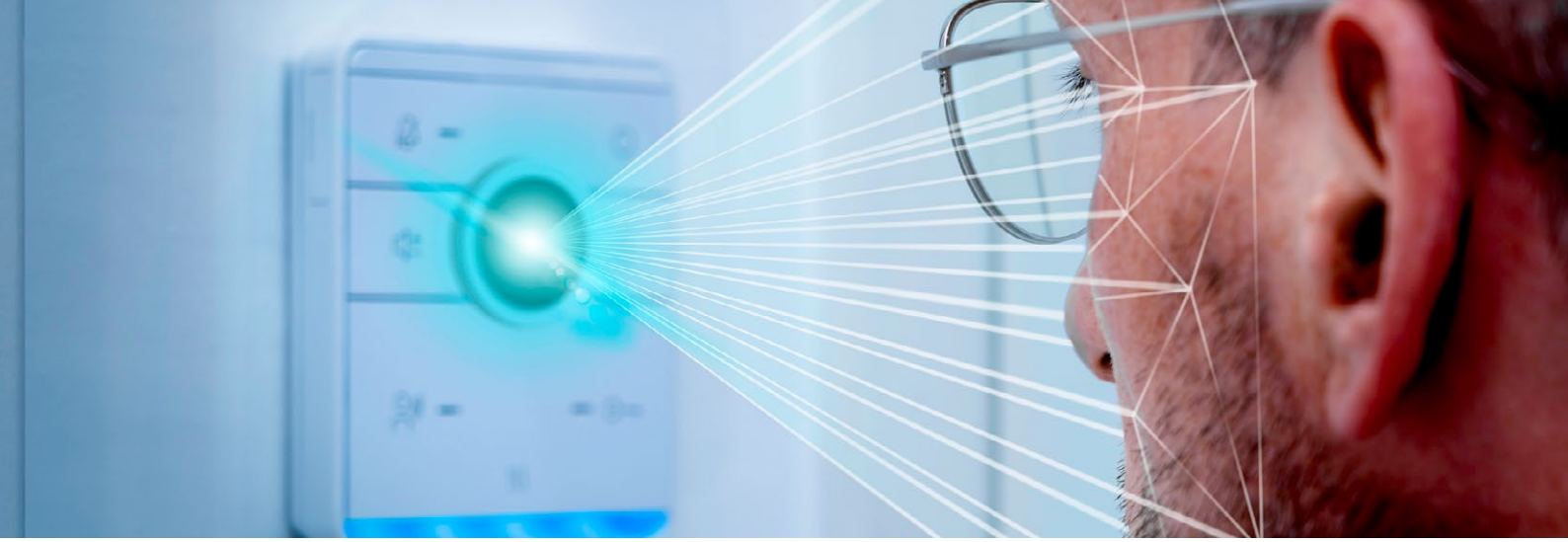
NXP Platinum Partner

WHITEPAPER

Gesichtserkennung

für Zugangskontroll- und Zeiterfassungssysteme



Schlüssel und Zugangscodes werden gelegentlich vergessen – Gesichtserkennung kennt dieses Problem nicht. (Bild: istock.com/IGphotography)

Automatisch mit dem Antlitz

TQ-Embedded hat eine Studie zur biometrischen Gesichtserkennung auf Embedded Hardware durchgeführt. Dazu wurden verschiedene Architekturen künstlicher neuronaler Netze mit einem synthetisch generierten Datensatz trainiert. Die Implementierung der trainierten Modelle erfolgte auf dem TQMa93xxLA-Modul von TQ-Embedded, das durch die integrierte Ethos-U65-NPU beschleunigt wird. Die mit den TQMa93xxLA-Modulen erzielte Leistung führt zu besseren KI-Inferenzergebnissen im Vergleich zu auf dem Papier leistungsstärkeren Modulen.
Abschließend wurden Möglichkeiten der Presentation Attack Detection evaluiert, um die Gesichtserkennung gegen Täuschungsversuche abzusichern.

Die biometrische Gesichtserkennung kann in Zeiterfassungs- und Zutrittskontrollsystemen grundsätzlich auf zwei Arten eingesetzt werden: Zum einen als zweiter Sicherheitsfaktor, um den Missbrauch verlorener oder gestohlener Zugangskarten zu verhindern oder zumindest zu erschweren und damit die Sicherheit weiter zu erhöhen. Die andere Einsatzmöglichkeit ist die Authentifizierung mittels Gesichtserkennung anstelle einer Zugangskarte für weniger sicherheitskritische Anwendungen. Beispiele hierfür sind die Zeiterfassung, die Zutrittskontrolle zu weniger kritischen Bereichen oder die Nutzung von Aufzügen und Maschinen, die nur von autorisiertem Personal bedient oder konfiguriert werden dürfen.

Das Datenschutzproblem

Um ein Gesichtserkennungsnetzwerk zu trainieren, werden Bilder von Gesichtern möglichst vieler Menschen benötigt. Die bisher gängige Methode, um möglichst schnell an große Datenmengen zu gelangen, war die Verwendung von Bildern, die in sozialen Medien hochgeladen wurden. 2018 wurde jedoch die Europäische Datenschutzgrundverordnung (DSGVO) zu EU-weit geltendem Recht und ist seitdem verpflichtend einzuhalten. Sie verbietet die Verarbeitung personenbezogener Daten von EU-Bürgern ohne deren vorherige ausdrückliche Einwilligung, was auch die bisher verwendeten Gesichtsdaten einschließt und deren rechtmäßige Verwendung für das Training von Gesichtserkennungsalgorithmen faktisch unmöglich macht.

