Action Recognition in Videos

Zineng Xu

A thesis submitted for the degree of Master in Artificial Intelligence

Facultat d'Informàtica de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA

June 18, 2018

Advisor: Prof. Josep Ramon Morros

Co-Advisor: Prof. Verónica Vilaplana

Declaration

I, Zineng Xu, declare that this thesis titled, “Action Recognition in Videos” and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a master degree at UPC.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at UPC or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Zineng Xu

June 2018

Acknowledgements

I would first like to thank my thesis advisor Prof. Josep Ramon Morros and Prof. Verónica Vilaplana of the Signal Theory and Communications Department at UPC. The door to their offices were always open whenever I ran into a trouble spot or had a question about my research or writing. They consistently allowed this paper to be my own work, but steered me in the right the direction whenever they thought I needed it.

I would also like to thank Albert Gil Moreno for giving us access to the UPC computing cluster, where I had access to two GPUs, and helped me with some troubles during the installation of the system.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

In this project, our work can be divided into two parts: RGB-D based action recognition in trimmed videos and temporal action detection in untrimmed videos.

For the action recognition part, we propose a novel action tube extractor for RGB-D action recognition in trimmed videos. The action tube extractor takes as input a video and outputs an action tube. The method consists of two parts: spatial tube extraction and temporal sampling. The first part is built upon MobileNet-SSD and its role is to define the spatial region where the action takes place. The second part is based on the structural similarity index (SSIM) and is designed to remove frames without obvious motion from the primary action tube. The final extracted action tube has two benefits: 1) a higher ratio of ROI (subjects of action) to background; 2) most frames contain obvious motion change. We propose to use a two-stream (RGB and Depth) I3D architecture as our 3D-CNN model. Our approach outperforms the state-of-the-art methods on the OA and NTU RGB-D datasets.

For the temporal action detection part, we follow the "proposal + classification" framework to propose a three-stage temporal action detection system: 1) multi-scale segment generation: construct multi-scale candidate segments with sliding window scheme and the SSIM based sampling method; 2) proposal generation: action proposals are generated via "multi-stream" I3D and Temporal Actionness Grouping (TAG); 3) action classification: classify each generated proposal to form final detection results. Due to the lack time of this project, a simplified system is tested on THUMOS14 dataset. The detection results on this challenging dataset demonstrate the effectiveness of the proposed action detection system.

Contents

1 Introduction

Human action recognition in videos has been an active research area in the last few years due to its potential applications, including intelligent surveillance, robotics, health-care monitoring, video retrieval, and interactive gaming. Compared to still image recognition, the temporal component of videos provides an additional clue for recognition, as many actions can be reliably recognized based on the motion information. Thus, the concept of action recognition can be simply defined as assigning a video to a set of predefined action classes.

In this report, we firstly discuss action recognition in temporally trimmed videos, which constitutes the main part of the thesis work, and then we present an extension to temporally untrimmed videos.

1.1 Trimmed Action Recognition

In past decades, research on human action recognition has been extensively explored in temporally trimmed videos (RGB frames). These collected videos are carefully trimmed to only contain the actions of interest. With the development of imaging devices (e.g. Microsoft Kinect), it is possible to capture low-cost and high sample rate depth images in real-time alongside color (RGB) images. Thus, RGB-D based trimmed action recognition has attracted much attention in recent years. Depth is insensitive to illumination changes and has rich 3D structural information of the scene. Therefore, fusing this multimodal information into feature sets can lead to methods that achieve higher performance. Compared to RGB videos (mostly collected from movies/sports lives/YouTube videos, for general action recognition), RGB-D videos are mostly collected from predefined activities, and most actions (e.g. falling, drink water, sneeze, pickup) are designed for the potential application of health-care monitoring or surveillance in the future. Therefore, RGB-D videos are used for trimmed action recognition in this work.

With the recent development of deep learning, the wide adoption of deep models has resulted in remarkable performance improvement over traditional approaches on action recognition. Therefore, we build our action recognition approach upon the deep models.

It is noteworthy that most deep learning based works are focused on the design of deep architectures and very few works focus on the frame preprocessing stage (extraction and rescaling). For most deep models (2D or 3D), it is necessary to extract/sample a fixed number of frames from each trimmed video. The general method is uniform sampling of a fixed number of frames [8]. However, this approach may miss some frames that contain an important amount of motion. This may affect the performance as motion is the most important clue for action recognition. Another common method [6] is to keep all frames, and to split them into several fixed-length clips. For video-based prediction, the model averages the predictions over all clips and provides the final prediction for the input video. This method has some weaknesses as it breaks the completeness of an action and may lead to a hard representation of actions. For frame rescaling, the common approach is to crop the center area from original frames and resize to a fixed resolution [6, 8]. However,

the subjects involved in the action are not always in the center of frames and their location could change through time. Another method of frame rescaling is to directly resize original frames to a fixed resolution. Whereas, the subjects can appear very small in the video frame when they are far from the camera. In this case, it will be difficult to recognize the action.

In this thesis, we propose a simple, yet effective, novel action tube extractor that takes as input a trimmed video containing one specific action, and outputs an action tube (with fixed number of frames). Here, an action tube is defined as a sequence of cropped frames through the video that contain the subjects of a given action. Then, the action tube can be directly fed into the action recognition model. Our proposed action tube extractor can solve the problems mentioned above. For the action recognition model, we propose to use I3D as our 3D-CNN model. The proposed approach is illustrated in Fig. 1.

Our approach has been evaluated on two challenging datasets, Office Activity (OA) [10] and NTU RGB-D [11] datasets. Experimental results achieved are state-of-the-art.

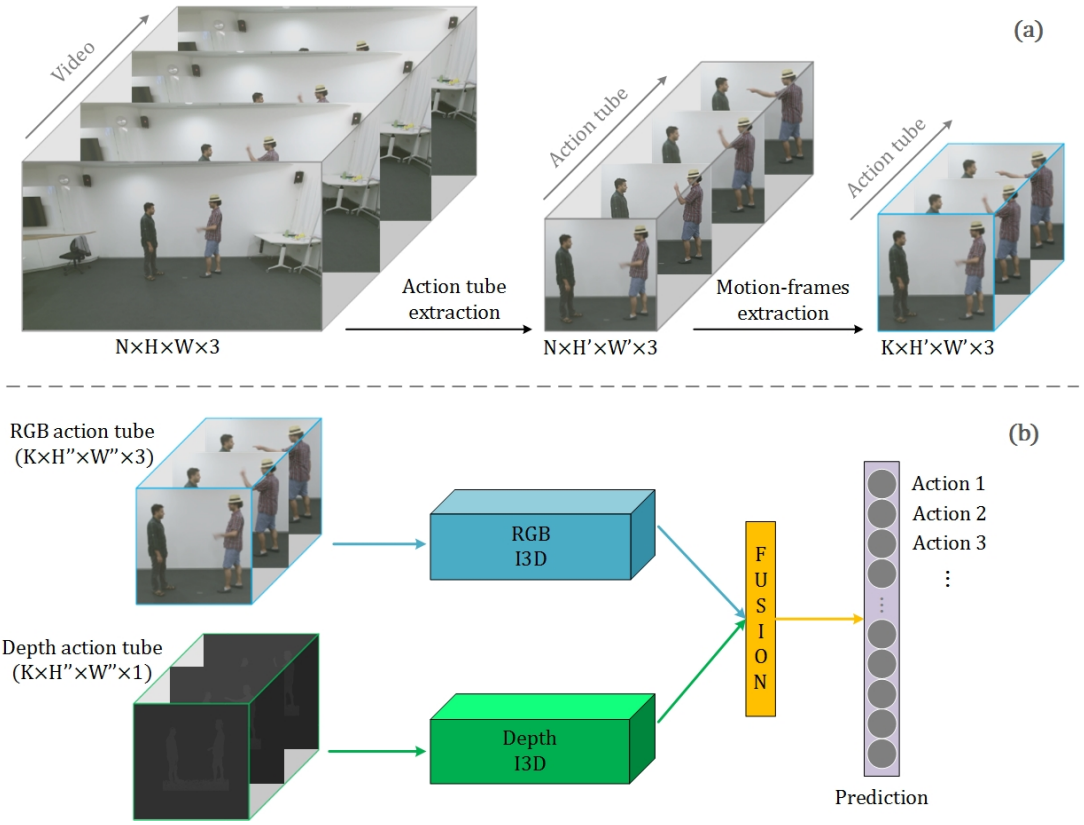


Figure 1: Illustration of our approach for RGB-D based human action recognition. (a) Given a trimmed video, we extract the action tube by human detection (MobileNet-SSD). Then, the structural similarity index (SSIM) is used to extract motion-frames (contains obvious motion) to form final action tube; (b) Extracted RGB-Depth action tubes are classified with two-stream I3D model. Late fusion is performed to combine RGB and Depth information.

1.2 Temporal Action Detection

Many of the existing action recognition schemes are devised for temporally trimmed videos. However, this is a significantly unrealistic assumption, since a real video often contains multiple action instances as well as irrelevant backgrounds. Therefore, it is more important and meaningful to do action recognition in untrimmed videos. Action recognition in untrimmed videos is called action detection/localization. It deals with the problem of identifying the exact spatio-temporal location where an action occurs. In this thesis, we just focus on temporal action detection.

In temporal action detection, we are given a long untrimmed video and aim to detect if and when a particular action takes place. Specifically, we answer three questions – "is there an action in the video?", "when does the action start and end?", and "what this action is?". These problems are very important because real applications usually involve long untrimmed videos, which can be highly unconstrained in space and time, and one video can contain multiple action instances plus background scenes or other activities. Compared to action recognition, it is more challenging, as it is expected to output not only the action category, but also the precise starting and ending time points.

