International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 29th Aug 2020. Link :http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-08

Title **Object classification and detection in multimedia using Tubelet Rendering Algorithm**

Volume 09, Issue 08, Pages: 155-162.

Paper Authors

Thella Sunitha, Dr. G. Lavanya Devi

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Object classification and detection in multimedia using Tubelet Rendering Algorithm

Thella Sunitha[1] , Dr. G. Lavanya Devi[2]

[1]Research Scholar, Department of Computer Science and System Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, Andhra Pradesh, India
sunikiranch@gmail.com

[2]Assistant Professor, Department of Computer Science and System Engineering, Andhra University College of Engineering(A), Andhra University, Visakhapatnam, Andhra Pradesh , India
lavanyadevig@yahoo.co.in

Abstract: Classification of images, audios and videos is tedious task. Object detection, a subdivision of computer perception, dynamic method for finding the objects in the image with respect to the background. Deep Learning is most widely used domain to perform object detection, text extraction from audio files, and object detection in video file. Previously, there are many existing methodologies that can be only object detection. Region-Based Convolutional Neural Networks (R-CNN) is the traditional deep learning techniques for addressing object localization and recognition tasks, designed for model performance. Classification of images can predict only one object in an image. Object localization is known as finding the location of one or more objects in an image and drawing abounding box around their extent. Object detection combines these two tasks and localizes and classifies one or more objects in an image. In this paper, an Optimized CNN with tubelet Rendering Algorithm is introduced for images, audio and video classification. Results show the performance of the proposed methodology.

Keywords: Deep Learning (DL), R-CNN, object detection, object localization.

Introduction

From the past many years, object detection is most active area of research in computer science. The aim of object detection is to find out the objects that are specifying within the image. Objects such as human, cars, keys, cups etc. Object detection can solves the complex tasks such as image segmentation, understanding the scene, tracking of object, captioning of image, detecting events, gender classification etc. Object detection supports all the applications such as camera surveillance, video surveillance, and security. Deep Learning represents the most powerful techniques for learning feature representations dynamically from any types of data. Some of the techniques have provided the major improvements in object detection. Object detection from videos is more complex for various ML algorithms, because the frame rate should be accurate to detect the objects in videos.

Traditional object detection system is done for static images and these are based on these networks which consist of three main stages. Bounding box is proposed to generate from the input image based on the every location consists of object as interest. The features that are present are extracted from every box to classify them as one of the object classes. Based on the bounding boxes and their related class scores are precise by post-processing techniques (e.g., Non-Maximal

Suppression) to extract the final detection results. Multiple frameworks, such as Fast R-CNN [1] and Faster R-CNN [2], followed this research direction and eventually formulated the object detection problem as training end-to-end deep neural networks.

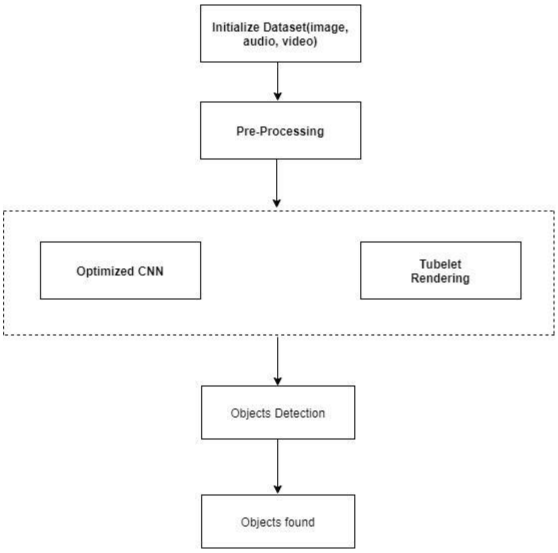


Figure 1: Functional diagram for object detection using Tubelet rendering

Object detection in videos is very difficult task because, detecting the accurate objects and finding the object based on the spatial information. Most of the challenges that does not exploit the temporal information available in videos to find challenges such as motion blur, occlusions or changes in objects appearance at certain frames.

In this paper, the enhanced Object detection frameworks consists of two main tasks: bounding box regression and object classification. The bounding box regression in a given frame is huge related with the spatial data available in that frame. However, the appearance of one object in previous frames might provide valuable information to classify the object in the present frame. This brings about the problem of how detections in different frames are linked and how the system aggregates this spatio-temporal information.

Literature Survey

The following are various existing object detection algorithms on static image, audio and video are discussed below.

The two stage techniques that are first developed by R-CNN [3]. This strategy takes a pre-determined object method set and then applies a deep CNN to get rid of per locale highlights to perform classification object. This strategy was improved in Fast R-CNN [4], adding a RoI pooling layer that allows to run a for each image CNN instead of per region. This work likewise modifies the header network to calculate classification and bounding box regression to filter the proposals. Thusly, all spine calculations are often reused, increasing the execution time. All of those techniques depend upon a neighborhood proposition technique autonomous of the organization. This issue is attended in Faster R-CNN [5], initializing a Region Proposal Network (RPN) done the object classification and bounding box regression with the same features are generated. The R-FCN object identifier [6] re-actualizes the network header avoiding the completely associated layers utilized by past work. All things being equal, it follows a totally convolutional approach changing the RoI pooling by an edge delicate RoI pooling.

Two-stream networks, for instance , [7], [8], [9] or [10] became the quality methodology in real world acknowledgment. One among the branches measures video outlines, while the opposite one takes pre-registered thick optical stream outlines as info. Diba et al. [11] proposes a end to end model able to extricate features that are temporal in deduction time via training the network with optical stream images. In spite of the very fact that activity acknowledgment may be a connected issue, the benefits of adding optical stream data to spatio-temporal object detectors might not be correct. To possess the choice to acknowledge some activity classes, for instance , "plunking down" and "getting up", movement data given by optical stream could also be urgent. This is not so obvious in object detection. This will be seen in [12], that utilizes an identical engineering for object identification and activity acknowledgment, showing how state-of-art fusion techniques work for actions but not for objects.

Tracking of Articles (MCMOT) [13] accomplishes 75.5% mAP, joining two detectors of objects with multi object following (MOT) procedures. The ILSVRC2016 champ [14] includes a 3 stage course R-FCN with a relationship tracker and setting surmising. they will improve the exactness up to 81.2% utilizing multi-scale testing and outfits. Our framework beats every past technique (78.2% mAP) with a solitary model execution.

Our network is ready start to end with no pre-registered recommendations, for instance , those utilized in [15]. That technique reuses the proposition set from [16], which actualizes a two-stage course RPN with multiscale testing adding to the proposition set those determined by the methodology attended in [17]. Likewise, that technique depends on the R-FCN system utilizing convolutional highlights to perform object identification and following all the while. This technique prompts 79.8% mAP.

Tang et al. [18] developed a brief tubelet identification system to acknowledge tubelets with fleeting covering. At that time , given two tubelets, they join those tubelets investigating the spatial cover between jumping encloses having an area with each tubelet the essential casing. They play out a multi-scale preparing and testing to support the accuracy to 80.6% mAP. This cannot be straightforwardly contrasted and our outcomes since we just test our framework with single scale pictures, making a more practical genuine testing environment.

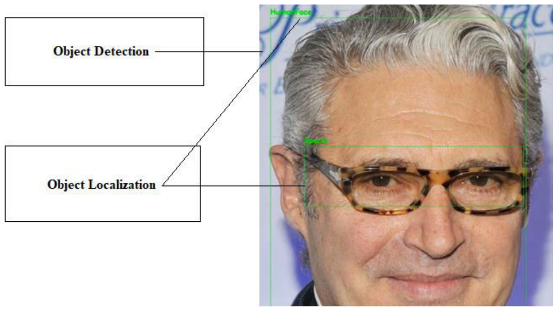


Figure 2: Objection detection using Optimized CNN

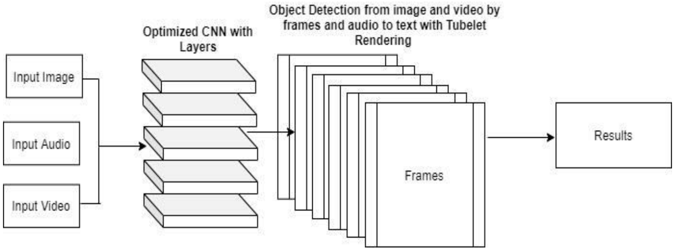

Object Detection from image and video by frames and audio to text with Tubelet Rendering
Optimized CNN with Layers
Input Image
Input Audio
Input Video
Frames
Results

