An Extensive Analysis of the Vision-based Deep Learning Techniques for Action Recognition

Manasa R[1], Ritika Shukla[2], Saranya KC[3]
School of Electronics Engineering
Vellore Institute of Technology
Vellore, India

***Abstract*—Action recognition involves the idea of localizing and classifying actions in a video over a sequence of frames. It can be thought of as an image classification task extended temporally. The information obtained over the multitude of frames is aggregated to comprehend the action classification output. Applications of action recognition systems range from assistance for healthcare systems to human-machine interaction. Action recognition has proven to be a challenging task as it poses many impediments including high computation cost, capturing extended context, designing complex architectures, and lack of benchmark datasets. Increasing the efficiency of algorithms in human action recognition can significantly improve the probability of implementing it in real-world scenarios. This paper has summarized the evolution of various action localization, classification, and detection algorithms applied to data from vision-based sensors. We have also reviewed the datasets that have been used for the action classification, localization, and detection process. We have further explored the areas of action classification, temporal and spatiotemporal action detection, which use convolution neural networks, recurrent neural networks, or a combination of both.**

Keywords*—Action recognition; deep learning; vision sensors; convolution neural networks (CNN); recurrent neural networks (RNN); action classification; temporal action detection; spatiotemporal action detection*

I. Introduction

There are two types of human action recognition systems - sensor-based and video-based [1]. Various on-body and ambient sensors are used to understand and label human actions performed in recorded videos or real-time video streaming. Video cameras are the essential wellsprings of new data on the Internet. A video is an organized arrangement of frames of a similar resolution taken at regular intervals of time. While developing the video processing algorithm, the video is partitioned into two classes-video streams and video sequences. Video stream is a continuous video for online processing as we are unaware of the information present in future frames. The video sequence is a fixed-length video where all frames are accessible without a moment's delay. Currently, most video cameras do not perform automated action recognition. Since the amount of video data available is extremely high, automatic action recognition has become a necessity. Furthermore, action recognition will facilitate efficient human-machine interactions, video surveillance, patient-care, smart homes, sports video analysis, gaming, and intelligent retail.

An action recognition process involves two tasks: action classification and action localization, as represented in Fig. 1. Action classification consists of assigning labels to various action instances in videos. Although it is possible to classify some actions using single frames, most actions occur in a series of adjacent frames. The motion in these frames must be captured to classify the actions. Video data brings a new feature that is absent in static images, which are motion. This motion characterizes actions in videos. To obtain these motion features, the motion field must be obtained. Optical flow, which represents the apparent motion between frames, is used to estimate the motion field.

The extensive input data, less availability of computational resources, and difficulty in obtaining the optical flow pose major problems while classifying actions. In action classification tasks, the model must run through multiple windows in search of action instances. This is computationally expensive and time-consuming. Temporal action detection models work on the data before action classification models to reduce computational costs. They define the temporal bounds of action instances and specify to the action classification model the actions' temporal location in any given video sequence. Spatiotemporal action detection models provide information on the spatial locations of the action in addition to the temporal bounds.

The field of computer vision and deep learning has already seen significant success in object detection, classification, and localization techniques, and now the area of study is moving towards efficient action detection and recognition tasks. Sliding window approaches were the earliest action localization approaches that scanned the videos exhaustively to get the video's actions' spatial and temporal coordinates. Some previous action recognition approaches like Silhouette and poses estimation were inspired by object detection frameworks [2]. These frameworks were directly extended to the spatiotemporal scale to localize action. Before Deep Learning approaches came into the picture, handcrafted techniques like Histogram of Oriented Gradient (HOG) [3], Histogram of Optical Flow (HOF) [4], Extended SURF(ESURF) were prevalent [5]. Although these approaches were robust to background noise, change in illumination, and video clutter, they lacked semantics and discriminative capacity.

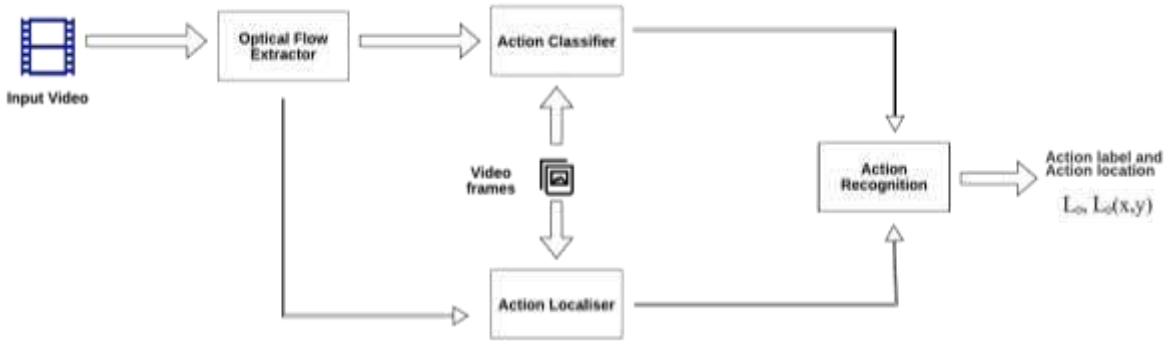

Fig. 1. Action Recognition - Steps Involved.

The purpose of this paper is to analyze the various deep learning architectures for action recognition techniques. It focuses on visual sensor-based methods. The paper has elaborated on action classification, temporal/spatiotemporal action detection, and localization techniques. Section 2 describes the various datasets available for action classification, recognition, detection, and localization. Section 3 explores the different proposed methodologies for action classification tasks. Section 4 delves into the existing approaches for temporal action detection, and Section 5 discusses the methods proposed for spatiotemporal action detection, respectively. Section 6 concludes the whole paper.

II. DATASETS

An estimation says that there are over 1000 human action categories. A variety of studies have been conducted to create datasets that can help us overcome the challenges posed by human action recognition. Action recognition and localization is a widely studied problem. The key challenges associated with this field have been variations in human posture, scaling, pixilation, speed, background clutter, and occlusion. Low-grade and insufficient datasets lead to challenges such as prediction of wrong action class, incorrect spatial or temporal action localization, and inability to detect more than one action in a frame. Table I lists some of the most used datasets for performing action localization and recognition tasks and compares them based on several action classes, data size, nature of video clips, and their aim.

Earlier datasets contained very few action classes. UCF sports has ten action classes: Golf Swing, Lifting, Running, SkateBoarding, Kicking, Diving, Swing-Bench, Swing-Side, Riding Horse, and Walking [6]. UCF sports is introduced, which mainly comprises the video sequences featured on television channels BBC and ESPN.

Various datasets are not realistic, and the action classes are also significantly less. K. Soomro, A. Zamir, and M. Shah [7] targeted these issues and proposed a new dataset, UCF101. It consists of 101 action classes, 13000 vid clips, 27 hours of video clips. Also, the video clips in this dataset are more realistic as they are not recorded in controlled environments, which is essential for training a model which performs well in the real world. However, there is not much variation in the video clips for a particular action class in UCF101.

TABLE I. DATASETS USED FOR ACTION RECOGNITION

Datasets	Number of action classes	Data size	Trimmed/Untrimmed	Year of release	Main Sources
UCF sports	51 action classes	6849 video clips	Trimmed	2008	BBC Motion Gallery and GettyImages
HMDB51	51 action classes	6849 video clips	Trimmed	2011	The Prelinger Archive, YouTube, and Google videos.
UCF101	101 action classes	13320 video clips	Trimmed	2012	YouTube
JHMDB	21 action classes	928 video clips	Trimmed	2013	The Prelinger Archive, YouTube, and Google videos
Thumos15/14	101 action classes	18,420(thumos15), 15,906(thumos14)	Untrimmed	2015(v15), 2014(v14)	YouTube
ActivityNet	200 action classes	9682 video clips(v1.2), 19,994 video clips(v1.3)	Untrimmed	2016(v1.3), 2015(v1.2)	
Kinetics 400	400 action classes	300k video clips	Trimmed (10s)	2017	YouTube
Kinetics 600	600 action classes	500k video clips	Trimmed(10s)	2018	YouTube
Kinetics 700	700 action classes	650k video clips	Trimmed(10s)	2019	YouTube

Some of the datasets focused on increasing the robustness of various action recognition models by exploring under numerous conditions like the movement of the camera, angle, and position of viewpoint, quality of the video, and occlusion. Human Motion Database (HMDB51) [8], dataset focuses on features mentioned above. At least two observers validate the clips of the datasets to establish consistency. The dataset also contains metadata like the number of actors involved, viewpoint, presence or absence of motion of the camera, and category labels.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black [9] proposed JHMDB, "joint-annotated HMDB." This dataset annotates human joints in the HMDB dataset. However, it contains lesser action categories as compared to HMDB51. Their main objective is to understand what features improve the efficiency of action recognition algorithms primarily. They find that high-level pose features are more efficient for capturing actions in videos than low/mid-level features. J-HMDB is beneficial for linking low-to-mid level features with high-level poses. As higher-level pose features need the information of joints. This provides richer information and enables more complex models.

Thumos14 [10] is a dataset used to detect and recognize actions in realistic untrimmed videos with a standard protocol for evaluation. This dataset's action classes are from UCF101, which are mainly divided into five categories - Body-Motion Only, Human-Object Interaction, Human-Human Interaction, Sports, and Playing Musical Instruments. After this, Thumos15 introduces background videos that do not contain the target action with multiple actions in the same video. This further increases the complexity of the dataset.

F. Caba, V. Escorcia, B. Ghanem, and J. Carlos [11] introduced ActivityNet, which has more action categories. Most significantly, ActivityNet has an organized set of activities according to social interactions and where they usually occur. Some of the classes of action in the dataset include - Household, Caring and helping, Personal care, Work-related, Eating and drinking, Socializing and leisure, Sports, and exercises. ActivityNet has the following applications - untrimmed video classification, trimmed activity classification, and activity detection. ActivityNet benchmark has rich semantic taxonomy and aims at covering daily activities performed by humans on an average. Results show that ActivityNet opens new challenges in understanding and recognizing human actions.

Just like various action recognition algorithms are inspired by multiple object detection algorithms. Similarly, some of the datasets are inspired by image datasets. ImageNet inspires kinetics dataset for action classification purposes. The kinetics project aimed to get the same number of action classes as image classes in ImageNet [12]. There are four versions of the kinetics dataset: kinetics 400, kinetics 600, and kinetics 700. Kinetics 400 contains 10 seconds trimmed video clips and a variation in resolution and frame rate having at least 400 clips of each action. Some of the parent action classes in kinetics 400 are arts and crafts, auto maintenance, ball sports, cleaning, dancing, electronics. This dataset can also be used for multi-modal analysis. Kinetics dataset is better than HMDB and UCF datasets due to more action classes and a wide range of actions.

The AVA-kinetics dataset [13] contains 624,430 unique frames and 238,906 unique videos. Some of the selected action classes include swimming, swimming backstroke, swimming breaststroke, swimming butterfly stroke, pushing a wheelchair, giving or receiving awards, punching bag.

III. Deep Learning for Action Recognition

Andrej Karpathy et al. [14] introduced Single Stream Deep Neural networks for action recognition. They proposed and tested four different single stream architectures: Single Frame, Late Fusion, Early Fusion, and Slow Fusion. Single Stream Networks can be induced with information from other models trained on larger datasets to obtain better results. Another significant advantage is that these models do not require the calculator of optical flow as the input includes only RGB images. Therefore, these models can be used for real-time purposes. However, these models were not able to effectively capture the motion features.

To overcome this shortcoming, K. Simonyan and A. Zisserman [15] brought forward the concept of Two-Stream Networks. The Two-Stream Network has two different architectures to individually process the temporal and spatial features. One network takes the single video frames as input, and the other will take the optical flow as input. The output of the two networks is then fused to obtain the class scores. Although this model produces state-of-the-art results in terms of accuracy, it has many drawbacks. As both the networks have to be trained separately, it is not end-to-end trainable. It cannot work with small datasets as transfer learning cannot be applied here. Even though the spatial network can derive features from large image datasets, the temporal model needs to be trained on a video dataset. It is also computationally expensive as the optical flow needs to be calculated before being fed into the temporal network.

Later works made use of LSTMs and 3D convolution networks for action recognition. These networks were not only end-to-end trainable but also worked in real-time. The LSTM architecture was first introduced by Jeffrey Donahue et al. [16]. The authors have taken inspiration from the encoder-decoder architecture and extended it for action recognition. The LSTM based network did not get results as good as the two-stream networks but surpassed the single-stream networks. D. Tran, L. Bourdev, R. Fergus, L. Torresani M. Paluri [17] introduced the concept of 3D convolution networks. This model surpassed the two-stream networks in terms of performance.

