World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:16, No:12, 2022

Improvement of Ground Truth Data for Eye Location on Infrared Driver Recordings

Sorin Valcan, Mihail Găianu

Abstract—Labeling is a very costly and time consuming process which aims to generate datasets for training neural networks in several functionalities and projects. For driver monitoring system projects, the need of labeled images has a significant impact on the budget and distribution of effort. This paper presents the modifications done to a ground truth data generation algorithm for 2D eyes location on infrared images with drivers in order to improve the quality of the data and performance of the trained neural networks. The algorithm restrictions become tougher which makes it more accurate but also less constant. The resulting dataset becomes smaller and shall not be altered by any kind of manual labels adjustment before being used in the neural networks training process. These changes resulted in a much better performance of the trained neural networks.

Keywords—Labeling automation, infrared camera, driver monitoring, eye detection, Convolutional Neural Networks.

I. Introduction

TRAINING neural networks requires large datasets with good accuracy in order to have a general and precise detection model. The generation of these datasets is a very expensive and time consuming aspect of this process, with big teams of people working in data marking. Of course, there are specific tasks, such as price prediction or sentiment analysis, where data already exist on the internet because they are naturally generated by people. For the problem addressed in this paper, which is eyes location in infrared images with drivers, these marked data can be very difficult to obtain.

Another big problem in the labeling process is that hours of marking eyes on images is a very strenuous task and it always leads to natural human errors. This usually leads to a double checking process for the labeling, which means additional time and effort and it does not guarantee the expected quality improvement.

Over the past period we realised that we have several themes which reveal from machine learning about data labeling. Data is the new oil and a fundamental component of training models.

Dark data like scientific articles or medical images are valuable but very hard to use as training data because it is not machine readable which makes it inaccessible. Only 20% of data is visible to computer systems at the moment and dark data will grow with 93% by 2023 according to IBM [1]. Data labeling will expand use cases for machine learning models and can also improve their precision.

In supervised learning, algorithms learn from labeled data. When the problems complexity increases and more dimensions are added to the model, small data volumes can result in overfitting. Training data are a bottleneck. Manual labeling can be very expensive in terms of time and budget. Data scientists often spend up to 80% of their time cleaning and preparing data instead of focusing on development of new methods [2]. Many large organizations bring data labeling in house. Manual labeling can become very expensive at scale (more than 100K images). Under these conditions, solutions for automated data labeling will be best positioned and feasible because they reduce most of the manual labeling effort, ideally all of it, including building of workforce, labelers training and quality control to the same degree.

Grounded theory is an incipient method for building hypothesis based on qualitative data. Instead of formulating hypothesis that leads to the data collection process, researchers are trying to generate ones that emerge from existing data. A following step is to sustain or reject them. A series of analytic practices has been developed since the foundation book by Glaser and Strauss, originally published in 1967 [3].

Ground truth data sets can be generated in different ways. First one that seems to be the simplest but also requires the most effort in practice is the human labeled data. The opposite side of human labeled data is the automated labeled data generated using specific algorithms that helps to avoid the very time consuming manual effort in practice. However, this method is very complex because developing such algorithms might require some innovative approaches. Somewhere in between is the synthetic produced data sets that do not require manual effort, the algorithms for data generation are not necessarily as complex as the fully automated method but they may not produce good enough data for real-life scenarios. Synthetic generated data may seem good enough to the human eye but a machine learning model trained on such data sets may behave very differently in practice.

This paper presents improvements done to the algorithm described in [4] which generates data sets containing images labeled for eyes location. The modifications result in a smaller size of the generated data set but a improved quality of the labels. Detection is carried out on grayscale images obtained from an infrared sensor with a resolution of one megapixel. This sensor is specific for the automotive industry since it offers good and similar visibility during day and night. The resolution may differ depending on the sensor generation. Implementing such an algorithm has many advantages such as a huge time improvement in data set generation and the removal of human errors.

S. Valcan and M. Găianu from West Univesity of Timişoara, Computer Science Department, Romania and Continental Automotive Romania, VNI HMI Department, Timişoara, Romania (e-mail: sorin.valcan96@e-uvt.ro, mihail.gaianu@e-uvt.ro).

