A Review on Action Recognition and Action Prediction of Human(s) using Deep Learning Approaches

Syed Abdussami
Sri Jayachamarajendra College of Engineering
Department of Electronics and Communication Engineering
Mysuru-06

Nagendraprasad S.
Sri Jayachamarajendra College of Engineering
Department of Electronics and Communication Engineering
Mysuru-06

Shivarajakumara K.
Sri Jayachamarajendra College of Engineering
Department of Electronics and Communication Engineering
Mysuru-06

Sanjeet Singh
Sri Jayachamarajendra College of Engineering
Department of Electronics and Communication Engineering
Mysuru-06

A. Thyagarajamurthy
Sri Jayachamarajendra College of Engineering
Department of Electronics and Communication Engineering
Mysuru-06

ABSTRACT

Human Action Recognition and Prediction are some of the hot topics in Computer Vision these days. It has its formidable contribution in the Anomaly detection. Many research scientists have been working in this field. Many new algorithms have been tried out in recent decades. In this paper, eight such approaches proposed in eight research papers have been reviewed. Compared to their counterparts for still images (the 2D CNNs for visual recognition), the 3D CNNs are considered to be comparatively less efficient, due to the limitations like high training complexity of spatio-temporal fusion and huge memory cost. So in the first referred paper the authors have proposed MiCT (Mixed Convolution Tube – for videos) with the right use of both 2D CNNs and 3D CNNs which reduces the training time. In the second research paper, the glimpse sequences in each frame correspond to interest points in the scene that are relevant to the classified activities. Unlike the last referred paper, the third referred paper presents a novel method to recognize human action as the evolution of pose estimation maps. The fourth referred paper presents a model for long term prediction of pedestrians from on-board observations. In the fifth research article referred, an attempt has been made to recognize the Human Rights Violation activities using the Deep Convolutional Neural Networks. In the sixth research article, Convolutional LSTM is used for the purpose of detecting violent videos. The seventh paper introduces a new Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation. In the eighth research paper, a new temporal transition layer (TTL) that models variable temporal convolution kernel depths is embedded into 3D CNN to form T3D (Temporal 3D Convnets). Transferring knowledge from a pre-trained 2D CNN to a 3D CNN reduces the number of training samples required for 3D CNNs.

Keywords

CNN, SVM,MiCT, Glimpse Clouds, Two-stream Bayesian Encoder-Decoder, Pose estimation, Heat Maps, ConvLSTM, Two-stream 3D CNN, TTL, T3D.

1. INTRODUCTION

Due to the rapidly increasing amount of video records and the large number of potential applications based on automatic video analysis such as visual surveillance, human-machine interfaces, sports video analysis and video retrieval, the Human action recognition and Human Action Prediction constitute active topics in the field of Computer Vision [5]. Though this is a fundamental task, but yet this is a challenging task with considerable efforts having been invested since decades [1].

The first paper referred presents a deep architecture model to address the problems of 3D CNNs used for action recognition in videos and improvement in the performance of 3D CNNs for action recognition with the proposed Mixed 2D/3D Convolution Tube (MiCT) which enables the feature map at each spatio-temporal level to be much deeper prior to the next spatio-temporal fusion, which in turn makes it possible for the network to achieve better performance with fewer spatio-temporal fusions, while reducing the complexity of each round of spatio-temporal fusion by using the cross-domain residual connection [1]. In contrast to the 3D CNNs that stack the 3D convolution layer by layer, the proposed MiCT integrates 3D CNNs with 2D CNNs to enhance the feature learning with negligible increase in memory usage and complexity [1]. Experiment results show that the proposed deep framework MiCT-Net with MiCT significantly enhances the performance of 3D CNNs for spatio-temporal feature learning and achieves state-of-the-art performance on three well-known benchmark datasets for action recognition [1].

The second paper referred presents a method for human activity recognition that does not require articulated pose during testing, and models activities using two attention processes; one extracting a set of glimpses per frame and one reasoning about entities over time [2]. The proposed model is evaluated on two datasets, NTU RGB-D and N – UCLA Multiview Action 3D, outperforming the state-of-the-art by a large margin [2].

