Video Event Specification using Programmatic Composition

Daniel Y. Fu, Will Crichton, James Hong, Xinwei Yao, Haotian Zhang, Anh Truong, Avanika Narayan, Maneesh Agrawala, Christopher Ré, Kayvon Fatahalian
Stanford University

Abstract

Many real-world video analysis applications require the ability to identify domain-specific events, such as interviews and commercials in TV news broadcasts, or action sequences in film. Unfortunately, pre-trained models to detect all the events of interest in video may not exist, and training new models from scratch can be costly and labor-intensive. In this paper, we explore the utility of specifying new events in video in a more traditional manner: by writing queries that compose the outputs of existing, pre-trained models. To write these queries, we have developed Rekall, a library that exposes a data model and programming model for compositional video event specification. Rekall represents video annotations from different sources (object detectors, transcripts, etc.) as spatiotemporal labels associated with continuous volumes of spacetime in a video, and provides operators for composing labels into queries that model new video events. We demonstrate the use of Rekall in analyzing video from cable TV news broadcasts and films. In these efforts, analysts were able to quickly (in a few hours to a day) author queries to detect new events. These queries were often on par with or more accurate than learned approaches (6.5 F1 points more accurate on average).

1 Introduction

Modern machine learning techniques can robustly annotate large video collections with basic information about their audiovisual contents (e.g., face bounding boxes, people/object locations, time-aligned transcripts). However, many real-world video applications require exploring a more diverse set of events in video. For example, our recent efforts to analyze cable TV news broadcasts required models to detect interview segments and commercials. A film production team may wish to quickly find common segments such as action sequences to put into a movie trailer. An autonomous vehicle development team may wish to mine video collections for events like traffic light changes or obstructed left turns to debug the car's prediction and control systems.

Unfortunately, pre-trained models to detect these domain-specific events often do not exist, given the large number and diversity of potential events of interest. Training models for new events can be difficult and expensive, due to the large cost of labeling a training set from scratch, and the computation time and human skill required to then train an accurate model. Our experiences suggest that it is important to enable more agile video analysis workflows where an analyst, faced with a video dataset and an idea for a new event of interest (but only a small number of labeled examples, if any), can quickly author an initial model for the event, immediately inspect the model's results, and then iteratively refine the model to meet the accuracy needs of the end task.

To enable these agile, human-in-the-loop video analysis workflows, we take a more traditional approach: *specifying novel events in video as queries that programmatically compose the outputs of existing, pre-trained models.* Since heuristic composition does not require additional model training and is cheap to evaluate, analysts can immediately inspect query results as they iteratively refine queries to overcome challenges such as modeling complex event structure and dealing with imperfect source video annotations (missed object detections, misaligned transcripts, etc.).

To facilitate a query-based approach for detecting novel events of interest in video, we developed Rekall, a library that exposes a data model and programming model for *compositional video event specification*, illustrated in Figure 1. Rekall adapts ideas from multimedia databases [2, 4, 7, 9–12] and complex event processing systems for temporal data streams [3, 5, 8] to the modern video analysis landscape.

Rekall adopts a unified representation of multi-modal video annotations, the *spatiotemporal label*, to compose video annotations from multiple data sources that may be sampled at different temporal resolutions (e.g., a car detection on a single frame from a deep neural network, the duration of a word over half a second in a transcript), and uses compositions of these labels to express complex event structure and define increasingly higher-level video events.

Rekall's data model and programming model have allowed for rapid, accurate detection of new events in real-world analyses of cable TV news broadcasts and films, and data mining vehicle logs for training data curation and debugging of vehicle perception and control systems at a major venture-backed commercial autonomous vehicle company. In a formal evaluation against learned approaches, Rekall queries achieved accuracy scores that were on average 6.5 F1 points more accurate than the learned baselines.

