TEXAS
The University of Texas at Austin

SCAN ME

Exploring Explainability in Video Action Recognition

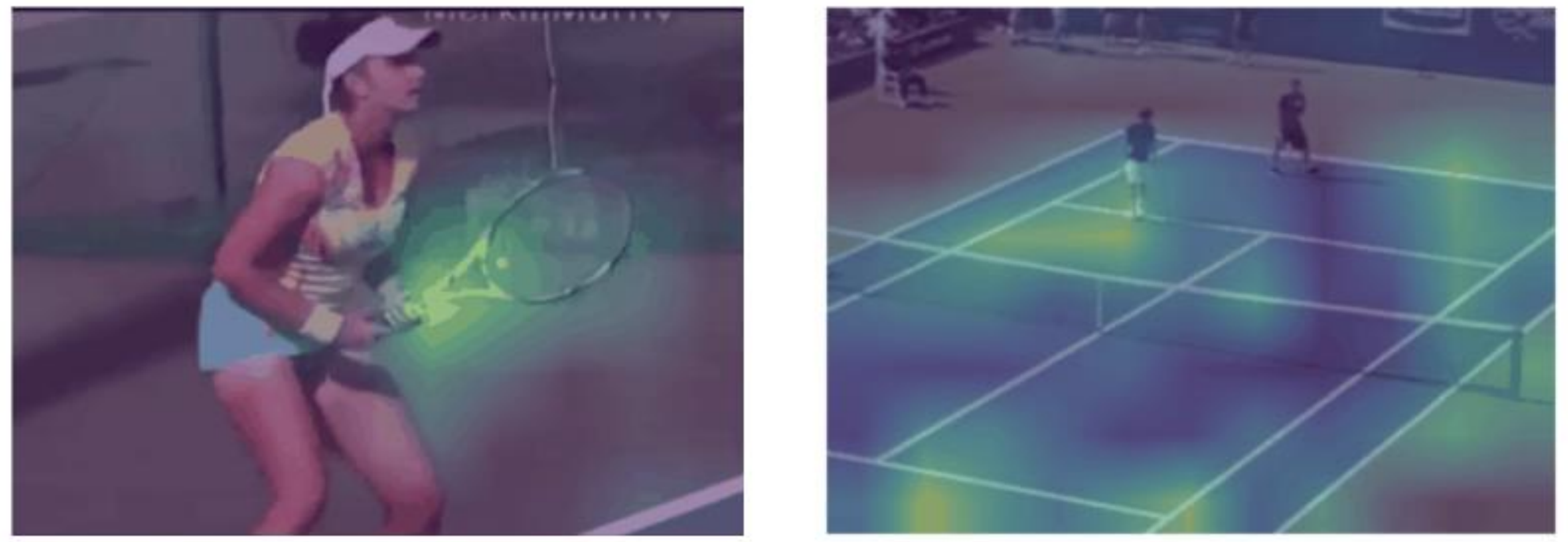
Avinab Saha*, Shashank Gupta*, Sravan Ankireddy*, Karl Chahine, Joydeep Ghosh

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Introduction

- While explainability methods have been designed for image tasks, there is **little work specifically for video**.
- We find that **attribution-based methods** such as GradCAM have **several issues** when applied to video tasks, such as flickering and ambiguous directionality.
- We propose an extension of TCAV called **Video-TCAV** and an automated method to **generate video concepts**.
- We generate two types of concepts – spatial-only and spatiotemporal, which reveal interesting properties about SOTA models such as Video Swin Transformer.

GradCAM Revisited



(a) Positive Illustration (b) Limitation

- Can be applied frame-by-frame or by frame batches.
- Framewise analysis **insensitive to temporal direction** – picking up a cup is the same as putting it down.
- Exhibits **temporal flickering** – attributions vary widely between frames and jump between elements.
- **Impractical** at scale – must be individually analyzed.

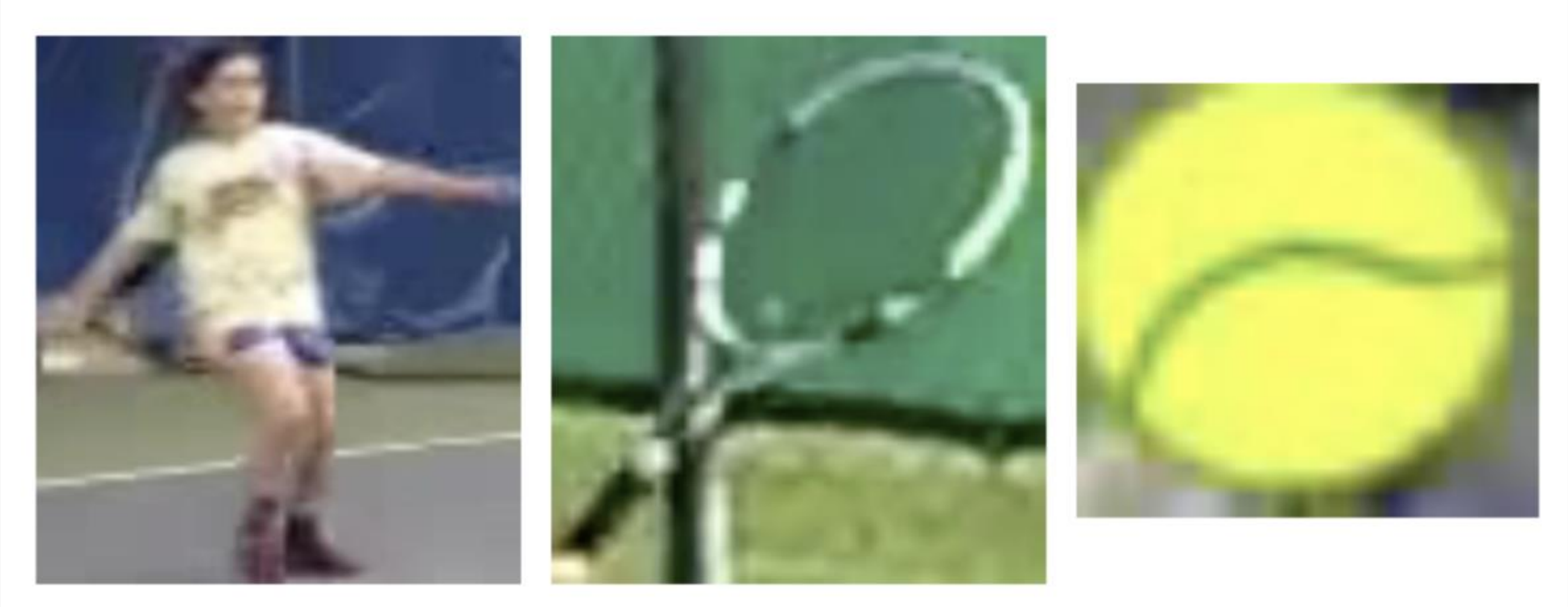
Video-TCAV

Concept: Collection of inputs sharing an implicit high-level property.
CAV: Hyperplane separating samples of concept and random inputs.
TCAV: Sensitivity of class samples to concept – directional derivative.
Pre-trained Model: Video Swin Transformer, action recognition SOTA.

Use YOLO-v7 to obtain **concepts based on objects and their behavior.**

Spatial Concept Generation

Crops of objects detected copied over several frames. Conceptually describe the **simple presence or absence – no temporal aspect.**



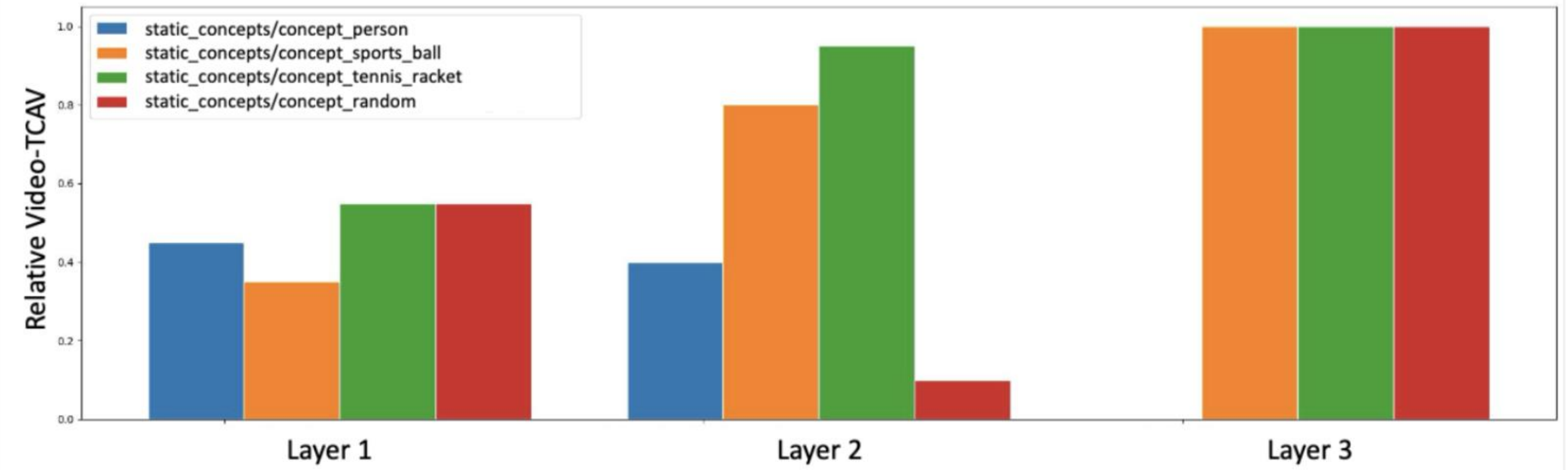
Spatiotemporal Concept Generation

Tracking of objects over several frames. Conceptually describe both the presence of an object and **how it moves in space in given context.**

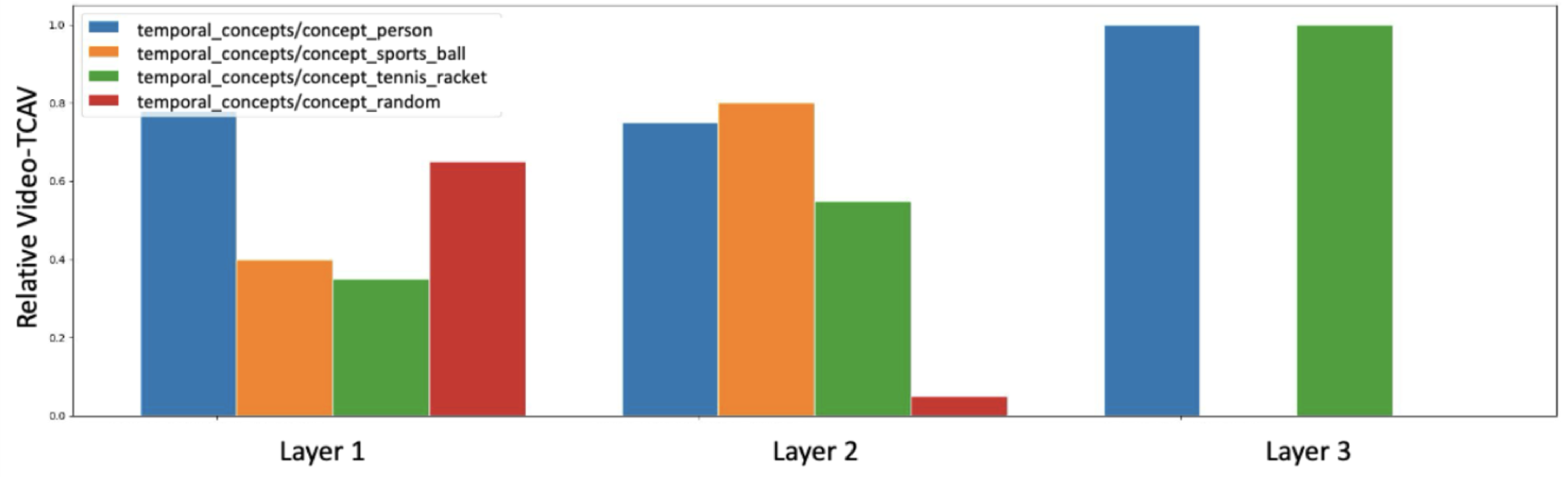
We study three levels of representation – early, middle and late stage in our pretrained model. We also manually verify concepts for sanity.

Results and Experiments

- **Static concepts are not very informative** when compared to 'random concept'. Effect persists till the last layer.



- **Dynamic concepts much more informative.** Motion sensitivity increases with depth, as in brain (V1 vs V5).



- Dynamic version of concept dominates static version with increase in depth. **Last layer exclusively prefers dynamic.**