Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Vakbijlage

Automatische gezichtsbeeldvergelijking

Inhoudsopgave

Inleiding

1. De vergelijking van de gezichtsfabeeldingen door software

2. De vergelijkingen van testafbeeldingen

3. Controle prestatie software

4. Interpretatie zaakscore en bepaling bewijskracht

5. Controle robuustheid

6. Naar de eindconclusie

Inleiding

In vergelijkend gezichtsbeeldonderzoek wordt meestal gevraagd een gezichtsbeeld van een onbekende persoon te vergelijken met een gezichtsbeeld van een bekende persoon, om uit te zoeken of de twee gezichtsbeelden van dezelfde persoon zijn of van twee verschillende personen. De afbeelding van het onbekende gezicht (vaak van de dader van een misdrijf) noemen we het betwiste beeld en de afbeelding van bekende gezicht (vaak een politiefoto van de verdachte) noemen we het referentiebeeld. Hieronder wordt de methode toegelicht waarbij software gebruikt wordt om de gezichtsbeelden te vergelijken. De andere methode, waarbij de mens kijkt en vergelijkt, wordt toegelicht in de vakbijlage ‘Algemene onderzoeksmethode vergelijking van gezichtsbeelden’. Uit meerdere onderzoeken is gebleken dat de combinatie van het menselijk oordeel met het op software gebaseerde oordeel duidelijk een voordeel heeft voor de accuraatheid van de gezichtsvergelijking (1, 2, 3).

De onderzoeksvraag die gesteld wordt komt meestal neer op: “Is de persoon zichtbaar op de camerabeelden de verdachte?”. Bij het NFI vertalen we die vraag naar twee hypothesen:

Zelfde-persoon-hypothese:
De persoon afgebeeld in het betwiste beeld is dezelfde als de persoon afgebeeld in het referentiebeeld.

Verschillende-personen-hypothese:
De persoon afgebeeld in betwiste beeld is een ander persoon dan de persoon afgebeeld in het referentiebeeld.

Het resultaat van vergelijkend beeldonderzoek drukt uit hoeveel beter de bevindingen passen bij de ene hypothese dan bij de andere.

Vergelijking met automatische gezichtsbeeldvergelijking-software houdt de volgende stappen in:

1. Bepaling van de beeldkwaliteit door de software
De software vergelijkt het betwiste beeld met test-gezichtsbeelden van verschillende kwaliteit. Daarmee wordt de kwaliteit en bruikbaarheid van het betwiste beeld bepaald (4).

2. De vergelijking van de twee beeldbestanden door de software
De software vergelijkt het betwiste beeld met het referentiebeeld en drukt de mate van overeenkomst uit in een ‘gelijkheidsscore’.

3. Controle van de prestatie van de software
Opnamen van gezichtsbeelden met een vergelijkbare kwaliteit uit een NFI-verzameling worden vergeleken met een set gezichtsbeelden van goede kwaliteit (de referentiedatabase) door de software. Dat levert zelfde-gezicht-scores en verschillend-gezicht-scores op. Met de zelfde-gezicht-scores en verschillend-gezicht-scores wordt gecontroleerd hoe goed de software werkt in de omstandigheden van de zaak.

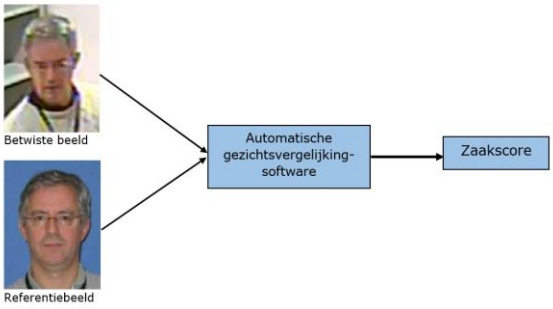
4. Interpretatie zaakscore en bepaling bewijskracht
De zaakscore wordt afgezet tegen de zelfde-gezicht-scores en verschillende-gezichten-scores van de referentiedatabase om de bewijskracht van de gelijkheidsscore te bepalen.

5. Controle robuustheid
Het hele onderzoek wordt uitgevoerd met telkens net verschillende keuzes van de ingeschatte kwaliteit van het beeldmateriaal. Daarmee wordt gecontroleerd of de uitkomst van het onderzoek robuust is voor kleine veranderingen in de bepaling van de kwaliteit van het beeldmateriaal.

6. Naar de eindconclusie
Het resultaat van de automatische gezichtsbeeldvergelijking wordt samengenomen met de uitkomst van de methode waarbij de mens kijkt en vergelijkt.

Gebruikte software
De software die op het NFI gebruikt wordt om te vergelijken is momenteel FaceVACS-DBScan-ID, versie 5.4.0.0 van Cognitec Systems GmbH. Van deze software is uit validatie-onderzoek gebleken dat die in principe geschikt is voor gebruik in NFI-zaakonderzoek. Als uit validatie-onderzoek in de toekomst blijkt dat andere software beter presteert, kan het gebeuren dat het NFI andere software zal gaan gebruiken. Dat zal naar verwachting geen grote gevolgen hebben voor hoe de software in zaakonderzoek wordt ingezet.

Werking van de software
Het betwiste gezichtsbeeld en het referentiebeeld worden aangeboden aan de software. De software haalt uit elk van de beelden kenmerken van het gezicht. Deze kenmerken beschrijven grofweg de vormen van het gezicht. Waarden van deze kenmerken worden – binnen de software – numeriek vastgesteld.
Vervolgens gebruikt de software de waarden van de kenmerken van het betwiste beeld en de waardenvan het referentiebeeld om een mate van overeenkomst te bepalen en drukt die uit in een score. Deze score geeft weer in hoeverre de gezichten in de twee beelden op elkaar lijken. Deze score – dus tussen het betwiste beeld en het referentiebeeld – noemen we hier de zaakscore, omdat het de score is die uit de vergelijking van het zaakmateriaal komt.



Figuur 1. Werking van automatische gezichtsvergelijking-software. Twee gezichtsbeelden worden ingeladen in de software. Die bepaalt een mate van overeenkomst tussen de gezichtsbeelden, en drukt die uit in de zaakscore.

1. Bepaling van de beeldkwaliteit door de software

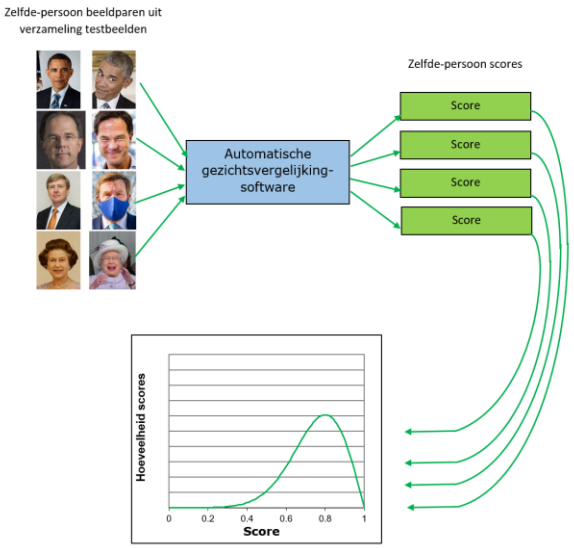
Opname-omstandigheden
De zaakscore uit een vergelijking met software van het zaakmateriaal is natuurlijk afhankelijk van hoeveel de gezichtsbeelden op elkaar lijken. Maar: eigen validatie-onderzoek en de wetenschappelijke literatuur op dit punt laten zien dat de uitkomst kan worden beïnvloed door vele factoren. Voorbeelden daarvan zijn: resolutie en compressie, scherpte van de afbeelding, pose, belichting, gezichtsuitdrukking, bedekking van het gezicht, etc.
Omdat er meer dan alleen gezichtskenmerken van invloed zijn op de zaakscore, kan de zaakscore niet op zichzelf geïnterpreteerd worden. Wat in een vergelijking tussen twee beelden van een bewakingscamera een hoge score is, kan in een vergelijking tussen een beeld van een bewakingscamera en een referentiebeeld een lage score zijn. Alleen de zaakscore is dus niet genoeg: die moet worden bekeken in het licht van hoe de software normaalgesproken scoort bij vergelijkingen met de betreffende kwaliteit van de beelden. Daarvoor is een speciale testdatabase aangelegd door het NFI.

Verzameling van testbeelden
Om vergelijkingen in de omstandigheden van de zaak mogelijk te maken wordt gebruik gemaakt van testbeelden van personen. Die beelden bestaan uit een verzameling forensisch realistisch video- en fotomateriaal die door het NFI is aangelegd (4). In die verzameling is bekend welke persoon is afgebeeld. Het is dus mogelijk om daar paren van opnamen uit te kiezen waarvan bekend is dat daarop dezelfde persoon is afgebeeld (zelfde-persoon-paar) en paren van beelden waarvan bekend is dat het om twee verschillende personen gaat (verschillende-personen-paar).

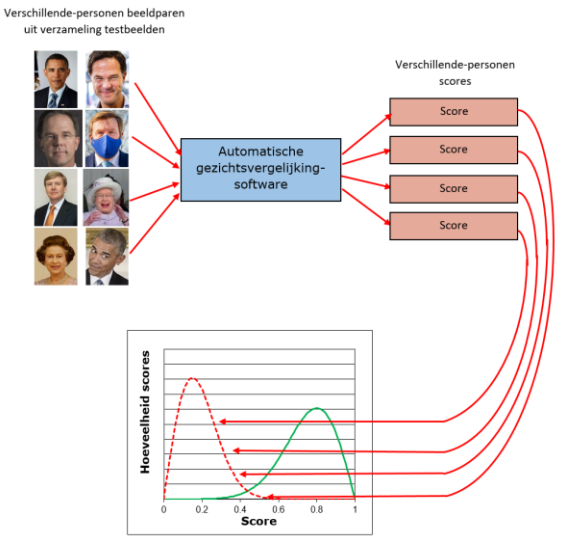
Deze paren van testbeelden worden geselecteerd op basis van de kwaliteit van de betwiste beelden. In een zaak kan het bijvoorbeeld gaan om vergelijking van slechte beelden van een bewakingscamera met een goede afbeelding van een politiefoto. Paren van vergelijkbare bewakingsbeelden en goede kwaliteit foto's worden dan geselecteerd om de omstandigheden van de zaak te simuleren.

Vergelijking testbeelden
Er zijn dus twee soorten paren van beelden: zelfde-persoon-paren en verschillende-personen-paren. Deze paren opnamen worden allemaal vergeleken door de software, op dezelfde manier zoals dat al gebeurde met het zaakmateriaal. Dat levert twee soorten scores op: scores uit vergelijkingen van een zelfde-persoon-paar (zelfde-persoon-scores) en scores uit vergelijkingen van een verschillende-personen-paar (verschillende-personen-scores).

Zelfde-persoon-scores en verschillende-personen-scores
Deze twee typen scores worden gevisualiseerd in twee scoredistributies. Zie figuren 2 en 3.



Figuur 2. Paren van gezichtsbeelden waarvan zeker is dat die van dezelfde persoon zijn (zelfde-persoon-paren) worden onderling vergeleken. Daaruit komen zelfde-persoon-scores, die worden weergegeven in een distributie (de groene lijn in de grafiek). In dit voorbeeld liggen de meeste zelfde-persoon-scores rond 0,8.



Figuur 3. Paren van gezichtsbeelden waarvan zeker is dat ze van verschillende personen zijn (verschillende-personen-paren) worden onderling vergeleken. Daaruit komen verschillende-personen-scores, die worden samengevat in een distributie (de rode lijn in de grafiek). In dit voorbeeld liggen de meeste verschillende-personen-scores rond 0,15.

De zelfde-persoon-scoredistributie (de groene curve) bestaat uit scores die waargenomen zijn als de zelfde-persoon hypothese waar is. De verschillende-personen-distributie (de rode curve) bestaat uit waargenomen scores als de verschillende-personen hypothese waar is. Voor de vergelijking met het zaakmateriaal vormen de groene en rode curve de scores die we kunnen verwachten als het een zelfde-persoon paar is of een verschillende-personen paar.

Met deze twee distributies worden twee dingen gedaan: Ten eerste wordt gekeken of de software wel goed presteert in de omstandigheden van de zaak (paragraaf 3 hieronder). Ten tweede kan worden bekeken of de zaakscore beter past bij de zelfde-persoon scores of bij de verschillende-persoon scores (paragraaf 4 hieronder).