In this work, we follow the "proposal + classification" framework to design an action detection system. Inspired by work in [33], we also apply a multi-scale segment generation scheme and a 3D ConvNet. We use I3D as the backbone net of the system. Instead of using uniform sampling as adopted in [33], we apply the SSIM based sampling method proposed by us for the action recognition part. The proposed temporal action detection system is illustrated in Fig. 2.

Due to the lack of time for this project and the time that took to train a single I3D network (around five days), we could not test the performance of the whole system. Therefore, we just use 32-frame length segments to obtain a baseline performance of the system. This simplified system is evaluated on a popular action detection dataset, THUMOS14 [39]. Experimental results demonstrate the effectiveness of our proposed system.

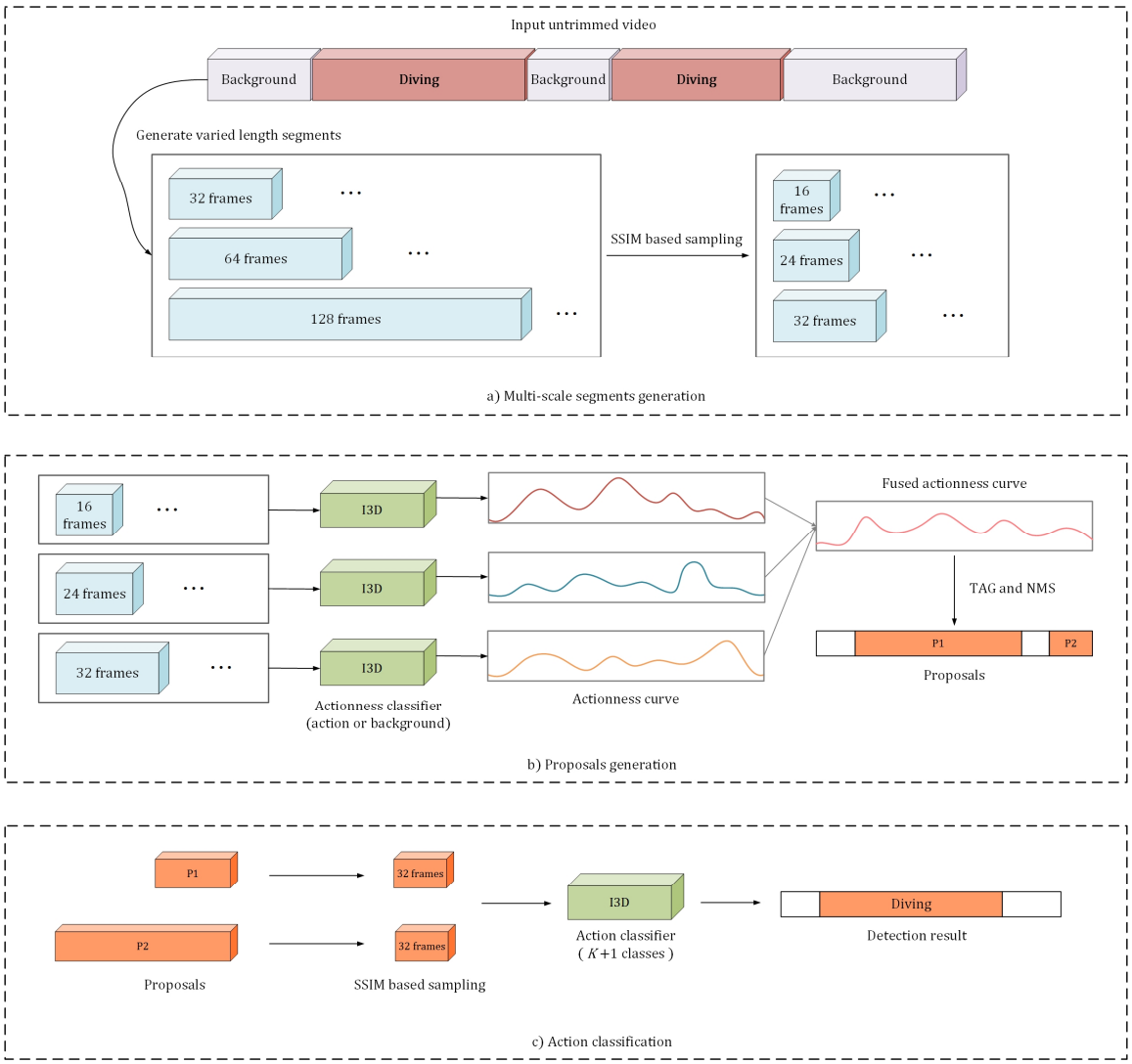


Figure 2: Illustration of the proposed system for temporal action detection. This system can be split into three parts: a) Multi-scale segments generation: given an untrimmed video, a set of varied length segments are generated. Then, these segments are sampled via SSIM method; b) Proposal generation: previous segments with different size are fed into three I3Ds. The generated three actionness curves are fused to form the final actionness curve. Then, temporal actionness grouping (TAG) and Non-maximal suppression (NMS) are applied to generate proposals; c) Action classification: each proposal is sampled to fixed number of frames via SSIM. Then, these sampled proposals are fed into I3D to output final detection result.

2 Related Works

In this section, several works related to action recognition and detection are reviewed. Considering that our proposed approaches rely on deep learning, we mainly list deep learning based works. The handcrafted features based works are also simply summarized in the following two subsections (Sect. 2.1 and Sect. 2.2).

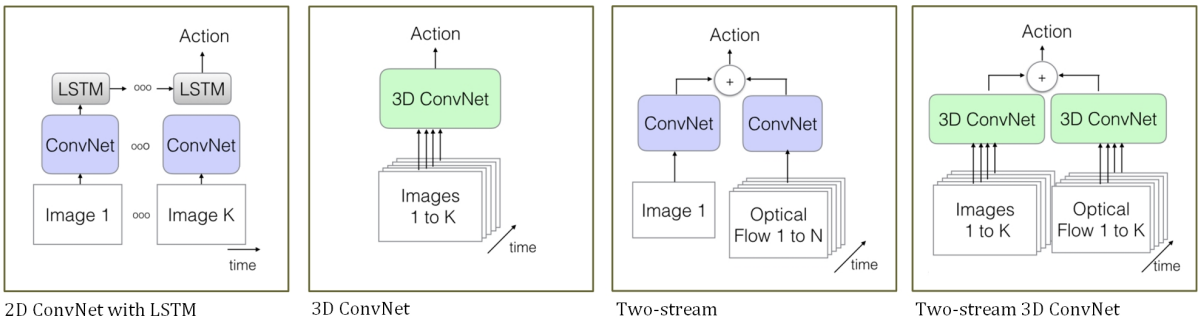
2.1 Deep Learning based Action Recognition

Traditional studies on (trimmed) action recognition use different kinds of methods [1, 2] to compute handcrafted features. Traditional handcrafted representation approaches can be split into two stages: 1) detectors which discover informative regions for action recognition; 2) descriptors which characterize the visual pattern of the detected regions. Among proposed handcrafted feature schemes for action recognition, dense trajectory (DT) [25] and improved dense trajectory (iDT) [26] have become very popular.

With the recent development of deep learning, a number of methods have been developed based on Convolutional Neural Networks (CNNs/ConvNets) [47] or Recurrent Neural Networks (RNNs) [47]. Unlike handcrafted approaches, deep learning based methods automatically learn features from raw data by utilizing a trainable feature extractor followed by a trainable classifier. In other words, deep learning is an end-to-end learning algorithm.

The wide adoption of ConvNets has resulted in remarkable performance improvement over traditional approaches on action recognition. These models used for action recognition can be categorized into four groups: 2D ConvNets, 3D ConvNets, two-stream networks and two-stream 3D ConvNets (see Fig 3, cited from [9]).

ConvNets were first introduced to this task in [39]. In this paper, ConvNets were first applied for video classification, where each video contains one specific activity. The idea of this paper is: using ConvNets (2D) to extract features independently from each frame then pooling their predictions across the whole video. This approach has an obvious drawback of ignoring temporal structure. Thus, there are some efforts to explore the long-range temporal structures via temporal pooling or RNNs [27, 28]. The architecture of these approaches can be visualized in Fig. 3 (first one). Later, 3D ConvNets [29, 30, 40] are proposed to deal with action recognition. 3D ConvNets seem like a natural approach to action modeling, and are just like standard convolutional networks, but with spatio-temporal filters. They have an important characteristic: they directly create hierarchical representations of spatio-temporal data. The 3D ConvNets models for action recognition is shown in Fig. 3 (second one). Two-stream architecture is another very practical approach, introduced by Simonyan and Zisserman [19]. This work firstly incorporated optical flow as additional input of CNNs for action recognition. It models short temporal snapshots of videos by averaging the predictions from a single RGB frame and a stack of 10 externally computed optical flow frames, after passing them through two ConvNets which were pre-trained on the ImageNet dataset. The two-stream networks have shown very high performance on many benchmarks, while being very efficient to train and test.



2D ConvNet with LSTM | 3D ConvNet | Two-stream | Two-stream 3D ConvNet

Figure 3: Four typical deep architectures for action recognition.

Fig 3 (third one) shows the basic architecture of two-stream networks. A recent extension [13] fuses the spatial and flow streams after the last network convolutional layer, showing some improvement over [19]. More recently, DeepMind proposed a new Two-Stream Inflated 3D ConvNet (I3D) [9]. The 3D ConvNet is inflated from the Inception architecture [31]. It replaces 2D convolutions with 3D convolutions. I3D achieves state-of-the-art performance on a wide range of video classification benchmarks. The Fig 3. (last one) shows the basic illustration of a two-stream 3D ConvNet.

