Motion Selectivity of Neurons in Self-Driving Networks

Baladitya Yellapragada[1,2], Alexander Anderson[1,3], Stella Yu[1,2], and Karl Zipser[1,3]

[1] University of California, Berkeley, Berkeley CA 94720, USA
{baladityay23,aga,stellayu,karlzipser}@berkeley.edu
[2] International Computer Science Institute, 1947 Center St, Berkeley, CA 94704, USA
[3] Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA 94720-3198, USA

Abstract. We investigated if optical flow filters were implicitly learned by a neural network trained to drive a vehicle. The network was not trained to predict optical flow across the frames, but, through a series of controlled experiments, we claim that optical flow filters are present in the network. However, this appears to be only the case for sideways flows more relevant for steering predictions. For motor throttle predictions, the network looks at the variance of the pixels over time rather than computing optical flow. In addition, the filters that are likely used for motor throttle predictions dominate primarily in the middle of the network.

Keywords: Optical Flow · Motion Selectivity · Self-Driving · Autonomous Driving · Convolutional Neural Network · Stereoscopic Disparity

1 Introduction and Relevant Work

Our novel contributions are (1) showing a neural network trained to output two separate driving tasks (ie, steering and motor throttle predictions) can yield different motion-sensitive neurons that contribute to different output behaviors, and (2) demonstrating that we can probe these hidden filters through controlled experiments inspired by psychology. The experiment results indicate that optical flow filters are used for steering decisions, but variance filters are used for motor throttle decisions.

Our self-driving network takes in video from left and right cameras to predict future steering and motor throttle values, so there are many possible spatiotemporal cues that our network could respond to.

We first tried reproducing receptive field visualizations [1], [7]. Shown in Fig 1, we generated gradient ascent visualizations on the layers for an early CNN (2 convolutional layers and 2 dense layers) taking in 2 frames at a times. Across frames and cameras for any given neuron filter, Layer 1 receptive fields appear sensitive to optical flow and natural stereoscopic disparity.

However, this is hard to quantify, and later layers are even noisier. Furthermore, our current convolutional network is primarily the SqueezeNet architecture from Iandola, et al. [2]. We did not want to interpret unstructured visualizations from 1x1 and 3x3 filters. Instead, though not semantic, we labeled and compared inputs by presumed relevant features, similar to Zhou, et al. [8]. We then took inspiration from the general feature manipulation of predictive modeling experiments in psychophysics [6].

We studied optical flow because they provide cues about depth and future trajectories [5], and there is early evidence for them through gradient ascent analysis.

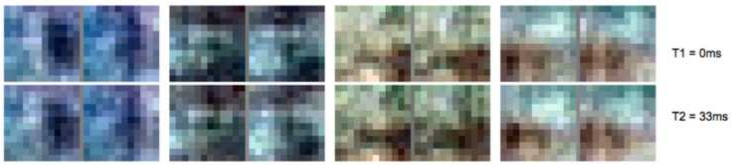

Fig. 1. Gradient Ascent Visualizations. Shown are four neurons' receptive fields from Layer 1 of our first self-driving network. Each neuron filter is divided into sub-filters, with one sub-filter per camera, per input frame – hence the 2x2 layout per neuron filter. These filters are appear sensitive to optical flow and stereoscopic disparity.

2 Experimental Setup

We labeled input videos by their average steer and motor throttle combinations. We only used videos whose current and future driving combinations had little variation, and the future ones had to be well predicted by the network. This allowed us to easily test on salient ego-motion videos containing one type of flow per video.

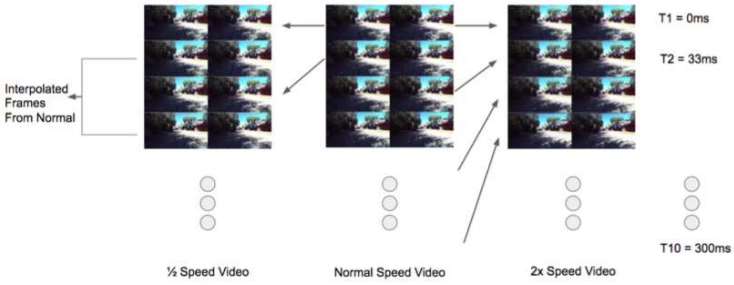

Fig. 2. Video Speed Manipulation. Natural videos are resampled for the optical flow experiment, to simulate optical flow changes invariant of other natural features. The network expects 10-frames of input video to the network, so each manipulated video samples the original frames to match the appropriate size. Sped up versions can just use future frames, but slowed down versions need the timepoints in between the normally captured frames, which are created using the interpolation method by Meyer et al. [4]

As seen in Fig 2, by speeding up and slowing down a given video, we created new videos with similar optical flow vectors across the visual field, but with more or less magnitude. We then compared how these affected output driving predictions to test the relevance of input video motion.