Since many of the existing methods do not directly distinguish human body from videos, these methods are easily affected by clutters and non-action motions from backgrounds [3]. To address this limitation, an alternative solution is proposed in this paper to detect human [9] and estimate the body pose in each frame. The success of 3D human pose inspires the authors to estimate 2D human poses from videos for action recognition [3].

Anticipation is key in preventing collisions by predicting future movements of dynamic agents e.g. people and cars in inner cities [4]. Without anticipation, domain knowledge and experience, drivers would have to maintain an equally large safety distance to all objects, which is clearly impractical in dense and cluttered inner city traffic [4]. Additionally, anticipation enables decision making [4]. The authors' contributions to this work include an approach to long-term prediction of pedestrian bounding box sequences from a mobile platform, a novel sequence to sequence model which provides a theoretically grounded approach to quantify uncertainty associated with each prediction and analysis of dependencies between uncertainty estimates and actual prediction error leading to an empirical error bound [4].

Human rights violations continue to take place in many parts of the world today, while they have been ongoing during the entire human history [5]. These days, organizations concerned with human rights are increasingly using digital images as a mechanism for supporting the exposure of human rights violations and international humanitarian law violations [5]. However, utilizing current advances in technology for studying, prosecuting and possibly preventing such misconduct from occurring have not yet made any progress [5]. From this perspective, supporting human rights is seen as one scientific domain that could be strengthened by the latest developments in computer vision [5]. To support the continued growth of images and videos in human rights and international humanitarian law monitoring campaigns, this study examines how vision-based systems can support human rights monitoring efforts by accurately detecting and identifying human rights violations utilizing digital images [5].

Human rights violations continue to take place in the world through human history [5]. To counter this, various human rights organization are increasingly using digital images as a mechanism for supporting the exposure of human rights humanitarian law violations [5]. Utilizing advances in technology for studying, prosecuting and possibly preventing such misconduct from occurring have not yet made any progress [5]. Hence supporting human rights is seen as one scientific domain that is strengthened by the developments in computer vision by accurately detecting and identifying human rights violations utilizing digital images [5]. The authors used 10 different CNN models fine-tuned them using transfer learning for feature extraction and a Support Vector Machine for classification [5].

Public violence is on the rise in modern times in new forms like terror attacks, mob lynching etc. This has resulted in heavy usage of surveillance to curb violence [6]. But all systems require manual human inspection of videos for identification of events, hence the authors aim to develop an end-to-end trainable deep neural network for violent video classification [6]. They used a combination of CNN and LSTM for encoding frame level changes which works better than a vanilla CNN [6]. They showed that a deep neural network trained on the frame difference performs better than a model trained on raw frames and experimentally validated the effectiveness of the proposed method using three widely used benchmarks for violent video classification [6].

In the seventh paper, the experimental strategy is to re implement a number of representative neural network architectures from the literature, and then analyze their transfer behavior by first pre-training each one on Kinetics and then fine-tuning each on HMDB-51 and UCF-101 datasets [7]. The results suggest that there is always a boost in performance by pre-training, but the extent of the boost varies significantly with the type of architecture [7]. Based on these findings, a new model is introduced that has the capacity to take advantage of pre-training on Kinetics, and can achieve a high performance [7]. The model termed as "Two-Stream Inflated 3D Conv Nets" (I3D), builds upon state-of-the-art image classification architectures, but inflates their filters and pooling kernels (and optionally their parameters) into 3D, leading to very deep, naturally spatio-temporal classifiers [7]. An I3D model obtains performance far exceeding the state-of-the art, after pre-training on Kinetics [7].

