Article published online: 2021-04-14

Novel artificial intelligence-driven software significantly shortens the time required for annotation in computer vision projects

Authors

Ulrik Stig Hansen[1], Eric Landau[1], Mehul Patel[2], Bu'Hussain Hayee[2]

Institutions

1 Cord Technologies Ltd, London, England, NW1 6NE
2 King's Health Partners Institute of Therapeutic Endoscopy, King's College Hospital NHS Foundation Trust, London SE5 9RS, United Kingdom

submitted 18.9.2020
accepted after revision 2.12.2020

Bibliography
Endosc Int Open 2021; 09: E621–E626
DOI 10.1055/a-1341-0689
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Corresponding author
Ulrik Stig Hansen, Cord Technologies Ltd, London, England, NW1 6NE, United Kingdom
Fax: +442032996474
ulrik.hansen@cord-ai.com

ABSTRACT

Background and study aims The contribution of artificial intelligence (AI) to endoscopy is rapidly expanding. Accurate labelling of source data (video frames) remains the rate-limiting step for such projects and is a painstaking, cost-inefficient, time-consuming process. A novel software platform, Cord Vision (CdV) allows automated annotation based on "embedded intelligence." The user manually labels a representative proportion of frames in a section of video (typically 5%), to create 'micro-models' which allow accurate propagation of the label throughout the remaining video frames. This could drastically reduce the time required for annotation.

Methods We conducted a comparative study with an open-source labelling platform (CVAT) to determine speed and accuracy of labelling.

Results Across 5 users, CdV resulted in a significant increase in labelling performance ($P<0.001$) compared to CVAT for bounding box placement.

Conclusions This advance represents a valuable first step in AI-image analysis projects.

Introduction

There is intense interest in artificial intelligence (AI) applications in gastroenterology [1–3]. Computer vision (CV) is a large part of the early activity in this field, assisting endoscopic detection or diagnosis [4]. To produce models of clinical utility, large amounts of high-quality labelled training data are required. This presents one of the main impediments to large-scale adoption of AI in healthcare [5], due to the cumbersome nature of labelling data coupled with the high cost of expert medical personnel performing this task (economic, societal and even environmental [6]).

A novel software platform, Cord Vision (CdV), works to address this problem by embedding automated labelling features and model functionality into the annotation process. The user is first required to manually label a relatively small number of frames in a given video sequence. These initial labels are sufficient to build a micro-model that can predict the annotations for the remaining frames, between those manually labelled. These annotations are then reviewed by the human and can be used to further train an even more accurate micro-model. CdV also includes state-of-the-art object detection and classification models to allow multiple regions of interest or abnormalities to be tracked within the same sequence. Thus, human focus shifts from performing annotations to reviewing those produced by a model. The time to complete annotation of large datasets should be significantly reduced. Here we assessed the

utility of CdV to handle the annotation of polyps in endoscopy videos.

