The Instantaneous Accuracy: a Novel Metric for the Problem of Online Human Behaviour Recognition in Untrimmed Videos

Marcos Baptista-Ríos[1], Roberto J. López-Sastre[1], Fabian Caba-Heilbron[2], Jan Van Gemert[3],
F. Javier Acevedo-Rodríguez[1], Saturnino Maldonado-Bascón[1]
[1] University of Alcalá, [2] Adobe Research, [3] Delft University of Technology
marcos.baptista@uah.es

Abstract

The problem of Online Human Behavior Recognition in untrimmed videos, aka On-line Action Detection (OAD), needs to be revisited. Unlike traditional off-line action detection approaches, where the evaluation metrics are clear and well established, in the OAD setting we find few works and no consensus on evaluation protocols to be used. In this paper we introduce a novel on-line metric, the Instantaneous Accuracy (IA), that exhibits an on-line *nature, solving most of the limitations of the previous (off-line) metrics. We conduct a thorough experimental evaluation on the TVSeries dataset, comparing the performance of various baseline methods with 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 on-line scenario for human behavior understanding.*

1. Introduction

In this work, we focus on the problem of recognizing human behaviors in untrimmed videos *as soon as* they happen, which has been coined as On-line Action Detection (OAD) by De Geest *et al*. [5].

The problem of action detection has been widely studied, but *mainly from an off-line perspective*, *e.g*. [1, 2, 3, 4, 8, 10, 11, 13, 14, 15], where it is assumed that all the video is available to make predictions. Few works address the *on-line* setting, *e.g*. [5, 6, 8, 9]. Think of a robotic platform that must interact with humans in a realistic scenario, recognizing their behaviors. *All* previous off-line methods make this application impossible as they will detect the action situations way later they have occurred. In contrast, OAD approaches must give detections over video streams, hence working with partial observations. However, among the on-line approaches there is an important weakness in a fundamental aspect: the evaluation metric. We have noticed that there is no consensus on the evaluation protocols, *i.e*. in each dataset a different metric is proposed for the same problem. Moreover, used metrics cannot be said to be of an *on-line nature*. In other words, the proposed metrics for recent *on-line* action detection models, such as the mean Average Precision (mAP) [7] or the Calibrated Average Precision (cAP) [5], do not provide information about the instantaneous performance of the solutions over time: they need to be computed entirely off-line, accessing the whole set of action annotations for a given test video.

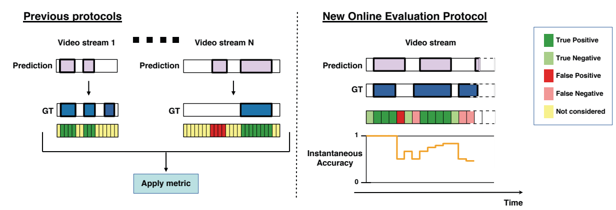


Figure 1: Previous evaluation protocols for OAD were based on: 1) running the on-line methods through all videos; 2) applying the off-line metric on the obtained results. We propose an on-line evaluation protocol, based on our new Instantaneous Accuracy metric (IA), where the approaches are evaluated on-line, considering the background, and regardless of the length of the video.

We introduce here an evaluation protocol with a novel *on-line* metric, the Instantaneous Accuracy (IA) (see Figure 1). This metric has been designed not only to overcome the described limitations, but to allow fair comparisons between OAD methods. A thorough experimental evaluation is conducted on the challenging TVSeries [5] dataset, offering a comparison between baselines and state-of-the-art approaches. The results show that an IA-based evaluation protocol is more adequate for 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.

2. The Instantaneous Accuracy

We argue an evaluation protocol for OAD must meet the following main condition: an on-line video-level metric is needed, with which method performances can be evaluated as a video grows over time.

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, on-line condition is directly violated. These protocols are mainly based on using the per-frame mean average precision (mAP) or its calibrated version (cAP).

Instantaneous Accuracy metric. Considering a set of $\mathcal{N}$ test videos, for each video, 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 action (including background) for every time slot Δt in the test video. Note Δt is the unique parameter of our IA metric and it measures how often the metric is computed. For a particular instant of time t', the $\text{IA}(t')$ is computed as the time slot-level accuracy for the action classification task as follows:

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

where $\mathbf{tp}$ and $\mathbf{tn}$ are two vectors encoding the true positives (actions) and true negatives (background), respectively, and K' represents the total population considered until time t', which is dynamically obtained as follows: $K' = \left\lfloor \left(\frac{t'}{\Delta t} \right) \right\rfloor$.

As working with untrimmed videos, where much more background than action frames appear, we propose a weighted version of the IA. Technically, we simply scale in Eq. 1 the *true* factors by the dynamic ratio between background and action slots until time t' in the ground truth.

To summarize the method's performance across a dataset for research purposes, we propose to use the mean average Instantaneous Accuracy (maIA) for every video:

$$\text{maIA} = \frac{1}{N} \sum_{i=1:N} \left(\frac{\Delta t}{T_i} \sum_{j=0:\Delta t:T_i} \text{IA}(j) \right). \quad (2)$$

3. Experiments and conclusions

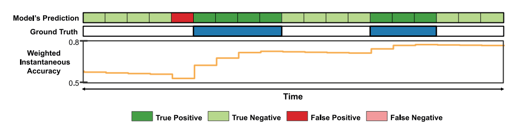
We use the challenging TVSeries [5] dataset, following the setup in [9] to analyze all the metrics considered in our study: mAP, cAP, and the novel on-line IA.

As baselines, we propose the following. All background (**All-BG**), which simply simulates a model that never generates an action class. Perfect Model (**PM**), that always assigns correct labels to action and background frames. PM helps to reveal the limitations of the previous evaluation protocols, showing their metrics cannot saturate to the maximum for which they have been designed. Finally, we propose a **3D-CNN** model, which consists in a C3D network [12] to recognize all actions and the background category.

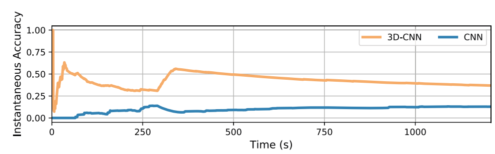
Table 1: Analysis of the metrics on TVSeries.

	CNN [5]	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

Table 1 shows the results for these baselines and state-of-the-art model in [5]. First, one observes that off-line protocols do not succeed in giving a 100% even for a perfect method. Like we explained, this is due to their incorrect way of managing the background category. It is true that the calibrated AP seems to alleviate this problem, but it still does not achieve a 100%. Second, results from All-BG baseline reveal the relevance of having a weighted metric, especially for so unbalanced problem. And third, 3D-CNN achieves competitive performance when compared to the state of the art, supporting its choice as a strong baseline for the OAD problem. It is only for the cAP where its performance decreases compared with CNN [5], but the reason is that CNN does not cast predictions for background category (while 3D-CNN does), and the cAP has been designed to minimize the importance of such errors.



(a)



(b)

Figure 2: (a) Evolution of the weighted IA for a video for the 3D-CNN baseline. (b) Weighted IA comparison between 3D-CNN and CNN [5].

Figure 2a shows the evolution of the weighted IA for a particular video. We use here for visualization 0.5 seconds for Δt. One can observe how the weighting mechanism works, dynamically adapting the IA to the proportion of the video observed by the model. Figure 2b also shows a comparison between CNN [5] and our 3D-CNN.

As a conclusion, on-line human behavior recognition in untrimmed videos is a challenging task with few contributions. We found limitations in the metrics used so far, so we have introduced a new on-line metric that complies with the on-line nature of the problem: the Instantaneous Accuracy (IA). Experimental results have proved both the limitations of previous used metrics and the robustness of IA.

References

[1] 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 *BMVC*, 2017. 1

[2] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Niebles. SST: single-stream temporal action proposals. In *CVPR*, 2017. 1

[3] 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 *CVPR*, 2018. 1

[4] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Q. Chen. Temporal context network for activity localization in videos. In *ICCV*, 2017. 1

[5] R. De Geest, E. Gavves, A. Ghodrati, C. Li, Z.and Snoek, and T. Tuytelaars. Online action detection. In *ECCV*, 2016. 1, 2

[6] R. De Geest and T. Tuytelaars. Modeling temporal structure with LSTM for online action detection. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 1549–1557, 2018. 1

[7] J. Gao, K. Chen, and R. Nevatia. CTAP: Complementary temporal action proposal generation. In *ECCV*, 2018. 1

[8] J. Gao, Z. Yang, and R. Nevatia. Cascaded boundary regression for temporal action detection. In *BMVC*, 2017. 1

[9] J. Gao, Z. Yang, and R. Nevatia. RED: Reinforced encoder-decoder networks for action anticipation. In *BMVC*, 2017. 1, 2

[10] 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 *CVPR*, 2017. 1

[11] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 1

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, Dec 2015. 2

[13] H. Xu, A. Das, and K. Saenko. R-C3D: Region convolutional 3D network for temporal activity detection. In *ICCV*, 2017. 1

[14] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. *arXiv preprint arXiv:1511.06984*, 2015. 1

[15] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *CVPR*, 2016. 1