IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2019, accepted December 16, 2019, date of publication December 23, 2019, date of current version January 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961789

Rethinking Online Action Detection in Untrimmed Videos: A Novel Online Evaluation Protocol

MARCOS BAPTISTA-RÍOS[1], ROBERTO J. LÓPEZ-SASTRE[1], FABIAN CABA HEILBRON[2], JAN C. VAN GEMERT[3], F. JAVIER ACEVEDO-RODRÍGUEZ[1], AND SATURNINO MALDONADO-BASCÓN[1], (Senior Member, IEEE)

[1]GRAM, Department of Signal Theory and Communications, University of Alcalá, Alcalá de Henares 314100, Spain
[2]Adobe Research, Media Intelligence Lab, Deep Learning Group, San Jose, CA 95110, USA
[3]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 Delft, The Netherlands

Corresponding author: Marcos Baptista-Ríos (marcos.baptista@uah.es)

This work was supported in part by the Project PREPEATE, Spanish Ministry of Economy, Industry and Competitiveness, under Grant TEC2016-80326-R, and in part by the NVIDIA Corporation with the donation of a GPU.

ABSTRACT The Online Action Detection (OAD) problem needs to be revisited. Unlike traditional offline action detection approaches, where the evaluation metrics are clear and well established, in the OAD setting we find very few works and no consensus on the evaluation protocols to be used. In this work we propose to rethink the OAD scenario, clearly defining the problem itself and the main characteristics that the models which are considered online must comply with. We also introduce a novel metric: the Instantaneous Accuracy (*IA*). This new metric exhibits an *online* nature and solves most of the limitations of the previous metrics. We conduct a thorough experimental evaluation on 3 challenging datasets, where the performance of various baseline methods is compared to that of the state-of-the-art. Our results confirm the problems of the previous evaluation protocols, and suggest that an IA-based protocol is more adequate to the online scenario. The baselines models and a development kit with the novel evaluation protocol will be made publicly available.

INDEX TERMS Computer vision, deep learning, evaluation, instantaneous accuracy, online action detection.

I. INTRODUCTION

In this work, we focus on the problem of localizing actions in untrimmed videos *as soon as* they happen, which was coined as Online Action Detection (OAD) by De Geest *et al.* [1].

Action detection in video has been widely studied, but *mainly from an offline perspective*, *e.g.* [2]–[12], where it is assumed that all the video is available to make predictions. Few works address the *online* setting, *e.g.* [1], [4], [13], [14]. Think of a robotic platform that must interact with humans in a realistic scenario, or an intelligent video surveillance application designed to raise an alarm when an action is detected. *All* previous offline methods make the described applications impossible because they would detect action situations way later they have occurred.

On the contrary, in an OAD approach, action detections must be given over video streams, hence working with partial observations, where the action segments are possibly the exception rather than the rule, compared with the background. Moreover, this online definition allows for an important property: the anticipation to the action. In other words, for an OAD model the objective is to anticipate the action even before the action is fully completed.

However, there are important weaknesses among the online approaches in two fundamental aspects: 1) the evaluation metric; and 2) the treatment of the background category by the models and in the evaluation. Regarding the former, we have noticed that there is no consensus on the evaluation protocols. In each dataset a different metric is proposed for the very same problem. Moreover, used metrics cannot be said to be of an *online nature*. In other words, metrics such as the mean Average Precision (mAP) [15] or the Calibrated Average Precision (cAP) [1], do not provide information about the instantaneous performance of the solutions over time. They need to be computed entirely offline, accessing the whole set of action annotations in a given test video, to sort, for instance, all frame predictions. Even the novel point-level Action Start detection mAP metric, proposed in [16] to

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao-Yu Zhang.

