Matthew Millendorf, Nathan Vollertsen, Matthew Ekins 20 February 2019

Professor Pengyu Hong

COSI 149b - Practical Machine Learning with Big Data

Room Number Detection Project - Phase 1 Report

The analysis of information from computer vision systems has seen a great improvement with the advent of deep learning techniques. The perception of an environment is integral to the success of robotics and autonomous transportation. In this project, the application of such models will be used to detect and read room numbers from class room placards. To perform this task, two separate models were strung together in the pipeline. In this paper, a detailing of each model will be provided as well as background information about their specific role in the pipeline. Results of the pipeline will follow as well as a discussion regarding the next steps in producing a program with optimal room number recognition accuracy.

The objective of the first stage in the pipeline is to identify regions in the video frames that contain text. While there exists several algorithms to detect text from natural scene imaging, the model proposed in [1] best various approaches on benchmarks. This EAST model consists of three main stages: feature extraction, feature-merging, and geometry map generation. In the first stage, features are extracted through convolutions and pooled to produce four feature maps from the original input. Each feature map is then passed onto the feature-merging stage where it is rescaled, concatenated with the other feature maps, and then convolved over by a 3x3 window. After the last feature map has been passed onto the merging stage, a 3x3 convolution is then used to produce the final feature map. The final feature map is passed into the output layer where the confidence score for the region of interested containing text is computed as well as the bounding box around the text. This model utilizes two different types of geometric mapping, one is RBOX (4 axis-aligned bounding boxes with a rotation) and QUAD.

In the second stage of the pipeline, an optical character recognition (OCR) model is fed the bounding boxes from EAST and outputs the characters found in that segment of the video frame. The OCR engine [2] developed by HP and later acquired by Google, Tesseract, was used for this portion. The pipeline of the OCR engine involves several steps. First, the (invisible) line in which the text resides along is computed, allowing the model to vertically distinguish between characters. These lines are referred to as “baselines”. Then, Tesseract segments the characters, which to the machine, appears to be random lines, whether they be of monospaced or non-monospaced font. Finally, blobs of line segments are then further segmented based on their line segment’s concavity and a static classifier is used to determine the class of the characters.

The results of our current pipeline are promising but insufficient. A key failure is the inability for the model to distinguish between areas with text and areas without. This leads to a high amount of false positives and decreases the effectiveness of the program. An additional concern is that the current model is unable to separate text found on posters from text found on the classroom door signs. To remedy these two ailments, our group proposes a “sandwiching” of the EAST and Tesseract model. The first model will implement at CNN, which is trained on classroom signs, to detect classroom signs. Training of this model will be done with images and with the use of transfer learning, will be applied to video, as shown by [3]. The fourth and final model will be an RNN trained on the classrooms found in [4]. This final stage will ensure that the output text from the OCR is that of a potential classroom, and not a false positive that found its way through the first model.

References:

[1] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[2] R. Smith, “An overview of the tesseract OCR engine,” in Proc. 9th Int. Conf. Doc. Anal. Recogn., Sep. 2007, vol. 2, pp. 629–633.

[3] Y.-C. Su, T.-H. Chiu, C.-Y. Yeh, H.-F. Huang, and W. H. Hsu, “Transfer learning for video recognition with scarce training data for deep convolutional neural network,” arXiv preprint arXiv:1409.4127, 2014.

[4] http://registrar-prod.unet.brandeis.edu/registrar/schedule/search?strm=1191&view=UGRD&status=&time=time&day=mon&day=tues&day=wed&day=thurs&day=fri&start\_time=07%3A00%3A00&end\_time=22%3A30%3A00&block=&keywords=&order=class&search=Search&subsequent=1&page=