Frame Selection Methods to Streamline Surgical Video Annotation for Tool Detection Tasks

1st Jianming Yang
School of Computing
Queens University
Kingston, Ontario, Canada
22bd23@queensu.ca

2nd Rebecca Hisey
School of Computing
Queens University
Kingston, Ontario, Canada
rebecca.hisey@queensu.ca

3rd Joshua Bierbrier
School of Medicine
Queens University
Kingston, Ontario, Canada
j.bierbrier@queensu.ca

4th Christine Law
Department of Ophthalmology
Queens University
Kingston, Ontario, Canada
christine.law@queensu.ca

5th Gabor Fichtinger
School of Computing
Queens University
Kingston, Ontario, Canada
fichting@queensu.ca

6th Matthew Holden
School of Computer Science
Carleton University
Ottawa, Ontario, Canada
Matthew.Holden@carleton.ca

***Abstract*—Given the growing volume of surgical data and the increasing demand for annotation, there is a pressing need to streamline the annotation process for surgical videos. Previously, annotation tools for object detection tasks have greatly evolved, reducing time expense and enhancing ease. There are also many initial frame selection approaches for Artificial Intelligence (AI) assisted annotation tasks to further reduce human effort. However, these methods have rarely been implemented and reported in the context of surgical datasets, especially in cataract surgery datasets. The identification of initial frames to annotate prior to the use of any tools or algorithms determines annotation efficiency. Therefore, in this paper, we chose to prioritize the development of a method for selecting initial frames to facilitate an automated annotation process. We propose a customized initial frames selection method based on feature clustering and compare it to commonly used temporal selection methods. In each method, initial frames are selected to train an object detection model. The model assists in the automated annotation process by predicting bounding boxes for the remaining frames. Evaluation matrices are built upon how many edits users need to perform when annotating the initial frames and how many edits users are expected to perform to correct all predictions. Additionally, the total annotation cost for each method is compared. Results indicate that on average the proposed cluster-based approach requires the fewest total edits and exhibits the lowest total cost compared to conventional methods. These results underscore a promising direction for streamlining AI-assisted annotation processes for surgical tool detection tasks.**

***Index Terms*—Streamline Annotation, Frame Selection, Surgical Video, Tool Detection**

I. Introduction

With the increasing availability of surgical video data, topics such as surgical instrument tracking, surgical skill assessment, and computer-assisted surgery have gained interest. Many researchers have practiced integrating machine learning and AI algorithms for efficient video analysis in these tasks. These types of analyses often require annotated data. Specifically, tasks involving surgical tool detection require numerous bounding box annotations and class assignments for each tool in each frame. This heavy workload poses considerable challenges for researchers, often leading to delayed progress or reliance on a limited selection of public annotation-ready datasets.

Researchers from various fields have made significant strides in the development of annotation tools and techniques, originating from basic frame-by-frame annotation which laid the foundation for annotation practices. Over time, the tool designs shifted towards leveraging AI and other advanced algorithms to assist in video annotation tasks, further reducing repetitive manual work by generating labels by prediction. While acknowledging the merits and applicability of these methods in natural video contexts, their application to surgical data remains underreported, leading to limited evidence to validate their performance in this domain. Besides, surgical data tasks pose unique challenges compared to natural video tasks.

For instance, because of confidentiality concerns associated with surgical videos, it is crucial to avoid indiscriminately outsourcing or uploading videos to online platforms. Additionally, surgical videos may capture instances of objects moving off the main viewing plane as a result of the tools being rotated by the surgeons during operations. This is a challenge not typically addressed by common tools that rely on interpolation methods primarily designed for handling in-plane translations. Moreover, for research teams with limited personnel, a method is only beneficial if it can still ensure a manageable workload. These constraints significantly limit the choices of annotation methods available for surgical tasks and highlight the need for an efficient way to select frames to annotate. Therefore, our goal is to develop a semi-automated annotation method for tool detection tasks in surgical data. We start by

choosing initial frames selection methods to minimize the initial annotation effort while also keeping the minimal editing effort for the remaining frames. An exploration of previous achievements in annotation tools and frame selection techniques can give insights into our method development.

