Software Impacts 12 (2022) 100278

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

HAVPTAT: A Human Activity Video Pose Tracking Annotation Tool

Hao Quan *, Andrea Bonarini

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy

ARTICLE INFO

Keywords:
Data labeling
Human activity recognition
Deep learning
Robotics

ABSTRACT

We propose a new semi-automatic annotation software: Human Activity Video Pose Tracking Annotation Tool (HAVPTAT). It can automatically detect and track multiple people and their pose in the video to improve work efficiency. HAVPTAT also provides the dynamical visualization of human pose, bounding boxes, person tracking ID, and possible prediction results together. The lightweight software can be launched in a few seconds and easily distributed. Its ease of use will allow non-professionals to get started quickly. This software will accelerate the development of human activity recognition models and service robots.

Code metadata

Current code version	v1
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2022-33
Permanent link to Reproducible Capsule	None
Legal Code License	GPL-3.0-or-later
Code versioning system used	Git
Software code languages, tools, and services used	C#, EmguCV (OpenCV), .NET Framework, Windows Forms
Compilation requirements, operating environments & dependencies	Visual Studio, MS Windows
If available Link to developer documentation/manual	https://github.com/AIRLab-POLIMI/HAVPTAT_annotation_tool/blob/master/README.md
Support email for questions	hao.quan@polimi.it, andrea.bonarini@polimi.it

1. Introduction

Human activity recognition is attracting increasing attention. There is a large number of publicly available datasets [1–17]. It requires a lot of labor-intensive work to annotate the video datasets. We have developed a new semi-automatic annotation tool: Human Activity Video Pose Tracking Annotation Tool (HAVPTAT). It can help dataset creators to efficiently annotate large-scale video datasets. The annotations include person tracking bounding boxes, person tracking 2-D skeleton, and activity labels. It also provides the dynamical visualization of human pose, bounding boxes, and person tracking ID, together. The prediction results obtained by an activity recognition model can also be visualized with this tool.

2. Positive impacts

There is a lack of adequate software to annotate large-scale human activity recognition video datasets collected in public spaces (*In The Wild–ITW*). People's actions are continuous and sequential in daily life, lasting at least a few seconds instead than single frames. Multiple persons or crowded scenes in a frame are often present in public spaces. The frame rate of RGB cameras in the current market is usually about 15 fps $\sim$ 30 fps. Manual annotation of the clips, person by person and frame by frame calls for enormous workload. Nowadays, skeleton-based human activity recognition from deep learning models is popular [18–24]. Single persons' tracking spatial–temporal skeletal data are essential for a model to learn and predict labels. The novel semi-automatic software HAVPTAT could fill the gap. Annotators do not need to spend time on spatial–temporal human pose detection and tracking; instead, they may work on multiple people and pose tracking data prepared by HAVPTAT, maximizing annotation efficiency.

3. Related work

Labeling is time and labor-intensive work. In general, the laboratory collected dataset like NUCLA, SYSU, NTU-RGB+D, PKU-MMD [1,2,

* Corresponding author.
E-mail addresses: hao.quan@polimi.it (H. Quan), andrea.bonarini@polimi.it (A. Bonarini).

https://doi.org/10.1016/j.simpa.2022.100278
Received 18 March 2022; Accepted 26 March 2022

8,9,25] may not have annotation problems, since the datasets are scripted, performed by actors and only contain single or few people performing single, pre-defined actions. Some researchers provide pre-defined labels and then use crowd-sourcing to label datasets, as it has been done for Charades and Something Something [26,27]. Labeling work would be painful and error-prone for datasets not collected in controlled settings. Some of them were annotated by hand, like Fine-Gym, UAV-Human, HOMAGE [11,13,28], etc. Other datasets (e.g., ActivityNet, AVA, Babel [12,17,29]) were labeled through commercial crowd-sourcing platforms like Amazon Mechanical Turk (AMT) [30], with a charge for dataset creators. Besides, annotators from crowd-sourcing platforms without formal training might skew the annotation quality. A crowd-sourcing method may leak confidentiality.