Methods

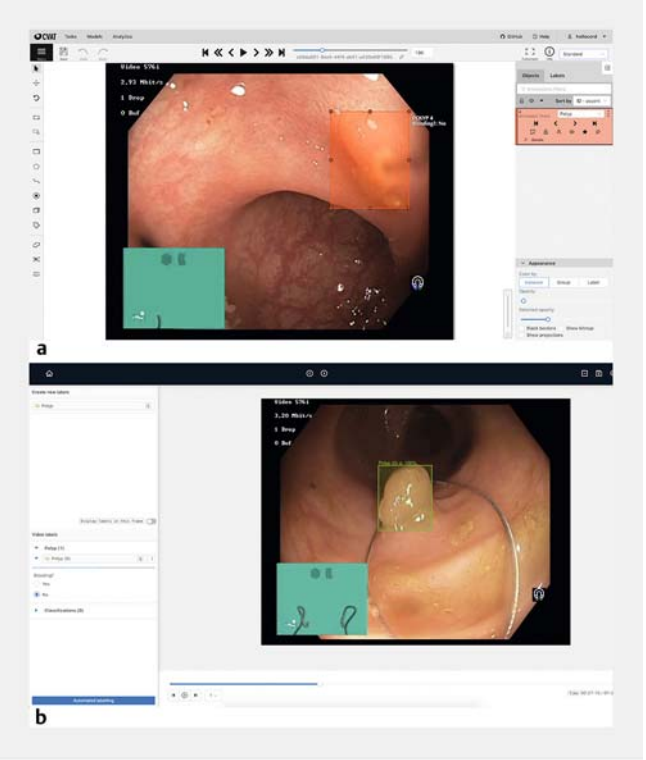
A study was conducted to compare CdV to the popular open-source Computer Vision Annotation Tool ('CVAT') developed by Intel. CVAT, like CdV, allows videos to be imported and annotated frame-by-frame (▶ **Fig. 1**). Using a subsample of polyp videos from the Hyper-Kvasir dataset [7], five independent annotators were asked to draw bounding boxes around polyps identified in videos from the dataset. A test set of 25,744 frames was used. The experiment was conducted by two annotators with good knowledge of both CdV and CVAT, and three annotators with little previous experience of CdV or CVAT. Analysis of paired labels and rate for each annotator was by Wilcoxon test with significance assumed at $P<0.05$.

Labelling experiment

An arbitrary time limit of 120 minutes was set for task completion on both platforms, to label the entire dataset following the same order of videos. The number of labelled frames completed by operators on both platforms was compared. If the dataset was exhausted before the time limit expired, the experiment was stopped.

Labellers were allowed to adopt their own labelling strategies with any functionality offered in each platform. With CVAT, this consisted of tools to draw bounding boxes and propagate them across frames using linear interpolation of box coordinates. With CdV, labellers had access to both hand labelling annotation tools and CdV's embedded intelligence features. This embedded intelligence was composed of object tracking algorithms and functionality to train and run convolutional neural networks (CNNs) to annotate the data. The object tracking algorithms track the motion of human-labelled objects through subsequent frames without requiring a prior model for those objects. They work not by simple interpolation but by trying to optimally match pixel and spatial information in regions from frames that have labels to ones that don't. They can be used in CdV starting with as little as one bounding box, but don't get more effective as more labels are added.

The CNNs, on the other hand, are models that can develop internal representations of objects by being trained with previous examples of those objects. For a maximally fair comparison to CVAT, the annotators started with no prior polyp models and set to bootstrap new models purely from the labels they themselves produced within the 120-minute time limit. While CdV allows users to integrate their own models and weights, all CNN training was done starting with the default training weights and a default Faster R-CNN architecture offered on CdV. Training parameters were also set to CdV defaults and not altered through the course of the experiment. Once models were trained, they would run inference on videos and render bounding boxes around areas they detected as polyps. Models would output a confidence score with each prediction, which could also be used to filter bounding box rendering.



▶ **Fig. 1** **a** Annotations in CVAT and **b** Cord Vision.

While the computation time for model training or inference was excluded from the timing count (as it did not require any human intervention after initiation), the time taken for the Annotators to correct or further annotate videos after micro-model labelling in CdV was included. Total frames labelled, average labelling speed (frames/min) and labelling kinetics (cumulative frames labelled every 10 minutes) were compared by Wilcoxon matched pairs sign-rank test with significance assumed at $P<0.05$.

Model quality experiment

Although micro-models produced in CdV are only used for labelling, and not for real-world medical applications, model quality is still important since higher-quality models save annotators more time in review and correction. After running all annotation experiments, we thus conducted a secondary analysis to assess the quality of micro-models trained on CdV. In particular, we were interested in how the model quality differed with the number of labelled frames it was trained with.

Because annotators used between ~500 and ~4000 labels when training their micro-models, the average precision of models as a function of total labels used in training was examined at different points around that range. To compute precision of the micro-models, we compared each model-produced bounding box with the ground-truth bounding box (if there was one) in that same frame. Ground truth boxes were from frames that had been reviewed by a medical professional. A prediction is considered accurate if the area of intersection divided by the area of union between these two boxes is over a specified

threshold (this is known as intersection-over-union or IoU). We looked at the precision averaged over IoU thresholds between 0.5 and 0.95.