The previously mentioned deep models for action recognition are proposed and tested on RGB data. However, these models are essentially suitable for RGB-D data. For action recognition, the only difference between RGB-D and RGB data is: how to effectively use the additional depth frames to obtain better recognition performance. A number of deep learning based approaches [3-6] are proposed for RGB-D based action recognition. These methods take as input either RGB, depth or both of them as independent streams and fuse the recognition scores of individual modalities. For RGB-D based action recognition, in addition to the design of deep architectures, another key point is the design of fusion method of RGB and depth information. According to our best knowledge, most RGB and depth fusion methods are based on hand-crafted features and tend to be dataset-dependent. Here, we summarize two kinds of deep learning based RGB+depth fusion approaches. The first one is from [41], they adopt a two-stream network and add depth stream and saliency stream to form a multi-stream network. The final score is fused from these streams. The second one is from [8], they propose to encode the depth and RGB video into structured dynamic images, and exploit the conjoint information of the heterogeneous modalities using one ConvNet. This approach achieves state-of-the-art performance on the NTU RGB-D dataset, which is the largest RGB-D action recognition dataset.

2.2 Deep Learning based Action Detection

Action detection/localization aims to predict where an action begins and ends in the untrimmed videos. The advances in ConvNets have led to remarkable progress in video analysis. Notably, the accuracy of action recognition has been significantly improved. However, the performances of action detection methods remain unsatisfactory. Existing

state-of-the-art approaches address this task as detection by classification, i.e. classifying temporal segments generated in the form of sliding windows [33, 34] or by an external proposal generation mechanism [32]. These methods can be divided into two categories: handcrafted representation and learning-based features. Among handcrafted representation approaches, improved Dense Trajectory (iDT) with Fisher Vector based methods [38] achieved best performance. For learning-based methods [33, 36, 37], most of them adopt the "proposal + classification" scheme in modern object detection architectures like Fast R-CNN [35]. Within this paradigm, a video is first processed to produce a set of candidate video segments or proposals, which are likely to contain a human action. These proposals are then used as a reduced candidate set, on which action classifiers can be applied for recognition. Therefore, high-quality temporal proposals are crucial following this framework. A promising temporal proposal candidate in action detection should contain the action of interest in accordance with high Intersection-over-Union (IoU) overlap with the groundtruth. In addition, the proposal generation algorithm should be robust enough to find candidates for any action or activity class, and simultaneously provide potential starting and ending times for each candidate action. The large variation in motion, scenes, and objects involved, styles of execution, camera viewpoints, camera motion, background clutter and occlusions impose additional burden to the proposal generation process.

Shou et al. propose a Segment-CNN (S-CNN) proposal network and address temporal action detection by using 3D ConvNets (C3D) features, which involve two stages, namely proposal network and localization network [33]. S-CNN is also the first work of proposing this two stage framework. Xu et al. [36] introduce the region C3D (R-C3D) model, which encodes the video streams using C3D model, then generates candidate temporal regions containing actions, and finally classifies selected regions into specific action. Zhao et al. [42] propose structured segment network (SSN) to model activities via structured temporal pyramid. On top of the pyramid, a decomposed discriminative model comprising two classifiers is introduced, respectively for classifying actions and determining completeness. Qiu et al. [37] propose a three-phase action detection framework, which is embedded with an Actionness Network to generate initial proposals through frame-wise similarity grouping, and then a Refinement Network to conduct boundary adjustment on these proposals. Finally, the refined proposals are sent to a Localization Network for further fine-grained location regression.

Although many state-of-the-art methods adopt the "proposal + classification" framework, this framework has drawbacks. The main drawback of this framework is that the boundaries of action instance proposals have been fixed during the classification step. Lin et al. [43] propose a Single Shot Action Detector (SSAD) network to address this issue. SSAD is based on 1D temporal convolutional layers to skip the proposal generation step via directly detecting action instances in untrimmed video.

3 Proposed Approaches

In this section, we describe the details of our proposed approaches for trimmed action recognition and temporal action detection.

3.1 Action Recognition

The approach presented here consists of two parts, as illustrated in Fig. 1. The first part is our action tube extractor. It takes as input a trimmed video (a sequence of N frames containing one specific action) and outputs an action tube. The second part is a RGB-D two-stream network. The inputs of the network are extracted action tubes using the method proposed in the first part. We propose to use the I3D architecture to model temporal context. It is designed based on the Inception architecture, but replaces 2D convolutions with 3D convolutions. Temporal information is kept throughout the network. At test time, late fusion [13] is applied to combine RGB and depth information. In this section, we first describe our proposed action tube extractor (Sect. 3.1-A), and then the two-stream I3D for action recognition (Sect. 3.1-B).

A. Action tube extractor

Our action tube extractor involves two steps (see Fig. 1-a). The first step performs the action tube extraction. An action tube is defined as a sequence of cropped frames through the video that contain the subjects of a given action. As the actions of interest are related to humans, in order to achieve this goal, human detection is applied on each frame to generate the action tube. The action tube extraction has two benefits: 1) removing most useless background information; 2) increasing the area of region-of-interest (subjects). The second part is designed to perform a temporal sampling of the video sequence to remove frames without obvious motion. This way, the video can be sampled using a fixed number of frames, as needed by the I3D model. Finally, we get an action tube with a fixed number (K) of frames. In the following we describe the action tube extraction and temporal sampling in detail.

Action tube extraction: The method is illustrated in Fig. 4. Considering efficiency, we propose to use MobileNet-SSD [14] as the human detection algorithm, which is a fast detection deep model. This model is pre-trained on VOC0712 (2007+2012) [15] dataset. For each input frame, if there is more than one person, the network outputs more than one bounding box. Here, we make a slight modification to output only one bounding box for each frame (see Fig. 4, second column). The final bounding box contains all the detected persons. Finally, we get N bounding boxes from N frames. As in some frames MobileNet-SSD can fail to detect humans, we use the detected bounding boxes of adjacent frames. Then, we generate a bounding box (Fig. 4, third column, black dashed box) that contains these Nbounding boxes. Finally, the expanded bounding box (Fig. 4, third column, black solid box) is applied on each video frame to generate the action tube.

Motion-frames extraction: Motion is the most important information for action

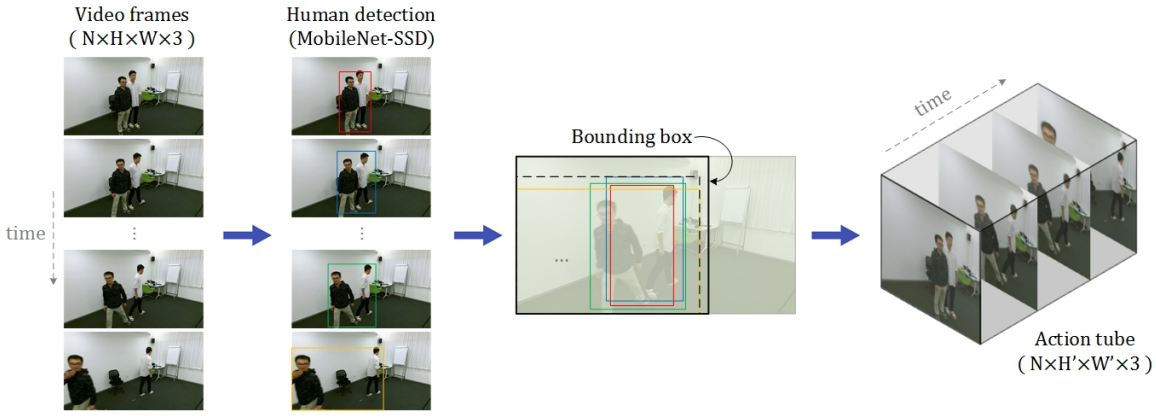


Figure 4: Overview of action tube extraction. Pre-trained MobileNet-SSD is performed to detect subjects in each video frame. The final bounding box (black solid box) is applied on every frame to form the action tube.

recognition. However, in most cases there are lots of similar frames (without motion change) in the extracted action tube. Thus, it is crucial to extract frames with obvious motion change. In order to extract those frames, we propose to use structural similarity index (SSIM) [16], which can be applied to measure the similarity between two consecutive frames. We choose SSIM as it can be computed very efficiently and when applied to successive frames gives a good indication of the amount of motion. Fig. 5 shows some examples. We can see that frames without motion have higher SSIM value (high similarity) than frames with obvious motion. In other words, lower SSIM value indicates the frames with more obvious motion. The motion-frames extraction is illustrated in Fig. 6. The first frame is always kept, and the other $K-1$ frames are extracted according to the SSIM values. The SSIM is calculated from every two consecutive frames. The extraction is performed in two steps: local extraction and global extraction. For the local extraction step, we extract one frame with the lowest SSIM value from every 16 frames. For the global extraction step, we extract first $K-1-N_{loc}$ frames with lowest SSIM values, where N_{loc} indicates the number of locally extracted frames. For mostly simple actions (see Fig. 6. falling), global extraction is enough. However, for some complex actions (i.e. actions that can be divided into several sub-actions, see Fig. 6. sleeping), local extraction is necessary because in some cases motion could mainly occur in one of the sub-actions so the remaining ones would not be represented in the final sampling. Our method combines a sort of uniform sampling with more detailed sampling where motion is present. At the end of this process, we obtain an action tube with only K frames.

B. Two-stream I3D

In [17], deep architectures used for action recognition are categorized in four groups: 2D models, motion-based input features, 3D models and temporal networks. In the first group, [18] uses a pre-trained model on one or more frames which are sampled from the whole

video. Then, the entire video is labeled by averaging the result of the sampled frames. To consider temporal information, in the second group, [19] and [20] compute 2D motion features like optical flow. Afterwards, these features are exploited as additional input streams of a 2D network to form two-stream network. The third group introduces 3D filters in the convolutional and pooling layers to learn discriminative features along both spatial and temporal dimensions [9, 21]. The input data of these networks are a fixed length sequence of frames. Finally in the fourth category, Recurrent Neural Networks (RNN) and variations [5, 11] are utilized to process temporal information. Among previous methods, two-stream (RGB frame and optical flow frames) 2D-CNN architecture achieved state-of-the-art results on many RGB datasets. More recently, Carreira and Zisserman proposed I3D architecture [9], and this model achieves state-of-the-art performance on a wide range of video classification benchmarks. Therefore, I3D has been selected in this work to be extended and analyzed for RGB-D data. Considering that the calculation of optical flow is very expensive, it is not adopted as additional input in our model. In this thesis, only two modalities (RGB and depth) are used as the input data for I3D to form the two-stream I3D architecture (see Fig. 1-b). The detailed architecture of the backbone net (inflated from 2D Inception-V1 architecture) of I3D is shown in Fig. 7 (cited from [9]). In trimmed activity recognition, the length of video is usually less than 10 seconds. As I3D needs a fixed number of frames as the input, we set the frame number $K = 32$. Many approaches [44, 45] have demonstrated that late fusion of both RGB and depth modalities is effective for action recognition. Therefore, we adopt late fusion as the fusion strategy in this work. For late fusion, we average scores from the RGB and depth streams.