We also controlled the frame order and stereoscopic disparity in the input videos, after manipulating the video speed. If optical flow is a relevant feature for our driving predictions, then we should see a change in response with or without properly ordered time frames, similar to the network in Zhou, et al. [9]. Furthermore, if the network is attempting to recover depth cues from motion, it could be also affected by stereoscopic disparity, another source of depth cues present with our network setup.

3 Results and Discussion

Theoretically, we expected lower frame rate sampling to push predictions toward zero, and for higher frame rate sampling to do the opposite.

As seen in Fig 3, input video speed manipulation affects both steering and motor throttle predictions. This suggests potential optical flow sensitivity, but will need to be explored further.

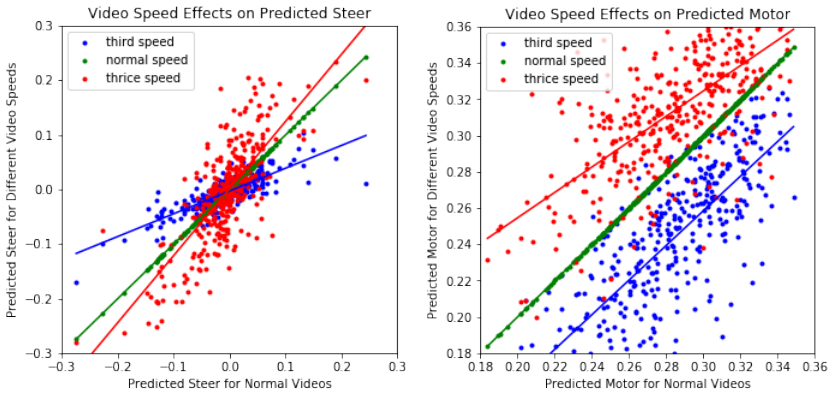

Fig. 3. Driving Predictions After Input Video Speed Manipulation. The output steer (left) and motor throttle (right) neurons' activations with respect to video speed changes are plotted. The X coordinates are normal video predictions, and the Y coordinates are changed-speed video predictions. Zero means no behavior for both plots. The fit lines indicate that speeding up the input video pushes steer predictions to become more extreme, as well as increasing throttle predictions. The opposite is also true for slower videos.

3.1 Temporal Controls

In Fig 4, steer and motor throttle predictions were plotted for input videos with different frame orders. Motor throttle predictions appear robust to frame order transformations, but the steering predictions are not.

As seen in Fig 5, changing around the frame order significantly impacts the video speed manipulation experiment for steer predictions. We need smooth flow of time, either forward or reverse, to get results similar to those from the video speed experiment in Fig 3. This implies optical flow filters are used for steer decisions.

For motor throttle predictions, changing around the frame order does not significantly impact the video speed manipulation experiment. Fig 6 shows motor throttle predictions are sensitive to input motion independent of frame order, implying that variance filters are used. Independent of frame order, little motion would yield little variance across the frames, whereas high motion would yield the opposite.