Results

We note that annotators employed various strategies with CdV's features. Some trained and ran a model for each video while others annotated multiple ones using hand-labelling and object tracking before training one model to cover the remaining videos. All found significant efficiency gains over CVAT.

In the 120-minute project, a mean ± SD of 2241 ± 810 frames (< 10 % of the total) were labelled with CVAT compared to 10674 ± 5388 with CdV ($P = 0.01$). Average labelling speeds were 18.7/minute and 121/minute, respectively (a 6.4-fold increase; $P = 0.04$) while labelling dynamics were also faster in CdV ($P < 0.001$; ▶ **Fig. 2**). The project dataset was exhausted by three of five annotators using CdV (in a mean time of 99.1 ± 15.2 minutes), but left incomplete by all in CVAT.

With CdV, only 3.44 % ± 2.71 % of labels produced were hand drawn by annotators. The remainder were generated through either models or tracking algorithms. Thus, with CdV, far more labels were produced with far less initial manual input (▶ **Fig. 3**). Automated labels still required manual time for review and/or adjustment. For model generated labels, a mean of 36.8 ± 12.8 minutes of the allocated annotator time was spent looking over them frame by frame and making corrections.

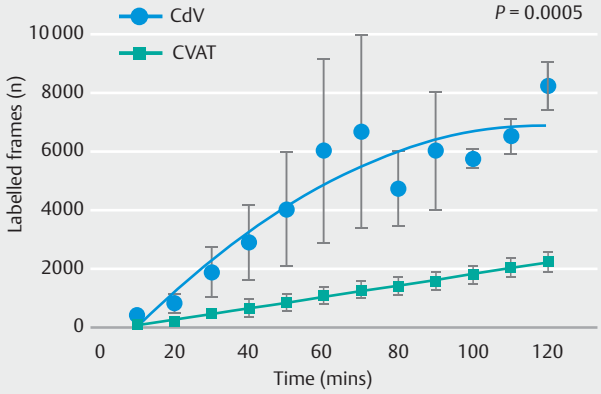
The model quality experiment demonstrated that, at around 3900 labelled frames, micro-model average precision saturated at between 61.2 % to 65.2 % (▶ **Fig. 4**). At an IoU threshold of 0.8 precision falls between 64.4 % to 72.7 % while at a threshold of 0.5 model precision rose to 95.9 % to 97.1 %. This indicates micro-models were rarely labelling completely errant regions as polyps when trained with close to 4000 frames, but that the boxes produced from these models might still require adjustment. Even adding in this additional process step, the overall time taken for a given task using CdV is still several-fold faster than a manual labelling platform.

With fewer labelled frames to train with, models could produce much worse quality outputs (▶ **Fig. 5**). We can also see this represented through how the mean and median IoU between model generated boxes and ground truth boxes changes as a function of the number of training examples (▶ **Fig. 6**). A histogram of the distribution of these IoU values for an example micro-model trained with 3972 labelled frames is also presented (▶ **Fig. 7**).

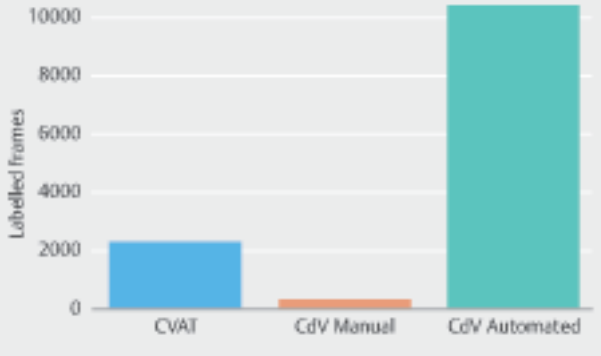
Discussion

This is the first description of software being used to annotate video frames for endoscopy, removing the need to have each frame labelled manually, and it showed a dramatic improvement in efficiency and task completion.

