a 2 step model to classify human actions. In the first step, a Conv 3D model is used to extract spatio-temporal features. In the second step, RNN with 1 hidden layer of LSTM cells is used to classify action sequences. Improvement is that a single-step model in which Conv3D and LSTM can be trained at once.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, Li Fei-Fei [12] in June 2017, introduced a MultiLSTM model that incorporates soft attention input-output temporal context for dense action labeling.

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Palur, in 2015 introduces a model similar to [5]. It uses 3D convolutional networks to learn spatio-temporal features. They demonstrate that 3x3x3 kernel works the best. But Long-range temporal modeling isn't addressed.

[14] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville in October 2015, proposed a 3D CNN RNN encoder-decoder model to capture local spatiotemporal information. They also propose the use of attention mechanisms for effective video description as it allows the usage features obtained using global analysis of static frames.

[15]In 2016, Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman proposed an architecture for two-stream networks with a convolutional fusion layer between the networks and a temporal fusion layer. Doesn't increase the number of parameters significantly.

[16] In August 2016, Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool introduced a model that improves the two-stream architecture. It introduces a network that extracts short snippets from the video by using sparse sampling (instead of dense sampling). The samples are distributed uniformly in the video. The snippets are fed to spatial stream ConvNets and Temporal stream ConvNets. The predictions from these ConvNets are combined to obtain a video level prediction. It also shows the usage of batch normalization, dropout, and pre-training as good practices.

[17] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell in 2017, proposed that samples frames

from the entire video and aggregates features from the appearance and motion streams into a single video level fixed-length vector. It is passed through a classifier that outputs the final classification scores. It explores multiple ways for combining the RGB and motion streams i.e. concat fusion, early fusion, and late fusion. The late fusion technique performs the best in their experiments.

[18] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann in 2018, introduced a pre-training layer(MotionNet) that generates optical flow from consecutive frames. The output from MotionNet is stacked with a temporal stream CNN to map the optical flows to target action labels. There's another spatial stream CNN which is combined with the temporal stream CNN using late fusion. Some improvements are Optical flow prediction can be improved based on smoothness loss, Using joint training instead of late fusion of spatial and temporal streams, Removing global camera motion and occlusion.

[19] starts where [13] left. In 2018 Joao Carreira, Andrew Zisserman proposed a 3D based models into two-stream architecture leveraging pre-training.

In 2017, [20] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, Luc Van Gool introduced an architecture to combine temporal information at variable depth. Prior methods used a fixed 3D homogeneous kernel depth. It also introduces a supervised transfer learning technique.

[21] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov proposed a soft attention-based model for action recognition. The model learns to focus selectively on the important parts of the video. Initially, the model takes a video frame as input and produces a feature cube. At each time step, the model predicts a softmax over K x K location(lt+1) and a softmax over the label classes(yt). Lt is the probability with which the model believes the corresponding region in the input frame is important.

In 2017, Rohit Girdhar and Deva Ramanan [22] proposed a modification to the networks by extending the existing architectures with attention maps that focus computation on specific parts of the input. The attention map doesn't require any additional supervision. It also provides a novel factorization of attention processing into bottom-up saliency combined with top-down attention. They also experiment with adding human pose as intermediate supervision to train the attention module. It looks for human-object interactions.

[23] Zhenxing Zheng, Gaoyun An, Dapeng Wu, Qiuqi Ruan in March 2020, proposed a novel global and local knowledge-aware attention network for action recognition. The proposed network incorporates two types of attention mechanisms called statistic-based attention (SA) and learning-based attention (LA) to attach higher importance to the crucial elements in each video frame. As global pooling (GP) models capture global information, while attention models focus on the significant details to make full use of their implicit complementary advantages, our network adopts a three-stream architecture, including two attention streams and a GP stream. Each attention stream employs a fusion layer to combine global and local information and produces composite features. Furthermore, global-attention (GA) regularization is proposed to guide two attention streams to better model dynamics of composite features with the reference to the global information. Fusion at the softmax layer is adopted to make better use of the implicit complementary advantages between SA, LA, and GP streams and get the final comprehensive predictions.

