Utilizing Driving Context to Increase the Annotation Efficiency of Imbalanced Gaze Image Data

Johannes Rehm, Odd Erik Gundersen, Kerstin Bach
Norwegian University of Science and Technology
{johannes.rehm,odderik,kerstin.bach}@ntnu.no

Irina Reshodko
Way AS
irina@way.no

Abstract

Knowing where the driver of a car is looking, whether in a mirror or through the windshield, is important for advanced driver assistance systems and driving education applications. This problem can be addressed as a supervised classification task. However, in a typical dataset of driver video recordings, some classes will dominate over others. We implemented a driving video annotation tool (DVAT) that uses automatically recognized driving situations to focus the human annotator's effort on snippets with a high likelihood of otherwise rarely occurring classes. By using DVAT, we reduced the number of frames that need human input by 87% while keeping the dataset more balanced and using human time efficiently.

1 Introduction

Acquiring annotated data of high quality that is representative of a specific domain is a critical part when developing supervised machine learning systems. Insufficient data can bottleneck the performance by a huge margin (Shao et al. [2019]). However, collecting and annotating data of high quality is both time-consuming and expensive, especially when this has to be done manually. One of the most common machine learning tasks is *classification*, which involves attributing a class label from a finite set of labels to new inputs. The workload and costs of data collection for classification problems increase further in situations where some of the classes are rare relative to the others (*imbalanced* distribution). In this case, very large amounts of data may need to be processed to get enough representative samples for each of the rare classes.

This is the exact problem we faced when developing a machine learning system that can classify which mirror or blind zone the driver of a car is looking at based on images of the driver's face. Similar to Ribeiro and Costa [2019], we call these target areas of the driver's attention *gaze zones*. Side mirrors, rear-view mirror, windshield, and left and right blind zones constitute the *gaze zone classes* in our classification problem. Observing the traffic environment, especially other road users, is an essential part of driving a car on public roads. However, the driver mostly looks out the front window of the car and will only look in the different mirrors from time to time, as what goes on in front of the car is in most situations considered the most important part of the environment for the driver. As the task of classifying where the driver is looking can easily be formulated as a supervised learning problem, labeled data is required. The data must be collected and annotated. The unlabeled source data consists of video recordings of the driver's face when driving a full-scale car simulator situated in a virtual environment.

The objective of the research presented here is to collect a balanced dataset for training supervised image classification algorithms while minimizing the time spent on searching for the sparse classes. The time of a human annotator should be spent on labeling a subset of the source data that is more balanced in regards to the classes than the source dataset. To speed up the annotation process, we use context information about the driving situation collected from the driving simulator. Our hypothesis is that the rare classes are most likely to appear during the situations where the parts of the environment

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

that are not in front of the car become important, such as when turning in an intersection or changing lanes. Automating the identification of these situations is relatively easy when driving a simulator in a virtual environment. The contributions of the research presented here include: 1) *a description of the problem of annotating an imbalanced gaze dataset using driving context*, 2) *a description of the Driving Video Annotation Tool (DVAT)* we developed for speeding up the annotation process of a highly imbalanced dataset, 3) *a description of the integration of the annotation workflow in a data-centric model development process*, and 4) *an analysis of the efficiency gains* from using DVAT.

2 Problem Description

Advanced driver assistance systems [Paul et al., 2016], educational car systems [Sharon et al., 2005], and virtual driving instructors [Weevers et al., 2003] require assessing the current situation and give feedback to the driver. A system that is capable of monitoring where the driver's gaze is directed can greatly improve the accuracy of the situation assessment and the relevance of the feedback. Eye-tracking systems (e.g. Selim et al. [2020]) and direct classification of the driver's face images (e.g. Rangesh et al. [2020]) are popular approaches to solve this issue. However, both validation and training of such systems require a dataset that includes ground truth annotations of the gaze zone classes. Obtaining such a dataset is time-consuming, and is most often performed semi-automatically, for example, by recording human participants with a camera and asking them to look at predefined gaze zones (Ortega et al. [2020], Ribeiro and Costa [2019]). The participants might be sitting in a car, but the car is usually standing still for safety reasons. The annotations in this approach are created automatically without requiring any additional manual work, but the scenario of sitting in a still car and being asked to focus on certain gaze zones does not approach realistic driving scenarios. The duration of the driver's attention to each zone, the rotation of the head and eyes can change significantly while driving in a realistic situation such as during an overtake. From our visual inspections, we see that head and eye movements are usually much faster, leading in some cases to motion blur (e.g. if side mirror and blind spot observation are performed within a continuous head movement), and there are bigger variations and larger extremes in head poses and positions of the pupils.