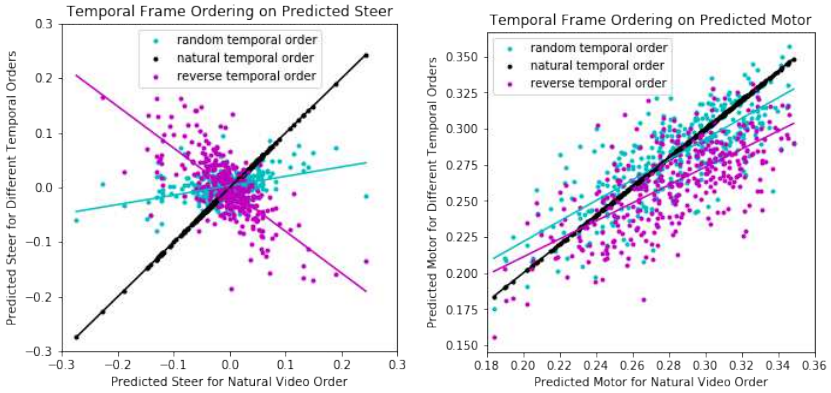

Fig. 4. Steer and Motor Throttle Prediction Changes From Temporal Frame Ordering. Changes to output steer (left) and motor throttle (right) neurons from input frame ordering are plotted. The X coordinates are naturally ordered video predictions, and the Y coordinates are predictions after temporal ordering. The fit lines for the steer plots indicate that randomizing the frame order nullifies any steering prediction, whereas reversing the order (not in the training set) reverses the steer prediction. The fit lines for the throttle plots indicate that randomizing and reversing the frame order had little impact on the throttle prediction.

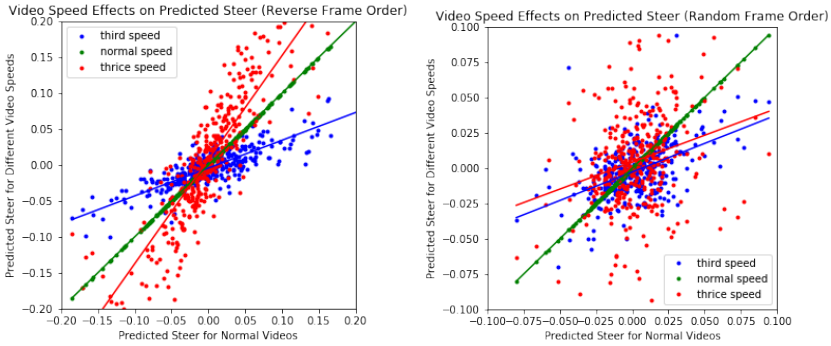

Fig. 5. Steer Predictions Changes From Temporal Frame Ordering After Video Speed Manipulation. Here, input videos are sped up and slowed down as in Fig 3, but also have their frame orders changed. We can see that reversing the frame order (left) maintains the natural steer changes correlated with video speed manipulation (as in Fig. 5), but randomizing the frame order (right) breaks the natural steer prediction changes after speeding up and slowing down the videos.

3.2 Steer and Motor Speed Results Across Stereo Controls

Lastly, for steer and motor speed predictions, stereoscopic disparity changes do not significantly impact the video speed experiment. Fig 7 shows that the motion selective filters for steer and motor speed predictions are independent of stereo features.

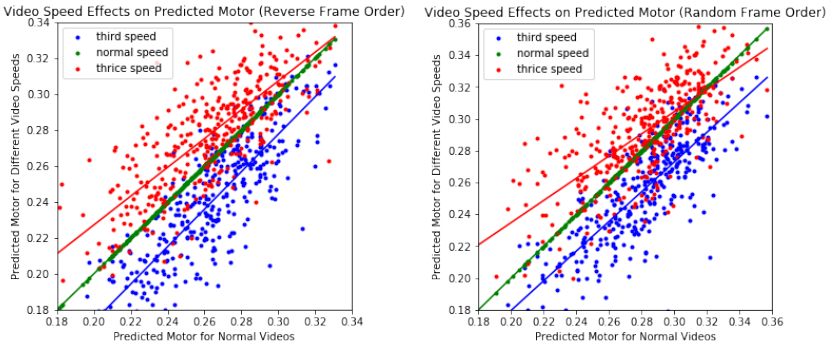

Fig. 6. Motor Throttle Prediction Changes From Temporal Frame Ordering After Video Speed Manipulation. Here, input videos are sped up and slowed down as in Fig 3, but also have their frame orders changed. We can see that both randomizing the frame order (left) and reversing the frame order (right) maintains the natural throttle prediction changes after changing video speed.