Figure 5: Examples of SSIM value of consecutive frames. Frames with obvious motion have lower values than those with no motion.

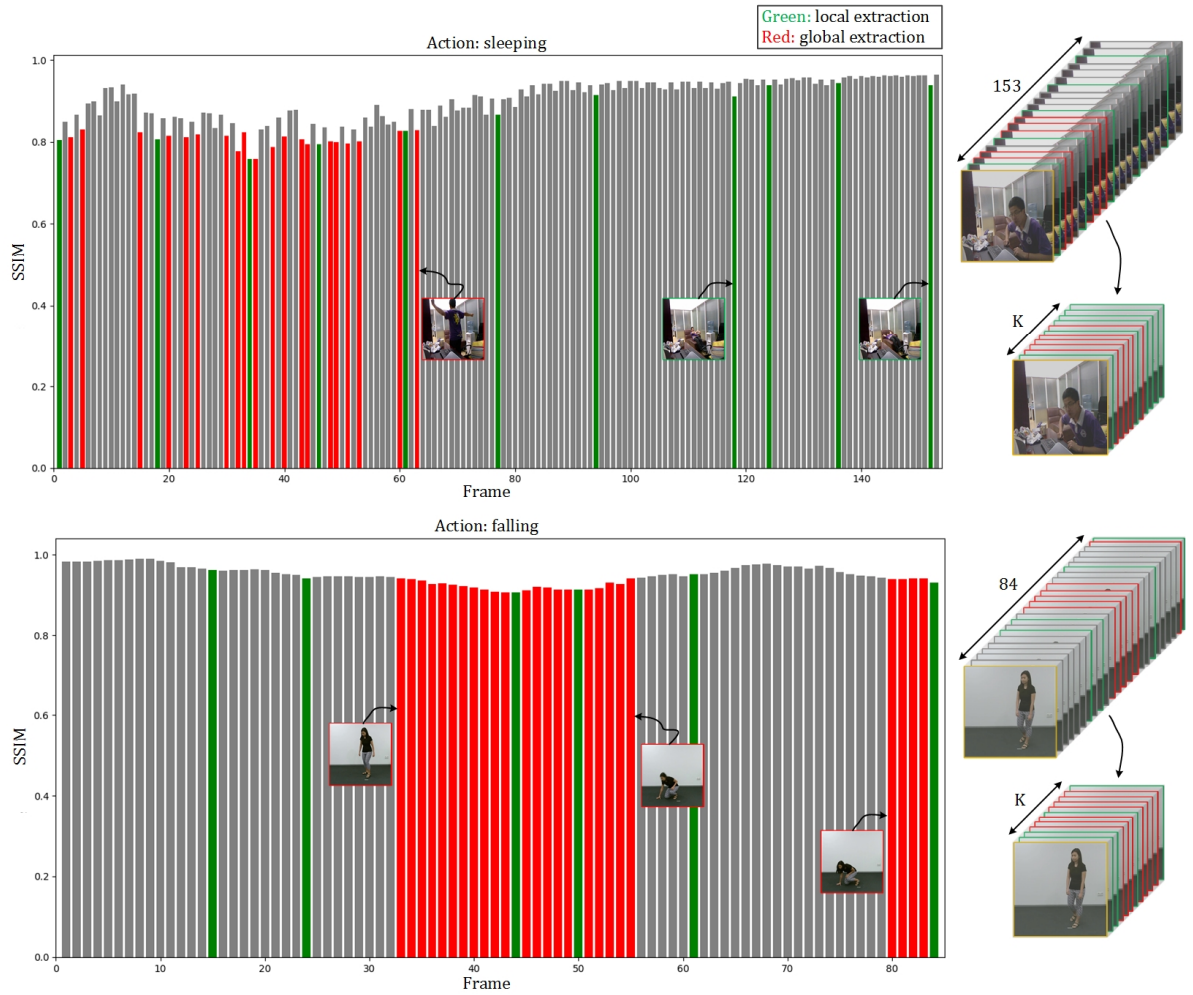


Figure 6: Overview of motion-frames extraction. The bar plot represents SSIM values of every two consecutive frames. Green frames are locally selected frames, red frames are globally selected. The K extracted frames consists of green frames, red frames and the first frame. (here, $K = 32$)

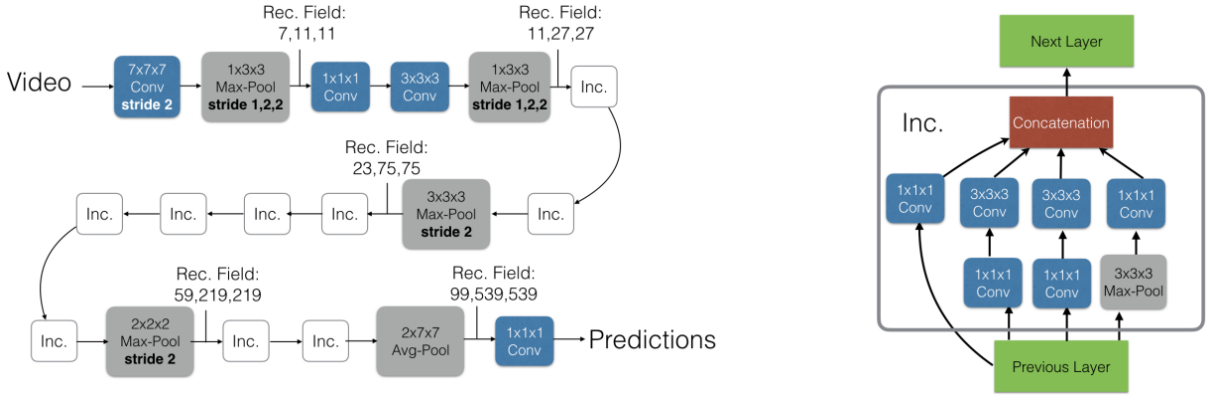


Figure 7: The backbone architecture (left) of I3D and its detailed inception submodule (right). The predictions are obtained convolutionally in time and averaged.

3.2 Action Detection

The duration of this project is limited to five months and the complexity of the work, we could successfully complete the action recognition part and we have some time to start working on the problem of action detection. Thus, we could not make a deep research on this topic. Our main efforts were focused on the action recognition part, which has been introduced in previous section. In this thesis, our proposed temporal action detection system follows the "proposal + classification" framework. Inspired by S-CNN, we adopt a multi-scale segment generation scheme and use 3D ConvNet as the basic network. In this work, we also use I3D architecture as the backbone 3D ConvNet of the action detection system. In this section, we first describe the multi-scale segment generation part (Sect. 3.2-A), and then the proposal generation part (Sect. 3.2-B). Finally, the action classification part (Sect. 3.2-C).

A. Multi-scale segment generation

Given an untrimmed video with N frames, we conduct temporal sliding windows of varied lengths as 32, 64, 128 frames with 75% overlap. For these three varied length windows, we construct segment S by sampling 16, 24, 32 frames, respectively. We adopt the SSIM based sampling method proposed by us in action recognition part (see Sect. 3.1-A). Consequently, for each untrimmed video, we generate three sets of candidates as input for proposal generation network. The process of this part is shown in Fig. 2-a. The number of each set of candidates can be calculated as follows:

$$N_{candi} = \frac{N+M-L}{SS} + 1 \tag{1}$$

where, L represents window length (32, 64 or 128); SS represents the stride size (8, 16 and 32); M is an integer ($0 \leq M < L$), it indicates the repetition of last frame of the given video. Here, we make an explanation of parameter M. Because the length (N) of videos is arbitrary, the number of frames could be less than L in the last window. In this case, we repeat the last frame to ensure there are L frames in the last window.

B. Proposal generation

Three sets of candidates that generated from last stage are fed into three I3D networks to output three actionness curves. In this stage, I3D network plays the role of binary actionness classifier to distinguish whether a candidate (snippet) contains human actions. For each segment the provided ground truth label (action classes and background) is converted into binary action/background labels. Accordingly, an actionness curve can be generated by accumulating all the actionness probabilities of snippets. As shown in Fig. 2-b, we get three actionness curves (one for each segment length or scale level) via training three I3D networks. Then, these three actionness curves are fused to generate final actionness curve. In this work, we average actionness probabilities of these three curves as the fusion method.

