Driving Scene-Retrieval by Example from Large-Scale Data

Sascha Hornauer, Baladitya Yellapragada, Arian Ranjbar, Stella X. Yu
UC Berkeley / ICSI

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

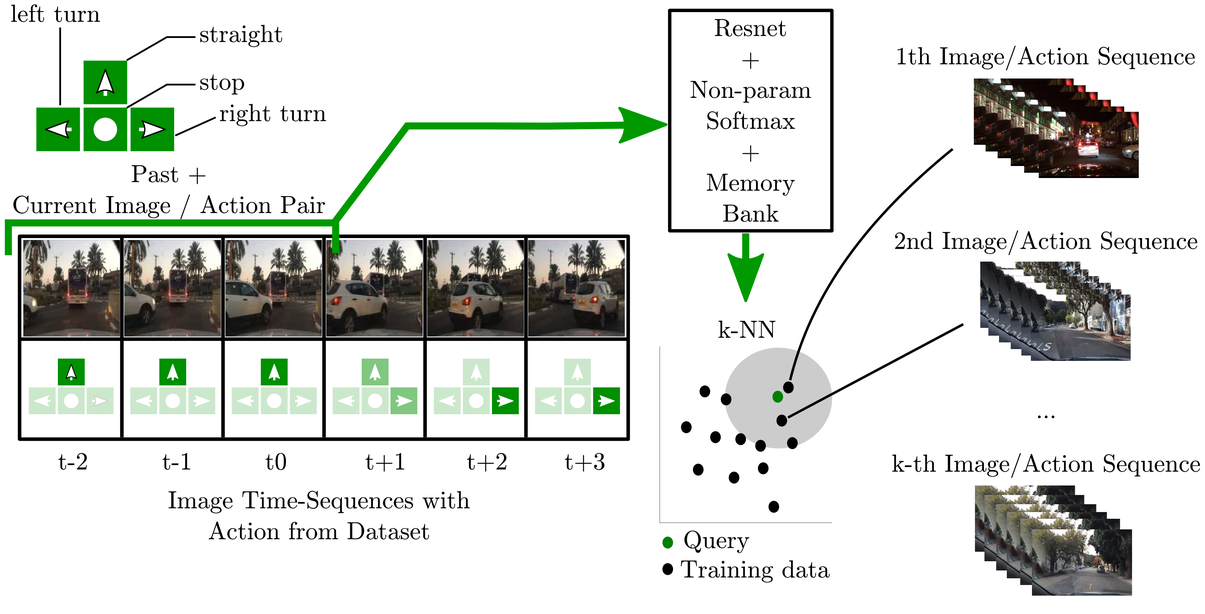
Overview

- Large video datasets for autonomous driving could be partitioned according to high-level concepts, such as *sunny*, *full of pedestrians* or *left turns in rural neighborhoods*.
- However, forming queries becomes very difficult or impossible without detailed labels reflecting those concepts.
- We train a network semi-supervised to embed image-action sequences in a feature space and then use it to retrieve those similar in appearance and action.
- Automatically logged vehicle actions are the only labels used during training.
- Querying video-sequences can retrieve similar sequences across object kind and number, weather, time of the day, and street layout.

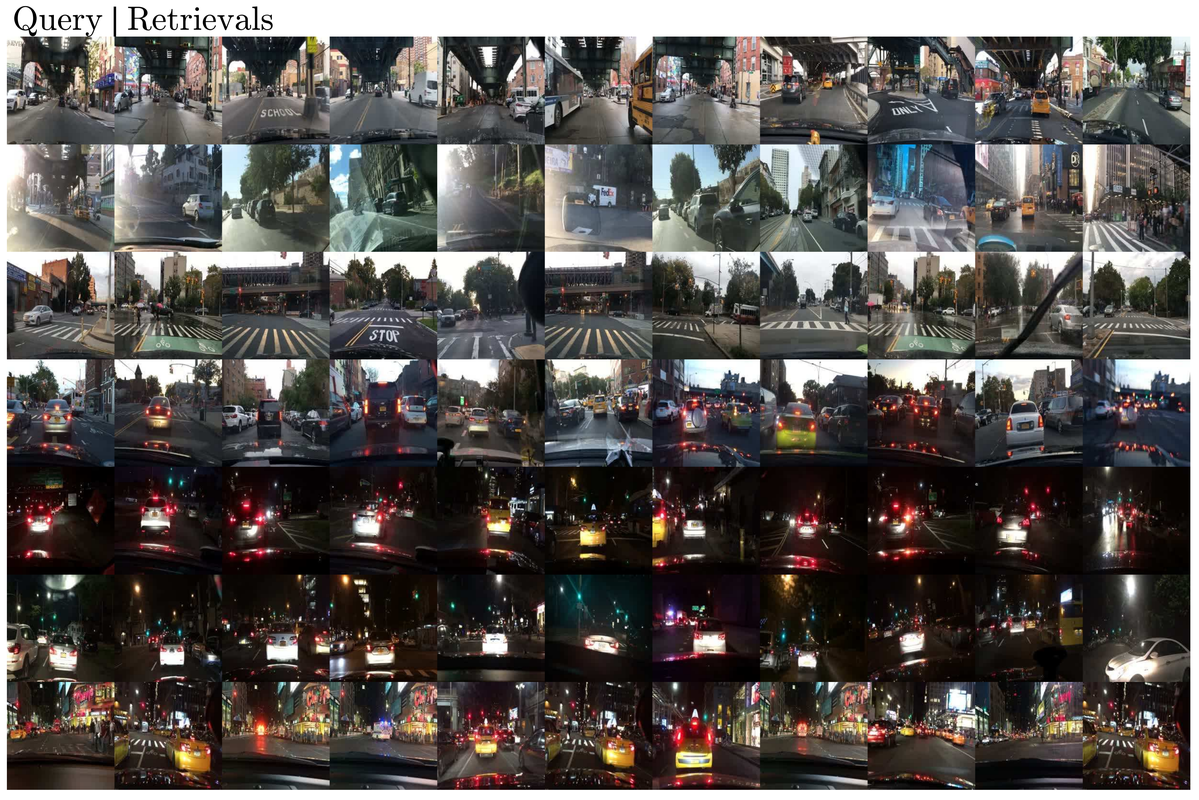
Method

- We compare individual images and short image-action sequences with 6 frames, sampled every 4 frames at the original rate of 30 frames per second, leading to 0.7 seconds per *driving scene* instance.
- A ResNet-18 mapping each driving scene to a feature representation is trained to maximally discriminate between different image-action sequences.
- We retrieve nearest neighbors in the learned feature space to get similar sequences to an input query.
- We compare learning on image-action pairs with learning on images alone, and find the network trained with image-action pairs retrieves more visually similar sequences.

Driving Scene Embedding and Retrieval



Our Results on Image-Action Scene Retrievals



Our Results on Image-Only Scene Retrievals