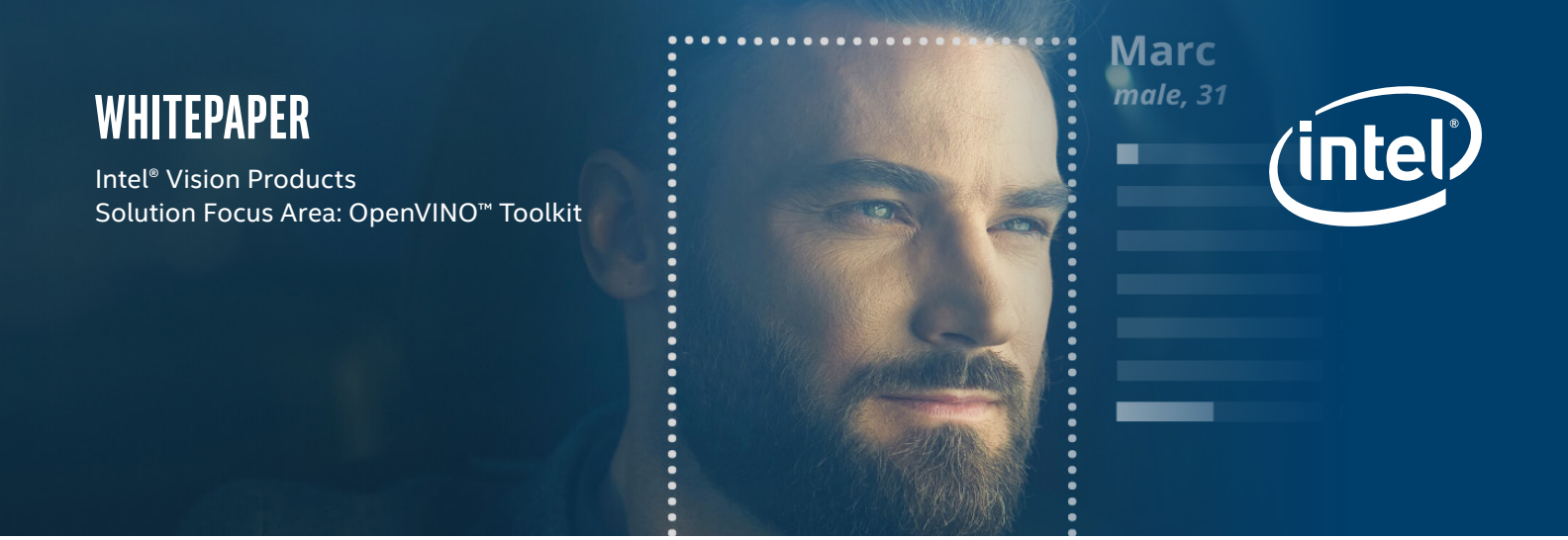


WHITEPAPER

Intel® Vision Products
Solution Focus Area: OpenVINO™ Toolkit

How Visage Technologies is Using OpenVINO™ Toolkit to Reach New Levels of Face Tracking Performance

Part of artificial intelligence (AI), face tracking applications have many important roles to play. Examples include driver drowsiness detection for improving road safety, and extending laptop and notebook battery life by monitoring user attention. In fact, a face tracking algorithm is the base component of any user awareness application and features in many use cases. It is therefore imperative to make this algorithm operate with little latency and as efficiently as possible.

Introduction

Visage Technologies is a computer vision company founded in 2002 in Linköping, Sweden. It is one of the leading providers of specialized face tracking and analysis solutions applied to various fields such as marketing research, biometrics, games and entertainment, marketing and sales, automotive safety, health, and assistive technologies. The company's primary product is the multiplatform software development kit visage|SDK that supports face and head tracking, gaze tracking, face recognition, and gender, age, and emotion estimation, along with advanced support for all major platforms and embedded systems.

Besides developing face tracking and analysis technology, the company has a separate division exclusively collaborating on a complex R&D project with a major system supplier for advanced driver-assistance systems and autonomous driving.

With more than 200 clients from 50 countries and a trend of strong growth, Visage Technologies has been listed among "Sweden Technologies Fast 50" companies by Deloitte since 2017. The company currently employs around 80 people with a substantial number of PhD researchers.