A. Object Detection Annotation Tools: Then and Now

The evolution of image annotation tools has been notable. Initially, static image labeling tools like Massachusetts Institute of Technology's LabelMe [1] model were employed. However, the labor-intensive nature of labeling every frame posed challenges, particularly for models like the Long short-term memory (LSTM) [2], which require input from each frame.

Efforts have also been made to automate annotation processes using tools like Intel's Computer Vision Annotation Tool [3] , Visual Geometry Group Image Annotator [4], and labelbox [5]. These tools leverage deep learning techniques and utilize pre-trained models for annotation tasks. However, in terms of ease of use, they may not always be ideal for video annotation since some of them still rely on frame-by-frame annotation methods.

Advancements by Gil-Jiménez et al. [6], introduced geometric bounding box interpolation methods. They proposed that automated annotating only requires a sparse set of frames for moving objects in a video. By providing the endpoints of the object and key intermediate frames, interpolation methods can propagate the annotations to the remaining frames.

This theory paved the way for the development of another tool genre: video-level annotation tools. They mitigate the need for extracting every individual frame and leverage interpolation algorithms to require only a few annotated frames to propagate the detection boxes to other frames. Tools like Video Annotation Tool from Irvine California (VATIC) [7] and its successor, BeaverDam [8], [9], are pioneering examples using video-level bounding box annotation methods. These tools have been used in conjunction with platforms like Amazon Mechanical Turk (crowd-sourcing marketplace) and have proven valuable for various annotating tasks [10], [11]. As mentioned, the interpolation method itself is challenged by non-linear, off-plane object movements in surgical video.

Another algorithm, Kalman filters [12], has significantly contributed to the field of object detection. Kalman filters model an object's state within a dynamic system, tracking the object's position, velocity, and acceleration over a sequence of frames. By incorporating measurement updates based on available information, the filter enhances its ability to predict an object's future state. Adapted Kalman filtering has been implemented to track the bounding box center of a target object, assisting in the annotation process. This has resulted in improvements in both labeling speed and accuracy [13].

One of the state-of-the-art automated annotator tools, V7 [14], implements both interpolation and tracking algorithms. This integration has led to promising annotation results on both natural images and videos. This success is primarily attributed to its advanced capability to account for object motions using temporal context. Temporal context refers to the sequential relationship and continuity of events over time, allowing V7 to accurately interpret object motions in images and videos by considering their past and future states. Fisher et al. [15] practiced the tool in their annotation workflow on pituitary tumor removal surgery videos. Unfortunately, similar to many online tools, some of the advanced features of V7 are only accessible through a paid version of the product.

Notably, despite the promising performance of these tools in other fields, there is a lack of explicit reporting on the annotation of any surgical dataset using the tools described above. Also, most of the advanced tools require an upload of the data, which raises confidentiality issues with proprietary surgical data. Instead, a local version would make them more suitable.

B. Frame Selection Techniques

AI-assisted annotation tools require a set of frames and their annotations for training predictive object detection models. The frame selection technique is crucial in initiating the annotation process and significantly impacts the performance of automated annotation [16]. Utilizing different selection methods results in different training sets, each representing the data differently. Training on these diverse representations should yield varying performance in predictive models, further influencing the efficiency of automated annotation processes. Therefore, this paper focuses on identifying the frame selection method that can best streamline tool bounding box annotation tasks for surgical data.