Figure 3: Optimized CNN with Tubelet Rendering

Extended based object linking

The Extended based object linking (EBOL) focuses on detections in network and methods that are extended object tubes. In action recognition, the linking network detections are done within the time to find single action/object occasion that has become a quality approach in action recognistion [30][31] and object detection[27],[23],[26]. All these techniques try to join the every single boxes or small tubelets into larger tubes and this will convert into video. The EBOL, detect the every object in every frame, if there are number of objects detected these are called as candidates. Within the frame to frame linking, the optimization problem identifies the assembled lining score in every tubelet. Network errors such as false negatives or missorted detections are try to break large tubes and it is not possible to find a detection to link in some frames. Thus all the video frames maintain linking and detect the objects in every frame.

The implementation is done with every detection is represented as $\mathrm{d}_t^i = \{\mathrm{a}_t^i, \mathrm{b}_t^i, \mathrm{c}_t^i, \mathrm{d}_t^i, \mathrm{e}_t^i\}$ in the set D_t which is indexed by i in the frame t, the box is collected at $(\mathrm{a}_t^i, \mathrm{b}_t^i)$ represents the axis coordinates. c_t^i as width, d_t^i as height, and the confidence of associated classification e_t^i for the object class. B represents the lower threshold that detects the set D_t. The proposed algorithm prevents the less confidence detections which negatively affect the tube creation. The linking score is calculated as $ls = (d^i, d^j)$ between two detections $d^i and\ d^j$ at various frames $t\ and\ t'$ is described as

$$ls = \left(d_t^i, d_t^j\right) = \mathrm{e}_t^i + \mathrm{e}_{t'}^j, +IoU\left(d_t^i, d_t^j\right) \quad (1)$$

The optimization problem is solved by applying the Viterbi algorithm:

The optimal tube represents as $\vartheta = \underset{v}{argmax} \sum_{t=2}^{T} ls\ (D_{t-1}, D_t)$ (2)

Optimized CNN with Tubelet Rendering Algorithm

The tubelet rendering algorithm (TRA) that works with Feature Pyramid Network (FPN) models that deals with multiple Region Proposal Networks (RPN) and extracting RoI features at various pyramid levels. From the starting of this research, this is the first spatio-temporal framework with a high level determination such as FPN. The proposed system focuses on generating the frames for the input video. Generating the frames and rendering the frames can be done by TRA. After creating the frames N, the object detection is done and highlighted with the bounding box. The proposed method summarizes the data in order to generate an output feature map with the same size as if the network was working with a single frame. With this, the processing time for every frame is reduced. The algorithm overcomes the issues by using Equation (2) to identify all the tubes that ends with frame i.

Algorithm-1 for creating tubes per frames:

```
Input: Object Detection for every frame
Input: All possible tubes: V
Output: All the object tubes.
Step 1: $V \leftarrow \emptyset$
Step 2: For i in T, ….2 do
Step 3: While $D_i \neq \emptyset\ do$
Step 4: v ← $\mathrm{argmax}_\mathrm{v} \sum_{\mathrm{t}=2}^{\mathrm{i}} \mathrm{ls}(\mathrm{D}_{\mathrm{t}-1}, \mathrm{D}_\mathrm{t})$
Step 5: $D \leftarrow D/v$
Step 6: $V \leftarrow V \cup v$
```

Tube rendering algorithm with object detection

```
Input: Per frame detection set
        $D = \{D_t = (d_t^t, \ldots\ldots, d_t^{nt})\}_{t=1}^{T}$
Input: Tubelet set $T = \{t_i = (b_i^1, \ldots.\, b_i^N)\}_{i=1}^{\theta}$
Input: Object tubes $V = \{v_i = (d^{i,1}, \ldots.., d^{i,m_i})\}_{i=1}^{\delta}$
Output: Merged object tubes $V$
Step 1: $\widetilde{\mathrm{V}} \leftarrow \overline{\mathrm{V}}$
Step 2: for $\mathrm{v_i}$ in $\overline{\mathrm{V}}$ do
Step 3: for $\mathrm{v_j}$ in $\overline{\mathrm{V}}$ do
Step 4: $ts_{max} = 0$
Step: 5 for $\mathrm{t_l}$ in T do
Step: 6 $if\ \exists b_l^k \in\ t_l \mid \gamma\left(b_l^k, d^{i,m_i}\right)$and
            $\exists b_l^{k'} \in\ t_l \mid \gamma\left(b_l^{k'}, d^{j,m_j}\right)$and
            time $(d^{i,m_i}) >$ time $\left(d^{j,m_j}\right) then$
Step: 7          if $ts\ (t_l) >\ ts_{max}\ then$
Step: 8        $ts_{max} = ts(t_l)$
Step: 9          $C_{ij} = ts_{max}$
Step: 10 $H \leftarrow Hungarian(C)$
Step: 11 $for\ h_i in\ H\ do$
Step: 12 $\tilde{V} \leftarrow\ \tilde{V} \backslash \tilde{v}_{h_i}$
Step: 13 $\tilde{v}_i \leftarrow\ \tilde{v}_i\ \cup\ \tilde{v}_{h_i}$
Step: 14 $for\ \tilde{v}_i in\ \tilde{V}\ do$
Step: 15 Update Scores $(\tilde{v}_i)$
```