A key supplier of technology to Visage Technologies and other pioneering companies, Intel can deliver one of the most comprehensive arrays of intelligent vision capabilities to developers. The Intel® Vision Portfolio comprises silicon, software tools, deep learning frameworks, and libraries that are uniquely positioned for the next generation of big data image capture and AI. Intel Vision Products help put data to work, from the edge to the cloud for real-time operations, decision speed, and new solutions for a connected world.

Visage Technologies

The Intel® Distribution of OpenVINO™ (Open Visual Inference and Neural Network Optimization) toolkit is a free, downloadable toolkit within the Intel Vision Products portfolio that accelerates the development of high-performance computer vision and deep learning inference into vision applications. The integration of OpenVINO toolkit with visage|SDK provides an excellent solution for computationally efficient face tracking and analysis applications. It also makes optimization and deployment easier for edge devices, an important point for a driver drowsiness detection application.

In addition, 10th generation Intel® processors support Vector Neural Network Instructions (VNNI) which improve AI performance by combining three instructions into one, thereby maximizing the use of compute resources, utilizing the cache better, and avoiding potential bandwidth bottlenecks. Based on Intel® Advanced Vector Extensions 512 (Intel® AVX-512), VNNI speeds the delivery of inference results, accelerating performance of AI applications such as face tracking.

Use Case 1: Driver Drowsiness Detection

Figure 1. Optalert driver drowsiness detection solution.

Face Tracking

The face tracking package of the visage|SDK is used to track or detect faces and facial landmarks in images and videos. The following output is available from the package: 2D and 3D landmark coordinates, 3D head pose, action units, gaze vector, iris diameter, and so on. The core algorithms used in the face tracking package are face detection and face alignment.

In order to get information about a face in an image or a video, the face must first be located. The face detection algorithm uses machine learning to create a model that can detect and locate the bounding boxes of faces in the image. The located faces are then tracked using the face alignment algorithm on each frame until the face is lost. The face alignment algorithm is used to detect 2D landmark locations (usually called face shape) that delineate prominent facial features such as eyes, nose, mouth, and so on. It uses the output of the face detection algorithm in the form of 2D face position and size. Machine learning is used again to create a model that can produce the 2D landmark locations from the image crop containing a face. Since this model is used in each frame, it is important to optimize its inference time in order to obtain real-time performance and low power consumption. This is especially important for applications requiring low latency, like the ones described below.

Optalert, one of the leading companies for drowsiness detection, uses this technology in its offering for driver drowsiness and attentiveness detection in the automotive industry. The company has several patents for early detection of drowsiness from a person's eyelid movements. For commercial products, they measure the eyelid movement with specialized hardware in the form of glasses with an integrated IR sensor. For their automotive offering, Optalert has developed an algorithm that detects drowsiness and attentiveness without specialized hardware by using cameras integrated into the cabin. As described in the patents, the minimum information input rate is 60 frames per second (FPS or Hz) for the solution to work accurately. On top of that, robustness of the face tracker in this specific use case is essential, as driver drowsiness must be detected during both day and night, and wide head angles and rotations must also be supported. To achieve suitable robustness, several different convolutional neural networks (CNNs) had to be introduced into the face tracker, significantly increasing the processing needs. With Intel® hardware and OpenVINO toolkit, all the requirements were met, and the solution is now operational.