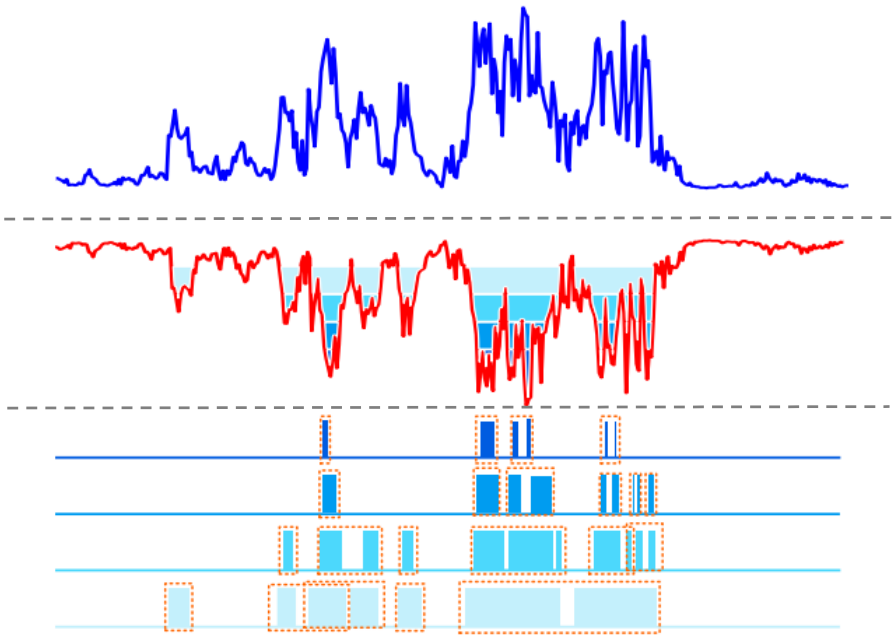


Figure 8. Visualization of the TAG process for proposal generation. Top: Actionness probabilities curve. Middle: The complement curve. It is flooded with different thresholds (water levels) γ. Bottom: Regions obtained by different water levels. By merging the regions according to the grouping criterion, we get the final set of proposals (in orange color).

Then, we adopt Temporal Actionness Grouping (TAG) method in [42] to generate temporal action proposals. Given an actionness curve, the classic watershed algorithm [49] with multiple thresholds γ is utilized to produce a set of "basins" corresponding to the temporal region with high actionness probability. Then, the TAG scheme is applied to connect small basins, resulting in proposals. The TAG works as follows (illustrated in Fig. 8, cited from [42]): it begins with a seed basin, and consecutively absorbs the basins that follow, until the fraction of the basin durations over the total duration drops below a certain threshold τ. The absorbed basins and the blank spaces between them are then grouped to form a single proposal. The values of γ and τ are uniformly sample from $\in (0, 1)$ with an even step of 0.1. The combination of these two thresholds leads to multiple sets of proposals. We then take the union of them. Finally, the highly overlapped proposals are filtered out via Non-maximal suppression (NMS) with Intersection-over-Union (IoU) threshold 0.95. Fig. 2-c shows the illustration of this part.

C. Action classification

After the generation of proposals, we train an action classification model (I3D) for K action categories as well as background. As I3D needs a fixed number of frames as the input, we adopt SSIM based sampling method to sample each proposal to 32 frames. For proposals containing less than 32 frames, we repeatedly add the last frame to the end of these proposals to ensure that they contain 32 frames. The output of this part will be the final action detection result.

4 Experiments

In this section, we evaluate the effectiveness of our proposed approaches on several challenging action recognition and detection benchmarks. The datasets are first introduced in this section, then the setups and parameter settings for the experiments are illustrated. We compare the results of the proposed models with the current best methods.

4.1 Datasets

A. Action recognition

NTU RGB+D dataset. To our best knowledge, it is currently the largest action recognition dataset in terms of training samples for each action. The dataset consists of 56,880 action videos and 4 million frames, which were collected by 3 Kinect V2 cameras from 40 distinct subjects, and divided into 60 different action classes including 40 daily (drinking, eating, reading, etc.), 9 health-related (sneezing, staggering, falling down, etc.), and 11 mutual (punching, kicking, hugging, etc.) actions. It has four major data modalities provided by the Kinect sensor: 3D coordinates of 25 joints for each person (skeleton), RGB frames, depth frames, and IR sequences. In this paper, we only use the RGB and depth frames. The large intra-class and view point variations make this dataset challenging. However, the large amount of action samples makes it highly suitable for data-driven methods. Fig. 9 shows some sample frames of this dataset.

This dataset has two standard evaluation criteria [11]. The first one is a cross-subject test, in which half of the subjects are used for training and the other half are used for testing. The second one is a cross-view test, in which two viewpoints are used for training and one is excluded for evaluation. According to previous works [3, 8, 11, 22], cross-subject is harder than cross-view. Therefore, we only focus on the cross-subject evaluation in this work. In the cross-subject evaluation, samples of subjects 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35 and 38 were used as training and samples of the remaining subjects were reserved for testing.

OA dataset. It covers the regular daily activities taken place in offices. The dataset consists of 1,180 sequences, containing 20 classes of activities performed by 10 subjects. Specifically, it is divided into two subsets, each of which contains 10 classes of activities: OA1 (complex activities by a single subject) and OA2 (complex interactions by two subjects). For fair comparison and evaluation, we follow the same protocol, and thus 5-fold cross validation is adopted by ensuring that the subjects in training set are different with those in testing set. This dataset consists of multiple camera views of same action. The high complexity of background clutter and occlusion makes this dataset challenging. Several sample frames of OA dataset are shown in Fig. 10.

Figure 9: Sample frames of the NTU RGB+D dataset. The last row illustrates RGB, RGB+joints, depth, depth+joints, and IR modalities of a sample frame.

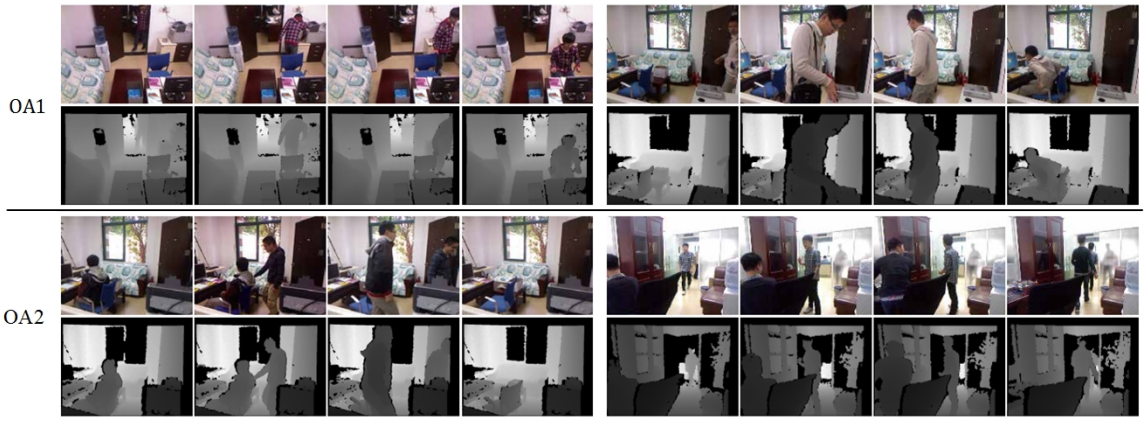


Figure 10: Sample frames (RGB and depth) of the OA dataset.

B. Action detection

THUMOS14. It contains 1010 untrimmed videos for validation and 1574 untrimmed videos for testing. This dataset does not provide the training set by itself. Instead, the UCF101 [46], a trimmed video dataset is appointed as the official training set. Following the standard practice, we train our models on the validation set and evaluate them on the testing set. On these two sets, 220 and 212 videos have temporal annotations in 20 classes, respectively. Two falsely annotated videos ("video_test_0000270", "video_test_0001496") in the testing set are excluded in evaluation.

4.2 Experimental Settings

A. Implementation details.

RGB-D based action recognition: the frame resolution of extracted action tube is resized to 300 × 300 for both datasets. For the RGB stream, the I3D networks are initialized with Kinetics [23] pre-trained models. Considering that the OA dataset contains only 1,180 videos, we adopted data augmentation. Concretely, we applied random left-right frame flipping consistently for each video during training. For very short videos ($N < 32$, where N is the number of frames), we looped the last frame 32-N times without motion-frames extraction.

Temporal action detection: the frame resolution is kept the same as that of original videos (320 × 180). All I3D networks used in the detection system are initialized with Kinetics pre-trained models. In part (b) and (c) of our proposed detection system (see Fig. 2), we need to label each segment (0 or 1) and proposal (0 ~ 20) for training. We use the following strategy: if its Intersection-over-Union (IoU) with ground truth is larger than 0.75, we assign a positive or a specific class; otherwise, we set it as the background. Furthermore, in part (b) and (c), the labeled instances (segments or proposals) are unbalanced (instances with negative label are larger than positive instances). In order to balance the number of training data for each class, we use the training set of THUMOS14 (UCF101) to produce additional positive instances. Due to the lack of time for this project and the time that took to train a single I3D network was around five days, we could not complete the training of the whole system. Therefore, we just use 32-frame length segments to obtain a baseline performance of the system.

B. Evaluation metrics.

On RGB-D dataset, we adopt cross-subject test, in which half of the subjects are used for training and the other half are used for testing. The evaluation metric used is classification accuracy.

On OA dataset, we follow the same protocol used in previous papers. Thus, 5-fold cross validation is adopted for both subsets (OA1 and OA2) by ensuring that the subjects in training set are different with those in testing set. The evaluation metric used is also classification accuracy.

On THUMOS dataset, we follow the conventional metrics to regard temporal action detection as a retrieval problem, and evaluate mean average precision (mAP) at 0.5 IoU threshold.

4.3 Comparison to the state-of-the-art

A. Action recognition

We compare our proposed approach to some state-of-the-art results on two challenging

datasets.

OA dataset. On this dataset, we apply our method on the two OA subsets. As shown in Table 1, our model performance is much better than the state-of-the-art method on both subsets, with improvements in accuracy larger than 20%. We see that using a combination of RGB and depth outperforms the individual modalities, as was expected. From the results, we can conclude that visual recognition of actions (interactions) by two subjects (OA2) is harder than recognition of actions by a single subject (OA1). In most cases, interactions by two subjects are more abstract/complex than actions performed by a single subject.

NTU RGB-D dataset. Table 2 lists the performance of the proposed method and previous works. The proposed method has been compared with some state-of-the-art skeleton-based, depth-based and RGB+Depth based methods that were previously reported on this dataset. We can see that the proposed method outperforms all these previous approaches.