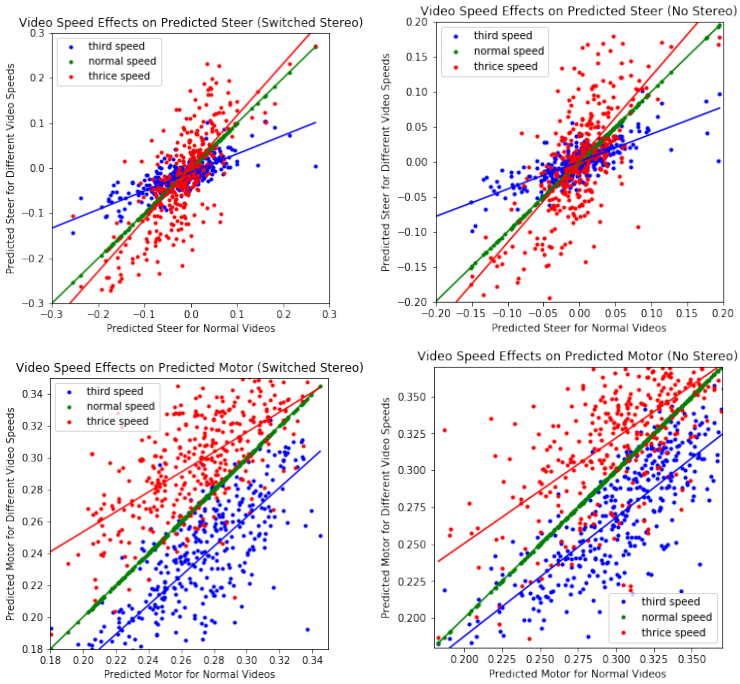

Fig. 7. Steer and Motor Prediction Changes From Stereo Effects After Video Speed Manipulation. Here, input videos are sped up and slowed down as in Fig 3, but also have their stereoscopic disparity changed. We can see that both switching the stereo (left) and removing the stereo (right) maintains the natural steer (top) and speed (bottom) prediction changes after speeding up and slowing down the videos, like in Fig 3.

4 Conclusion

We show that our network trained to predict steering and motor throttle from stereo video exhibits different motion-selective behavior for steering and throttle. Through a series of controlled psychophysical experiments, we demonstrated that both the steer and motor throttle predictions are correctly affected by varying the motion in the input video. However, even though both behaviors look similar on the surface, correct steer predictions are dependent on smooth frame order, whereas motor throttle predictions are not.

We show that steer decisions are based on optical flow filters in the hidden layers, whereas motor throttle decisions are based on variance filters.

Even though we did not present this in the paper, we did the same video speed experiments on hidden layer neurons as we did for the output neurons. By plotting average neuron activation for changed-speed videos versus normal speed videos, we can generate the same steer-like and motor-like profiles as in Fig 3. We further found the distribution of steer-like and motor-like neurons across the layers, arguing that these ultimately contribute to the final steer and motor throttle predictions. Linear SVMs were used to find the motor-like neurons based on their activation profiles, with the middle layers of our network having the most motor-like neurons.

From a theoretical standpoint, motor throttle only affects radially-dependent optical flow, but steering creates optical flow consistent throughout the visual field. The latter optical flow is easier for convolutional filters to capture, which we see in our results.

Lastly, consistent with Lundquist et al. [3], depth-sensitive stereo features are more difficult for convolutional networks to learn than other features. Our results appear to be robust to changes in stereoscopic disparity. It seems as though motion cues were more relevant than stereo cues in deciding changes in steer or motor throttle predictions.

References

1. Erhan, D., Courville, A., Bengio Y.: Understanding representations learned in deep architectures. Techreport (2010)
2. Iandola, F., Han, S., Moskewicz, M., Ashraf, K., Dally, W., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5MB model size. International Conference on Learning Representations (2017)
3. Lundquist, S., Paiton, D., Schultz, P., Kenyon, G.: Sparse Encoding of Binocular Images for Depth Inference. IEEE (2016)
4. Meyer, S., Wang, O., Zimmer H., Grosse, M., Sorkine-Hornung, A.: Phase-based frame interpolation for video. IEEE Conference on Computer Vision and Pattern Recognition (2015)
5. Saunders. J.: View rotation is used to perceive path curvature from optic flow. Journal of Vision (2010)
6. Yarkoni, T., Westfall, J.: Choosing prediction over explanation in psychology: lessons from machine learning. Perspectives on Psychological Science (2017)
7. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. European Conference on Computer Vision (2014)
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. International Conference on Learning Representations (2015)
9. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal Relational Reasoning in Videos. arXiv (2018)