Video annotation tools: A Review

Eshan Gaur, Vikas Saxena, Sandeep K Singh
Jaypee Institute of Information Technology
Noida, India
96eshan@gmail.com, vikas.saxena@jiit.ac.in, sandeepk.singh@jiit.ac.in

***Abstract*—Due to the advancements of Machine learning and vast usage of multimedia data, the image and video annotation has become quite popular now. In the area of image annotation, the researchers and specially Microsoft research lab have achieved nearly 98% accuracy. But same is not true for the video annotation. This paper presents a review of the state of the art tools being used for video annotation.**

Keywords—video annotation; video processing, machine learning;

I. Introduction

In this digital age, video content is becoming more and more popular with the increasing use of social media among the masses and exponential increase in the bandwidth allocation by the Internet Service Providers (ISPs). With the increasing consumption of video, arises one new problem of identifying the content in the said videos. Here comes the problem of Video Annotation. Unlike images, video processing consumes significantly more resources and hence can be quite a tedious task. As mentioned earlier, video annotation is a super labor-intensive task. By modern computing standards, each hour of data collected takes an upwards of 500 hours of human intervention to annotate the images separately [1]. In these exciting times of Artificial Intelligence (AI) algorithms like Deep Learning, acceleration of development in the field of Computer Vision is tremendous.

II. The current state-of-art

A. Modern computational problems

The process of annotating video or just video frames is significantly more challenging than the annotation of images. For example, treating each video frame like an image, a 10-minute video contains between 18,000 and 36,000 frames, on an average at a rate of 30–60 frames per second [2]. Therefore, going by the general consensus, frame-by-frame annotation of video is very time consuming in an age of AI and moreover it is cost prohibiting [2-7]. Hence, this approach becomes a significant roadblock for tech innovators.

B. Introduction of Crowdsourcing

One approach to counter this problem is with the implementation of Crowdsourcing [8]. Crowdsourcing platforms have proven to be a viable option when it comes to providing access to a scalable workforce and off-the shelf and ready to use annotation tools [8]. However, one of the significantdisadvantage that cannot be overlooked on behalf of the crowdsourcing platforms are the use of anonymous workers that annotate on the behalf of the crowdsourcing unit with limited annotation functionality being the second major pain point when it comes to meeting the standards of modern computer vision requirements that demand ultra-precision accuracy to minimize human fatality.

Considering a large dataset of managed workforce providers in the market with trained workers who have extensive experience doing annotation tasks and produce higher-quality training data, almost all of them require their clients to use proprietary annotation tools within their platform and restrict clients from using the annotation tool of their choice [2].

In the next section we have mentioned some of the annotation tools that that provides the user a right understanding of the actions and interactions like individuals and groups in each video frame. Ideally, the right video annotation tool maximizes annotation quality and minimizes the human effort.

III. Video annotation tools

There are a vast variety of annotation tools at our disposal and these different tools can be characterized by the operation and functionalities they can support. One very significant and helpful survey done [14] has been able to efficiently characterize and differentiate between various semantic video and image annotation tools. All of the tools have target annotation schema that they implement like individual targeting or group targeting. Then there are the shapes that the tools can process while annotating. Most of the tools can easily draw a rectangle or an eclipse but some also go for a polygon when convenient. The next major classification is the type of interface that the annotation tool provides, be it manual, semi or automatic or the ability to switch between various traceable objects. It can also include the ability to do to efficient cloud-based video annotation. There is also the classification of the platforms the annotation tool support and it can be a strong decisive factor in choosing a particular tool based on the parameters of the project that it is required for. Last but not the least is the ability of the tool to customize the result and parameters of the annotation used by the users.

Below is the overview of five of the leading video annotation tools that are currently used worldwide:

ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE

A. iVAT

iVAT [9] is an interactive video annotation tool based on the C/C++, Qt libraries for GUI and Open Computer vision libraries for the CV algorithm used throughout the tool. iVAT is open source tool and the libraries are available on the project web page. Due to the use of Qt libraries, the tool is independent and can be used as a mobile application too.

In the analysis part of the tool, it handles a video annotation session as a project and its analysis module separates the input video shots using either of the two algorithms:

- It automatically detects the input video shots using the Edit Decision File (EDL), provided as an input, or,
- Using a shot detection algorithm [8].

Once the annotation session is complete, it is associated to one or more objects of interest to be annotated as text files.

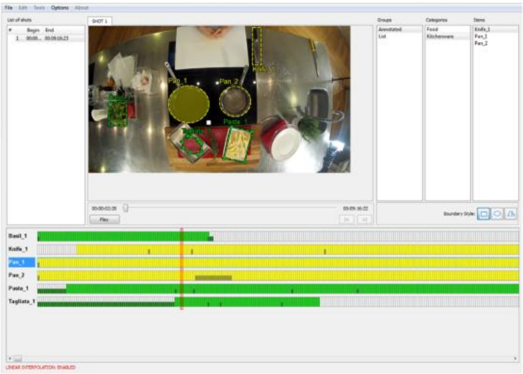


Fig 1. iVAT GUI (source: [9])

Among its main features is the support of three different annotation modules: manual, semi-automatic and automatic. These features make this tool highly flexible and suitable to be used in a variety of different applications.

B. ViTBAT

Video **T**racking and Behavior**A**nnotation **T**ool (ViTBAT) [10] is a reliable video annotation tools that generates both ground-truth information low-level tracking for individual tacking and high-level behavior recognition and analysis task for high level group tracking. ViTBAT primarily works on the states and behavior at both individual and group level. Implementation of ViTBAT is done on MATLAB with the help of its Computer Vision toolbox.

As a result of using state behavior, ViTBAT has improved tracking of targets in crowded and dynamic scenes [8]. ViTBAToffers two options when comes to annotating target state:

- point-based representation (x; y position),
- area based representation (x; y, width, height of a rectangular bounding box or an ellipse).

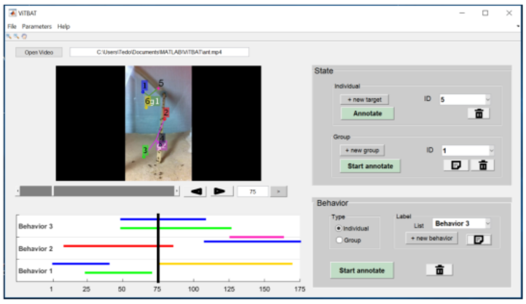


Fig 2. ViTBAT GUI (source: [10])

Furthermore, it is able to switch between different targets and the user inputs.

C. MViPER-GT

MViPER-GT [11] is the improved version of ViPER [4] ground truth control that includes a tracking system to improve the real time video analysis.MViPER-GT is a Java Open Source project that is supported by the ViPER API [4]. Users are able to select a variety of cells for the spatial attribute, i.e., point, box, or a circle.

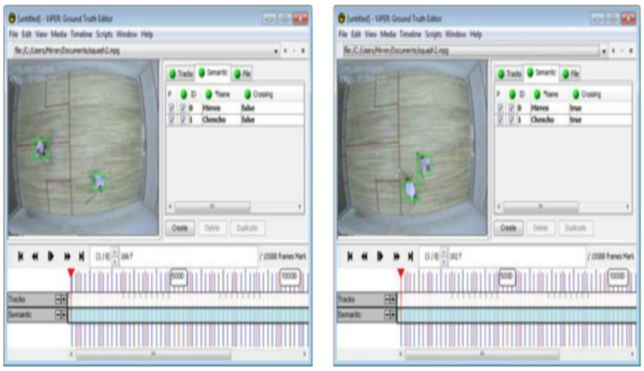

ViPER Ground Truth Editor
File Edit View Media Timeline Scripts Window Help
Tracks
Semantic
File
Create
Delete
Duplicate

Fig 3. MViPER-GT GUI (source: [11])

The MViPER-GT system works upon the already established model of ViPER and integrates a tracking module. Tracking module is used to update, establish and delete tracks automatically. Tracks here can be raw images or the annotation updates. The data from the tracking system goes to the user supervision modules which enable the user to select particular tracks. Due to this trajectory prediction algorithm [11],MViPER-GT can compute nonlinear trajectories of each track or object in a frame. After the completion of this analysis, the Annotation module annotates them and starts a new cycle.

D. BeaverDam

BeaverDam [12] is an annotation tool used for frame-by-frame box annotations. Unlike, ViTBAT [10] and MViPER-GT[11], it does not support group tracking and hence it is only useful for individual targeting. It borrows many of its concepts from the initial VATIC [8] tool hence supporting both statesand behavior annotation types. BeaverDam aims to tackle the main problem of VATIC of long setting up time and hence making it a researcher happy tool.

In comparison to the difficult and tiresome installation of VATIC, BeaverDam provides a setup script that is thoroughly tested on Ubuntu 16.04. Its script also allows the automatic configuration of Nginx and TLS to database configuration and backups, which is missing in VATIC. However, the selling point of BeaverDam is the inclusion of a web-based interface that is much superior to the command-based interface of VATIC efficiently making BeaverDam a cloud-based video annotation tool.

BeaverDam is based on Python shell interface, backed by Django Framework exposing every functionality of the python framework [12].

E. VATIC

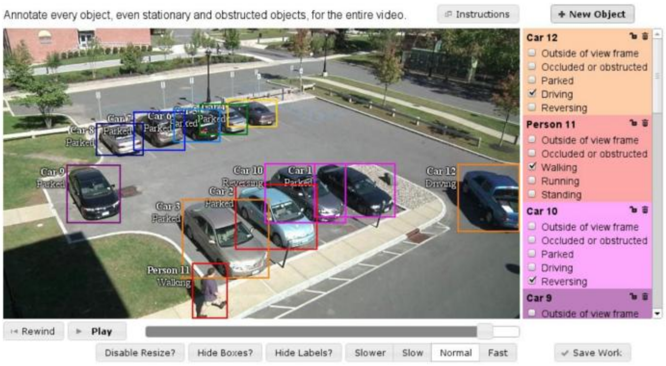
VATIC [8] is a large-scale platform independent video annotation tool that works on high quality, monetized, crowdsourced video labeling for complex videos. VATIC is employed on the Amazon's Mechanical Turk [13] where contrary to using the mainstream crowdsourcing techniques and relying on the performance of the crowd, it is able to collect and extract high quality labels by pinpointing to an expert but small group of workers who can deliver the best results.Results associated with VATIC timely indicate that the "Turk Philosophy" [8] is not able to hold in every case and that computers should assist human and not vice-versa. However, to achieve crowdsourced annotations, VATIC cannot solely utilize low -wage crowdsourcing only. The distinguishable feature of VATIC is that its users are allowed to have more than one attribute to further extend its actions. The option to seek forward and backward along with the replay speed with locking onto objects are added bonuses.

IV. ANALYSIS

As presented in Table-1, we compared the five tools based on platforms; shape of the objects; targets; interface design and kind of machine learning was used. One can compare the results from this table.In our opinion, out of these five tools, if we keep AI related issues at prime, the tool iVAT is suitable. For overall performance and ease of use the tool ViBAT is suggested.

V. CONCLUSION

Although many survey papers exist for the images and video annotation tool, in this paper we tried to fill the gap by providing the survey of only latest few tools which are not addressed or covered so far. We are only complimenting the existing survey paper and hence authors are required to read the survey papers as mentioned in the references. Based on current scenario, there is a pressing need for tools, new algorithms and very high end computing machines which can annotate videos of all domains with improved accuracy and acceptable real time performance.



Annotate every object, even stationary and obstructed objects, for the entire video.

Car 12
☐ Outside of view frame
☐ Occluded or obstructed
☐ Parked
☑ Driving
☐ Reversing

Person 11
☐ Outside of view frame
☐ Occluded or obstructed
☑ Walking
☐ Running
☐ Standing

Car 10
☐ Outside of view frame
☐ Occluded or obstructed
☐ Parked
☐ Driving
☑ Reversing

Car 9
☐ Outside of view frame

Fig 4. VATIC GUI (source: [8])

REFERENCES

[1] M. Kipp, "ANVIL: A Generic Annotation Tool for Multimodal Dialogue" Proceedings of INTERSPEECH, 2001.

[2] J. Yuen, B. C. Russell, C. Liu, and A. Torralba, "LabelMe video: Building a video database with human annotations," Proceedings of International Conference of Computer Vision, pp. 1–8, 2009.

[3] K. Ali, D. Hasler, F. Fleuret, " Flowboost: Appearance Learning from Sparsely Annotated Video", Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1433–1440, 2011.

[4] D. Mihalcik , D. Doermann, "The design and implementation of VIPER", Technical Report, 2003.

[5] I. Kavasidis, S. Palazzo, R. Di Salvo, D. Giordano, C. Spampinato, "A Semi- Automatic Tool for Detection and Tracking Ground Truth Generation in Videos" , Proceedings of 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications,pp. 6:1– 6:5, 2012.

[6] A. Ambardekar, M. Nicolescu, S. Dascalu, "Ground Truth Verification Tool (GTVT) for Video Surveillance Systems ", Proceedings of Second International Conferences on Advances in Computer–Human Interactions (ACHI), pp. 354–359, 2009.

[7] A. Yao, J. Gall, C. Leistner, L. Van Gool, "Interactive Object Detection, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pp. 3242–3249, 2012.

[8] C. Vondrick, D. Patterson, D. Ramanan, "Efficiently Scaling Up Crowdsourced Video Annotation", Proceedings of International Journal of Computer Vision, pp. 184–204, 2013.

[9] S. Bianco, G. Ciocca, P., Napoletano, R.Schettni , "An interactive tool for manual, semi-automatic and automatic video annotation", Proceedings of Computer Vision and Image Understanding, vol. 131, pp. 88-99, 2015.

[10] T. A. Biresaw, T. Nawaz, J. Ferryman and A. I. Dell, "ViTBAT: Video tracking and behavior annotation tool", Proceedings of 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 295-301, 2016.

[11] M. A. Serrano, J. García,M. A. Patricio, J. M. Molina, "Interactive Video Annotation Tool", Proceedings of Advances in Intelligent and Soft Computing (AINSC) on Distributed Computing and Artificial Intelligence, vol. 79, Springer, Berlin, Heidelberg.

[12] A. Shen, "BeaverDam: Video Annotation Tool for Computer Vision Training Labels", Technical Report, UCB/EECS-2016-193, December 8, 2016.

[13] Crowston K., A Research Tool for Organizations and Information Systems Scholars. In: Bhattacherjee A., Fitzgerald B. (eds) Shaping the Future of ICT Research. Methods and Approaches. IFIP Advances in Information and Communication Technology, vol 389. Springer, Berlin, Heidelberg., 2012.

[14] Dasiopoulou S. et.al., " A Survey of Semantic Image and Video Annotation Tools". Knowledge-Driven Multimedia Information

Extraction and Ontology Evolution. Lecture Notes in Computer Science, vol 6050. Springer, Berlin, Heidelberg, 2011.

Table.1: Comparison of Video annotation tools surveyed

Parameters vs Tool	iVAT	ViTBAT	MViPER-GT	BeaverDam	VATIC
Platforms/Programming Language used	C,C++ libraries/Qt library/Open Source Computer Vision Library	MATLAB/Computer Vision Toolbox	Java Open Source Project/ ViPER API	Cloud based/ Python GUI/Django	Amazon Mechanical Turk
Boundary/ Shape of the Object	Rectangle/Eclipse/ Polygon	Rectangle/Eclipse/ Polygon	Rectangle/Eclipse/Polygon	Rectangle	Rectangle
Targetting	Individual	Individual, Group	Individual, Group	Individual	Individual
Interface	Three modules: semi, semi-automatic and automatic detection using various detection algorithms	Improved behaviour tracking of crowds	Improved tracking for enhanced real time video analysis	Cloud interface removes the problems of command line interface of ViPER	Crowd sourced Video annotation
Customisability	Automated Tracking with the option of manula assist	Can target individual human behaviour	Capable of developing automatic annotations at the semantic level	User friendly GUI and easy to install	Suitable to work in different application domains
Machine Learning Algorithms used, if any	Supervised Obeject Detector Algoritms	No	Semi-Supervised Learning	No	No