Detailed results, including per class accuracies can be found in the Additional Material document [24], or you can see them in the Appendix.

OA1			
Method	RGB	Depth	RGB + Depth
R-SVM-LCNN [10] (2016)	60.4 %	65.2 %	69.3 %
Ours	**87.7 %**	**84.8 %**	**91.9 %**
OA2			
Method	RGB	Depth	RGB + Depth
R-SVM-LCNN [10] (2016)	46.3 %	51.1 %	54.5 %
Ours	**77.5 %**	**72.8 %**	**82.2 %**

Table 1: Comparison of the proposed method with state-of-the-art approach on OA dataset (OA1, OA2).

Method	Skeleton	RGB	Depth	RGB + Depth
SSSCA-SSLM [7] (2017)	-	-	-	74.86 %
HCN [22] (2018)	86.50 %	-	-	-
c-ConvNet [8] (2018)	-	-	-	86.42 %
D-CNN [3] (2018)	-	-	**87.08 %**	-
Ours	-	**91.95 %**	86.02 %	**93.56 %**

Table 2: Comparative accuracies of the proposed method and state-of-the-art methods on NTU RGB-D dataset (cross-subject evaluation). " - " indicates the result is not available.

B. Action detection

On THUMOS14, we compare our system with the recent state-of-the-art approaches. From Table 3 we can see that the performance of our system is higher than S-CNN, but worse than other three works. Although our system's performance is not very good, this result is

satisfactory. As this result is just the baseline performance of our whole system, without the multi-scale approach. Therefore, this result can be a demonstration of the effectiveness of our proposed action detection system. In Fig. 11, we show example detections on the THUMOS14 test set.

Year	2016	2017	2017	2018	2018
Method	S-CNN [33]	R-C3D [36]	SSN [42]	ETP [37]	Ours
mAP@0.5	19.0	28.9	29.8	**34.2**	20.5

Table 3: Comparison of the proposed method with state-of-the-art approaches on THUMOS14, measured by mAP at IoU thresholds 0.5.

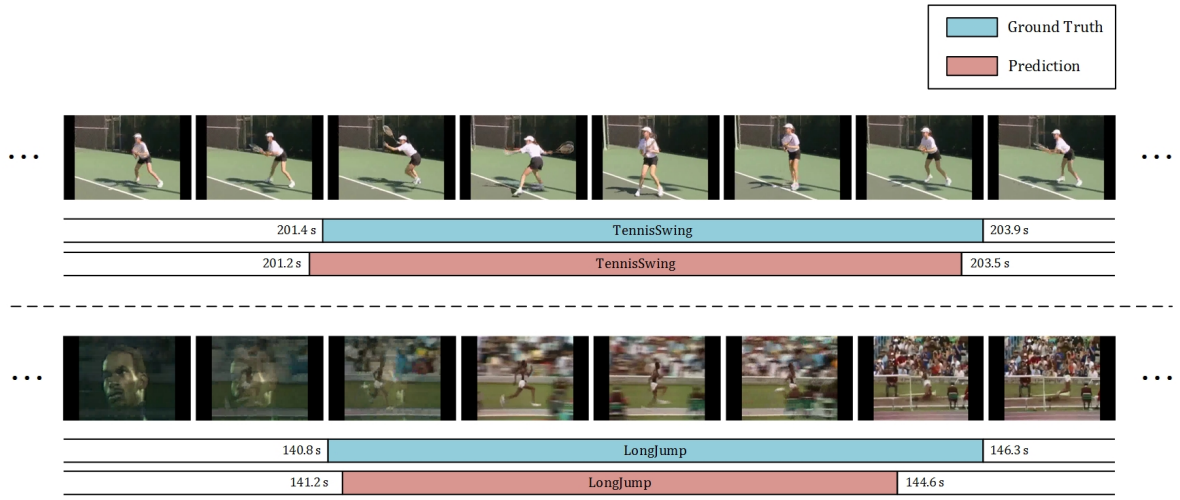


Figure 11: Example detections from our system.

4.4 Discussion

To better analyze the performance of the proposed model for action recognition, we take a closer look at actions that are highly confusing to the two-stream I3D structure (Fig. 13 shows the confusion matrices for the NTU RGB-D and OA datasets). As presented in Fig. 14, such action pairs include reading vs. writing, nod head/bow vs. pickup, nausea or vomiting condition vs. nod head/bow, showing object vs. shaking hands, chatting vs. chatting and eating, and arranging files vs. looking for objects. From these samples, we can observe that these misclassified actions are inherently confusing. In order to deal with such actions, we may need to obtain fine-grained motion information. This will be our future work.

In order to further demonstrate the effectiveness of the action tube extractor, we compare the results of our method against a similar system where the action tube extractor has been replaced by a more traditional approach consisting in cropping the center region and using uniform sampling (illustrated in Fig. 12). A region of size $H \times H$ is cropped from the original frame, where H is the size of shorter side of frame. The extracted

frames are resized to 300 × 300 pixels. Finally, these resized frames are fed into I3D model. For this test, we used only the RGB modality for simplicity. The comparisons are shown in Table 4. We can see that our proposed action tube extractor provides an improvement in accuracy around 3% on both OA and NTU RGB-D datasets. This is a strong demonstration of the effectiveness of our proposed action tube extractor.

Dataset	with ATE	w/o ATE
NTU RGB-D	**91.95 %**	89.29 %
OA1	**87.7 %**	84.2 %
OA2	**77.5 %**	73.9 %

Table 4: Comparison of performance with and without action tube extractor (ATE) on NTU RGB-D and OA datasets. (RGB modality)

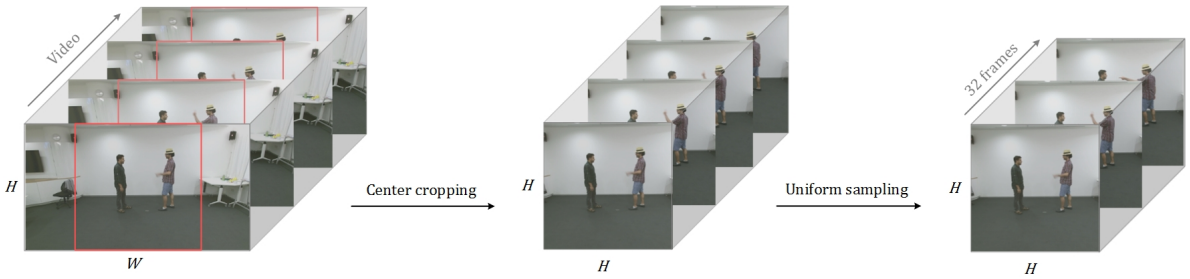
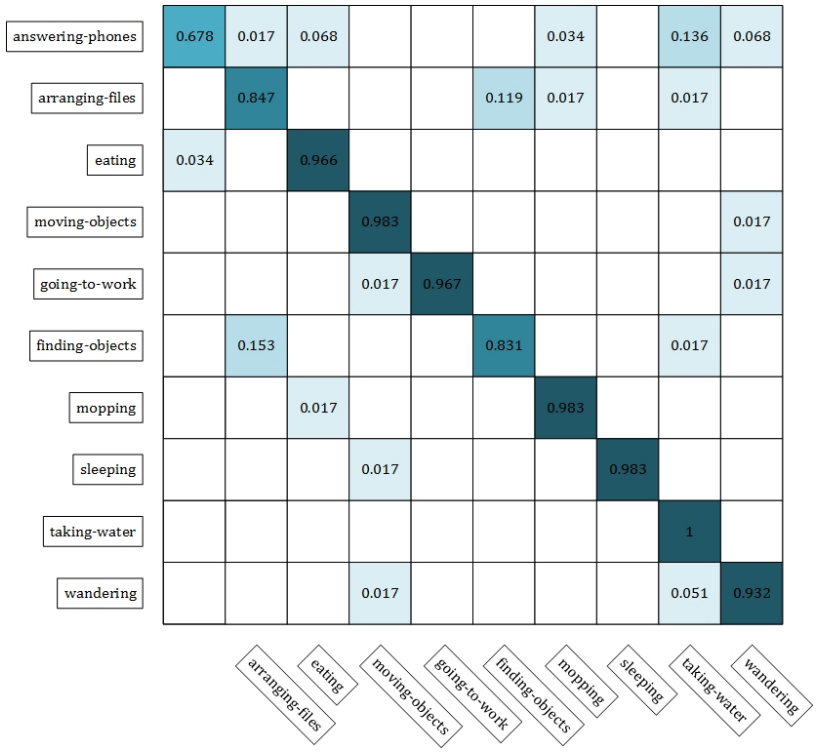
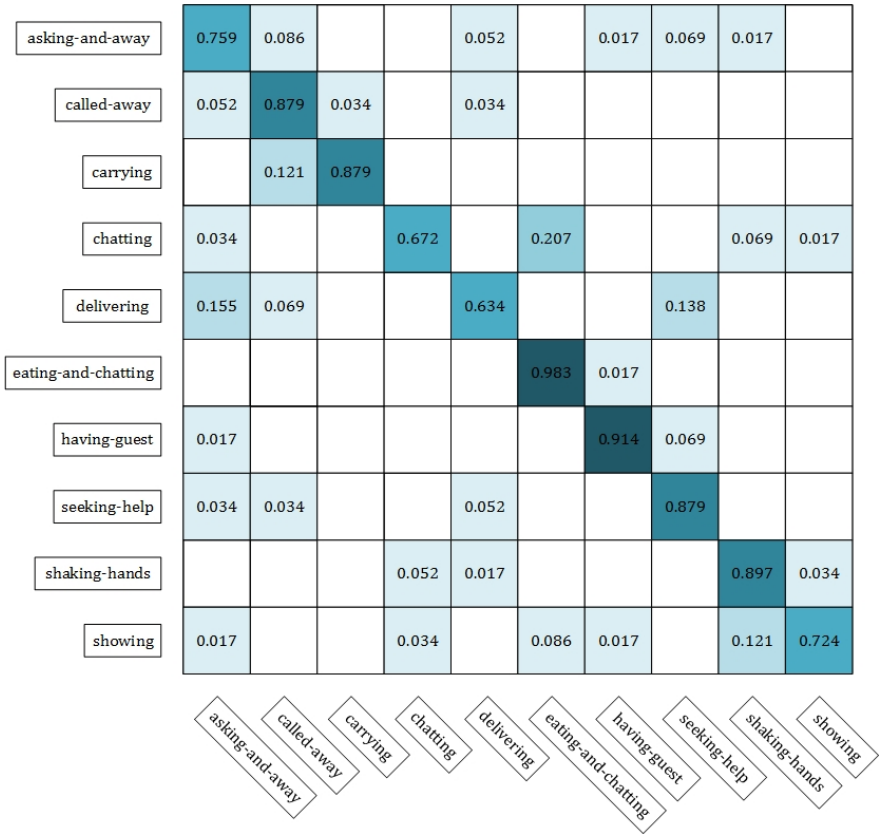


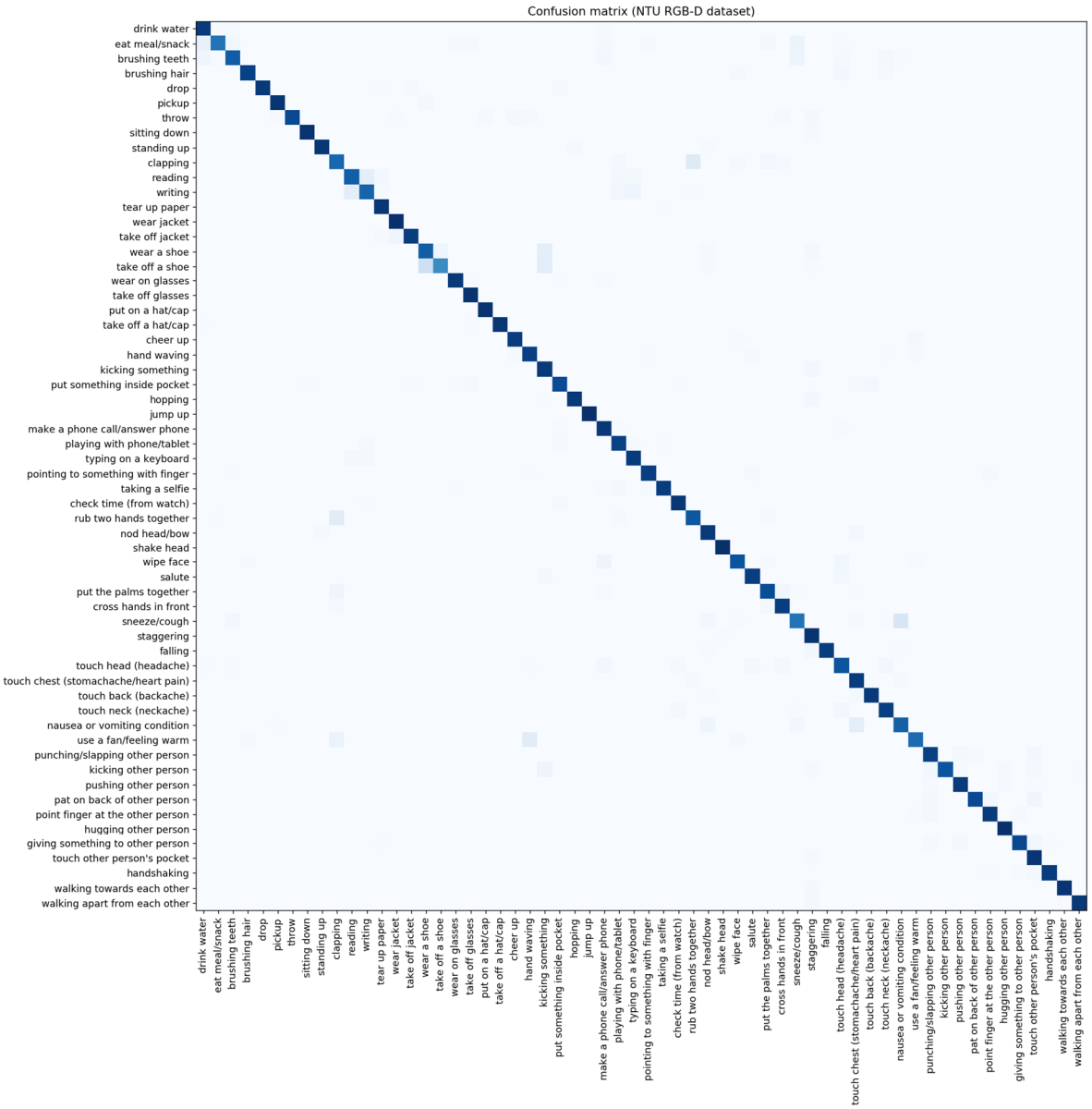
Figure 12: The replacement of our proposed action tube extractor (ATE).



a) Confusion matrix of OA1 dataset



b) Confusion matrix of OA2 dataset



c) Confusion matrix of NTU RGB-D dataset

Figure 13: Confusion matrices of NTU RGB-D and OA datasets.

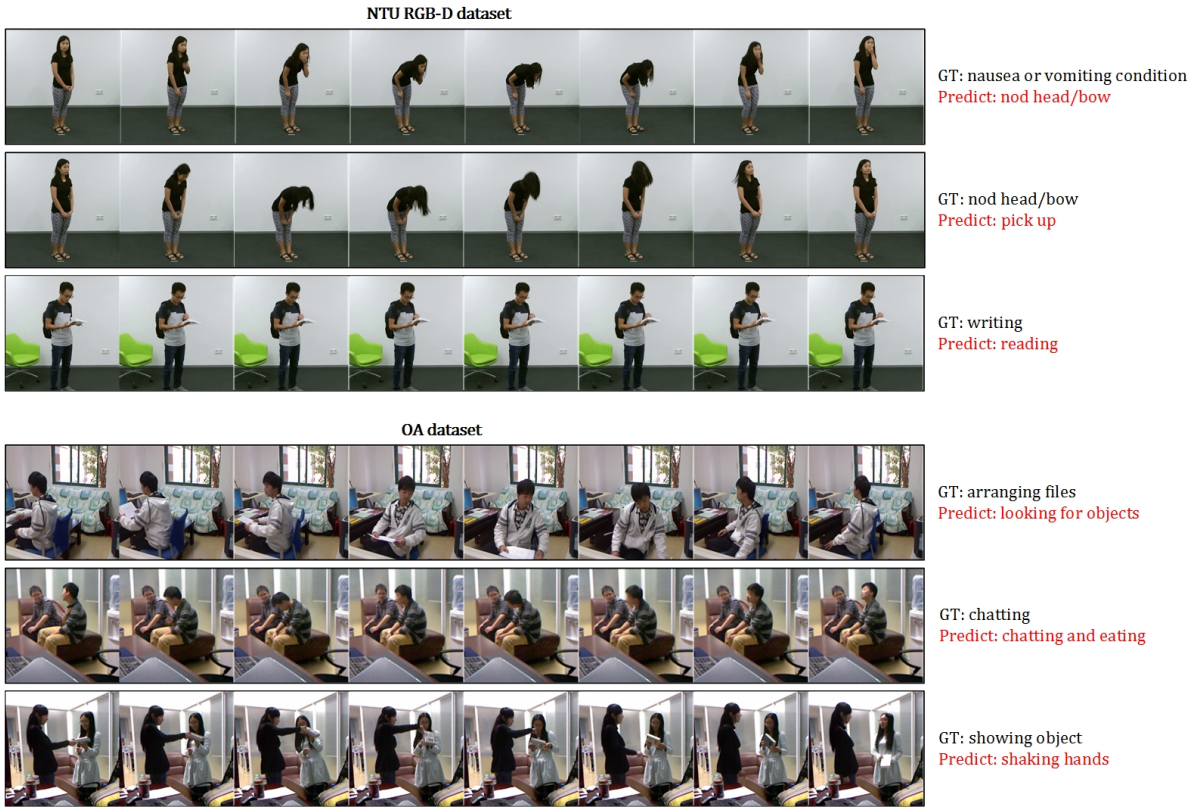


Figure 14: Some incorrect action recognition results on the test set of OA and NTU RGB-D datasets.

5 Conclusions and Future Work

5.1 Conclusions

In this project, we firstly introduced the problem of action recognition in videos, which has many potential applications (e.g. intelligent surveillance, robotics, health-care monitoring, and interactive gaming) in our daily life. Then, we made detailed descriptions of related works on two main topics: action recognition in trimmed videos and temporal action recognition in untrimmed videos.

One of the main contributions of our work is to propose a novel action tube extractor for 3D action recognition. It takes as input a trimmed video and outputs an action tube. The action tube contains much less background information, and has higher ratio of ROI (subjects) to background. Besides, every frame of the extracted action tube contains obvious motion change. Then the extracted RGB/Depth action tubes are directly fed into two-stream I3D model. An extensive experimental analysis shows the benefits of our proposed approach, which achieves state-of-the-art results on both OA and NTU RGB-D datasets. This work has also been submitted to the International Conference on Content-Based Multimedia Indexing (CBMI 2018) and us currently under review.

Considering that the action recognition in trimmed videos is a significantly unrealistic

assumption, since a real video often contains multiple action instances as well as irrelevant backgrounds. We extended our work to temporal action detection. A new action detection system is proposed. It contains three main stages: 1) multi-scale segment generation; 2) proposal generation; 3) action classification. Due to the lack of time for this project, we just trained and tested a simplified system. The experimental result demonstrates the effectiveness of our proposed action detection system.