The coming sections describe the works that use deep learning techniques for action classification, temporal action detection spatiotemporal action detection.

IV. Action Classification

Action classification is the identification of the type of action in a trimmed or untrimmed video. There has been ongoing research on producing efficient methods of classifying actions in a video clip. L. Wang, Y. Qiao, and X. Tang [18] have put forward a novel video representation

known as Trajectory Pooled Deep Convolutional Descriptor (TDD), which considers the advantages of both deep-learned features as well as handcrafted features. Deep architectures are used to learn discriminative Conv feature maps. Trajectory constrained pooling is conducted to concentrate these convolutional features into effectual descriptors. The accuracy of TDDs is enhanced by using two normalization methods, namely channel normalization, and spatiotemporal normalization, to transform convolutional feature maps. This approach has several advantages. The learning process in TDDs is automatic, and the discriminative capacity is higher when compared to handcrafted features. The plans of action of trajectory-constrained pooling and sampling are introduced by considering the temporal dimension's intrinsic characteristics for aggregating the deep-learned features. The shortcoming of this method is that it is computationally expensive.

Many researchers have made efforts to make the process of action classification less computationally expensive. Although two-stream CNNs are quite efficient and are state-of-the-art when it comes to action recognition, they are computationally costly. One of the main reasons for this is the requirement to calculate the optical flow, which has very high computational needs. The two-stream networks consist of two CNN networks. One is the spatial network that takes as input RGB images, and the other is the temporal network that takes the optical flow as input. This process is not only high on computation but is also time taking. To address this problem, B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang [19] introduced Real-time action recognition with enhanced motion vector CNNs. They have replaced optical flows with motion vectors. Like optical flow, motion vectors describe the motion in a video, but unlike optical flow, they are easily obtained directly in the video decoding process. Hence, they can be used alongside deep convolutional frameworks for action recognition tasks. The authors have proposed a mechanism where the RGB images and motion vectors are obtained from the video decoding process and fed into two-stream CNN. Optical flows are very dense and hence are entirely accurate with fewer noise features. Motion vectors are not very precise and consist of a lot of inaccurate movements and noise. To increase the motion vector CNN's performance, the knowledge learned from an optical flow CNN is transferred into a motion vector CNN. Although optical flow needs to be calculated for this procedure, it is still efficient as this calculation is done only while training and not while testing.

H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould [20] introduced dynamic images for action recognition to further reduce the computational costs. Dynamic images are a novel compact representation of the video, which is based on the rank pooling idea and are acquired through the parameters of a ranking system that encrypts the temporal evolution of the video frames. Since it is an image, CNN models can directly be applied to the video data with fine-tuning allowing end-to-end training for action recognition. This approach is efficient and is not time-consuming as the whole video is summarized to an amount of data equivalent to a single frame.

To further reduce the computation costs while maintaining accuracy, Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann [21] have proposed a Hidden two-stream CNN for action recognition. A two-stream network takes as input RGB images as well as optical flows. The hidden two-stream network is designed to input only the video frames and not the optical flow. This allows a 10x faster performance when compared to the traditional two-stream architecture. This approach uses unsupervised methods to predict the optical flow. The flow field between two adjacent frames is generated using CNN. This indicated flow field and a frame are used to reconstruct the previous frame using backward warping. The idea is that if one frame can rebuild the last frame, then the network has learned the representations of some underlying motions of a video.

While some research works focus on reducing the computational costs of a system, others have attempted to increase the networks' accuracy. W. Byeon, Q. Wang, R. Kumar, and P. Koumoutsakos [22] have proposed a fully context-aware system that produces sharp predictions of high visual quality. The previous prediction models based on CNNs, RNNs, or a combination of both, tend to produce blurry results. Some efforts have attempted to address this issue by separating the foreground from the background, adversarial training, or motion flow learning, but have mainly failed to consider the issue that the model is unaware of the complete information. To solve this shortcoming, the authors have proposed a fully context-aware architecture that captures past information using parallel multidimensional LSTM units.

R. Girdhar and D. Ramanan [23] have also tried to improve action recognition accuracy while ensuring that the network size and computational cost will remain unchanged. They have proposed an Attentional Pooling module that can be used as a replacement for the normal pooling operation in any convolutional network. This model is built over a base Resnet architecture. The proposed Attention layer is plugged into the last layer after generating spatial feature maps, which need to be average pooled.