Angesichts dieser Schwierigkeiten beschränken sich die für das Training verwendeten öffentlichen Datensätze auf die mittels Computergrafik synthetisch erzeugte Digi-Face-1M Datenbank von Microsoft. Es handelt sich dabei gegenwärtig um den effektivsten, für Forschungszwecke öffentlich zugänglichen synthetischen Datensatz für Gesichtserkennung. Dennoch eignet er sich lediglich für das Vortrainieren des Modells, für das Finetuning werden immer noch reale Daten verwendet, die dabei benötigte Menge von Gesichtsbildern realer Personen ist aber vergleichsweise gering.

Zur Validierung während des Trainings wird das Benchmark-Protokoll von LFW („Labeled Faces in the Wild") verwendet. Als Test und Vergleich der fertig trainierten Modelle dient zudem das Testprotokoll von YouTube Faces DB, kurz YTF. Die Genauigkeit errechnet sich dabei aus der Summe der Wahr-Positiv- (TPR) und Wahr-Negativ-Anteile (TNR), bei der optimalen Entscheidungsgrenze. In den Anwendungsfällen der Zugangskontrolle und Zeiterfassung ist eine geringe Falsch-Positiv-Rate (FPR) von besonderem Interesse, um unberechtigten Personen nicht den Zugang zu gewähren. Daher werden auch die Erkennungsraten bei FPR von 0,1 % und 0,01 % ausgewertet.

Finetuning der vortrainierten Modelle

Für das Finetuning der Modelle wurde aus dem TQ-internen Medienarchiv eine eigene kleine Datenbank aufgebaut. Die Inhalte des Archivs sind kommerziell nutzbar. Der DigiFace-1M-Datensatz ist ausschließlich für Forschungszwecke bestimmt. Da für die Erstellung nur Techniken verwendet wurden, die in der VFX- (Visual Effects) und Computerspielindustrie üblich sind, ist davon auszugehen, dass die Beschaffung vergleichbarer Daten für eine kommerzielle Nutzung kein größeres Problem darstellt.

Die synthetisch erzeugten Gesichter unterscheiden sich in ihrem Aussehen von echten Gesichtern. Da die trainierten Modelle bisher nur synthetische Daten gesehen haben, sollen sie nun durch Finetuning auf echte Daten optimiert werden. Dabei sollen die relevanten, durch das Vortraining gelernten Informationen möglichst nicht verloren gehen. Die Autoren von DigiFace-1M verwenden unterschiedliche Anzahlen an realen Identitäten, um zu ermitteln wie viele für gute Ergebnisse notwendig sind. Die Anzahl der Bilder je Person beträgt 20, die geringste getestete Anzahl an Identitäten beträgt 200, wobei mit dem LFW-Testprotokoll eine Genauigkeit von etwa 97 % erreicht wird. Sie empfehlen dabei die Lernrate des Netzwerks gegenüber dem Vortraining um den Faktor 100 und die des Klassifizierungs-Layer um den Faktor 10 zu senken, damit das Netzwerk zuvor Gelerntes nicht vergisst.

Die TQ-Datenbank wurde mithilfe des „BlazeFace"-Detektors nach Gesichtern in den verfügbaren Bildern gesucht und anschließend mit einem der DigiFace-1M vortrainierten „ResNet50" (Residual-Netz mit 50 Schichten) in Identitäten vorgruppiert. Letztendlich wurden alle Bilder manuell begutachtet und falsche Zuordnungen korrigiert. Auch Bilder ohne Zuordnung wurden mit den jeweils drei wahrscheinlichsten Zugehörigkeiten verglichen und anschließend von Hand richtig zugeordnet. Am Ende wurden nur die Identitäten behalten, die aus mindestens zwei Bildern bestehen. Auf diese Weise ist ein Datensatz mit 207 Klassen und 1151 Samples entstanden. Das ist etwa um den Faktor 3,6 kleiner als die kleinste getestete Menge an realen Daten in der Vorlage.

Um den größtmöglichen Nutzen aus den wenigen verfügbaren Daten zu ziehen, wird nach Möglichkeiten gesucht, das Finetuning zu verbessern. Ziel ist es, mehr Training zu ermöglichen, bevor das Netz durch das Auswendiglernen der begrenzten Daten beschädigt wird, was als Overfitting bezeichnet wird.

Neben den etablierten Regularisierungsmethoden, wie Data Augmentation, L2-Regularisierung und Dropout, wird hier versucht bestimmte Information aus dem Vortraining gezielt vor dem Verlust zu bewahren.

Der große Vorteil synthetischer Daten liegt darin, dass die Verteilung von Geschlecht, Hautfarbe, ethnischer Herkunft und Alter der im Datensatz repräsentierten Identitäten steuerbar ist und somit leicht an die tatsächliche Verteilung in der Weltbevölkerung angepasst werden kann. Im Gegensatz dazu ist die Verteilung der für die Feinabstimmung verfügbaren Daten typischerweise suboptimal. Die Identitäten im Datensatz verteilen sich während des Trainings gleichmäßig auf der durch das Embedding gebildeten Hypersphäre. So maximiert sich die Distanz zwischen den Klassenzentren und auch die Zuverlässigkeit der Erkennung. Die gleichmäßige und faire Verteilung der Bevölkerungsgruppen, die durch die synthetischen Daten antrainiert wird, ist also auch im Embedding gespeichert, welches am Modell-Output erzeugt wird. Optische Merkmale, bei denen sich reale und synthetische Daten unterscheiden, liegen hingegen eher in der Bildebene und werden typischerweise in den vorderen Netzwerkschichten nahe dem Modell-Input verarbeitet.

Statt das gesamte Netzwerk mit derselben Lernrate zu trainieren, wird von Eingangs- zur Ausgangsschicht des Netzwerks hin eine exponentiell kleiner werdende Lernrate realisiert. Die Hoffnung dabei ist, dem Netzwerk die optischen Eigenschaften der realen Daten schneller und die unfaire Verteilung der Identitäten langsamer anzutrainieren.

Und schließlich das Deployment auf TQMa93xxLA

Vor der Implementierung der trainierten Modelle auf dem Target müssen diese zunächst in das erforderliche Format gebracht werden. Typischerweise müssen die Netzwerkparameter für die Inferenz mit NPU-Architekturen von 32-Bit Fließkommazahlen in 8-Bit Ganzzahlen quantisiert werden. Wie im i.MX Machine Leaning User's Guide beschrieben, muss das trainierte Netz außerdem in das Tensorflow Lite Format übertragen und anschließend mit einer von NXP bereitgestellten Software für die NPU des i.MX 93 kompiliert werden.

Ergebnisse

In der folgenden Tabelle 1 sind Ergebnisse mit den verschiedenen, nur mit DigiFace-1M vortrainierten Netzarchitekturen abgebildet. Die Protokolle LFW (Labeled Faces in the Wild) und YTF (YouTube Faces DB) wurden zum Testen und Vergleichen der Modelle verwendet.

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 %

Tabelle 1

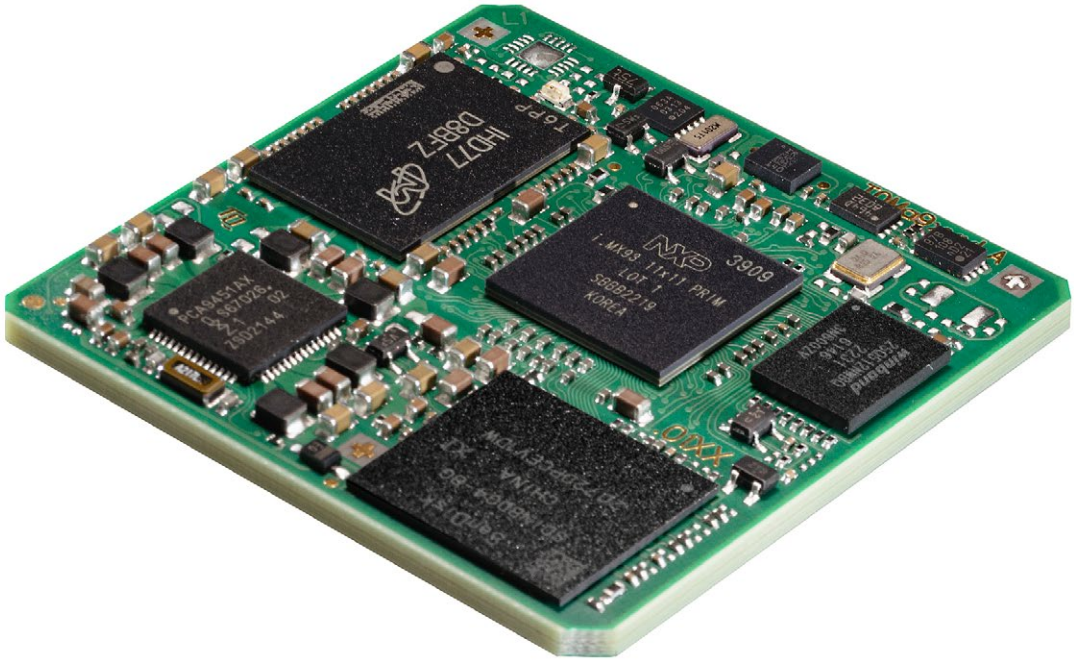
Mit ResNet50 wird ein Ergebnis von 93,50 % mit LFW erreicht, was dem in der Veröffentlichung zu DigiFace-1M mit demselben Netz erreichten Wert von 94,55 % sehr nahekommt. Es ist anzunehmen, dass der leicht geringere Wert auch auf die etwas niedrigere, hier verwendete Inputauflösung zurückzuführen ist. Das EfficientNet-lite0 Modell, das Testweise mit der höheren Auflösung trainiert wurde, erzielt mit beiden Protokollen etwa 0,9 Prozentpunkte mehr, die Erfolgsrate bei FPR<0,01 % nimmt durch die Erhöhung der Auflösung besonders zu.
Bei den effizienteren Modellen erreichen besonders MobileNetV3 Large und die Ausbaustufe EfficientNet-lite1 vergleichsweise hohe Erkennungsraten bei FPR von <0,1 % und <0,01 %.