- Manual Selection - Originally, researchers relied on manual annotation to select frames. In Krenzer et al. [17], the researchers hoped to build an annotation tool for polyp (a small growth in the gastrointestinal tract) detection in Gastroenterology. They utilized a freeze-frame detection algorithm. Because experts froze videos when they detected a polyp to capture photos, they took advantage of this to select relevant frames based on the "frozen" part of the video. While effective for specific annotation tasks, this method requires the recording to have the "frozen" parts as its first layer of manual labels, making it bias-prone and impractical for different datasets.
- Temporal Selection - Selection based on the temporal order of frames is a straightforward method with one frame chosen at every fixed interval. Many current annotation tools employ this method, as it is easy to implement and requires minimal computational overhead. This approach provides a systematic and evenly spaced representation of the video.
- Selection by Feature Clustering - Feature extractors are used to convert frames into high-dimensional feature vectors containing rich spatial and semantic information. Through dimension reduction and clustering, each group of frames has a representative centroid, capturing the

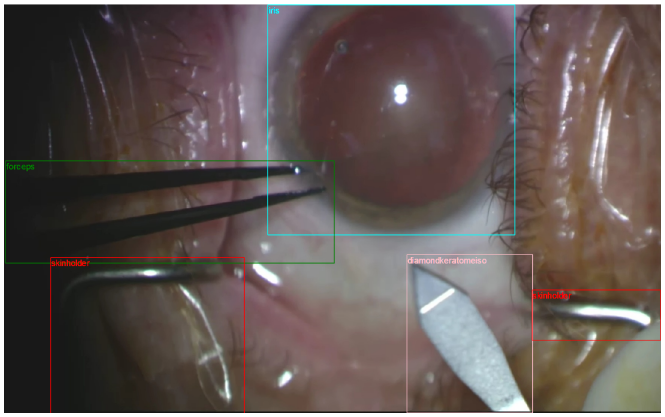

Fig. 1. Annotated Surgical Video Frame. Objects are labeled with bounding boxes corresponding to their class, location, and size.

relationship among frames and the diversity of visual content present in the video. This method has been employed on various natural videos [12], [18], [19], and laparoscopic surgery videos [20]; however, not for cataract surgery videos. Furthermore, they often report that due to the clustering methods requiring a preset number of clusters, they encounter limitations in summarizing or representing the entire dataset most efficiently.

Various other techniques have been reported for selecting representative frames. For instance, some attempts have been made to minimize energy between consecutive frames to capture significant changes [21]. Others have used autoencoders to create thumbnails for movies [22], while [16] have advanced sequential processing to extract temporal features and identify significant events in surveillance camera footage. While these methods are effective for video analysis, not all the computational efforts are necessary for AI-assisted annotation for a frame-by-frame tool detection task.

II. Materials and Methods

In this study, cataract surgery videos were acquired and annotated. While acknowledging the potential value of interpolation methods and tracking filters in the initial frame selection process, our primary objective was to establish a baseline model that considers frame selection methods as the only independent variable. To avoid the impracticality of the manual selection method and computational inefficiencies, we opted to use the temporal selection method as the benchmark method given its widespread use. We then compared it with a customized cluster-based method. Frames were extracted from the acquired videos and subjected to the selection methods. Selected frames were used to train an object detection model and generate predictions for the remaining frames. Each method's predicted bounding boxes will be compared to the ground truth bounding boxes to evaluate the performance of each automated annotation process. The goal is to assess whether a custom-designed clustering approach outperforms a conventional temporal selection method in the context of surgical data annotation.

TABLE I
All-Frame Set Videos: Number of Frames and Bounding Boxes

Video ID	Video Frames	Bounding Box Instances
1	6301	24825
2	5601	24276
3	4301	16543
4	4401	17465
5	8001	21421
6	4001	17087
7	4901	22928
8	4201	16409
9	6111	22469
10	5301	21921
11	5301	18876
12	6401	24510

A. Data Preparation

The dataset used in this study comprises 12 cataract surgery videos recorded in the operation room of Kingston Health Sciences Centers. These videos were captured using a binocular surgical microscope with monocular recording. Each video is recorded a 30 frames per second and includes the first four phases of the procedure. The videos were divided into individual frames starting from the frame with index 0, terminating after the 4th phase capsulorhexis (at frame indexes that can be completely divided by 10). Every frame was manually annotated with bounding box locations for 8 object classes: iris, eye speculum, forceps, diamond keratome straight, viscoelastic cannula, cystotome needle, diamond keratome iso, and capsulorhexis forceps (Figure 1). The data were organized into two sets: the set only includes every 10th frame (preliminary set), serving as a smaller representation of the videos to quickly test the feasibility of the workflow; and the all-frame set, encompassing every single frame, which is ideal for making the most comprehensive result (Table I).

B. Frame Selection

The first method involves frame selection by feature clustering. This method is customized to overcome the limitation in other clustering methods where the number of clusters must be predefined. The details are shown in Figure 2. We utilized a base-sized Vision Transformer (ViT) model pretrained on ImageNet-1k with mean absolute error (MAE) to extract features from all frames. These features were then dimensionally reduced using Principal Component Analysis (PCA) to facilitate clustering. We employed the affinity propagation method for clustering, which computes matrices to evaluate the similarity between data points and their suitability as exemplars for each other. Through iterative updates of these message-passing matrices, the affinity propagation method autonomously determines the number of clusters (N). Without needing a preset N like other clustering methods, the affinity propagation method has the advantage of balancing the selection size and the retained

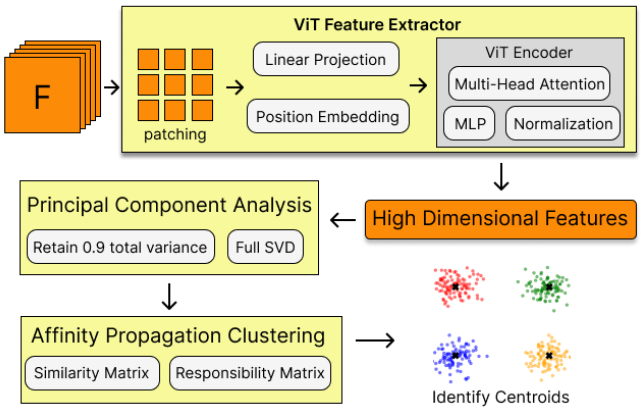

Fig. 2. Selection by Feature Clustering. In the Vision Transformer model, each frame, "F", is divided into patches, and these patches are encoded along with their positional information. The encoding process involves normalization, multi-head attention mechanisms, and multiple layer perceptrons (MLP). Instead of directly passing the encoded features to a classification layer, they are exported to a Principal Component Analysis for dimensionality reduction. The PCA threshold is set at 0.9, indicating that components capturing 90% of the explained variance are retained. This threshold is chosen with a full Singular Value Decomposition (SVD) to ensure optimal reduction. Then the Affinity Propagation algorithm utilizes similarity and responsibility matrices to select the most representative points as centroids (shown as black cross marks in colored example clusters).

information size. The centroids, which serve as exemplars, are chosen to minimize dissimilarity within their respective clusters while representing the most distinct features. These centroids are meaningful in the frame selection context, as each one not only signifies a unique event but also efficiently captures the most representative content within its cluster. Subsequently, the frames corresponding to these centroids were selected as the initial set for training.

The second method employs a common temporal selection technique, which involves selecting every 10th frame from the video sequence. This is to simulate the selection method commonly practiced by existing video annotation tools. The temporal every-10th-frame method offers the advantage of allowing the model to have a larger training dataset while seeing the full video as if it were operating at a lower frame rate. This approach enables the model to capture more details across various temporal locations within the video, enhancing its learning ability.

The third method—temporal same-size method—selects the same number of frames (N) as proposed by the first method. It divides the whole data set by N intervals and selects the first image of each interval to construct the initial set for training. This approach allows for a direct comparison with the feature clustering method because the training data size matches that of the feature clustering method.

These three methods complete the first part of the experiment workflow (Figure 3.A).

C. Training and Prediction

For both datasets (preliminary set and all-frame set) and each method, the selected frames in the initial set, along with their annotation files, are used to train a pretrained small-sized You Only Look Once Version 8 object detection network (YOLOv8) [23]. All three sets—training, validation, and test—are the same as the initial set. This is to intentionally encourage overfitting the model to the selected representative frames of the current data, with the aim of improving its predictive accuracy on the remaining frames. This approach is based on the assumption that each method considers the selected set to be representative of the entire dataset.

The remaining frames and their annotations serve as the evaluation set. The trained YOLOv8 model generates predicted bounding boxes for the frames in the evaluation set. This serves as the automated annotation process and prepares predicted results for evaluation (Figure 3.B).

TABLE II
EDIT CALCULATION METRIC FOR DIFFERENT SCENARIOS

Scenarios	Edit Count and Definition
Initial Addition	Count of bounding boxes in initial set
Correct	0 for each predicted box passing IOU threshold and with the correct class label
Deletion	1 each for predicted box not passing IOU threshold and not matching any class in the remaining classes
Addition	1 for each true truth box not matched by any predicted box
Renaming	1 for each predicted box passing IOU threshold but having the wrong class label
Reposition & Resizing	2 for each predicted box not passing IOU threshold but existing in the remaining classes

D. Evaluation

Our metric for evaluating annotation methods was inspired by Shen et al. [8], where VATIC and BeaverDam were compared. The concept originated from the notion of minimizing user interactions, quantified by the number of clicks, which served as their fundamental evaluation metric. For instance, in their study, in a standard 15-second video with 30 cars and 5 keyframes, annotating using BeaverDam required 270 click-and-drag actions compared to VATICS, which required 270 individual clicks plus 240 click-and-drag actions. Notably, the click-and-drag action was observed to take twice as long as a single click.

Our metric also considers the number of user interactions, which is quantified by the number of edits. Firstly, it considers the initial edits, which refer to the addition of bounding boxes during the annotation of the initial frame set. The number of initial edits is equal to the number of ground truth bounding boxes. Creating initial annotations involves one click for class selection and one click-and-drag to create the box.

Then, the predicted bounding boxes were compared to the ground truth bounding boxes, along with each box's class. The degree of overlap between bounding boxes is measured by the intersection over union (IOU) metric, calculated as the intersection area of the ground truth and predicted boxes divided by their union area. To assess model performance at different levels of strictness, IOU thresholds ranging from 50% to 90% were established at intervals of 5%.