Another major factor affecting the accurate classification of actions on how much information we can gather from the temporal cues available in the video. Ali Diba et al. [24] have introduced new architecture and transfer learning for video classification. The computer vision community has mainly focused on spatiotemporal approaches where the temporal convolutional kernel depths are fixed. This paper has introduced a new temporal layer that models various kernel depths of temporal convolutions, which are embedded into a proposed 3D CNN. The 3D CNN is extended from the 2D DenseNet by including 3D filters and pooling kernels. Most of the researchers working on 3D convnets tend to train them from scratch. This can prove inefficient as they fail to consider the knowledge gained by the 2D convnets. To overcome this issue, this paper has done an effective transfer of knowledge from 2D convnets to 3D convnets. This not only diminishes the computational cost but also makes the system more accurate.

V. Temporal Action Detection

Temporal action detection is another significant yet testing problem that goes one step beyond action classification. Since recordings in real-world applications are generally long, untrimmed, and contain numerous action instances, this issue

requires perceiving action classifications and recognizing each activity occasion's start time and end time. Temporal action detection can help define the temporal bounds of an action sequence and reduce the computation of action classification tasks. Researchers have tried to solve this problem in various ways. G. Yu and J. Yuan [25] proposed a Fast action proposal for human action search and detection. The action proposal is quite challenging as both the appearance and motion cues have to be considered. This paper is targeted at producing action proposals in unconstrained videos. An action proposal is represented by a temporal series of spatial bounding boxes (spatiotemporal video tube) which can locate a single human action. They have established the action proposal generation as a max set coverage problem, and greedy search is employed to maximize the actionness score. Actionness is a measure that quantifies the likelihood of the presence of an action instance at specified locations. This method can be used before the process of action classification to ensure limited computational costs. The action classification system can now focus only on the action proposals rather than on the whole video. This algorithm works well with moving cameras and can detect actions even in cluttered backgrounds.

Numerous researchers make consistent efforts to facilitate accurate and efficient estimation of actionness. L. Wang, Y. Qiao, X. Tang, and L. Van Goo [26] proposed a hybrid fully convolutional network for actionness estimation. They have introduced a novel convolutional network consisting of an appearance FCN(A-FCN), which takes as input RGB images, and a motion FCN(M-FCN) which takes optical flow fields input. These two networks derive information from static appearance and dynamic motion, respectively. The completely convolutional nature of H-FCN permits it to productively handle recordings with subjective sizes. Each FCN is a discriminative system prepared in a start to finish and pixel-to-pixel way. These estimated actionness maps are then fed into detection frameworks for the action detection process.

Previous temporal action localization strategies depend on applying action classifiers at each time area and different transient scales in a temporally designed sliding window. While most approaches for activity detection find it quite hard to produce high accuracy on large-scale video collections due to their high computational complexity, F. Caba, J. Carlos, and B. Ghanem [27] devised a method to extract temporal segments from untrimmed videos with high recall and good precision at a fast rate. A sparse learning frame is generated for scoring transient frameworks as indicated by the fact that they are prone to contain an action. This proposal is then merged into an activity detection framework to enhance the overall performance.

Many researchers understood the importance of performing temporal action localization in untrimmed videos as recordings in genuine applications are typically unconstrained and contain numerous activity cases in addition to background clutter. To address this issue, Z. Shou, D. Wang, and S. Chang [28] proposed an action localization framework using three-segment-based 3D ConvNets. The framework contains three networks, namely, localization network, classification network, and network. The proposal network is used for identifying action sequences in an untrimmed video. The classification network serves as an initiation for the localization network, which fine-tunes the classification network to localize action temporally.

Single-Stream Temporal Action proposals are another method for obtaining temporal action proposals in long, untrimmed videos [29]. While most methods require the video to be divided into short overlapping clips for temporal action localization, SSTs can process a long video in a single stream. Hence, they are much faster than previous models where temporal action proposals are identified from temporal windows and then independently classified. Applying windows at multiple scales is computationally expensive. Hence, SSTs are less exhaustive and generate action proposals in long videos with just a single video pass through the network.

Single-Stream Temporal Action Detection [30] is another example of a network that incorporates Single-Stream Networks. It draws inspiration from object detection algorithms like YOLO and Faster RCNN. It provided an end-to-end approach of action detection in untrimmed videos, claiming that everything happens in a single pass network. Hence, it is very efficient which can operate at 701 frames/sec. The network was trained for thumos14. This model also outperforms other models in detection performance and fps, just like YOLO.