[24] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, Shilei Wen in 2017, introduced us to Attention Clusters where first, multiple feature sets are extracted from the video. For each feature set, we apply independent attention clusters with shifting operations to obtain a modality-specific representation vector. Next, the output of all attention clusters are concatenated to form a global representation vector of the video. Finally, the global representation vector is used for classification through a fully-connected layer. Their idea is to focus on local features instead of trying to capture global features. They argue that using long term temporal information isn't always needed for video classification. They use multiple attention mechanisms units (called attention clusters) to capture information from multiple modalities. They use shifting operations to increase the diversity between attention units. Some improvements are when applied to low-level local features and assess to what extent it can uncover relationships between features in different spatial coordinate, integrate it into end-to end-trained networks.

[25] Rohit Girdhar, Joao Carreira, Carl Doersch, Andrew Zisserman, May 2019. This paper introduces a transformer-based architecture for classifying actions. They use class agnostic queries by sharing features across different classes. The supporting argument is that the features for all the action classes need not be learned from scratch as the object is always a human. Their attention model learns to focus on hands and faces which is often crucial in differentiating between actions. They use an action transformer as input for the video feature representation and the box proposal from RPN and maps it into query and memory features. Issues are that the network doesn't seem to perform well for all action classes (for e.g. smoking) even though there is enough training data for some of those classes; If the size of the person in the clip is large, then the model is able to predict the classes correctly but for smaller sized objects, it performs poorly; It fails for modes such as similar action/interaction, identity, and temporal positions.

[39] In June 2019, Du Tran et. al. proposed Channel Separated convolution Networks (CSN) for the task of action recognition. They build on the ideas of group convolution and depth-wise convolution that received great success in Xception and MobileNet models.

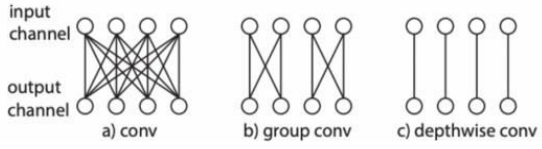

Fig-1: (a) A conventional convolution, which has only one group. (b) A group convolution with 2 groups. c) A depth-wise convolution where the number of groups matches the number of input/output filters.

Fundamentally, group convolutions introduce regularisation and less computations by not being fully connected. Depth-wise convolutions are the extreme case of group convolutions where the input and output channels equal the number of groups, as seen in Fig-1. Conventional convolutional networks model channel interactions and local interactions (both spatial and spatiotemporal) jointly in their 3D convolutions. This network effectively captures spatial and spatiotemporal

features in their own distinct layers. The channel separated convolution blocks learns these features distinctly but combines them locally at all stages of convolution. This alleviates the need to perform slow fusion of temporal and spatial two stream networks. The network also does not need to decide between learning spatial or temporal features as in C3D where the network can decide to learn features that are mixed between the two dimensions.

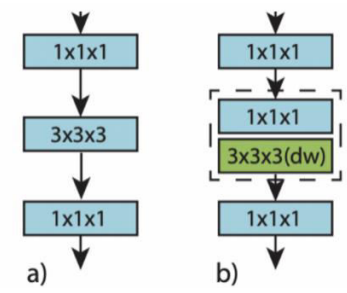

Fig-2: (a) A standard ResNet bottleneck block. (b) An interaction preserved bottleneck block.

The researchers propose to decompose 3x3x3 convolution kernels into two distinct layers, where the first layer is a 1x1x1 convolution for local channel interaction and the second layer is a 3x3x3 depth-wise convolution for local spatiotemporal interactions. By using these blocks, the researchers significantly decrease the number of parameters in the network and introduce a strong form of regularisation. The channel separated blocks allow for the network to locally learn spatial and spatiotemporal features in distinct layers.

This network effectively captures the bias that 2D spatial slices should form a natural image, whereas a 2D slice in the temporal direction has different temporal properties and does not fall in the natural manifold. In this way, the researchers enforce this bias by creating two separate distinct layers to process each direction. It is also capable of real time inference.

Method	input	video@1	video@5
C3D [30]	RGB	61.1	85.2
P3D [24]	RGB	66.4	87.4
Conv pool [40]	RGB+OF	71.7	90.4
R(2+1)D [31]	RGB	73.0	91.5
R(2+1)D [31]	RGB+OF	73.3	91.9
ir-CSN-101	RGB	74.8	92.6
ip-CSN-101	RGB	74.9	92.6
ir-CSN-152	RGB	**75.5**	**92.7**
ip-CSN-152	RGB	**75.5**	**92.8**

Table-1:Comparisons with state-of-the-art architectures on Sports-1M

As shown in Table-1, The CSN improves on state of the art RGB methods like R(2+1)D, C3D, and P3D on the Sports-1M dataset. The network is also 2–4x faster during inference. The model is also trained from scratch, where the rest of the models in the table are pre-trained on ImageNet or Kinetics dataset. This novel architecture improves on previous factorized networks while reducing over-fitting, being exceptionally fast, and producing state of the art accuracy on benchmark datasets.

3. BENCHMARK DATASETS

Although there is not a standard benchmark in activity recognition, there are some datasets that are being considered as references. As it has been mentioned before, due to the complexity of collecting data, the available datasets are limited. The following are the most used datasets:

3.1. UCF-101

UCF101 [32] is an action recognition dataset of realistic action videos. It is composed of 13,320 videos with 101 action categories and 27 h of video data. This dataset is an extension of the UCF50 dataset that has 50 action categories. The videos have been collected from YouTube, making the dataset realistic, and it provides a great variety of videos with different objects, camera motion, background, lighting, viewpoint, etc. Based on those features, videos are gathered into 25 groups (4–7 videos per action in each group) with videos sharing some of the features, as background, for example. The 101 categories can be divided in five main groups:

1. Human–Object Interaction: twenty categories.
2. Body-Motion Only: sixteen categories.
3. Human–Human Interaction: five categories.
4. Playing Musical Instruments: ten categories.
5. Sports: fifty categories.

3.2. HMDB51

HMDB51 [33] is another action recognition database that collects videos from various sources, mainly from movies but also from public databases such as YouTube, Google and Prelinger Archives. It consists of 6849 videos with 51 action categories and a minimum of 101 clips belong to each category. The action categories can be divided as well in five main groups:

1. General facial actions: smile, laugh, chew, talk.
2. Facial actions with object manipulation: smoke, eat, drink.
3. General body movements: cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave.
4. Body movements with object interaction: brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw.
5. Body movements for human interaction: fencing, hug, kick someone, kiss, punch, shake hands, sword fight. Apart from the action label, other meta-labels are indicated in each clip.

3.3. Weizmann

Before the two previous databases were created, many methods used the Weizmann [34] to evaluate the performance of their contributions. It provides 90 low-resolution (180 × 144, deinterlaced 50 fps) video sequences. These clips show 10 different actions performed by nine different people. These are the actions that appear in the database: run, walk, skip, jumping-jack (jack), jump-forward-on-two-kegs (jump), jump-in-place-on-two-legs (pjump), side-gallop (side), wave-two-hands (wave2), wave-one-hand (wave1) and bend. Background and the viewpoint are statics.

3.4. MSRAction3D

In 2010, as there was no public benchmark database, the authors published the database called MSRAction3D [35] which provided the sequences of depth maps captured by a depth camera. The dataset contains twenty actions: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap,

two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw. Seven different individuals performed each action three times, facing the camera during the performance. The depth maps have a size of 640 × 480 and they were captured at about 15 frames per second (fps) by a depth camera with infra-red light structure.

3.5. ActivityNet

The authors of [36] presented in 2015 the ActivityNet database. It is composed of 203 different classes with an average of 137 videos per class and a total of 648 video hours. The videos were obtained from online video sharing sites and they are around 5–10 min long. Half of the videos are in HD resolution (1280 × 720) and most of them have a frame rate of 30 fps. The aim of this database is to collect activities of human's daily life and it has a hierarchical structure, organizing the activities according to social interactions and where they take place.

3.6. Something Something

Later, in 2017, the authors of [37] introduced the "Something Something" dataset. The first version of the database consists of 108,499 videos belonging to 174 different labels with 23,137 distinct object names. The length of the videos varies between 2 and 6 s and they have a height of 100px and variable width. Labels are textual descriptions such as "Putting something next to something" where something refers to an object name. This database is already split into train, validation and test, containing 86,017, 11,522 and 10,960 videos, respectively. However, there has been a second release of the dataset and now it contains 220,847 videos, 168,913 for the training set, 24,777 for the validation set and 27,157 for the test set. The number of labels remains the same, but there are additional object annotations now. Moreover, the pixel resolution has increased from 100px to 240px.

3.7. Sports-1M

In [6], Karpathy et al. presented a new database, Sports-1M, which contains 1,133,158 video URLs with 487 automatically annotated different labels. YouTube Topics API was used to do the annotation. There are around 1000–3000 videos per class and some of them, nearly the 5%, are labelled with more than one class. Nowadays, the YouTube-8M dataset is also available and the Sports-1M dataset is included in it. This dataset is composed of videos from 3862 labels and it contains 350,000 h of video. In this case, each video has an average of three labels.

3.8. AVA

The authors of [38] presented AVA, a video dataset of spatio-temporally localized Atomic Visual Actions. This dataset consists of 430 movie clips of 15 min length annotated with 80 actions (14 poses, 17 person–person, 49 person–object). There are 386,000 labelled segments, 614,000 labelled bounding boxes and 81,000 person tracks, with a total of 1.58M labelled actions, with multiple labels per person occurring frequently. Every person of the scene is localized by a bounding box and labels are assigned according to the action performed by the actor. Each scene can have more than a label, one of them corresponds to the actor's pose and additional labels which correspond to person–object or person–person interactions can be assigned. A frame containing more than one actor is labelled separately for each person of the scene.

4. SUMMARY

Deep learning has revolutionized the way we process videos for action recognition. Deep learning literature has come a long way from using improved Dense Trajectories. Many learnings from the problem of image classification has been used in advancing deep networks for action recognition. Specifically, the usage of convolution layers, pooling layers, batch normalization, and residual connections have been borrowed from the 2D space and applied in 3D with substantial success. Many models that use a spatial stream are pre-trained on extensive image datasets. Optical flow has also had an important role in representing temporal features in early deep video architectures like the two stream networks and fusion networks. Optical flow is our mathematical definition of how we believe movement in subsequent frames can be described as densely calculated flow vectors for all pixels. Originally, networks bolstered performance by using optical flow. However, this made networks unable to be end-to-end trained and limited real-time capabilities. In modern deep learning, we have moved beyond optical flow, and we instead architect networks that are able to natively learn temporal embedding and are end-to-end trainable.

Action recognition is a truly unique problem with its own set of complications. The first source of friction is the high computation and memory cost associated with 3D convolutions. Some models take over 2 months to train on Sports-1M on modern GPU's. The second source of friction is that there is no standard benchmark for video architecture search [1]. Sports-1M and UCF101 are highly correlated and false-label assignment is common when a portion of a video is selected to be trained on but actually may not contain the actual action as it may be in another part of the video. The last source of friction is that designing a video deep neural network is nontrivial. The choice of layers, how to preprocess the input, and how to model the temporal dimension is an open problem. The authors of the papers above attempt to tackle these issues in an empirical fashion and propose novel architectures that resolve temporal modelling in videos.

5. CONCLUSIONS

There has been plenty of research in the area of human action recognition (HAR) and video analysis. It has come a long way in past few years after the advent of neural networks. Initially, CNNs applied frame by frame helped in improving the accuracies as compared to the manual feature extraction techniques. Later 3D-CNNs further improved the accuracies of CNNs by processing multiple frames at a time. Recent architectures focused on RNNs and LSTMs to factor in the temporal component of the videos. Also, architectures started incorporating attention mechanisms to focus on the salient parts of the videos, in the recent years. Most recently,

Channel-Separated convolution Networks was built on group convolution and depth-wise convolution models.

5.1. State-of-the-art method

The current state-of-the-art for action recognition, in my opinion, is the channel separated network (CSN). This network effectively captures spatial and spatiotemporal features in their own distinct layers. The channel separated convolution blocks learns these features distinctly but combines them locally at all stages of convolution. This alleviates the need to perform slow fusion of temporal and spatial two stream networks. The network also does not need to decide between learning spatial or temporal features as in C3D where the network can decide to learn features that are mixed between the two dimensions. This network effectively captures the bias that 2D spatial slices should form a natural image, whereas a 2D slice in the temporal direction has different temporal properties and does not fall in the natural manifold. In this way, the researchers enforce this bias by creating two separate distinct layers to process each direction. It is also capable of real time inference. Channel separation is an important step forward in action recognition and beats the state-of-the-art results even when trained from scratch.

5.2. Future Enhancements

Human action recognition is still a very active research area and new approaches are still trying to solve the issues with the current approaches. Some of the existing issues are background clutter or fast irregular motion in videos, occlusion, viewpoint changes, high computational complexity, and responsiveness to illumination changes.

For future research, it is also recommended to look into including more biases we have of the real world in deep video network architecture. Another good area to explore is how depth modelling can relate to better video classifications. Also, it is observed that any spatial changes in a video come from either a transformation of an external object we are observing, or change in the observer's view point i.e., angle or position. Both of these sources of movement have to be learned by the current networks. It would be interesting to study how depth fields could be used to model either sources of change.

REFERENCES

1. Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, Manohar Paluri. ConvNet Architecture Search for Spatiotemporal Feature Learning. In arXiv:1708.05038v1, 16 Aug 2017.
2. Heng Wang, Alexander Kläser, Cordelia Schmid, Liu Cheng-Lin. Action Recognition by Dense Trajectories. In HAL, IEEE CVPR 2011.
3. Piotr Dollar, Vincent Rabaud, Garrison Cottrell, Serge Belongie. Behavior Recognition via Sparse Spatio-Temporal Features.
4. Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In ICCV, 2013.
5. Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. In IEEE, March 2013.
6. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In IEEE, June 2014.
7. Quoc V. Le, Will Y. Zou, Serena Y. Yeung, Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis.
8. Karen, Simonyan, Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In arXiv:1406.2199v2 [cs.CV] 12 Nov 2014.
9. Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici. Beyond Short Snippets: Deep Networks for Video Classification. In arXiv:1503.08909v2 [cs.CV] 13 Apr 2015.
10. Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In arXiv:1411.4389v4 [cs.CV] 31 May 2016.
11. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, Atilla Baskurt. Sequential Deep Learning for Human Action Recognition. In Springer-Verlag Berlin Heidelberg 2011.
12. Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, Li Fei-Fei. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. In arXiv:1507.05738v3 [cs.CV] 9 Jun 2017.
13. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Palur. Learning Spatiotemporal Features with 3D Convolutional Networks. In arXiv:1412.0767v4 [cs.CV] 7 Oct 2015.
14. Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville. Describing Videos by Exploiting Temporal Structure. In arXiv:1502.08029v5 [stat.ML] 1 Oct 2015.
15. Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In arXiv:1604.06573v2 [cs.CV] 26 Sep 2016.
16. Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In arXiv:1608.00859v1 [cs.CV] 2 Aug 2016.
17. Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell. ActionVLAD: Learning spatio-temporal aggregation for action classification. In arXiv:1704.02895v1 [cs.CV] 10 Apr 2017.
18. Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden Two-Stream Convolutional Networks for Action Recognition. In arXiv:1704.00389v4 [cs.CV] 30 Oct 2018.
19. Joao Carreira, Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics

Dataset. In arXiv:1705.07750v3 [cs.CV] 12 Feb 2018.
20. Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, Luc Van Gool. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. In arXiv:1711.08200v1 [cs.CV] 22 Nov 2017.
21. Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action Recognition using Visual Attention. In arXiv:1511.04119v3 [cs.LG] 14 Feb 2016.
22. Rohit Girdhar, Deva Ramanan. Attentional Pooling for Action Recognition. In arXiv:1711.01467v3 [cs.CV] 30 Dec 2017.
23. Zhenxing Zheng, Gaoyun An, Dapeng Wu, Qiuqi Ruan. Global and Local Knowledge-Aware Attention Network for Action Recognition. In IEEE, March 2020.
24. Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, Shilei Wen. Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification. In arXiv:1711.09550v1 [cs.CV] 27 Nov 2017.
25. Rohit Girdhar, Joao Carreira, Carl Doersch, Andrew Zisserman. Video Action Transformer Network. In arXiv:1812.02707v2 [cs.CV] 17 May 2019.
26. Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, Cordelia Schmid. PoTion: Pose MoTion Representation for Action Recognition. In HAL, 11 Apr 2018.
27. Lucas Smaira, Joao Carreira, Eric Noland, Ellen Clancy, Amy Wu, Andrew Zisserman. A Short Note on the Kinetics-700-2020 Human Action Dataset. In arXiv:2010.10864v1 [cs.CV] 21 Oct 2020.
28. IVAN LAPTEV. On Space-Time Interest Points. In Springer International Journal of Computer Vision 64(2/3), 107–123, 2005.
29. Gul Varol, Ivan Laptev, and Cordelia Schmid, Fellow, IEEE. Long-term Temporal Convolutions for Action Recognition. In arXiv:1604.04494v2 [cs.CV] 2 Jun 2017.
30. Lin Sun, Kui Jia, Dit-Yan Yeung, Bertram E. Shi. Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks. In arXiv:1510.00562v1 [cs.CV] 2 Oct 2015.
31. Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger. Densely Connected Convolutional Networks. In arXiv:1608.06993v5 [cs.CV] 28 Jan 2018.
32. Soomro, K Zamir, A.R Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. In arXiv:1212.0402, 2012.
33. Kuehne, Jhuang, Garrote, Poggio, Serre. HMDB: A large video database for human motion recognition. In ICCV, November 2011.
34. Blank, Gorelick, Shechtman, Irani, Basri. Actions as space-time shapes. In IEEE ICCV'05, 21 October 2005.
35. Li, Zhang, Liu, Z. Action recognition based on a bag of 3D points. In IEEE, June 2010.
36. Caba Heilbron, Escorcia, Ghanem, Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In IEEE, June 2015.
37. Goyal, Kahou, Michalski, Materzynska, Westphal, Kim, Haenel, Fruend, Yianilos, MuellerFreitag. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In IEEE ICCV, October 2017.
38. Gu, Sun, Ross, D.A. Vondrick, Pantofaru, Li, Y. Vijayanarasimhan, Toderici, Ricco, Sukthankar. AVA: A video dataset of spatio-temporally localized atomic visual actions. In IEEE, June 2018.
39. Tran, Du, et al. "Video Classification with Channel-Separated Convolutional Networks." arXiv preprint arXiv:1904.02811, November 2019.