In the eighth paper, a novel deep feature extractor network – TTL with the aim to model variable temporal 3D convolution kernel depths over shorter and longer time ranges is introduced firstly [8]. Observation has been made that training 3D CNNs from scratch takes two months for them to learn a good feature representation from a large scale dataset like Sports1M [8]. So the authors have employed supervision transfer across architectures, thus avoiding the need to train 3D CNNs from scratch [8]. Specifically, a 2D CNN pre-trained on ImageNet can act as 'a teacher' for supervision transfer to a randomly initialized 3D CNN for a stable weight initialization [8]. In this way the excessive computational workload and training time are avoided [8]. T3D achieves the state-of-the-art performance on HMDB51 and UCF101 datasets and also competitive results on Kinetics dataset [8].

2. METHODOLOGIES

In this research article, to overcome the difficulties of using only 3D CNNs, the 2D CNNs are combined with them because they can be trained effectively, constructed deeply, and learned with huge datasets [1]. This combination is known as MiCT and this new model empowers effective feature learning [1]. It integrates 2D convolutions with 3D convolutions to output much deeper feature maps at each round of spatio-temporal fusion [1]. The authors propose mixing 3D and 2D convolutions in two ways, i.e. concatenating connections and cross-domain residual connections [1]. In concatenating connections of 2D and 3D Convolution, feature maps of a 3D input are achieved by coupling the 3D convolution with the 2D convolution block serially in which the 3D convolution enables spatio-temporal information fusion while the 2D convolution block deepens feature learning for each 2D output of the 3D convolution [1]. The MiCT with Cross-Domain Residual Connections introduces a 2D convolution between the input and output of the 3D convolution to further reduce spatio-temporal fusion complexity and facilitate the optimization of the whole network [1]. A simple but yet efficient deep MiCT Network (MiCT-Net in short) is proposed by stacking the MiCT together [1]. The MiCT-Net takes the RGB video sequences as inputs and is end-to-end trainable [1]. Compared to the baseline C3D architecture, the MiCT-Net contains fewer 3D convolutions for spatio-temporal fusion while producing deeper feature maps and limiting the complexity of the entire deep model [1]. Moreover, unlike traditional 3D CNNs, MiCTNet framework is able to take advantage of 2D models pre-trained on large image datasets [1]. The pre-trained parameters on large image datasets potentially provide MiCT with more advanced initialization in 2D convolution blocks for feature learning [1].

In this research article, the Glimpse Cloud model processes videos using several key components: a recurrent spatial attention model that extracts features from different local glimpses following an attention path in each video over the frames and multiple glimpses in each frame; distributed soft-tracking workers, which process these spatial features sequentially [2]. As the input data is unstructured, the spatial glimpses are soft-assigned to the workers, such that no hard decisions are made at any point [2]. To this end, an external memory module keeps track of the glimpses seen in the past, their features, and past soft-assignments, and produces new soft-assignments optimizing spatio-temporal consistency [2]. The approach is fully-differentiable, allowing end-to-end training of the full model [2]. The activities are recognized jointly based on global and local features [2]. In order to speed up calculations and to avoid extracting redundant calculations, a single feature space computed is used by a global model [2]. Inspired by human behavior when scrutinizing a scene, a fixed number of features are extracted from a series of glimpses within each frame [2]. The process of moving from one glimpse to another is achieved with a recurrent model [2]. A differentiable glimpse function, a Spatial Transformer Network (STN) allows the attention process to perform a differentiable crop operation on each feature map [2]. Features are extracted using a transformed ROI average pooling, resulting in a 1D feature vector [2]. In order to make the spatial attention process aware of frame transitions a context vector is introduced which contains high level information about humans present in the current frame [2]. The context vector is obtained by global average pooling over the spatial domain of the penultimate feature maps of a given time step [2]. Using this type of model along with unstructured feature points makes it possible for an effective and efficient Human Action Recognition [2].