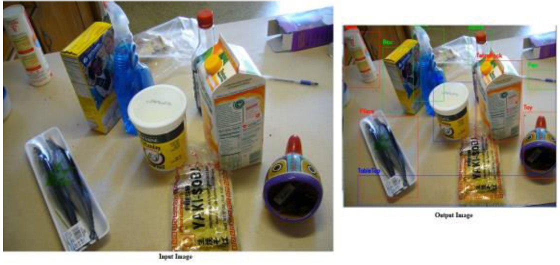


Figure 4: The proposed output for static image detecting the objects in the image.

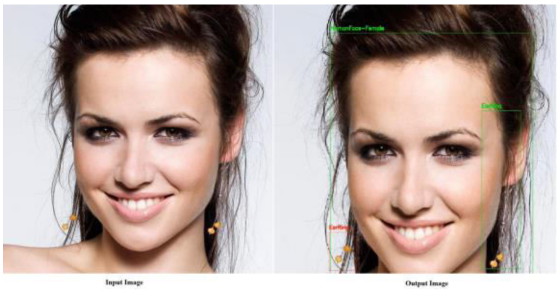
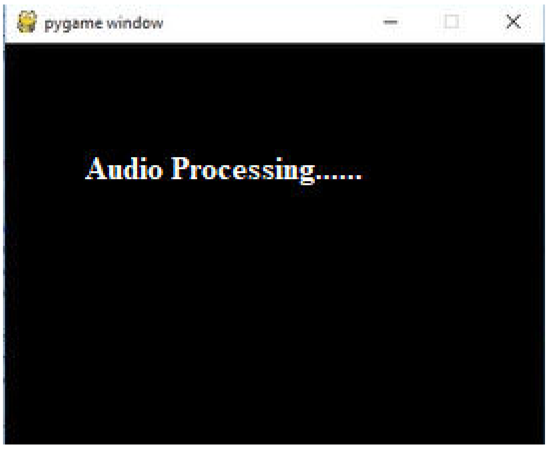
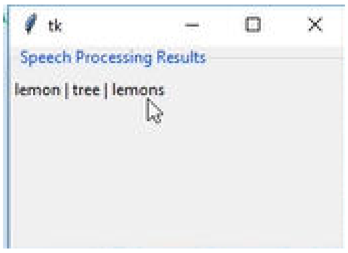


Figure 5: the proposed tubelet rendering to detect the objects and gender classification



Audio Processing......