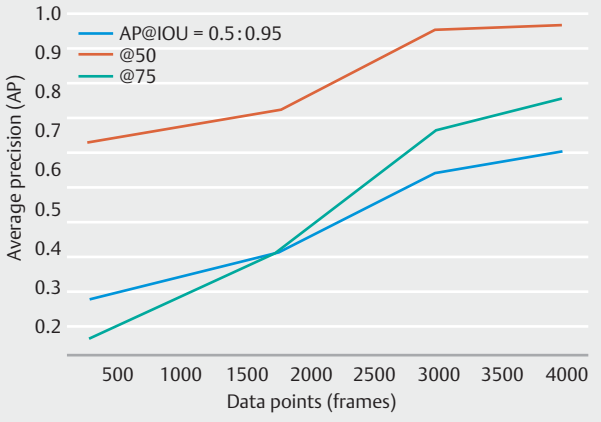
There were a few key reasons that drove efficiency gains over CVAT. Because polyp movement in videos rarely followed simple linear trajectories over long periods of time, CdV's object tracking algorithm outperformed CVAT's propagation features



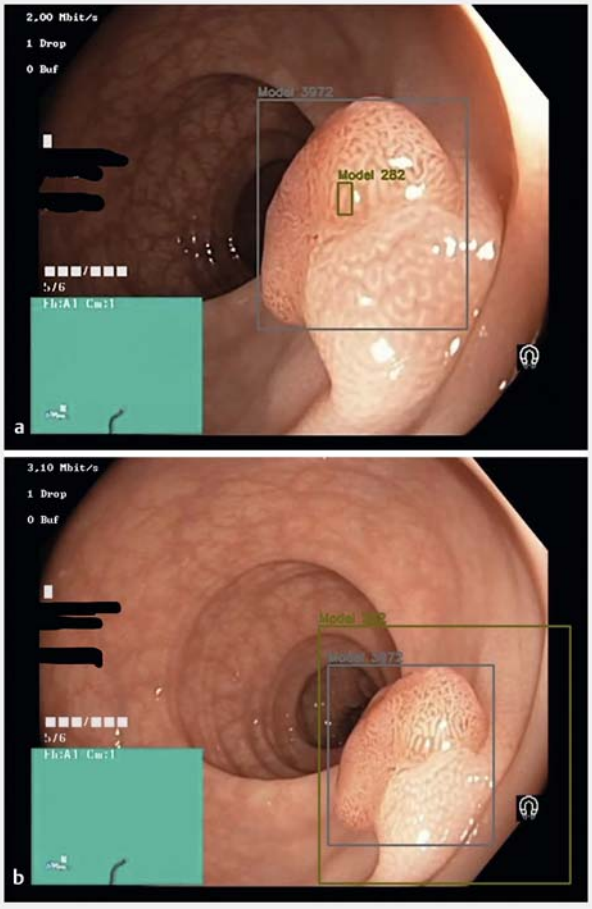
▶ **Fig. 2** Labelling kinetics in CdV and CVAT showing significant difference in speed across all annotators in the two platforms.



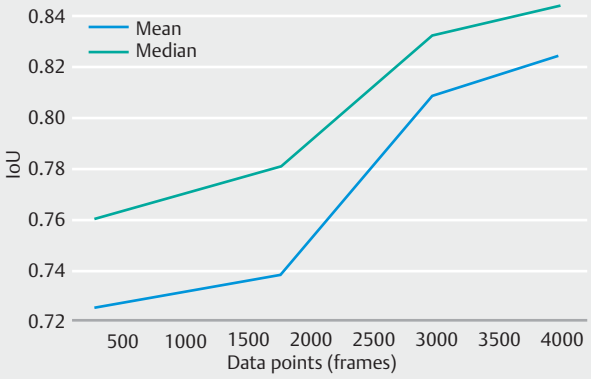
▶ **Fig. 3** Average number of labelled frames produced by annotators from CVAT, manually drawn CdV labels, and automated CdV labels.



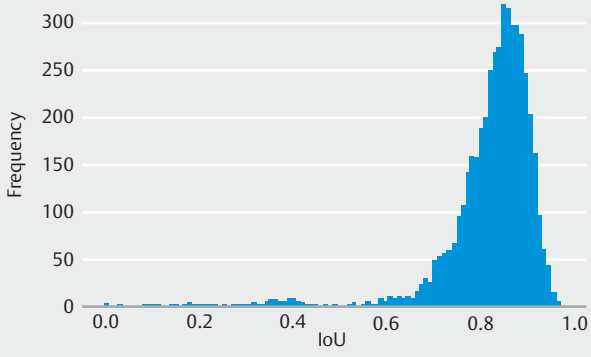
▶ **Fig. 4** Average precision (AP) for IOU threshold 50:95 and precision with threshold @50 and @75.



▶ **Fig. 5** Model outputs from micro-model trained on **a** 282 and **b** 3972 data points, respectively.



▶ **Fig. 6** Mean and median IoU between ground truth and boxes produced from micro-models trained with varying number of data points.



▶ **Fig. 7** Histogram of IoU between ground truth and boxes produced from micro-model trained with 3972 data points.

(which use linear interpolation) where propagated boxes often required frequent adjustment. CdV's object tracking algorithm, in contrast, tended to accurately follow the polyp movement relative to the scope through new frames.

With the model assistance, we found a much higher increase in efficiency within CdV simply because most labels were produced by a trained model and did not require correction. The main detriment to the model was the production of false positives, such as with labelling bubbles as polyps, that then required manual deletion. Frames that were also very blurry could produce poor results from the model.

We found the main benefit of the approach using CdV was that the first set of annotations informed the next, thus lowering the marginal time per label as the number of annotations increased. With more time, we envision training even more powerful models with the model-produced labels to assist in auto-labelling new videos.

The Cord Vision platform offers significant efficiency increases compared to CVAT for annotation of polyps. By decreasing the amount of time a gastroenterologist needs to annotate data for an AI model, the hope is that both more labels are produced to train superior models and that time is freed up for more productive activities. Although we tested for polyp bounding box annotation speed, AI models now are going beyond just detection of polyps. Future work can be done with similar studies comparing more complex labelling structures and classifications.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Gulati S, Patel M, Emmanuel A et al. The future of endoscopy: Advances in endoscopic image innovations. Dig. Endosc 2020; 32: 512–522

[2] Le Berre C, Sandborn WJ, Aridhi S et al. Application of Artificial Intelligence to Gastroenterology and Hepatology. Gastroenterology 2020; 158: 76–94.e2

[3] Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. World J. Gastroenterol 2019; 25: 1666–1683

[4] Min JK, Kwak MS, Cha JM. Overview of deep learning in gastrointestinal endoscopy. Gut Liver 2019; 13: 388–393

[5] Qayyum A, Qadir J, Bilal M et al. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 2021; 14: 156–180

[6] Dhar P. The carbon impact of artificial intelligence. Nat Mach Intell 2020; 2: 423–425

[7] Borgli H, Thambawita V, Smedsrud PH et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci Data 2020; 7: 16

Appendix

Identification of subsamples of videos from the Hyper-Kvasir dataset:

1. 0220d11b-ab12–4b02–93ce-5d7c205c7043
2. cb9da601–6ee5–44f4-ab41-a420e69f1895
3. c472275e-c791–4911-aeb2–065c4b1940b3
4. 54c32c85–21a8–4917–93e2-dfcbf4fa6cbe
5. 7821b294-f676–4bea-92c3-fd91486b18f0
6. 5fbcae8c-17d7–46c6–9cfa-e05a73586a2 d
7. 7cfdbb45-a132–4a04–8e6e-72270e3c7792