While most works usually involve building frame-level classifiers and passing the video through them multiple times, S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei [31] have designed a methodology for end-to-end learning of action detection that learns to predict the temporal bounds of actions. An RNN based agent decides which frame to analyze next and when to send forth a prediction. This paper puts forth a single network that takes an untrimmed video for input and gives as output the temporal bounds of any detected actions.

F. Heidarivincheh, M. Mirmehdi, and D. Damen [32] proposed an approach wherein, despite localizing the action, they focused on localizing the moment of completion, where it localizes the completed action along with localizing the action. Hidden Markov Model (HMM) and Long-Short Term Memory (LSTM) are used to assess six kinds of actions - switch, plug, open, pull, pick and drink. The model uses supervised learning. Therefore, the annotations of pre-completion and post-completion frames are already available. They also concluded that fine-tuned CNN features give better results than handcrafted features. An action may often be localized in the video, even if it was an incomplete action. In this approach, by targeting the completion of the action, they successfully overcome this problem.

Some other works also focused on detecting complete actions. An end-to-end trainable network proposed by Yue Zhao et al. [33] Structured Segment Networks focused on untrimmed videos does this by implementing both action classifiers and detecting the complete action. This increases the overall accuracy of the model. Their model also includes detecting high-quality proposal generation termed - Dubbed Actionness Grouping (DAG). The limitation which comes to this model is the existence of a large number of unfinished action snippets in temporal boundaries. To overcome this

issue, the model must understand the various stages of an action. They have introduced structured temporal pyramid pooling that produces a global portrayal of the whole proposal and a broken-down discriminative model to order action classifications together, finding whether a particular action is complete. The model is also computationally efficient because they have used a sparse snippet sampling strategy.

VI. Spatiotemporal Action Detection

Spatiotemporal detection is the process of detecting coordinates of action on a spatial as well as temporal scale. Various algorithms devised for 2D images were directly extended to check their accuracy for 3D actions. One such method is Spatiotemporal Deformable Part Models (SDPM) for Action Detection [34]. This approach explores the generalization of deformable part models from 2D images to 3D spatiotemporal volumes to study their effectiveness for action detection in video. In this paper, a deformable part model is generated for each action (spatiotemporal patterns) and from a collection of examples. The proposed spatiotemporal deformable part model (SDPM) stays true to the structure of the original DPM. This model employs volumetric parts that displace in both time and space, which allows it to perform better for intra-class variation in terms of execution and better performance in clutter.

Another approach that extends a two-dimensional object proposal technique is adopted in spatiotemporal object detection methods [35]. This paper presents spatial, temporal, and spatiotemporal pairwise super voxel features to manage the blending process. Also, they propose another effective super voxel method. Experimental evaluation of the complete model shows that this super voxel approach leads to more precise recommendations than utilizing existing cutting-edge super voxel methods. They have built on the approach of S. Manen, M. Guillaumin, and L. Van Gool [36] that uses a randomized superpixel consolidating methodology to get object proposals.

K. Soomro, H. Igrees, N, and M. Shah [37] proposed early predication and localization of action by taking input at relatively more minor video lengths. Action prediction and online localization accuracies improve over time as the number of frames available increases.

Action localization with tubelets from motion [38] considered super voxels instead of super-pixels to produce spatiotemporal shapes, which directly gives us 2D+t sequences of bounding boxes as tubelets in this paper. Their contributions include investigating the selective search sampling strategy for videos and incorporating motion information in various analysis stages. The singularity of the motion is encoded in a feature vector associated with each super-voxel.

G. Gkioxari and J. Malik [39], inspired by the field of object detection in images, propose an approach where motion and appearance are incorporated in two different ways. In this paper, they select the frames with a higher probability of containing a motion or are more useful for detecting the motion in the video. They select candidate regions and employ CNNs to classify them. The idea of eliminating the regions with lower motion saliency significantly decreases the computation time. The two networks - spatial-CNN and motion CNN operate on static cues and motion cues, respectively.

Other approaches adopted for spatiotemporal action localization include techniques employing dense trajectories. APT: Action localization Proposals from dense Trajectories [40] proposes an efficient generation algorithm to handle many trajectories in a video. The dense trajectories are computed for the video's representation; this paper focuses on re-using them for proposal generation. Therefore, this paper introduces the use of dense trajectories for classification as well.

