Chapter 2

Related work

In this section, we provide an overview of recent research on deep learning for human action recognition. The publications have been categorized according to a proposed taxonomy, which includes the following approaches: human action recognition using Convolutional Neural Networks (CNNs), Recurrent Neural Network-Long Short-Term Memory (RNN-LSTMs), and various other architectures

2.1 CNN-based for Action Recognition

There have been numerous studies on video-based human action recognition that utilize Deep Learning (DL) models. Among these studies, CNNs and their extensions have emerged as one of the most widely utilized DL models. Researchers have demonstrated the effectiveness of CNN-based architectures for various tasks, such as detecting and tracking people [8–10], action recognition [11–16], pose estimation [17–19], event detection and crowded scene understanding [20–22]. In 1995, Nowlan et al. [17] made early attempts to use CNNs for hand tracking and recognition. They proposed a CNN model that could accurately locate the hand and determine whether it was open or closed. Despite the impressive results, the complexity of the image backgrounds was found to significantly affect the recognition accuracy. Nevertheless, recognition accu-

racy can be significantly affected by the intricately structured backgrounds in images. Building upon the research of [23], introduced a hierarchical feed-forward architecture for recognizing biological movements, such as walking, running, or other full-body actions. In [24], a hierarchical model is created for detecting walkers by utilizing neural detectors that can extract motion features at different levels of complexity. In [25], authors extended the work of [26] by proposing a model for recognizing actions from video sequences. Kim et al. [27] employed a modified CNN model and a weighted fuzzy min-max neural network (WFMM) for human action recognition. Typically, CNNs have been applied only to two-dimensional data (2D-CNN), computing features from the spatial dimensions alone. To make use of the temporal information of human motion, Ji et al. [28] introduced a novel three-dimensional convolutional neural network (3D-CNN) architecture for recognizing human actions. The design employs 3D kernels in the convolution stages to extract motion features from both the spatial and temporal domains, enhancing the ability to gather multiple features from consecutive frames in a video. Motivated by Ji et al. [28] and Wang et al. [29] have also used 3D-CNN for building a deep architecture for human activity understanding using RGB-D data. Ji et al. introduced a groundbreaking 3D convolutional neural network (3D-CNN) architecture for recognizing human actions in [28] that leverages the temporal information of human motion. This design uses 3D kernels in the convolutional layers to extract motion features from both the spatial and temporal dimensions, thereby increasing the ability to gather multiple features from consecutive frames in a video. Following in the footsteps of Ji et al. [28] and Wang et al. [29] also utilized 3D-CNN to construct a deep architecture for human activity understanding using RGB-D data. Furthermore, Tran et al. conducted a comprehensive investigation of the 3D-CNN model in [12] and demonstrated that it surpasses the 2D-CNN in capturing human motion information for various recognition tasks. Moreover, [12] found that the best kernel size for 3D-CNN is $3 \times 3 \times 3$. In [30], the authors utilized 3D-CNN to learn action representation

in videos but with long-term temporal convolutions at the input layer. This research showed that this approach can significantly enhance performance on state-of-the-art action recognition datasets. A drawback of the 3D-CNN model is the growing number of network parameters. To mitigate the complexity of the model, the authors proposed a factorized spatiotemporal convolutional network in [31]. The factorized spatiotemporal convolutional network breaks down the 3D convolution kernels into 2D spatial kernels, followed by 1D temporal kernels, to reduce the complexity of the model. In [32], Ijjina et al. identified human actions in videos by employing the standard action bank [33] as a feature detector and a CNN as a classifier. In [34], the authors have proposed state-of-the-art performance for predicting actions on the PASCAL VOC 2012 detection and action train set by using the CNN architecture. Chéron et al. [35] developed a CNN architecture for recognizing human actions from RGB and optical flow data. The two-stream convolutional network proposed by Simonyan and Zisserman is a two-stream architecture that includes a spatial stream and a temporal stream, each executed by a CNN, and has shown strong performance for human action recognition in videos [11]. The spatial stream recognizes actions from a single frame, while the temporal stream recognizes actions from motion information of multi-frame optical flow, and the two streams are combined for the classification task. Using multi-frame optical flow for training allows for achieving very good performance with limited training data, making it the most effective approach for applying DL to action recognition with limited training data. Many authors have developed two-stream convolutional networks for action recognition [36–38]. Inspired by the work of Simonyan and Zisserman et al. [11], many different authors have developed two-stream convolutional networks for solving action recognition problems, e.g., Wang et al. [36–38], Xiong et al. [39]. Inspired by the two-stream architecture Simonyan and Zisserman et al. [11], Liu et al. [40] proposed stCNN (Spatio-Temporal Convolutional Neural Network) to the standard CNN model for exploiting motion and content-dependent features concurrently. Experiments on

KTH [41] and UCF 101 [1] datasets showed the improvement in recognition accuracy for motion-content. The motion-content combined was better when compared with motion alone. Singh et al. [42] addressed the problem of understanding egocentric activities by using a three-stream CNN architecture. More specifically, the authors proposed a framework for the recognition of wearer's actions. First, a CNN model called "Ego Convnet" is trained for learning features from egocentric cues including hand mask, head motion, and a saliency map. Then, Ego Convnet is extended by adding two more streams corresponding to spatial and temporal streams as the model proposed by Simonyan and Zisserman et al. [11]. Experiments showed that the model with the Ego Convnet stream alone achieved state-of-the-art accuracy on different egocentric videos datasets. In addition, the three-stream architecture. Advances of 3D sensors such as Microsoft Kinect [43] brings up new opportunities in computer vision, even though they tend to be limited to small indoor environments. RGB-D data is able to provide additional information about human motion. Take advantage of depth maps provided by Kinect sensors, Wang et al. [44] proposed the use of CNNs to learn actions from sequences of depth maps. Given a sequence of depth maps, 3D points are created and three Depth Motion Maps (DMMs) are constructed by projecting the 3D points to the three orthogonal planes. Three CNNs are constructed based on AlexNet architecture [45] to extract motion features from each DMM and then classify them into classes. This study is extended in [46] and [47]. State-of-the-art results have been shown on MSR Action3D Dataset [48], an extension of the MSR Action3D Dataset, UTKinect-Action Dataset [49], and MSR-Daily-Activity3D Dataset [50]. Dobhal et al. [51] also used depth information and a CNN for recognizing human activities. Given a sequence of 2D images, background subtraction is performed. All binary frames are then stacked into a single image called Binary Motion Image (BMI) which contains the flow of the action and is used as the input for the CNN in training and testing phases. The CNN's architecture is same the architecture introduced by LeCun and Bengio [52].

Their approach is extended for extracting BMI from 3D depth maps and achieved competitive performance on Weizmann [53] and MSR Action3D Dataset [48]. The key ideas behind CNNs such as "local connections" or "shared weights" and the improvements on GPU computing technology have enabled CNNs to train on very large scale datasets. Karpathy et al. [54] studied the performance of CNNs by trying to predict and classify on Sports-1M [55] dataset which consists of more than one million sport videos. Multi-resolution CNN architecture with two separate streams of processing has been proposed for reducing training time. The results show that CNNs are capable of learning powerful features and significantly outperform the feature-based baseline. Some examples of predictions on Sports-1M dataset [55] with fine-tuning technique. The authors also proposed to train CNN from scratch by generating more dynamic images from video segments. Experiments on HMDB-51 and UCF-101 datasets shown the effectiveness of the "Dynamic Image" representation. In addition to RGB-D information, the acquisition of the skeleton data has become easier with the support of RGB-D sensor. Mo et al. [56] presented a deep model which combines a CNN with a multilayer perceptron [57] for recognizing the human activities based on skeleton data acquired from a Kinect sensor [43]. The method achieves a recognition accuracy of 81.8% on the CAD-60 dataset [55]. Skeleton data has been used by Wang et al. [58]. Firstly, the spatio-temporal information of the joint trajectories is encoded into color images. Then, a CNN based on the AlexNet architecture [45] is used to learn the color distribution and to classify actions. The idea of encoding the spatio-temporal information of a skeleton sequence into color texture images and using a standard CNN architeture such as AlexNet [45] can also be found in the work of Hou et al. [59]. Very deep convolutional neural networks such as VGGNet [60], GoogLeNet [61] have achieved significant success for object recognition and classification tasks. Several authors started to exploit these architectures for action recognition problems. Wang et al. [62] introduced very deep two-stream CNNs for action recognition based on VGG-16 (VGGNet

C with 13 convolutional layers and 3 fully-connected layers) and GoogLeNet [61] with 22-layers network. Feichtenhofer et al. [63] proposed a CNNs-based novel architecture for spatio-temporal fusion of two stream networks in which the deep CNN model VGG-M-2048 [64] and very deep model VGG-16 [60] have been used. The performance comparison between deep (VGG-M-2048) and very deep (VGG-16) models on UCF-101 and HMDB-51 datasets shown that the use of deeper networks improves performance. In addition, GoogLeNet [61] and VGGNet [60] have also been used to design the two-stream CNNs in the work of Wang et al. [65]. Fernando et al. trained VGG-16 [60] on HMDB-51 [2], UCF-101 [1] and Hollywood2 [66] datasets for obtaining VGG-16 CNN features. The CNN feature vectors are then encoded by a method called "hierarchical rank pooling". This method allows encoding the temporal dynamics of a video sequence for action recognition. A video sequence is encoded at multiple levels in which the output of the each level is a sequence of vectors that captures higher-order dynamics of its previous level. The final representation can be used to learn an SVM classifier for activity recognition as descriptors. Among the local space-time features, trajectories are one of the best ways to describe motion [67–69]. Wang et al. [46] combined the benefits of improved trajectories [67] and two-stream CNN architecture from the work of Simonyan and Zisserman et al. [11] for designing an effective representation of video feature called "Trajectory-Pooled Deepconvolutional Descriptor (TDD)". The experimental results show that this framework has obtained state-of-the-art performance for recognizing action on the UCF-101 [1] and HMDB51 datasets [2]. Inspired by the work of Wang et al. [46], Cao et al. [70] proposed a novel 3D deep convolutional descriptor based on joint positions named "Joints-Pooled 3D Deep Convolutional Descriptors (JDD)". Promising experimental results on sub-JHMDB [71], Penn Action [72], and Composable Activities [73] have shown that using joint-based descriptor with deep model is an effective and robust way for understanding human action. A new powerful and simple representation of videos for action recognition based on DL, especially CNNs,

called "Dynamic Image" has been presented in the work of Bilen et al. [74]. The idea of this work is summarizing the video content in a single standard RGB image, then using a pre-trained CNN model such as AlexNet [45] on a dataset of dynamic images. Feichtenhofer et al. employed the residual learning framework (ResNets) [75], which is a state-of-the-art and the deepest CNN model currently available, for human action recognition [76]. The underlying network with 50 layer ResNet has been used in the work of Feichtenhofer et al. [76] to design a two-stream network [75]. The experimental results demonstrate that the system is capable of detecting objects, predicting events, and achieving state-of-the-art performance on a very large dataset.

2.2 Human action recognition based on RNN-LSTMs

The main advantage of RNN-LSTMs is the capacity to model the long-term contextual information of temporal sequences. This advantage puts RNN-LSTM as one of the best sequence learners for time-series data including visual information of human action. CNNs has been shown its effectiveness in learning features from raw data. Therefore, the works of Baccouche et al. [77], Ng et al. [78], Donahue et al. [79], Giel et al. [80], Sharma et al. [207], Ibrahim et al. [81], Singh et al. [82], Li et al. [83], Wu et al. [84], Wang et al. [212], Chen et al. [85] tackle the question of understanding human actions by combining a CNN and an RNN-LSTM network. The general idea of these works is to use the standard CNN models such as AlexNet [45], VGGNet [60], or GoogLeNet [61] for extracting motion features from input video. Then, RNN-LSTM network is connected to the output of the CNN to classify sequences using learned features. CNN and RNN-LSTM for human action recognition from the work of Donahue et al. [79] is mentioned. While all the work above just uses RNN-LSTMs as a sequence classification, several studies have proposed the use of RNN-LSTMs as an end-to-end learning framework for skeleton. The Siamese network architecture has been designed for learning action features. In fact, this is a two-stream CNN models where the first stream is trained

on the precondition state frames and the second is trained on the effect state frames. E.g., the work of Du et al. [86], Song et al. [87], Zhu et al. [88], Li et al. [89], Liu et al. [90]. RNN LSTMs learn directly motion features and classify them into classes from 3D human-skeleton sequences provided by depth sensors. Experiments on the state-of-the-art datasets demonstrate the effectiveness of these methods. In another study of Mahasseni et al. [91] used a parallel architecture to recognize actions with multi-source data. A RNN-LSTM is trained in unsupervised manner on 3D human-skeleton sequences. In the same time, another RNN-LSTM with a CNN is trained on 2D videos. The outputs are then compared to improve the ability of the system.