2 System Overview

To better understand compositional video event specification with Rekall, consider a situation where an analyst, seeking to understand sources of bias in TV political coverage, wishes

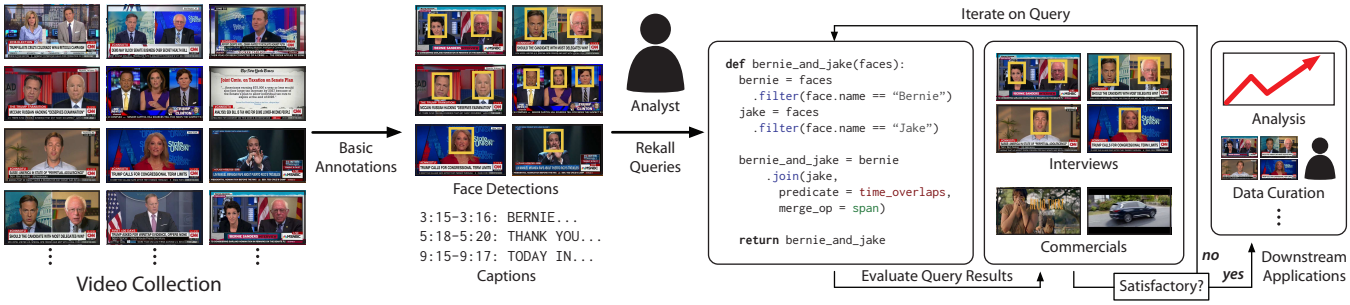

Figure 1. Overview of a compositional video event specification workflow. An analyst pre-processes a video collection to extract basic annotations about its contents (e.g., face detections from an off-the-shelf deep neural network and audio-aligned transcripts). The analyst then writes and iteratively refines Rekall queries that compose these annotations to specify new events of interest, until query outputs are satisfactory for use by downstream analysis applications.

to tabulate the total time spent interviewing a political candidate in a large collection of TV news video. Performing this analysis requires identifying video segments that contain interviews of the candidate. Since extracting TV news interviews is a unique task, we assume a pre-trained computer vision model is not available to the analyst. However, it is reasonable to expect that an analyst does have access to widely available tools for detecting and identifying faces in the video, and to the video's time-aligned text transcripts.

Common knowledge of TV news broadcasts suggests that interview segments tend to feature shots containing faces of the candidate and the show's host framed together, interleaved with headshots of just the candidate. Therefore, a first try at an interview detector query with Rekall might attempt to find segments featuring this temporal pattern of face detections. Refinements to this initial query might require parts of the sequence to align with common phrases like "welcome" and "thank for you being here." As illustrated in Figure 1, arriving at an accurate query for a dataset often requires multiple iterations of the analyst reviewing query results and modifying the query; Rekall's programmatic interface allows rapid iteration and interpretable fine-tuning.

3 Preliminary Results

Task	Learned Baseline	Rekall
Interview	87.3 ± 2.4	**95.5**
Commercial	90.9 ± 1.0	**94.9**
Conversation	**79.4 ± 2.3**	71.8
Shot Detect	83.2 ± 1.0	**84.1**
Shot Scale	70.1 ± 0.8	**96.2**

Table 1. In three representative tasks drawn from video analysis of cable TV news broadcasts and film, Rekall queries are more accurate than learned baselines.

We have written Rekall queries for video analysis tasks from media bias studies of cable TV news broadcasts and cinematography studies of Hollywood films. Table 1 shows F1 scores for some of these queries compared to learned baselines for interview detection and commercial detection in TV news, and conversation detection, shot transition detection, and shot scale classification in film. For the learned baselines, we trained three learning approaches and report the best score for each task: ResNet-50 image classification (pre-trained on ImageNet) [1, 13], with and without temporal smoothing over the outputs, and Conv3D ResNet-34-backed action recognition pre-trained on Kinetics [6].

These Rekall queries composed the outputs of face detectors, pose estimations, and color histograms, and were developed by analysts with little prior Rekall experience in a short amount of time – ranging from an afternoon to two days – but they often achieved higher accuracies than the learned baselines (6.5 F1 points more accurate on average).

Rekall queries have also been used to drive human-in-the-loop workflows. At a major venture-backed commercial autonomous vehicle company, Rekall queries are used to focus human labeler effort by surfacing rare but important scenarios, such as situations where traffic lights are behaving in unexpected ways. We have also used Rekall queries to drive video content retrieval tasks, such as supercuts of film idioms or movie trailer creation; see http://www.danfu.org/projects/rekall-aisystems2019/ for examples and a technical report with more details on the experiments and use cases.

4 Discussion

By adapting ideas from complex event analysis over temporal streams to the video domain, Rekall gives analysts a new tool for quickly specifying video events of interest. We believe productive systems for compositional video event specification stand to play an important role in the development of traditional machine learning pipelines by, for example, helping engineers write programs that surface a more diverse set of training examples for better generalization, or enabling search for anomalous model outputs (feeding active learning loops). We hope that our experiences encourage the community to explore techniques that allow video analysis efforts to more effectively utilize human domain expertise and more seamlessly provide solutions that move along a spectrum between traditional query programs and learned models.

Acknowledgments

We thank Ines Chami, Tri Dao, Jared Dunnmon, Sarah Hooper, Megan Leszczynski, Bill Mark, Avner May, and Paroma Varma for their valuable feedback. We gratefully acknowledge the support of DARPA under Nos. FA87501720095 (D3M), FA86501827865 (SDH), and FA86501827882 (ASED), NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), 1937301 (RTML), III-1908727 (A Query System for Rapid Audiovisual Analysis of Large-Scale Video Collections), and III-1714647 (Extracting Data and Structure from Charts and Graphs for Analysis Reuse and Indexing), ONR under No. N000141712266 (Unifying Weak Supervision), the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, and American Family Insurance, Google Cloud, Swiss Re, Brown Institute for Media Innovation, Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, SAP, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

[1] Pytorch: Transfer leraning tutorial. https://pytorch.org/ tutorials/beginner/transfer_learning_tutorial.html, 2019.

[2] Adali, S., Candan, K. S., Chen, S.-S., Erol, K., and Subrahmanian, V. The advanced video information system: data structures and query processing. *Multimedia Systems 4*, 4 (Aug 1996), 172–186.

[3] Chandramouli, B., Goldstein, J., Barnett, M., DeLine, R., Fisher, D., Platt, J., Terwilliger, J., Wernsing, J., and DeLine, R. Trill: A high-performance incremental query processor for diverse analytics. VLDB - Very Large Data Bases.

[4] Dönderler, M. E., Ulusoy, O., and Güdükbay, U. Rule-based spatiotemporal query processing for video databases. *The VLDB Journal 13*, 1 (Jan. 2004), 86–103.

[5] Friedman, E., and Tzoumas, K. *Introduction to Apache Flink: Stream Processing for Real Time and Beyond*, 1st ed. O'Reilly Media, Inc., 2016.

[6] Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 6546–6555.

[7] Hibino, S., and Rundensteiner, E. A. A visual query language for identifying temporal trends in video data. In *Proceedings. International Workshop on Multi-Media Database Management Systems* (Aug 1995), pp. 74–81.

[8] Jayasinghe, M., Jayawardena, A., Rupasinghe, B., Dayarathna, M., Perera, S., Suhothayan, S., and Perera, I. Continuous analytics on graph data streams using wso2 complex event processor. In *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems* (New York, NY, USA, 2016), DEBS '16, ACM, pp. 301–308.

[9] Köprülü, M., Cicekli, N. K., and Yazici, A. Spatio-temporal querying in video databases. In *Proceedings of the 5th International Conference on Flexible Query Answering Systems* (London, UK, UK, 2002), FQAS '02, Springer-Verlag, pp. 251–262.

[10] Kuo, T. C. T., and Chen, A. L. P. A content-based query language for video databases. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems* (June 1996), pp. 209–214.

[11] Li, J. Z., Özsu, M. T., and Szafron, D. Modeling of moving objects in a video database. *Proceedings of IEEE International Conference on Multimedia Computing and Systems* (1997), 336–343.

[12] Oomoto, E., and Tanaka, K. Ovid: design and implementation of a video-object database system. *IEEE Transactions on Knowledge and Data Engineering 5*, 4 (Aug 1993), 629–643.

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*, 3 (2015), 211–252.