II. Methods

Creating a ground truth data set may include consideration of some major steps. The model design must be defined for the input shape and size and also for the output structure. Ground truth labels of the data set are very closely related with the output of the model. Training and testing sets must be generated in a correct way for every specific problem, with equal number of labels for each class in case of classifiers and different samples in testing set compared to training set. In case of eye detection labels it is important to separate subjects in training and testing sets because the learning is closely related to the eye shapes of every person.

Paper [4] presents an eye location selection algorithm on an infrared driver image with resolution 1280×800. The modifications presented here are related to some searching parameter adaptation and a different approach in the interpretation of the possible eye patches especially in pupil reflection computation and final patch selection. Some steps have also been removed.

The main idea of conversion to black-gray-white of the grayscale images and eye patches remains unchanged. We can achieve very good precision of eye location labels even if we reduce the grayscale details to only three possible colors.

A. Eye Search Area Traversing

The eye search area traversing process parameters have been adapted to allow a more precise eyes location area to be computed.

In the previous version the window that traversed the eye search area had a width of 5.8% of the picture width and a height of 8.7% of the picture height [4]. The new window size has a width of 8.75% of the picture width and a height of 5.46875% of the picture height. This modification results in all possible eye patches having a size of 70×70.

The adaptation that has a crucial impact for the labels precision is the modification of the stride with which the eye search area is traversed by the defined window. In the previous version the stride was 1.1% of the picture width (14 pixels) in x direction and 1.2% of the picture height (10 pixels) in y direction. Those values limited the precision of the resulted eye location because the jump between two closest possible eye patches was too big. The data generated with the previous stride were acceptable and usable for the neural network training but the jumps were too obvious and it affected the KPI computation based on the labeled ground truth data from this algorithm. The KPI used for our neural network is described in [5].

Because of all these reasons we reduced the stride to 0.234375% of the picture width in x direction (3 pixels) and 0.375% of the picture height in y direction (3 pixels). By doing this we obtained more stable eye location labels for similar frames, it improved the testing precision of the neural networks and it also improved the KPI results since we were facing the situation where analysis was computed between a good neural network output and an unstable ground truth generated label.

B. Eye Ratio Map Score

The eye ratio map score process has been completely removed from the possible eye patches processing. The reason behind this decision is the modification of the criteria for choosing between two overlapping eye patches.

In the previous version, if two possible eye patches were overlapping, the score defined in [4] Section 2.4.3 was used to choose which one is better. The eye ratio map score was the fourth parameter that made up the score. In the improved version from this paper, the removal of overlapping eye patches is done only using the pupil reflection detection.

C. Pupil Reflection Check

The pupil reflection check is the most important step in the entire algorithm that differentiates between a correct eye patch and another one which respects all previous steps of the algorithm.

Here we have done significant changes in order to improve the stability of the detected eye patches around the pupil reflection and also remove even more wrong possible eye patches that may pass the test in the previous version of the algorithm.

In the previous version the eye patch was processed in order to search for possible pupil reflections. Each detected white spot was checked to meet the conditions described in [4] Section 2.4.2. The first white spot that met the conditions was considered as a valid pupil reflection and the eye patch was accepted for the next selection step.

In the new version the following process applies:

- the white spots in the possible eye patch continue to be processed until the end of the data instead of stopping after the first one that met all conditions to be considered as a pupil reflection;
- for each white spot that met all conditions for a valid pupil reflection the distance to the eye patch center is computed. The minimum distance that was obtained for the current eye patch is saved in order to be used later;
- in case there are 0 or more than 3 valid pupil reflections the eye patch will be discarded from further processing. This check was added in order to discard possible patches that have nothing to do with a valid eye detection but because they are very noisy it may result in white spots that could pass all the previous checks.