Nach Finetuning

Wie in Tabelle 2 zu sehen, ist der Unterschied zwischen den Ergebnissen der kleineren Modelle und denen des deutlich größeren ResNet50 durch das Finetuning weiter geschrumpft. Auch der Vorsprung des mit höherer Auflösung trainierten EfficientNet-lite0 ist hier nun größtenteils verschwunden. Auffällig ist aber, dass nach dem Finetuning die Werte für FPR<0,01 mit YTF in fast allen Fällen gegenüber den vortrainierten Versionen in Tabelle 1 eingebrochen sind.

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 %

Tabelle 2



Einlötbare Module wie das TQMa93xxLA ermöglichen durch ihre kompakten Abmessungen neue Anwendungsmöglichkeiten. (Bild: TQ-Systems GmbH)

Bei oberflächlicher Betrachtung des Problems fällt auf, dass die in YTF enthaltenen Daten häufig besonders schlechte Bildqualität aufweisen. In manchen Fällen scheitert aufgrund dessen schon das Alignment (Festlegung des Bildausschnitts) durch BlazeFace, was dem Modell die Klassifizierung zusätzlich erschwert. Zudem müssen für korrekte Klassifizierung oft starke Unterschiede bei Perspektive, Gesichtsausdruck, Frisur und Accessoires, Beleuchtung, partieller Überdeckung und in einigen Fällen auch Alter überwunden werden. Auch LFW testet Modelle auf diese typischen Hürden, allerdings deutlich weniger stark ausgeprägt.

In den DigiFace-1M Datensatz wurden diesbezüglich sehr starke Variationen künstlich eingebaut, während die Daten aus dem TQ-Medienarchiv fast nichts dergleichen enthalten, da hier für die meisten Personen alle Bilder am selben Tag unter ähnlichen Bedingungen gemacht wurden. Es liegt also nahe, dass durch das Finetuning etwas an Robustheit gegenüber diesen Variationen verloren gegangen ist, was sich nun besonders bei YTF in den Ergebnissen mit effektiv Null Toleranz gegenüber Falsch Positiv äußert.

Die hier erreichten Ergebnisse liegen im Vergleich zu großen Datensätzen mit realen Identitäten weit zurück. So erreicht „ArcFace" 99,83 % mit LFW und 98,02 % mit YTF. Auch in „FaceNet" werden bereits 99,63 % mit LFW und 95,10 % mit YTF ermöglicht. Dennoch ist das wie beschrieben trainierte Netzwerk in einer Messedemo für das TQMa93xxLA zum Einsatz gekommen.

Für die Demo wurde das EfficientNet-lite0 verwendet, die Entscheidungsschwelle wurde auf etwa denselben Wert gesetzt, bei dem laut LFW ein FPR von kleiner 0,01 % möglich ist. Obwohl mit dieser Entscheidungsschwelle etwa die Hälfte aller berechtigten Zugangsversuche scheitern müsste, hat das System im Verhältnis dazu sowohl bei Tests, als auch auf der Messe sehr zuverlässig funktioniert.

Der Grund dafür ist wohl die anwendungsbedingt deutlich reduzierte Anforderung gegenüber unterschiedlichen Perspektiven, Gesichtsausdrücken und partieller Überdeckung robust sein zu müssen. Bei Zugangskontrollsystemen und dem verwendeten Demoaufbau kommt nur ein Kameratyp zum Einsatz, die Personen stehen immer in ähnlichen Abständen frontal vor der Kamera und haben auch in den meisten Fällen einen neutralen Gesichtsausdruck.

Performance mit i.MX 8M Plus und i.MX 93 NPU

Die NPU des i.MX 8M Plus liefert laut NXP bis zu 2,3 TOPS an Leistung, die Ethos-U 65 NPU des eine Leistungsklasse niedriger angesiedelten i.MX 93 soll hingegen nur 0,5 bis 1 TOPS leisten können. Demensprechend ist die Erwartungshaltung, dass die Inferenzzeit der Modelle sich im Vergleich zum Vorgänger in etwa verdoppelt.

Getestet wurde mithilfe eines vorkompilierten Benchmark-Programms, das in den von NXP bereitgestellten Softwarekomponenten, die für den Betrieb der NPU notwendig sind, enthalten ist. Es wurde die Latenzzeit von jeweils 100 Inferenzen gemessen und gemittelt.

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

Tabelle 3

Wie es sich zeigte, ist der i.MX 93 entgegen der Erwartung in den meisten Fällen sogar schneller als der Vorgänger. Die EfficientNet-lite-Modelle sind etwa 40 Prozent schneller, ResNet50 dagegen nur 20 Prozent langsamer. MobileNetV3 Large läuft mit dem i.MX 8M Plus nicht besonders schnell, da die Version scheinbar einige Operationen enthält, die mit der NPU des i.MX 8M Plus nicht kompatibel sind und daher auf der CPU berechnet werden müssen. Hier ist der i.MX 93 sogar mehr als doppelt so schnell.

Täuschungsversuche erkennen

Neben der korrekten Erkennung der Personen im Rahmen der gestellten Aufgabe, müssen biometrische Systeme auch gegen Betrugsversuche gewappnet sein. Eine besondere Bedeutung hat dabei die Presentation Attack Detection (PAD): Bei Presentation Attacks oder auch Spoofing Attacks handelt es sich um Angriffe auf Zugangskontrollsysteme, durch Vorhalten gefälschter biometrischer Daten. Da kamerabasierte Gesichtserkennungssysteme nur mit zweidimensionalen Bilddaten arbeiten, macht sie das besonders anfällig für solche Angriffe.

