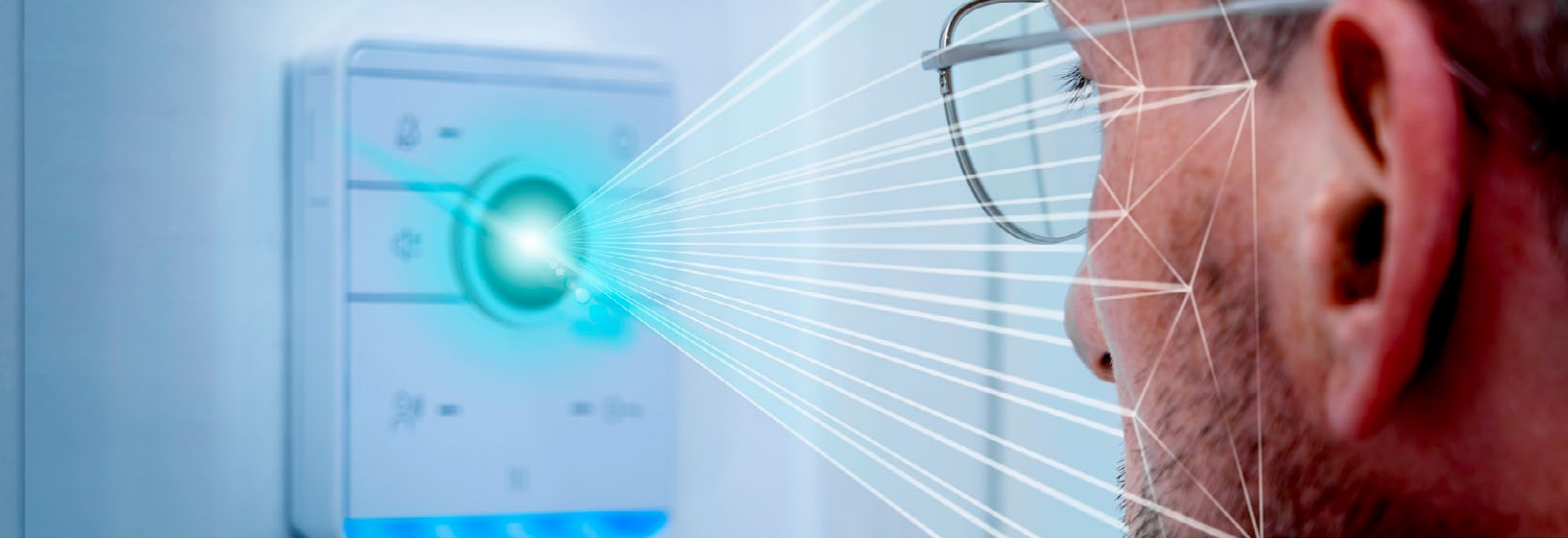
WHITEPAPER

Facial recognition

for access control and time & attendance systems



Keys and access codes are occasionally forgotten - facial recognition does not have this problem. (Image: istock.com/IGphotography)

Automatic Identification with Facial Recognition

TQ-Embedded has conducted a study on biometric face recognition on embedded hardware. Different deep neural network architectures were trained with a synthetically generated data set. The trained models were implemented on the TQMa93xxLA module from TQ-Embedded, which is accelerated by the integrated Ethos U65 NPU. The performance achieved with the TQMa93xxLA modules yields better AI inference results when compared to more powerful modules on paper. Finally, presentation attack detection options were evaluated to protect facial recognition against spoofing attempts.

Biometric facial recognition can be used in time & attendance and access control systems in two ways: First, as a secondary security factor to prevent, or at least make more difficult, the misuse of lost or stolen access cards, thus further increasing security. The second category of use cases is to use facial recognition for authentication instead of an access card for less security-critical applications. Examples include time and attendance, access control to less critical areas, or the use of elevators and machines that can only be operated or configured by authorized personnel.

