Contextual Action Recognition using Tube Convolutional Neural Network (T-CNN)

S. Venkata Kiran¹, Dr. R.P. Singh²

¹ Research Scholar (SSSEC1527), ECE Dept., SSSUTMS, Bhopal, India
(e-mail: venkatkiran95@gmail.com)
² Vice-Chancellor & Professor, ECE Dept., SSSUTMS, Bhopal, India

Abstract: Deep learning has been shown to accomplish exceed expectations loaned comes about for picture characterization and protest identification. In any case, the effect of Deep learning on video examination has been constrained because of multifaceted nature of video information and absence of a documentations. Past convolutional neural systems (CNN) based video activity identification approaches ordinarily comprise of two noteworthy advances: outline level activity proposition age and relationship of recommendations crosswise over edges. Additionally, the greater part of these techniques utilize two-stream CNN system to han-dle spatial and fleeting element independently. In this paper, we propose a conclusion to-end Deep learning system called Tube Con-volutional Neural Network (T-CNN) for activity identification in recordings. The proposed design is a bound together profound net-work that can perceive and confine activity in light of 3D convolution highlights. A video is first isolated into measure up to length cuts and next for each clasp an arrangement of tube expert posals are produced in light of 3D Convolutional Network (ConvNet) highlights. At last, the tube proposition of contrast ent cuts are connected together utilizing system stream and spatio-transient activity identification is performed utilizing these connected video recommendations. Broad analyses on a few video datasets show the unrivaled execution of T-CNN for grouping and restricting activities in both trimmed and untrimmed recordings contrasted with condition of human expressions.

Keywords: action recognition; T-CNN

I. INTRODUCTION

The objective of activity location is to identify each event of a given activity inside a long video, and to restrict every identification both in space and time. Deep learning getting the hang of learning based methodologies have essentially enhanced video activity acknowledgment execution. Contrasted with activity acknowledgment, activity location is an all the more difficult assignment because of adaptable volume shape and huge spatio-fleeting inquiry space. Past Deep learning based activity identification ap-proaches first recognize outline level activity proposition by pop-ular proposition calculations [5, 30] or via preparing proposition net-works [19]. At that point the casing level activity proposition is associated crosswise over edges to shape last activity identification through

Following based methodologies [32]. In addition, so as to top true both spatial and fleeting data of an activity, two-stream arranges (a spatial CNN and a movement CNN) are utilized. In this way, the spatial and movement data are prepared independently. Region Convolution Neural Network (R-CNN) for protest discovery in pictures was proposed by Girshick et al. [4]. It was trailed by a quick R-CNN proposed in [3], which incorporates the classifier too. Afterward, speedier R-CNN [20] was created by presenting an area proposition organize. It has been widely used to create great outcomes for question recognition in pictures. A characteristic speculation of the RCNN from 2D pictures to 3D spatio-fleeting volumes is to contemplate their adequacy for the issue of activity discovery in recordings. A direct spatio-fleeting speculation of the R-CNN approach is treat activity recognition in recordings as an arrangement of 2D picture location utilizing quicker RCNN. Nonetheless, tragically, this approach does not take the worldly data into account and isn't adequately expressive to recognize activities.

Motivated by the spearheading work of speedier R-CNN, we propose Tube Convolutional Neural Network (T-CNN) for activity discovery. To better catch the spatio-worldly data of video, we misuse 3D ConvNet for activity recognition, since it can catch movement attributes in recordings and shows promising outcome on video activity acknowledgment. We propose a novel system by utilizing the clear energy of 3D ConvNet. In our approach, an info video is partitioned into approach length cuts first. At that point, the cuts are nourished into Tube Proposal Network (TPN) and an arrangement of tube proposition are gotten. Next, tube proposition from each video cut are connected by their actionness scores what's more, cover between nearby proposition to shape a total tube proposition for spatio-transient activity restriction in the video. At long last, the Tube-of-Interest (ToI) pooling is connected to the connected activity tube proposition to produce a settled length include vector for activity mark expectation.