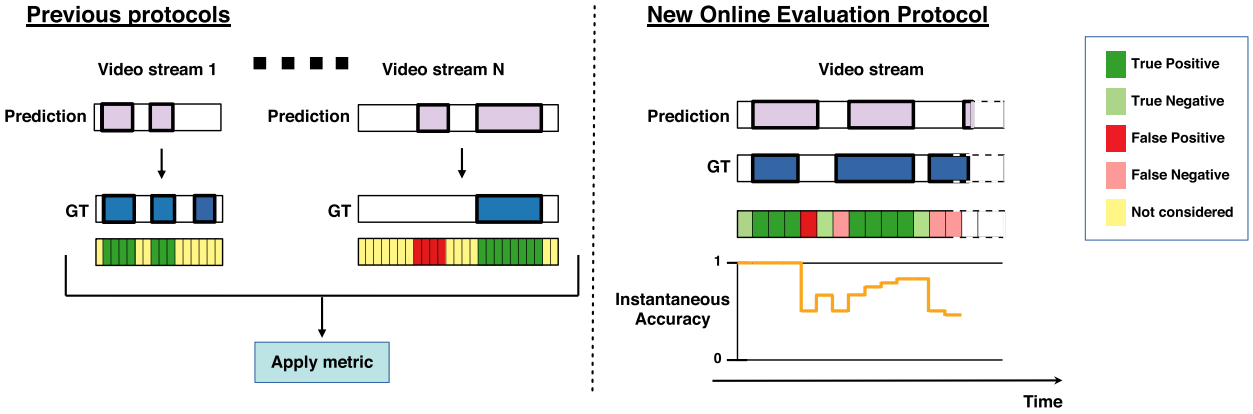

FIGURE 1. Online Evaluation Protocol. Previous evaluation protocols for Online Action Detection (OAD) were based on: 1) running the online methods through all videos; 2) applying the offline metric on the obtained results. Additionally, offline metrics proposed so far do not consider the background in their evaluation. We propose an Online Evaluation Protocol based on our new Instantaneous Accuracy metric (IA). OAD approaches are evaluated online considering the background and regardless of the length of the video.

evaluate the different problem of online detection of action start, has the same limitation.

Regarding the second aspect, the OAD setting is characterized by long untrimmed videos where actions appear sparsely and the background predominates. Consequently, the online problem should demand the background category to be treated as a first-class citizen. However, if we analyze the online methods published to date, almost all have been designed to cast a specific prediction for the background category: given a test video, every frame is categorized with an action class. For this reason, some propose to modify the evaluation metric, as in [1], where a calibrated version of the average precision is proposed to mitigate the penalty with the background frames. Furthermore, when the background class is not considered in the evaluation, but it is considered in the annotation, all the proposed metrics cannot saturate to the maximum which they have been designed for. In other words, the maximum of a precision-based metric will never be of 100% even if the method cast for every action frame the correct category.

In this paper we address all the described limitations. Our scientific contributions are as follows:

- First, we introduce an evaluation protocol, with a novel *online* metric: the Instantaneous Accuracy (*IA*) (see Figure 1). This metric has been designed not only to overcome the limitations, but to allow for fair and effective comparisons between OAD methods.
- Second, we propose a thorough experimental evaluation on three challenging datasets (Thumos'14 [17], TVSeries [1] and ActivityNet [18]), where a comparison between baselines and the state-of-the-art approaches is offered. The results show that an *IA*-based evaluation protocol is more adequate to the OAD problem, because it is able to offer a detailed evolution of the performance of OAD models when the video stream grows over time.
- We will publicly release the implementations of the baseline models as well as a development kit with the novel evaluation protocol.

II. RELATED WORK

We summarize here some contributions on related problems: offline action detection, early action detection and online action detection.

A. OFFLINE ACTION DETECTION

In offline action detection, the whole video is known beforehand and the goal is to detect when and where actions occur. There are works that apply classification on action proposals segments, *e.g.* [2]–[4], [8], [19], [20]. However, other works [5]–[7], [21] train models to directly detect action segments, without the proposal stage. All the previously mentioned works propose fully supervised approaches. Since it is more complicated each day to have labels for such big amount of videos, the community is exploring also weakly supervised alternatives [22]–[29]. In any case, our analysis focuses on the different problem of *online* action detection.

B. EARLY ACTION DETECTION

In this setting, the objective of the approaches (*e.g.* [25], [30], [31]) is to predict the action label of an action video before the ongoing action execution ends. They assume the video stream contains only one action instance and once the video has ended, they decide start and end frames. An Online Action Detection (OAD) scenario makes no assumption on the video and actions must be detected as soon as they happen. F1-score is used for evaluation, but this metric does not meet our online evaluation protocol conditions since it is a class-level metric and background is not considered.

C. ONLINE ACTION DETECTION

There exist few recent works on OAD [1], [13], [14], [32]. De Geest *et al.* [1] set the OAD conditions and introduced some frame-level baselines models which do not explicitly discriminate action from background. They also proposed two evaluation metrics: a per-frame mean Average Precision and a calibrated Average Precision. In their follow-up work [14], they designed a two-stream LSTM network to

capture better temporal dependencies. Li *et al.* [32] trained a method which predicts skeleton-based action classes (plus background) and regresses the start and end frames. This scenario is simpler and in the lack of an established evaluation protocol, they evaluate their method by adapting traditional offline action detection metrics. Gao *et al.* [13] present a LSTM-based Reinforced Encoder Decoder network which anticipates future frame labels and representations. As a side experiment, they address the OAD task as a special case of anticipation where the anticipation time is set to zero. Therefore this type of network cannot be considered as a pure OAD approach. In our work, we explain why the metrics proposed so far are not suitable for online evaluation in streaming videos and propose a new protocol. We also implement a simple method for OAD capable of explicitly distinguishing action and background. There is also the work of Shou *et al.* [16] which focuses on the problem of Online Detection of Action Start (ODAS). ODAS can be seen as a variant of OAD where only the starting point of actions is of interest. An OAD method must always find the start and end of an action. The OAD and ODAS evaluation protocols have in common that both use class-level metrics that are computed offline.

