Spot On: Action Localization from Pointly-Supervised Proposals

Mettes, P.; van Gemert, J.C.; Snoek, C.G.M.

DOI
10.1007/978-3-319-46454-1_27

Publication date
2016

Document Version
Other version

Published in
Computer Vision – ECCV 2016

Spot On: Action Localization from Pointly-Supervised Proposals

Supplementary Materials

Pascal Mettes*, Jan C. van Gemert‡, and Cees G. M. Snoek*

*University of Amsterdam
‡Delft University of Technology

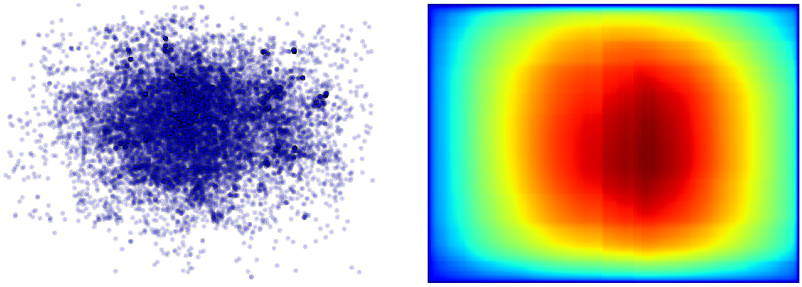
The supplementary materials for the ECCV paper "Spot On: Action Localization from Pointly-Supervised Proposals" contain the following elements regarding *Hollywood2Tubes*:

- The annotation protocol for the dataset.
- Annotation statistics for the train and test sets.
- Visualization of box annotations for each action.

1 Annotation protocol

Below, we outline how each action is specifically annotated using a bounding box. The protocol is the same for the point annotations, but only the center of the box is annotated, rather than the complete box.

- **AnswerPhone:** A box is drawn around both the head of the person answering the phone and the hand holding the phone (including the phone itself), from the moment the phone is picked up.
- **DriveCar:** A box is drawn around the person in the driver seat, including the upper part of the stear itself. In case of a video clip with of a driving car in the distance, rather than a close-up of the people in the car, the whole car is annotated as the driver can hardly be distinguished.
- **Eat:** A single box is drawn around the union of the people who are joinly eating.
- **FightPerson:** A box is drawn around both people fighting for the duration of the fight. If only a single person is visible, no annotation is made. In case of a chaotic brawl with more than two people, a single box is drawn around the union of the fight.
- **GotOutCar:** A box is drawn around the person starting from the moment that the first body parts exists the car until the person is standing complete outside the car, beyond the car door.
- **HandShake:** A box is drawn around the complete arms (the area between the union of the shoulders, ellbows, and hands) of the people shaking hands.
- **HugPerson:** A box is drawn around the heads and upper torso (until the waist, if visible) of both hugging people.
- **Kiss:** A box is drawn around the heads of both kissing people.
- **Run:** A box is drawn around the running person.



(a) Points (train). (b) Boxes (test).

Fig. 1: Annotation aggregations for the point and box annotations on *Hollywood2Tubes*. The annotations are overall center-oriented, but we do note a bias towards the rule-of-thirds principle, given the higher number of annotations on $\frac{2}{3}$-th the width of the frame.

	Training set	Test set
Number of videos	823	884
Number of action instances	1,026	1,086
Numbers of frames evaluated	29,802	31,295
Number of annotations	16,411	15,835

Table 1: Annotation statistics for *Hollywood2Tubes*. The large difference between the number of frames evaluated and the number of annotations is because the actions in Hollwood2 are not trimmed.

- **SitDown:** A box is drawn around the complete person from the moment the person starts moving down until the person is complete seated at rest.
- **SitUp:** A box is drawn around the complete person from the moment the person starts to move upwards from a laid down position until the person no longer moves upwards..
- **StandUp:** Vice versa to SitDown.

2 Annotation statistics

In Figure 1, we show the aggregated point annotations (training set) and box annotations (test set). The aggregation shows that the localization is center oriented. The heatmap for the box annotations do show the rule-of-thirds principle, given the the higher number of annotations on $\frac{2}{3}$-th the width of the frame.

In Table 1, we show a number of statistics on the annotations performed on the dataset.

3 Annotation examples

In Figure 2 we show an example frame of each of the 12 actions, showing the diversity and complexity of the videos for action localization.

(a) Answer Phone. (b) Drive Car. (c) Eat.

(d) Fight Person. (e) Get out of Car. (f) Hand Shake.

(g) Hug. (h) Kiss. (i) Run.

(j) Sit down. (k) Sit up. (l) Stand up.

Fig. 2: Example box annotations of test videos for *Hollywood2Tubes*.