VIVA: Visual Exploration and Analysis of Videos with Interactive Annotation

Anita Ruangrotsakun
ruangroc@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Dayeon Oh
ohda@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Thuy-Vy Nguyen
nguythu2@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Kristina Lee
leekr@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Mark Ser
serm@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Arthur Hiew
hiewa@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Rogers Ngo
dongrog@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Zeyad Shureih
shureihz@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Roli Khanna
khannaro@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Minsuk Kahng*
minsuk.kahng@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

ABSTRACT

This paper presents VIVA, a novel interactive tool for visually exploring long videos and searching for specific moments. Previous work on video data exploration and analytics often assumes that manually-created, rich annotations are available. However, such metadata may not be easily obtained. We design an interactively machine learning workflow for users to rapidly create annotations along a timeline. Combined with VIVA's focus+context visualization that effectively displays frame snapshots in the context of a video stream, VIVA enables users to explore and analyze long video clips by incrementally make sense of them. We present usage scenarios that demonstrate how users would use VIVA for video-related tasks.

CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools*; *Visual analytics.*

ACM Reference Format:
Anita Ruangrotsakun, Dayeon Oh, Thuy-Vy Nguyen, Kristina Lee, Mark Ser, Arthur Hiew, Rogers Ngo, Zeyad Shureih, Roli Khanna, and Minsuk Kahng. 2023. VIVA: Visual Exploration and Analysis of Videos with Interactive Annotation. In *28th International Conference on Intelligent User Interfaces (IUI '23 Companion), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3581754.3584160

*Minsuk Kahng is now at Google Research.

IUI '23 Companion, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0107-8/23/03.
https://doi.org/10.1145/3581754.3584160

1 INTRODUCTION

Videos are being produced at an unprecedented scale. These include long multi-hour videos, such as live-streams and event recordings [11, 28–30, 33]. A critical challenge exists in exploring these videos and searching for their contents because they are *unstructured and multi-modal data*—a video consists of a long sequence of images and a continuous audio stream. Thus the first step for video analysis often involves extracting metadata, such as entities tagged to time segments (e.g., persons, places). However, existing work on video interfaces often assumes that manually annotated data are already available [16, 18, 23, 24, 32], which is not always the case. While the problem of obtaining annotated data interactively from users has been widely studied in the literature of interactive machine learning [3, 8–10, 15, 17, 34], applying it to the domain of video exploration and analysis remains a challenging problem.

In this paper, we present a new interactive tool that supports visual information seeking and analytic tasks for videos when annotations only partially exist. This problem poses the following two questions:

(1) How can we design user interfaces for exploring a long video? Can we effectively display the snapshots of video frames (which are always available from raw files)?
(2) How can we develop user workflows for quickly and interactively creating annotations to support their analysis? Can we integrate modern deep learning techniques into the process?

1.1 Design Considerations

Overview and Filtering of Frame Images (Q1). We aim to apply Shneiderman's information seeking mantra, *"Overview first, Zoom and filter, then Details-on-Demand,"* [26] to videos. If we consider a video as a sequence of images, then our Q1 can be turned into