The Privacy Problem

To train a facial recognition network, you need images of the faces of as many people as possible. In the past, the most common way to obtain large amounts of data as quickly as possible was to use images uploaded to social media. In 2018, however, the European General Data Protection Regulation (GDPR) became EU-wide law and has been mandatory ever since. It prohibits the processing of EU citizens' personal data without their prior explicit consent, which includes the previously used facial data, making its lawful use for training facial recognition algorithms virtually impossible.

In light of this legislation, the public datasets used for training are limited to Microsoft's DigiFace-1M database, which is synthetically generated using computer graphics. This is currently the most effective synthetic face recognition dataset publicly available for research purposes. However, it is only suitable for pre-training the model; real data is still used for ffine-tuning. Fortunately, the amount of real face images required is comparatively small.

The LFW (Labeled Faces in the Wild) benchmark protocol is used for validation during training. The YouTube Faces DB (YTF) test protocol is also used to test and compare the trained models. Accuracy is calculated from the sum of the true-positive (TPR) and true-negative (TNR) components at the optimal decision threshold. In access control and time and attendance applications, a low false positive rate (FPR) is of particular interest to prevent unauthorized access. Therefore, detection rates at FPRs of 0.1% and 0.01% are also evaluated.

Fine-tuning the pre-trained models

To fine-tune the models, a small database was created from the TQ internal media archive. The contents of the archive can be used commercially. The DigiFace-1M data set is intended for research purposes only. Since it was created using only techniques commonly used in the VFX (visual effects) and computer game industries, it can be assumed that obtaining comparable data for commercial use will not be a major problem.

The synthetically generated faces differ in appearance from real faces. Since the trained models have only seen synthetic data so far, they will now be optimized for real data through fine-tuning. The relevant information learned during pre-training should not be lost. The authors of DigiFace-1M use different numbers of real identities to determine how many are needed for good results. The number of images per person is 20, the lowest number of identities tested is 200, with the LFW test protocol achieving an accuracy of about 97%. They recommend to reduce the learning rate of the network by a factor of 100 compared to the pre-training, and the learning rate of the classification layer by a factor of 10, so that the network does not forget what it has learned before.

The TQ database was searched for faces in the available images using the "BlazeFace" detector and then pre-grouped into identities using one of the DigiFace-1M pre-trained "ResNet50" (residual network with 50 layers). Finally, all images were manually inspected and incorrect matches were corrected. Images without a match were also compared to the three most likely matches and then manually matched. In the end, only the identities consisting of at least two images were retained. This resulted in a dataset with 207 classes and 1151 samples. This is about 3.6 times smaller than the smallest amount of real data tested in the template.

To make the most of the limited data available, we are looking for ways to improve the fine tuning. The goal is to allow more training before the network is damaged by memorizing the limited data, which is known as overfitting. In addition to established regularization methods such as data augmentation, L2 regularization, and dropout, this project tries to prevent the loss of certain information from pre-training.

The great advantage of synthetic data is that the distribution of gender, color, ethnicity, and age of the identities represented in the dataset is controllable and can be easily adjusted to reflect the actual distribution in the world population. In contrast, the distribution of the data available for fine-tuning is typically suboptimal.

During training, the identities in the data set are evenly distributed over the hypersphere created by the embedding. This maximizes the distance between the class centers and the reliability of the recognition. The even and fair distribution of the population groups trained on the synthetic data is therefore also stored in the embedding generated on the model output. Optical features, on the other hand, where real and synthetic data differ, tend to be in the image layer and are typically processed in the front layers of the network, close to the model input.

Instead of training the whole network with the same learning rate, an exponentially decreasing learning rate is implemented from the input to the output layer of the network. The hope is to train the network to recognize the optical properties of the real data faster and the inequitable distribution of identities slower.

And finally, the deployment on TQMa93xxLA

Before the trained models can be deployed on the target, they must first be converted to the required format. Typically, for inference with NPU architectures, the network parameters must be quantized from 32-bit floats to 8-bit integers. As described in the i.MX Machine Leaning User's Guide, the trained network must also be converted to Tensorflow Lite format and then compiled for the i.MX 93 NPU using software provided by NXP.

Results

Table 1 below shows the results with the different network architectures pre-trained with DigiFace-1M only. The protocols LFW (Labeled Faces in the Wild) and YouTube Faces DB (YTF) are used to test and compare the models.

Version	LFW (acc. / FPR<0.1 % / FPR<0.01 %)	YTF (acc. / FPR<0.1 % / FPR<0.01 %)
MobileNetV3 Large	90.45 % / 52.53 % / 43.90 %	84.48 % / 45.52 % / 38.40 %
EfficientNet-lite0	90.97 % / 48.60 % / 34.33 %	85.14 % / 33.48 % / 25.84 %
EfficientNet-lite0 (112x112 res.)	91.82 % / 53.00 % / 49.27 %	86.06 % / 39.32 % / 34.52 %
EfficientNet-lite1	91.73 % / 46.53 % / 40.70 %	85.90 % / 45.52 % / 41.52 %
EfficientNet-lite2	91.87 % / 48.06 % / 33.50 %	85.26 % / 41.08 % / 35.28 %
EfficientNet-lite3	91.97 % / 46.60 % / 39.36 %	85.72 % / 45.60 % / 39.84 %
EfficientNet-lite4	92.63 % / 48.73 % / 29.00 %	86.00 % / 42.28 % / 35.80 %
ResNet50	93.50 % / 57.27 % / 38.56 %	88.16 % / 47.20 % / 41.04 %

Table 1

With ResNet50, a result of 93.50% is achieved with LFW, which is very close to the value of 94.55% achieved with the same network in the DigiFace-1M publication. It can be assumed that the slightly lower value is also due to the slightly lower input resolution used here. The EfficientNet-lite0 model, tested with the higher resolution, achieves about 0.9 percentage points more with both protocols; the success rate for FPR<0.01% in particular increases due to the higher resolution.
Among the more efficient models, MobileNetV3 Large and EfficientNet-lite1 in particular achieve comparatively high detection rates at FPRs of <0.1% and <0.01%.