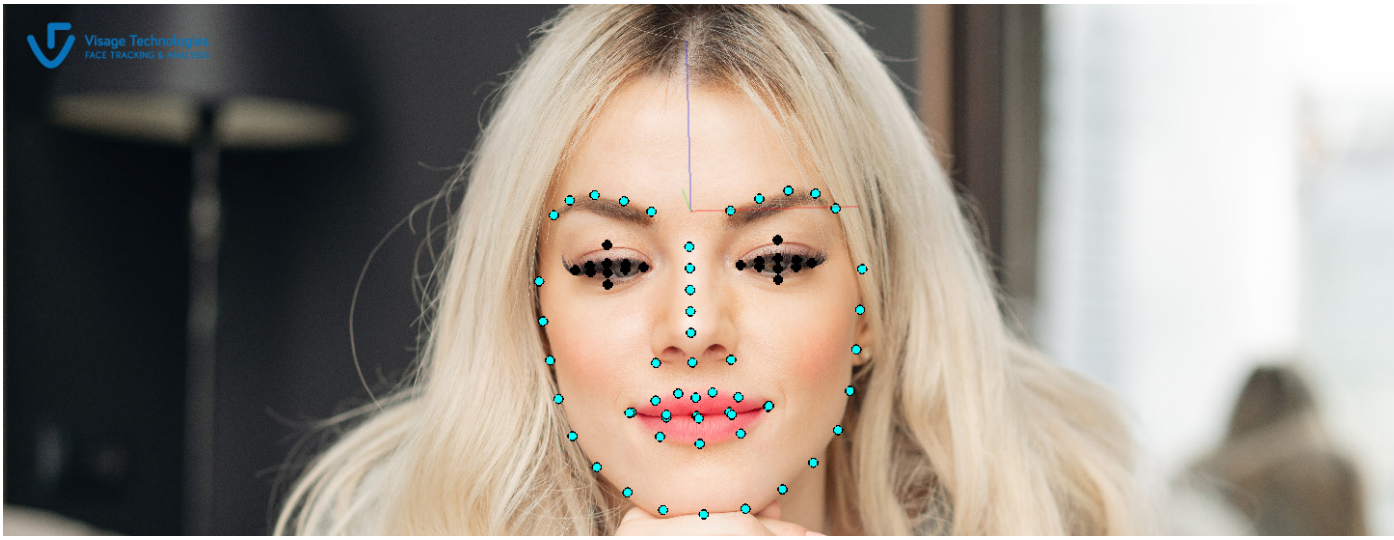


Figure 2. Face tracking with 2D landmarks and head pose axis displayed.

Use Case 2: Laptop User Awareness

As already announced by Project Athena from Intel, modern laptops and notebooks are expected to have a long battery life (nine hours) with normal usage[1]. One of the most power-hungry components of the laptop is its screen. Face tracking and detection can be used to detect whether a user is present in front of the screen and automatically dim the screen to conserve battery life. Since many laptops will employ multiple screens, face and head pose tracking can be used to determine which screen the user is focused on in order to dim the currently unused screens. For this application to be power-efficient, the face tracking solution needs to be as fast as possible. By employing the OpenVINO toolkit inference engine, the visage|SDK face tracking solution achieves extremely fast run-time performance, making it a viable solution for such user awareness applications.

Methodology

When Visage Technologies realized that its in-house inference engine was becoming too slow, the company switched its focus to finding an appropriate substitute. Since different platforms have different optimizations, Visage Technologies was aware that one solution would not cover all its run-time needs. The company decided that it would support multiple inference engines by creating a wrapper around them, allowing it to have a single, unified application programming interface (API).

OpenVINO toolkit was selected for desktop and laptop platforms due to Intel's extensive documentation and support. Intel engineers worked with Visage Technologies developers over a period of 18 months to help educate them on OpenVINO toolkit and assist with integration. Intel also provided a laptop with a 10th generation Intel® Core™ processor for tests.

For everything to work with OpenVINO toolkit, specific prerequisites had to be met. Visage Technologies' models needed to be converted into the OpenVINO toolkit format. The OpenVINO toolkit inference engine had to be wrapped under a consistent API before integration into the SDK to allow convenient inference engine switching without modifying the SDK code. Finally, in order to unlock the full optimization capabilities of OpenVINO toolkit, the models needed to be calibrated and quantized to INT8 (8-bit integer) for computing performance improvements. Using the INT8 data type instead of a floating-point data format like FP32 can boost performance in deep learning inference with negligible loss of accuracy, in most cases ~<1% accuracy loss. INT8 is also needed to take advantage of VNNI. Each step is described in more detail in the following paragraphs.

Visage Technologies' R&D team uses PyTorch for model prototyping and training because of the considerable flexibility and convenience provided by this machine learning framework. As PyTorch is supported by the Open Neural Network Exchange format (ONNX), Visage was able to utilize it as an intermediate format before converting to an OpenVINO toolkit format. Some minor problems in this step, such as unsupported layers, were easily solved, either by a slight modification of the layer (for example, view layer with inferred batch size), or by adding support for layers with OpenVINO toolkit development.