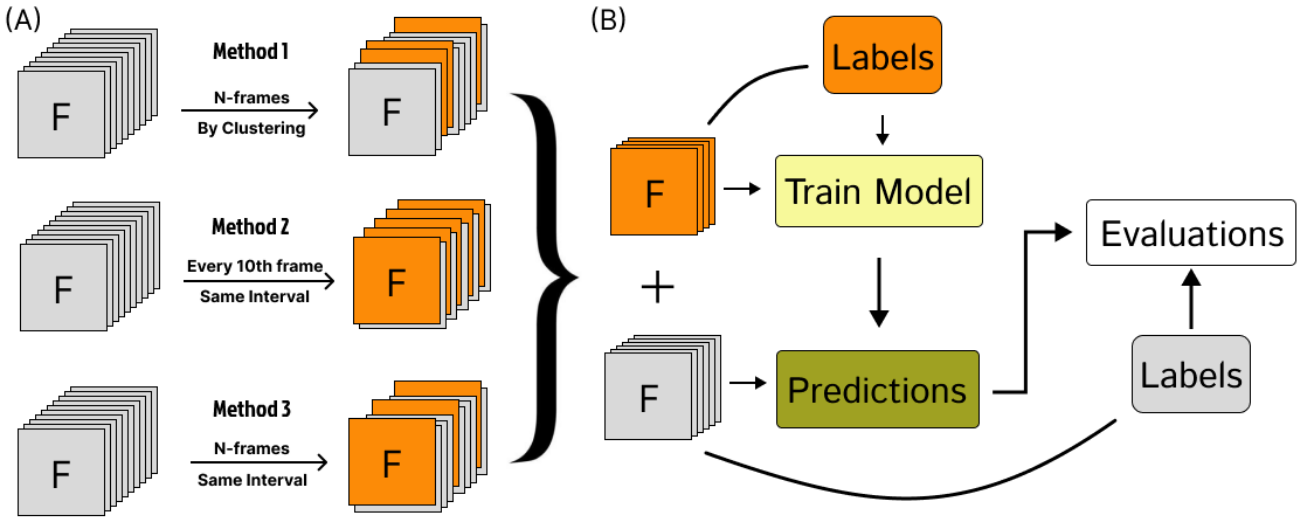

Fig. 3. General Workflow. The "F" squares denote frames, with orange squares indicating the selected frames and gray squares representing the remaining frames. N is the number of frames determined by the feature clustering method. Selected frames with their labels will be used to train the object detection model, whereas the remaining frames and their labels are responsible for prediction and final evaluation.

The edit classification workflow follows: first, identify frames containing empty truth boxes and remove all predicted boxes associated with those frames; next, identify correctly matched boxes between the predicted and ground truth sets and remove them; then find highly overlapping boxes that required renaming of the predicted box and remove the pair; following this, find boxes with matched class that need repositioning and resizing to align and remove the pair; finally, delete remaining unmatched predicted boxes and perform addition for the remaining unmatched truth boxes. This classifies the post-prediction edits in five scenarios: correct, deletion, addition, renaming, reposition and resizing. (see Table II)

While correctness does not add a count to edits or action time, deletion adds one edit for each deletion click; addition, similar to what was mentioned above, adds one edit for the two actions; renaming requires one edit for each renaming click; and repositioning and resizing together are two edits for double click-and-drag actions.

In terms of the cost of labeling, we standardize the unit as the unit of time. As previously reported [8], the cost of click-and-drags is twice that of clicks. All action costs are converted with the click time as the standard unit 1. The summation of the total action cost for each method is then quantified.

With this approach, we compute various metrics including the number of correctly labeled boxes, the total number of edits, performance on each type of edit, and the estimated time in standardized units required for all editing actions. Additionally, box accuracy is calculated using the number of correctly predicted boxes divided by the total number of boxes that need to be predicted. The average accuracy is a more comprehensive way to evaluate the performance. These metrics serve as the basis for evaluating the effectiveness of the methods.

III. Results

During the fast workflow testing using the preliminary dataset, the temporal every-10th-frame selection method generally outperformed the other two methods, with an average of 50 fewer total edit counts and a prediction accuracy of 90.3%. The feature clustering method and the temporal same-size method showed similar performances, achieving accuracies of approximately 88.5% and 89.1%, respectively.

However, averaged results from 12 all-frame videos show different outcomes (Table III). Except for manual labeling, the every-10th-frame selection results in the largest number of total edits. When also considering the edits during the initial set annotation, the every-10th-frame selection fares worse, resulting in a higher total action cost. Conversely, the Feature Clustering method exhibited the lowest total edit counts for all post-prediction criteria.

Notably, the feature clustering method demonstrated an advantage in accuracy but not significantly higher than the temporal same—size method. On the average threshold, the feature clustering method costs only 6112 units of time, saving 85.2% of time compared to manual labeling and being 9.9% faster than the temporal method with the same training data size on average.

IV. Discussion

This paper aims to streamline the annotation process for surgical videos. Through research, we found that most existing tools and methods have not been optimized or tested for their efficacy on surgical data. We also found that there are existing initial frames selection techniques that can impact AI-assisted annotation performance, so we focused on developing an approach that begins with the initial frame selection step.