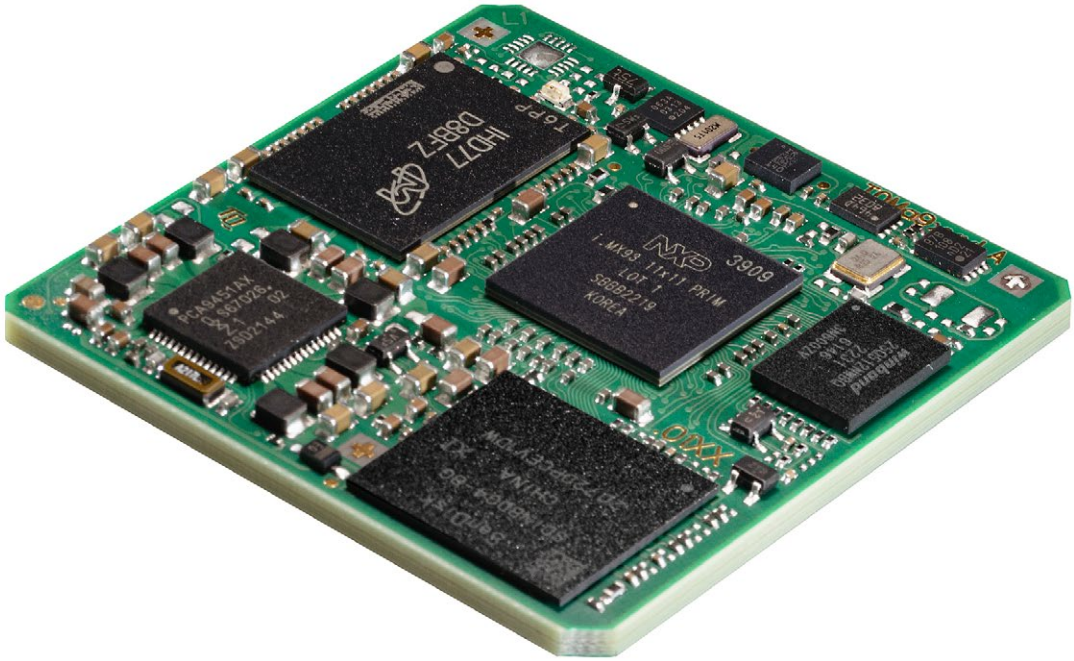
After fine tuning

As can be seen in Table 2, the difference between the results of the smaller models and those of the much larger ResNet50 has been further reduced by fine-tuning. The advantage of EfficientNet-lite0, which was trained with a higher resolution, has also largely disappeared.

However, it is noticeable that the values for FPR<0.01 with YTF have collapsed in almost all cases after fine tuning compared to the pre-trained versions in Table 1.

Version	LFW (acc. / FPR<0.1 % / FPR<0.01 %)	YTF (acc. / FPR<0.1 % / FPR<0.01 %)
MobileNetV3 Large	93.33 % / 59.30 % / 52.03 %	87.42 % / 41.80 % / 31.20 %
EfficientNet-lite0	93.60 % / 58.73 % / 52.36 %	87.84 % / 46.84 % / 33.92 %
EfficientNet-lite0 (112x112 res.)	94.40 % / 63.23 % / 41.63 %	87.82 % / 39.84 % / 27.28 %
EfficientNet-lite1	94.00 % / 60.43 % / 50.60 %	88.24 % / 49.92 / 24.44 %
EfficientNet-lite2	94.16 % / 58.73 % / 49.00 %	87.72 % / 45.08 % / 21.28 %
EfficientNet-lite3	94.41 % / 63.16 % / 45.80 %	87.88 % / 47.40 % / 23.64 %
EfficientNet-lite4	94.73 % / 69.50 % / 38.43 %	88.08 % / 45.32 % / 28.52 %
ResNet50	95.53 % / 64.17 % / 52.73 %	89.72 % / 52.36 % / 41.96 %

Table 2



Solder-in modules such as the TQMa93xxLA open up new application possibilities thanks to their compact dimensions. (Image: TQ-Systems GmbH)

A superficial look at the problem reveals that the data contained in YTF is often of particularly poor image quality. In some cases, this causes BlazeFace to be unable to align (determine the section of the image), making classification even more difficult for the model. In addition, there are often strong differences in perspective, facial expression, hairstyle and accessories, lighting, partial overlap and, in some cases, age to overcome for correct classification. LFW also tests models for these typical hurdles, but to a much lesser extent.

In the DigiFace-1M data set, very strong variations have been artificially incorporated in this respect, while the data from the TQ Media Archive contains almost nothing of the kind, as here all pictures were taken on the same day under similar conditions for most people. Therefore, it stands to reason that some of the robustness against these variations has been lost as a result of the fine-tuning. This is particularly evident in the YTF results, which now have virtually zero tolerance for false positives.

The results achieved here are far below those of large datasets with real identities. For example, "ArcFace" achieves 99.83% with LFW and 98.02% with YTF. "FaceNet" also achieves 99.63% with LFW and 95.10% with YTF. Nevertheless, the network trained as described was used in a trade show demo for the TQMa93xxLA.

EfficientNet-lite0 was used for the demo, and the decision threshold was set to approximately the same value at which an FPR of less than 0.01% is possible according to LFW. Although about half of all authorized access attempts should fail with this decision threshold, the system worked very reliably both in testing and at the show.

The reason for this is likely the significantly reduced need for robustness against different perspectives, facial expressions and partial overlapping. For access control systems and the demo setup used, only one type of camera is used, where people always stand at similar distances in front of the camera, and in most cases have neutral facial expressions.

Performance of i.MX 8M Plus and i.MX 93 NPU

According to NXP, the NPU of the i.MX 8M Plus delivers up to 2.3 TOPS of performance, while the Ethos-U 65 NPU of the i.MX 93, which is in a lower performance class, should only be able to deliver 0.5 to 1 TOPS. Accordingly, the inference time of the models is expected to be approximately double that of its predecessor. Testing was performed using a pre-compiled benchmark program that is included in the software components provided by NXP that are required to operate the NPU. The latency of 100 inferences was measured and averaged for reporting.

Version	i.MX 8MP	i.MX 93
BlazeFace (Front, 128x128)	2.18 ms	2.01 ms
MobileNetV3 Large	3.95 ms	1.68 ms
EfficientNet-lite0	2.46 ms	1.72 ms
EfficientNet-lite0 (112x112 res.)	2.88 ms	1.95 ms
EfficientNet-lite0 (256x256 res.)	10.44 ms	6.13 ms
EfficientNet-lite1	3.06 ms	2.04 ms
EfficientNet-lite2	3.23 ms	2.24 ms
EfficientNet-lite3	4.02 ms	2.72 ms
EfficientNet-lite4	5.09 ms	3.73 ms
ResNet50	8.38 ms	3.73 ms

Table 3

It turns out that, contrary to expectations, the i.MX 93 is actually faster than its predecessor in most cases. The EfficientNet-lite models are about 40 percent faster, while ResNet50 is only 20 percent slower. MobileNetV3 Large does not run very fast on the i.MX 8M Plus, as the version contains some operations that are not compatible with the i.MX 8M Plus NPU and therefore have to be computed on the CPU. The i.MX 93 is more than twice as fast in this setup.

Detecting deception

In addition to correctly recognizing people in the context of the task at hand, biometric systems must also be able to defend against deception. Presentation attack detection (PAD) is particularly important: Presentation attacks or spoofing attacks are attacks on access control systems by presenting false biometric data. As camera-based facial recognition systems only work with two-dimensional image data, they are particularly vulnerable to such attacks. In most cases, presenting an image on the screen of a mobile device or printed on a piece of paper is sufficient to impersonate another person.

In its biometric requirements, the FIDO Alliance defines three different types of attacks (Level A, B and C), which are categorized according to the time, expertise and access required to the source of the biometric data. Below are examples of facial recognition attacks that FIDO has identified for each level of attack.

	Source of biometrics	Difficulty	Type of attack
Level A (simple)	Social media photos	**Time:** < 1 day **Expertise:** Layman **Equipment:** Standard	Image of a face printed on paper / displayed on mobile device
Level B (moderate)	High-resolution photo, video of target	**Time:** < 7 days **Expertise:** Practiced **Equipment:** Standard + special	Paper mask, moving image of face played on screen
Level C (difficult)	High-resolution photo, 3D information of the target's face	**Time:** > 7 days **Expertise:** Expert **Equipment:** Special + customized	Silicone mask, theater mask

Table 4

There are several approaches to protect a system against such attacks. One option is to use additional sensors to collect additional information, such as depth information, to make the system more robust against attacks. An example of this is the FaceID built into Apple's iPhones, which uses TrueDepth infrared sensor system to scan the user's face in three dimensions[1]. Other options include infrared, thermal, light-field, multispectral, and stereo cameras. However, the use of additional sensors is often associated with very high development and material costs and is not feasible in all designs. They are also rarely an option for improving existing systems. In addition, advances in 3D printing are increasingly threatening the security of systems equipped with depth sensors or 3D cameras.

There are a number of ways to make purely camera-based face recognition systems more robust against all three types of FIDO attacks. A distinction is made between static and dynamic analysis, where static methods evaluate only one image at a time, while dynamic methods process information from multiple images simultaneously.