In this research article, Deep Convolution Neural Network is used to predict action labels [3]. After Feature Extraction, Convolution Pose Machine is used to predict pose estimation maps for each body part in each frame [3]. Then Spatial Map pooling method is used to compress the heat map. 3D pose provides direct physical interpretation for human actions from depth videos [3]. To better utilize the pose estimation maps, instead of relying on the inaccurate 2D pose estimated from the pose estimation maps, the authors propose to directly model the evolution of pose estimation maps for action recognition [3]. As a robust and compact video representation method, temporal rank pooling has the ability to aggregate the temporal relevant information throughout a video via a learning to rank approach [3]. The encoded temporal information denotes the temporal order among frames, which is a robust feature showing less sensitivity to different types of input data [3]. Spatial Rank Pooling reduces features in each heat map to generate a compact feature. Body guided sampling does not differentiate different joints. Complex body occlusions and large appearance in the video sequences collected make it challenging as well as difficult for action recognition [3].

In this research article, the authors work on an approach that jointly predicts 1 second beforehand, the ego motions and people trajectories over large time horizons [4]. Automatic scene understanding is used by this model for the prediction of future movements of possibly vehicles and pedestrians in traffic scenarios so that it can be further used for safety distance prediction [4]. The vehicles especially cars must travel at a conservative and careful driving speed of 25 miles/ hr in residential areas, so that the distance traveled in 1 second corresponds roughly to the breaking distance [4]. This method does not depend upon 3D co-ordinates, it just requires one camera [4]. This approach uses social LSTM to find the trajectory of each person in a scene [4]. This model detects the motion patterns of humans and vehicles & this deep learning architecture does not model uncertainty, assuming uniform content observation noise [4]. Bayesian neural networks offer a probabilistic view of deep learning and provide model (epistemic) uncertainty estimates [4]. The odomtery prediction stream predicts a mean estimate of the future vehicle ego-motion and here the authors use an RNN encoder-decoder architecture used for bounding box prediction, but without the embedding layers [4]. A basic sequence is described to sequence RNN first and then extend it to predict distributions and provide uncertainty estimates [4]. During training and prediction, Monte-Carlo integration was used [4]. Bayesian approach and long-term prediction of odometry based novel two-stream Bayesian LSTM encoder-decoder is proposed and evaluated by the authors in their work [4].

In this paper the authors have introduced well sampled human rights-centric dataset, called the Human Rights Understanding (HRUN) dataset, which consists of 4 different categories of human rights violations and 100 diverse images per category [5]. Images were downloaded for each class using a Python interface to the Google and Bing application programming interfaces (APIs), with the maximum number of images permitted by their respective API for each query term [5]. Secondly the authors conducted a large set of rigorous experiments for the task of recognizing human rights violations [5]. The authors created a pipeline using n CNN models and m SVM for every m categories for image classification where every block is fixed except the feature extractor as different deep convolutional networks are plugged in, one at a time, to compare their performance utilizing the mean average precision (mAP) metric [5]. The training dataset is used as input to the first CNN architecture C1 [5]. The output of C1, as described above, is then utilized to train m SVM classifiers [5]. Once trained, the test dataset Ts is employed to assess the performance of the pipeline using mAP [5]. Since the entire pipeline is fixed (including the training and test datasets, learning procedure and evaluation protocol) for all n CNN architectures, the differences in the performance of the classification pipeline can be attributed to the specific CNN architectures used [5].

In this research article the authors developed an end-to-end trainable deep neural network models for classifying videos into violent and non-violent ones [6]. The authors developed a unique model featuring convolutional layers followed by max pooling operations for extracting discriminant features and convolutional long short memory (convLSTM) for encoding the frame level changes, which characterizes violent scenes, existing in the video [6]. For a system to identify a video as violent or non-violent, it should be capable of encoding localized spatial features and the manner in which they change with time [6]. The convolutional layers are trained to extract hierarchical features from the video frames and are then aggregated using the convLSTM layer [6]. The frames of the video under consideration are applied sequentially to the model [6]. Once all the frames are applied, the hidden state of the convLSTM layer in this final time step contains the representation of the input video frames applied [6]. This video representation, in the hidden state of the convLSTM, is then applied to a series of fully-connected layers for classification [6]. In the network, instead of applying the input frames as such, the difference between adjacent frames are given as input [6]. In this way, the network is forced to model the changes taking place in adjacent frames rather than the frames itself [6].

The seventh research article referred demonstrates that with the new Two Stream Inflated 3D ConvNet architecture, the 3D ConvNets can benefit from ImageNet 2D ConvNet designs and, optionally, from their learned parameters [7]. Image (2D) classification models can be converted into 3D ConvNets by starting with a 2D architecture, and inflating all the filters and pooling kernels – endowing them with an additional temporal dimension [7]. Filters are typically square and authors have just made them cubic – N×N filters become N×N×N [7]. To bootstrap parameters from the pre-trained ImageNet models, an image can be converted into a (boring) video by copying it repeatedly into a video sequence [7]. The 3D models can then be implicitly pre-trained on ImageNet, by satisfying the boring-video fixed point: the pooled activations on a boring video should be the same as on the original single-image input [7]. The boring video fixed-point leaves ample freedom on how to inflate pooling operators along the time dimension and on how to set convolutional/pooling temporal stride [7]. The authors have paced receptive field growth in space, time and network depth by training the model using 64-frame snippets and testing using the whole videos, averaging predictions temporally [7]. Because of the absence of the recurrence behavior in 3D ConvNets when compared to Optical flow algorithms, a two-stream configuration - with one I3D network trained on RGB inputs, and another on flow inputs which carry optimized and smooth flow information would be valuable [7]. The two networks are trained separately and their predictions averaged at test time [7]. Adoption of a two stream configuration makes sure that while 3D ConvNets can directly learn about temporal patterns from an RGB stream, their performance can still be greatly improved by including an optical-flow stream [7].

In the eighth research article referred, inspired by GoogleNet, the TTL (Temporal Transition Layer) has been proposed that consists of several 3D Convolutional kernels, with diverse temporal depths, thus allowing to capture short, mid and long term dynamics for a video representation that embodies more semantic information not captured when working with some fixed temporal depth homogeneously throughout the network [8]. The TTL layer has been employed in a DenseNet3D architecture and the resulting network named as Temporal 3D ConvNets (T3D), which here operates on 32 RGB frames [8]. The feature maps from the layers preceding the 3D TTL layer are simply concatenated into a single tensor and then fed into the 3D Pooling Layer, resulting to the output TTL Feature map [8]. The TTL output feature maps are densely fed forward to all the subsequent layers, and are learnt end-to-end [8]. The output of the network is a video-level prediction [8]. Although T3D model has 1.3 times more model parameters than DenseNet3D, but it is worth to have it because of its outstanding performance [8]. Instead of using fixed 3D Convolutions homogeneously throughout the network, one can readily employ TTL in other architectures too such as in Res3D or I3D [8]. Supervision and knowledge transfer between cross architectures (2D ConvNets and 3D ConvNets) avoids the need to train 3D ConvNets from scratch [8]. Another important aspect of proposing the transfer learning for 3D ConvNets is for finding a cheaper way to train 3D ConvNets when availability of large datasets is at scarce, because the knowledge transfer reduces the need for more labeled data and very large datasets [8].

3. DISCUSSIONS

After studying different research articles, it can be observed that the performance of the experiment obtained in each of the paper differs because of the methodologies employed and the datasets used. The below table 1 tabulates the individual research articles of human action recognition and action prediction, reviewed in this paper:

Table 1: Tabulation of the research articles related to action recognition and action prediction.

Sl. No .	Title of the paper and the author	Methodology Used	Dataset used to implement
1	MiCT : Mixed 3D/2D Convolutional Tube for Human action Recognition by Yizhou Zhou, Xiaoyansum, Zheng-Jung-Jan Zha, Wenjun Zeng	MiCT, MiCTNet	1.HMDB51 2.UCF101 3.Sports Dataset
2	Glimpse Clouds : Human Activity Recognition from Unstructured Feature points by Fabien Baradel, Christian Wolf, Julien Mile, Graham W.Taylor	Glimpse Clouds, STN	1.NTU RGB+D Datasets 2.Northwestern UCLA Multiview Action 3D Dataset
3	Recognizing Activity Human Actions as the Evolutions of Pose of Estimation Maps by Mengyuan Liu, Junsong Yuan	2D Pose estimation and Heat Maps	1.NTU RGB+D dataset 2.UTD-MHAD dataset
4	Long-Term On-Board Prediction of people in Traffic Scenes under Uncertainty by Aprathim Bhattacharya, Mario Fritz, Brent Schiele	Bayesian Neural Network, Bounding Box Prediction and Odometry Prediction	1.Cityscapes Datasets
5	Detection of Human Rights Violation in Images: can Convolutional Neural Networks help? By Grigorios Kalliatakis, Shoaib Ehsan, Maria Fasli, Ales Leonardis, Juergen Gall and Kluas D.McDonald-Maier	CNN, ResNet, GoogleNet, VGG	1.HRUN dataset
6	Learning to Detect Violent Videos using	Convolutional LSTM	1.Hockey 2.Fight Dataset 3.Movies

	Convolutional Long Short-Term Memory by Swathikiran , Sudhakarn and Oswld Lanz		Dataset 4.Violent-Flows 5.Crowd Violence Datasets
7	Quo Vadis, Action Recognitions? A New Model and the Kinetics Dataset by Joao Carreira and Andrew Zisserman	Two-stream I3D (Inflated 3D ConvNets)	1. HMDB51 2.UCF101 3.Kinetics
8	Temporal 3D ConvNets : New Architecture and Transfer Learning for Video Classification by Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossain Karami, Mohammad Mahdi Arzani, Rahman Yosefzadeh, Luc Van Gool	T3D(Temporal 3D ConvNets)	1. HMDB51 2.UCF101 3.Kinetics

4. CONCLUSION

In this review paper, the recent studies in Human Action Recognition and Action Prediction that employ Deep Learning Approaches have been reviewed. The authors have appropriately tailored the methodologies, or in some cases have appropriately integrated the old ones, in order to meet the demands of the problem in hand. The proposed methodologies have been evaluated by them on state-of-the-art popular datasets, related to the respective fields of their research. The future scope of these reviewed papers is that their employment in anomaly detection in video footages from surveillance systems significantly enhances the systems' effectiveness, efficiency and area of influence.

5. REFERENCES

[1] Y. Zhou, X. Sun, Z. Zha and W. Zeng, "MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 449-458

[2] F. Baradel, C. Wolf, J. Mille and G. W. Taylor, "Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 469-478

[3] M. Liu and J. Yuan, "Recognizing Human Actions as the Evolution of Pose Estimation Maps," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 1159-1168.

[4] A. Bhattacharyya, M. Fritz and B. Schiele, "Long-Term On-board Prediction of People in Traffic Scenes Under Uncertainty," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 4194-4202.

[5] Kalliatakis, Grigorios & Ehsan, Shoaib & Fasli, Maria & Leonardis, Ales & Gall, Juergen & McDonald-Maier, Klaus. (2016). Detection of Human Rights Violations in Images: Can Convolutional Neural Networks help?.

[6] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, 2017, pp. 1-6.

[7] Carreira, J & Zisserman, Andrew. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 4724-4733. 10.1109/CVPR.2017.502.

[8] Diba, Ali & Fayyaz, Mohsen & Sharma, Vivek & Hossein Karami, Amir & Mahdi Arzani, Mohammad & Van Gool, Luc & Yousefzadeh, Rahman. (2017). Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification.

[9] Tu, Zhigang&Xie, Wei & Qin, Qianqing&Poppe, Ronald &Veltkamp, Remco & Li, Baoxin& Yuan, Junsong. (2018). Multi-stream CNN: Learning representations based on human-related regions for action recognition. Pattern Recognition. 79. 32-43. 10.1016/j.patcog.2018.01.020.