2.3 Other deep architectures for human action recognition

Other deep architectures have been utilized for tasks similar to human action recognition, such as group activity analysis or physical interaction prediction. Sparse coding has been identified as a potential deep model for identifying human action, with successful applications in various fields including pattern recognition and image classification [92–94]. Sparse coding has been leveraged by numerous authors to solve human action recognition problems, using the sparse representations of signals as image features for classifiers. In addition to sparse coding, other deep architectures have been utilized for similar tasks, such as group activity analysis or predicting physical interactions. Sparse coding is a potential deep model for identifying human action, and its success in various fields, including pattern recognition, has demonstrated its ability to adapt flexibly to diverse low-level natural signals [95, 96] or image classification [97] has shown that it could flexibly adapt to diverse low-level natural signals. The image features extracted through sparse representations of signals are directly inputted into classifiers. Therefore, many authors [98–102] Authors have taken advantage of the benefits of sparse coding for solving human action recognition problems. In recent literature, novel deep architectures for recognizing human action have been

published [103–105]. For instance, Ullah and Petrosino [103] employed a CNN and a pyramidal neural network (PyraNet) [106] to recognize human action. The construction of a strict 3D pyramidal neural network (3DPyraNet) enables the learning of spatio-temporal features of human motion. These works continued to be expanded by the same authors [104] and achieved competitive results on some action datasets. Rahmani et al. [105] presented the "Robust Non-Linear Knowledge Transfer Model" (R-NKTM), a deep fully-connected neural network that is capable of understanding human action from cross-view by learning features from dense trajectories of synthetic 3D human models and real motion capture data. The work published by Le et al. [107] reports that we can combine the different network models to build a single deep architecture for improving its performance. Based on two key ideas, "convolution" and "stacking" in CNN architecture, the authors constructed a deep model by using the Independent Subspace Analysis (ISA) [108] and Principal Component Analysis (PCA) [109]. The ISA is trained on small input patches for learning features directly from unlabeled video data. It is then convolved with a larger region of the input image. The PCA algorithm is applied on top of ISA for reducing dimensions. The responses are then used as the input layer for another ISA. The method is evaluated on KTH [41], Hollywood2 [66], UCF sports [110] and YouTube datasets [111]. Srivastava et al. [112] constructed a model which consists of two LSTMs- the encoder LSTM and the decoder LSTM to learn representations of sequences of images. The state of the LSTM encoder is the representation of the input video. Then, the LSTM decoder will reconstruct the input sequence from this representation. It can be used for reconstructing the input sequence as well as predicting the future sequence. Very recently, Luo et al. [113] combined many different models to build a deep-learning framework for recognizing human motion in Videos. The idea is to design a network that is able to predict future 3D motions in videos. Given input frames, the model will predict 3D flows in future frames, then use the features to recognize activities. To do that, a Recurrent Neural Network based

Encoder-Decoder framework has been proposed. During the encoding process, CNNs (the standard VGG-16 networks) are used for extracting a low-dimensionality feature from the input frames.

2.4 Conclusion

The aim of our research is to provide readers with a detailed insight into the development process and current progress of deep learning models used for human action recognition in videos. We have conducted a comprehensive review of over 200 publications on various deep learning architectures and their applications in action recognition and related tasks. Our analysis and comparison of recognition accuracy between deep learning-based approaches and other techniques show that, currently, deep learning is the best choice for recognizing and classifying human action and predicting human behavior. We have also analyzed the characteristics of the most important deep learning architectures for action recognition to provide current trends and identify open problems for future works in this field. Furthermore, we have compiled a list of datasets with varying complexity levels to assist readers in selecting appropriate algorithms and datasets for developing new solutions. Despite significant progress made in recent years, challenges remain in applying deep learning models to build vision-based action recognition, such as two-stream CNN models where one stream is trained on precondition state frames and the other on effect state frames. We hope this survey will be valuable to researchers in this field.