Static Analysis

Static methods are based on the fact that false face data is produced using masks, screens or paper printing, and that the products differ in quality and appearance from real faces. The key features are differences in texture, but also in reflection and absorption, as well as scattering and refraction of light by the material under consideration. A disadvantage is the strong dependence on the quality of the image, which is mainly influenced by the resolution of the camera and the exposure conditions.

Thanks to the ever-increasing availability of sample data for this task and machine learning, the results of these methods are now very promising - with the caveat that they only work really well for known types of attack under known circumstances.

Ultimately, by training a model with images of known deception attempts, it is possible to protect against all three levels of attack, but it is always possible that attackers will develop new methods. For this reason, it is essential that the system can be updated to ensure long-term security.

[1]https://support.apple.com/en-ca/guide/security/sec067eb0c9e/web

Dynamic analysis

Dynamic methods use information from multiple frames of the camera, i.e. they are based on movements that indicate a real person. They can be further subdivided into passive methods, which react to the natural movements of the person, and active methods, which require a specific action by the user. For the user-friendliness of an access control system, or more specifically a time and attendance system, the focus is on methods that require little or no active cooperation from the user.

The first method focused on the detection of level B and especially level A attacks. One approach uses optical flow to determine whether the visible movement of pixels corresponds to the rotation of a flat surface around itself, as would be the case in a photograph. Other optical flow methods look for a correlation between the movement of the face and the immediate background. Synchronous movement of the face and background, such as in handheld photos or mobile devices, would be classified as an attack, and purely uncorrelated movement as a real person. These methods require a minimum amount of movement by the user to be effective.

Another approach is to use the focus of the camera. By slightly shifting the focus distance past the recognised face, a depth profile can be created by changing the pixel values of the focused image. Accuracy depends on the size of the camera's focus area, and therefore on its aperture, focal length and sensor size. This method also assumes that there has been no significant movement in the scene between the two images.

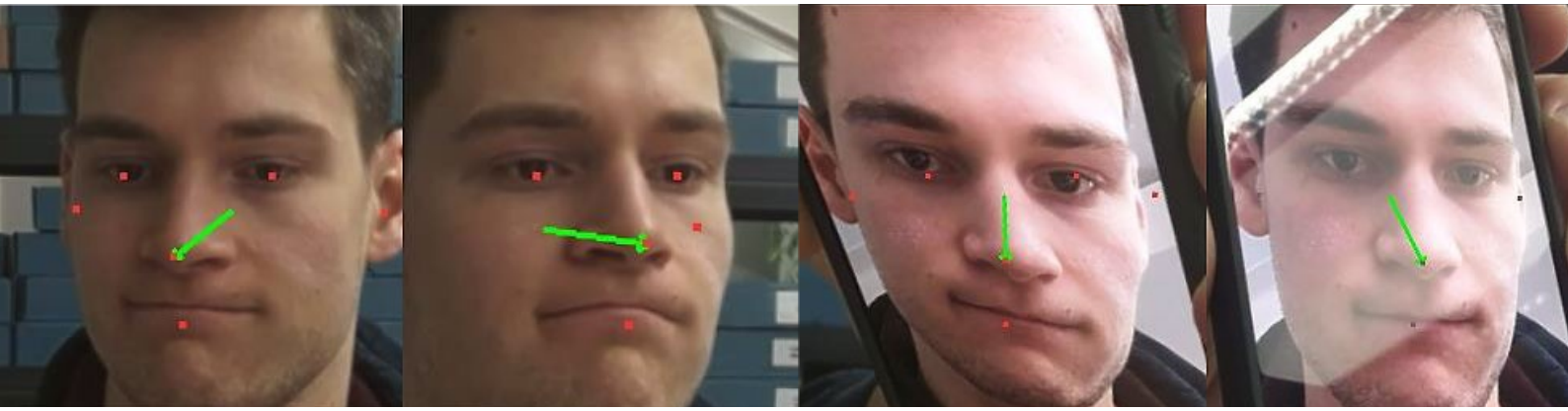
These methods would have limited, if any, success against level C masking attacks. Methods are now available that can detect the human heartbeat from the slight colour changes in short sequences of images from an RGB camera. The disadvantages of this method are the negative influence of movement and the minimum observation time of about five seconds.

Conclusion PAD

The methods presented all have different advantages and disadvantages for presentation attack detection. It is therefore imperative to use different approaches of dynamic and static analysis in parallel to provide sufficient security and usability in all situations. As facial recognition and the PAD system can usually run in parallel, it is common to combine the results of both systems when deciding on the authenticity of the authentication attempt, which can further improve accuracy. Google has demonstrated that it is indeed realistic in practice to make a purely camera-based system sufficiently secure against attacks: The camera-only Face Unlock on the Google Pixel 8 (Pro) meets the highest biometric security class in Android, and users can use it to authenticate themselves in banking apps[1].

To protect facial recognition systems against attempts at deception, the combination of static and dynamic analyses is recommended.

[1]https://blog.google/products/pixel/google-pixel-8-pro/



Attempts to deceive with photos can be recognized with the help of trigonometry. (Image: TQ-Systems GmbH)

Level A PAD system with facial recognition only

In practice, not every facial recognition application needs to be equally security-critical. For example, time and attendance systems are usually located in restricted areas. The system should be protected against level A attacks that can be carried out with simple office equipment, as attempts at deception, e.g. by pranksters in the workplace, are still conceivable.

An experiment should therefore demonstrate the potential of embedded hardware for countermeasures: The face detector already required for face recognition and alignment recognises several key points such as eyes, nose, mouth and ears. The BlazeFace model used in this work is therefore already capable of determining the position of the head in three-dimensional space behind the two-dimensional projection plane of the camera image. This information can be used in a PAD system to determine whether the recognised object is moving in space like a real head or just like a two-dimensional image.

The basic idea is similar to the optical flow-based method described above, which uses motion to determine the approximate geometry of the subject.

In the example above, the virtual point between the ears and the tip of the nose was used to visualize the position of the head in space. The real distance between the tip of the nose and the center of both ears is a known constant. Therefore, the relative position angle of the head to the projection plane can be calculated using trigonometric functions. Although the displayed face is distorted by rotating the cell phone photo, the calculated position angle of the head is not changed significantly. Due to the slight misinterpretation of the BlazeFace model caused by the distortion, relative angle differences of less than 10° occur. With a real head, differences of more than 20° can usually be observed within a short period of time, resulting spontaneously from natural movements.

This method is an effective protection against simple photo attacks, as defined in Level A according to FIDO. However, it is not possible to go beyond Level A, as this system can be fooled by a video recording. However, since only the already implemented BlazeFace model is used, the method is almost free in terms of additional computing power, especially in contrast to the high computational effort that Optical Flow would cause.

Conclusion and Outlook

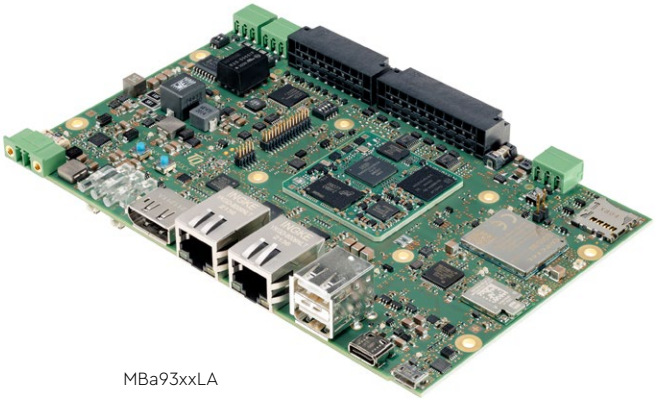
The study has successfully trained a face recognition system that works well for demonstration purposes and performs well on the target hardware, despite a very unfavorable starting condition for the training data. Access control and time & attendance systems place much lower demands on the face recognition algorithm itself than e.g. surveillance systems. It is expected that a better adaptation of the training data to the application can further improve the results. If the training patterns are generated synthetically, as in DigiFace-1M, no additional effort is required. In order to reliably test the security and robustness of the system, an application-specific test protocol should be used.