Meist genügt das präsentieren eines Bildes auf dem Bildschirm eines Mobilgerätes oder ausgedruckt auf einem Blatt Papier, um sich für eine andere Person auszugeben. Die FIDO-Allianz definiert in ihren Biometric Requirements drei verschiedene Angriffsarten (Level A, B und C), die nach Zeitaufwand, der notwendigen Expertise und des Zugangs zur Quelle der biometrischen Daten geordnet sind. Die auf Gesichtserkennung bezogenen Beispiele, die durch FIDO für die jeweiligen Angriffslevel genannt werden, sind im Folgenden dargestellt.

	Quelle der biometrischen Charakteristika	Schwierigkeit	Art des Angriffs
Level A (leicht)	Fotos aus sozialen Medien	**Zeit:** < 1 Tag **Expertise:** Laie **Equipment:** Standard	Bild eines Gesichts auf Papier gedruckt / auf Mobilgerät angezeigt
Level B (moderat)	Hochauflösendes Foto, Videoaufnahme der Zielperson	**Zeit:** < 7 Tage **Expertise:** Geübt **Equipment:** Standard + Speziell	Papiermaske, Bewegtbild des Gesichts auf Bildschirm abgespielt
Level C (schwer)	Hochauflösendes Foto, 3D-Informationen des Gesichts der Zielperson	**Zeit:** > 7 Tage **Expertise:** Experte **Equipment:** Speziell + maßgefertigt	Silikonmaske, Theatermaske

Tabelle 4

Um ein System gegen solche Angriffe zu schützen, gibt es verschiedene Ansätze. Eine Möglichkeit ist es, mit weiteren Sensoren zusätzliche Informationen zu erfassen, wie Tiefeninformationen, um das System robuster gegen Angriffe zu machen. Ein Beispiel hierfür ist das in Apple iPhones verbaute FaceID, das das TrueDepth Infrarot Sensorsystem für die dreidimensionale Abtastung des Nutzergesichts verwendet[1]. Weitere Möglichkeiten sind Infrarot-, Wärmebild-, Lichtfeld-, Multispektral- und Stereokameras. Der Einsatz von zusätzlichen Sensoren ist jedoch häufig mit sehr hohen Kosten bei Entwicklung und Material verbunden und nicht in allen Designs realisierbar. Auch für das Verbessern bereits bestehender Systeme sind sie selten ein Option. Zusätzlich bedroht der Fortschritt im 3D-Druck zunehmend auch die Sicherheit von Systemen, die mit Tiefensensoren oder 3D-Kameras ausgestattet sind.

Es gibt eine Reihe von Möglichkeiten, rein kamerabasierte Gesichtserkennungssysteme robuster gegen Angriffe aller drei FIDO-Arten zu machen. Dabei wird zwischen statischer und dynamischer Analyse unterschieden, wobei statische Verfahren jeweils nur ein Bild auswerten, während dynamische Verfahren Informationen aus mehreren Bildern gleichzeitig verarbeiten.

Statische Analyse

Statische Methoden basieren darauf, dass gefälschten Gesichtsdaten Masken, Bildschirme oder Papierdruck nutzen und sich die Erzeugnisse in ihrer Qualität und ihrem Aussehen von echten Gesichtern unterscheiden. Die entscheidenden Eigenschaften sind dabei die Unterschiede in der Textur, aber auch bei der Reflexion und Absorption, sowie der Streuung und Brechung des Lichts durch das betrachtete Material. Ein Nachteil ist die starke Abhängigkeit von der Qualität der Aufnahme, die vor allem durch die Kameraauflösung und die Belichtungsbedingungen beeinflusst wird.

Dank der sich stetig verbessernden Verfügbarkeit von Beispieldaten für diese Aufgabe und Maschinellem Lernen sind die Resultate dieser Verfahren mittlerweile sehr vielversprechend - mit der Einschränkung, dass sie nur bei bekannten Angriffsarten unter bekannten Umständen wirklich gut funktionieren.

Da letztlich nur ein Modell mit Bildern bekannter Täuschungsversuche trainiert wird, ist ein Schutz gegen alle drei Angriffsebenen zwar möglich, es muss jedoch immer damit gerechnet werden, dass Angreifer neue Methoden entwickeln. Aus diesem Grund ist eine Updatefähigkeit des Systems für eine dauerhafte Sicherheit unbedingt erforderlich.

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

Dynamische Analyse

Dynamische Methoden verwenden Informationen aus mehreren Frames der Kamera, basieren also auf Bewegungen, die auf eine echte Person schließen lassen. Sie lassen sich noch weiter unterteilen in passive Methoden, die auf natürliche Bewegungen der Person reagieren, und aktive Methoden, die eine bestimmte Aktion des Nutzers fordern. Für die Nutzerfreundlichkeit eines Zugangskontroll- oder besonders einem Zeiterfassungssystem, beschränkt man sich auf Methoden, die keine bis nur minimale aktive Kooperation des Nutzers erfordern.