5.2 Future Work

The tests of whole action detection system have been left for the future. Thus, temporal action detection will be our main topic in the future. The first work will be the training and testing of our detection system. Then, we will make a deeper analysis of the system to improve its performance. Actually, this system has an obvious drawback: it cannot obtain the precise boundary of actions (e.g. a detected action segment could contain small background fragments). For each proposal generated from stage two of the system, the classification part assigns it an action class or background. However, a proposal could contain both action and background fragments. Therefore, the action classification part of the system will be redesigned and improved to obtain better localization performance in the future.

Besides, as discussed in Sect. 4.4, our proposed approach is still confused by actions (e.g. reading vs. writing, showing object vs. shaking hands, chatting vs. chatting and eating) with similar motion. This result indicates the motion representation ability of our approach is still weak. How to extract better motion information for action recognition is also the main focus of our future work.

References

[1]. G. Evangelidis, G. Singh, and R. Horaud, “Skeletal Quads: Human Action Recognition Using Joint Quadruples,” in 2014 22nd International Conference on Pattern Recognition, 2014, pp. 4513–4518.

[2]. C. van Gemeren, R. T. Tan, R. Poppe, and R. C. Veltkamp, “Dyadic interaction detection from pose and flow,” in Human Behavior Understanding, H. S. Park, A. A. Salah, Y. J. Lee, L.-P. Morency, Y. Sheikh, and R. Cucchiara, Eds. Springer International Publishing, 2014, pp. 101–115.

[3]. P. Wang, W. Li, Z. Gao, C. Tang, and P. Ogunbona, “Depth Pooling Based Large-scale 3D Action Recognition with Convolutional Neural Networks,” IEEE Transactions on Multimedia, vol. 20, no. 5, pp. 1051– 1061, 2018.

[4]. P. Wang, W. Li, Z. Gao, and P. O. Ogunbona, “Action Recognition From Depth Maps Using Deep Convolutional Neural Networks,” IEEE Transactions on Human-Machine Systems, Vol. 46, No. 4, vol. 46, no. 4, pp. 498–509, 2016.

[5]. Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in CVPR 2015, vol. 07-12-June, 2015, pp. 1110–1118.

[6]. R. Zhao, H. Ali, and P. Van Der Smagt, “Two-stream RNN/CNN for action recognition in 3D videos,” in IEEE International Conference on Intelligent Robots and Systems, vol. 2017-September, 2017, pp. 4260–4267.

[7]. A. Shahroudy, S. Member, and T.-t. Ng, “Deep Multimodal Feature Analysis for Action Recognition in RGB + D Videos,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1045–1058, 2017.

[8]. P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, “Cooperative Training of Deep Aggregation Networks for RGB-D Action Recognition,” in AAAI, 2018.

[9]. J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[10]. L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, “A Deep Structured Model with RadiusMargin Bound for 3D Human Activity Recognition,” International Journal of Computer Vision, vol. 118, no. 2, pp. 256–273, 2016.

[11]. A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,” in CVPR 2016, 2016.

[12]. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826, 2016.

[13]. C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional TwoStream Network Fusion for Video Action Recognition,” in CVPR, 2016.

[14]. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” ArXiv, p. 9, 2017.

[15]. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge 2007 (voc 2007) results (2007),” 2008.

[16]. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error

visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[17]. M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, ´ V. Ponce-Lopez, X. Bar ´ o, I. Guyon, S. Kasaei, and S. Escalera, "Deep learning for action and gesture recognition in image sequences: A survey," in Gesture Recognition. Springer, 2017, pp. 539–578.

[18]. 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.

[19]. 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.

[20]. P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3164–3172.

[21]. 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.

[22]. C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," arXiv preprint arXiv:1804.06055, 2018.

[23]. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," CVPR, 2017.

[24]. Z. Xu, V. Vilaplana, and J. R. Morros, "Additional results for action tube extraction based 3D-CNN for RGB-D action recognition," Tech. Rep. [Online]. Available: https://imatge.upc.edu/web/sites/default/files/resources/1972/Additional_material_action_recognition.pdf.

[25]. Heng Wang, Alexander Klaser, Cordelia Schmid, and ¨ Cheng-Lin Liu. Action recognition by dense trajectories. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3169–3176. IEEE, 2011.

[26]. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In Proceedings of the IEEE international conference on computer vision, pages 3551–3558, 2013.

[27]. J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4694–4702, 2015.

[28]. J. Donahue, L. Anne Hendricks, 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, pages 2625–2634, 2015.

[29]. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. IEEE TPAMI, 35(1):221–231, 2013.

[30]. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497. IEEE, 2015.

[31]. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016.

[32]. H. Wang and C. Schmid. Action recognition with improved trajectories. In IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, pages 3551–3558, 2013.

[33]. Z. Shou, D. Wang, and S.-F. Chang. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[34]. L. Wang, Y. Yu Qiao, and X. Tang. Action Recognition and Detection by Combining Motion and Appearance Features. ECCV THUMOS Workshop, 1, 2014.

[35]. R. Girshick. Fast R-CNN. In Proceedings of the International Conference on Computer Vision (ICCV), 2015.

[36]. Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In International Conference on Computer Vision (ICCV).

[37]. Qiu, Haonan, et al. "Precise Temporal Action Localization by Evolving Temporal Proposals." arXiv preprint arXiv:1804.04803 (2018).

[38]. J. Yuan, Y. Pei, B. Ni, P. Moulin, and A. Kassim. Adsc submission at thumos challenge 2015. In CVPR THUMOS Workshop, 2015.

[39]. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, pages 1725–1732, 2014.

[40]. G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence, 2017.

[41]. Duan, Jiali, et al. "Multi-modality fusion based on consensus-voting and 3D convolution for isolated gesture recognition." arXiv preprint arXiv:1611.06689 (2016).

[42]. Zhao, Yue, et al. "Temporal action detection with structured segment networks." The IEEE International Conference on Computer Vision (ICCV). Vol. 8. 2017.

[43]. Lin, Tianwei, Xu Zhao, and Zheng Shou. "Single shot temporal action detection." Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017.

[44]. S Mohsen Amiri, Mahsa T Pourazad, Panos Nasiopoulos, and Victor CM Leung. Human action recognition using meta learning for rgb and depth information. In Computing, Networking and Communications (ICNC), 2014 International Conference on, pages 363–367. IEEE, 2014.

[45]. Chen Chen, Baochang Zhang, Zhenjie Hou, Junjun Jiang, Mengyuan Liu, and Yun Yang. Action recognition from depth sequences using weighted fusion of 2d and 3d autocorrelation of gradients features. Multimedia Tools and Applications, 76(3):4651–4669, 2017.

[46]. K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.

[47]. Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.

[48]. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[49]. J. B. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. Fundamenta informaticae, 41(1, 2):187–228, 2000.

Appendix

1. Per-class accuracies for the OA1/OA2 datasets:

OA1 – action class	Accuracy		OA2 – action class	Accuracy
answering-phones	67.8%		asking-and-away	75.9 %
arranging-files	84.7 %		called-away	87.9 %
eating	96.6 %		carrying	87.9 %
moving-objects	98.3 %		chatting	67.2 %
going-to-work	96.7 %		delivering	63.4 %
finding-objects	83.1 %		eating-and-chatting	98.3 %
mopping	98.3 %		having-guest	91.4 %
sleeping	98.3 %		seeking-help	87.9 %
taking-water	100 %		shaking-hands	89.7 %
wandering	93.2 %		showing	72.4 %

2. Per-class accuracies for the NTU RGB-D dataset:

NTU RGB-D action class	Accuracy
drink water	96.65%
eat meal/snack	80.91%
brushing teeth	87.61%
brushing hair	94.75%
drop	97.38%
pickup	98.46%
throw	92.86%
sitting down	98.55%
standing up (from sitting position)	98.10%
clapping	83.62%
reading	85.88%
writing	85.43%
tear up paper	97.10%
wear jacket	100%
take off jacket	95.93%
wear a shoe	88.25%
take off a shoe	75.49%
wear on glasses	96.38%
take off glasses	98. 55 %
put on a hat/cap	98. 55 %
take off a hat/cap	97.46%

cheer up	96.29%
hand waving	94.48%
kicking something	96.01%
put something inside pocket	94.13%
hopping (one foot jumping)	96.38%
jump up	99.64%
make a phone call/answer phone	96.74%
playing with phone/tablet	94.48%
typing on a keyboard	96.65%
pointing to something with finger	94.84%
taking a selfie	95.29%
check time (from watch)	97.10%
rub two hands together	87.14%
nod head/bow	96.01%
shake head	100%
wipe face	88.23%
salute	95.20%
put the palms together	89.41%
cross hands in front (say stop)	96.20%
sneeze/cough	82.91%
staggering	98.19%
falling	96.29%
touch head (headache)	86.87%
touch chest (stomachache/heart pain)	96.29%
touch back (backache)	96.38%
touch neck (neckache)	94.84%
nausea or vomiting condition	86.16%
use a fan/feeling warm	85.17%
punching/slapping other person	95.48%
kicking other person	87.87%
pushing other person	96.65%
pat on back of other person	91.22%
point finger at the other person	96.65%
hugging other person	99.64%
giving something to other person	92.94%
touch other person's pocket	97.46%
handshaking	95.29%
walking towards each other	97.83%
walking apart from each other	96.74%

3. Per-class AP@0.5 on THUMOS14 (in %):

Action class	AP@0.5
BaseballPitch	37.8
BasketballDunk	25.7
Billiards	26.0
CleanAndJerk	10.8
CliffDiving	17.7
CricketBowling	18.1
CricketShot	11.6
Diving	17.6
FrisbeeCatch	17.8
GolfSwing	10.7
HammerThrow	29.9
HighJump	20.9
JavelinThrow	20.4
LongJump	25.6
PoleVault	10.3
Shotput	11.5
SoccerPenalty	27.5
TennisSwing	39.3
ThrowDiscus	19.5
VolleyballSpiking	11.7