Following Gaze in Video

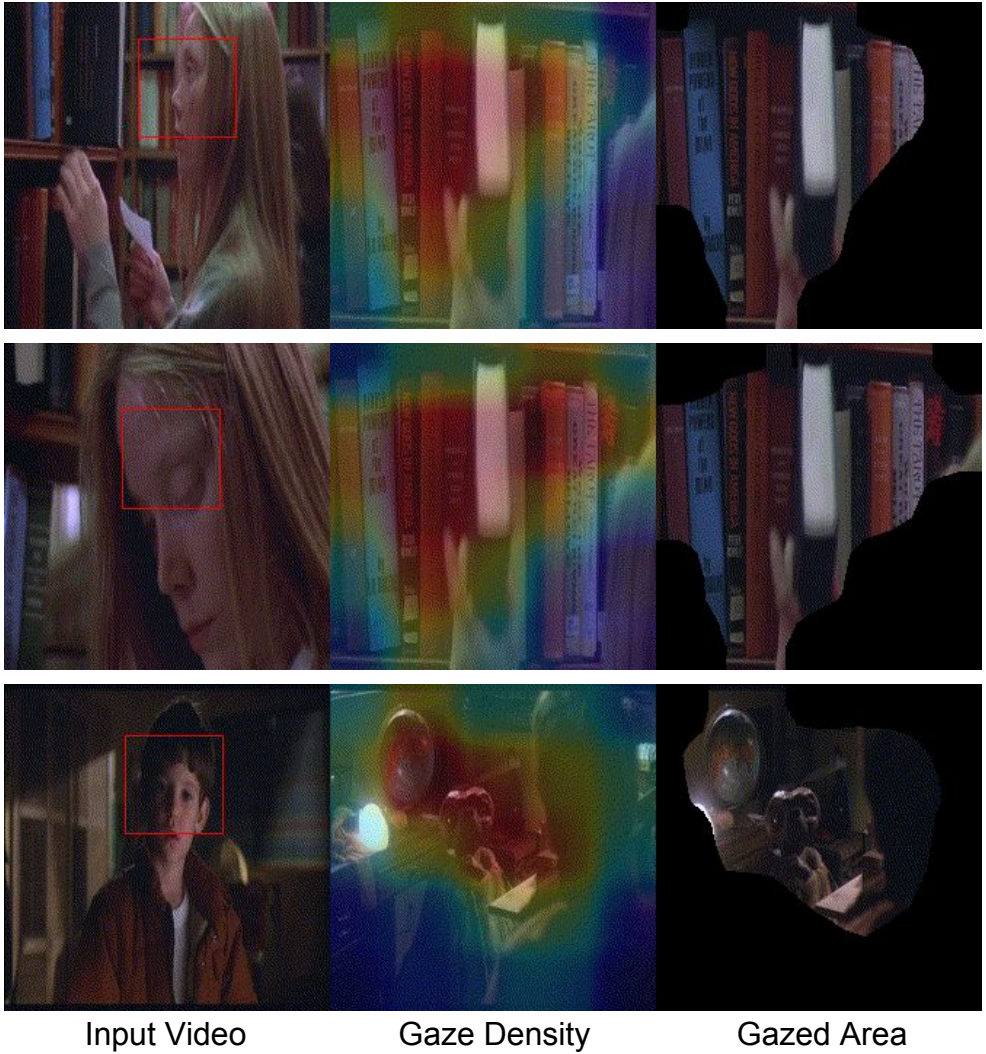
A. Recasens et al.