Die ersten Methoden fokussierten auf die Erkennung von Level-B- und besonders auf Level-A-Angriffe. Ein Ansatz verwendet Optical Flow um festzustellen, ob die sichtbare Bewegung die Pixel der Rotation einer ebenen Fläche um sich selbst entsprechen, wie es bei einem Foto der Fall wäre. Andere auf Optical Flow basierende Methoden suchen eine Korrelation zwischen der Bewegung des Gesichts und des unmittelbaren Hintergrunds. Synchrone Bewegung von Gesicht und Hintergrund, wie bei Bewegungen von handgehaltenen Fotos oder Mobilgeräten, würden so als Angriff, und die rein unkorrelierte Bewegung als echte Person klassifiziert werden.

Diese Methoden erfordern ein gewisses Mindestmaß an Nutzerbewegung, um effektiv zu sein.

Ein weiterer Ansatz ist das Nutzen des Fokus der Kamera. Durch leichtes Verschieben der Fokusdistanz am erkannten Gesicht vorbei, kann durch Änderung der Pixelwerte zum fokussierten Bild ein Tiefenprofil erstellt werden. Die Genauigkeit hängt dabei von der Größe des Fokusbereichs der Kamera, und damit von dessen Blende, Brennweite und Sensorgröße ab. Bei dieser Methode wird zudem davon ausgegangen, dass zwischen den beiden Bildern keine wesentliche Bewegung in der Szene stattgefunden hat.

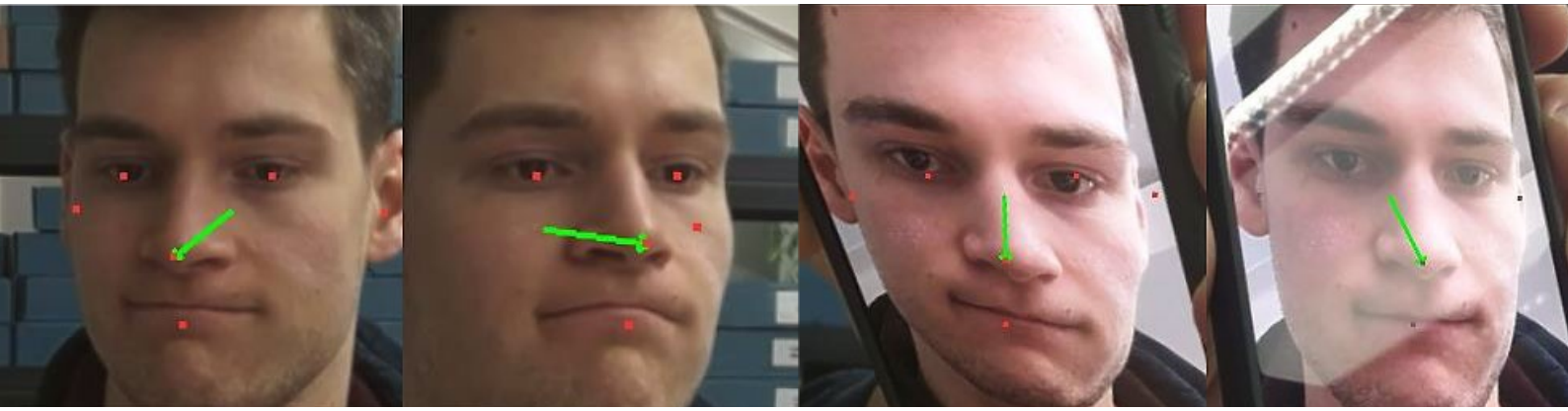
Diese Methoden wären nur bedingt oder überhaupt nicht in der Lage, Level-C-Angriffe mittels Masken erfolgreich abzuwehren. Hierfür existieren mittlerweile Methoden, die den menschlichen Herzschlag über die leichten Farbänderungen in kurzen Bildabfolgen einer RGB-Kamera ermitteln können. Die Nachteile dieser Methode sind der negative Einfluss von Bewegung, sowie der etwa fünfsekündige Betrachtungszeitraum den das System mindestens benötigt.

Schlussfolgerung PAD

Die vorgestellten Methoden haben alle verschiedene Vor- und Nachteile für die Presentation Attack Detection. Es ist daher naheliegend verschiedene Ansätze dynamischer sowie statischer Analyse parallel einzusetzen, um in allen Situationen hinreichende
Sicherheit und Nutzbarkeit bieten zu können. Da die Gesichtserkennung und das PAD-System in der Regel parallel laufen können, ist es üblich die Ergebnisse beider Systeme in der Entscheidung über die Echtheit des Authentifizierungsversuchs zu vereinigen, was die Genauigkeit weiter verbessern kann. Dass es in der Praxis auch wirklich realistisch ist, ein rein kamerabasiertes System ausreichend gegen Angriffe zu sichern, wurde von Google demonstriert: Das rein kamerabasierte „Face Unlock" des Google Pixel 8 (Pro) genügt der höchsten biometrischen Sicherheitsklasse in Android und Nutzer können sich damit in Banking Apps authentifizieren[1].

Um Gesichtserkennungssysteme gegen Täuschungsversuche abzusichern, empfiehlt sich die Kombination von statischen und dynamischen Analysen.

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



Täuschungsversuche mit Fotos können mit Hilfe der Trigonometrie erkannt werden. (Bild: TQ-Systems GmbH)

Level-A PAD-System nur mit Face Detector

In der Praxis muss nicht zwangsläufig jeder Anwendungsfall für Gesichtserkennung gleichermaßen sicherheitskritisch sein. So befinden sich Zeiterfassungssysteme meistens bereits in zugangsbeschränkten Bereichen. Da Täuschungsversuche, beispielsweise durch Schabernack treibende Arbeitskollegen, dennoch denkbar sind, sollte das System, gegen mit einfachem Büromaterial durchführbare Angriffe auf Level A, gewappnet sein.

Ein Experiment soll daher das Potenzial von Embedded-Hardware für Gegenmaßnahme aufzeigen: Der für die Erkennung und das Alignment der Gesichter bereits benötigte Face Detector erkennt mehrere Schlüsselpunkte, wie Augen, Nase, Mund und Ohren. Das bei dieser Arbeit hierfür verwendete BlazeFace-Modell ist also bereits dazu fähig, die Lage des Kopfes im dreidimensionalen Raum hinter der zweidimensionalen Projektionsebene des Kamerabilds zu ermitteln. Diese Information kann in einem PAD-System genutzt werden, um festzustellen, ob sich das erkannte Objekt im Raum bewegt wie ein echter Kopf oder nur wie ein zweidimensionales Abbild.

Die Grundidee ähnelt der oben beschriebenen, auf Optical Flow basierenden Methode, anhand von Bewegung die ungefähre Geometrie des Subjekts zu ermitteln.

Im obigen Bild gezeigten Beispiel wurde der virtuelle Punkt zwischen den Ohren und der Nasenspitze verwendet, um die Lage des Kopfes im Raum zu visualisieren. Da der reale Abstand zwischen Nasenspitze und Mittelpunkt beider Ohren eine bekannte konstante Größe ist, kann mithilfe trigonometrischer Funktionen der relative Lagewinkel des Kopfes zur Projektionsebene berechnet werden. Durch gezieltes Drehen des Handyfotos wird das dargestellte Gesicht zwar verzerrt, der errechnete Lagewinkel des Kopfes wird dadurch jedoch nur unwesentlich verändert. Aufgrund der leichten Fehlerkennungen des BlazeFace-Modells durch die Verzerrung treten dabei relative Winkeldifferenzen von weniger als 10° auf. Bei einem echten Kopf sind meist schon innerhalb eines kurzen Zeitraums Unterschiede von über 20° zu beobachten, die unaufgefordert durch natürliche Bewegungen entstehen.

Diese Methode ist ein effektiver Schutz gegen einfache Foto-Angriffe, wie in Level A nach FIDO definiert. Mehr als Level A ist damit aber in keinem Fall möglich, da sich dieses System durch eine Videoaufnahme täuschen lässt. Da lediglich das bereits implementierte BlazeFace-Modell genutzt wird, ist die Methode allerdings nahezu kostenlos in Bezug auf die zusätzlich benötigte Rechenleistung, besonders im Gegensatz zu dem hohen Rechenaufwand den Optical Flow verursachen würde.

Fazit und Ausblick

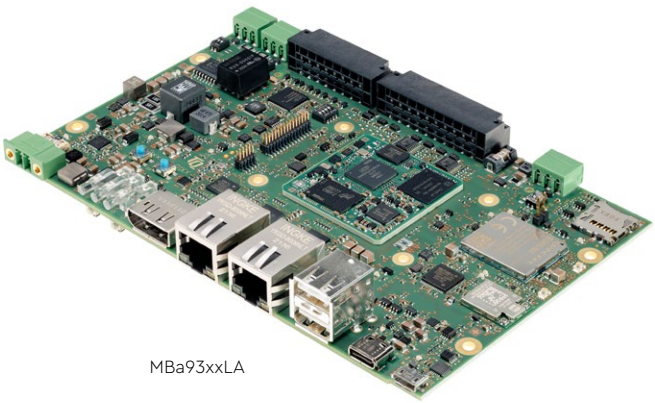
Im Rahmen der Studie konnte ein für Demonstrationszwecke gut funktionierendes und auf der Zielhardware performant laufendes Gesichtserkennungssystem erfolgreich trainiert werden, trotz einer sehr ungünstigen Ausgangslage bei den Trainingsdaten. Zugangskontroll- und Zeiterfassungssysteme stellen deutlich geringere Anforderungen an den Gesichtserkennungsalgorithmus selbst als z. B. Überwachungssysteme. Es ist zu erwarten, dass eine bessere Anpassung der Trainingsdaten an die Anwendung die Ergebnisse weiter verbessern kann.
Bei einer synthetischen Generierung der Trainingsmuster nach dem Vorbild von DigiFace-1M ist dies mit keinem zusätzlichen Aufwand verbunden. Um die Sicherheit und Robustheit des Systems zuverlässig testen zu können, sollte ein anwendungsspezifisches Testprotokoll verwendet werden.

