Detecting Eye Contact From Body Pose With Machine Learning

Originally published at https://www.unite.ai/detecting-eye-contact-from-body-pose-with-machine-learning/

Researchers from France and Switzerland have developed a computer vision system that can estimate whether a person is looking directly at the 'ego' camera of an AI system based solely on the way the person is standing or moving.

The new framework uses very reductive information to make this assessment, in the form of semantic keypoints (see image below), rather than attempting primarily to analyze eye position in images of faces. This makes the resulting detection method very lightweight and agile, in comparison to more data-intensive object detection architectures, such as YOLO.

The new framework evaluates whether or not a person in the street is looking at the AI's capture sensor, based solely on the disposition of their body. Here, people highlighted in green are likely to be looking at the camera, while those in red are more likely to be looking away. Source: https://arxiv.org/pdf/2112.04212.pdf

Though the work is motivated by the development of better safety systems for autonomous vehicles, the authors of the new paper concede that it could have more general applications across other industries, observing *'even in smart cities, eye contact detection can be useful to better understand pedestrians' behaviors, e.g., identify where their attentions go or what public signs they are looking at'*.

To aid further development of this and subsequent systems, the researchers have compiled a new and comprehensive dataset called LOOK, which directly addresses the specific challenges of eye-contact detection in arbitrary scenarios such as street scenes perceived from the roving camera of a self-driving vehicle, or casual crowd scenes through which a robot may need to navigate and defer to the path of pedestrians.

Results from the framework, with 'lookers' identified in green.

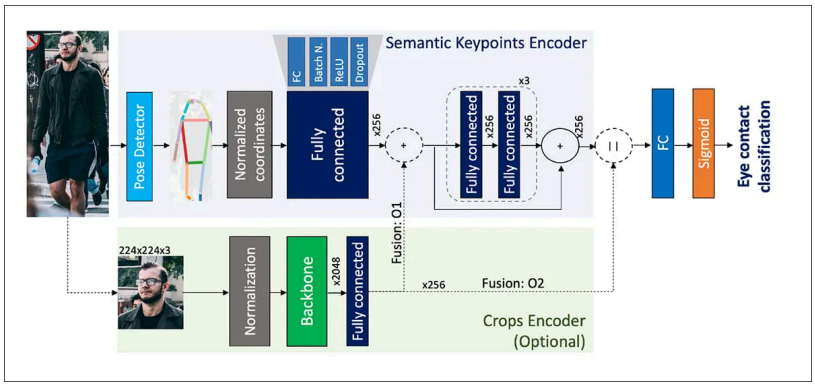
The research is titled *Do Pedestrians Pay Attention? Eye Contact Detection in the Wild*, and comes from four researchers at the Visual Intelligence for Transportation (VITA) research initiative in Switzerland, and one at Sorbonne Université.

Architecture

Most prior work in this field has been centered on driver attention, using machine learning to analyze the output of driver-facing cameras, and relying on a constant, fixed, and close view of the driver – a luxury that's unlikely to be available in the often low-resolution feeds of public TV cameras, where people may be too distant for a facial-analysis system to resolve their eye disposition, and where other occlusions (such as sunglasses) also get in the way.

More central to the project's stated aim, the outward-facing cameras in autonomous vehicles will not necessarily be in an optimal scenario either, making 'low-level' keypoint information ideal as the basis for a gaze-analysis framework. Autonomous vehicle systems need a highly responsive and lightning-fast way to understand if a pedestrian – who may step off the sidewalk into the path of the car – has seen the AV. In such a situation, latency could mean the difference between life and death.

The modular architecture developed by the researchers takes in a (usually) full-body image of a person from which 2D joints are extracted into a base, skeletal form.



The architecture of the new French/Swiss eye contact detection system.

The pose is normalized to remove information on the Y axis, to create a 'flat' representation of the pose that puts it into parity with the thousands of known poses learned by the algorithm (which have likewise been 'flattened'), and their associated binary flags/labels (i.e. *0: Not Looking* or *1:Looking*).

The pose is compared against the algorithm's internal knowledge of how well that posture corresponds to images of other pedestrians that have been identified as 'looking at camera' – annotations made using custom browser tools developed by the authors for the Amazon Mechanical Turk workers who participated in the development of the LOOK dataset.

Each image in LOOK was subject to scrutiny by four AMT workers, and only images where three out of four agreed on the outcome were included in the final collection.

Head crop information, the core of much previous work, is among the least reliable indicators of gaze in arbitrary urban scenarios, and is incorporated as an optional data stream in the architecture where the capture quality and coverage is sufficient to support a decision about whether the person is looking at the camera or not. In the case of very distant people, this is not going to be helpful data.

Data

The researchers derived LOOK from several prior datasets that are not by default suited to this task. The only two datasets which directly share the project's ambit are JAAD and PIE, and each have limitations.

JAAD is a 2017 offering from York University in Toronto, containing 390,000 labeled examples of pedestrians, including bounding boxes and behavior annotation. Of these, only 17,000 are labeled as *Looking at the driver* (i.e. the ego camera). The dataset features 346 30fps clips running at 5-10 seconds of on-board camera footage recorded in North America and Europe. JAAD has a high incident of repeats, and the total number of unique pedestrians is only 686.

The more recent (2019) PIE, from York University at Toronto, is similar to JAAD, in that it features on-board 30fps footage, this time derived from six hours' driving through downtown Toronto, which yields 700,000 annotated pedestrians and 1,842 unique pedestrians, only 180 of which are looking to camera.

Instead, the researchers for the new paper compiled the most apt data from three prior autonomous driving datasets: KITTI, JRDB, and NuScenes, respectively from the Karlsruhe Institute of Technology in Germany, Stanford and Monash University in Australia, and one-time MIT spin-off Nutonomy.

This curation resulted in a widely diverse set of captures from four cities – Boston, Singapore, Tübingen, and Palo Alto. With around 8000 labeled pedestrian perspectives, the authors contend that LOOK is the most diverse dataset for 'in the wild' eye contact detection.

Training and Results

Extraction, training and evaluation were all performed on a single NVIDIA GeForce GTX 1080ti with 11gb of VRAM, operating on an Intel Core i7-8700 CPU running at 3.20GHz.

The authors found that not only does their method improve on SOTA baselines by at least 5%, but also that the resulting models trained on JAAD generalize very well to unseen data, a scenario tested by cross-mixing a range of datasets.

Since the testing performed was complex, and had to make provision for crop-based models (while face isolation and cropping are not central to the new initiative's architecture), see the paper for detailed results.

	JAAD [25]				
Method / Box Height [px]	240+	160-240	110-160	0-110	All
Crops (ResNet-18)	80.4	79.5	80.4	73.4	78.1
Crops (ResNeXt-50)	79.0	81.2	83.0	75.3	79.7
Eyes Crops	74.4	79.1	81.3	74.7	77.4
Keypoints	87.9	88.7	**87.0**	78.7	85.9
Head Keypoints	**90.4**	**89.7**	86.7	76.5	**86.3**
Keypoints & Crops	80.5	82.2	83.6	75.9	80.6
Keypoints & Eyes Crops	85.3	86.4	85.0	**79.4**	83.9

Results for average precision (AP) as a percentage and function of bounding box height in pixels for testing across the JAAD dataset, with authors' results in bold.

The researchers have released their code publicly, with the dataset available here, and the source code at GitHub.

The authors conclude with hopes that their work will inspire further research endeavors in what they describe as an *'important but overlooked topic'*.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai