Eye Spy

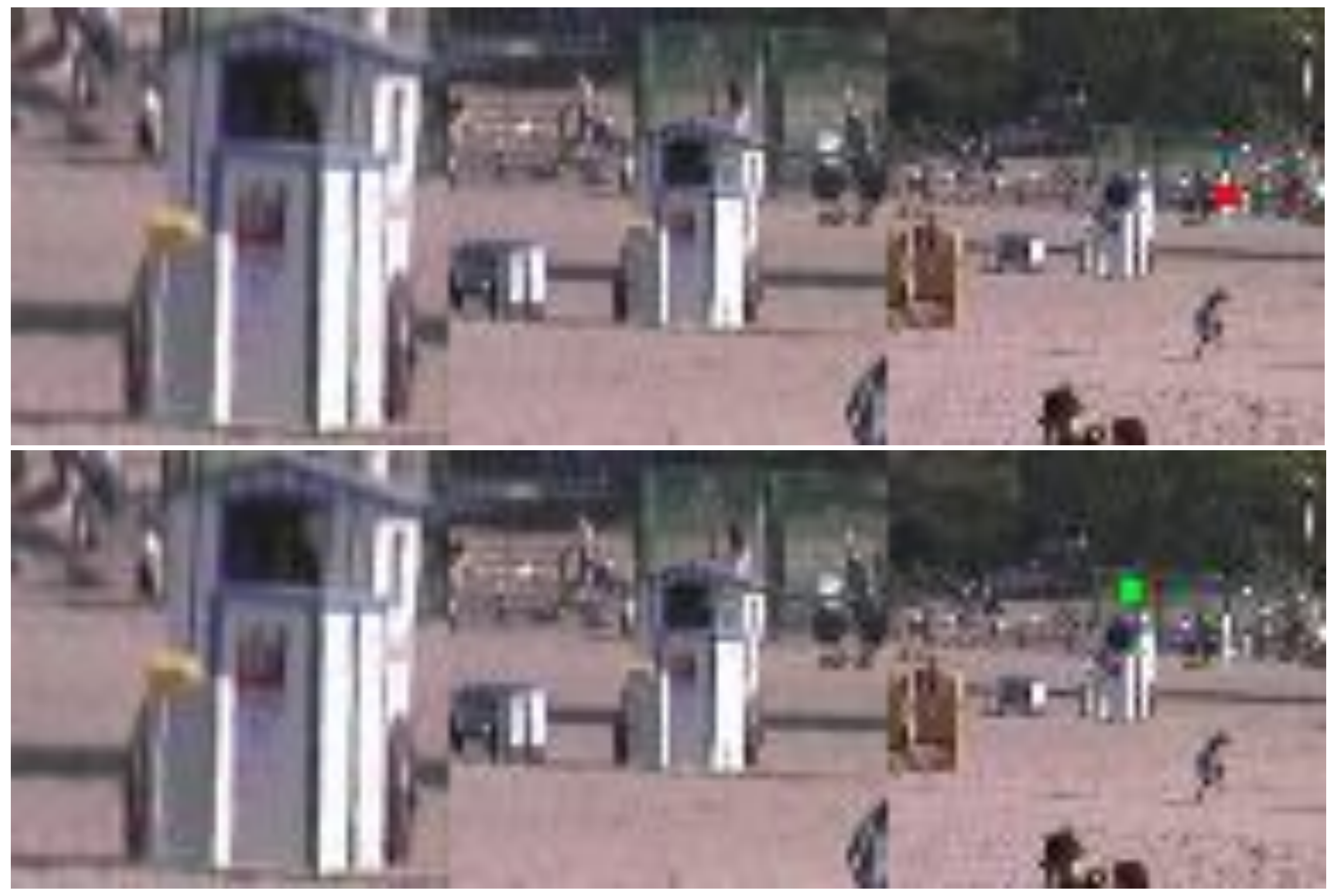
Are saccadic eye movements triggered by frame content?

