Direct Image Mono-Camera Localization Using Deep Learning

Matthew W. Powelson **Stephen L. Canfield**
Department of Mechanical Engineering

Introduction

Mono-camera localization is an active area of robotics research. Historically, localization has been carried out in a 2D space using sensors such as lidar. However, with the prevalence of monocular RGB cameras in highly mobile systems – for example cell phones or drones – localizing from a camera image is an area of increasing interest. While typical methods involve creating a 3D map or comparing subsequent images to extract odometry information, this poster presents a deep learning approach that estimates camera pose directly from a single RGB image.

Approach: The Model

The Inception V3 architecture [1] built in Tensorflow [2] is used as the model feature detector. This is followed by a global average pooling layer and two fully connected layers with 1024 nodes each and tanh activations. These layers act as a regressor that predicts the pose based on the features in the image.

Approach: The Data

Two videos were taken using a Google Pixel cell phone while the 6 DOF pose of the phone was tracked using an OptiTrack motion capture system. Each video frame is used as an example and is scaled to 216 x 384 x 3. In this initial test, only a single angle is considered.

The video was captured by rotating the camera in one axis while keeping the location within 10 cm of the starting location. In Figure 1, the frames of the validation video are stitched together for visualization.

Learning the Weights

A transfer learning approach is used to leverage highly trained models that are integrated with TensorFlow and have been trained on the ImageNet dataset. Since the same features that are useful for identifying images for classification are useful for characterizing a location, these features are fed into the custom pose regressor to predict the output.

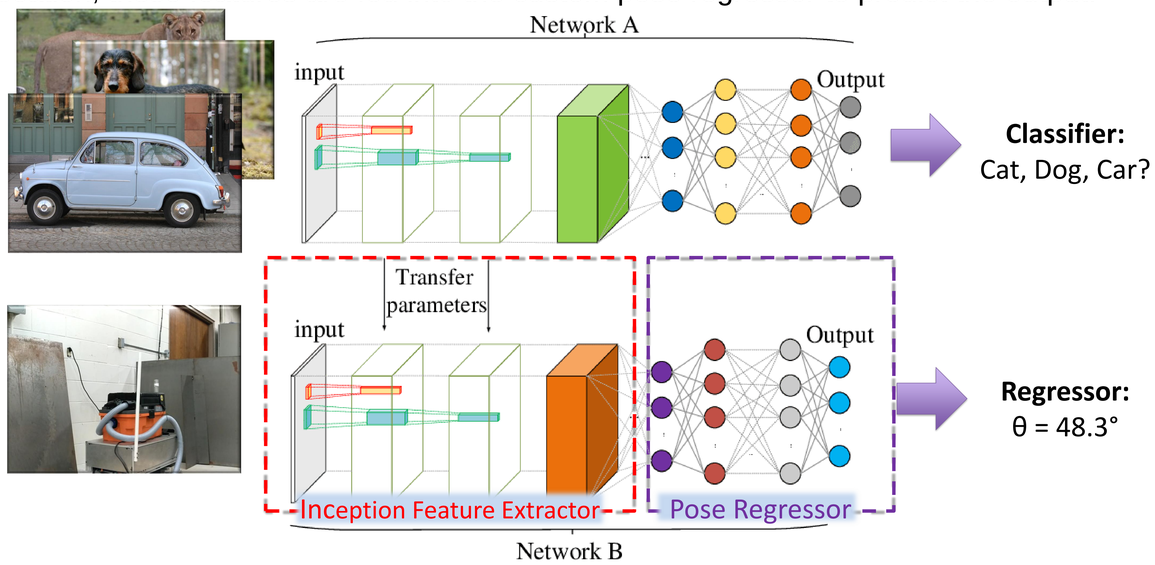

Figure 2: Illustration of transfer learning and the use of Inception V3 as a feature extractor for regression [3]

Prediction

As the model is trained, the regressor "learns" the room and the features associated with each pose. The training image on the left (Figure 3) shows the performance of the model on the video used to train the model. The validation image on the right (Figure 4) shows the model's performance on a separate video used only for evaluation.

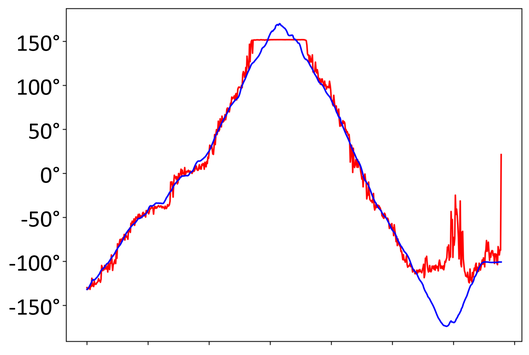

Figure 3: Training video with angle vs frame. Blue is OptiTrack pose and red is the model prediction.

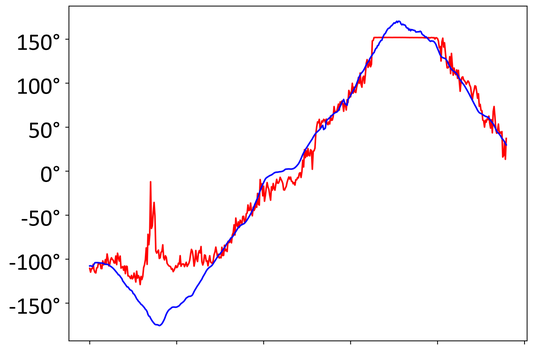

Figure 4: Validation video with angle vs frame. Blue is OptiTrack pose and red is the model prediction.

Figure 1: Validation video stitched into a panorama image with pose angles labeled

Discussion of Results

Both the training and validation sets show reasonable agreement between the OptiTrack pose and the predicted pose from the model. Further, regions near the pose singularities exhibit the highest errors, but these are found in both the training and validation sets – indicating the need for additional training or architecture changes rather than suggesting overfitting.

Conclusions

This poster has shown that one dimensional localization using a single RGB image is possible using the deep learning technique proposed. By using a highly trained feature extractor, a relatively low resolution image can be used to localize with reasonable precision over the majority of the workspace. Applications include indoor mapping, aids for individuals with visual impairments, and swarm robot navigation.

References

[1] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826).

[2] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). TensorFlow: A System for Large-Scale Machine Learning. In OSDI (Vol. 16, pp. 265-283).

[3] Lemley, J., & Corcoran, P. (2017). Transfer Learning of Temporal Information for Driver Action Classification. In The 28th Modern Artificial Intelligence and Cognitive Science Conference (MAICS).

Acknowledgments

The authors would like to acknowledge the use of facilities provided by the Denso Foundation grant for the Intelligent Vehicle Development Cluster