III. ONLINE EVALUATION PROTOCOL FOR ONLINE ACTION DETECTION

Despite the many practical applications Online Action Detection (OAD) offers, it has been barely explored. As the pioneer work of De Geest *et al.* [1] stated, OAD needs a solid definition and a strong evaluation protocol, which we revisit in this section.

A. ONLINE ACTION DETECTION

The established properties of the OAD task in realistic scenarios are summarized as follows:

1) **Streaming videos** are assumed, where neither length nor content are known.
2) Actions must be **detected as soon as they happen**, ideally in real-time.
3) **Detections must be causal.** Future cannot be used, simply because it is not known.

Note that even though OAD is naturally characterized by untrimmed streaming videos where actions appear sparsely, we found state-of-the-art models that do not consider the background as a category. They treat the OAD problem as a per-frame labeling task where detecting ground truth action frames is what only matters. Mislabeled background frames are dismissed. This means that these methods will not achieve the maximum of a precision-based metric even if the method cast for every action frame the correct category, as we show later in Section IV.

In our exercise of revisting the OAD problem, we propose to add the following properties for OAD methods:

- Methods will explicitly discriminate action from background.
- No post-processing or posterior thresholding to action label scores can be applied.
- Methods cannot revisit past detections.

B. ONLINE EVALUATION PROTOCOL

A true online evaluation protocol is needed. It is necessary to revisit the evaluation protocol and establish a new one that is in line with the online nature of the OAD problem. We argue an evaluation protocol for OAD must comply with the following conditions:

(**C1**): An online video-level metric is needed. So method's performance can be evaluated as a video grows without having to wait to an unknown end.
(**C2**): If the OAD task requires methods that are able to detect background, the evaluation protocol must measure such ability.
(**C3**): The value of a *true*, true positive (action) and true negative (background), should be conditioned to the negatives vs. positives ratio, which must be dynamic and based only on the seen portion of the video.

1) PREVIOUS METRICS

All previous evaluation protocols use class-level metrics which have to be applied offline, *i.e.* at the end of the test time, accessing the whole set of action annotations in a given test video. Hence, condition (C1) is directly violated. These protocols are mainly based on using the per-frame mean average precision (mAP) or its calibrated version (cAP).

Regarding mAP, it measures the precision, defined by $Prec = \frac{TP}{TP+FP}$, across all classes. As can be seen in its definition, only positives factors (actions) are considered and their value is always the same regardless of any ratio. Conditions (C2) and (C3) are not complied.

Precision in cAP is defined by $cPrec = \frac{wTP}{wTP+FP}$. This metric was introduced in [1] and balances the precision with the w parameter, which is the ratio between negative vs. positive frames. It is basically a modification of mAP metric so conditions (C1) and (C2) are still not complied. It would solve condition (C3) but w is computed a priori (not dynamically) using previous information about all videos and action categories.

2) INSTANTANEOUS ACCURACY METRIC

We introduce a new metric which meets all the aforementioned conditions: the Instantaneous Accuracy (IA(t)). Considering a set of $\mathcal{N}$ test streaming videos, for each video $\mathcal{V}_i$, where $i = 1 \ldots N$, an OAD method generates a set of action detections defined by their initial and ending times. IA metric takes as input these detections to build a dense temporal prediction of background or action for every time slot Δt in the test video. Note Δ is the unique parameter of our IA metric and it measures how often the metric is computed. In section IV we give details on choosing this value.

For a particular instant of time $0 < t' \leq T_i$, the IA(t') is computed as the time slot-level accuracy for the classification

between action and background:

$$\mathrm{IA}(t') = \frac{\sum_{j=0:\Delta t:t'} \vec{tp}(j) + \sum_{j=0:\Delta t:t'} \vec{tn}(j)}{K'}, \quad (1)$$

where $\vec{tp}$ and $\vec{tn}$ are two vectors encoding the true positives (action) and true negatives (background), respectively, according to the predictions and ground truth. K' represents the total population considered until time t', which is dynamically obtained as follows:

$$K' = \left\lfloor \left(\frac{t'}{\Delta t} \right) \right\rfloor. \quad (2)$$

To meet condition (C3), and to enable easy and fair comparisons across different OAD approaches, we propose a weighted version of the IA: the wIA. Technically, we scale the *true* factors by the background vs. action slots as follows:

$$\mathrm{wIA}(t') = \frac{\sum_{j=0:\Delta t:t'} w(t') \cdot \vec{tp}(j) + \sum_{j=0:\Delta t:t'} \frac{1}{w(t')} \cdot \vec{tn}(j)}{K'} \quad (3)$$

where $w(t')$ represents the dynamic ratio between background and action slots until time t' in the ground truth, *i.e.* in $\mathcal{V}_i(0:t')$.