Our work makes the following contributions:

• We present a Tube Proposal Network, which influences skip pooling in worldly space to safeguard transient data for activity confinement in 3D volumes.
• We propose another pooling layer – Tube-of-Interest (ToI) pooling layer in T-CNN. The ToI pooling layer is a 3D speculation of Region-of-Interest (RoI) pooling layer of R-CNN. It successfully eases the issue with variable spatial and transient sizes of tube proposition. We demonstrate that ToI pooling can enormously enhance the acknowledgment comes about.
• We widely assess our T-CNN for activity location in both trimmed recordings from UCF-Sports, J-HMDB what’s more, UCF-101 datasets and untrimmed recordings from THUMOS’14 dataset and accomplish best in class execution.

II. LITERATURE SURVEY/RELATED WORK
Convolutional Neural Networks (CNN) have been shown to accomplish brilliant outcomes for activity acknowledgment [17, 18]. Karpathy et al. [14] investigate different frame level combination techniques after some time. Ng et al. [31] utilize repetitive neural system utilizing the CNN include. Since these methodologies just utilize outline based CNN highlights, the worldly data is dismissed. Simonyan et al. [22] propose the two-stream CNN approach for activity acknowledgment. Other than an exemplary CNN which takes pictures as an info, it has a different system for optical stream. Additionally, Wang et al. meld the directions and CNN highlights. In spite of the fact that these techniques, which grasp hand-created worldly element as a different stream, indicate promising execution on activity acknowledgment, in any case, they don’t utilize end to end profound system and require isolate calculation of optical stream and improvement of the parameters. 3D CNN is a coherent answer for this issue. Ji et al. [9] propose a 3D CNN based human indicator and go to fragment human subjects in recordings. Tran et al. [27] use 3D CNN for extensive scale activity acknowledgment issue. Sun et al. [25] propose a factorization of 3D CNN and endeavor numerous approaches to decay Convolutional portions. In any case, to the best of our insight, we are the initial ones to misuse 3D CNN for activity discovery. Contrasted with activity acknowledgment, activity location is an additionally difficult issue [2, 7, 29], which has been a dynamic territory of research. Ke et al. [15] exhibit an approach for occasion identification in swarmed recordings. Tian et al. [26] create Spatio-transient Deformable Parts Model [1] to recognize activities in recordings. Jain et al. [6] and Soomro et al. [23] utilize supervoxel and particular inquiry to restrict the activity limits. As of late, specialists have utilized the energy of profound learning for activity recognition. Creators in [5] remove outline level activity proposition utilizing particular hut and connect them utilizing Viterbi calculation. While in [30] outline level activity proposition are acquired by EdgeBox and connected by a following calculation. Two-stream R-CNNs for activity identification is proposed in [19], where a spatial Region Proposal Network (RPN) and a movement RPN are utilized to create outline level activity recommendations. Nonetheless, these profound learning-based approaches recognize activities by connecting outline level activity recommendations and treat the spatial and transient highlights of a video independently via preparing two-stream CNN. In this manner, the fleeting consistency in recordings isn’t very much investigated in the system. Conversely, we decide activity tube recommendations specifically from input recordings and concentrate conservative and more successful spatio-transient highlights utilizing 3D CNN. For protest discovery in pictures, Girshick et al. propose Region CNN (R-CNN) [4]. In their approach locale recommendations are removed utilizing particular hut. At that point the competitor areas are distorted to a settled size and encouraged into ConvNet to remove CNN highlights. At long last, SVM display is prepared for question grouping. A quick form of R-CNN, Fast R-CNN, is exhibited in [3]. Contrasted with the multi-arrange pipeline of R-CNN, quick R-CNN joins protest classifier in the system and trains question classifier and jumping box regressor at the same time. Area of intrigue (RoI) pooling layer is acquainted with extricate settled length highlight vectors for jumping boxes with various sizes. As of late, quicker R-CNN is proposed in [20]. It presents a RPN (Region Proposal Network) to trade specific look for proposition age. RPN shares full picture convolutional highlights with the identification organize, subsequently the proposition age is nearly taken a toll free. Speedier R-CNN accomplishes state-of-the-art craftsmanship question discovery execution while being effective amid testing. Roused by its superior, in this paper we investigate summing up quicker R-CNN from 2D picture areas to 3D video volumes for activity identification.

III. GENERALIZING R-CNN FROM 2D TO 3D
To better catch the spatio-fleeting data in video, we misuse 3D CNN for activity proposition age also, activity acknowledgment. One favorable position of 3D CNN over 2D CNN is that it catches movement data by applying convolution in both time and space. Since 3D convolution what’s more, 3D max pooling are used in our approach, not just in the spatial measurement yet in addition in the fleeting measurement, the span of video cut is decreased while recognizable data is concentrated. To deliver a settled length include vector, we propose another pooling layer – Tube-of-Intrigue (ToI) pooling layer. The ToI pooling layer is a 3D speculation of Region-of-Interest (RoI) pooling layer of R-CNN. The exemplary max pooling layer characterizes the part size, walk and cushioning which decides the state of the yield. Interestingly, for RoI pooling layer, the yield shape is settled initially, at that point the portion size and walk are resolved in like manner. Contrasted with RoI pooling which takes 2D highlight guide and 2D districts as information, ToI pooling manages highlight 3D shape and 3D tubes. Signify the span of a component 3D square as d3hww, where d, h and w individually speak to the profundity, stature and width of the element solid shape. Back-propagation of ToI pooling layer routes the derivatives from output back to the input. Assume xi is the i-th activation to the ToI pooling layer, and yj is the j-th output. Then the partial derivative of the loss function (L) with respect to each input variable xi can be expressed as:

$$\frac{\partial L}{\partial x_i} = \frac{\partial y_j}{\partial x_i} \frac{\partial L}{\partial y_j} = \sum_i [i = f(j)] \frac{\partial L}{\partial y_j}.$$  \hspace{1cm} (1)
IV. T-CNN PIPELINE

As shown in Figure 1, our T-CNN is an end-to-end deep learning framework that takes video clips as input. The core component is the Tube Proposal Network (TPN) (see Figure 3) to produce tube proposals for each clip. Linked tube proposal sequence represents spatio-temporal action detection in the video and is also used for action recognition.

A. Tube Proposal Network

Our 3D ConvNet comprises of seven 3D convolution layers and four 3D max-pooling layers. We signify the piece state of 3D convolution/pooling by d×h×w, where d, h, w are profundity, stature and width, individually. In all convolution layers, the bit sizes are 3×3×3, cushioning and walk stay as 1. The quantities of channels are 64, 128 and 256 individually in the initial 3 convolution layers also, 512 in the rest of the convolution layers. The part measure is set to 1 × 2 × 2 for the initial 3D max-pooling layer, also, 2×2×2 for the rest of the 3D max-pooling layers.

Figure 1: Overview of the proposed Tube Convolutional Neural Network (T-CNN).
B. Linking Tube Proposals
We obtain a set of tube proposals for each video clip after the TPN. We then link these tube proposals to form a proposal sequence for spatio-temporal action localization of the entire video. Each tube proposal from different clips can be linked in a tube proposal sequence for action detection.

Figure 2: An example of linking tube proposals in each video clips using network flow. In this example, there are three video clips and each has two tube proposals, resulting in 8 video proposals.

C. Action Detection
In the wake of connecting tube recommendations, we get an arrangement of connected tube proposition arrangements, which speak to potential activity cases.

The following stage is to group these connected tube proposition arrangements. The tube proposition in the connected groupings may have distinctive sizes. Keeping in mind the end goal to extricate a settled length include vector from every one of the connected proposition grouping, our proposed ToI pooling is used. At that point the ToI pooling layer is trailed by two completely associated layers and a dropout layer. The measurement of the last completely associated layer is N + 1 (N activity classes and 1 foundation class).

V. EXPERIMENTS
To verify the effectiveness of the proposed T-CNN for action detection, we evaluate T-CNN on three trimmed video datasets including UCF-Sports [21], J-HMDB [8], UCF-101 [11] and one untrimmed video dataset – THUMOS’ 14 [12].

