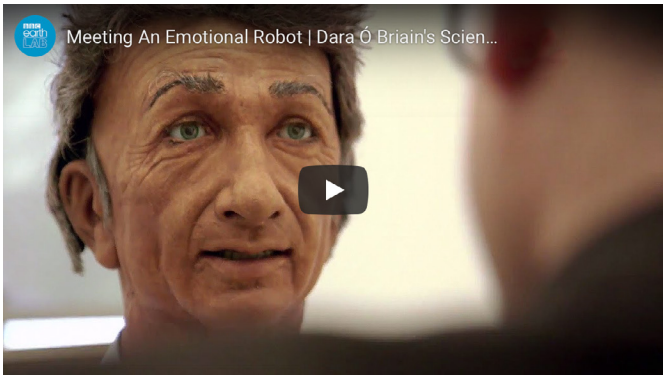
insight.tech POWERING IoT AND EMBEDDED INNOVATION

Is Facial Recognition AI's First "Killer App"?

Every new paradigm-shifting technology needs a killer app. For artificial intelligence, the first one may be facial recognition.

Security and access control—as we see today in many airports—are primary applications of deep learning-based facial recognition. But more advanced use cases are already on the horizon. Spod, for example, is a robot shopping assistant prototype that uses facial recognition to identify a customer's age and gender to make product recommendations.

Even more sophisticated are AI-enabled robots like Cambridge University's "Charles," an intelligent robot prototype. Designed by students, "he" can read and interpret human facial expressions as shown in the *Meeting an Emotional Robot* video below.



"Charles" is just one of a number of prototypes capable of detecting human feelings and emotions, enabling numerous applications. These include personal companionship, health diagnosis, emergency intervention, and so on. But to be effective and believable in these tasks, intelligent robots need natural human response times.

These real-time reactions are possible only by executing AI technology directly at the edge. One company, CyberLink, demonstrated significant leadership in this area as part of the University of Washington facial recognition algorithm competition, the MegaFace Challenge.

High-Performance Facial Recognition for the Edge

The MegaFace Challenge is a series of benchmark tests that invites facial recognition developers to test their algorithms against a large training data set.

Images of celebrities and people with extreme age differences are sprinkled in among anywhere from 10 to 1,000,000 "distractors," which are images of other faces and even some photos that the AI training set as incorrectly identified as faces.

The competition was tough, as algorithms from companies like Google, Tencent, and others achieved identification accuracies in the high 90 percentiles on smaller data sets. But they performed progressively worse as the number of distractors increased, as shown in **Figure 1**.