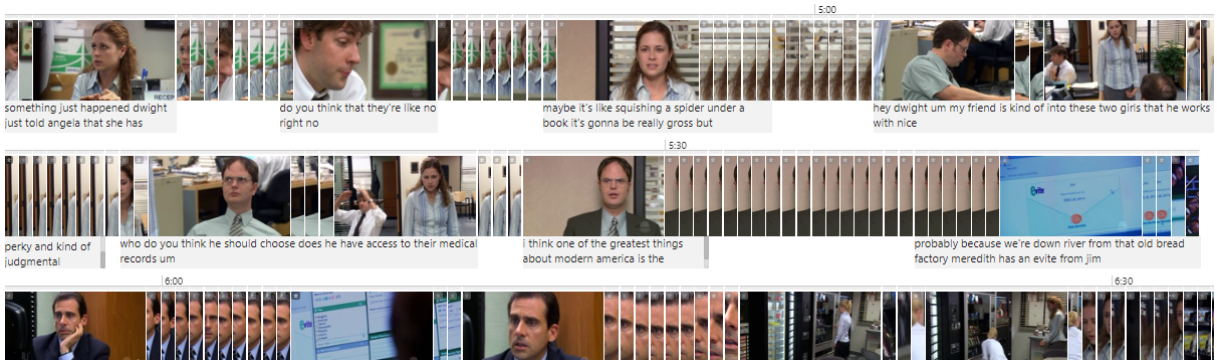

Figure 1: VIVA's focus+context view applied to a sequence of video frames for overview. Frames that are visually very different from the previous frame are fully shown (i.e., focus), while the others are collapsed to provide the context around the fully shown frames. In this example, VIVA ingests the video of the episode 'E-Mail Surveillance' in Season 2 of The Office TV series.

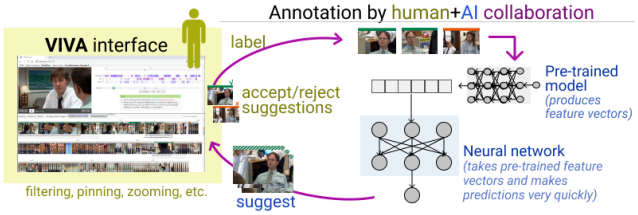

Figure 2: VIVA's human-AI collaboration approaches to rapidly creating entity annotations

a matter of supporting *overview* and *zooming/filtering* for a long sequence of frame images that may be tagged with entities. To address this problem, we adapt the *focus+context* technique [6], a classic but effective technique for visualizing data where sequential context is crucial. This brings several design considerations, such as how we determine frames to "focus", how we display "context", and how we let users "filter" frames.

Deeper Exploration by Annotating High-level Concepts (Q2). To support the *filtering* operations for video frames, the frames need to be annotated. While some of them may be annotated without human inputs by using pretrained models (e.g., known types of objects like cars and traffic lights [19]), *high-level concepts* that are specific to users may not be fully-automated (e.g., a family member appearing at a special occasion) [9, 34]. We take an interactive machine learning (iML) workflow to minimally take human inputs and let users refine machine predictions, which is illustrated in Figure 2.

Design Goals. With these considerations in mind, we derive the following design goals:

- G1. Provide an *overview* of a video using focus+context techniques applied to frame snapshots;
- G2. Support *filtering* with various annotated information;
- G3. Allow users to annotate entities using a machine-in-the-loop interactive workflow; and
- G4. Enable incremental sensemaking of and faceted searching within videos through a tight integration of the frame overview and annotation workflow.

2 INTRODUCING VIVA

We present VIVA, a novel interactive tool for visually exploring and analyzing videos, consisting of a focus+context visualization of videos and an iterative annotation workflow for further filtering.

2.1 Video Frame Overview with Focus+Context

The frame view, the main panel of VIVA, features a new *focus+context* visualization of video frames. As depicted in Figure 1, it displays video frames, video snapshots sampled once per second, along with their captions. Most frames are cropped to 10% of the width of the original image, adapted from the literature on cropping video frames [7, 13, 20, 31]. Frames that are fully shown serve as focal points (focus) while the cropped frames serve as context within a video stream (context), for an overview without losing context. To determine a set of default focal points, we adapted temporal segmentation algorithms to detect when a scene changes and we determine the very first frame of each segment as a keypoint [5, 14].

User interactions to maximize and filter frames. We design multiple ways for users to interact with the frame view. Users can maximize collapsed frames (with hover; click to pin), zoom into multiple collapsed frames (with context menu), filter frames by captions or entity annotations, and find similar frames by using pretrained image embeddings, as in Figure 3.

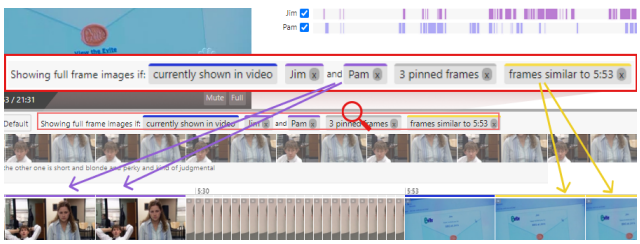

Figure 3: Multiple filters applied to frames. In this example, fully shown frames include the frame shown in the video player, those in which both Jim and Pam are present, those that are pinned by a user, and those that are similar to the frame at 5:53.

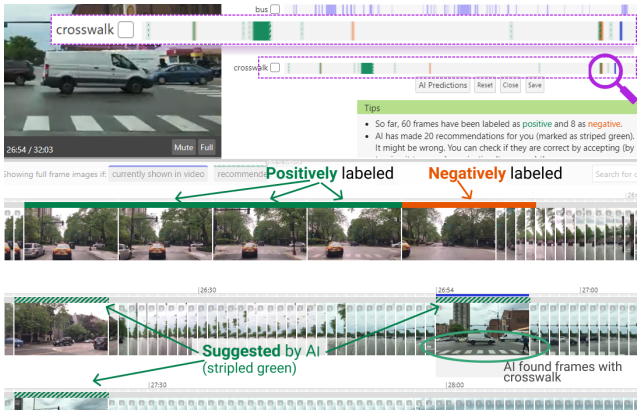

Figure 4: Using VIVA's interactive annotation workflow, a user creates a "crosswalk" while exploring driving videos. AI-suggested frames are shown as full images and highlighted with striped green, and the user can accept/reject them. The example video used is from https://youtu.be/wdIu8XdqtWM.

2.2 Deeper Exploration with Interactive Annotation

Even with the focus+context view described above, making sense of long videos can still be challenging for users because of the limited number of available attributes. To address this issue, VIVA enables users to rapidly annotate video frames (e.g., person A appearing at 2:00-2:15, 3:10-3:40), so that users can *filter* frames using this information. We design an iterative labeling workflow that use an interactive deep ML model with the following steps:

- **Step 1: Labeling frames.** For a specified entity (e.g., crosswalk), a user applies binary labels to a set of frames (positive shown as green at the top of each frame or negative as dark orange). Figure 4 illustrates our design.
- **Step 2: Training AI to make suggestions.** Once the user labels several frames, a deep model is trained and recommends additional frames that can possibly be annotated from the rest of all the frames (sorted by prediction scores).
- **Step 3: Validating results.** The predictions made by the model can be incorrect. VIVA allows users to accept or reject these suggestions by clicking the frames (changing from striped green to either green or orange).
- **Step 4: Repeating the process.** Users can label more frames and retrain the model to refine the results.

Incremental sensemaking of videos. With this workflow, users can incrementally make sense of videos and perform high-level exploration and complex tasks by gradually adding more annotations to VIVA [22, 25]. When first launching VIVA, no annotations may exist and users must rely on the frame view, which makes it relatively difficult to perform complex searching tasks. However, by creating annotations for entities, they will have more options for filtering, enabling them to explore and analyze videos in multiple ways and gain a deeper understanding of video contents.

2.3 Implementation Details

VIVA is a web application implemented with Flask, a Python web framework, for server-side, and Svelte [1], a JavaScript web framework, for client-side. For the binary classifier used in the annotation workflow, a neural network with two hidden layers takes as input a pre-trained representation of a frame obtained from VGG16 [27]. When training data contains more positively labeled frames than negative ones, an additional randomly sampled set of unlabeled images are labeled as negative to balance the numbers of positive and negative images, which improves performance. A single annotation cycle consisting of training on tens of labeled frames and inference on a video with thousands of frames takes often less than 10 seconds on a modern desktop that does not have a GPU, which enables rapid iterations, but it varies depending on many factors.

3 USAGE SCENARIOS

This section presents a scenario of how VIVA can be used. Our accompanying video demo showcases additional use cases (e.g., exploring academic talks in conference recordings; curating self-driving datasets from driving footage).

Finding Special Guests on TV Show Episodes. Consider Jane, a video editor in the entertainment industry who is tasked with creating a teaser clip of Melvina, a surprise guest, from the episode 'The Dinner Party' in Season 2 of *The Office* series. Melvina is not a main character in this TV series, but she does appear as Dwight's plus-one. Thus, Jane decides to create a label for "Melvina" by using the "Dwight" label available which has already been created by her colleague since Dwight is one of the main characters. Jane filters the frames using the "Dwight" label to collapse frames that do not include Dwight. Then she quickly skims through the filtered frames until she finds frames that include both Dwight and Melvina. Jane then creates a new "Melvina" label by using VIVA's annotation workflow. She first identifies some frames depicting both of these characters and then verifies other frames recommended by AI. After Jane believes she has found most of the segments in which Melvina appears, she now proceeds with splicing together a short clip of Melvina's appearances by using her "Melvina" label as a reference. This saves much time for Jane. Without VIVA, she would potentially have to watch the entire video to find relevant segments.

4 PRELIMINARY QUALITATIVE STUDY

We conducted a preliminary qualitative study to investigate how users would use VIVA. We recruited 12 student participants from our university (6 female, 6 male) and each had an one-hour Zoom session with us. They are asked to use VIVA for two videos (i.e., The Office TV series episode and driving footage), and for each video, perform four searching tasks (e.g., find when Dwight arrives at party with a lady), one of which preceded by a labeling task (e.g., annotate the lady).

We interestingly observed how successful participants annotated entities and how VIVA's frame view affects their annotation process. Successful participants first labeled a sufficient number of frames, which allowed the AI to provide good results on the first iteration. They also selected frames from various scenes by quickly inspecting through collapsed frames with the help of VIVA's focus+context frame view. This provides some diversity in the AI's input, which

is known to improve the performance. While reviewing the AI's suggestions, the participants could successfully identify frames that were missed in the first iteration. They initially missed some frames that were collapsed; however, the AI determines these frames to be relevant and the tool fully displays them for review. This demonstrates how VIVA's approach to adaptively selecting a set of frames to be fully displayed based on the context aids users in identifying video segments of their interest.

5 CONCLUSION AND FUTURE WORK

In this paper, we present an interactive prototype that integrates information visualization principles and interactive machine learning workflows into the domain of video exploration and analysis. This work leaves several future research directions. The effectiveness of both the focus+context technique and the annotation workflow can be rigorously evaluated with quantitative evaluations. Furthermore, many variations of the interactive ML workflow exist in optimizing the user's time and the model's accuracy. For example, a model may suggest video frames which it is uncertain about (like in *active learning*) [4], possibly with explanations [12]. Lastly, on the technical side, future work can investigate the use of more complex deep models that have been actively developed by the AI community, such as object tracking, multimodal models [2, 21].

ACKNOWLEDGMENTS

We thank Giuseppe Raffa, Rahul Khanna, and Kai Ishikawa for their feedback. This work was supported in part by NSF Industry-University Collaborative Research Center on Pervasive Personalized Intelligence, Google Cloud (GCP19980904), NSF and USDA-NIFA AI Institute (2021-67021-35344), DARPA (N66001-17-2-4030), and NAVER AI Lab.

REFERENCES

[1] 2016. Svelte: Cybernetically enhanced web apps. https://svelte.dev/. Accessed on February 16, 2023.
[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems* 34 (2021), 24206–24221.
[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
[4] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2017. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 298–308.
[5] Tat-Seng Chua, Shih-Fu Chang, Lekha Chaisorn, and Winston Hsu. 2004. Story boundary detection in large broadcast news video archives: techniques, experience and trends. In *Proceedings of the 12th annual ACM international conference on Multimedia*. 656–659.
[6] Andy Cockburn, Amy Karlson, and Benjamin B Bederson. 2009. A review of overview+ detail, zooming, and focus+context interfaces. *ACM Computing Surveys (CSUR)* 41, 1 (2009), 1–31.
[7] Ork de Rooij, Jarke van Wijk, and Marcel Worring. 2010. Mediatable: Interactive categorization of multimedia collections. *IEEE Computer Graphics and Applications* 30, 5 (2010), 42–51.
[8] Ork de Rooij and Marcel Worring. 2013. Active bucket categorization for high recall video retrieval. *IEEE Transactions on Multimedia* 15, 4 (2013), 898–907.
[9] Dazhen Deng, Jiang Wu, Jiachen Wang, Yihong Wu, Xiao Xie, Zheng Zhou, Hui Zhang, Xiaolong Zhang, and Yingcai Wu. 2021. EventAnchor: Reducing Human Interactions in Event Annotation of Racket Sports Videos. In *CHI Conference on Human Factors in Computing Systems*.
[10] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
[11] C Ailie Fraser, Joy O Kim, Hijung Valentina Shin, Joel Brandt, and Mira Dontcheva. 2020. Temporal Segmentation of Creative Live Streams. In *CHI Conference on Human Factors in Computing Systems*. 1–12.
[12] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.
[13] Gaudenz Halter, Rafael Ballester-Ripoll, Barbara Flueckiger, and Renato Pajarola. 2019. VIAN: A visual annotation tool for film analysis. *Computer Graphics Forum* 38, 3 (2019), 119–129.
[14] Marti A Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*. 9–16.
[15] Benjamin Höferlin, Rudolf Netzel, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann. 2012. Inter-active learning of ad-hoc classifiers for video visual analytics. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 23–32.
[16] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus Gross, and Hanspeter Pfister. 2017. Visualizing nonlinear narratives with story curves. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 595–604.
[17] Kuno Kurzhals, Marcel Hlawatsch, Christof Seeger, and Daniel Weiskopf. 2016. Visual analytics for mobile eye tracking. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 301–310.
[18] Kuno Kurzhals, Markus John, Florian Heimerl, Paul Kuznecov, and Daniel Weiskopf. 2016. Visual movie analytics. *IEEE Transactions on Multimedia* 18, 11 (2016), 2149–2160.
[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[20] A Chris Long, Brad Myers, Juan Casares, Scott Stevens, and Albert Corbett. [n. d.]. Video Editing Using Lenses and Semantic Zooming. *Carnegie Mellon University* ([n. d.]).
[21] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 202–211.
[22] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
[23] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video lens: rapid playback and exploration of large video collections and associated metadata. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 541–550.
[24] Amy Pavel, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 181–190.
[25] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. 269–276.
[26] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.
[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *preprint arXiv:1409.1556* (2014).
[28] Tan Tang, Yanhong Wu, Yingcai Wu, Lingyun Yu, and Yuhong Li. 2021. VideoModerator: A Risk-aware Framework for Multimodal Video Moderation in E-Commerce. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 846–856.
[29] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *CHI Conference on Human Factors in Computing Systems*. 1–16.
[30] Saelyne Yang, Jisu Yim, Juho Kim, and Hijung Valentina Shin. 2022. CatchLive: Real-time Summarization of Live Streams with Stream Content and Interaction Data. In *CHI Conference on Human Factors in Computing Systems*. 1–20.
[31] Minerva M Yeung and Boon-Lock Yeo. 1997. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Transactions on circuits and Systems for Video Technology* 7, 5 (1997), 771–785.
[32] Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert, and Huamin Qu. 2019. EmoCo: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 927–937.
[33] Jian Zhao, Chidansh Bhatt, Matthew Cooper, and David A Shamma. 2018. Flexible learning with semantic visual exploration and sequence-based recommendation of MOOC videos. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[34] Zhenge Zhao, Panpan Xu, Carlos Scheidegger, and Liu Ren. 2021. Human-in-the-loop extraction of interpretable concepts in deep learning models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 780–790.