Presented by: Keivaun Waugh and Kapil Krishnakumar

Background

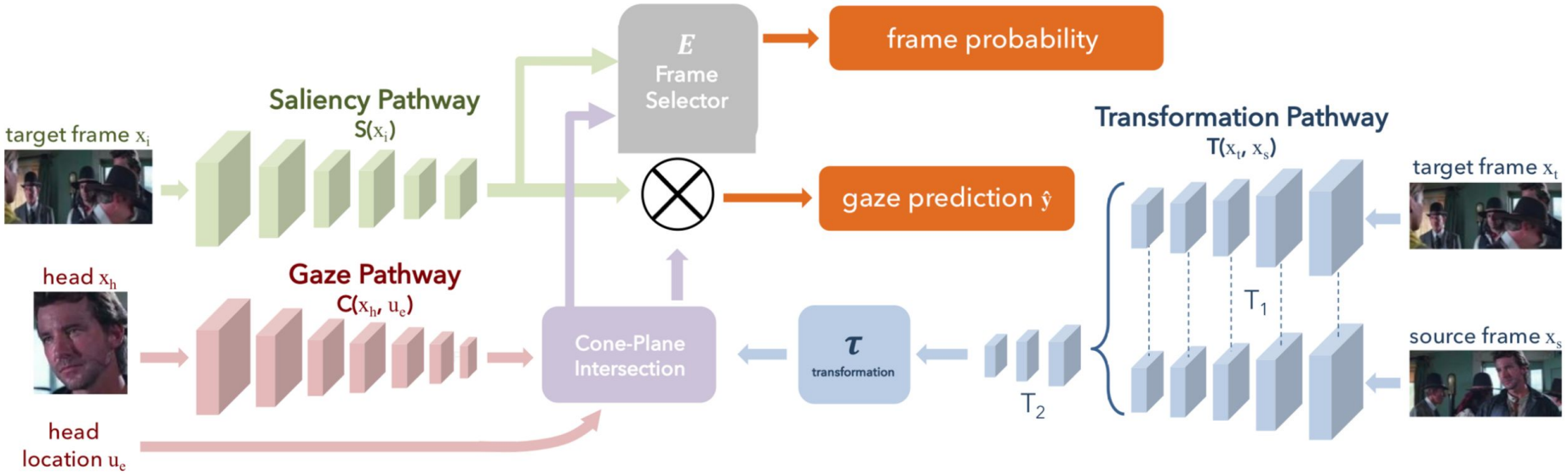
- Given face in one frame, how can we figure out where that person is looking?
- Target object might not be in the same frame

Sample Results



Input Video | Gaze Density | Gazed Area

Architecture



VideoGaze Dataset

- 160k annotations of video frames from MoviesQA dataset
- Annotations:
  - Source Frame
  - Head Location
  - Body
  - Target Frame (5 per source frame)
    - Gaze Location
    - Time difference between Source and Target

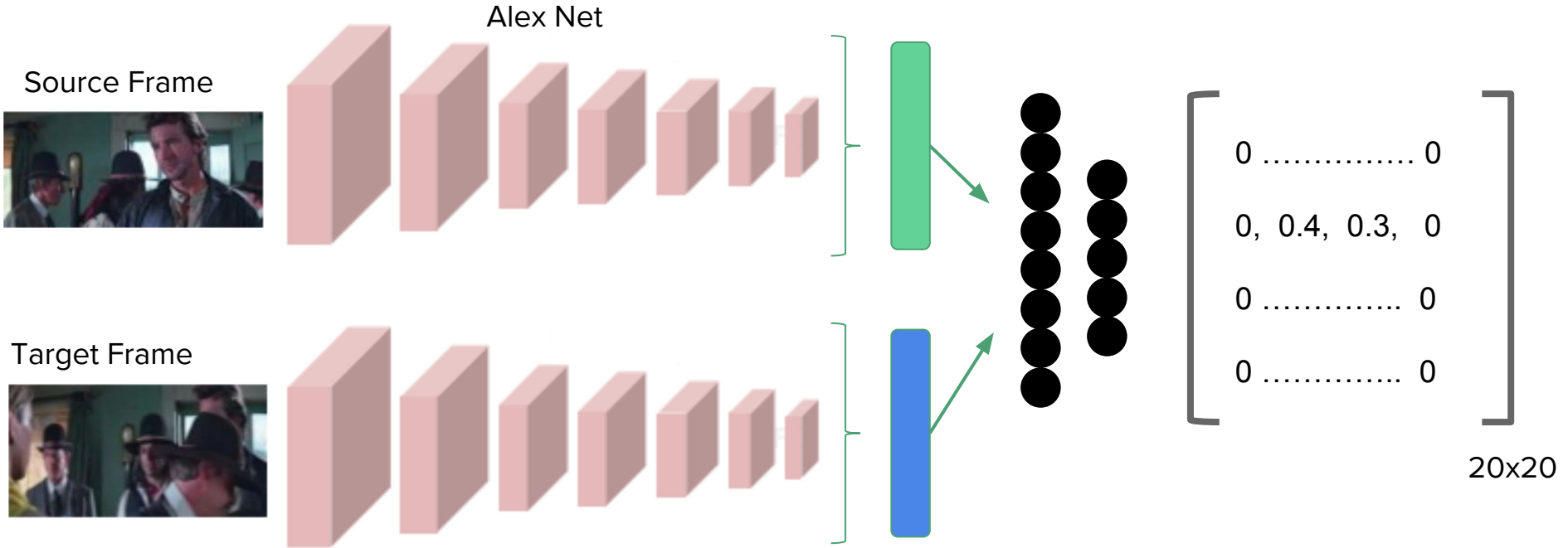
Experiments

- Naive network architecture
  - Don’t segment network into different into different pathways
  - Concatenate all inputs and predict directly
- Replace transformation pathway with SIFT+RANSAC affine fit finding
- Various neighboring frame prediction windows
- Examine failure cases
  - “Look cone” doesn’t take into account the eye position
  - Other failures

Naive Model

Naive Architecture

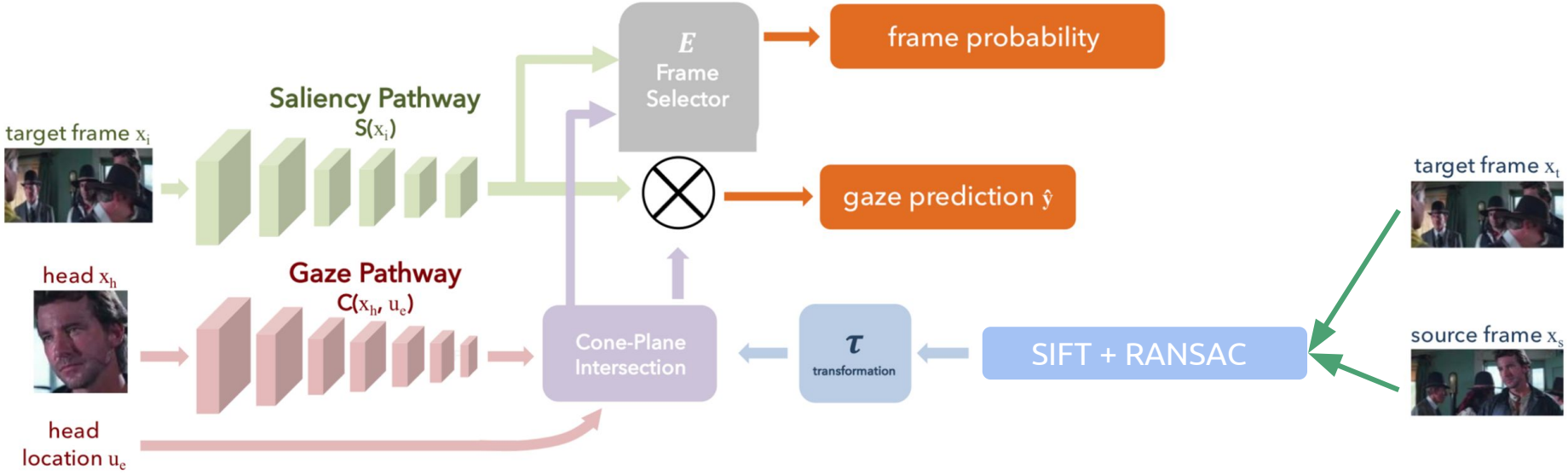
- Use fusion of target frame and source frame to predict gaze location



Alternate Transformation Pathway

Architecture

- Replace deep CNN pathway with traditional SIFT+RANSAC affine warp



Quantitative Results

Results

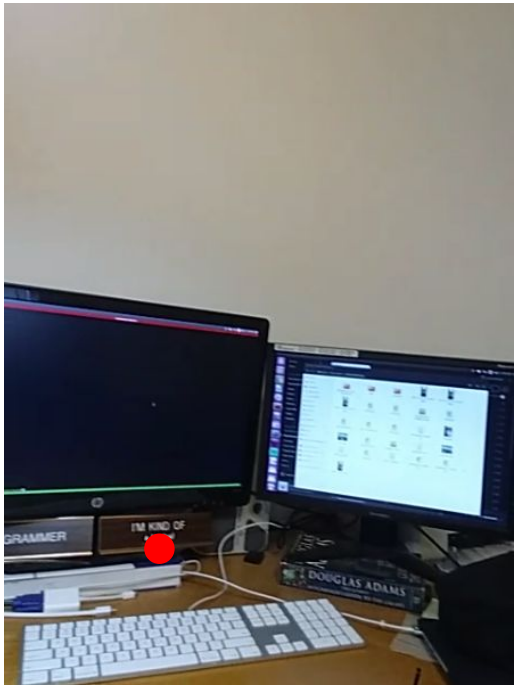
AUC (higher better)	KL Divergence (lower better)	L2 Dist (lower better)	Description
73.7	8.048	0.225	Normal model with transformation pathway
60.2	6.604	0.294	Normal model with sparse affine
60.2	6.6604	0.294	Normal model with dense affine
60.9	6.641	0.242	Naive model
56.9	28.39	0.437	Random

Qualitative Results

Results

- Input video is 150 frames long

Cropped Head

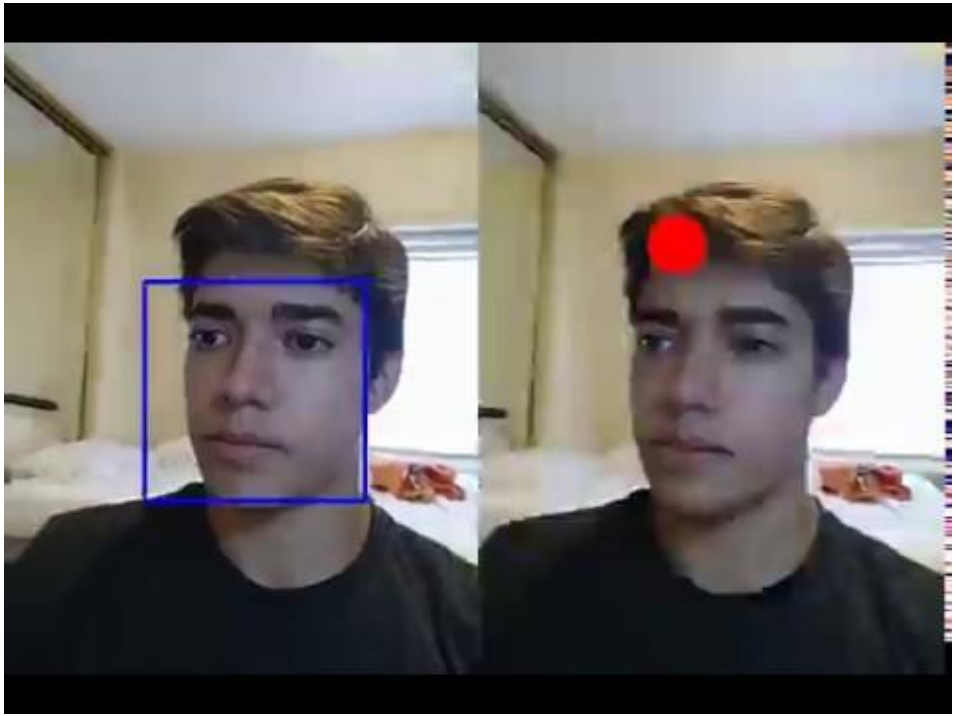


What I'm looking at

Full Video

Results - Search 150 Neighboring Frames

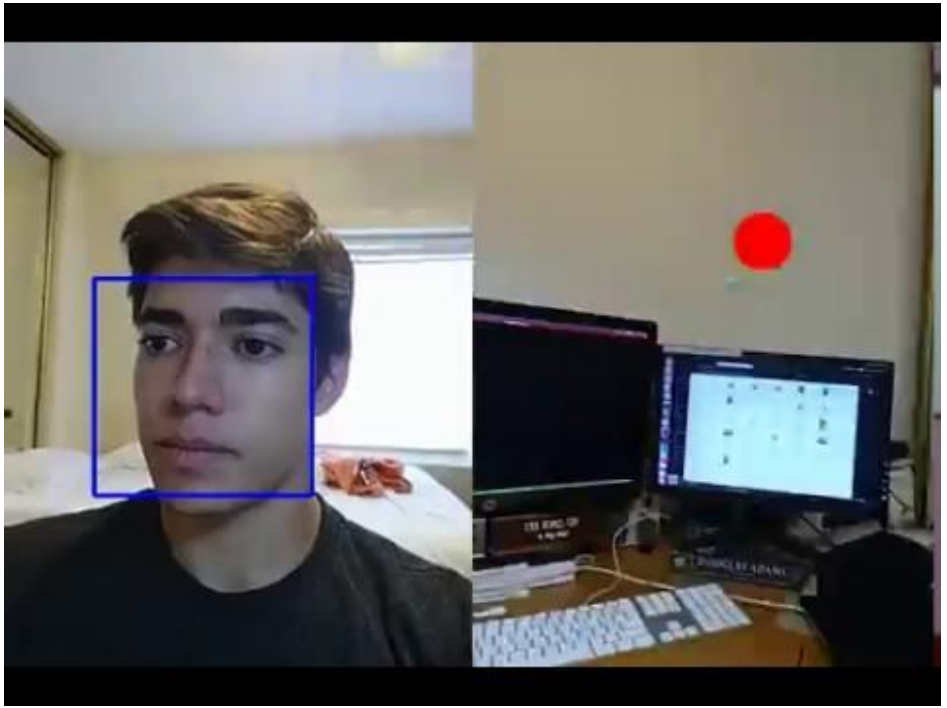
Original Transformation Pathway



Naive Model

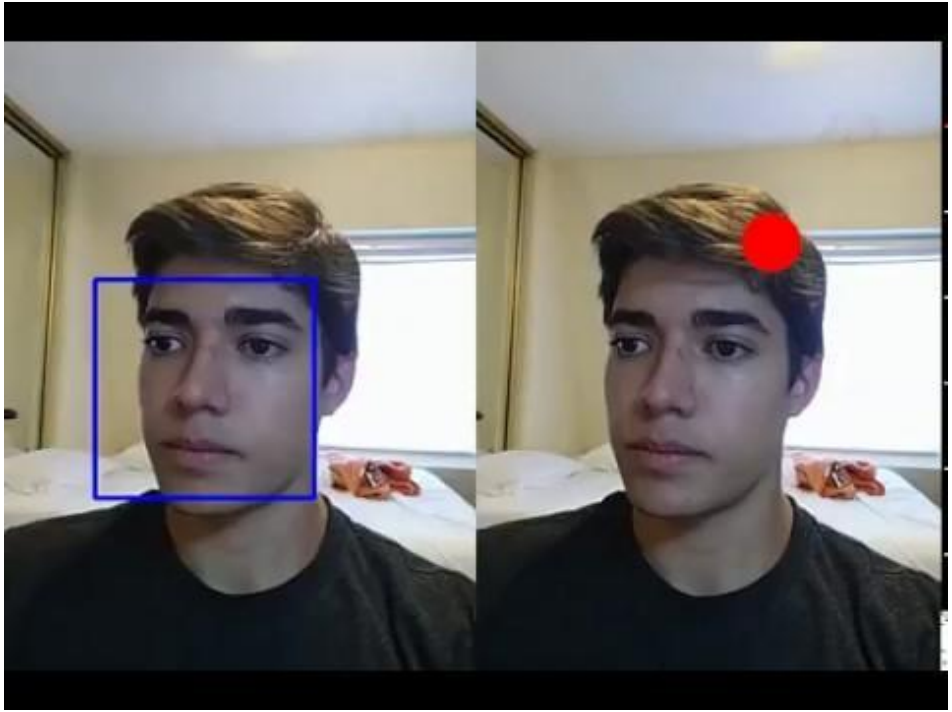
Results - Search 150 Neighboring Frames

Sparse SIFT Affine Warp

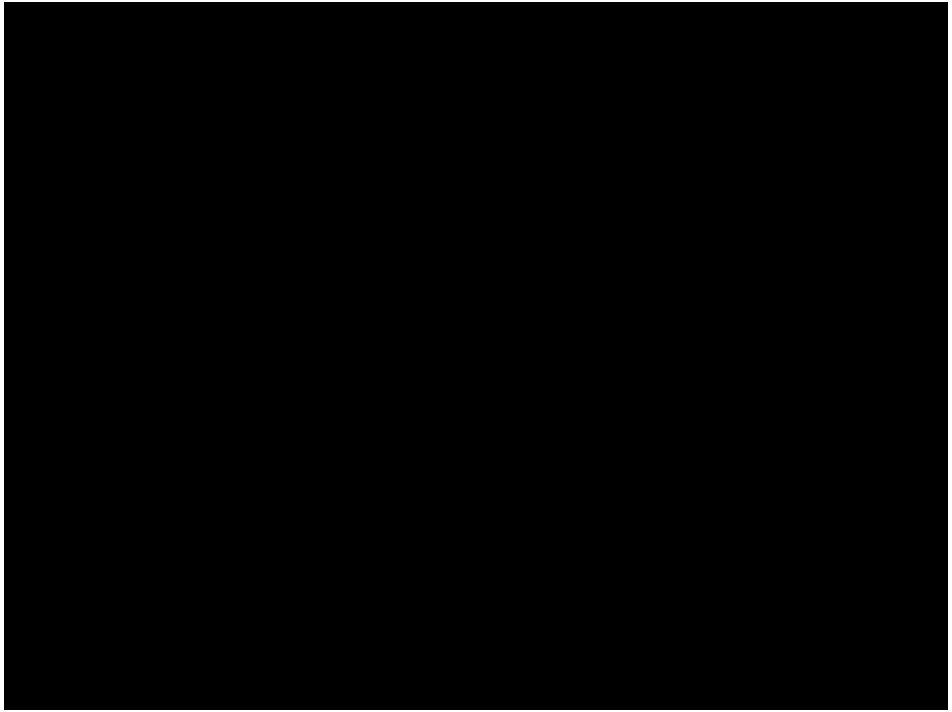


Dense SIFT Affine Warp

Results - Search 25 Neighboring Frames

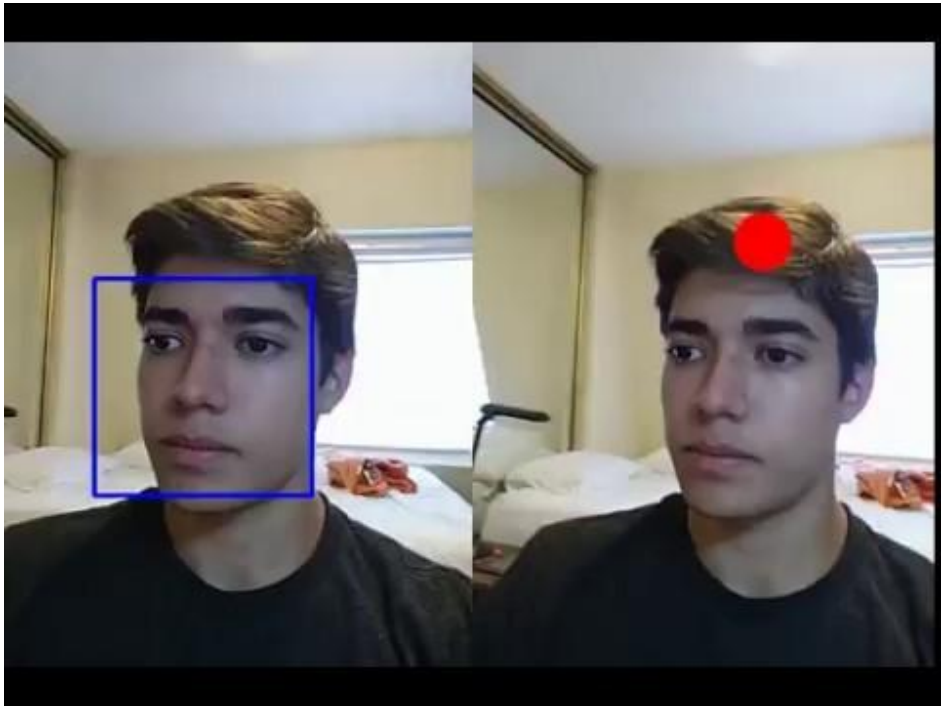


Original Transformation Pathway



Naive Model

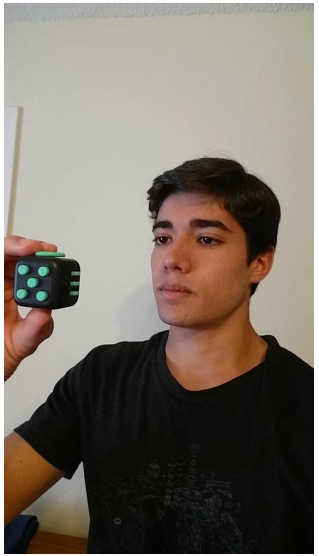
Results - Search 25 Neighboring Frames



Sparse SIFT Affine Warp

Dense SIFT Affine Warp

Target in Same Frame



Original Video

Original Transformation Pathway

Naive Model

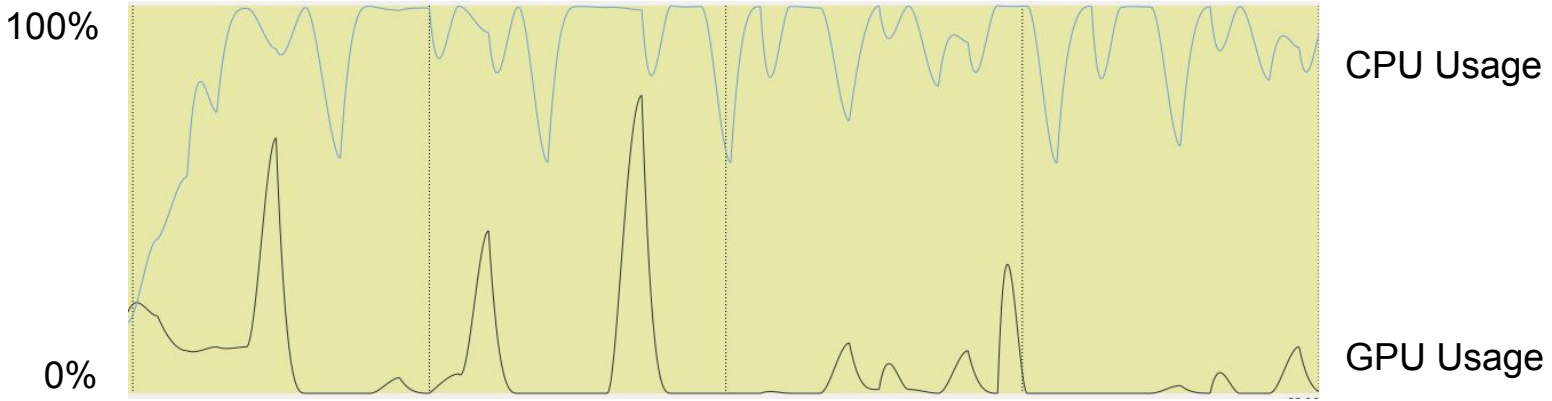
Target in Same Frame

Sparse SIFT Affine Warp

Dense SIFT Affine Warp

Runtimes

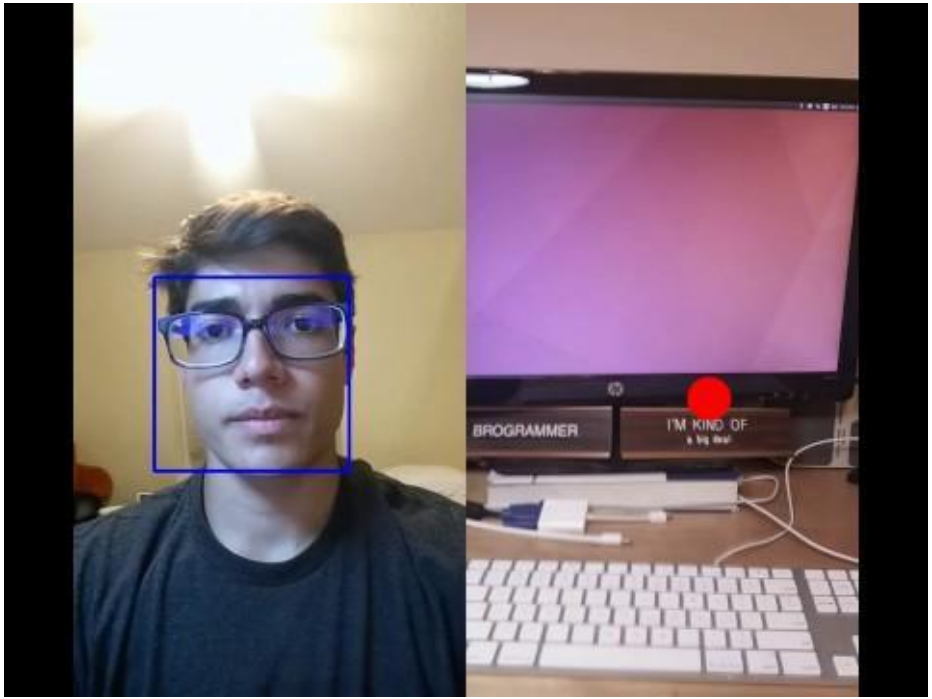
- GTX 1070 and Haswell Core i5
- Generating results is CPU bound
- 5 second video with 150 frame search width
  - Deep transformation pathway: 6.5 minutes
  - Sparse affine: 10.5 minutes
  - Dense affine: 32 minutes



Usage when running model with transformation pathway

Failure Cases

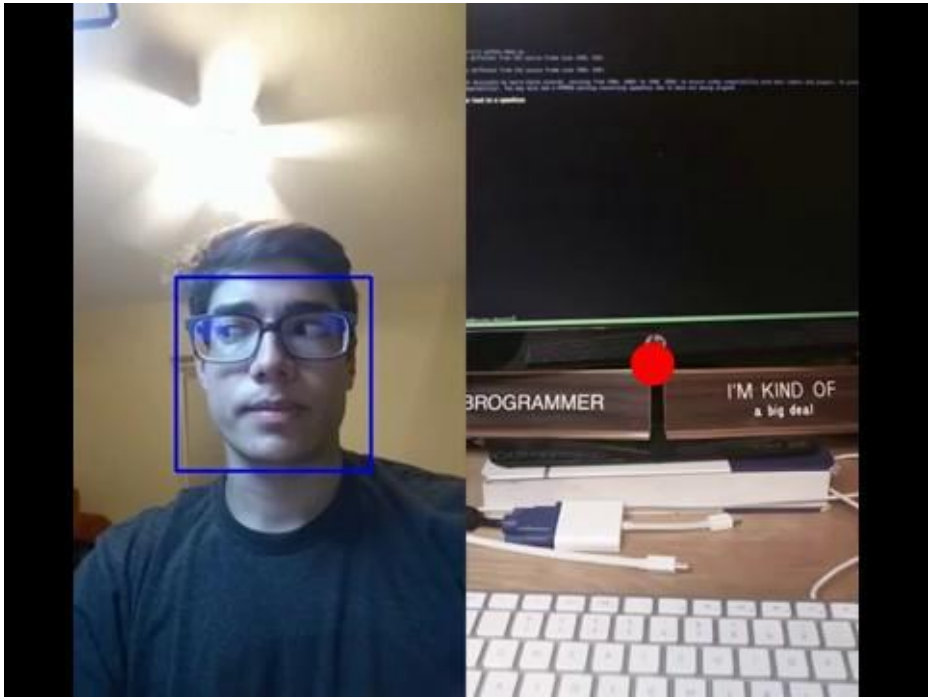
Input Video



Original Transformation Pathway

Failure Cases

Input Video



Original Transformation Pathway

Conclusions

- Separating input modalities for Saliency and Head Pose provides significant information to the model.
    - Illustrates importance of hand-crafted architecture even though features are automatically discovered
- Head Direction != Eye Direction
- Frame Predictor window selection determines whether match can be found or not.