The development of an access control or time and attendance system based on facial recognition can therefore be realized with manageable resources, whereby the required effort essentially depends on the desired security level. As soon as a higher level of security is desired, more effort must be invested in the detection of facial attacks in order to effectively prevent unauthorized access. As attackers may constantly develop new methods to circumvent the usual PAD methods, the system must be able to be updated with software updates during operation in order to be sufficiently protected.

The performance of the TQMa93xxLA hardware accelerator has far exceeded expectations. The fact that sufficient hardware resources are available makes the choice of network architecture for the face recognition algorithm almost trivial. At the same time, the available performance also allows the implementation of more complex and hardware-intensive methods for PAD or the use of larger networks for training the PAD algorithms to achieve better results. This covers a wide range of hardware requirements.

Summary

1. Demonstrated the feasibility of using synthetic data as a basis for AI models for facial recognition.
2. Measured and compared the inference performance of different image classification models on the i.MX8M Plus and i.MX93 NPUs, showing a significant improvement using the newer Ethos-U65 NPU on the i.MX93.
3. An analysis of proven methods to prevent tampering and spoofing attacks for secure access control and biometric identification.
4. The combination of synthetic data, NPU-accelerated inference and PAD methods enables commercially viable facial recognition on NXP SoCs.



MBa93xxLA

Other possible uses

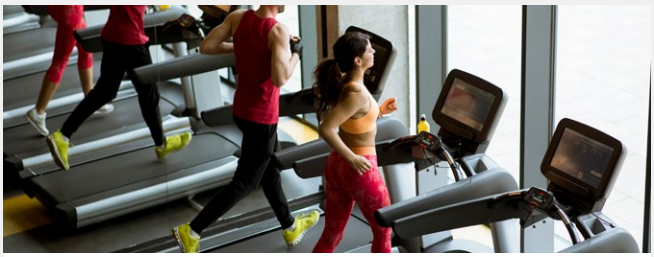
In addition to access control and time & attendance systems, facial recognition can be used for many other applications:

Recognizing recurring customers
Beverage, snack, and other vending machines can identify customers and optimize menu selections

Operator recognition
The vehicle recognizes the operator and automatically adjusts seat settings and other parameters

Smart Building
Devices for shared use will recognize the user and automatically bill the user

Supervisor mode
Users with enhanced authorisation (e.g. to cancel a transaction) are automatically recognised by the machine



Activate user profiles
Automatically activate recurring applications (such as training profiles) for recurring users

Age verification
Age characteristics are also stored in the face databases - this prevents incorrect operation by children

Parcel machines and e-commerce
Residents can conveniently receive and send parcels

About the author

Konrad Zöpf is product manager for Arm-based embedded modules and systems at TQ-Systems GmbH in Seefeld near Munich. He is also deputy division manager of TQ Embedded. He is the author of several technical articles on Arm modules and systems in connection with IOT, security and wireless.

The **technology company TQ-Group** offers the complete range of services from development, production and service to product lifecycle management. The services cover assemblies, devices and systems including hardware, software and mechanics. Customers can obtain all services from TQ on a modular basis as individual services or as a complete package according to their individual requirements. Standard products such as prefabricated microcontroller modules (mini modules), drive and automationsolutions complete the range of services.

The TQ Group employs a total of around 2,000 people at its locations in Delling, Seefeld, Inning, Murnau, Peißenberg, Peiting, Durach im Allgäu, Wetter an der Ruhr, Chemnitz, Leipzig, Fontaines (Switzerland), Shanghai (China) and Chesapeake (USA).

Your contact to TQ

Would you like to learn more about how TQ-Systems can support you with Face-AI?

info@tq-embedded.com
www.tq-group.com/en/