TABLE III
AVERAGED EVALUATION MATRICES ON AVERAGE 50%-90% IOU RESULTS

Matirces	Feature Clustering	Temporal every-10th-frame	Temporal same-size	Manual
Initial Edit	1313	2725	1316	20727
Correct	17858	15138	17491	20727
Deletion	244	594	324	0
Addition	276	305	294	0
Renaming	2668	5309	3138	0
Reposition & Resizing	22	57	55	0
Accuracy	92.0%	84.1%	90.1%	100%
Total Edits	4523	6240	5124	20727
Action Cost	6112	11690	6786	41454

Throughout the experimentation, the feature clustering method demonstrated advantages over common temporal selection methods on the cataract surgical dataset when performing AI-assisted object detection annotation tasks. One key factor contributing to its success could be its ability to extract semantic information from each frame using ViT. This semantic representation enables more precise localization of objects and reduces the occurrence of false negatives (tool not labeled) or misclassifications (tool wrongly labeled) demonstrated by the lowest counts for addition and renaming.

Feature clustering methods can reduce annotation efforts compared to temporal selection approaches. While temporal methods sample frames at fixed intervals, feature clustering selects frames based on their distinctiveness and relevance to the video content. Consequently, fewer frames need to be annotated, resulting in a reduction in annotation effort and overall labeling cost.

The higher accuracy of the temporal same-size method compared to the temporal every-10th-frame method is likely attributed to the latter's susceptibility to overfitting. With the temporal every-10th-frame method employing a larger training set, it introduces more variance and noise, potentially leading the model to overfit to these factors, thereby resulting in the observed difference in performance.

One potential limitation to consider is the potential impact of variables such as video length, frame rate, or the total number of frames. This is observed through the comparison between the test results on the preliminary set and the all-frame set. The discrepancy could derive from the fact that in scenarios with fewer frames, the feature clustering method is constrained to selecting the most representative frames from a limited pool of information. Consequently, this results in fewer clusters, diminishing the informational richness of the centroids compared to the selection by the every-10th-frame method. Such variations could influence the choice of selection method under different conditions, requiring further investigation to determine the threshold or cutoff point for method selection.

In future research, it would be beneficial to compare the feature clustering method with all other frame selection techniques to assess their performance in terms of labeling cost, computational complexity, and user-friendliness. This report can systematically review each method and better inform researchers of the choices. It would also be interesting to test the performance of this approach using data from other types of surgery. This will provide an understanding of the generalizability of such approach across all types of surgical data. Additionally, this feature clustering method holds the potential for integration into the development of task-specific auto-annotation tools.

This research may help establish a standard methodology for evaluating frame selection methods aimed at streamlining the bounding box annotation processes. Introducing metrics using edit counts and labeling costs could offer researchers in this field a viable framework for comparing different annotation strategies.

V. CONCLUSION

Overall, the feature clustering approach offers a promising solution for surgical video object detection annotation tasks, combining advanced feature engineering techniques with adaptive clustering algorithms. Its ability to achieve high prediction accuracy and reduce annotation effort makes it a compelling choice for AI-assisted annotation for surgical videos.

REFERENCES

[1] B. C. Russell, A. Torralba, K. P. Murphy, and et al., "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, May 2008, received 06 September 2005, Accepted 11 September 2007, Published 31 October 2007. [Online]. Available: https://doi.org/10.1007/s11263-007-0090-8

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[3] B. Sekachev *et al.*, "opencv/cvat: v1.1.0," Aug. 2020. [Online]. Available: https://github.com/opencv/cvat

[4] A. Dutta and A. Zisserman, "The via annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2276–2279. [Online]. Available: https://doi.org/10.1145/3343031.3350535

[5] M. Sharma, D. Rasmuson, B. Rieger, D. Kjelkerud *et al.*, "Labelbox: The best way to create and manage training data." 2019. [Online]. Available: https://labelbox.com/

[6] P. Gil-Jiménez, H. Gómez-Moreno, R. López-Sastre *et al.*, "Geometric bounding box interpolation: an alternative for efficient video annotation," *J. Image Video Proc.*, vol. 8, 2016. [Online]. Available: https://doi.org/10.1186/s13640-016-0108-7