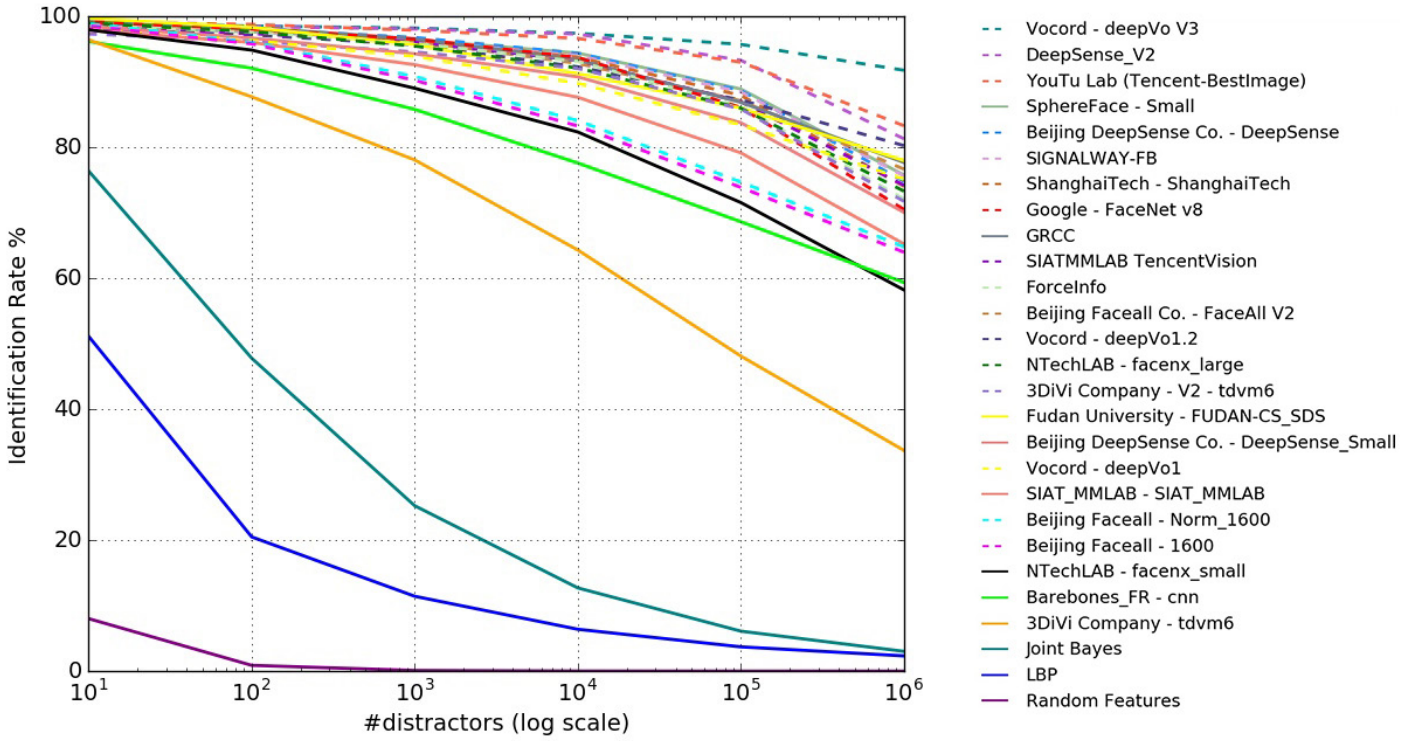


Figure 1. The MegaFace Challenge benchmarks measure the accuracy of facial recognition algorithms against large data sets. (Source: University of Washington)

Exceeding Google's FaceNet v8 and SIATMMLAB TencentVision, in performance was CyberLink's FaceMe AI facial recognition engine, based on the SphereFace algorithm. It registered better than 98.4 percent accuracy while running completely at the edge.

Designed for facial recognition engineers, FaceMe is a cross-platform Software Development Kit (SDK), capable of analyzing attributes such as age, gender, and emotion. Its neural network has been pre-trained on large image databases and is compatible with frameworks such as TensorFlow and Core ML—enabling developers to integrate their own training data.

After being trained, the resulting inference engine is combined with pre-processing algorithms that compress it to as small as 4 MB. These algorithms also make the SDK compatible with edge systems ranging from mobile devices to Linux-based digital signage solutions to Windows PCs. This level of hardware and software compatibility means that FaceMe is applicable in use cases such as smart city surveillance, intelligent retail marketing, and personalized smart home robots **(Figure 2)**.

"Right now most facial recognition solutions, like those from AWS, Google, and Microsoft, are running in the cloud," said Steven Lien, Marketing Director at CyberLink. "Cloud-based facial recognition isn't ideal in certain scenarios, such as a door security system, for example. When you walk in front of the system, you would have to wait 20 or 30 seconds for the image or video to be uploaded to the cloud and to get a response back on the local device."

"FaceMe is optimized for edge devices. It could run on a personal computer, mobile devices like Windows, iOS, or Android phones, and so on. It is also optimized for different hardware, like CPUs and GPUs," Lien continued. "There are a lot of different parameters you can adjust inside the FaceMe environment for different deployments, such as the size of face capture (in pixels), frame rates, and false acceptance rates."

Figure 2. CyberLink's FaceMe facial recognition engine allows developers to interpret age, gender, and emotion. (Source: CyberLink)

As Lien pointed out, some facial recognition systems will require faster response times, higher accuracy, or the ability to capture more or fewer facial images at once. FaceMe allows developers to tailor their algorithms to use less computing power (in the case of a mobile application), reduce latency (a smart door lock), or capture many images at once (smart city surveillance).

