Diagnosing Error in Temporal Action Detectors

Humam Alwassel(✉), Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
{humam.alwassel,fabian.caba,victor.escorcia, bernard.ghanem}@kaust.edu.sa,
http://www.humamalwassel.com/publication/detad/

Abstract. Despite the recent progress in video understanding and the continuous rate of improvement in temporal action localization throughout the years, it is still unclear how far (or close?) we are to solving the problem. To this end, we introduce a new diagnostic tool to analyze the performance of temporal action detectors in videos and compare different methods beyond a single scalar metric. We exemplify the use of our tool by analyzing the performance of the top rewarded entries in the latest ActivityNet action localization challenge. Our analysis shows that the most impactful areas to work on are: strategies to better handle temporal context around the instances, improving the robustness w.r.t. the instance absolute and relative size, and strategies to reduce the localization errors. Moreover, our experimental analysis finds the lack of agreement among annotator is not a major roadblock to attain progress in the field. Our diagnostic tool is publicly available to keep fueling the minds of other researchers with additional insights about their algorithms.

Keywords: Temporal action detection · Error analysis
Diagnosis tool · Action localization

1 Introduction

We are in the *Renaissance* period of video understanding. Encouraged by the advances in the image domain through representation learning [14,17,23], large scale datasets have emerged over the last couple of years to challenge existing ideas and enrich our understanding of visual streams [4,15,16,19,21,22,28,37].

The first three authors contributed equally to this work. Authors ordering was determined using Python's random.shuffle() seeded with the authors' birthday dates.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01219-9_16) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11207, pp. 264–280, 2018.
https://doi.org/10.1007/978-3-030-01219-9_16

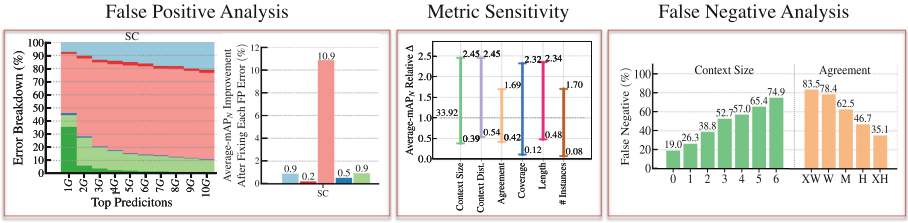

Fig. 1. Illustration of the three types of analyses that our diagnostic tool provides for action localization algorithms. Left: We analyze the false positive error sources and their impact on the performance. Middle: We investigate the localization metric sensitivity to different characteristics of the ground truth instances. Right: We inspect the influence of ground truth instance characteristics on the miss detection rate.

Recent work has already shown novel algorithms [41] and disproved misconceptions associated with underrated 3D representations for video data [7]. However, we are still awaiting the breakthrough that allows us to temporally localize the occurrence of actions in long untrimmed videos [13,19]. In this paper, we propose to step back and analyze the recent progress on the temporal action localization as a means to fuel the next generation with the right directions to pursue.

Currently, researchers have appealing intuitions to tackle video action localization problem [1,10,34,45], they are equipped with large datasets to validate their hypothesis [4,5,19], and they have access to appropriate computational power. Undoubtedly, these aspects helped materialize an increasing performance throughout the years [3,13,45]. Yet, such improvements are not enough to describe the whole picture. For example, we are still not able to answer the following questions: How close are we to achieve our goal of delimiting the start and end of actions? What makes an algorithm more effective than another? What makes an action hard to localize? Is the uncertainty of the temporal boundaries impeding the development of new algorithms? Inspired by similar studies in other areas [18,30,36,44], we take a deep look at the problem beyond a single scalar metric and perform a quantitative analysis that: (i) informs us about the kind of errors a given algorithm makes, and measures the impact of fixing them; (ii) describes which action characteristics impact the performance of a given algorithm the most; and (iii) gives insights into the action characteristics a proposed solution struggles to retrieve. Figure 1 shows a brief glimpse of our diagnostic analysis applied to a state-of-the-art method on ActivityNet version 1.3 [4,26].

Relation to Existing Studies. The seminal work of Hoiem *et al.* showcased the relevance of diagnosing the failure modes of object detectors in still images [18]. Inspired by this work, [11,19,27,30–32,36,44] provided insightful analysis of algorithms for multiple localization tasks such as human pose estimation, object detection at large-scale and multi-label action classification, in still images in most of the cases. In contrast with them, our work contributes to the understanding of temporal action localization in untrimmed videos.

Object Detection and Human Pose Estimation. [18] pioneered the categorization of localization errors as a mean to get more insights about the performance of object detection algorithms. [30,44] extended the diagnostic tools to the context of human pose estimation showing the relevance of this approach to quantitatively identify the failure modes and to recommend ways to improve existing algorithms for body parts localization. In a similar spirit, our work is the first that characterizes the localization errors for temporal action localization in videos.

Multi-label Action Classification. Sigurdsson *et al.* [36] provides an insightful diagnosis of algorithms and relevant directions needed for understanding actions in videos. [36] studies the influence of different attributes, such as verbs, objects, human poses, and the interactions between actions in the scope of video action recognition. Most of the study is centered around action classification at the frame level or the entire video and is carried on relatively short streams of 30 seconds on average. Our work contributes with an orthogonal perspective to this study, performing an in-depth analysis of the problem of delimiting temporal boundaries for actions in long videos.

Contributions. Our contributions in this paper are threefold. **(i)** We collect additional annotation data for action context and temporal agreement in ActivityNet. We use the collected data to categorize the ground truth instances into six action characteristics: context size, context distance, agreement, coverage, length, and the number of instances (Sect. 3). **(ii)** We investigate and classify the most relevant error types to temporal action localization (Sect. 4). **(iii)** We provide a complete analysis tool (annotations, software, and techniques) that facilitates detailed and insightful investigation of action detectors performance. We exemplify the use and capabilities of our diagnosis tool on the top four action detectors in the recent ActivityNet 2017 challenge (Sects. 5–7).

2 Preliminaries

Evaluation Framework. We use the ActivityNet dataset v1.3 [4] as a test bed for our diagnostic analysis of the progress in temporal action localization in videos. The choice of this dataset obeys multiple reasons, (i) it is a large scale dataset of 20K videos with an average length of four minutes; (ii) it consists of a diverse set of human actions ranging from household activities, such as *washing dishes*, to sports activities, like *beach volleyball*. This allow us to make conclusions about a diverse type of actions; (iii) it is an active non-saturated benchmark with a held-out test set and an additional validation set, ensuring good machine learning practices and limiting over-fitting risk; (iv) it provides an open-source evaluation framework and runs an annual competition, which safeguards good progress on the community. Additionally, we extend our analysis to the widely used THUMOS14 dataset [20] in the *supplementary material*. In this way, we cover the most relevant benchmarks used to dictate the progress in this area.

The action localization problem measures the trade-off that an algorithm consistently retrieves the occurrence of true action instances, from different classes,

Table 1. Localization performance as measured by average-mAP and average-mAP$_N$ on ActivityNet [4]. We show the two metrics for all predictions and for the top-10G predictions, where G is the number of ground truth instances. Using average-mAP$_N$ gives slightly higher values. Notably, limiting the number of predictions to the top-10G gives performance values similar to those when considering all predictions.

Method	Average-mAP (%)		Average-mAP$_N$ (%)	
	All	top-10G	All	top-10G
SC	33.42	32.99	33.92	33.45
CES	31.87	31.83	32.24	32.20
IC	31.84	31.70	32.14	32.00
BU	16.75	16.52	17.26	17.02

without increasing the numbers of spurious predictions. This task is evaluated by measuring the precision and recall of the algorithms. The metric used to trade-off precision and recall for retrieving the segments of a particular action is the Average Precision (AP), which corresponds to an interpolated area under the precision-recall curve [11]. To evaluate the contribution of multiple action classes, the AP is computed independently for each category and averaged to form the mean AP (mAP). Given the continuous nature of the problem, a prediction segment is considered a true positive if its temporal Intersection over Union (tIoU) with a ground truth segment meets a given threshold. To account for the varied diversity of action duration, the public evaluation framework employs the average-mAP, which is the mean of all mAP values computed with tIoU thresholds between 0.5 and 0.95 (inclusive) with a step size of 0.05.

To establish a middle ground between multiple algorithms that is robust to variations of ratio between true and false positives across multiple classes, we employ the normalized mean AP [18]. In this way, we can compare the average-mAP between uneven subsets of ground truth instances, *e.g.* when the number of instances of a given category doubles the number of instances of another category for a given detection rate. We compute the normalized mAP (mAP$_N$) in terms of the normalized precision $P_N(c) = \frac{R(c) \cdot N}{R(c) \cdot N + F(c)}$, where c is the confidence level, $R(c)$ is the recall of positive samples with confidence at least c, $F(c)$ is the false positive rate for predictions with confidence at least c, and N is a constant number. We report average-mAP$_N$ as the action localization metric, and set N to the average number of ground truth segments per class.

Algorithms. We exemplify the use of our diagnostic tool by studying the four rewarded approaches in the latest action localization task in the ActivityNet challenge [13] (Table 1 summarizes the methods' performances). Interestingly, all the methods tackled the problem in a two-stage fashion, using a proposal method [2,6,9,12,34] followed by a classification scheme [38–40]. However, there are subtle design differences which are relevant to highlight.

SC [26]. It was the winner of the latest action localization challenge with a margin of 2% average-mAP. The key ingredient for its success relies on improving the action proposals stage. To this end, this work re-formulates the fully convolutional action detection network SSAD [25] as a class-agnostic detector. The detector generates a dense grid of segments with multiple durations, but only those near the occurrence of an instance receive a high score. In addition to the multi-scale proposal network, this work refines proposals' boundaries based on the outputs of the TAG grouping approach [42]. Finally, the classification stage is performed at the video level independently of the proposal stage results.

CES [13,45]. This work achieved the runner-up entry in the challenge [13], and held the state-of-the-art approach on THUMOS14 at that time. It employs a temporal grouping heuristic for generating actions proposals [42] from dense actioness predictions. The proposals are classified and refined in a subsequent stage by the SSN network [45]. Most of its effort involves enhancing the SSN network to a diverse set of actions. SSN applies a temporal pyramid pooling around the region spanned by a proposal segment, and then it classifies the segment by balancing the information inside the segment and the context information around it. This work found consistent improvement in the validation set through the use of deeper architecture and fine tuning on the larger Kinetics dataset [22].

IC [13]. This approach ranks third by employing a similar strategy to the CES submission. Its main distinction relies on using a sliding window-based proposal scheme as well as employing human pose estimation to influence the classification decisions of the SSN network [45].

BU [43]. It was awarded the challenge most innovative solution. This work extended the Faster RCNN architecture [29] to the problem of temporal action localization. It designed a temporal proposal network coupled with a multi-layer fully connected network for action classification and boundary refinement. In comparison with the top-ranked submissions that exploit optical flow or human pose estimation, this work relies solely on RGB streams to learn a temporal representation via 3D convolutions pretrained on Sports-1M dataset [21].

3 Dataset Characterization

Our first goal is to describe datasets with inherent characteristics such as coverage, length, and the number of instances. Moreover, we are interested in augmenting the dataset with two additional characteristics, temporal context and temporal boundary agreement, which we argue are critical for understanding the current status of action localization. Let's play some games to motivate our selection (Jump to Fig. 2). The first game, *guess-the-action*, consists of watching a series of frames to guess what action happens next. The second game, *let-us-agree*, asks you to pick the instant when a given action ends. We invite you to play the game and check your answers afterwards in the footnote[1]. We relate

[1] (1) The action that happens is *Bungee Jumping*. (2) There is not a unanimous answer for this game. 67% of our lab colleagues picked frame B as the correct answer.

the first game with whether an action instance has temporal context or not. If an action instance is in temporal context, the player should be able to exploit semantic information such as objects, scenes, or motions to guess what action happens either before or after. The second game explores how humans agree on defining temporal boundaries of an action instance. Surprisingly, this toy example reveals that defining an action's temporal boundaries is hard. Intrigued, we decided two conduct two formal online user studies with the aim of quantifying the amount of temporal context and temporal boundaries agreement for temporal action localization. In this section, we first present the online user studies that allow us to augment ActivityNet v1.3 with the temporal context and temporal boundaries agreement attributes. Then, we provide a detailed definition of each action characteristics studied in this work.

Fig. 2. Left: *guess-the-action* game. In this game you have to guess what action (one out of three options) could happen in the context of the depicted video clip. Right: *let-us-agree* game. Here, the goal is to pick the frame that best represents when the action *Long Jump* ends. To check your answers read the footnote 1.

3.1 Online User Studies

User Study I: Temporal Context of Actions. Our goal is to quantify the amount of temporal context around an action instance. To that end, we conduct an online user study that resembles the *guess-the-action* game described earlier. We choose Amazon Mechanical Turk as a test bed to hold the user study. Each participant's task is to watch a 5-second video clip and pick, from a given list, all the human actions that they believe could happen in the context of the video clip. We revisit our definition of temporal context, which describes that an action instance is in temporal context if semantic information *around* the instance helps a person to guess the action class of such instance. Thus, we investigate the temporal context of an instance by sampling six non-overlapping 5-second clips around the action's temporal boundaries. We present each user with three different candidate classes, one of the options is the correct action class, and the other two options are either similar or dissimilar class to the ground truth class. Following the findings of [3], we use objects and scene information to form sets of similar and dissimilar actions. Given that multiple selections are allowed, we consider an answer as correct if the participant chooses the correct action only, or if they pick the correct action and the option that is similar to

it. If a temporal segment allows the participant to guess the action, we call that segment a *context glimpse*.

Our study involved 53 Amazon Mechanical Turk workers (Turkers), who spent a median time of 21 seconds to complete a single task. In total, we submitted a total of 30K tasks to cover the existing instances of ActivityNet. Interestingly, Turkers were able to correctly guess the action in 90.8% of the tasks. While that result can be interpreted as a signal of dataset bias towards action-centric videos, it also suggests that action localization methods would require temporal reasoning to provide accurate predictions in such scenario. For instance, most probably you used information about scene (bridge, river) and objects (elastic cord, helmet) to predict the *Bungee Jumping* answering when playing *guess-the-action*. However, such high-level information did not help you to provide the ending time of *Long Jump* in the *let-us-agree* game. In short, for each ActivityNet temporal instance, we conducted 6 temporal context experiments, which we use later when defining action characteristics.

User Study II: Finding Temporal Boundaries of Actions. After playing *let-us-agree*, the question naturally arises, can we precisely localize actions in time? To address this question, we followed [36] and designed an instance-wise procedure that helped us characterize the level of human agreement achieved after annotating temporal bounds of a given action.

We relied on 168 Turkers to *re-annotate* temporal boundaries of actions from ActivityNet. The median time to complete the task was three minutes. The task consisted in defining the boundaries of an already spotted action. Additionally, we asked the participants to annotate each temporal boundary *individually*. For each action instance, we collected three new annotations from different Turkers. We measure agreement as the median of the pairwise tIoU between all the four annotations (the original annotation and the three newly collected ones). As a result, Turkers exhibited an agreement score of 64.1% over the whole dataset. The obtained results suggests that it is hard to agree, even for humans, about the temporal boundaries of actions, which matches with previously reported conclusions [36]. In summary, we collected three additional annotations for each action instance from ActivityNet, enabling future discussions about the effect of ambiguous boundaries on action detectors.

3.2 Definitions of Action Characteristics

We annotate each instance of the ActivityNet v1.3 dataset with six different characteristics: context size, context distance, agreement, coverage, length, and number of instances. Here, we define these characteristic and discuss their distribution (Fig. 3).

Context Size. We use the collected data from User Study I to characterize the amount of temporal context around an instance. We define context size as the number of *context glimpses* associated with an instance. Thus, values of context size range from 0 to 6. Interestingly, we find that only 6.9% of instances do

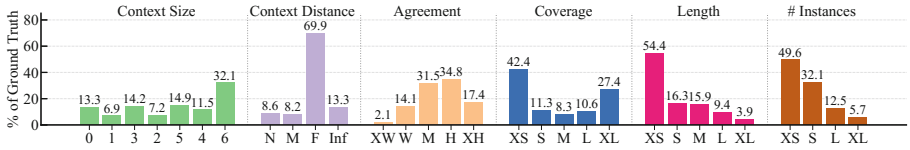

Fig. 3. Distribution of instance per action characteristic. We report the percentage of ground truth instances belonging to each characteristic bucket.

not have temporal context. Additionally, many instances have large temporal context, *e.g.* 58.4% of instances have more than 3 context glimpses.

Context Distance. We use the results from User Study I to characterize the furthest distance away from the instance where a *context glimpse* exists. We define four types of context distance: Inf, which indicates that no temporal context exists; Far (F); Middle (M); Near (N). Notably, We see that most instances (69.9%) have *context glimpses* far away.

Agreement. Our goal is to characterize an instance based on how difficult it is to agree on its temporal boundaries. To this end, we exploit the data collected from User Study II. We measure agreement as the median tIoU between all annotation pairs for an instance. We form five groups based on agreement score (median tIoU): Extra Weak (XW: $(0, 0.2]$), Weak (W: $(0.2, 0.4]$)), Mid (M: $(0.4, 0.6]$), High (H: $(0.6, 0.8]$), and Extra High (XH: $(0.8, 1.0]$). We discover that a relatively small number of instances have extremely weak agreement (2.1%). On the other hand, most of the dataset (83.8% of instances) exhibit at least Mid agreement.

Coverage. To measure coverage, we normalize the length of the instance by the duration of the video. We categorize coverage values into five buckets: Extra Small (XS: $(0, 0.2]$), Small (S: $(0.2, 0.4]$), Medium (M: $(0.4, 0.6]$), Large (L: $(0.6, 0.8]$), and Extra Large (XL: $(0.8, 1.0]$). Interestingly, Extra Small and Extra Large instances compose most of the dataset with 42.4% and 27.4% of instances assigned to each bucket, respectively.

Length. We measure length as the instance duration in seconds. We create five different length groups: Extra Small (XS: $(0, 30]$), Small (S: $(30, 60]$), Medium (M: $(60, 120]$), Long (L: $(120, 180]$), and Extra Long (XL: > 180). We find that more than half (54.4%) of the instances are small. We also observe that the instance count gradually decrease with length size.

Number of Instances (# Instances). We assign each instance the total count of instances (from the same class) in its video. We create four categories for this characteristic: Extra Small (XS: 1); Small (S: $[2, 4]$); Medium (M: $[5, 8]$); Large (L: > 8). We find half of the dataset contains a single instance per video.

4 Categorization of Temporal Localization Errors

When designing new methods, researchers in the field often identify an error source current algorithms fail to fully address. For example, [33] identifies the

problem of localization errors at high tIoU thresholds and devises the CDC network to predict actions at frame-level granularity. However, the field lacks a detailed categorization of the errors of specific relevance to the temporal localization problem. A thorough classification of error types and analysis of their impact on the temporal localization performance would help guide the next generation of localization algorithms to focus on the most significant errors. To this end, we propose in this section a taxonomy of the errors relevant to action localization, and we analyze the impact of these errors in Sects. 5 and 7.

Let $\mathcal{G}$ be the set of ground truth instances such that an instance $g^{(k)} = (g_l^{(k)}, g_t^{(k)})$ consists of a label $g_l^{(k)}$ and temporal bounds $g_t^{(k)}$. Let $\mathcal{P}$ be the set of prediction segments such that a prediction $p^{(i)} = (p_s^{(i)}, p_l^{(i)}, p_t^{(i)})$ consists of a score $p_s^{(i)}$, a label $p_l^{(i)}$, and a temporal extent $p_t^{(i)}$. A prediction $p^{(i)}$ is a **True Positive (TP)** if and only if $\exists g^{(k)} \in \mathcal{G}$ such that $p^{(i)}$ is the highest scoring prediction with $tIoU(g_t^{(k)}, p_t^{(i)}) \geq \alpha$ and $p_l^{(i)} = g_l^{(k)}$, where α is the tIoU threshold. Otherwise, the prediction is a **False Positive (FP)**. Suppose that $p^{(i)}$ is an FP prediction and $g^{(k)}$ is the ground truth instance with the highest tIoU with $p^{(i)}$. We classify this FP prediction into five categories (see Fig. 4).

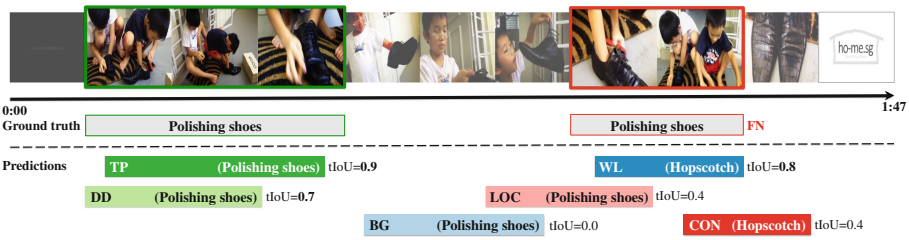

Fig. 4. Illustration of the most relevant action localization errors (Sect. 4). Predictions with bold tIoU values meet the tIoU threshold (0.55 in this example). The left action instance is correctly matched, while the right instance is miss detected (false negative). Each prediction shows a case that exhibit one of the error types we categorize.

Double Detection Error (DD). A prediction that satisfies the tIoU threshold with a ground truth instance with the correct label, however, the ground truth instance is already matched with another prediction of a higher score. We identify this error due to the nature of the ActivityNet evaluation framework, which measures performance at high tIoU thresholds and penalizes double detections.

$$tIoU(g_t^{(k)}, p_t^{(i)}) \geq \alpha,\ g_l^{(k)} = p_l^{(i)}; \exists p^{(j)} \in \mathcal{P}, tIoU(g_t^{(k)}, p_t^{(j)}) \geq \alpha,\ p_s^{(j)} \geq p_s^{(i)} \quad (1)$$

Wrong Label Error (WL). A prediction that meets the tIoU threshold but incorrectly predicts the label of the ground truth instance. The source of this error is often a weakness in the action classification module.

$$tIoU(g_t^{(k)}, p_t^{(i)}) \geq \alpha \text{ and } g_l^{(k)} \neq p_l^{(i)} \quad (2)$$

Localization Error (LOC). A prediction with the correct label that has a minimum 0.1 tIoU and fails to meet the α tIoU threshold with the ground truth instance. The source of this error is typically a weakness in the localization module and/or the temporal feature representation.

$$0.1 \leq tIoU(g_t^{(k)}, p_t^{(i)}) < \alpha \text{ and } g_l^{(k)} = p_l^{(i)} \tag{3}$$

Confusion Error (CON). A prediction of the wrong label that has a minimum 0.1 tIoU but does not meet the α tIoU threshold with the ground truth instance. This error is due to a combination of the same error sources in WL and LOC.

$$0.1 \leq tIoU(g_t^{(k)}, p_t^{(i)}) < \alpha \text{ and } g_l^{(k)} \neq p_l^{(i)} \tag{4}$$

Background Error (BG). A prediction that does not meet a minimum 0.1 tIoU with any ground truth instance. This error could arise in large percentages due to a weakness in the prediction scoring scheme.

$$tIoU(g_t^{(k)}, p_t^{(i)}) < 0.1 \tag{5}$$

Another error source of relevance to our analysis is the miss detection of ground truth instances, *i.e.* **False Negative (FN)**. In Sect. 7, we analyze why some type of instances are typically miss detected by current algorithms.

5 False Positive Analysis

In this section, we take the four state-of-the-art methods (SC, CES, IC, and BU) as an example to showcase our FP analysis procedure. First, we introduce the concept of a *False Positive Profile*, the mechanism we employ to dissect a method's FP errors. Then, we present insights gathered from the methods' FP profiles. Finally, we investigate each error type's impact on the average-mAP$_N$.

False Positive Profile. The computation of average-mAP$_N$ relies inherently on the ranking of predictions. Thus, it is important to take the prediction score into account when analyzing FP errors. Thus, we execute our analysis on the error profile of the top-10G predictions, where G is the number of ground truth instances. We pick the top predictions in a per-class manner, *i.e.* we select the top-10G_j predictions form class j, where G_j is the number of instances in class j. Moreover, to see the trend of each error type, we split the top-10G predictions into ten equal splits and investigate the breakdown of the five FP error types (defined in Sect. 4) in each split. The collection of these error breakdowns allow us to model the error rate of each type as a function of the prediction score. This intuitively allows us to examine the behaviors of the different detector components such as the classifier and scoring function.

We choose to focus on the top-10G predictions instead of all predictions for the following four reasons: (i) 10G is sufficiently large to show trends in error types; (ii) Current state-of-the-art methods exhibit an extremely low normalized

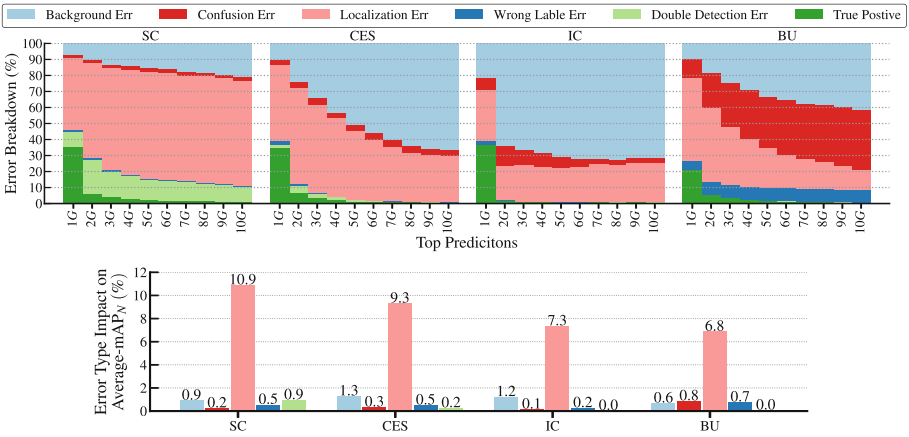

Fig. 5. Top: The false positive profiles of the four methods. Each profile demonstrates the FP error breakdown in the top-10G predictions. Bottom: The impact of error types on the average-mAP$_N$, *i.e.* the improvement gained from removing all predictions that cause each type of error. The *Localization Error* (pink bar) has the most impact. (Color figure online)

precision ($<0.05P_N$) beyond this large number of predictions. An error analysis at such low precision point is not insightful as predictions' quality degrades and the background errors dominate; (iii) The average-mAP$_N$ for the top-10G is very close to the performance of all predictions (Table 1); and (iv) It is easier to compare the FP profiles of multiple methods on the same number of predictions.

What can FP Profiles Tell Us? Figure 5 (Top) shows the four methods' FP profiles. The top-G predictions in all methods contain the majority of TP. SC is the best in terms of background error rate, while IC is the worst since the majority of its predictions beyond the top-G are background errors. This indicates a shortcoming in IC's scoring scheme. On the other hand, SC has a relatively high double detection error rate. We attribute this to the fact that SC is purely a proposal method (*i.e.* it is optimized for a high recall) combined with a video-level classifier that is independent of the proposal generation. However, this double detection rate can be fixed by applying a stricter *non-maximum-suppression* (NMS). Notably, errors due to incorrect labels (*i.e.* wrong label and confusion errors) are relatively small for the top three methods. This signals the strength of these methods' classifiers. At the same time, we can see a high wrong label and confusion errors for BU, indicating a weakness in BU's classifier.

FP Categories Impact on the Average-mAP$_N$. The insights we get from the FP profile help us to identify problems in algorithms, however, they do not tell us which problem we should prioritize fixing. In order to address this, we quantify the impact of an error type by measuring the average-mAP$_N$ after fixing that error, *i.e.* we calculate the metric after removing all predictions causing the given error type. Figure 5 (Bottom) shows the impact of the five errors on

the performance of the four methods. Fixing localization errors gives a significant boost to average-mAP$_N$, while fixing other error types provides limited improvements. This is a compelling evidence that localization error is the most significant error to tackle and that the research field should focus on addressing this error in order to advance detection algorithms.

6 Average-mAP$_N$ Sensitivity

Typically, researchers design localization algorithms to tackle certain action characteristics. For example, multiple works have tried to capture the temporal context along the video [8,12,24,35,45] as a proxy for the localization of the actions. Indeed, the recent SSN architecture, presented in this study as CES, is the latest successful work in this school of thought. In this case, the architecture not only describes each segment proposal, as typically done by template based methods, but it also represents the adjacent segments around to influence the instances localization. Although, these ideas are well motivated, it is unclear if changes in performance with respect to a single metric actually corresponds to representative changes on instances with the characteristics of interest. In that sense, another important component in the diagnosis of localization algorithms is the analysis of AP variation with respect to the actions characteristics.

Figure 6 (Left) shows CES's performance variations over all the action characteristics described in Sect. 3. Each bar represents the performance after dropping all the instances that do not exhibit a particular characteristic, and the dash bar represents the performance of the method over all the instances in the dataset. In contrast with the analysis of FP profiles (Sect. 5), all four methods exhibit similar variation trends across multiple action characteristics. Refer to the *supplementary material* for the other methods' sensitivity figures.

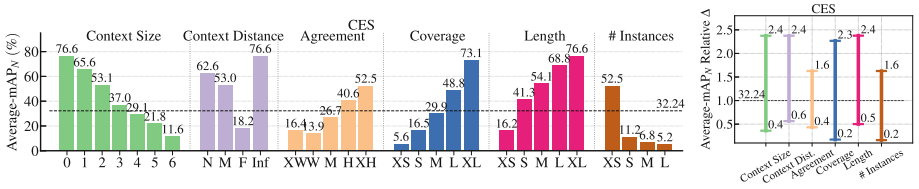

Fig. 6. Left: The detailed sensitivity CES's average-mAP$_N$ to action characteristics. The dashed line is the overall performance. Each bar measures the average-mAP$_N$ on a subset of ActivityNet for which a particular action characteristic holds. Right: The sensitivity profile summarizing the left figure. The difference between the max and min average-mAP$_N$ represents the sensitivity, while the difference between the max and the overall average-mAP$_N$ denotes the impact of the characteristic.

Among the interesting patterns to highlight, we find instances where humans tend to agree more on the starting and end of the action translates to gains in

performance (H-XH agreement in the figure), while the opposite behavior shows a drop in performance. This correlation is a bit surprising considering that the models are not trained with multiple annotations per instance. Unfortunately, we do not find any concluding evidence to explain this interesting correlation besides the nature of the instances or the bias of the dataset. Similarly, instances where an action occurs naturally surrounded by enough temporal evidence that reinforces its presence are associated with drops in performance (context size of 5–6 in the figure). We argue that this is due to the presence of similar actions around the instances which creates a confusion and impedes precise boundaries positioning around the instance. In terms of coverage and instance length, the results are intuitive and easy to interpret. Short instances, either absolute in time or in relationship to the video length, tend to be more difficult to detect. This could be a consequence of the coarse temporal structure used by the algorithms to accumulate temporal evidence across the video.

Figure 6 (Right) summarizes, in a sensitivity profile, the variations in CES's average-mAP$_N$ for each characteristic group, as well as the potential impact of improving the robustness. According to our study, all the methods exhibit a similar trend, they are more sensitive to variations in the temporal context, coverage, and length compared to variations in the agreement and number of instances. Based on experiments with ideal classifiers, [36] hypothesizes the temporal agreement between annotators is not a major roadblock for action localization. Interestingly, our diagnostic analysis shows the first experimental evidence corroborating this hypothesis. Considering the small positive and negative impacts of these instances, efforts to improve in this area must be validated carefully such that improvements do not come from more common and easier cases. Our analysis also justifies researchers' focus on designing algorithms that exploit temporal context for localization. Three out of the four models studied here would benefit the most by improving the contextual reasoning along the temporal scale.

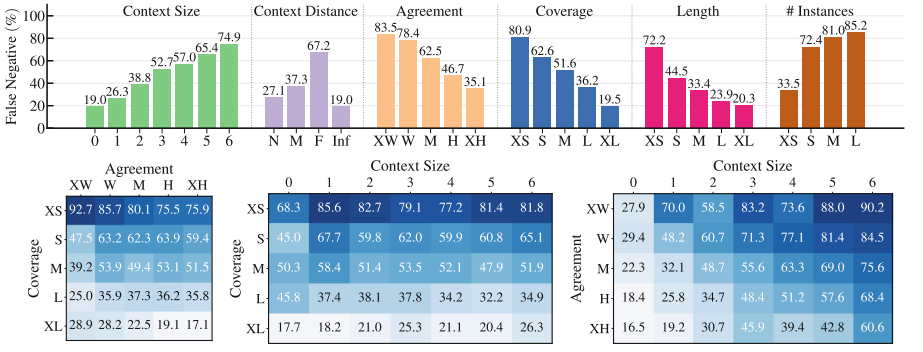

Fig. 7. Average false negative rate across algorithms for each characteristic (Top) and three pairs of characteristics (Bottom). We observe characteristics such as XS coverage and XW agreement are hard to detect individually and when paired with others. Differently, instances with XL coverage and XH agreement are relatively easy.

7 False Negative Analysis

So far, we have only considered the types of FP errors introduced by the detections algorithms, and the characteristics of the actions that introduce more variations in the performance. On the other hand, it is insightful to study what makes an action instance difficult to detect, even at minuscule confidence threshold. Towards this end, we compute the percentage of missed detections instances for each algorithm and group them according to the action characteristics defined in Sect. 3. For this purpose, we consider that an action instance is miss detected if we do not find a matching detection at a precision level higher than $0.05P_N$. Figure 7 (Top) summarizes our findings. In the interest of saving space, we average the results across multiple algorithms (refer to the *supplementary material* for the results of each algorithm by itself). The first observation that we can grasp from the results is its inverse relationship with the sensitivity profiles shown in Fig. 6. For example, the drops in performance we observed for instances with extremely weak agreement, low coverage, short length, or high temporal context size match the evidence that the algorithms struggle to retrieve such instances. On the other hand, we can appreciate that algorithms are struggling to find multiple instances per videos. Note how the amount of missed detections increase more than double due to the presence of another instance in the video. This is definitely an area where methods should focus on to mitigate the negative impact in performance. For context distance, the pattern is intuitive given that the increase in cumulative size of the *context glimpses* correlates with the spread of context and confusion in time. Thus, the chance of delimiting the start and end of the instance get worse.

Finally, we also find some interesting patterns in the FN rate at the intersection between two groups of characteristics. Figure 7 (Bottom) compactly summarizes those in a similar fashion. It is interesting how particular pairwise combinations such as low coverage (XS) - large context size (6), extremely weak agreement (XW) - large context size (6), and low coverage (XS) - extremely weak agreement (XW) are very difficult to detect, even when some are well represented in the dataset. Similarly, we find pairs involving high agreement (XH), small context (0), and high coverage (XL) relatively easy to detect. Finally, we find some interesting contours in the pairwise interactions, *e.g.* the percentage of FN diffuses in the matrix of agreement v.s. context size as we move from the top right corner to the bottom left in a non-smooth way.

8 Discussion and Conclusion

We introduced a novel diagnostic tool for temporal action localization and demoed its application by analyzing four approaches in the latest ActivityNet action localization challenge. We showed how our proposed methodology helps detection methods not only to identify their primary sources of FP errors but also to reason about their miss detections. We provided a detailed categorization of FP errors, which is tailored for action localization specifically. Using this categorization, we later defined our proposed *False Positive Profile* analysis. We found that the FP profile varies across methods. Some of the techniques exhibited shortcomings in their scoring function, while others showed weaknesses in

their action classifier. We also investigated the impact of each error type, finding that *all* detectors are strongly hurt by localization errors. We conducted an extensive dataset characterization, which empowered us with a deeper understanding of what makes an action instance harder to localize. We introduced and collected six new action characteristics for the ActivityNet dataset, namely, context size, context distance, agreement, coverage, length, and the number of instances. We measured methods' sensitivities to these action characteristics. We observed all methods are very sensitive to temporal context. Also, we showed that temporal agreement between annotators is not a significant barrier towards improving action localization. For future work, we plan to explore new metrics for action localization that incorporate the inherent ambiguity of temporal action boundaries. In our *supplementary material*, we present a preliminary study that exploits the newly collected temporal annotations to ease the strict performance computation current evaluation frameworks use. With the release of our diagnostic tool, we aim to empower the temporal action localization community with a more in-depth understanding of their methods' failure modes. Most importantly, we hope that our work inspires the development of innovative models that tackle the current flaws of contemporary action localization approaches.

Acknowledgments. This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2017-3405.

References

1. Alwassel, H., Caba Heilbron, F., Ghanem, B.: Action search: spotting actions in videos and its application to temporal action localization. In: Ferrari, V. (ed.) ECCV 2018, Part IX. LNCS, vol. 11213, pp. 253–269. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_16
2. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Niebles, J.C.: SST: single-stream temporal action proposals. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 6373–6382 (2017)
3. Caba Heilbron, F., Barrios, W., Escorcia, V., Ghanem, B.: SCC: semantic context cascade for efficient action detection. In: CVPR (2017)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Niebles, J.C.: ActivityNet: a large-scale video benchmark for human activity understanding. In: CVPR 2015, pp. 961–970 (2015)
5. Caba Heilbron, F., Lee, J.Y., Jin, H., Ghanem, B.: What do I annotate next? An empirical study of active learning for action localization. In: Ferrari, V., et al. (eds.) ECCV 2018, Part XI. LNCS, vol. 11215, pp. 212–229. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_13
6. Caba Heilbron, F., Niebles, J.C., Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 1914–1923 (2016)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July, 2017, pp. 4724–4733 (2017)

8. Dai, X., Singh, B., Zhang, G., Davis, L.S., Chen, Y.Q.: Temporal context network for activity localization in videos. In: ICCV, pp. 5727–5736 (2017)
9. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: DAPs: deep action proposals for action understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 768–784. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_47
10. Escorcia, V., Dao, C.D., Jain, M., Ghanem, B., Snoek, C.: Guess where? Actor-supervision for spatiotemporal action localization. CoRR abs/1804.01824 (2018)
11. Everingham, M., Eslami, S.M.A., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The Pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. IJCV **111**(1), 98–136 (2015)
12. Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R.: Turn tap: temporal unit regression network for temporal action proposals. In: ICCV (2017)
13. Ghanem, B., et al.: ActivityNet challenge 2017 summary. CoRR abs/1710.08011 (2017)
14. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016)
15. Goyal, R., et al.: The "something something" video database for learning and evaluating visual common sense. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 5843–5851 (2017). https://doi.org/10.1109/ICCV.2017.622
16. Gu, C., et al.: AVA: a video dataset of spatio-temporally localized atomic visual actions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018 (2018)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
18. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 340–353. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_25
19. Idrees, H., et al.: The THUMOS challenge on action recognition for videos "in the wild". Comput. Vis. Image Underst. **155**, 1–23 (2017)
20. Jiang, Y.G., et al.: THUMOS challenge: action recognition with a large number of classes (2014). http://crcv.ucf.edu/THUMOS14/
21. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
22. Kay, W., et al.: The kinetics human action video dataset. CoRR abs/1705.06950 (2017)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
24. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
25. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: ACM on Multimedia Conference, MM 2017 (2017)

26. Lin, T., Zhao, X., Shou, Z.: Temporal convolution based action proposal: submission to ActivityNet 2017. CoRR abs/1707.06750 (2017)
27. Moltisanti, D., Wray, M., Mayol-Cuevas, W.W., Damen, D.: Trespassing the boundaries: labeling temporal bounds for object interactions in egocentric video. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 2905–2913 (2017)
28. Monfort, M., et al.: Moments in time dataset: one million videos for event understanding
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
30. Ronchi, M.R., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: ICCV 2017, pp. 369–378 (2017)
31. Russakovsky, O., Deng, J., Huang, Z., Berg, A.C., Li, F.: Detecting avocados to zucchinis: what have we done, and where are we going? ICCV 2013, pp. 2064–2071 (2013)
32. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. IJCV **115**(3), 211–252 (2015)
33. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: CVPR (2017)
34. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage CNNs. In: CVPR (2016)
35. Sigurdsson, G.A., Divvala, S., Farhadi, A., Gupta, A.: Asynchronous temporal fields for action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
36. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? In: ICCV 2017, pp. 2156–2165 (2017)
37. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 510–526. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_31
38. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
39. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV (2015)
40. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
41. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. CoRR abs/1711.07971 (2017)
42. Xiong, Y., Zhao, Y., Wang, L., Lin, D., Tang, X.: A pursuit of temporal accuracy in general activity detection. CoRR abs/1703.02716 (2017)
43. Xu, H., Das, A., Saenko, K.: R-C3D: region convolutional 3D network for temporal activity detection. In: ICCV (2017)
44. Zhang, S., Benenson, R., Omran, M., Hosang, J.H., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR 2016, pp. 1259–1267 (2016)
45. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV 2017, October 2017