The metric described so far only uses information from the past and is capable of adapting its parameters in each iteration. It would be sufficient to evaluate an OAD method on a single video stream of any length. Additionally, we introduce the mean average Instantaneous Accuracy (maIA) shown in equation 4 to summarize a method's performance across a large dataset. In this way, researchers can compare their methods.

$$\mathrm{maIA} = \frac{1}{N} \sum_{i=1:N} \left(\frac{\Delta t}{T_i} \sum_{j=0:\Delta t:T_i} \mathrm{IA}(j) \right). \quad (4)$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASETS

We use three datasets for all our experiments. All of them provide untrimmed videos where action and background segments coexist, suiting our OAD scenario. Thumos'14 [17] dataset has temporal annotations for a set of 413 videos, covering 20 sport classes. On average, every video contains 15 action annotations. For training, we use the 200 videos from the validation set, while the remaining 213 from the test set are used for testing. TVSeries [1] is an OAD-specific dataset. It contains 27 episodes from 6 popular TV series with 30 realistic action categories annotated. Its large variability (occluded, multiple persons or non-relevant actions, among others), makes it a really challenging dataset. Finally, we also integrate in the OAD experiments, ActivityNet v1.3 [18], which is a large scale dataset specifically designed for Temporal Action Localization. It contains about 20K untrimmed videos for 200 action classes. The average number of action instances per video is of 1.5. For this dataset, we follow the standard procedure: we use the training set and the validation set during training and test respectively. While both Thumos'14 and TVSeries have been already used within the OAD context, we are the first in integrating the challenging ActivityNet into the online setting.

2) EVALUATION METRICS

On all datasets, we report our novel *IA* for each video and provide *maIA* to evaluate methods across each dataset. Following the setup detailed in [13], we analyze the per-frame mAP on Thumos'14 and TVSeries datasets. And finally, for the TVSeries, we analyze the proposed Calibrated Average Precision (cAP) [1].

B. BASELINES

In our study, we use three baselines:

1) All background (**All-BG**). It simply simulates a model which never outputs an action class, helping to understand the difficulty of the datasets.
2) Perfect Model (**PM**), that always assigns correct labels to ground truth action frames and produces a random action label for every background frame. PM helps to reveal the limitations of the mAP and cAP evaluation metrics for OAD, showing they cannot saturate to the maximum.
3) **3D-CNN**. As shown in Figure 2, it consists of a 3D CNN network trained to discriminate between all action labels plus the background category. Our goal is to establish baseline results for the new online evaluation protocol for OAD with a model capable of explicitly detecting actions and background for the first time.

Our **3D-CNN** is based on the C3D network [33]. Technically, we adapted the dimension of the last fully connected layer of C3D model so that it coincides with the number of classes of interest plus the background category. The architecture is fed with 16-frame length chunks. For training, we extract 16-frame length contiguous chunks. Those whose intersection with ground truth is greater than 0.8 are marked as positive, otherwise they are considered negative (background). The training data $\mathcal{T}$ is balanced by matching the number of samples in each class: $N_{\mathcal{T}} = \frac{N_{chunks}}{\mathcal{C}}$, being $\mathcal{C}$ the total number of classes including background. We initialize our network with Sports-1M [34] weights and SGD is configured with learning rates 10^{-3}, 10^{-4} and 10^{-5} for Thumos'14, ActivityNet and TVSeries respectively. For all datasets, momentum is 0.9 and learning rate decreases every 2 epochs. The model is trained for 15 epochs. During test, we simulate the online process on each video by gathering 16 non-overlapping frames and input them to the network, which will cast a prediction. We take the softmax value corresponding to the background class and if it is above 0.8 we consider the detection as background. Otherwise, the detection will be the action class with highest softmax score.

3D-CNN baseline not only is simple but it also requires neither refinement nor post-processing, and can run in

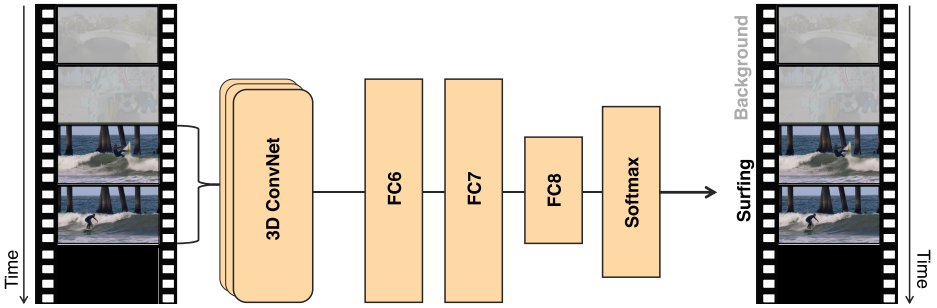

FIGURE 2. 3D-CNN baseline model. Our model closely follows C3D [33] but trained to discriminate between all action categories plus background. We simply adapted the dimension of the last fully connected layer so that it coincides with the number of categories of interest and the background class. The model makes predictions in an online fashion, avoiding to peek into the future for adjusting or post-processing these detections. In short, 3D-CNN generates action and background predictions as the video evolves.

real-time (at more than 100 fps). The experimental evaluation shows that it is a strong baseline. Caffe [35] is used for its implementation and it will be publicly available.

C. A COMPARISON BETWEEN THE METRICS

It is important for us that the reader understands the main weaknesses of the previous evaluation protocols. For this reason, and though it might be unfair, we make a comparison in this section of the performance of all methods with all the metrics on TVSeries dataset.

Table 1 shows the results for all the baselines and the state-of-the-art model in [1]. First, the fact that the PM baseline performance is not the maximum for both cAP and mAP, confirms that using methods and metrics that are not capable of managing the background category is not appropriate for the OAD task. Even though the cAP metric seems to alleviate this problem, it is not enough to achieve a 100% and it is based on diminishing background errors. Second, All-BG baseline reveals: a) that previous metrics are not able to measure method's ability to distinguish both action and background and b) the need of having a metric such as IA, capable of weighting the relevance of errors in both action and background. This last fact is especially important when dealing with very unbalanced datasets like TVSeries. Third, results from 3D-CNN are competitive when compared to the state-of-the-art. So it is confirmed as a strong baseline for the OAD problem. It is only with the cAP metric that CNN [1] really outperforms it. The reason is that this method does not cast predictions of background category (while 3D-CNN does) and, as said before, cAP has been designed to minimize the importance of such errors.

TABLE 1. Analysis of all the metrics on TVSeries.

	CNN [1]	All-BG	3D-CNN	PM
mAP (%)	1.9	0	1.6	30.9
cAP (%)	60.8	0	10.8	96.9
maIA (%)	3.51	78.3	71.9	100
weighted maIA (%)	12.46	22.9	28.9	100

In the OAD problem, it is fundamental to consider the background as one more category in the video. While our 3D-CNN baseline does explicitly consider it, most state-of-the-art online models do not. How does this fact affect performances? We analyze this on Thumos'14.

Table 2, shows the per-frame mAP achieved by all state-of-the-art models and our 3D-CNN. The poor performance of the perfect model confirms again the limitation of the mAP metric. Additionally, 3D-CNN results on this dataset also demonstrate this model is a good baseline for OAD. It is important noticing that all state-of-the-art methods assign an action category to every frame in the video, including those frames that belong to background segments. Moreover, the metric is not considering background errors. This means that mAP does not encourage methods to correctly discriminate background segments. To be precise, RED [13] does use the background during training to predict sequences of labels. But at test time, in no case is the background separated from the action. Furthermore, RED is designed for anticipation and these results are obtained when taking a very short anticipation time. So, it cannot be considered as a pure online action detection method, because it violates the causality condition. In any case, from this perspective, the performance reported by 3D-CNN is even more relevant: while our model has been trained to deal with a harder problem, it is able to maintain a state-of-the-art performance.

TABLE 2. Per-frame mAP performance on Thumos'14.

	TS-CNN [36]	MultiLSTM [36]	RED [13]	3D-CNN	PM
mAP (%)	36.2	41.3	45.3	30.1	57.0

Finally, we want to emphasize that neither mAP nor cAP are *online* metrics. Results in Tables 2 and 1 for these two metrics can only be reported once the methods have been executed on all the videos. Instead, our IA metric is online. It can perform a true online comparison between OAD models, as we show in the next section.

Overall, we conclude that a novel online metric with an adequate evaluation protocol is needed.

D. EVALUATION WITH INSTANTANEOUS ACCURACY

We analyze our IA metric with the 3D-CNN baseline and the CNN [1] approach. We have not found any code or results of

other OAD state-of-the-art methods, except for CNN [1] and LSTM [1]. Since the performance of CNN [1] and LSTM [1] are similar, we have decided to generate the results of the first for its simplicity. We have exactly reproduced the code provided by the authors. Note that these methods do not recognize background. Additionally, we tried to use results from offline temporal action detection methods but did not find a fair adaptation of them.

1) INSTANTANEOUS ACCURACY FOR EVALUATION IN ONLINE STREAMING VIDEOS

In a nutshell, our novel IA measures in an online way how accurate a certain OAD method is being along the streaming video, based only on what has been seen up to the instant of evaluation.

Regarding the parameters of the metric, the slot duration represents how often the evaluation is applied and it should be 0. Since such an ideal duration is technically not achievable, we have configured it to be the shortest possible. Most action detection approaches use chunk-level features. The chunk length is typically in the interval [16, 64] frames, with a 25/30 frame rate, representing each chunk about 0.6 to 2 seconds. Thus, we choose a 0.5 seconds for slot duration parameter Δt. The IA metric considers correct predictions of both true positives (action categories) and true negatives (background). The value of a true prediction is dynamically weighted according to the ratio of negative/positive slots seen so far. Figure 3 shows this dynamic behaviour of the weights. The weight applied to TP (the weight of the TN is the inverse) changes throughout the video. Action and background are not always balanced at each instant of evaluation during the video stream. For this reason, the weight of the true positive predictions (finding action) increases in those portions of the video in which there is no action annotated. This fact represents how the IA metric is modulating the importance of a good prediction and it is a very relevant difference with previous protocols.

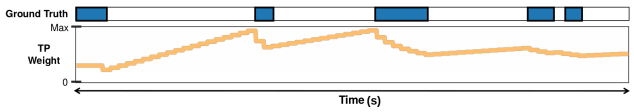

FIGURE 3. Dynamic weights. It can be seen how the value of a True Positive (TP) is weighted according to the ratio of negative/positive slots seen up to the instant of evaluation. As Equation 3 shows, TN weights offer the inverse effect.

In Figure 4 we show qualitative results for two different videos. Only a section of the videos is shown. Note that an accuracy value for a certain slot does not depend on that of the previous slot. Those values rely only on the predictions and the weights for each correct prediction. These dynamic weights can lead to situations where the accuracy value decreases (not much) even if a method is getting right predictions. This effect is seen in the upper example of Figure 4, in the segment between the ground truth annotations. However, that is exactly what we want: since nothing about the video is known before, the importance of detecting action or background must vary throughout the streaming.

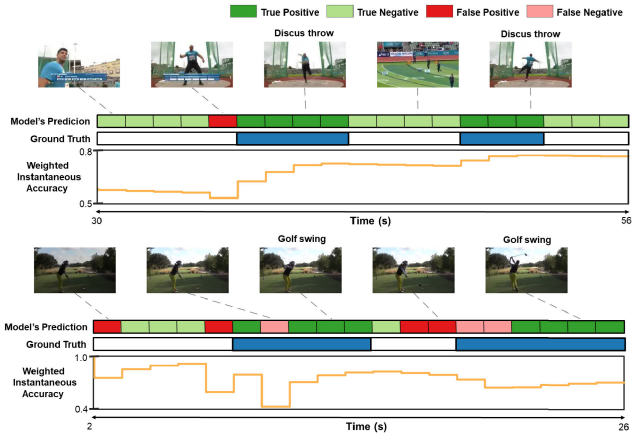

FIGURE 4. Qualitative Results. We showcase the evolution of the weighted IA on two different Thumos'14 videos. Each instant of evaluation depends on the current model's prediction. IA metric is an online video-level metric which measures the ability of methods to discriminate actions and background.

Figure 5 shows the evolution of the weighted IA on the 7 videos of the test subset of TVSeries dataset. Note how it allows for a true online comparison between OAD models, in this case, CNN [1] and 3D-CNN.

2) maIA AS IA CONSOLIDATION FOR EVALUATION ACROSS DATASETS

Despite the fact that IA metric can be directly used as it is in a video stream, we propose also the maIA to compare methods on a certain dataset.

Table 3 presents the performance with the weighted and non-weighted versions of maIA on the three datasets. Results from All-BG baseline reveal the relevance of having a weighted metric. Thumos'14 and TVSeries are very unbalanced datasets and when introducing the weighting, the performance drops a lot. On ActivityNet All-BG performs similar with the two versions of the metric due to the fact that the dataset is more balanced. These results confirm the consistency of our metric, which is capable of making a fair evaluation in all kind of datasets.

TABLE 3. Weighted an non-weighted maIA on Thumos'14, TVSeries and ActivityNet.