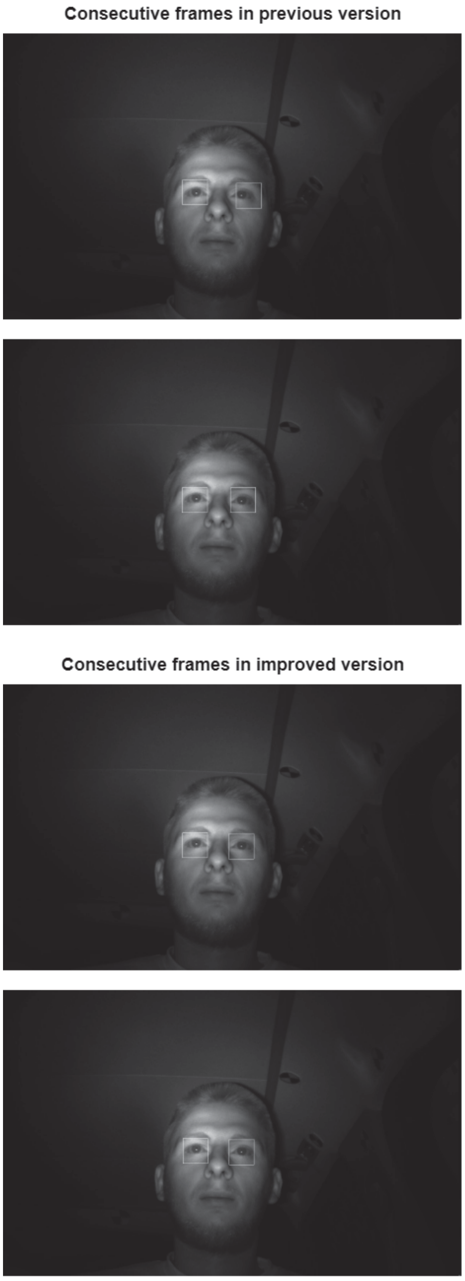
The process described above improved very much the detection on recordings where subjects wear glasses. Glasses reflection are very difficult to treat in the pupil reflection check. Another similar problem occurs sporadically in recordings with subjects having light hair color. Multiple hair areas can be converted to white in the conversion from grayscale to black-gray-white creating noisy patches with multiple white spots.

D. Removal of Overlapping Patches

This selection process which chooses the best option between two overlapping eye patches has remained similar as described in [4] Section 2.4.3. The major difference in the improved version of the algorithm is the score used to choose between two patches.

In the past we used the score as a sum between the black percentages rule from [4] Section 2.4 and the ratio map score which has been completely removed as described in II-B.

In the current improved version of the algorithm that score has been replaced by the minimum distance between a pupil reflection white spot and the center of the eye patch computed in II-C. This change helps for a better efficiency in selecting between overlapping eye patches since the current distance is a much better indicator of how good each is fitted on the eye. The stability of detection between consecutive frames has improved since the eye patches do not jump in similar situations. This case is exemplified in Fig. 1.



Consecutive frames in previous version

Consecutive frames in improved version

Fig. 1 Example of detection stability improvement for consecutive similar frames

E. Computing Possible Eye Pairs

The process of computing possible eye pairs from a list of valid eye patches obtained from the previous steps has remained unchanged from the previous version described in [4] Section 2.4.4. There have been adjustments made to the parameters that allow two possible eye patches to be considered as a valid eye pair.

The maxY value for the maximum distance in y direction between two possible eye patches has been reduced from 8.7% of the picture width (69.6 pixels) to the window height defined in II-A divided by 4 (17.5 pixels). The reduction is intended to have a stabilization effect on the selected patches in order to have smaller jumps and differences between the left and right eye.

In the previous version, the big value for maxY was intended to allow detections also for frames where head has a bigger roll rotation (tilt left and right) but it is not the case in our current dataset.

The minX value for minimum distance in x direction between two possible eye patches has been increased from 150% of the eye patch width to 180%. This increase has the main purpose to disallow selection of an eye pair with one patch being set correct on the eye and the second one being on the nose where a reflection from the glasses frame is selected. This problem has also been reduced by the new pupil reflection check process described in II-C.

Together these two improvements have increased significantly the quality of the generated data on recordings where the subjects wear glasses.

III. Obtained Results

The algorithm was run on 762 recordings with 113 subjects. These recordings have different scenarios where subjects are required to look at specific markers on the board, get out of the car and back in and also rotate their head in all directions. Data were recorded in different times of the day with different environmental light conditions depending also on the weather.

For some subjects a single set of recordings have been performed with their normal outfit while others have done multiple sets with additional glasses and/or surgical masks. Using this recordings database, there are 2.067.779 frames processed with the improved version of the algorithm presented in this paper. The output of ground truth data for eyes location was computed for 263.713 frames. This version generated more than 30% less frames with ground truth data compared to the previous one but the stability of the labels has improved significantly. The main problem of light reflection from the frame of the glasses has been solved thanks to the new tougher restrictions and parameter adaptation, especially for cases where one eye was detected correct but the other label was on the nose. There are still few frames where drivers with glasses rotate their head and two separate reflections from the frame of the glass generate wrong eye labels. This major improvement makes the small number of wrong labels to have a negligible effect on the neural network training process because the automatic generation of training data subsets described in [5] reduces the chances to have wrong labels selected.

There are no frames where the algorithm computed an eye location when the driver is not present at the steering wheel. Compared to the previous version of the algorithm where about 95% of the labels contained eyes inside the generated bounding box, using the improvements presented in this paper we increased that percentage to approximately 98%.

IV. Discussions

In the process of manual ground truth labels creation, one or more people label data through a process subsequently dim into invisibility. The term ”ground truth” implies that these labels represent a neutral description of reality. We tend to forget the series of human actions that led to their creation. We end up using the generated ”ground truth” labels containing an element of subjectivity as a base for testing the quality of a model, while the series of human decisions that produced them, becomes less and less visible [6], [7], [8], [9].

V. Future Work

The future work of this project is focusing on improvement of the face area selection algorithm, development of nostrils and mouth selection functionalities and training of neural networks using the generated datasets.

A. Face Area Selection Improvements

The current algorithm for face area selection presented in [4] Section 2.2 is currently good enough for the generation of eye labels and probably also for nostrils and mouth labels but it is not usable as training data for a neural network. It also generates useless bounding boxes for frames where the driver is not behind the steering wheel or when different movements appear. There are also recordings with very short or very tall subjects where the face area is not computed correctly at all, which results in zero eye labels.

It is not a main priority to have a perfect face area selection since it does not have a major impact on face feature ground truth data generation but there are improvements that can be developed.

B. Nostrils and Mouth Selection

Nostrils and mouth areas are face features we want to be able to detect in future. There are some initial implementations for generation of ground truth data but they are not at a very precise level.

Both areas should be based on the eye ground truth labels but we are still searching for the best solution.

The implementations will use the same principle of converting grayscale images to black-gray-white level and will compute one bounding box for each area.

C. Neural Networks Training

We are training multiple neural networks in order to search for the best architecture for eyes detection using the automatic system described in [5].

There have been significant improvements using the ground truth data generated with the algorithm presented in this paper and we will present our newest results in a future paper.

VI. Conclusions

This paper presented the improvements done for an automatic ground truth data generator for eye locations on infrared driver recordings. Using this modifications the quality and stability of the generated dataset has improved very much which helps the entire system used for training of the neural networks.

The purpose of generating ground truth data using an automatic algorithm has been achieved and is becoming more and more precise with every improvement that is implemented. This can help future projects from different fields to have a more effective approach to neural networks training when data is not directly available.

References

[1] https://blog.datumize.com/evolution-dark-data

[2] https://www.ibm.com/blogs/cloud-archive/2017/08/ibm-data-catalog-data-scientists-productivity/

[3] Barney G Glaser and Anselm L Strauss. 2017. Discovery of grounded theory:Strategies for qualitative research. Routledge.

[4] Valcan, S.; Gaianu, M. Ground Truth Data Generator for Eye Location on Infrared Driver Recordings. J. Imaging 2021, 7, 162. https://doi.org/10.3390/jimaging7090162.

[5] S. Valcan, ”Convolutional Neural Network Training System For Eye Location On Infrared Driver Recordings Using Automatically Generated Ground Truth Data,” 2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2021, pp. 222-226, doi: 10.1109/SYNASC54541.2021.00045.

[6] Ian Hampson and Anne Junor. 2005. Invisible work, invisible skills: interactivecustomer service as articulation work. New Technology, Work and Employment20, 2 (2005), 166–181.

[7] Kjeld Schmidt. 2002. Remarks on the complexity of cooperative work. Revued'intelligence articielle 16, 4-5 (2002), 443–483.

[8] Susan Leigh Star. 1999. The ethnography of infrastructure. American behavioral scientist 43, 3 (1999), 377–391.

[9] Susan Leigh Star and Karen Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. Information systems research 7, 1 (1996), 111–134.