[7] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision (IJCV)*, June 2012. [Online]. Available: https://doi.org/10.1007/s11263-012-0564-1

[8] A. Shen, "Beaverdam: Video annotation tool for computer vision training labels," Dec. 2016. [Online]. Available: https://digitalassets.lib.berkeley.edu/techreports/ucb/text/EECS-2016-193.pdf

[9] E. Gaur, V. Saxena, and S. K. Singh, "Video annotation tools: A review," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018, pp. 911–914. [Online]. Available: https://doi.org/10.1109/ICACCCN.2018.8748669

[10] "Detecting motorcycle helmet use with deep learning," *Accident Analysis Prevention*, vol. 134, p. 105319, 2020. [Online]. Available: https://doi.org/10.1016/j.aap.2019.105319

[11] V. Stepanyants, M. Andzhusheva, and A. Romanov, "A pipeline for traffic accident dataset development," in *2023 International Russian Smart Industry Conference (SmartIndustryCon)*, 2023, pp. 621–626. [Online]. Available: https://doi.org/10.1109/SmartIndustryCon57312.2023.10110794

[12] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Fluids Engineering*, vol. 82, no. Series D, pp. 35–45, 1960. [Online]. Available: https://doi.org/10.1115/1.3662552

[13] B. Wang, V. Wu, B. Wu, and K. Keutzer, “Latte: Accelerating lidar point cloud annotation via sensor fusion, one-click annotation, and tracking,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 265–272. [Online]. Available: https://doi.org/10.1109/ITSC.2019.8916980

[14] V7 Labs, “V7 Labs.” [Online]. Available: https://www.v7labs.com/

[15] E. Fischer, K. J. Jawed, K. Cleary, and et al., “A methodology for the annotation of surgical videos for supervised machine learning applications,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 9, pp. 1673–1678, September 2023, received 12 January 2023, Accepted 14 April 2023, Published 28 May 2023. [Online]. Available: https://doi.org/10.1007/s11548-023-02923-0

[16] K. Dai, J. Zhao, L. Wang, D. Wang, J. Li, H. Lu, X. Qian, and X. Yang, “Video annotation for visual tracking via selection and refinement,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 276–10 285. [Online]. Available: http://doi.org/10.1109/ICCV48922.2021.01013

[17] A. Krenzer, K. Makowski, A. Hekalo *et al.*, “Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists,” *Biomed Eng Online*, vol. 21, no. 1, p. 33, May 2022. [Online]. Available: https://doi.org/10.1186/s12938-022-01001-x

[18] M. Lux, O. Marques, K. Schöffmann *et al.*, “A novel tool for summarization of arthroscopic videos,” *Multimedia Tools and Applications*, vol. 46, no. 2-3, pp. 521–544, 2010. [Online]. Available: https://doi.org/10.1007/s11042-009-0353-1

[19] Y. Sun, P. Li, Z. Jiang, and S. Hu, “Feature fusion and clustering for key frame extraction,” *Mathematical Biosciences and Engineering*, vol. 18, no. 6, pp. 9294–9311, 2021. [Online]. Available: https://doi.org/10.3934/mbe.2021457

[20] M. Ma, S. Mei, S. Wan *et al.*, “Keyframe extraction from laparoscopic videos via diverse and weighted dictionary selection,” *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, Aug. 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2020.3019198

[21] C. Panagiotakis, A. Doulamis, and G. Tziritas, “Equivalent key frames selection based on iso-content distance and iso-distortion principles,” in *Proceedings of the 8th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS07)*, 2007, p. 29. [Online]. Available: https://doi.org/10.1109/WIAMIS.2007.41

[22] Y. Xu, F. Bai, Y. Shi *et al.*, “Gif thumbnails: Attract more clicks to your videos,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, Feb. 2–9 2021, pp. 3074–3082. [Online]. Available: https://doi.org/10.1609/aaai.v35i4.16416

[23] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLO,” Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics