Action Search: Spotting Actions in Videos and Its Application to Temporal Action Localization (Supplementary Material)

Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem

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

1 Action Spotting: What do Humans do?

In this section, we expand two analysis presented in Section 3 of the main paper: (i) Single vs. multiple target search, and (ii) Spotting vs. localization. Additionally, we show the user interface used to collect the *Human Searches* datasets and present statistics for the collected data.

Single vs. multiple target search. Our aim is to investigate whether humans are distracted when finding one instance among multiple targets as opposed to finding an instance from a single target. Towards this end, we study the performance of Amazon Mechanical Turk workers (Turkers) on conducting both modalities of the search task: single vs. multiple target search. We conduct experiments on 20 videos from the AVA [8] training set and randomly pick the target actions from the dataset ground truth. To avoid the bias of participants remembering video content, we only allow each Turker to do one task type for a given video. 105 Turkers participated in our experiments and completed a total of 400 HITs, for each of which we paid $0.1. As described in the main paper, Turkers observe a larger amount of the video when finding one instance among multiple targets. We measure the total number of steps (observed frames) a Turker requires at different target set sizes. Figure 1 (**Left)** shows that when the targets set size is 10 and 20, Turkers make 1.9 and 2.1, respectively, times the search steps in the single target search.

Spotting vs. localization. The experiment's goal is to compare the efficiency of human annotators when asked to spot or localize actions in videos. To execute this experiment, we asked Turkers to annotate the start of a target action. We employ 88 Turkers to annotate 30 videos from the THUMOS14 [9] training set. Turkers completed a total of 400 HITs, each of which was paid $0.1. As described in the paper, we define spotting as finding any temporal instance within the boundaries of an action. Conversely, localization (in this experiment) refers to finding the exact start time of the target action. Figure 1 (**Right**) compares the average number of steps the Turkers require for the two different search task

The first two authors contributed equally to this work.

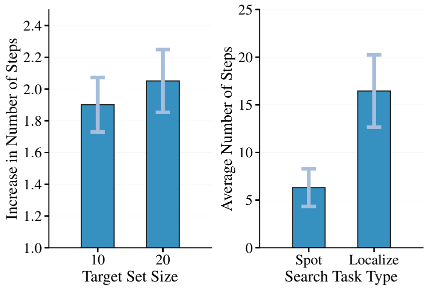
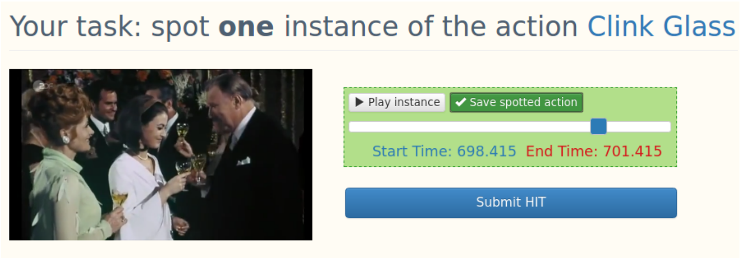


Fig. 1: **Left:** We report the increase in the number of search steps (wrt. the single target search) human annotators make when finding actions among multiple search targets. The number of steps significantly increases when the human annotators are presented with multiple targets to choose from. **Right:** We compare the average number of search steps human annotators make to spot an action (*i.e.* land anywhere inside the action temporal bounds) vs. localize the same action (*i.e.* define the start time of the action). The number of steps triples when the human annotators are asked to localize the same action.


Your task: spot **one** instance of the action Clink Glass

Start Time: 698.415 End Time: 701.415

Fig. 2: User interface screen-shot. Our user interface includes a time bar, which allows Turkers to navigate over the video quickly until the action is found.

types. Our experiments reveal that human annotators observes roughly *three times* more of the video to localize than simply spot an action.

***Human Searches* annotation details.** To collect the *Human Searches* datasets, we design a user interface (Figure 2) that includes a time bar, which allows human annotators to navigate over the video quickly until the target action is found. To collect *AVA searches* and *THUMOS14 searches*, we employ more than 800 Turkers which were paid \$0.1 for completing a single HIT. We provide per action category statistics of the collection for each dataset in Table 1.

Table 1: AVA and THUMOS14 searches datasets annotation details. For each activity class, we count the number of collected search sequences (# sequences), average number of steps over the collected search sequences (# Steps), and the number of videos annotated (# Videos).