Figure 6: Audio file processing for text recognition



Speech Processing Results

lemon | tree | lemons

Figure 7: speech recognition using proposed tubelet rendering

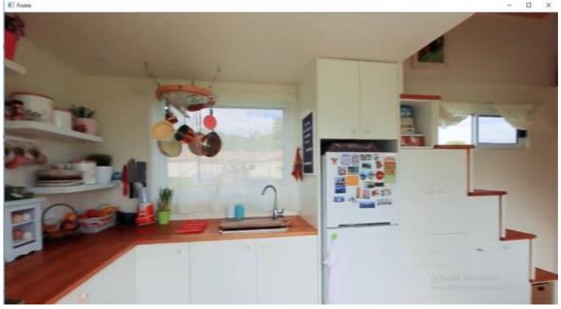


Figure 8: input video for object detection

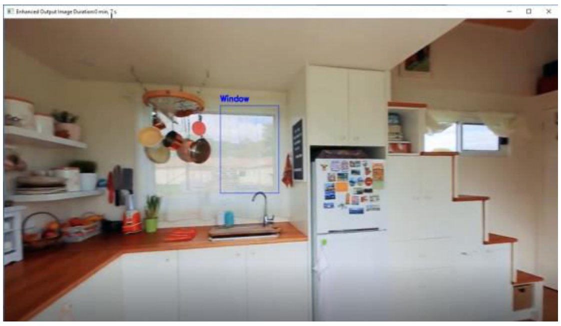


Figure 9: object detection for input video

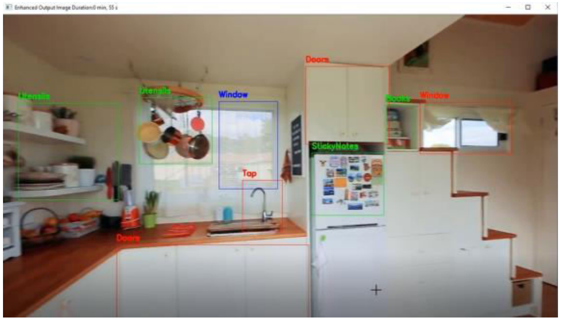


Figure 10: all the objects detected for input video

Performance metrics:

The performance of the tubelet rendering is calculated for image, audio and video. The overall performance is calculated by False Positive Rate, False Negative Rate, Sensitivity, Specificity and Accuracy; all these are calculated based on the obtained output. The basic count values such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used by these measures.

False Positive Rate (FPR)

The percentage of total number of objects detected.

$$\mathbf{FPR} = \frac{\mathbf{FP}}{\mathbf{FP + TN}}$$

False Negative Rate (FNR)

The total number of objects that are not detected.

$$\mathbf{FNR} = \frac{\mathbf{FN}}{\mathbf{FN + TN}}$$

Sensitivity

This is the parameter that correctly finds the detected objects, text in audio, objects in every frame of the video.

$$\mathbf{Sensitivity} = \frac{\mathbf{No.\,of\ TP}}{\mathbf{No.\,of\ TP + No.\,of\ TN}}$$

Specificity

This is the parameter that correctly finds the undetected objects, text in audio, objects in every frame of the video.

$$\mathbf{Specificity} = \frac{\mathbf{No.\,of\ TN}}{\mathbf{No.\,of\ TN + No.\,of\ FP}}$$

Accuracy: This will calculate the overall accuracy of the object detection

$$\mathbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

Experimental Results

The experimental results are conducted on synthetic dataset which is collected from various UCI repository sources. The dataset consists of 50 images, 15 audio files and 10 video files for classification and object detection, object localization. Pyhton is used as programming language.

	R-CNN	Optimized CNN	Optimized CNN + Tublet Rendering
Sensitivity	78.23	83.21	96.56
Specificity	78.98	85.34	95.67
Accuracy	74.12	78.12	97.32
Time (Sec)	15	9	4

Table 1: the overall performance of the existing and proposed methodologies for video frames classification

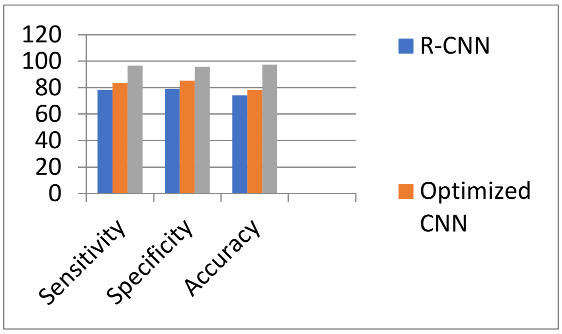


Figure 11: performance computation parameters for object detection in videos

	R-CNN	Optimized CNN	Optimized CNN + Tublet Rendering
Sensitivity	79.12	80.32	97.32
Specificity	79.67	86.12	96.12
Accuracy	78.32	78.87	97.56
Time (Sec)	9	5	3

Table 2: the performance of existing and proposed methodologies for image object detection

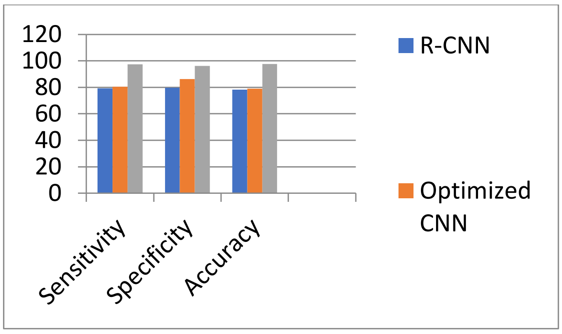


Figure 10: performance computation parameters for image object detection

Table 3: the performance of existing and proposed methodologies for single audio file to detect the text.

	R-CNN	Optimized CNN	Optimized CNN + Tublet Rendering
Sensitivity	67.21	68.12	91.32
Specificity	66.54	67.65	90.12
Accuracy	59.12	60.21	88.67
Time (Sec)	7	5	3

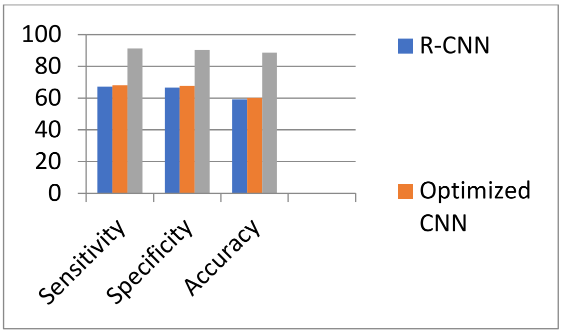


Figure 11: performance computation parameters for audio file to text detection.

Conclusion

This paper focuses on developing the spatio-temporal data to increase object detection accuracy in videos, images, and audios. The proposed methodology effectively detects the objects in the still images, videos and finding the text in the audios. This is the hybrid model because it is merged with optimized CNN with tublet rendering algorithm. Object localization is also one of the advantage to detect the location of the object with bounding box representation.

References

[1] R. Girshick. Fast r-cnn. ICCV, 2015.

[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS, 2015.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, 'Rich feature hierarchies for accurate object detection and semantic segmentation', in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587, (2014).

[4] Ross Girshick, 'Fast R-CNN', in IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, (2015).

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster RCNN: Towards real-time object detection with region proposal networks', in Advances in Neural Information Processing Systems (NIPS),pp. 91–99, (2015).

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, 'R-FCN: Object detection via region-based fully convolutional networks', in Advances in Neural Information Processing Systems (NIPS), pp. 379–387, (2016).

[7] Joao Carreira and Andrew Zisserman, 'Quo vadis, action recognition? a new model and the kinetics dataset', in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299–6308, (2017).

[8] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes, 'Spatiotemporal residual networks for video action recognition', in Advances in Neural Information Processing Systems (NIPS), pp. 3468–3476,(2016).

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, 'Convolutional two-stream network fusion for video action recognition', in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1933–1941, (2016).

[10] Karen Simonyan and Andrew Zisserman, 'Two-stream convolutional networks for action recognition in videos', in Advances in Neural Information Processing Systems (NIPS), pp. 568–576, (2014).

[11] Ali Diba, Ali Mohammad Pazandeh, and Luc Van Gool, 'Efficient twostream motion and appearance 3D CNNs for video classification', arXiv preprint arXiv:1608.08851, (2016).

[12] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid, 'Joint learning of object and action detectors', in IEEE International Conference on Computer Vision (ICCV), pp. 4163–4172,(2017).

[13] Byungjae Lee, Enkhbayar Erdenee, Songguo Jin, Mi Young Nam,Young Giu Jung, and Phill Kyu Rhee, ‘Multi-class multi-object tracking using changing point detection’, in European Conference on Computer Vision (ECCV), pp. 68–83, (2016).

[14] Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, and Jiankang Deng. ILSVRC2016 object detection from video: Team NUIST.

[15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, ‘Detect to track and track to detect’, in IEEE International Conference on Computer Vision (ICCV), pp. 3038–3046, (2017).

[16] Xingyu Zeng, Wanli Ouyang, Junjie Yan, Hongsheng Li, Tong Xiao,Kun Wang, Yu Liu, Yucong Zhou, Bin Yang, Zhe Wang, et al., ‘Crafting GBD-Net for object detection’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(9), 2109–2123, (2017).

[17] Spyros Gidaris and Nikos Komodakis, ‘LocNet: Improving localization accuracy for object detection’, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 789–798, (2016).

[18] Peng Tang, Chunyu Wang, Xinggang Wang, Wenyu Liu, Wenjun Zeng,and Jingdong Wang, ‘Object detection in videos by high quality object linking’, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2019).