CSC2515 Project Proposal: Extended Video Stream Queries with Semantic Analysis

Daren Chao

October 2019

1 Problem Statement

Recent advances in computer vision - in the form of deep neural networks - have made it possible to query increasing volumes of video data with high accuracy. However, neural network inference is computationally expensive at scale. Although some have proposed systems for accelerating neural network queries over the video, there are still many limitations for them to be applied to real-world scenarios.

To understand the visual world, a machine must not only recognize individual object instances but also how they interact. Most of the state-of-the-art video query processing models focus on the queries about the count of categorized objects and their location relationships. Semantic information, such as human-object interaction information, is often ignored.

I address the task of queries for video stream database system, about not only the detection of objects and categories and their locations but also semantic information such as human actions/poses. For queries involving actions/poses, the algorithm can, firstly, quickly detect the humans and objects that appear in each frame, as well as their location information, and then conduct the possibly human actions, i.e. the most likely combination of triplet, (human, action, object[1]), in these frames.

2 Prior work and Background

2.1 Data Management on Video Streams

Recently there has been increased interest in the application of Deep Learning techniques in data management. Many works tried to apply image classification and object detection algorithms to fast query processing on video streams.

NoScope [4] uses a modified form of distillation [2] to train an extremely lightweight specialized model at the cost of generalization to other videos. Given a target video, object to detect, and reference neural network, NoScope automatically searches for and trains a sequence of models that preserves the accuracy of the reference network but is specialized to the target video and is therefore far less computationally expensive.

Based on NoScope, [3] has proposed a system, BlazeIt, that optimizes queries over video for spatiotemporal information of objects. BlazeIt accepts queries via a declarative language, FrameQL, and new query optimization techniques including an aggregation algorithm, a scrubbing algorithm, and a selection algorithm to leverage imprecise specialized NNs, find rare events and apply content-based selection.

[5] presented a series of filters to estimate the number of objects in a frame, the number of objects of specific classes in a frame as well as to assess an estimate of the spatial position of an object in a frame.

[1] Some actions may have no object.

Although these algorithms have achieved good accuracy for counting and location estimation purposes and dramatic speedups by several orders of magnitude in real video datasets, there are still many additional query types that need to be considered.

2.2 Detecting and Recognizing Human-object Interaction

Visual recognition of individual instances, e.g., detecting objects and estimating human actions has witnessed significant improvements thanks to deep learning visual representations. [1] proposed a novel model that is driven by a human-centric approach, which learns to predict an action-specific density over target object locations based on the appearance of a detected person using the appearance of a person as a cue for localizing the objects they are interacting with.

Some works used video clips for action/pose recognition. I will review them in the future, as I might use them in my approach. however, these NN-based algorithms cannot be applied to video stream queries directly due to their expensive at scale. As a result, one of my contribution to this project is to accelerate these algorithms and apply them to the video semantic-query scenario.

3 Approach

3.1 Proposed Techniques

For queries involving actions/poses of humans, the algorithm can, first, quickly detect the humans and objects that appear in each frame, as well as their location information, and then return the most likely combination of triplet, (human, action, object), in these frames.

There are many ways to implement the above algorithm of conducting possible actions in each frame. First, a neural network similar to human-centric branch [1] can be applied to obtain the score of assigning an action for the person and whether the object is the actual target of the action; then, the whole model obtains the actions that may exist in each frame, based on the above scores. Second, actions are made up of many consecutive frames; thus, detecting human-object interaction in the video stream can be based on extracted video clips. Based on this idea, we can propose a video clips extraction algorithm and an action detection algorithm on video clips, so that the return of queries can be multiple video clips, each of which belongs to a (human, action, object) triplet.

3.2 To what degree the project will repeat existing works

Xarchakos, I., and Koudas, N., et, al. have proposed the method of dealing with video stream queries involving actions based on the [1]. I will implement the code in their under-reviewing paper and help Xarchakos, I. to create a demo to show how a specific video query is processed, and also the final result (accuracy and time).

After that, I will study new kinds of video stream queries involving semantic information and their more efficient processing methods. I will refer to the definition of SQL-like language of video stream queries in prior works. When implementing object detection and the action detection part, we will learn from the models in [3–6] and [1] respectively. However, the network of object detection and action recognition in previous work is complicated and can not be directly applied to the current scenario. Therefore, the modification of the model so that it can better meet the needs of video stream queries will be the main contribution of this project.

4 Plan of Action

i. **Oct 21 - Oct 27** Reading papers of Xarchakos, I., and Koudas, N., et, al. and its codes, and preparing for the Demo. Requirements: To learn Flask, D3.js, etc.

ii. **Oct 28** Discussing the thoughts of the Demo with Professor Koudas.

iii. **Oct 28 - Nov 6** Writing and run the code of previous works. Build the Demo.

iv. **Nov 3 - Nov 10** Reviewing literature about detecting and recognizing semantic information in images. The main task at this stage is to find models that are suitable for use in video queries involving semantic information and think about how to improve them.

v. **Nov 11** Discussing the new model with professor Koudas.

vi. **Nov 11 - Nov 21** Writing the code of the improved model. Also, try to find related data sets.

vii. **Nov 21 - Nov 27** Experiments. The model should be evaluated and adjusted during this time.

viii. **Nov 27 - Dec 1** Preparing the reports and presentation.

References

[1] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[3] Daniel Kang, Peter Bailis, and Matei Zaharia. Blazeit: Fast exploratory video queries using neural networks. *arXiv preprint arXiv:1805.01046*, 2018.

[4] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. Noscope: optimizing neural network queries over video at scale. *Proceedings of the VLDB Endowment*, 10(11):1586–1597, 2017.

[5] Ioannis Xarchakos Nick Koudas, Raymond Li. Video monitoring queries. *Under Review*, 2020.

[6] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.