Rekall: Specifying Video Events using Compositions of Spatiotemporal Labels

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 in video, 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 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, domain experts were able to quickly (in a few hours to a day) author queries that enabled the accurate detection of new events.

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.

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. We seek 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 propose taking 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 explore the utility of 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.* Rekall adapts ideas from multimedia databases [1, 3, 5, 7–10] to the modern video analysis landscape, where using the outputs of modern machine learning techniques allows for more powerful and expressive queries, and adapts ideas from complex event processing systems for temporal data streams [2, 4, 6] to the spatiotemporal domain of video.

The primary technical challenge in building Rekall was defining the appropriate abstractions and compositional primitives for users to write queries over video. In order 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), Rekall's data model adopts a unified representation of multi-modal video annotations, the *spatiotemporal label*, that is associated with a continuous volume of spacetime in a video. Rekall's programming model uses hierarchical composition of these labels to express complex event structure and define increasingly higher-level video events.

2 An Analysis Example

To better understand the thought process underlying our video analysis tasks, consider a situation where an analyst, seeking to understand sources of bias in TV political coverage, wishes 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

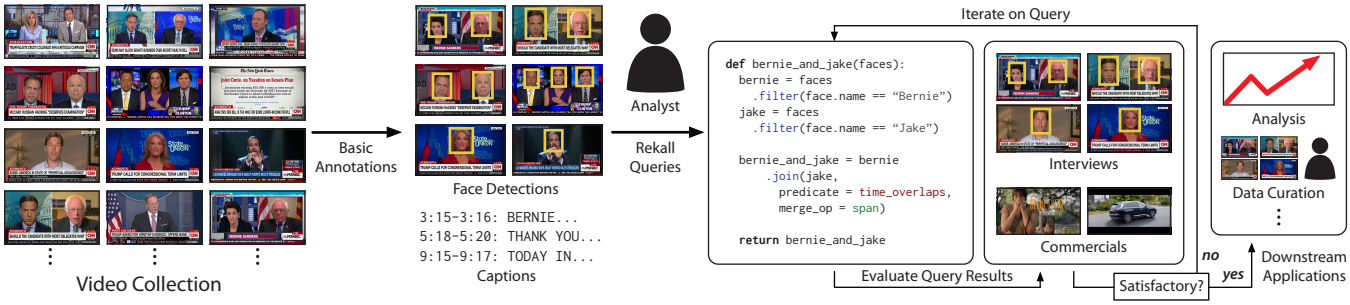

Figure 1. Overview of a compositional video event specification workflow. An analyst pre-processes a video collection to extract basic annotations about its contents (e.g., face detections from an off-the-shelf deep neural network and audio-aligned transcripts). The analyst then writes and iteratively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

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 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 might attempt to find segments featuring this temporal pattern of face detections. Refinements to this initial query might permit the desired pattern to contain brief periods where neither individual is on screen (e.g., display of B-roll footage for the candidate to comment on), or require parts of the sequence to align with utterences of the candidate's name in the transcript or 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 adding additional heuristics as necessary until a desired level of accuracy is achieved.

3 Preliminary Results

Task	Learned Baseline	Rekall
Interview	88.6 ± 5.3	**95.5**
Commercial	90.0 ± 0.9	**94.9**
Conversation	66.1 ± 3.5	**71.8**

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 report average F1 scores and standard deviations over five random weight initializations.

We have implemented 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 Rekall queries compared to a learned baseline for three tasks – interview detection and commercial detection in TV news, and conversation detection in film. For the learned baselines, we fine-tuned a ResNet-50 image classifier (pre-trained on ImageNet) for each task, and performed "temporal smoothing" on the results by taking the mode of model predictions over a window of seven frames.

These Rekall queries were developed by domain experts with little prior Rekall experience in a short amount of time – ranging from an afternoon to two days (time to learn how to use Rekall) – but they achieved accuracies more accurate than the learned baselines (6.5 F1 points more accurate on average). Rekall queries have also been used to drive human-in-the-loop video content retrieval tasks, such as supercuts of film idioms or movie trailer creation; see http://www.danfu.org/projects/rekall-aisystems2019/ for examples.

4 Discussion

Rekall is intended to give analysts a new tool for quickly specifying video events of interest using heuristic composition. Constructing queries through procedural composition lets users go from an idea to a set of video event detection results rapidly, does not incur the costs of large-scale human annotation and model training, and allows a user to express heuristic domain knowledge, modularly build on existing labels, and more intuitively debug failure modes.

We believe productive systems for compositional video event specification stand to play an important role in the development of traditional machine learning pipelines by helping engineers write programs that surface a more diverse set of training examples for better generalization, enabling search for anomalous model outputs (feeding active learning loops), or as a source of weak supervision to bootstrap model training. 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.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.