For efficient INT8 computations, the models needed to be calibrated using the official Python tools from the OpenVINO toolkit repository. Since Visage Technologies models for the face alignment algorithm solve a regression problem (estimating numerical values by minimizing mean squared error loss) with the company's proprietary data set and annotations, their developers needed to implement adapters and annotation converters for the validation data set used for calibration. This was done using available documentation and the source code of the official Python pipeline for calibration with OpenVINO toolkit.

[1] https://www.intel.com/content/www/us/en/products/docs/devices-systems/laptops/laptop-innovation-program.html

One of the biggest challenges concerning quantization came from the fact that the outputs of one model (deep features) are the inputs for the other models. The calibration tool optimizes the quantized model performance iteratively with all the weights quantized at the beginning of the process. Afterwards, the model performance is evaluated using a relative metric, which is used to measure the drop in performance due to quantization with respect to the original model. The tool then iteratively removes quantized layers until a threshold on the relative metric is met.

The original model used by Visage Technologies was already as precise as possible, since its output was the ground truth that the quantized model was trying to match. This was a problem since these outputs gave no room for achieving the targeted results (defined by a desirable threshold), resulting in an infinite loop of optimization iterations. Therefore, the calibration step had to be stopped immediately after the first run, which was a departure from the way the calibration tool normally works. Fortunately, the fully quantized model from the first iteration resulted with no significant drops in accuracy confirmed by measurements on annotated test set outside of the calibration tool.

Results from the Tests

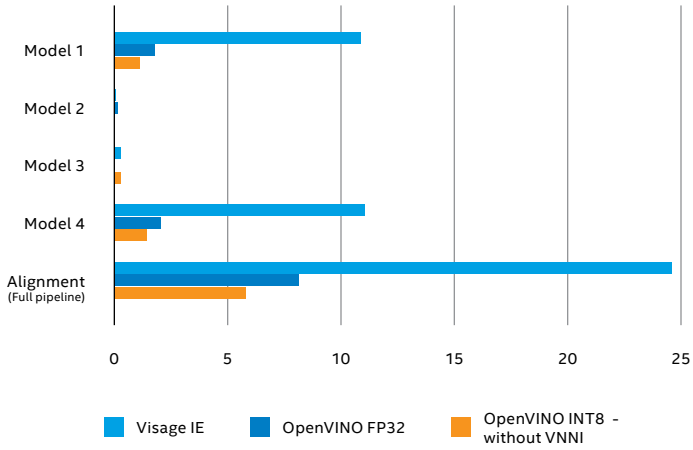
The face alignment algorithm is divided into stages, which solve the problem in a coarse-to-fine manner. This means that the algorithm uses multiple models across stages, sharing deep features for different purposes. Each model can be named and briefly described as follows to better understand the measurements:

- M1 – First stage deep convolutional network model
- M2 – First stage fully connected network model sharing M1 features
- M3 – First stage fully connected network model sharing M1 features (separate network and different functionality compared to M2)
- M4 – Second stage deep convolutional network model with a fully connected layer

The tests were run on two laptops, one being powered by an Intel® Core™ i7-8750H processor and the other by a 10th generation Intel® Core™ processor. The Visage Technologies team compared performances using OpenVINO toolkit FP32 (32-bit floating points) and INT8 models with its in-house developed inference engine based on OpenBLAS. Google's open-source benchmark project was used to measure inference times of each model individually and the full alignment algorithm using all models. The measurements were performed using a single thread.

The following results were obtained on the Intel® Core™ i7-8750H processor (without VNNI capability).