These drawbacks motivated us to create a dataset from real driving sessions of drivers in a virtual world using a car simulator. To our knowledge, to date, there is no work published on creating a gaze zone dataset using real driving sessions. In this work, participants are recorded while using a high-fidelity driving simulator (Allen et al. [2007]). In realistic situations, drivers tend to look into the mirrors for a short time (by our measurements mainly in the range of 200ms - 500ms) and mostly during specific maneuvers. This makes it very tedious to manually locate all relevant frames where the driver's attention is on the right gaze zone. Another issue with this approach is that in a realistic scenario, the driver mostly looks straight ahead through the windshield, and thus the data becomes imbalanced.

Formally, supervised learning of a set of image classes can be formulated as follows: Given a training set of N example pairs $(x_1, y_{x_1}), (x_2, y_{x_2}), ..., (x_N, y_{x_N})$ where $x_i \in \mathcal{X}$ are images that belong to a finite set of classes $y_j \in \mathcal{Y}$, for a new input $x \in \mathcal{X}$ assign $y \in \mathcal{Y}$. In our case, $\mathcal{X}$ is a set of images of the upper body of drivers captured by a fixed driver-facing camera mounted inside a car simulator. An example image is shown in Figure 2 (5). The set of classes $\mathcal{Y}$ contains the *gaze zones* $\{\text{windshield}, \text{rearMirror}, \text{leftMirror}, \text{rightMirror}, \text{leftBlindspot}, \text{rightBlindspot}\}$ indicating when the driver looks through the windshield, in the rear-view mirror, left-side mirror, right-side mirror, left blind-zone or right blind-zone.

In a perfectly balanced dataset $\mathcal{Q}$, each of the M classes of $\mathcal{Y}$ is equally probable: $P(y_1|\mathcal{Q}) \approx P(y_2|\mathcal{Q}) \approx ... \approx P(y_i|\mathcal{Q})$. However, our problem is that the source data $\mathcal{W}$ is highly unbalanced in favor of the gaze zone *windshield*: $P(\text{windshield}\,|\mathcal{W}) \gg P(\text{rear}\,|\mathcal{W}) \approx P(\text{left}\,|\mathcal{W}) \approx P(\text{right}\,|\mathcal{W})$. *Shannon entropy* $H_{\mathcal{W}} = \sum_i P(y_i|\mathcal{W}) \log_2(P(y_i|\mathcal{W}))$ is a common measure of balance in a dataset (Mitchell [1997]). The maximum possible entropy $H_{\mathcal{Q}} = \log_2(M)$ corresponds to the perfectly balanced dataset, while $H_{\mathcal{I}} = 0$ corresponds to a dataset with only one class present. The goal of this work is to create a video annotation tool that efficiently uses the annotator's time to create a dataset that is more balanced compared to the brute-force labeling of each video frame. We have implemented a two-step process where the first step is an automated filter $f(C, S, \mathcal{W}) = \mathcal{V}$ which uses the context C about the driving situation S to retrieve a subset $\mathcal{V} \subset \mathcal{W}$ that has higher entropy than $\mathcal{W}$ and thus more balanced. The second step g is not automated but implemented as a

tool that allows a user to manually specify and annotate the subset of $g(\mathcal{V}) = (\mathcal{V}, \mathcal{Y})$. This dataset will still hold many images of the windshield class. In the most labor-intensive case of the annotator labeling every single frame we present, the resulting dataset $g(\mathcal{V})$ is still more balanced than $\mathcal{W}$. In the less labor-intensive case where the annotator is guided by the counts of samples that were assigned to each of the classes, we get a dataset that is even more balanced than $g(\mathcal{V})$ but may contain fewer samples overall.

3 Gaze Zone Annotation Workflow and Model Development Process

We follow a data-centric model development process shown in Figure 1. We receive new video recordings from the simulators multiple times a day, allowing us to include the data collection process in the model development process. The question of how much data of what quality is required to be annotated is continuously evaluated after each new deployment and, if needed, adjusted.

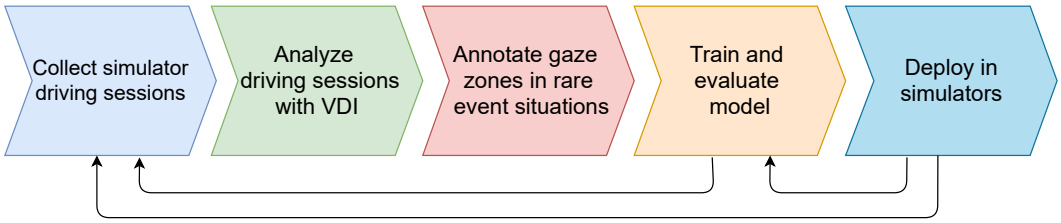

Figure 1: Annotation workflow integrated in model development process (derived from Ng [2021]).

The first step is to collect new simulator driving sessions. The driving sessions include all data required to replay and re-render the complete driving. The driver-facing camera stream is recorded in sync with the simulation time. The gaze zones for these sessions are then annotated using the two-step process described above. The session is automatically analyzed by the virtual driving instructor (VDI) described in Sandberg et al. [2020]. This analysis creates context information C, e.g. timestamps of events like a lane change. DVAT uses C to filter out a set of situations S in which it expects the driver to look into specific gaze zones, e.g. a few seconds before a lane change event occurred. S is used in turn to create an annotation task. This task is then assigned to an annotator. The annotator does not need to check the complete video but can just annotate one situation and jump to the next one. It is also possible to semi-automate the annotation process by pre-annotating the data using the current model and encode it in the annotation task, similar to Russell et al. [2008] and Vondrick et al. [2013]. This restricts the labeling task to correcting model errors.

The annotators can perform the task at their schedule over the web frontend depicted in Figure 2. The frontend shows a video of both the driver's face and the driving scene itself. This flexible setup allows the use of DVAT for any kind of driving or traffic situation annotation. Being able to playback the video makes it easier to distinguish where the driver is looking, as the whole head movement is visible, which is a clear advantage over a single still image. Gaze zone observations are strongly correlated to driving decisions. Observing a driving maneuver gives additional context on what the driver is currently doing. This allows us to create highly accurate annotations. The annotator can label a frame with a gaze zone class by marking the corresponding checkbox as can be seen in Figure 2 (4). A single frame can be marked by just clicking on the checkbox. If the annotator presses the *Ctrl* key and clicks on the checkbox, the label is locked, and the checkbox state will be propagated to the next frame as long as the lock is engaged. The number next to the label is the count on how often the class was labeled in this session. The annotator can use this count information to focus on gaze zones that were labeled less often so far. They can also just skip frames by not putting any label. The model performance and the quality of the annotations are monitored by domain experts. Domain experts are in our case, simulator-trained driving teachers. To reduce cost, annotators do not need to be domain experts. However, to ensure high-quality labels and to train them, domain experts control a random subset of the labels.

This approach restricts the annotation task from processing a complete video to a couple of situations with a duration of a few seconds each. After the labeling process is finished, the newly acquired data is added to either the training or the validation set using a random assignment which itself launches a new training and validation process. This allows to track model accuracy improvements over time.

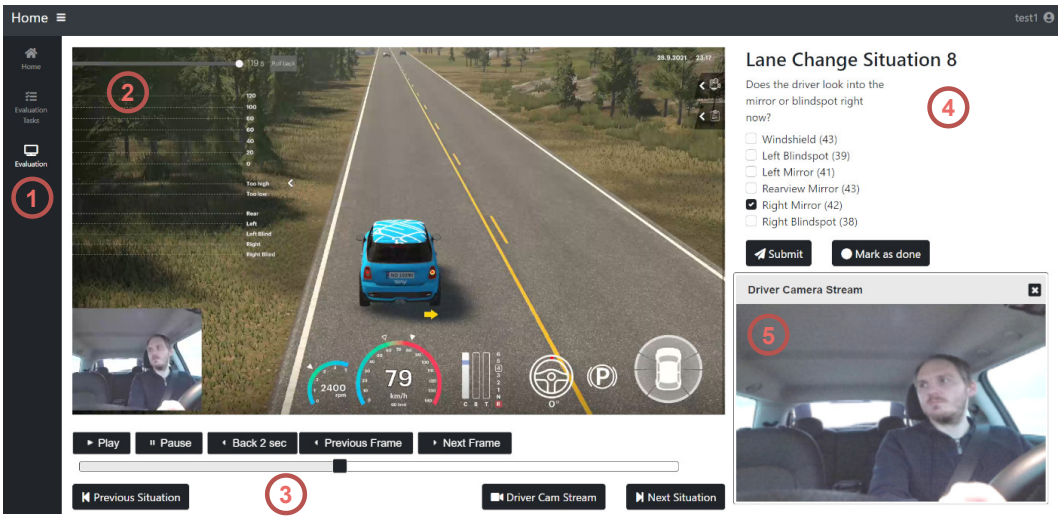

Figure 2: Screenshot of Annotation Framework Web Frontend. *1. Navigation Menu 2. Simulation visuals streamed from streaming server 3. Controls which allow to jump from situation to situation 4. Annotation input form 5. Driver camera stream received from streaming server*

Table 1: Results Analysis

Gaze Zone Class	**Source** $\mathcal{W}$ (frequency / $P(y_i\|X)$)	**Filtered Subset** $\mathcal{V}$
Windshield	10583 / 90.8%	857 / 56.5%
Left Blindspot	125 / 1.1%	125 / 8.2%
Left Mirror	120 / 1.0%	103 / 6.8%
Rearview Mirror	507 / 4.3%	216 / 14.2%
Right Mirror	204 / 1.8%	99 / 6.5%
Right Blindspot	118 / 1.0%	118 / 7.8%
Entropy (bits) [0, 2.58]	0.63	1.97

The task of the domain experts is also to analyze the deployed model performance to steer further data collection and annotation requirements.

4 Results and Conclusion

We annotated the gaze zones of a simulator driving session for every frame. This gives us an approximation of how imbalanced such a dataset would be and how more balanced the dataset is if we apply our context filtering approach. The distribution and entropy of the resulting source dataset $\mathcal{W}$ can be seen in the second column of Table 1. We use lane changes as the context information C to extract a more balanced subset of the dataset $f(C, S, W) = \mathcal{V}$ as they had a high occurrence rate in this particular session. Furthermore, we define the start time of the corresponding situations S as $n = 5$ seconds before the lane change and the end time of the situation as the time of the lane change. As can be seen in Table 1, this increases the entropy of the dataset to 1.97 bits while a perfectly balanced dataset with 6 classes would have an entropy of 2.58 bits. The number of images that need to be processed by a human annotator are reduced by 87% in this example. However, most of the samples from the rare classes remain in the filtered subset.

We presented an efficient method to label gaze image data. By utilizing context information about driving situations, we automatically get a subset of the complete data which includes a higher proportion of otherwise rarely occurring classes. This considerably reduces the effort for human annotators to label a fairly balanced dataset from very unbalanced source data.

References

R Wade Allen, George D Park, Marcia L Cook, and Dary Fiorentino. The effect of driving simulator fidelity on training effectiveness. *DSC 2007 North America*, 2007.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

Andrew Ng. MLOps: From Model-centric to Data-centric AI, 2021. URL `https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf`.

Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-an Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. DMD : A Large-Scale Multi-Modal Driver Monitoring Dataset for Attention and Alertness Analysis. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops (Accepted)*, 2020.

Aneesh Paul, Rohan Chauhan, Rituraj Srivastava, and Mriganka Baruah. Advanced driver assistance systems. Technical report, SAE Technical Paper, 2016.

Akshay Rangesh, Bowen Zhang, and Mohan M Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1054–1059. IEEE, 2020.

Rafael F Ribeiro and Paula DP Costa. Driver gaze zone dataset with depth data. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3): 157–173, 2008.

Martin KH Sandberg, Johannes Rehm, Matej Mnoucek, Irina Reshodko, and Odd Erik Gundersen. Explaining traffic situations–architecture of a virtual driving instructor. In *International Conference on Intelligent Tutoring Systems*, pages 115–124. Springer, 2020.

Mohamed Selim, Ahmet Firintepe, Alain Pagani, and Didier Stricker. Autopose: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2020. URL `http://autopose.dfki.de`.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.

T. Sharon, T. Selker, L. Wagner, and A.J. Frank. Carcoach: a generalized layered architecture for educational car systems. In *IEEE International Conference on Software - Science, Technology Engineering (SwSTE'05)*, pages 13–22, 2005. doi: 10.1109/SWSTE.2005.9.

Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013.

Ivo Weevers, Jorrit Kuipers, Arnd O. Brugman, Job Zwiers, Elisabeth M.A.G. van Dijk, and Anton Nijholt. The virtual driving instructor creating awareness in a multiagent system. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 2671, pages 596–602, 2003. ISBN 3540403000. doi: 10.1007/3-540-44886-1_56.