In literature, there are different open-source video annotation tools [31–36]. Only some of them have the “interpolation” functionality (e.g., VATIC [37] and CVAT [38]) to track moving targets. This functionality facilitates annotators to save a lot of time by automatically tracking annotations instead than giving labels frame by frame. These tools usually split object detection and object tracking in two different phases. Before the “interpolation” functionality could be used, it requires the annotator manually, or by using other object detection methods, to identify the interested target(s) by drawing bounding box(es) on parts of the video key frames to perform the interpolation. The quality of object detection depends on the individual annotators/detectors. The annotators may easily miss some subjects in a crowded scene. Moreover, the interpolation performance is often not perfect, so that additional effort should be spent in manual adjustments of bounding boxes. Furthermore, they do not provide human pose tracking data. Hence, it is necessary for dataset creators to separately use other pose estimation methods to extract skeletal data [39,40,40–50]. Finally, an additional tedious elaboration should be made to integrate persons’ tracking and skeletal data to obtain persons’ pose tracking data. Except for the drawbacks mentioned above, if a video has multiple targets or crowded scenes, it will be a challenge for them to execute detection and interpolation on the video which requires a lot of hardware resources, which an average PC on the current market could not afford. To the best of our knowledge, there is no annotation tool which could meet the needs of annotation for large-scale skeleton-based, human activity, video datasets.

4. HAVPTAT functionality

HAVPTAT amends most of the issues of the current open-source labeling tools. It could automatically detect and track multiple people and their pose in the video without the need of manually setting bounding box(es) and key frame(s). The annotator does not need to give action labels frame by frame, but may label just once for a person with the same action along the whole clip. HAVPTAT requires the annotator to give only a label for each different action if the same person performs multiple actions in a clip. The pose tracking with annotation data are ready without the need of further integration work. Besides, it also provides the dynamical visualization of human pose, bounding boxes, person tracking ID, and possible prediction results together. Its ease of use and efficiency will allow non-professionals to quickly get started.

The interface of HAVPTAT is shown in Fig. 1. It has been developed by .NET Framework version 4.6.1, using C# programming language for coding, *Windows Forms* library for UI, the EmguCV library (OpenCV library for .NET version) for image/video processing. It is based on JSON format data produced by the OpenPifPaf [42] model. It runs as an offline desktop application in *MS Windows*.

The upper part of the interface contains the menus composed of the available actions labels corresponding to coarse macro actions: *“Walking”*, *“Standing”*, *“Sitting”* and *“Other Actions”*. Each macro action menu contains the fine-grained detailed action such as *“WalkingWhileCalling”*, *“StandingWhileWatchingPhone”*, *“SittingWhileEating”*, etc. Users can also add customized action label button(s) by clicking the “Add” button,

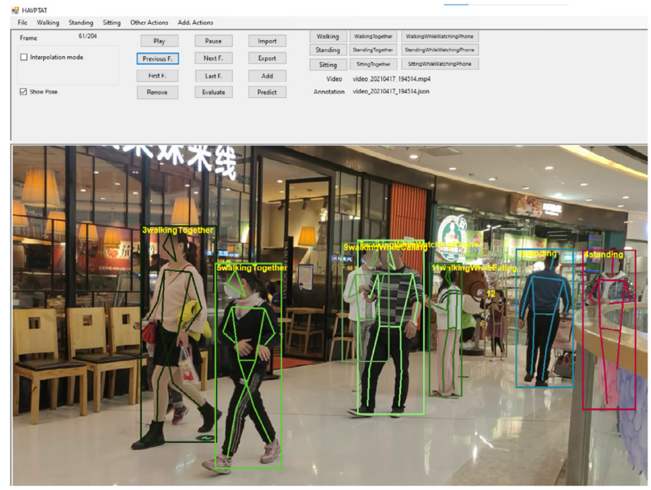

Fig. 1. The Human Activity Video Pose Annotation Tool (HAVPTAT) interface and a snapshot of a clip from POLIMI-ITW-S.[2]

writing the desired action names. The new added label button(s) will be available on the menu.

The middle part of the screen is dedicated to the main functionalities of the tool. The left region shows the current number of the frame. The interpolation mode gives the users the possibility to assign activity labels to persons for multiple frames. The “Show Pose” checkbox gives the option to show or hide the human’s poses. In the middle, there are the video control buttons such as “Play”, “Pause”, “Previous frame”, “Next Frame”, “First Frame” and “Last Frame”. The “Import” button imports the JSON file for a specific video clip that was generated by OpenPifPaf [42] from the file system. Users can click the “Export” button to export the final annotated JSON file to the file system. Moreover, users can add customized label(s) by using “Add” button. The wrong label(s) can be removed by the “Remove” button. Users may review the results of the annotation by clicking the“Evaluate” button. The “Predict” button is used to comparably visualize ground truth(s), predicted label(s) and decision of a service robot about whether to approach a person or not, a decision among “NEED SERVICE”, “MAYBE NOT NEED SERVICE”, “and NOT DISTURB”. We have defined these three decision instructions for developing a service robot application. Users could also define instruction(s) by modifying the source code for their specific application(s).

On the right, some high frequently used labels’ buttons are placed on for improving productivity. The currently used video and annotation JSON file’s names are shown in the “Video” and “Annotation” text fields.

Moreover, the tool also provides a set of keyboard shortcuts to manipulate videos such as Play/Pause (CTRL+Space), Next/Previous frame (CTRL+Right/Left Arrow).

The typical use of the tool develops along the phases described below.

First of all, the user should use OpenPifPaf [42] to generate the original videos’ keypoint annotations (human body pose estimation and tracking) and store them on file.

Then, the user uses the annotation tool to open a video file, clicking “Import” button to import the corresponding keypoint annotations previously generated by the OpenPifPaf from the file system. Then the user can start to associate action labels to persons who appear in the video by clicking bounding boxes and buttons of action names.

After having finished the action labels association for a video, the user clicks the “Export” button to store the final annotated JSON format file back to the file system.

[2] https://airlab.deib.polimi.it/polimi-itw-s-a-shopping-mall-dataset-in-the-wild.

5. Lightweight & easy to use

The layout of HAPVTAT is very similar to the major part of the *MS Windows* desktop applications. It could be directly used without a complex setup. Non-technical users can learn how to use it quickly. The complete software size is about 110 megabytes. The lightweight software could be easily deployed and distributed.

The software could reduce a large amount of annotation cost and time for large-scale video dataset creators, especially for skeleton-based human activity recognition task, which is useful to advance the development of human activity recognition models. It also supports the production of a service robotic system which could be deployed such a type of models.

6. Use case

We have used HAVPTAT to annotate a large-scale In The Wild video dataset for human activity recognition.

7. Future work

The current version of HAVPTAT is semi-automatic. Once having trained a reliable activity recognition model, we would like to update the software so that it can become a fully automatic labeling tool. We believe that it will decrease further the cost and time for dataset annotation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by a research grant from the China Scholarship Council.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.simpa.2022.100278.

References

[1] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[2] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, A.C. Kot, NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. (2019) http://dx.doi.org/10.1109/TPAMI.2019.2916873.

[3] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017, arXiv preprint arXiv:1705.06950.

[4] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, 2018, arXiv preprint arXiv:1808.01340.

[5] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, 2019, arXiv preprint arXiv:1907.06987.

[6] L. Smaira, J.a. Carreira, E. Noland, E. Clancy, A. Wu, A. Zisserman, A short note on the kinetics-700-2020 human action dataset, 2020, arXiv preprint arXiv:2010.10864.

[7] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, J. Kim, ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2020, pp. 10990–10997.

[8] J. Wang, X. Nie, Y. Xia, Y. Wu, S. Zhu, Cross-view action modeling, learning, and recognition, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656, http://dx.doi.org/10.1109/CVPR.2014.339.

[9] J. Hu, W. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2017) 2186–2200.

[10] D. Damen, H. Doughty, G.M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100, Int. J. Comput. Vis. 130 (1) (2022) 33–55.

[11] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, J.C. Niebles, Home action genome: Cooperative compositional action understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11184–11193.

[12] A.R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, M.J. Black, BABEL: Bodies, action and behavior with english labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 722–731.

[13] D. Shao, Y. Zhao, B. Dai, D. Lin, Finegym: A hierarchical video dataset for fine-grained action understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2616–2625.

[14] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, G. Francesca, Toyota smarthome: Real-world activities of daily living, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019.

[15] H. Zhao, A. Torralba, L. Torresani, Z. Yan, Hacs: Human action clips and segments dataset for recognition and temporal localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8668–8678.

[16] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, T. Murakami, MMAct: A large-scale dataset for cross modal human action understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019.

[17] C. Gu, C. Sun, D.A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, J. Malik, AVA: A video dataset of spatio-temporally localized atomic visual actions, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6047–6056, http://dx.doi.org/10.1109/CVPR.2018.00633.

[18] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 143–152.

[19] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026–12035.

[20] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.

[21] W. Peng, X. Hong, H. Chen, G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 2669–2676.

[22] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2019) 1963–1978.

[23] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1112–1121.

[24] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[25] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, PKU-MMD: A large scale benchmark for skeleton-based human action understanding, in: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, in: VSCC '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1–8, http://dx.doi.org/10.1145/3132734.3132739.

[26] G.A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, in: European Conference on Computer Vision, Springer, 2016, pp. 510–526.

[27] R. Goyal, S.E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al., The" something something" video database for learning and evaluating visual common sense, in: ICCV, Vol. 1, 2017, p. 5.

[28] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li, UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16266–16275.

[29] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[30] Amazon, Amazon mechanical turk (MTurk), https://www.mturk.com/.

[31] J.L. da Silva, A.N. Tabata, L.C. Broto, M.P. Cocron, A. Zimmer, T. Brandmeier, Open source multipurpose multimedia annotation tool, in: International Conference on Image Analysis and Recognition, Springer, 2020, pp. 356–367.

[32] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2276–2279.
[33] T.A. Biresaw, T. Nawaz, J. Ferryman, A.I. Dell, Vitbat: Video tracking and behavior annotation tool, in: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, IEEE, 2016, pp. 295–301.
[34] S. Bianco, G. Ciocca, P. Napoletano, R. Schettini, An interactive tool for manual, semi-automatic and automatic video annotation, Comput. Vis. Image Underst. 131 (2015) 88–99.
[35] M. Riegler, M. Lux, V. Charvillat, A. Carlier, R. Vliegendhart, M. Larson, Videojot: A multifunctional video annotation tool, in: Proceedings of International Conference on Multimedia Retrieval, 2014, pp. 534–537.
[36] J. Yuen, B. Russell, C. Liu, A. Torralba, Labelme video: Building a video database with human annotations, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1451–1458.
[37] C. Vondrick, D. Patterson, D. Ramanan, Efficiently scaling up crowdsourced video annotation, Int. J. Comput. Vis. 101 (1) (2013) 184–204.
[38] Intel, Computer vision annotation tool, https://github.com/openvinotoolkit/cvat.
[39] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: CVPR, 2017.
[40] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, P. Luo, Whole-body human pose estimation in the wild, in: European Conference on Computer Vision, Springer, 2020, pp. 196–214.
[41] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
[42] S. Kreiss, L. Bertoni, A. Alahi, OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association, 2021, arXiv preprint arXiv:2103.02440.
[43] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 466–481.
[44] M. Kocabas, S. Karagoz, E. Akbas, Multiposenet: Fast multi-person pose estimation using pose residual network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 417–433.
[45] R.A. Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.
[46] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, N.M. Thalmann, Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272–2281.
[47] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, B. Schiele, Posetrack: A benchmark for human pose estimation and tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5167–5176.
[48] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5386–5395.
[49] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10863–10872.
[50] M. Contributors, OpenMMLab pose estimation toolbox and benchmark, 2020, https://github.com/open-mmlab/mmpose.