Implementation Details
We implement our method based on the Caffe toolbox [10]. The TPN and recognition network share weights in their common layers. Due to memory limitation, in training phase, each video is divided into overlapping 8-frame clips with resolution 300 × 400 and temporal stride 1. When training the TPN network, each anchor box is assigned a binary label. Either the anchor box which has the highest IoU overlap with a ground-truth box, or an anchor box that has an IoU overlap higher than 0.7 with any round-truth box is assigned a positive label, the rest are assigned negative label. In each iteration, 4 clips are fed into the network. Since the number of background boxes is much more than that of action boxes, to well model the action, we randomly select some of the negative boxes to balance the number of positive and negative samples in a batch. For recognition network training, we choose 40 linked proposal sequences with highest scores in a video as Tubes of Interest. Our model is trained in an alternative manner. First, Initialize TPN based on the pre-trained model in [27], then using the generated proposals to initialize recognition networks. Next, the weights tuned by recognition network are used to update TPN. Finally, the tuned weights and proposals from TPN are used for finalizing recognition network. For all the networks for UCF-Sports and J-HMDB, the learning rate is initialized as 10\(^{-3}\) and decreased to 10\(^{-4}\) after 30k batches. Training terminates after 50k batches. For UCF-101 and THUMOS’14, the learning rate is initialized as 10\(^{-3}\) and decreased to 10\(^{-4}\) after 60k batches. Training terminates after 100k batches. During testing, each video is divided into nonoverlapping 8-frame clips. If the number of frames in video cannot be divided by 8, we pad zeros after the last frame to make it dividable. 40 tube proposals with highest actionness confidence through TPN are chosen for the linking process. Non-maximum suppression (NMS) is applied to linked proposals to get the final action detection results.

Datasets and Experimental Results
UCF-Sports. This dataset contains 150 short videos of 10 different sport classes. Videos are trimmed and bounding boxes annotations are provided for all frames. We follow the training and test split defined in [16]. We use the IoU criterion and generate ROC curve in Figure 5(a) when overlap criterion equals to \(\alpha = 0.2\). Figure 5(b) illustrates AUC (Area-Under-Curve) measured with different overlap criterion. In direct comparison, our T-CNN clearly outperforms all the competing methods shown in the plot. We are unable to directly compare the detection accuracy against Peng et al. [19] in the plot.
since they do not provide the ROC and AUC curves. As shown in Table 2, the frame level mAP of our approach outperforms theirs in 8 actions out of 10. Moreover, by using the same metric, the video mAP of our approach reaches 95.2 ($\alpha = 0.2$ and 0.5), while they report 94.8 ($\alpha = 0.2$) and 94.7 ($\alpha = 0.5$). J-HMDB. This dataset consists of 928 videos with 21 different actions. All the video clips are well trimmed. There are three train-test splits and the evaluation is done on the average results over the three splits. The experiment results comparison is shown in Table 3. We report our results with 3 metrics: frame-mAP, the average precision of detection at frame level as in [5]; video-mAP, the average precision at video level as in [5] with IoU threshold $\alpha = 0.2$ and $\alpha = 0.5$. It is evident that our T-CNN consistently outperforms the state-of-the-art approaches in terms of all three evaluation metrics. UCF101. This dataset has 101 actions. For action detection task, a subset of 24 action classes and 3, 207 videos have spatio-temporal annotations. Similar to other methods, we perform the experiments on the first train/test split only. We report our results in Table 4 with 3 metrics: frame-mAP, video-mAP ($\alpha = 0.2$) and video-mAP ($\alpha = 0.5$). Our approach again yields the best performance. Moreover, we also report the action recognition results of T-CNN on the above three datasets in Table 5. THUMOS’14. To further validate the effectiveness of our proposed T-CNN approach for action detection, we evaluate it using the untrimmed videos from the THUMOS’14 dataset [12]. The THUMOS’14 spatio-temporal localization task consists of 4 classes of sports actions: BaseballPitch, golfSwing, TennisSwing and ThrowDiscus. There are about 20 videos per class and each video contains 500 to 3,000 frames. The videos are divided into validation set and test set, but only video in the test set have spatial annotations provided by [24]. Therefore, we use samples corresponding to those 4 actions in UCF-101 with spatial annotations to train our model. In untrimmed videos, there often exist other unrelated actions besides the action of interest. For example, “walking” and “picking up a golf ball” are considered as unrelated actions when detecting “GolfSwing” in video. We denote clips which have positive ground truth annotation as positive clips, and the other clips as negative clips (i.e. clips contain only unrelated actions). If we randomly select negative samples for training, the number of boxes on unrelated actions is much smaller than that of background boxes (i.e. boxes capturing only image background). Thus the trained model will have no capability to distinguish action of interest and unrelated actions. To this end, we introduce a so called negative sample mining process. Specifically, when initializing the TPN, we only use positive clips. Then we apply the model on the whole training video both positive clips and negative clips). Most false positives in negative clips should include unrelated actions to help our model learn the correlation between action of interest and unrelated actions. Therefore we select boxes in negative clips with highest scores as hard negatives because low scores probably infer image background. In updating TPN procedure, we choose 32 boxes which have IoU with any round truth greater than 0.7 as positive samples and randomly pick another 16 samples as negative. We also select 16 samples from hard negative pool as negative. Therefore, we efficiently train a model, which is able to distinguish not only action of interest from background, but also action of interest from unrelated actions. The mean ROC curves of different methods on THUMOS’14 action detection are plotted in Figure 5(c). Our method without negative mining performs better than the baseline method Sultani et al. [24]. Additionally, with negative mining, the performance is further boosted. For qualitative results, we shows examples of detected action tubes in videos from UCF-Sports, JHMDB, UCF-101 (24 actions) and THUMOS’14 datasets (see Figure 6). Each block corresponds to a different video that is selected from the test set. We show the highest scoring action tube for each video.

![Figure 5: The ROC and AUC curves for UCF-Sports Dataset [21] are shown in (a) and (b), respectively. The results are shown for Jain et al. [6] (green), Tian et al. [26] (purple), Soomro et al. [23] (blue), Wang et al. [28] (yellow), Gkioxari et al. [5] (cyan) and Proposed Method (red). (c) shows the mean ROC curves for four actions of THUMOS’14. The results are shown for Sultani et al. [24] (green), proposed method (red) and proposed method without negative mining (blue).](image-url)

VI. CONCLUSION

In this paper we propose a conclusion to-end Tube Convolutional Neural Network (T-CNN) for activity recognition in recordings. It abuses 3D convolutional system to extricate viable spatio-worldly highlights and perform activity confinement and acknowledgment in a bound together structure. Coarse proposition boxes are thickly tested in view of the 3D convolutional include 3D square and connected for activity acknowledgment and limitation. Broad examinations on a few benchmark datasets show the quality of T-CNN for spatiotemporal confining activities, even in untrimmed recordings.

REFERENCES


Authors

S. Venkata kiran is a Research Scholar (SSSEC1527) in the Department of Electronics & Communication Engineering at Sri Satya Sai University of Technology and Medical Sciences, Sehore, Madhya Pradesh. His areas of interest are Image processing, Artificial Neural Networks and signal Processing. His Publications includes 8 research articles in International Journals / conferences.

Shri. Dr. R.P. Singh is vice chancellor of Sri Satya Sai University of Technology and Medical Sciences, Sehore, Madhya Pradesh. He is former Director and Prof. Electronics and Communication at Maulana Azad National Institute of Technology, (MANIT) Bhopal. Dr. Singh Graduated and Post Graduated in Electronic Engineering from Institute of Technology (now IIT), B.H.U. Varanasi in 1971 and 1973, respectively. He did his Ph.D. from Barakatullah University Bhopal in 1991. He has 39 years of teaching, research, and administrative experience in Maulana Azad College of Technology (MACT)/MANIT out of which 22 years as Professor.