# Capstone Proposal
# Why don't you focus on the road?

Ennio Nasca
nascaennio@gmail.com

April 16, 2020

## 1 Domain Background

According to the World Health Organization, road traffic injuries cause more than 1M deaths worldwide every year[15]. In particular, among the different causes that may lead to car accidents (e.g. speeding and driving under the use of alcohol/drugs) *distracted driving* is one of the most common. It suffices to say that, according to the WHO, drivers that use their phone are 4 times more likely to get involved in an accident. These are shocking numbers, that not only make you pause and reflect, but need to be a call to action, because, being able to detect if a driver is absent-minded or, even worst, with a hand on their phone may save real lives. Thanks to all the recent developments in Computer Vision we can perform tasks such as real-time object detection[10, 7] and segmentation[5, 3] that can be useful for the problem at hand.

## 2 Problem Statement

*Activity recognition* is defined as the task of recognizing the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. The goal of this project is to identify the action that a person is carrying out while sitting on the driver's seat of a car. The problem can be formulated as a multi-label classification task. Our goal is to develop a model $f : \mathbf{X} \rightarrow \mathbf{N}$ that is able to predict, starting from a 2D dashboard camera image $\mathbf{X}$, the driver's action from a finite set of discrete labels $\mathbf{N} = \{1, \ldots, n\}$ (e.g texting, talking or safe driving).

## 3 Datasets and Inputs

The data that will be used in this project is provided by State Farm, an insurance company that created a Kaggle competition[12] back in 2016. The company mounted dashboard cameras on different vehicles and took 2D images picturing drivers while performing different type of actions. Information related to the dataset dimensionality and base statistics are reported in table 1.

| Categories | Units |
|---|---|
| training images | 22424 |
| testing images | 79726 |
| drivers | 26 |
| classes | 10 |

Table 1: Dataset information

After a first look at the images we can see that this dataset is particularly suitable for our task, it contains many differnt real scenarios for which a driver may get distracted (e.g. operating the radio, talking to the passanger and texting). Nonetheless, we will need to tackle some difficulties that may arise due to the presence of multiple drivers and trickier classes, namely texting with right or left hand. In general, the dataset is well structured and documented and provides the necessary images $\mathbf{X}$ and labels $\mathbf{N}$ required to train our previously defined model.

## 4 Solution Statement

When it comes to image classification Convolutional Neural Network[6] are the fundamental building blocks of state of the art models. For this project we will build on top of an existing family of CNN models that are used for Human Pose estimation (HPE)[14, 2, 13], a task that has enjoyed the attention of the Computer Vision community for the past few decades. Neural Netowrks trained for HPE on dataset such as COCO[8] and MPII[1] can act as feature extractor and provide crucial information useful to better understand people, and in particular drivers, behaviour in images and videos. The goal of HPE is to localize of human joints (also known as keypoints - elbows, wrists, hands, etc) in images or videos.

We can formalize our solution for the State Farm challenge as follows: our model $f$ becomes a composed function $f = g(h(\mathbf{X}))$ where $h$ is the pre-trained HPE model and $g$ is the classifier (e.g. XGBoost) that uses the extracted localized information to carry out the prediction.

## 5 Benchmark Model

As stated in previous sections, Convolutional Neural Networks have proved to be extremely effective for the task of image classification. For this reason, the results obtained with our model $f$ will be compared to a CNN baseline: a VGG-16[11] architecture trained using transfer-learning[9], i.e. the process of using knowledge gained while solving one problem and applying it to a different but related problem. In our case, the original VGG-16 architecture has been trained on the ImageNet[4] dataset, and fine-tuned on the State Farm dataset.

Additionally, since the data we will be using comes from a Kaggle competition, we can compare our result to the top-ranking ones available in the Public Leaderboard.

## 6 Evaluation Metrics

This project treats the activity recognition problem as a supervised multi-label classification problem. For this project we decided to assess the goodness of a model adopting the same metric used to evaluate submissions on Kaggle, i.e. multi-class logarithmic loss. The formula is then:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

where $N$ is the number of images in the test set, $M$ is the number of image class labels, $log$ is the natural logarithm, $y_{ij}$ is 1 if observation $i$ belongs to class $j$and 0 otherwise, and $p_{ij}$ is the predicted probability that observation $i$ belongs to class $j$

## 7 Project Design

In order to obtain a model $f = g(h(\mathbf{X}))$ that is able to recognize the activity that a driver is carrying out we can split this complex task into two smaller and simpler subproblems, namely feature extraction and classification, as shown in figure 1.


Figure 1: The proposed workflow: from left to rigth, the image $\mathbf{X}$ is given as input to the pipeline. The Human Pose Estimation model $h$ (OpenPose) estimates the position of the keypoints. Finally the classifier $g$ (XGBoost) uses these keypoints to predict the image label.

The former part of the workflow will explore the effectiveness of a Human Pose Estimation model like OpenPose[2] to extract body keypoints from an input image $\mathbf{X}$. Besides, based on the performance results achieved, we may consider extending the use of HPE models to extract additional keypoints associated to the driver's face and/or hands position.

The latter step of the workflow will build a classifier $g : h(\mathbf{X}) \rightarrow \mathcal{R}^N$ that will map the new image representation $h(\mathbf{X})$ to a probability vector $x = [x_0, \ldots, x_9]$ and the image label will be assigned to the class with the highest probability.

## References


[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[3] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2019.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[9] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] State Farm. State farm distracted driver detection, 2016. `https://www.kaggle.com/c/state-farm-distracted-driver-detection/overview`, Last accessed on 2020-04-14.

[13] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[14] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[15] World Health Organisation. Road traffic injuries, 2020. `https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries`, Last accessed on 2020-04-14.