# Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review

Di Wu, Nabin Sharma, and Michael Blumenstein
School of Software, University of Technology Sydney
New South Wales, Australia
Email: Di.Wu-16@student.uts.edu.au, {Nabin.Sharma,Michael.Blumenstein}@uts.edu.au

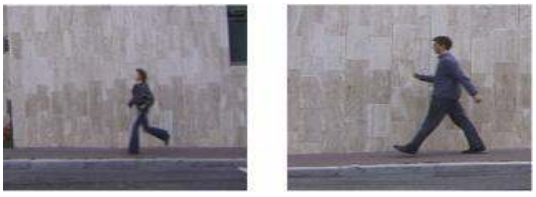
***Abstract*—Video-based human action recognition has become one of the most popular research areas in the field of computer vision and pattern recognition in recent years. It has a wide variety of applications such as surveillance, robotics, health care, video searching and human-computer interaction. There are many challenges involved in human action recognition in videos, such as cluttered backgrounds, occlusions, viewpoint variation, execution rate, and camera motion. A large number of techniques have been proposed to address the challenges over the decades. Three different types of datasets namely, single viewpoint, multiple viewpoint and RGB-depth videos, are used for research. This paper presents a review of various state-of-the-art deep learning-based techniques proposed for human action recognition on the three types of datasets. In light of the growing popularity and the recent developments in video-based human action recognition, this review imparts details of current trends and potential directions for future work to assist researchers.**

## I. Introduction

Recognizing human actions from a video stream is a challenging task and has received significant attention from the computer vision research community recently. Analysing a human action is not merely a matter of presenting patterns of motion of different parts of the body, rather, it is also a description of a person's intention, emotion and thoughts. Hence, it has become a crucial component in human behavior analysis and understanding, which are essential in various domains including surveillance, robotics, health care, video searching, human-computer interaction, etc. Different from still image classification, video data contains temporal information which plays an important role in action recognition. Additionally, video data includes natural data augmentation, e.g. jittering for video frame classification.

In recent years, much work has been done in different areas in the computer vision research area, such as video classification [5], resolution [6] and segmentation [7] etc. However, the research on video-based human activity recognition has not been explored much, due to the challenges in processing temporal information from the video stream. Action recognition from a video stream can be defined as recognizing human actions automatically using a pattern recognition system with minimal human-computer interaction. Typically, an action recognition system analyzes certain video sequences or frames to learn the patterns of a particular human action in the training process and use the learnt knowledge to classify similar actions during the testing phase [8]–[23]. Among the early state-of-the-art approaches [2], [24]–[27] for human action recognition, all of these investigations use motion and texture descriptors calculated based on the spatio-temporal interest points, which are built manually. Subsequently, they compute features from raw video frames and classifiers are trained based on the features obtained. Thus, even the features can be fully extracted automatically, and these hand-crafted features are used for specific problems. Therefore, the main drawback of these approaches is that they are problem-dependent, which is challenging to apply in the real-world, even though they may achieve high performance in action recognition.

(a) Gestures: Waving and Bend [1]



(b) Actions: Running and Walking [1]

(c) Interactions: Pickup phone call and Hugging [2]

(d) Group activities: Volleyball [3] and Basketball [4]

Fig. 1: Different classes of human activities

Over the last few years, deep learning-based approaches have become very popular in the video-based human action recognition research area as they have the ability to learn features from multiple layers hierarchically and build a high-level representation of the raw inputs, automatically. Therefore, unlike the traditional approaches, the feature extraction process is fully automated. For example, a deep learning system uses methods such as local perception, weight sharing, a multi-convolution kernel, down-pooling etc. to learn local features from a part of an image instead of the whole image. The final recognition output is determined by the result of multiple convolution layers. One of the popular deep learning approaches used for images/frames is a Convolutional Neural Network (CNN). A 3D CNN architecture [28] has been applied to generate multiple channels of information and perform convolution and sub-sampling in each channel from adjacent video frames [28]. The main advantage of deep learning approaches compared with traditional approaches is their ability to recognize high-level activities with complex structures. Hence, researchers prefer to use deep learning methods to incorporate the representation of features, such as time-space interest points, frequency, local descriptors and body modeling from video, depth video or RGB video datasets. The promising performance, robustness in feature extraction and the generalization capabilities of deep learning approaches are the major reason behind their increasing popularity and success.

Previous surveys have defined human activities into four classes: gestures, actions, interactions with objects and group activities [29]. Gestures [29] can be a static or dynamic elementary physical movement of a persons body parts, most commonly defined as the atomic components that represent the meaningful motion of a person such as "shaking hands" or "swinging arms". Unlike gestures, actions [29] are a combination of multiple gestures, for example, "running", which is a combination of arm and leg gestures. Two or more persons and/or objects involved in a human activity are defined as human interactions (human-human or human-object) [29]. For instance, "a boxing game" is an interaction between two persons and "a person picks up a cup" from a table is a human-object interaction. Lastly, group activities [29] can be defined as multiple persons/objects forming a group and participating in one activity such as "a meeting" or "a soccer match". Sample images for each of these types of human activities are shown in Figure 1.1 to Figure 1.4. Ramanathan et al. [30] discussed the various approaches and challenges involved in human action recognition with video data. Li and Kuai [31] focused on specific features, that is spatio-temporal interest points, whereas, the challenges of image representation and classification algorithms was discussed by Poppe [32]. Approaches on view-invariant pose detection and behavior understanding was reviewed by Ji and Liu [33]. Weinland et al. [34] presented a review on the approaches for action representation, segmentation and recognition. In addition, some of the datasets in human action and activity recognition were surveyed by Chaquet et al. [35]. Unlike other reviews/surveys, this review does not focus on classifying the existing approaches for different issues. In contrast, the paper reviews the recent developments in the use of deep learning techniques which have been applied in the human action recognition research area.

In this review, the main focus is on video-based human action recognition systems proposed for different types of video datasets in the past five years. The rest of the paper is organized as follows. Section 2 details the classification and differences between various datasets available for research. Deep learning-based human action recognition approaches applied to different datasets are discussed in Section 3. Sections 4 & 5 suggest the potential research opportunities and provide a conclusion, respectively.

## II. Datasets

With the development of human action recognition technology, many different types of datasets have been prepared and released recently. These datasets are widely used for experimental purposes to evaluate the performance and accuracy of existing/new approaches and to ensure appropriate comparison with other approaches. Generally, deep learning can be applied to different types of datasets with raw input data. In addition, the complexity of the networks may be determined by the different types of the datasets. For example, single viewpoint data may require less steps than multiple viewpoint data, which needs to generate multiple networks to obtain the final output. A depth camera may provide depth and RGB features using different technologies. Therefore, we classify the datasets as single viewpoint, multiple viewpoints and depth camera and RGB camera videos. These datasets offer dedicated features for different research purposes, such as gestures, 3D body modeling and joints etc. In this section, we review the popular public datasets on which deep learning techniques have been successfully applied. Table 1 lists the various datasets which are popularly used for research.

### *A. Single Viewpoint Datasets*

The single viewpoint datasets normally use a single camera recording human actions from a certain invariant angle without camera movement. These datasets were used for the analysis of human actions in the early stage of research, as shown in Figure 2. The earliest single viewpoint dataset was released in 2001 by Weizmann Institute [1]. This dataset recorded ten actions and each action was performed by ten persons. The foreground silhouettes are included in the dataset and the backgrounds are static as the viewpoints are static. In 2004, another dataset named KTH [36] was published. The KTH dataset contains six actions with four different scenarios, performed by twenty five actors. Similar to the Weizmann dataset, the backgrounds are static as well, except in the zooming scenarios. These early datasets have some drawbacks, such as videos are recorded in constrained environments and the actors perform simple identical actions in the video clips which are not the representative of human actions in the real world. To consider real scenarios, several other datasets were