AVA Searches

Action Class	**# Sequences**	**# Steps**	**# Videos**
Clink Glass	51	221.4	10
Dance	520	95.3	42
Drink	320	179.8	73
Drive (*e.g.* a Car, a Truck)	175	130.3	47
Eat	355	148.7	67
Hand Clap	242	149.1	34
Hand Shake	159	194.4	39
Jump/Leap	96	164.2	41
Kiss (a Person)	230	148.1	48
Martial Art	193	63.4	14
Play Musical Instrument	228	121.6	31
Push (Another Person)	82	186.6	37
Shoot	42	182.6	10
Smoke	422	115.4	59
Work on Computer	47	182.8	10
All	3162	152.2	124

THUMOS14 Searches

Action Class	**# Sequences**	**# Steps**	**# Videos**
Baseball Pitch	89	31.1	11
Basketball Dunk	67	34.6	10
Billiards	100	37.9	10
Clean and Jerk	67	27.7	10
Cliff Diving	93	33.0	10
Cricket Bowl	106	39.2	17
Cricket Shot	72	34.9	16
Diving	79	31.2	20
Frisbee Catch	94	38.2	10
Golf Swing	105	36.4	11
Hammer Throw	98	27.7	10
High Jump	64	44.2	10
Javelin Throw	97	24.5	10
Long Jump	94	39.1	11
Pole Vault	85	28.9	10
Shot put	98	34.8	10
Soccer Penalty	103	34.0	10
Tennis Swing	91	34.2	10
Throw Discuss	92	21.0	11
Volleyball Spike	67	22.8	10
All	1761	31.0	200

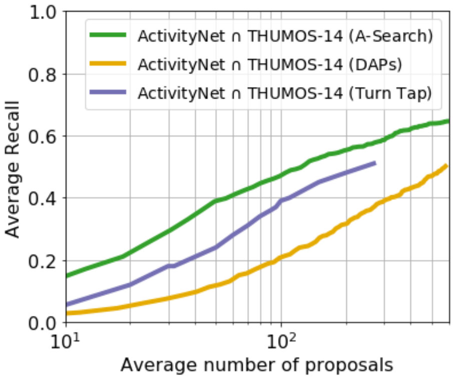


Fig. 3: Proposal metric performance of *Action Search*, DAPs [5], and Turn Tap [7] on ActivityNet ∩ THUMOS14.

2 Action Search for Action Spotting

Additional Baseline Methods. Here we give two additional baselines: *Binary Search Baseline* and the ideal *Perfect Binary Search*.
Binary Search Baseline (BSB): This model is similar to the *Direction Baseline* model, but instead of picking the next search location randomly from a uniform distribution on the search interval predicted by the *direction network*, *BSB* always picks the middle point in the search interval.
Perfect Binary Search (PSB): This model is similar to the *BSB* model, but it uses a perfect *direction network*, *i.e.* it uses the ground truth direction. This is an ideal model as it is very difficult to train a *direction network* with perfect testing accuracy.
Results. On average, *Action Search*, *BSB*, and *PBS* spot an action in 109, 115, 76 observations. While the ideal *PBS* requires less observations, it is quite the challenge to improve *BSB*'s direction algorithm to reach the ideal *PBS*. *Action Search* is the first model of its kind to predict *both* the search direction and step size, which are both key components for efficient action spotting.

3 Action Search for Action Localization

Action Search on ActivityNet. To demonstrate that our model's performance generalizes to other temporal action localization datasets beyond THUMOS14 [9], we conduct an additional experiment on ActivityNet v1.2 [2]. We evaluate *Action Search* (trained only on THUMOS14 [9]) on the ActivityNet validation videos with the same THUMOS14 classes (*i.e.* ActivityNet ∩ THUMOS14). Fig. 3 shows *Action Search* outperforming DAPs [5] and TURN TAP [7] in terms of the ActivityNet proposal metric. Moreover, *Action Search* achieves

Table 2: Temporal localization results (mAP at tIoU) on the THUMOS14 [9] testing set. We assign '–' to unavailable mAP values. We report the average percentage of observed frames (**S**) for each approach. **Our method** (*Action Search + Priors* + Res3D + S-CNN) achieves state-of-the-art results while observing only 17.3% of the video.

	mAP at tIoU							
Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	**S**
Karaman *et al.* [10]	1.5	0.9	0.5	0.3	0.2	–	–	100
Wang *et al.* [16]	19.2	17.8	14.6	12.1	8.5	–	–	100
Caba *et al.* [3]	36.1	32.9	25.7	18.2	13.5	–	–	100
Escorcia *et al.* [5]	–	–	–	–	13.9	–	–	100
Oneata *et al.* [11]	39.8	36.2	28.8	21.8	15.0	–	–	100
Richard *et al.* [12]	39.7	35.7	30.0	23.2	15.2	–	–	100
Yeung *et al.* [18]	48.9	44.0	36.0	26.4	17.1	–	–	40 †
Yuan *et al.* [19]	51.4	42.6	33.6	26.1	18.8	–	–	100
Shou *et al.* [14]	47.7	43.5	36.3	28.7	19.0	–	–	100
Tran *et al.* [15]	–	–	40.6	32.6	22.5	12.3	6.4	100
Shou *et al.* [13]	–	–	40.1	29.4	23.3	13.1	7.9	100
Gao *et al.* [7]	54.0	50.9	44.1	34.9	25.6	–	–	100
Dai *et al.* [4]	–	–	–	33.3	25.6	15.9	9.0	100
Xu *et al.* [17]	54.5	51.4	44.8	35.6	28.9	–	–	100
Buch *et al.* [1]	–	–	45.7	–	29.2	–	9.6	100
Zhao *et al.* [20]	**66.0**	**59.4**	**51.9**	41.0	29.8	–	–	100
Gao *et al.* [6]	60.1	56.7	50.1	41.3	**31.0**	19.1	9.9	100
Our method	60.8	57.9	51.8	**42.4**	30.8	**20.2**	**11.1**	**17.3**

mAP=19.59% (0.5 tIoU) while observing **S**=23.2% of the frames, outperforming not only DAPs with mAP=17.04% and **S**=100% but also TURN TAP with mAP=21.8% and **S**=100%.

Extended Comparison Results on THUMOS14. We extend the state-of-the-art comparison by adding the mean Average Precision (mAP) at several temporal Intersection over Union (tIoU) thresholds (see Table 2).

[†] We assume each of the 20 models in Yeung *et al.* [18] approach observes 2% of the video and report an upper-bound of 40% (20 models times 2%) of video frames observed.

References

1. Buch, S., Escorcia, V., Ghanem, B., Fei-Fei, L., Niebles, J.C.: End-to-end, single-stream temporal action detection in untrimmed videos. In: BMVC (2017)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
3. Caba Heilbron, F., Niebles, J.C., Ghanem, B.: Fast temporal activity proposals for efficitent detection of human actions in untrimmed videos. In: CVPR (2016)
4. Dai, X., Singh, B., Zhang, G., Davis, L.S., Qiu Chen, Y.: Temporal context network for activity localization in videos. In: ICCV (2017)
5. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: Daps: Deep action proposals for action understanding. In: ECCV (2016)
6. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. In: BMVC (2017)
7. Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. In: ICCV (2017)
8. Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D.A., Toderici, G., Li, Y., Ricco, S., Sukthankar, R., Schmid, C., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
9. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014)
10. Karaman, S., Seidenari, L., Del Bimbo, A.: Fast saliency based pooling of fisher encoded dense trajectories
11. Oneata, D., Verbeek, J., Schmid, C.: The lear submission at thumos 2014 (2014)
12. Richard, A., Gall, J.: Temporal action detection using a statistical language model. In: CVPR (2016)
13. 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)
14. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR (2016)
15. Tran, D., Ray, J., Shou, Z., Chang, S., Paluri, M.: Convnet architecture search for spatiotemporal feature learning. CoRR **abs/1708.05038** (2017), http://arxiv.org/abs/1708.05038
16. Wang, L., Qiao, Y., Tang, X.: Action recognition and detection by combining motion and appearance features
17. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection (2017)
18. Yeung, S., Russakovsky, O., Mori, G., Fei-Fei, L.: End-to-end learning of action detection from frame glimpses in videos. In: CVPR (2016)
19. Yuan, J., Ni, B., Yang, X., Kassim, A.A.: Temporal action localization with pyramid of score distribution features. In: CVPR (2016)
20. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Lin, D., Tang, X.: Temporal action detection with structured segment networks. In: ICCV (2017)