Die Entwicklung eines auf Gesichtserkennung basierenden Zutrittskontroll- oder Zeiterfassungssystems ist somit mit überschaubaren Ressourcen realisierbar, wobei der Aufwand im Wesentlichen vom gewünschten Sicherheitsniveau abhängt. Sobald ein höheres Sicherheitsniveau erreicht werden soll, muss entsprechend mehr Aufwand in die Presentation Attack Detection investiert werden, um das Eindringen Unbefugter effektiv verhindern zu können. Da Angreifer potenziell ständig neue Methoden entwickeln können, um die gängigen PAD-Methoden zu umgehen, muss das System im laufenden Betrieb mit Software-Updates aktualisiert werden können, um ausreichend gewappnet zu sein.

Der Hardwarebeschleuniger des TQMa93xxLA hat die Erwartungen an seine Leistungsfähigkeit bei weitem übertroffen. Die Tatsache, dass mehr als ausreichend Hardware-Ressourcen zur Verfügung stehen, macht die Wahl der Netzarchitektur für den Gesichtserkennungsalgorithmus nahezu trivial. Gleichzeitig erlaubt die verfügbare Leistung auch die Implementierung komplexerer und hardwareintensiverer Methoden für PAD oder die Verwendung größerer Netze für das Training der PAD-Algorithmen, um bessere Ergebnisse zu erzielen. Damit wird ein breites Spektrum an Anforderungen an die Hardware abgedeckt.

Zusammenfassung

1. Demonstration der Nutzbarkeit synthetischer Daten als Grundlage für KI-Modelle zur Gesichtserkennung.
2. Messung und Vergleich der Inferenzleistung verschiedener Bildklassifizierungsmodelle auf den i.MX8M Plus und i.MX93 NPUs, wobei eine signifikante Verbesserung bei Verwendung der neueren Ethos-U65 NPU auf dem i.MX93 festgestellt wurde.
3. Eine Analyse bewährter Methoden zur Verhinderung von Manipulations- und Spoofing-Angriffen für eine sichere Zugangskontrolle und biometrische Identifikation.
4. Die Kombination aus synthetischen Daten, NPU-beschleunigter Inferenz und PAD-Methoden ermöglicht eine kommerziell nutzbare Gesichtserkennung auf NXP SoCs.



MBa93xxLA

Weitere Einsatzgebiete

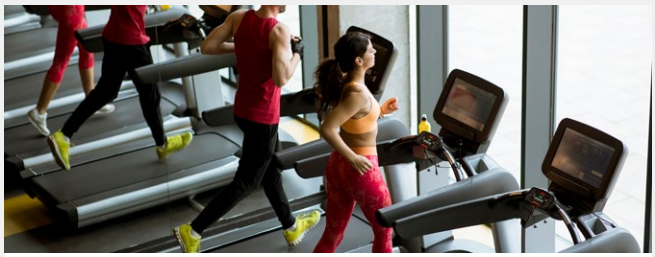
Gesichtserkennung kann neben Zugangskontroll- und Zeiterfassungssysteme auch für zahlreiche weitere Anwendungen zum Einsatz kommen:

Stammkundenerkennung
Getränke-, Snack- und andere Verkaufsautomaten erkennen die Stammkunden und optimieren die Menüauswahl

Fahrererkennung
Das Fahrzeug erkennt seinen Fahrer und passt die Sitzeinstellungen und andere Parameter automatisch an.

Smart Building
Geräte zur Gemeinschaftsnutzung erkennen die Nutzer und rechnen automatisch ab.

Supervisor-Modus
Anwender mit erhöhten Berechtigungen (z. B. für Parameteränderungen) werden von der Maschine automatisch erkannt.



User-Profil aktivieren
Wiederkehrende Anwendungen (z. B. Trainingsprofile) von wiederkehrenden Usern automatisch aktivieren.

Altersfreigabe
In Gesichtsdatenbanken sind auch Altersmerkmale hinterlegt – damit lässt sich eine Fehlbedienung durch Kinder verhindern

Paketautomaten und e-commerce
Hausbewohner können bequem Pakete Empfangen und Versenden

Über den Autor

Konrad Zöpf ist Produkt Manager für ARM-basierte Embedded-Module und -Systeme bei TQ-Systems GmbH in Seefeld bei München. Zudem ist er stellvertretender Geschäftsbereichsleiter von TQ Embedded.
Er ist Autor mehrerer Fachartikel zu den Themen ARM-Module und -Systeme in Verbindung zu IOT, Security und Wireless.

Das **Technologie-Unternehmen TQ-Group** bietet das komplette Leistungsspektrum von der Entwicklung, Produktion und Service bis hin zum Produktlebens-zyklusmanagement. Die Dienstleistungen umfassen dabei Baugruppen, Geräte und Systeme inklusive Hardware, Software und Mechanik. Kunden können bei TQ sämtliche Leistungen modular als Einzelleistungen wie auch im Komplettpaket entsprechend ihrer individuellen Anforderungen beziehen. Standard-produkte wie fertige Mikrocontrollermodule (Mini-module), Antriebs- und Automatisierungslösungen ergänzen das Dienstleistungsspektrum.

Die TQ-Group beschäftigt an den Standorten Delling, Seefeld, Inning, Murnau, Peißenberg, Peiting, Durach im Allgäu, Wetter an der Ruhr, Chemnitz, Leipzig, Fontaines (Schweiz), Shanghai (China) und Chesapeake (USA) insgesamt rund 2.000 Mitarbeiter.

Ihr Kontakt zu TQ

Sie möchten mehr darüber erfahren, wie TQ-Systems Sie beim Thema Face-KI unterstützen kann?

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