TABLE I: Comparison of the datasets

| Name | Type | No. of View | No. of Actions | Website Link |
|---|---|---|---|---|
| Weizmann [1] | single-view | 1 | 10 | http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html |
| KTH [36] | single-view | 1 | 6 | http://www.nada.kth.se/cvap/actions/ |
| UCF sports [37] | single-view | 1 | 150 | http://crcv.ucf.edu/data/UCF_Sports_Action.php |
| Hollywood [2] | single-view | 1 | 8 | http://www.di.ens.fr/ laptev/actions/ |
| IXMAS [38] | multi-view | 5 | 14 | http://cvlab.epfl.ch/data/ixmas10/ |
| i3DPost [39] | multi-view | 8 | 12 | http://kahlan.eps.surrey.ac.uk/i3dpost_action/ |
| MuHAVi [40] | multi-view | 8 | 17 | http://dipersec.king.ac.uk/MuHAVi-MAS/ |
| Videoweb [41] | multi-view | 4-8 | 51 | http://www.ee.ucr.edu/ amitrc/datasets.php |
| CASIA Action [42] | multi-view | 3 | 8 | http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp |
| MSR-Action3D [43] | RGB-D | 1 | 20 | http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/ |
| DailyActivity3D [45] | RGB-D | 1 | 16 | http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/ |
| Multiview 3D [46] | RGB-D | 1 | 10 | http://users.eecs.northwestern.edu/ jwa368/my_data.html |
| CAD-60 [47] | RGB-D | 1 | 12 | http://pr.cs.cornell.edu/humanactivities/data.php |

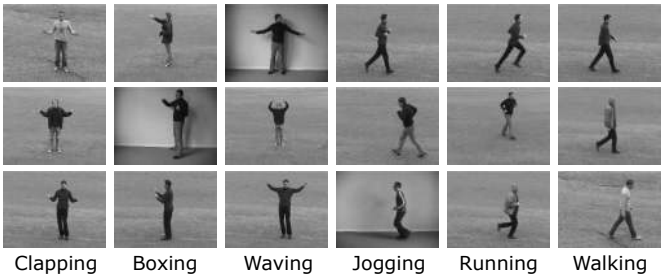



Fig. 2: Samples from a typical single viewpoint dataset KTH [36], where 25 actors perform six actions under different scenarios.

introduced, including UCF sports [37] and Hollywood datasets [2] which are extracted from YouTube or from movies. The UCF sports dataset contains 150 sports motions considering human appearance, camera movement, viewpoint change, illumination and background. The Hollywood dataset proposes eight actions to address the challenges of occlusions, camera movements and dynamic backgrounds. These datasets have a fixed viewpoint to monitor the actions in the video stream.

### *B. Multiple Viewpoint Datasets*

In a real-world scenario, multiple cameras are used for monitoring large public spaces, such as shopping malls, airports, trains and bus stations. Some multi-view datasets have been created specifically for studying the problem of processing multiple views of the same human. The advantages of these datasets is that they model a 3D human body shape from different angles and occlusion problems are avoided in contrast with single viewpoint streams.

Weinland et al. [38] released the IXMAS dataset which contains 14 actions performed by 11 persons. For each action, there are five cameras capturing the action from five angles with a static background and illumination settings. Sample images taken from the IXMAS dataset are shown in Figure 3, where multiple views of the same human actions are captured by different cameras placed at different viewpoints. Another indoor dataset, the i3DPost Multi-view dataset [39] was published in 2009. Eight high definition cameras were used to capture twelve actions performed by eight persons. Kingston University released their dataset in 2010 which was called MuHAVi [40]. They used eight non-synchronized cameras to capture 17 actions performed by 14 actors and it was designed to test different action recognition algorithms. Unlike the indoor datasets with static backgrounds, several datasets captured actions under real conditions, such as Videoweb [41] and the CASIA Action datasets [42]. In the Videoweb dataset, four groups of actors perform actions, which were captured by four to eight cameras tailored for group activity recognition. The CASIA Action dataset mainly focuses on interactions between persons and it contains eight types of single person actions performed by 24 people and seven types of interactions captured by three static cameras from different angles. These multi-view datasets can provide multiple streams as inputs for researchers.