Faster Recognition on More Platforms

To make inferencing engines compatible with different edge processors and operating systems, FaceMe SDKs rely on enabling technologies like the Intel® Distribution of OpenVINO™ Toolkit **(Figure 3)**. The OpenVINO Toolkit is a development tool that optimizes computer vision inferencing algorithms for use on CPUs, GPUs, FPGAs, and machine learning accelerators like Intel® Movidius™ vision processing units (VPUs).

"There are a lot of different tools for neural network processing. And currently there are many different inference engines, so FaceMe provides converters between each of them," Lien said. "With so many different components, the Intel OpenVINO Toolkit is a good platform for aggregating them into one complete SDK."

When enabling the OpenVINO Toolkit for FaceMe-based algorithms, facial recognition can be accelerated by as much as 500 percent. Meanwhile, FaceMe retains its accuracy, with gender detection accuracy of 98 percent, emotion detection of up to 86 percent, and just a mean average error of 5.8 years when analyzing age.

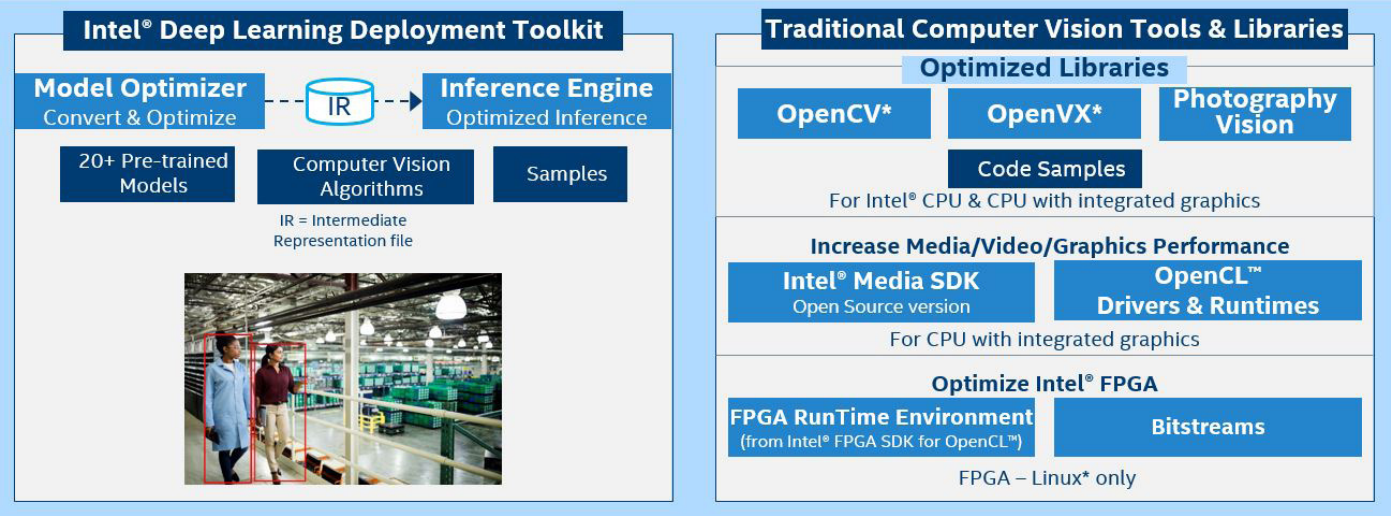


Figure 3. The Intel® OpenVINO™ Toolkit helps optimize the FaceMe facial recognition engine for a variety of neural network processors and accelerators. (Source: Intel® Corp.)

AI's Killer App is Just Around the Corner

The ability to scale performance and power utilization across so many different platforms means that facial recognition could be more pervasive, sooner. The FaceMe SDK is also easily portable, meaning developers can reuse application components across their development environments to accelerate deployment.

With access to facial recognition technology that rivals the accuracy of technology giants, AI engineers can now provide commercially viable facial recognition directly at the edge. In time, this will enable more precision retail marketing, better surveillance and access control, and a generation of "Charles"-like robots that provide believable human interactions.