Nicholas Seay, Emma Spellman, Katie Lamb | {nseay, espell, katlamb}@stanford.edu

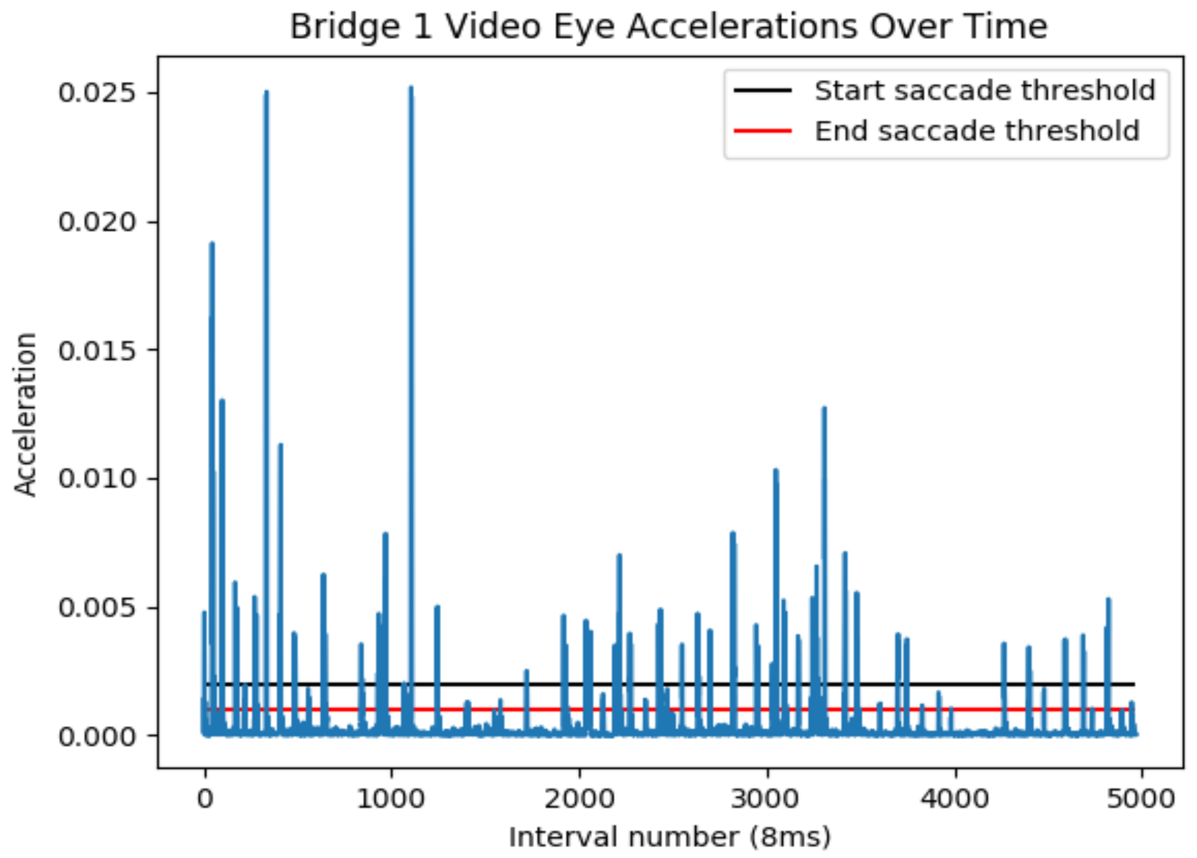
Problem Motivation

Humans either fixate on a point or moving object in an image, or their eyes will saccade, meaning they make a rapid jump to a new location. It is not well understood what motivates saccades, whether it's aspects of the image or signals from the brain independent from the image. With help from Dan Birman, a PhD candidate in Psychology, we sought to find if a machine learning model could predict where one will saccade to, which would indicate there are features of an image that motivate saccades.

Results



Top: A training entry of 3 down sampled crops. Red indicates where eyes saccade to
Bottom: Outputted prediction, green indicates where model predicts saccade location to



Right: Acceleration in pixel/ms^2 from start to end of 8ms intervals of a viewer's eye data

Approach

Data Sets

– Eye movement data for ~20 subjects on various natural movie clips [1]

– Frames from the videos subjects watched [1]

Extracting Saccades

– Found acceleration between pixels at start and end of 8 ms intervals

– When acceleration passed threshold, marked start of saccade [2][3]

– Pixel location at end of saccade is label value

Producing Images

- Cropped 3 squares of video frame from that timeframe centered at saccade start (512 x 512, 256 x 256, 128 x 128)
- Down sampled each crop to 64 x 64, simulates increasing blurriness at periphery
- Combined 3 crops into 1 (64 x 192) to feed into network, about 10K examples
- Label value moved into 512 x 512 area range if necessary, down sampled with crop to 64 x 64 image

Model

Architecture:

- *SingleCropEncoder*: 3 SingleCropEncoders (one for each crop) consisting of 3 ConvReluMaxPool layers
- *ThreeLayerConvTranspose Net*: Decoder utilizes two transpose convolutions. The first uses a ReLU activation while the second uses softmax.

Pretraining:

- We pretrained our SingleCropEncoders using data from CIFAR-10 which we up-sampled from 32x32 to 64x64 pixels [4]

Loss:

- The output of the model is a Gaussian distribution across the image frame, indicating the probability of saccading to each pixel. We use cross entropy loss with the probability outputted at the location of the true label.

Improvements:

We experimented with using pretrained VGG to obtain input for the decoder, but since its output is very deep with 512 channel width we need more time and data for better results.

Results and Analysis

The final test set accuracy we obtained was 22.39% when we divided pixels into 16 areas to predict from. This is certainly better than randomly guessing a location in the image, but not high enough to indicate from this model that saccades are triggered by features of the image. Perhaps the model must be more complex to learn more about the images, or it may be that saccades aren't determined by the images being looked at, but external factors in the brain. Our model is just a starting point and more research on the topic may provide indication or specific steps or layers to add that simulate eye perception better.

Looking Forward:

– More experimentation with networks like VGG with more data

– Our pretrain accuracy was 60.19% and improving this accuracy may boost model accuracy

References

[1]Universitat zu Ľubeck: Institut Ḟur Neuro–Und Bioinformat

[2] C. Araujo, E. Kowler, M. Pavel. ”Eye movements during visual search: The costs of choosing the optimal path.” 2001.

[3] K. Ehinger, et al. ”Modeling Search for People in 900 Scenes: A combined source model of eye guidance.” 2009.

[4] CIFAR 10 dataset collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, 2009.