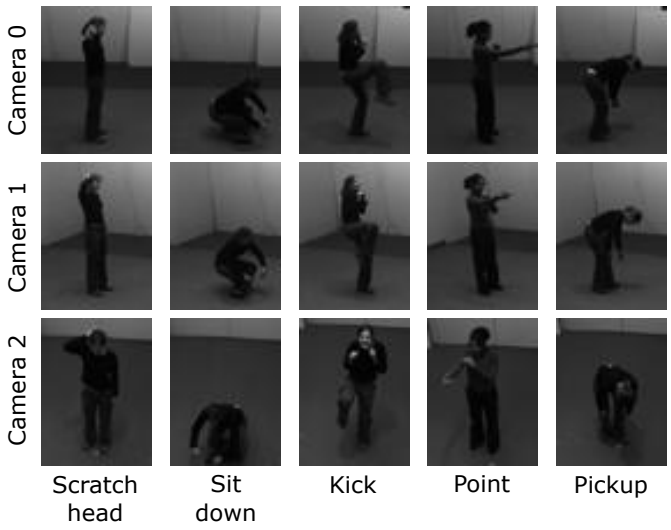



Fig. 3: Samples from the IXMAS dataset [38] where five cameras were used to capture the same activities from different angles at the same time.

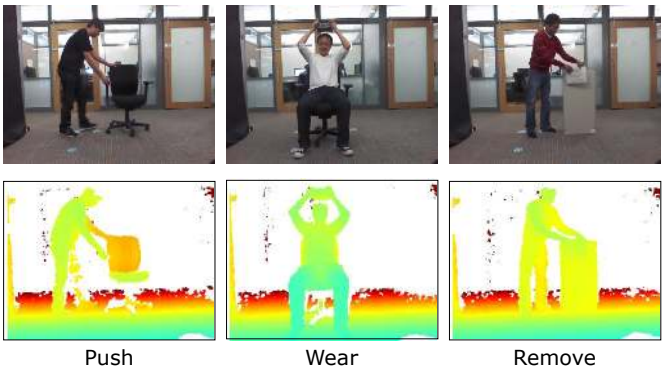



Fig. 4: Samples of RGB frames (Row 1) and corresponding depth maps (Row 2) from the MSR-Pairs dataset [44] which were acquired through a Kinect device. This device is used to collect depth and RGB videos with both on depth camera and an RGB camera.

### C. Depth and RGB Datasets

Depth and RGB video datasets are normally generated by specific devices, such as a depth camera or an RGB camera. In recent years, a device manufactured by Microsoft, named Kinect, has become quite popular as it contains features of both depth and RGB cameras. This device was specifically designed to capture human motion, and has been widely used in human recognition research. The depth and RGB videos not only contain video frames, but also have special data called depth maps, which are used to measure the depth of the objects from the observation point.

The MSR-Action3D dataset [43] uses a depth camera to capture depth sequences. It contains 20 action types performed by 10 subjects and each action was performed two to three times. This dataset is used to generate skeleton motions which can be used to describe the action precisely. Figure 4 shows some samples of RGB frames and the corresponding depth maps from the MSR-Pairs dataset [44] which were captured using Kinect. The depth maps shown in Figure 4 (row 2) were converted to color images for better visualization purposes. The DailyActivity3D dataset [45] is a daily activity dataset captured by a Kinect device which comprises 16 activity types. The Multiview 3D event dataset [46] contains RGB, depth and human skeleton data captured simultaneously by three Kinect cameras from different viewpoints and consists of ten action categories performed by ten actors. The Cornell Activity Datasets [47] recorded a RGB-D video sequence of human activities using Kinect. It has two sub-datasets CAD-60 and CAD-120, which are comprised of 60 RGB-D videos and 120 RGB-D videos, respectively. These RGB-D datasets are able to generate skeleton and 3D models from the video in order to describe human action accurately.

## III. Approaches for Datasets

In order to recognize high-level activities hierarchically, the multi-layered Hidden Markov Model (HMM) was introduced in the early stages of human action recognition research. Most HMM-based work has been performed on single viewpoint datasets. A fundamental form of the multi-layer approach was presented by Oliver et al. [48]. At the lower level, HMMs were used to recognize various sub-events, such as stretching and withdrawing. The upper level treats the result from the lower level as input and recognizes the punching activity when stretching and withdrawing occurred in a certain sequence. However, by nature, HMMs require strict sequences in each layer. Therefore, HMM approaches may not be able to meet the expectations of processing speed and system performance. This section focuses on the use of deep learning techniques with raw input data used by researchers for human action recognition on three types of video datasets. Since the approaches proposed were on different datasets and testing strategies, it is difficult to make a quantitative performance comparison. Even deep learning is still relatively new in this research area, however, it is crucial that these approaches have the ability to undertake high-level action recognition with high performance.

### A. Approaches for Single Viewpoint Datasets

In the last five years, different types of deep learning techniques have been applied on single viewpoint datasets. This is not only because single viewpoint human action recognition is the foundation of the research area and provides large scale datasets, but also the framework which has been developed for single viewpoints can be directly extended to multiple view points by generating multiple networks.

CNNs became a popular deep learning technique in the human action recognition area due to their ability to learn visual patterns directly from the image pixels without any pre-processing step. Baccouche et al. [49] introduced a two-step neural network-based deep learning model. The first step uses CNNs to learn the spatio-temporal features automatically and the following step uses a Recurrent Neural Network (RNN) to classify the sequence. Similarly, Ji et al. [28] proposed a 3D CNN for human action recognition. In 3D CNN architecture, they applied multiple convolution operations at the same location on the input which could extract multiple types of features. Then, multiple channels were generated to perform convolution and subsampling each channel from adjacent video frames. The final feature representation can be obtained by combining information from all channels. Factorized spatio-temporal CNNs [51] were designed to handle the spatial and temporal kernels in different layers which could reduce the number of learning parameters of the network. With the transformation and permutation operator, a training and inference strategy along with a sparsity concentration index scheme produced the final result, which outperformed existing CNN-based methods. Another work [49] shared a similar idea, the only difference being that they extracted the spatial and temporal information as a single frame and a multi-frame optical flow. This spatial and temporal information was fed into a spatial and temporal stream CNN, respectively. Ballas et al. [65] used a convolutional GRU-RNN (GRU-RCN) to process the visualization of convolutional maps on successive

TABLE II: Comparison of Single/Multiple View Approaches

| Author | Methods | Datasets | Performance (%) |
|---|---|---|---|
| Baccouche et al. [49] | CNN & RNN | KTH [36] | 94.39 |
| Ji et al. [28] | 3DCNN | TRECVID [50], KTH [36] | 78.24, 90.02 |
| Sun et al. [51] | Factorized spatio-temporal CNN (S-FTCN) | UCF-101 [52], HMDB-51 [53] | 88.1, 59.1 |
| Simonyan et al. [54] | two stream CNN, SVM | UCF-101 [52], HMDB-51 [53] | 88.0, 59.4 |
| Grushin et al. [55] | LSTM | KTH [36] | 90.7 |
| Veeriah et al. [56] | Differential RNN | KTH [36], MSR-Action3D [43] | 93.96, 92.03 |
| Shu et al. [57] | SNN | Weizmann [1], KTH [36] | 98.63, 92.3 |
| Ali and Wang [58] | DBN & SVM | KTH [36] | 94.3 |
| Shi et al. [59] | DTD, DNN | KTH [36], UCF50 [60] | 95.60, 95.24 |
| Wang et al. [61] | TDD, CNN | UCF-101 [52], HMDB-51 [53] | 95.1, 65.9 |
| Chéron et al. [62] | Pose-based CNN, IDT-FV | JHMDB [63], MPII Cooking [64] | 79.5, 71.4 |
| Ballas et al. [65] | GRU-RCN | UCF-101 [52] | 80.7 |

frames in a video. The results show that the Bi-Directional GRU-RCN Encoder outperforms the VGG-16 Encoder by 3.4% and 10% for action recognition compared to both RGB and Flow inputs, respectively.

Another architecture called Long Short Term Memory (LSTM), a variation of RNN, also received increasing attention in sequence processing. LSTMs use memory blocks to replace the regular network units. The gate neurons of the LSTM determine when it should remember, forget or output the value. It was previously used to recognize speech and handwriting. A robust LSTM [55] with recurrent cell connections was tested for action recognition to show that classification accuracy may be affected by training set size, length of the video sequence and quality of the video. Veeriah et al. [56] delivered a different gating scheme to address the problem of conventional LSTMs which emphasizes the change in information gain caused by the salient motions between successive frames. Then, the LSTM model was termed as differential RNN. This model can recognize actions from a single view or a depth dataset automatically.

Unlike other neural networks, Spiking Neural Networks (SNNs) work similarly to their biological counterparts. A special model based on SNNs was designed by Shu et al. [57], which is a hierarchical architecture of the feed-forward spiking neural networks modeling two visual cortical areas: primary visual cortex (VI) and middle temporal area (MT), neurobiologically dedicated to motion processing. It simulates the working mechanism from the VI and MT. After detecting the motion energy, the information is processed by the VI layer and MT layer. The motion energy is first transformed by the spiking neuron model in the VI layer, then the MT cell pools the information received from the VI cell according to the mapping connection between the two layers. Features are extracted from the spike trains which are generated by MT spiking neurons. The final output is recognized by an SVM classifier. Ali and Wang [58] built a Deep Brief Network (DBN) which is another variant of deep neural networks. It is composed of multiple hidden unit layers with connections between the layers to the learning feature for action recognition.

Some of the methods prefer to extract different descriptors as input before using deep learning techniques. In [59], researchers firstly extract dense trajectories from raw data with multiple consecutive frames and then project the trajectories onto a canvas. In this way, they can transfer the raw 3D space into a 2D space and import them, hence, the complexity of the data is reduced. Subsequently, they input the data into a Deep Neural Network (DNN) which is utilized to learn a more macroscopical representation of dense trajectories. Some additional features are extracted and used as inputs to the classifier. Wang et al. [61] claimed that their trajectory-pooled deep-convolutional descriptor (TDD) outperformed the hand-crafted features with higher discriminative capacity. A posed-based CNN [62] descriptor was used for action recognition which was generated based on human poses. The input data was divided into five part patches. For each patch, two kinds of frames were extracted from the video, namely RGB and flow frames. The P-CNN features are generated by both frames and processed in the CNN, respectively after aggregation and normalization stages. Table 2.2 presents a comparative study of different single/multiple view approaches.

### *B. Approaches for Multiple Viewpoint, Depth and RGB Datasets*

Multiple viewpoint datasets contain information from multiple cameras from different directions which naturally avoids the drawback of occlusion and it also captures different views of the same gestures from different angles, thereby, it provides more information for better performance. In addition, the advantage of depth and RGB datasets is that the skeleton information or trajectories can be generated directly. This information could represent a human action. Researchers are paying more attention to these datasets to achieve high performance in relation to recognition based on skeleton and trajectory information.

A fuzzy CNN was presented to deal with motion capture information (MOCAP) [66], The MOCAP is widely used for human-skeletal prediction from depth and multi-view videos. They use the ability of CNNs to recognize local patterns and an analysis of MOCAP information can achieve high

TABLE III: Comparison of Depth and RGB View Approaches

| Author | Methods | Datasets | Performance (%) |
|---|---|---|---|
| Ijjina et al. [66] | MOCAP, CNN | Berkeley MHAD [67] | 99.248 |
| Wang et al. [68] | WHDMM, Deep CNN | MSR-Action3D [43], MSRDailyActivity3D [45], UTKinect-Action [69] | 100.00, 85.00, 90.91 |
| Du et al. [70] | RNN, LSTM | MSR-Action3D [43], Berkeley MHAD [67], Motion Capture Dataset HDM05 [71] | 94.49, 100.00, 96.92 |
| Zhang et al. [72] | MTRL | SARCO [73] | Mean = 0.5156 |
| Yang et al. [74] | MTL | MSR-Action3D [43], UTKinect-Action [69], Florence3D-Action [75] | 95.62, 98.80, 93.42 |
| Liu et al. [76] | MTSL | TJU(self constructed), MV-TJU(self constructed), KTH [36] | 97.6, 95.8, 96.7 |

classification accuracy. Wang et al. [68] applied three channel deep CNNs to recognize human actions using weighted hierarchical depth motion maps. They evaluated their algorithm on some popular depth datasets using cross-subject protocols and the results achieved 2-9% performance improvement on most of the individual datasets. Du et al. [70] proposed an RNN combined with an LSTM architecture. It divides the human skeleton into five parts, based on the human body structure and feeds them into five subnets, called bidirectional RNNs (BRNNs). The LSTM neurons are adopted in the last BRNN layer to overcome the vanishing gradient problem.

The multi-task learning approach demonstrates its effectiveness as a hierarchical method to learn several tasks to capture intrinsic correlations [72]. A latent max-margin multi-task learning model [74] has been proposed to address flexibility for incorporating latent "skelet". It is a combination of a subset of joints and it achieves maximum margin separation among the action classes. Liu et al. [76] also tested their part-regularized multi-task structural learning framework with the hierarchical part-wise bag-of-words representation on single-view, multi-view and depth datasets. They generated three levels of classifiers, each level focussing on the visual saliency of different body parts. Consequently, the performance of all the three kinds of datasets significantly improved in comparison to the standard bag-of-words methods. Table III shows the comparison of different depth and RGB approaches.

## IV. Future Work and Discussion

There are many challenges which need to be addressed in this research area. The first one is multi-view human action recognition. In our review, it was found that few researchers have pursued this scenario. We may be able to generate multiple networks for different streams to monitor detailed human actions. Recently, the datasets all perform simple actions, those actions being non-emotional and intentional which makes it challenging to describe why people act in a certain way. For example, a "punch" action may be defined as a fight activity or a greeting between friends depending on the strength, speed of the punch and other atomic-level actions, perhaps a smile. To recognize such actions, we need to analyse multiple networks from different streams. For each stream, different networks may monitor different objects such as the face, body parts and motion etc. The final output could cover different results from all the streams. Classifying multiple activities from single view video frames may be another challenge. The problem of tracking or detecting multiple persons in a certain video is a problem that has been solved for years [77]. However, most of the datasets are still concerned with the performance of an activity by a single person. This is because to classify multiple activities requires multiple networks, however, one input stream normally can generate only one network. If we want to generate multiple networks, we need to undertake the pre-processing step on the dataset to extract the inputs for the networks, similar to the work mentioned in this review. Hence, to generate multiple networks automatically based on the detected regions in one single view stream would be the second step. This review gives an overview of the current developments in hierarchical statistical approaches in the area of video-based human action recognition. This will help researchers to focus their research effort on the pressing challenges, which will most likely advance knowledge in this area.

## V. Conclusion

Deep learning techniques have recently been introduced in the video-based human action recognition research area. They have been widely used in other areas, such as speech recognition, language processing and recommendation systems etc. There are many advantages to hierarchical statistical approaches, such as raw data input, self-learnt features and a high-level or complex action recognition, hence, deep learning techniques have received much interest. Based on these advantages, researchers could design a real-time, adaptive and high performing recognition system. However, these approaches also have several drawbacks, such as the need to generate large datasets, the performance depends on the scale of the network weights and hyper-parameter tuning is non-trivial etc.

In this review, we presented techniques mainly focusing on developments in deep learning over the past five years. Many investigations have been conducted to deal with different types of datasets. For single/multiple viewpoint approaches, the inputs are normally frames, so researchers have performed 3D convolution operations to add the temporal information in order to recognize videos. Additionally some of the approaches

could also be used to generate features for different classifiers. In depth and RGB datasets, skeleton structure, gestures and body motion are the main descriptors for hierarchical statistical approaches to recognize or predict human actions.

## REFERENCES


[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1395–1402.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[3] Z. D. A. V. G. M. Mostafa S. Ibrahim, Srikanth Muralidharan, "A hierarchical deep temporal model for group activity recognition," in *CVPR*, 2016.

[4] S. A.-E.-H. A. G. K. M. Vignesh Ramanathan, Jonathan Huang and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *CVPR*, 2016.

[5] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *International Conference of Learning Representations*, 2016.

[6] A. A. J. Johnson and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *In European Conference on Computer Vision (ECCV)*, 2016.

[7] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *In European Conference on Computer Vision (ECCV)*, 2016.

[8] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 527–540, 2013.

[9] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Neural representation and learning for multi-view human action recognition," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–6.

[11] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[12] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[13] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "Human action recognition using robust power spectrum features," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 753–756.

[14] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1860–1870, 2013.

[15] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[16] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "View-independent behavior analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 4, pp. 1028–1035, 2009.

[17] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[18] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1635–1648, 2013.

[19] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. IEEE, 1992, pp. 379–385.

[20] P. Natarajan and R. Nevatia, "Coupled hidden semi markov models for activity recognition," in *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*. IEEE, 2007, pp. 10–10.

[21] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 533–547, 2012.

[22] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[23] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

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

[25] M.-y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," *CMU-CS-09-161*, 2009.

[26] Z. Gao, M.-Y. Chen, A. G. Hauptmann, and A. Cai, "Comparing evaluation protocols on the kth dataset," in *International Workshop on Human Behavior Understanding*. Springer, 2010, pp. 88–100.

[27] J. Liu and M. Shah, "Learning human actions via information maximization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[28] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[29] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," in *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*. IEEE, 1997, pp. 90–102.

[30] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

[31] Y. Li and Y. Kuai, "Action recognition based on spatio-temporal interest points," in *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*. IEEE, 2012, pp. 181–185.

[32] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[33] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 13–24, 2010.

[34] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[35] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[36] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[37] M. Rodriguez, "Spatio-temporal maximum average correlation height templates in action recognition and video summarization," 2010.

[38] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[39] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *Visual Media Production, 2009. CVMP'09. Conference for*. IEEE, 2009, pp. 159–168.

[40] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 48–55.

[41] G. Denina, B. Bhanu, H. T. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, and B. Varda, "Videoweb dataset for multi-camera activities and non-verbal communication," in *Distributed Video Sensor Networks*. Springer, 2011, pp. 335–347.

[42] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on r transform," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[43] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 9–14.

[44] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[45] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2649–2656.

[46] D. R. Faria, C. Premebida, and U. Nunes, "A probabilistic approach for human everyday activities recognition using body motion from rgb-d images," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 732–737.

[47] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[48] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*. IEEE, 2002, pp. 3–8.

[49] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[50] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

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

[52] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 2012.

[53] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[54] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[55] A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra, "Robust human action recognition via long short-term memory," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–8.

[56] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4041–4049.

[57] N. Shu, Q. Tang, and H. Liu, "A bio-inspired approach modeling spiking neural networks of visual cortex for human action recognition," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 3450–3457.

[58] K. H. Ali and T. Wang, "Learning features for action recognition and identity with deep belief networks," in *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*. IEEE, 2014, pp. 129–132.

[59] Y. Shi, W. Zeng, T. Huang, and Y. Wang, "Learning deep trajectory descriptor for action recognition in videos using deep neural networks," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015, pp. 1–6.

[60] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[61] 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*, 2015, pp. 4305–4314.

[62] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3218–3226.

[63] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3192–3199.

[64] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1194–1201.

[65] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *International Conference of Learning Representations*, 2016.

[66] E. P. Ijjina and C. K. Mohan, "Human action recognition based on motion capture information using fuzzy convolution neural networks," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015, pp. 1–6.

[67] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 53–60.

[68] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, 2015.

[69] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[70] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[71] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," Universität Bonn, Tech. Rep. CG-2007-2, June 2007.

[72] Y. Zhang and D.-Y. Yeung, "A regularization approach to learning task relationships in multitask learning," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 3, p. 12, 2014.

[73] S. Vijayakumar, A. D'souza, T. Shibata, J. Conradt, and S. Schaal, "Statistical learning for humanoid robots," *Autonomous Robots*, vol. 12, no. 1, pp. 55–69, 2002.

[74] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao, "Latent max-margin multitask learning with skelets for 3-d action recognition," *IEEE Transactions on Cybernetics*, 2016.

[75] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 479–485.

[76] A.-A. Liu, Y.-T. Su, P.-P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multipe/single-view human action recognition via part-induced multitask structural learning," *IEEE transactions on cybernetics*, vol. 45, no. 6, pp. 1194–1208, 2015.

[77] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.