M. Zolfaghari, G. Oliveira, N. Sedaghat, and T. Brox [41] exploits pose, motion, and appearance for action recognition. To integrate them Markov chain model is utilized, which adds cues successively. This helps in the sequential refinement of action labels.

Action Detection by Implicit Intentional Motion Clustering [42] is based on using spatiotemporal trajectory clustering by leveraging intentional movement properties. The calculated movement clusters are then utilized as action proposals for detection. They find that trajectories from deliberate motion are appreciably densely localized in space and time.

Another group of approaches is based on using two-stream networks for spatiotemporal action detection or localization. Various two-stream networks have been tested successfully for action detection and localization. Two-stream networks consist of a spatial network that models appearance, whose input is RGB frames, and a temporal network that models motion. Optical flow or dense trajectories can be used as input for these networks. Real-Time End-to-End Action Detection with Two-Stream Networks [43] proposes a model that integrates the optical flow computation using Flownet2 and then, applying early fusion for the two streams and training the whole pipeline jointly end-to-end. Experimental results prove that training the pipeline together end-to-end with fine-tuning the optical flow for the objective of action detection improves detection performance appreciably. This model is inspired by YOLOv2.

VII. Conclusion

This paper has presented an expanded overview of various works done in action classification, temporal action detection, and spatiotemporal action detection. Although various on-body sensors are used to understand and label human action recognitions, this paper focuses on visual sensor inputs. Video data is available in abundance and can be effectively utilized for action recognition. The process of action recognition comprises two main tasks, namely, action classification and action localization. The former involves assigning labels to instances of action in a video, and the latter defines the temporal and spatial bounds. Action recognition tasks are challenging due to the lack of complete datasets and high computational cost levels. Significant research has made action recognition a less cumbersome process. A concise summary of multiple datasets employed for action recognition has been presented in the paper. The most used datasets are

compared based on several acting classes, data size, nature of video clips, and their aim. Among the available datasets, the Kinetics 600 dataset has the maximum number of action classes. Although this dataset offers high variation in action types, the videos are trimmed and do not depict real-life scenarios. Contrarily, the ActivityNet dataset offers 200 action classes with untrimmed videos and is a better depiction of real-life activities.

Most of the recent algorithms can localize action in long untrimmed videos with limited computational capacities. The creation of better datasets can significantly improve the performance of these algorithms. The introduction of Single Stream Deep Neural Networks profoundly enhanced the performance of action recognition algorithms. Although this was a considerable breakthrough, these networks had trouble capturing the motion features. It was after this invention that deep learning started to be widely used for action recognition purposes. Later, the introduction of Two Stream Networks made it possible to capture the motion features effectively. Even then, these networks still had a shortcoming of not being end-to-end trainable and fast. LSTMs and 3D convolution networks' proposal made it possible to develop end-to-end trainable, real-time action recognition systems. In the future, the performance of action recognition systems can be significantly increased with the creation of publicly available datasets that contain more action classes with untrimmed videos. Recognizing actions for specific use cases would be much more comfortable with the availability of task-specific datasets. Apparent and standardized documentation of the action recognition methodology would further help make more robust models. Considering a broader set of features and input from multiple sensors while creating models will also significantly improve action recognition systems' performance. The utilization of a range of sensors alongside vision based sensors will drastically improve the performance of deep learning models for action recognition purposes.

REFERENCES

[1] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu and Y. Liu, "Deep learning for sensor-based human activity recognition: overview, challenges and opportunities," arXiv preprint, vol. 37, August 2018.

[2] M. Zolfaghari, G. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2904-2913, 2018.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In IEEE computer society conference on computer vision and pattern recognition, vol. 1, pp. 886-893, 2005.

[4] N. Dalal, B. Triggs and C. Schmid, "Human detection using oriented histograms of flow and appearance," In European conference on computer vision, pp. 428-441, 2006.

[5] H. Wangl, M. Ullah, A. Klaser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," pp. 124.1-124.11, 2009.

[6] M. Rodriguez, J. Ahmed and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," In IEEE conference on computer vision and pattern recognition, pp. 1-8, 2008.

[7] K. Soomro, A. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," In IEEE conference on computer vision and pattern recognition, arXiv preprint, November 2012.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Piggio and T. Serre, "HMDB: a large video database for human motion recognition", In 2011 International Conference on Computer Vision," pp. 2556-2563, 2011.

[9] H. Jhuang, J. Gall, S. Zuffi, C. Schmid and M. Black, "Towards understanding action recognition," In Proceedings of the IEEE international conference on computer vision, pp. 3192-3199, 2013.

[10] Y. Jiang, J. Liu, A. Zamir, G. Toderici, I. Laptev, M. Shah and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes,", 2014.

[11] F. Caba, V. Escorcia, B. Ghanem, and J. Carlos, "Activitynet: A large-scale video benchmark for human activity understanding," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 961-970, 2015.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, C. Vijayanarasimhan and M. Suleyman, "The kinetics human action video dataset," In Proceedings of the IEEE conference on computer vision and pattern recognition, arXiv preprint, 2017.

[13] C. Gu, C. Sun, D. Ross, C. Vondrick, C. Pantofaru, Y. Li and C. Shmid, "Ava: A video dataset of spatio-temporally localized atomic visual actions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6047-6056, 2018.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems, vol. 27, pp. 568-576, 2014.

[16] J. Donahue, L. Anne, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625-2634, 2015.

[17] 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, pp. 4489-4497, 2015.

[18] L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4305-4314, 2015.

[19] B. Zhang, L. Wang, Z. Wang, Y. Qiao and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2718-2726, 2016.

[20] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, "Dynamic image networks for action recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3034-3042, 2016.

[21] Y. Zhu, Z. Lan, S. Newsam and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," In Asian Conference on Computer Vision, pp. 363-378, 2018.

[22] W. Byeon, Q. Wang, R. Kumar and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," In Proceedings of the European Conference on Computer Vision, pp. 753-769, 2018.

[23] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," In Advances in Neural Information Processing Systems, pp. 34-45, 2017.

[24] A. Diba, M. Fayyaz, V. Sharma, A. Karami, M. Arzani, R. Yousefzadeh and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification, arXiv preprint, 2018.

[25] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1302-1311, 2015.

[26] L. Wang, Y. Qiao, X. Tang and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2708-2717, 2016.

[27] F. Caba, J. Carlos and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1914-1923, 2016.

[28] Z. Shou, D. Wang and S. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1049-1058, 2016.

[29] S. Buch, V. Escorcia, C. Shen, B. Ghanem and J. Carlos, "Sst: Single-stream temporal action proposals," In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2911-2920, 2017.

[30] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei and J. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," 2019.

[31] S. Yeung, O. Russakovsky, G. Mori and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2678-2687, 2016.

[32] F. Heidarivincheh, M. Mirmehdi and D. Damen, Detecting the moment of completion: temporal models for localising action completion," arXiv preprint, 2017.

[33] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang and D. Lin, "Temporal action detection with structured segment networks," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2914-2923, 2017.

[34] Y. Tian, R. Sukthankar and M. Shah, "Spatiotemporal deformable part models for action detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2642-2649, 2013.

[35] D. Oneata, J. Revaud, J. Verbeek and C. Schmid, "Spatio-temporal object detection proposals," In European conference on computer vision, pp. 737-752, 2014.

[36] S. Manen, M. Guillaumin and L. Van Gool, "Prime object proposals with randomized prim's algorithm," In Proceedings of the IEEE international conference on computer vision, pp. 2536-2543, 2013.

[37] K. Soomro, H. Igrees, N and M. Shah, "Predicting the where and what of actors and actions through online action localization," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2648-2657, 2016.

[38] M. Jain, J. Van Gemert, H. Jegou, P. Bouthemy and C. Snoek, "Action localization with tubelets from motion," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 740-747, 2014.

[39] G. Gkioxari and J. Malik, "Finding action tubes," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 759-768, 2015.

[40] J. Van Germert, M. Jain, E. Garti and C. Snoek, "APT: Action localization proposals from dense trajectories," pp. 2-4, 2015.

[41] M. Zolfaghari, G. Oliveira, N. Sedaghat and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2904-2913, 2017.

[42] W. Chen and J. Corso, "Action detection by implicit intentional motion clustering," In Proceedings of the IEEE international conference on computer vision, pp. 3298-3306, 2015.

[43] A. El-Nouby and G. Taylor, "Real-time end-to-end action detection with two-stream networks," arXiv preprint, 2018.