Interactive Labeling for Human Pose Estimation in Surveillance Videos

[1] Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB
[2] Karlsruhe Institute for Technology KIT

Mickael Cormier[1,2], Fabian Röpke[2], Thomas Golda[1,2], and Jürgen Beyerer[1,2]

Problem Statement

- Process of annotating video data is time consuming and expensive
- Depending on skeleton model, one to two dozen context sensitive keypoints have to be annotated for each person in an image
- Especially for surveillance there are no suitable datasets
- Existing datasets come with annotation errors

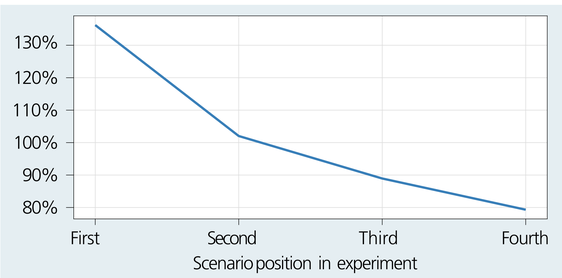


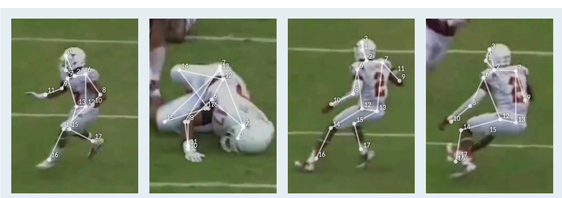
Figure 1. Relative annotation time for each scenario, depending on its random position during the experiment, compared to the mean time across all positions. 10 samples per position.

Web-based Annotation Tool

- Sloth [2] and VATIC [1] are the only open-source software for human pose estimation (require installation)
- Developed to overcome drawbacks and restrictions encountered using Sloth
  - Simple browser-side methods like inter- and extrapolation
  - Machine learning server-side backend for more complex annotation proposals
- Interactive system to improve annotation quality and speed by using technical assistance

(a) Box annotation (b) Pose suggestions (c) Final annotations

Figure 2. Provided bounding boxes, such as in (a), the pose estimation processing tool suggests pose annotations (b), that annotators accept or correct.



(a) Accurate pose estim. (b) Inaccurate pose estim. (c) Accurate interpolation (d) Inaccurate interpolation

Figure 3. Situations in which tool assistance is applicable and in which the tool does not provide a useful suggestion.

Table 1. Per-joint AP. Due to helmets and noise in the GT the AP for the three head keypoints remains quite low.

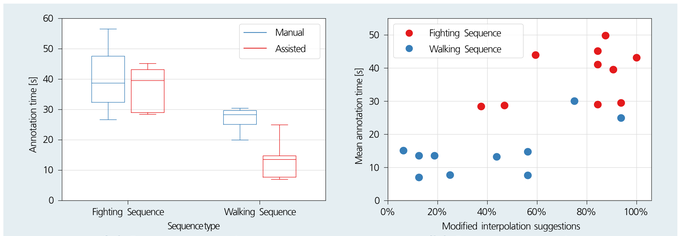
	Average Precision							
Annotation Method	**Head**	**Shou**	**Elb**	**Wri**	**Hip**	**Knee**	**Ankle**	**Total**
Mean	55.3	82.0	76.9	65.9	75.0	80.1	74.9	71.7
Most accurate (manual w. occ.)	63.3	90.0	86.9	77.7	90.2	85.1	87.5	81.7
Least accurate (manual w/o. occ.)	49.6	63.4	62.0	57.2	57.7	69.7	57.6	59.0

User Study: Aided Human Pose Annotation

- Task duration of 2 hours
- 15 body kepoints, person ID, and a full bounding box
- 27 annotators with differing levels of experience
- Over 60 experiments
  - Simple scenarios with few people and no occlusions
  - Complex scenes with multiple persons and many mutual and self-occlusions
- Results
  - 55% decreased annotation time for simple surveillance scenarios
  - Perceived workload decreased as well

Conclusion & Outlook

- Simple framework for video human pose annotation
- Improvement of annotation speed and workload for annotators
- In the future: focus on complex scenarios with dynamic occlusions



(a) Box annotation (b) Pose suggestions

Figure 4. (a) Annotation time per bounding box and pose for each of the four scenarios. Sample size is 10 per scenario. (b) Relationship between the percentage of modified interpolated annotation suggestions and mean annotation time per annotator for the walking and fighting sequences.

(a) Walking (b) Fighting

Figure 5. Frames from surveillance footage on a public place. In (a) the two persons are walking towards each other. The scene is simple: the movements are clear, monotonous and there are no occlusions. In contrast, (b) is a quite complex scene with several (self-)occlusions and unpredictable motions.

1 Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. International journal of computer vision, 101(1):184–204, 2013.
2 Sloth, https://github.com/cvhciKIT/sloth, 2011.