	Thumos'14		TVSeries			ActivityNet	
	All-BG	3D-CNN	All-BG	3D-CNN	CNN [1]	All-BG	3D-CNN
maIA (%)	70.9	77.8	78.3	71.9	3.51	40.1	21.7
weighted maIA (%)	41.8	69.0	22.9	28.9	12.46	53.6	27.4

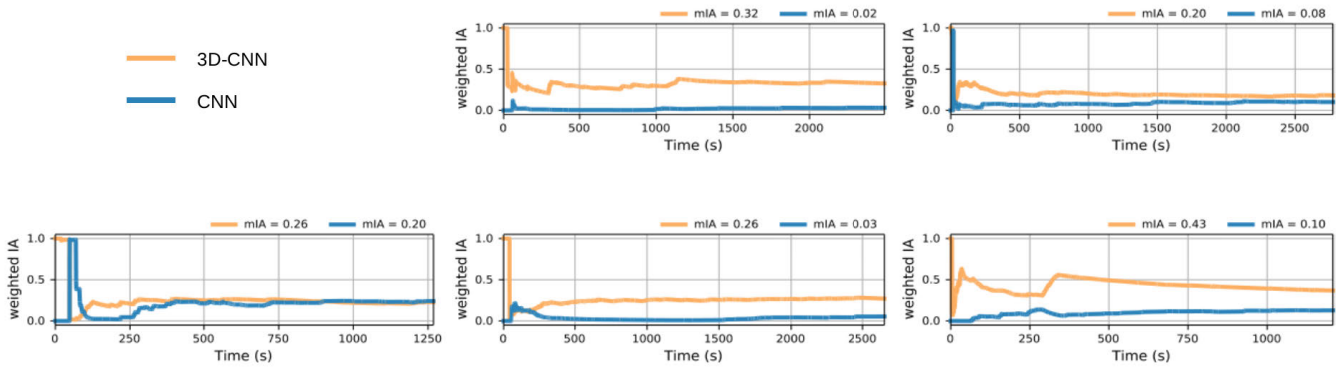

FIGURE 5. **Online IA based evaluation in videos of the TVSeries dataset.**

The low numbers of 3D-CNN on TVSeries and ActivityNet are caused by different reasons. TVSeries is a specially very unbalanced and challenging dataset. With such a lot of background, a model as simple as 3D-CNN is not able to learn well to discriminate action from background. ActivityNet is balanced but has many classes to distinguish. Finally, our reproduced CNN [1] performs poorly according to maIA due to it does not handle background. Thus, its performance is alleviated when weighted with the positive/negative ratio.

V. CONCLUSION

Online Action Detection in untrimmed streaming videos is a challenging task with few contributions. We have found that a) the task itself needs a solid definition of its properties, b) there is no clear consensus on how the methods should deal with this type of videos and c) a proper evaluation protocol is not defined.

In this work, we solved the first two problems by revising and establishing the properties of the OAD task itself as well as those for the methods designed for it. Regarding the third, we noticed that there are limitations in the metrics used so far. Therefore, we have clearly defined the conditions of a proper evaluation protocol: i) it has to be online, for consistency with the metric; ii) it must measure the ability of methods to discriminate both action and background and iii) it must be based only on the seen portion of video.

Since none of the previously used metrics complies with these conditions, we have introduced a new metric: the Instantaneous Accuracy (IA). IA is an online video-level metric which computes the accuracy for every instant of evaluation. Our results have proved the limitations of the previous metrics and the robustness of our novel IA.

We expect in the future more methods will be analyzed with our IA. Thanks to its characteristics, it will be possible to study the situations in which methods should perform better.

The baseline models and a development kit with the novel evaluation protocol will be made publicly available.

REFERENCES

[1] R. De Geest, E. Gavves, A. Ghodrati, C. Li, Z. Snoek, and T. Tuytelaars, "Online action detection," in *Proc. ECCV*, 2016.
[2] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. CVPR*, 2016.
[3] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. CVPR*, 2017.
[4] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *Proc. BMVC*, 2017.
[5] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," Nov. 2015, *arXiv:1511.06984*. [Online]. Available: https://arxiv.org/abs/1511.06984
[6] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proc. BMVC*, 2017.
[7] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Q. Chen, "Temporal context network for activity localization in videos," in *Proc. ICCV*, 2017.
[8] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Niebles, "SST: Single-stream temporal action proposals," in *Proc. CVPR*, 2017.
[9] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. CVPR*, 2018.
[10] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. ICCV*, 2017.
[11] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. CVPR*, 2016.
[12] Y. Wu, J. Yin, L. Wang, H. Liu, Q. Dang, Z. Li, and Y. Yin, "Temporal action detection based on action temporal semantic continuity," *IEEE Access*, vol. 6, pp. 31677–31684, 2018.
[13] J. Gao, Z. Yang, and R. Nevatia, "RED: Reinforced encoder-decoder networks for action anticipation," in *Proc. BMVC*, 2017.
[14] R. De Geest and T. Tuytelaars, "Modeling temporal structure with LSTM for online action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1549–1557.
[15] J. Gao, K. Chen, and R. Nevatia, "CTAP: Complementary temporal action proposal generation," in *Proc. ECCV*, 2018.
[16] Z. Shou, J. Pan, J. Chan, K. Miyazawa, H. Mansour, A. Vetro, X. G.-I. Nieto, and S.-F. Chang, "Online detection of action start in untrimmed streaming videos," in *Proc. ECCV*, 2018.
[17] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. (2014). *THUMOS Challenge: Action Recognition With a Large Number of Classes*. [Online]. Available: http://crcv.ucf.edu/THUMOS14/
[18] F. Caba-Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. CVPR*, 2015, pp. 961–970.
[19] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. ECCV*, 2018.
[20] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. ICCV*, Oct. 2017.

[21] N. Li, H.-W. Guo, Y. Zhao, T. Li, and G. Li, "Active temporal action detection in untrimmed videos via deep reinforcement learning," *IEEE Access*, vol. 6, p. 59126–59140, 2018.
[22] X. Zhang, H. Shi, C. Li, K. Zheng, X. Zhu, and L. Duan, "Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision," in *Proc. AAAI*, 2019.
[23] S. Narayan, H. Cholakkal, F. Shahbaz Khan, and L. Shao, "3C-Net: Category count and center loss for weakly-supervised action localization," in *Proc. ICCV*, 2019.
[24] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. ICCV*, 2019.
[25] D. Wang, Y. Yuan, and Q. Wang, "Early action prediction with generative adversarial networks," *IEEE Access*, vol. 7, pp. 35795–35804, 2019.
[26] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proc. CVPR*, 2018.
[27] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proc. ECCV*, 2018.
[28] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proc. ECCV*, 2018.
[29] X.-Y. Zhang, C. Li, H. Shi, X. Zhu, P. Li, and J. Dong, "AdapNet: Adaptability decomposing encoder-decoder network for weakly supervised action recognition and localization," Nov. 2019, *arXiv:1911.11961*. [Online]. Available: https://arxiv.org/abs/1911.11961
[30] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *Proc. IJCV*, 2014.
[31] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *Proc. CVPR*, 2017.
[32] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016.
[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[34] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. CVPR*, 2014.
[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," Jun. 2014, *arXiv:1408.5093*. [Online]. Available: https://arxiv.org/abs/1408.5093
[36] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 375–389, 2018, doi: 10.1007/s11263-017-1013-y.

MARCOS BAPTISTA-RÍOS received the Telecommunication Engineering degree from the University of Alcalá, in 2016, where he is currently pursuing the Ph.D. degree. His research interest focuses on image and video understanding, particularly in the development of deep/machine learning models capable of analyzing streaming videos in an online fashion.

ROBERTO J. LÓPEZ-SASTRE is currently an Associate Professor with the Department of Signal Theory and Communications, University of Alcalá. His research interests are centered around computer vision, machine learning and robotics, focusing on object detection, scene understanding, and action recognition. He serves as a Program Committee Member/Reviewer of the major computer vision conferences ICCV, ECCV, and CVPR.

FABIAN CABA HEILBRON received the Ph.D. degree from the King Abdullah University of Science and Technology, in 2019. He is currently a Research Scientist with Adobe Research. His research interest is in video understanding, specifically his delving deep into machine learning models to speed up video editing workflows. He has organized the ActivityNet challenge during four consecutive years from 2016 to 2019, which has attracted a large number of participants and industry sponsors.

JAN C. VAN GEMERT received the Ph.D. degree from the University of Amsterdam, in 2010. He was a Postdoctoral Fellow with École Normale Supérieure in Paris. He currently leads the Computer Vision Lab, Delft University of Technology, where he teaches the Computer Vision and Deep learning M.Sc. courses. His current research interest is adding visual inductive priors to deep learning. He published over 75 peer-reviewed articles with more than 4500 citations.

F. JAVIER ACEVEDO-RODRÍGUEZ received the M.Sc. degree in electronic engineering from the Univesity of Alcalá, in 1998, and the Ph.D. degree from the University of Alcalá, in 2009. He currently leads the Department of Signal Theory and Communications, University of Alcalá. His research interests are in the field of pattern recognition and signal processing, especially on those projects in the electrochemical field and those dedicated to people with special needs.

SATURNINO MALDONADO-BASCÓN (Senior Member, IEEE) received the Telecommunication Engineering degree from Universidad Politécnica de Madrid, in 1996 and the Ph.D. degree from the University of Alcalá, in 1999. Since 1997, he has been a Faculty Member with the University of Alcalá, becoming a Full Professor with the Department of Signal Theory and Communications, in 2011. His research interests are image processing and pattern recognition.