Inference Time in ms (the lower the figure, the better)



Inference Time in ms	Visage Technologies IE	FP32 (Optimized by OpenVINO)	OpenVINO Toolkit INT8
Model 1	10.93	1.76	1.18
Model 2	0.04	0.09	0.1
Model 3	0.15	0.1	0.15
Model 4	11.38	2.69	1.29
Alignment (full pipeline)	24.6	8.28	5.95

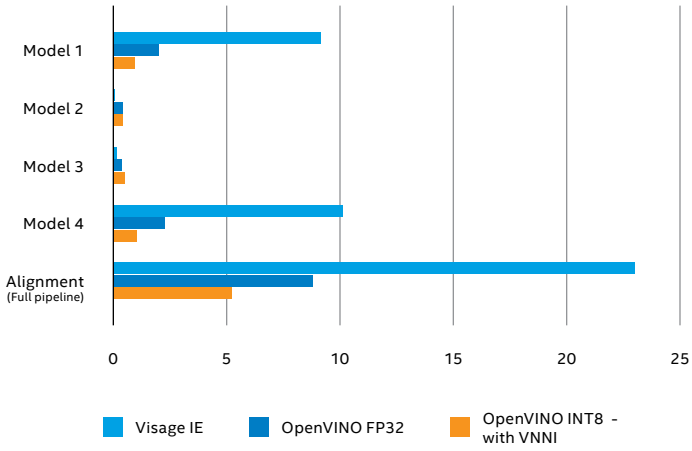
As can be seen from the results, the most significant improvements are obtained for deep convolutional networks (M1, M4). This is expected, since these are the computationally heaviest parts of the algorithm. For small, fully connected networks, there are practically no gains, since the computational load is already quite low.

By comparison, the following results were obtained on the 10th generation Intel Core processor (with instruction set architecture (ISA) VNNI capability).

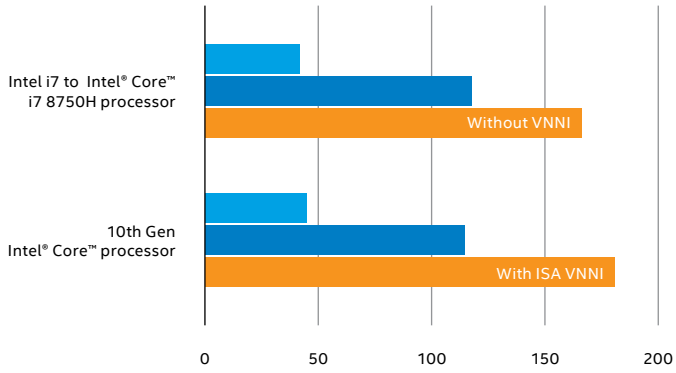
The 10th generation Intel Core processor outperforms the previous generation CPU (Intel Core i7-8750H processor) using INT8 models. Part of this performance improvement is due to the support of VNNI by the 10th generation Intel Core processor.

From another perspective, a comparison of performance in FPS is shown in the following chart.

Inference Time in ms (the lower the figure, the better)



FPS results (the higher the figure, the better)



Inference Time in ms	Visage Technologies IE	OpenVINO™ Toolkit FP32	OpenVINO™ Toolkit INT8
Model 1	9.3	1.92	0.81
Model 2	0.03	0.24	0.25
Model 3	0.12	0.22	0.27
Model 4	10.37	2.63	0.93
Alignment (full pipeline)	23.1	8.66	5.52

Conclusion

The results show that the use of a highly optimized inference engine such as OpenVINO toolkit as an enhancement to the existing solution greatly improves the inference speed of the face tracking algorithm when compared to an implementation based on OpenBLAS. More specifically, it provides approximately four times faster performance on both tested platforms[2].

Additionally, the Intel Core i7-1065G7 processor supports the new specialized VNNI, which provides even greater efficiency. This enables power-efficient solutions such as driver drowsiness detection and laptop user awareness applications where energy consumption is critical.

Overall, the combination of the OpenVINO toolkit inference engine, support for VNNI, and the latest generation of Intel processors allows the Visage Technologies face tracking algorithm to perform above 180 Hz. This greatly surpasses the target frame rate for Optalert, allowing the Visage Technologies' client to exploit the full potential of its patented driver drowsiness detection solution. It also demonstrates that laptop user awareness applications based on Visage Technologies' technology are compatible with long laptop battery life expectations, such as those specified in the next-generation Project Athena program from Intel.

Authors

Ivan Gogić
Director R&D,
Visage Technologies AB

Nikola Mrzljak
R&D Engineer,
Visage Technologies AB

Fredrick Odhiambo
Application Engineer, Intel

Jonas Kollberg
Global Account Manager, Intel

[2] See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization