Spatio-temporal activity detection and recognition in untrimmed surveillance videos

Konstantinos Gkountakos
ITI - CERTH
Thermi-Thessaloniki, Greece
gountakos@iti.gr

Despoina Touska
ITI - CERTH
Thermi-Thessaloniki, Greece
destousok@iti.gr

Konstantinos Ioannidis
ITI - CERTH
Thermi-Thessaloniki, Greece
kioannid@iti.gr

Theodora Tsikrika
ITI - CERTH
Thermi-Thessaloniki, Greece
theodora.tsikrika@iti.gr

Stefanos Vrochidis
ITI - CERTH
Thermi-Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
ITI - CERTH
Thermi-Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

This work presents a spatio-temporal activity detection and recognition framework for untrimmed surveillance videos consisting of a three-step pipeline: object detection, tracking, and activity recognition. The framework relies on the YOLO v4 architecture for object detection, Euclidean distance for tracking, while the activity recognizer uses a 3D Convolutional Deep learning architecture employing spatio-temporal boundaries and addressing it as multi-label classification. The evaluation experiments on the VIRAT dataset achieve accurate detections of the temporal boundaries and recognitions of activities in untrimmed videos, with better performance for the multi-label compared to the multi-class activity recognition.

KEYWORDS

Activity detection, activity recognition, spatiotemporal boundaries detection, 3D-convolutional neural networks

ACM Reference Format:
Konstantinos Gkountakos, Despoina Touska, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2021. Spatio-temporal activity detection and recognition in untrimmed surveillance videos. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460426.3463591

1 INTRODUCTION

Activity recognition has been widely explored in computer vision problems when using trimmed videos [4, 5, 9, 17], while the temporal localization and activity recognition in untrimmed video footage has been less studied and thus, remains a big challenge [12, 14]. The main challenges stem from the visual footage's untrimmed nature, the number of involved activities, the varying length of co-occurring activities, and the multiple actors/objects involved within each activity. Moreover, the interaction between/among objects and the diversity of the spatial regions where each activity occurs, compared to the camera's projection plane, is also a challenge in itself. These issues are further compounded in the context of real-life applications, such as surveillance systems, that involve multiple similar objects of interest (e.g. trucks, cars, and vans) that are independently detected, while some of the detected objects of interest (e.g., parked cars) may not actually participate in the target activities, and thus should be removed from consideration.

Spatio-temporal activity detection and recognition in such real-life applications is a complex problem and typically involves a pipeline of several computer vision components executed in sequence [12, 14]. First, an object detection component analyzes the untrimmed video footage to identify objects of interest. An object tracker then follows, intending to retain the track of the objects identified in the previous step. Finally, the activity localization and recognition component considers the object tracking information to predict the spatio-temporal boundaries and classify the latter to specific categories of activities.

This work presents a spatio-temporal activity detection and recognition pipeline for untrimmed surveillance videos focusing on human- and vehicle-related activities. The framework relies on the YOLO v4 [2] architecture for recognizing the detected objects, uses the Euclidean distance for the tracking process, and finally employs a 3D-Convolutional Neural Network (3D-CNN) architecture [7], in particular, a 3D-ResNet with 50 layers, for recognizing the target human- and vehicle-related activities.

Specifically, this work considers activity recognition as a multi-label, rather than a multi-class problem, to effectively recognize concurrent activities performed by the same object. To this end, it proposes the use of spatio-temporal boundaries generated by the union of the individual bounding boxes of the detected objects over a time window. To the best of our knowledge, this is the first time that such spatio-temporal boundaries are considered in activity recognition, both by multi-label and multi-class approaches. Moreover, this work fine-tunes the object detection as part of the pipeline to reduce the amount of post-processing required for the removal of specific instances of the objects of interest that should not be considered by the activity recognition (such as static objects) according to the annotations of the dataset, as well as for the merging of similar objects (such as trucks, cars, and vans merged into the 'vehicle' class). The enhanced framework is evaluated using the

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463591

VIRAT dataset [13] and the experimental results indicate accurate detections of the temporal boundaries and recognitions of activities on processing untrimmed videos.

2 RELATED WORK

2.1 Object Detection & Tracking

There has been a plethora of object detection methods introduced during the past decade [11, 16], but only a small minority has reported real-time performance in object recognition and tracking when executed in real operational scenarios. Tan et al. [18] proposed a real-time object detector, named EfficientDet, that focuses on identifying specific objects efficiently by using a deep neural network architecture that consists of a weighted bidirectional design topology that enables the learning of the "importance" features at different scales. Moreover, Bochkovskiy et al. proposed YOLO v4 [2], the upgraded version of YOLO v3 [15], which enables fast and accurate training processes using common GPUs. They also verified the importance of utilizing Bag-of-Freebies and Bag-of-Special approaches for object detection. EfficientDet [18] and YOLO v4 [2] approaches indicate a trade-off between time-efficiency and accuracy compared to the rest of state-of-the-art approaches.

2.2 Activity Detection & Recognition

Gao et al. [6] proposed the TURN TAP overlapping sliding window approach. Calculation time was decreased without significantly affecting the accuracy of temporal localization. By considering the novel characteristics of Faster-RCNN [16], they transformed the problem of boundary detection to a problem of boundary regression. Each video is divided into 16-frames length clips and the features are extracted using a 3D-CNN, namely the 3CD network [19]. Finally, 1-D feature vectors are computed to calculate the temporal coordinates offset and the confidence score, similarly to Faster-RCNN. Lin et al. [10] proposed the Boundary-Sensitive Network (BSN) architecture that proved efficient for both short and long duration activities, as it firstly generates possible boundary points and selects the points that have been predicted with higher scores. Subsequently, starting and ending points are combined to select the activities' boundaries and generate the sequence for each activity, especially the start, mid, and the end parts.

Rana et al. [14] proposed a real-time method that comprises three stages: the detection of activities tubelets, their classification, and, finally, their merging using the Tubelet-Merge Action-Split (TMAS) algorithm to generate spatio-temporal detections with high speed. Liu et al. [12] proposed a method that first generates video proposals by applying object detection and tracking, then classifies the proposals' features and eliminates inaccurate events, and lastly fuses the predictions. However, these methods do not consider the spatio-temporal boundaries of an object which includes the entire field of action during its trajectory.

3 SPATIO-TEMPORAL ACTIVITY DETECTION AND RECOGNITION

This work presents a spatio-temporal activity detection and recognition framework that comprises the following three-stage pipeline: detection of objects of interest, object tracking, and activity recognition. Based on this sequential framework, the desired activities

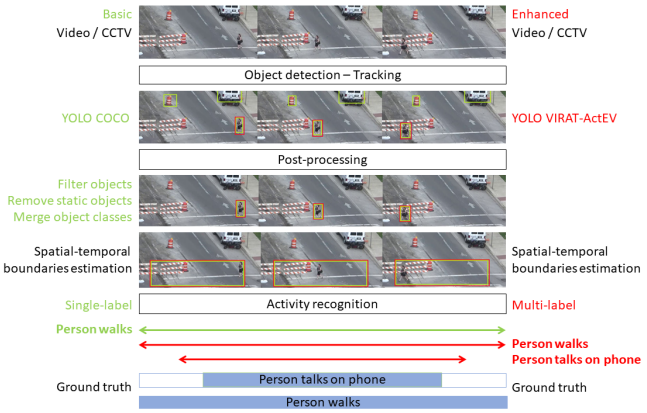

Figure 1: Illustration of the pipeline for the basic (left) and enhanced configuration of the framework (right).

can be identified in untrimmed video footage by exploiting the extracted object boundaries and by tracking their course. We propose that the estimated bounding boxes comprise both temporal and spatial object boundaries, where "temporal" refers to the start and ending timestamps of the activities, while "spatial" refers to the localization of the activities within the acquisition camera's projection plane and field of view.

We first consider a basic configuration of the framework (denoted as "B") as follows. **For object detection (B-OD)**, a deep CNN architecture, namely YOLO v4, is employed for (near) real-time object identification. In particular, a pre-trained model under the COCO dataset is considered as it involves target objects, such as vehicles and persons, that are relevant to the activities of interest. **For object tracking (B-OT)**, a frame-by-frame comparison technique is deployed where the Euclidean distance is exploited as the metric among the detected objects in two subsequent frames. Thus, adjacent objects between two frames are considered to correspond to the same object when their distance satisfies a threshold. Since the COCO dataset considers various classes of objects, further to the person and vehicle classes of interest, post-processing steps to filter any redundant detections should be introduced. These steps are summarized as the filtering of the target objects, the exclusion of the static ones, and the merging of similar classes as trucks, buses, and cars to one unified 'vehicle' class. At this step, i.e., during the tracking of each specific object, the spatio-temporal boundaries are also generated by the union of the separate bounding boxes. **For activity recognition (B-AR)**, a 50-layer 3D-ResNet [7] architecture is used. The architecture consists of four sequential bottleneck blocks, three 3D-convolution (with variant kernel sizes), batch normalization and ReLU activation layers, while the temporal dimension is set equal to 16. As the initial step, the weights of the Kinetics dataset [4] were pre-loaded, a common practice in the activity recognition problem [7]. This architecture involves a single-label assignment mechanism to each identified trimmed activity.

An enhanced configuration of the framework (denoted as "E") improves the basic approaches as follows. **For object detection (E-OD)**, the YOLO v4 model is fine-tuned for 20 epochs to target only the detection of vehicle and person objects; as a result, we avoid the time-consuming post-processing steps as the generated predictions are more accurate, outputting the desired objects' types.

For object tracking (E-OT), the Euclidean distance as well as the calculation of the spatio-temporal boundaries, which are generated by the union of the separated bounding boxes, are used, similarly to the basic approach. Figure 1 illustrates with green the post-processing steps of object filtering that are skipped. Specifically, the object filtering post-processing step is avoided as the object detector returns only labels of objects related to vehicles and humans. Also, as the various vehicle-related objects such as cars, trucks, vans have been merged into a single object label 'vehicle', the post-processing step that merges objects classes is skipped. Finally, the post-processing step of removing static objects, especially parked vehicles, is avoided as the enhanced framework deals with detecting vehicles with no occlusions and captured by multivariate poses during detection and tracking.

For **activity recognition (E-AR)**, the main difference between the basic and the enhanced approach is the transformation of the multi-class problem to a multi-label one. Specifically, a binary vector is generated for each detected object, sized equal to the number of frames during its tracking. Zeros indicate activities non-occurrence, while ones indicate activities occurrence. Thus, given a sequence of 16-frames, multi-label annotations are learned for each calculated spatio-temporal activity boundary. Regarding the recognizer, the same architecture to the basic has been incorporated by replacing the categorical cross-entropy loss with the weighted binary cross-entropy loss function to transform the problem into a multi-label objective and deal with unbalanced datasets.

4 EXPERIMENTAL EVALUATION

4.1 Datasets

Several datasets are available for evaluating activity recognition, such as THUMOS14 [8], ActivityNet 1.3 [3], and Kinetics-700 [4]. These datasets consider trimmed videos and thus do not need temporal boundaries annotations. Although there are some datasets with temporal boundaries annotations in long untrimmed videos, such as the MEXaction2 [20], these only include a small number of activity classes; e.g., MEXaction2 considers only two classes.

As the aforementioned datasets are not appropriate for the discussed problem, we considered part of the VIRAT [13] dataset related to a CCTV surveillance system, namely the VIRAT-ActEV dataset that is annotated for both activities and objects by the National Institue of Standards and Technology (NIST) in the scope of the Activities in Extended Videos (ActEV) TRECVID challenge[1].

Three sub-sets for training, validating, and testing compose the dataset, with annotations provided only for the training and validation sets. The test set of the dataset is not accompanied by annotations, as it constitutes material for upcoming challenges. The training set consists of 64 videos that describe 4311 activities samples, while the validation set comprises 54 videos with 3521 activities samples. The unique number of activities is equal to 35, with 19 being human-related, 6 being vehicle-related, and the rest 10 describing human-vehicle interactions. The field of view of the cameras used is extremely-wide and is characterized by high resolution (e.g. 1920 × 1080), while the annotations include varying length activities, human and vehicle interaction, and the simultaneous occurrence of activities.

4.2 Experimental setup

For **enhanced object detection (E-OD)**, the YOLO v4 detector was fine-tuned using the VIRAT-ActEV dataset. In particular, due to the lack of publicly available annotations of its testing set that prevents any experiment outside the scope of TRECVid challenge, the training set of VIRAT-ActEV was split into three sub-parts, with the commonly used 60:30:10 ratio, to train, validate, and test the performance of the fine-tuned model. For **activity recognition (B-AR and E-AR)**, the original training and validation sets of the VIRAT-ActEV are used for training and validation/testing, respectively.

4.3 Evaluation Metrics

For evaluating object detection, the mean Average Precision (mAP) at different levels of Intersection over Union (IoU) is used. Evaluation measures for activity detection and temporal localization include the Probability of Missed Detection (Pmiss) and the Time-based false alarm (Tfa) measures. Pmiss (Equation 1) is the fraction of reference activity instances not detected by the system. Tfa (Equation 2) is the fraction of non-activity instance time as indicated by the ground truth references, for which the proposed system falsely identified an instance.

In this work, the primary measure for activity detection and recognition will be the normalized calculation of the partial Area Under the Detection Error Tradeoff (DET) Curve (nAUDC) from 0 to a fixed Tfa value a, denoted nAUDCa (Equation 3). The partial nAUDC is computed separately for each activity. A system is scored on both of these, measured by Pmiss and Tfa. The use of multiple thresholds creates the DET curve. The final evaluation summarises the DET curve: the nAUDC across the Tfa ranges between [0, a] and normalizes the value to [0:1]. In brief, less errors are reflected with lower values for the aforementioned metrics.

$$P_{miss}(x) = \frac{N_{md}(x)}{N_{TrueInstance}}, \tag{1}$$

where $N_{md}(x)$ = The number of missed detections at the confidence score threshold that results in $Tfa(x)$, $N_{TrueInstance}$ = The number of true instances in the sequence of reference, and $P_{miss}(x)$ = The probability of missed detections (instance-based) value for $Tfa(x)$.

$$T_{fa} = \frac{1}{NR} \sum_{i=1}^{N_{frames}} max(0, S_i^{'} - R_i^{'}) \tag{2}$$

where N_{frames} = The duration (frame-based) of the video, NR = The duration of the video without the target activity occurring, $S_i^{'}$ = The probability of missed detections, and $R_i^{'}$ = The total count of reference instances for frame i.

$$nAUDC_a = \frac{1}{a} \int_{x=0}^{a} P_{miss}(x)dx, x = T_{fa} \tag{3}$$

4.4 Evaluation Results

For evaluating both configurations of the framework, the following experiments were performed.

1. For object detection: (B-OD) vs (E-OD)

Figure 2 illustrates the loss of the training set along with the mAP of the validation set during the training of the fine-tuned

(E-OD) classifier on the VIRAT-ActEV dataset. As it is presented, the training loss is reduced as the step number is increased, while the mAP increases, confirming the model's successful fine-tuning.

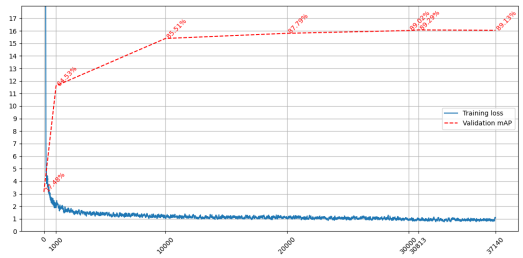

Figure 2: Loss and mAP values for the training and validation sets, respectively, during object detector training.

We evaluate both the native object detector (B-OD) trained using the COCO dataset and the fine-tuned (E-OD) using the VIRAT-ActEV dataset. For region proposals, we set the confidence threshold to 0.25 (as suggested by the YOLO v4 developers[1]) and report the mAP with different IoU thresholds in Table 1. The best mAP is achieved by the fine-tuned object detector; however, the main difference between the models relies on the irrelevant detections that B-OD inserts, thus increasing the number of false positives.

Table 1: Object detection mAP for different IoU values.

	COCO	VIRAT-ActEV
mAP@0.5	17.68%	89.42%
mAP@0.6	11.59%	87.29%
mAP@0.7	5.76%	81.42%

2. For object tracking: (B-OD) & (B-OT) vs (E-OD) & (E-OT)

After finalizing the fine-tuning using the VIRAT-ActEV dataset, many post-processing steps can be ignored by the process. In particular, predictions can be performed on objects belonging to classes 'person' and 'vehicle' without the necessity of filtering and merging the objects excluded from the VIRAT-ActEV dataset. Also, Table 1 indicates highly accurate object detections, particularly in terms of *mAP*@0.50. Finally, no discarding of the static vehicles is required, as, in the fine-tuning process, the annotation consists only of non-static vehicles; hence, the predictions of the object detector for the class vehicle corresponds only to moving objects. This is due to the different background and captured angles of static/non-static vehicles and the absence of occlusions among the non-static objects.

To assess the activity localization performance, especially the temporal boundaries detection, pairs (B-OD) & (B-OT) and (E-OD) & (E-OT) were compared. The discussed evaluation metrics are used with the assumption that all activities belong to the same and only class "temporal evaluation boundaries". Therefore, we create a DET curve for the different threshold values. For different ranges of Tfa between 0 and 1, we calculate the partial nAUDC for all videos' temporal activity boundaries in the validation set of VIRAT-ActEV and their corresponding normalized values in that range.

The normalized values for every range are illustrated in Figure 3 where E-OD displays significantly lower values for the partial nAUDC, than the B-OD, especially in the Tfa range with a right limit at 0.25 until 0.85. This indicates the fewer errors of the fine-tuned object detector to proposing temporal boundaries, in contrast to the basic. Considering that the fine-tuned detector does not require any additional filtering, the results indicate an accuracy improvement, while the E-OD extracts faster estimations in inference mode.

[1]https://github.com/AlexeyAB/darknet

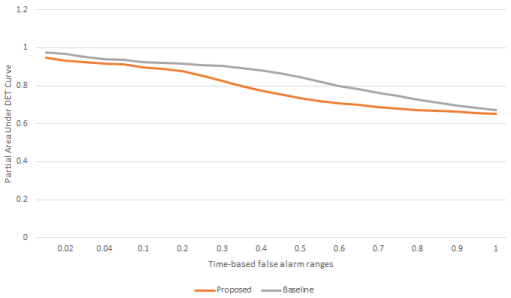

Figure 3: Temporal boundaries reported performance.

3. For activity classification: (E-OD) & (E-OT) & (B-AR) vs (E-OD) & (E-OT) & (B-AR)

An essential adaptation of of the framework was the conversion of the 3D-ResNet classifier to support multi-label predictions. Our experiments above showed that the (E-OD) & (E-OT) approach generates more accurate temporal boundaries. Hence, in order to perform a fair comparison we selected E-OD and tracker to predict the temporal boundaries to evaluate the effectiveness of the different (B-AR) and (E-AR) classifiers. For (E-AR), the activities predicted with a confidence score above different threshold values (60%, 70%, and 80%) were retained for evaluation; this was applied to objects involved in one or many activities.

As Table 2 shows, the results using multi-label predictions reflect higher effectiveness (lower nAUDC) compared to the multi-class approach. This indicates that multi-label classification captures an object's multiple activities simultaneously and more accurately than the multi-class approach. Moreover, the best performance is reported for the threshold with the lowest value.

Table 2: nAUDC for various threshold values.

Threshold	60%	70%	80%
Multi-class	0.8739		
Multi-label	**0.8240**	0.8373	0.8495

5 CONCLUSIONS

In this work, we address the problem of detecting and recognising activities in untrimmed surveillance videos by employing sequential modules consisting of an object detector and tracker accompanied by an activity classifier. For object detection and tracking, the YOLO v4 architecture and the Euclidean distance, are respectively, used, followed by the calculation of the spatio-temporal boundaries based on the union of the individual bounding boxes of the detected objects over a time window. For activity classification, a multi-class and a multi-label approach are examined. The detailed experimental evaluation indicates that the proposed improvements are advantageous. Future work includes the investigation of alternative ways to manage multiple resulting activities.

ACKNOWLEDGMENTS

This work was supported by the Horizon 2020 projects PREVISION (H2020-833115), CREST (H2020-833464), CONNEXIONs (H2020-786731) and funded by the European Commission.

REFERENCES

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. 2020. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[6] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE international conference on computer vision*. 3628–3636.

[7] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d CNNs retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[8] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[10] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[12] Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. 2020. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*. 126–133.

[13] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. IEEE, 3153–3160.

[14] Aayush Jung Rana, Praveen Tirupattur, Mamshad Nayeem Rizve, Kevin Duarte, Ugur Demir, Yogesh Singh Rawat, and Mubarak Shah. 2019. An Online System for Real-Time Activity Detection in Untrimmed Surveillance Videos.. In *TRECVID*.

[15] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[17] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

[18] Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10781–10790.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[20] Huifen Xia and Yongzhao Zhan. 2020. A Survey on Temporal Action Localization. *IEEE Access* 8 (2020), 70477–70487.