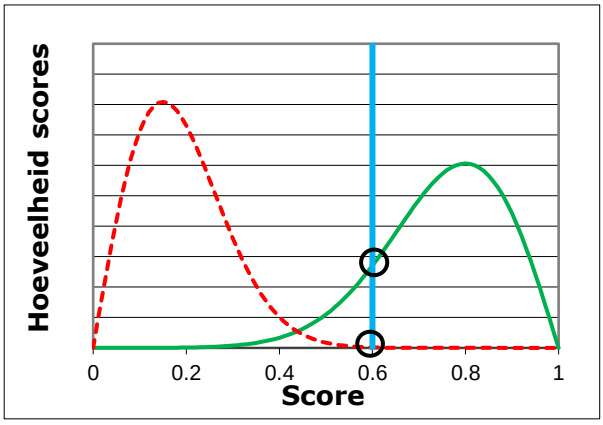
2. De vergelijking van de twee beeldbestanden door de software

De zaakscore is een getal dat tussen de 0 en 1 ligt. Het getal zelf heeft geen betekenis anders dan: hoe hoger, des te meer overeenkomst tussen de beelden is gevonden door de software. Om dat getal goed te kunnen interpreteren als bewijs dat in de richting van de zelfde-persoon hypothese wijst of in de richting van de verschillende-personen hypothese, moet dit getal nog worden afgezet tegen scores waarvan bekend is of die afkomstig zijn uit vergelijkingen van dezelfde persoon of uit vergelijkingen van verschillende personen voor beelden met een vergelijkbare bruikbaarheid voor gezichtsvergelijking.

Meer zaakmateriaal
In veel zaken is er meer dan één beeld aan betwist materiaal of meer dan één beeld aan referentiemateriaal beschikbaar. Indien er meerdere beelden aan referentiemateriaal beschikbaar zijn dan worden die zoveel mogelijk gezamenlijk aan de software aangeboden, zodat die een zo compleet mogelijk beeld van het gezicht van de persoon kan gebruiken. Als er meer dan genoeg referentiemateriaal is, kan het voorkomen dat de onderzoeker een aantal opnamen uit het referentiemateriaal kiest. Daarbij wordt gekeken naar de kwaliteit en opnamedatum van de beelden, zodat de beelden met hoogste kwaliteit en opnamedatum dichtst bij de opnamedatum van de betwiste beelden kunnen worden gekozen.
Omdat van het betwiste materiaal niet bekend is wie is afgebeeld, kan in zaken met meerdere betwiste opnamen meestal niet worden aangenomen dat het allemaal beelden van dezelfde persoon betreffen. Daarom worden de betwiste opnamen apart behandeld, dus in een zaak met vijf betwiste opnamen wordt het gehele onderzoek vijf keer uitgevoerd: eenmaal voor elke betwiste opname.

3. Controle van de prestatie van de software

Om te controleren of de software voldoende onderscheid kan maken tussen de beelden van verschillende personen in de omstandigheden van de zaak, wordt gekeken naar de twee scoredistributies. Als de twee distributies ver uit elkaar liggen, betekent dat dat de software in de omstandigheden van de zaak goed onderscheid kan maken tussen zelfde-persoon paren en verschillende-personen paren. Zelfde-persoon vergelijkingen leveren dan immers meestal hogere scores dan verschillende-personen paren, en de uitkomst van de software bij de zaakvergelijking geeft dan waarschijnlijk bruikbare informatie.
Als de twee distributies voor een groot deel overlappen, betekent dat dat de software in de omstandigheden van de zaak niet goed onderscheid kan maken. Of er nu tweemaal dezelfde persoon of twee verschillende personen worden vergeleken: het levert altijd ongeveer dezelfde score op. De zaakscore levert dan dus geen bruikbare informatie op. Als dit het geval blijkt te zijn in een zaak, zal in het rapport kortweg worden vermeld dat uit onderzoek is gebleken dat het materiaal zich niet leent voor automatische gezichtsbeeldvergelijking en zal de methode verder niet worden toegepast.



Figuur 4. De scores uit zelfde-persoon vergelijkingen (groene lijn) en de scores uit verschillend-personen vergelijkingen (rode, onderbroken lijn) worden gebruikt om de bewijskracht van de zaakscore (verticale blauwe lijn) te bepalen. In dit voorbeeld past de zaakscore veel beter bij een zelfde-persoon score dan bij een verschillende-personen score. De zaakscore (van ca. 0.6) kruist de groene lijn in dit voorbeeld op een hoger punt dan de rode lijn, wat betekent dat de zaakscore veel meer voorkomt bij zelfde-persoon vergelijkingen dan bij verschillende-personen vergelijkingen. De likelihood ratio (de maat voor bewijskracht) wordt berekend door de hoogtes van het kruispunt van de blauwe en groene lijn en van het kruispunt van de blauwe en rode lijn (aangegeven met cirkels) op elkaar te delen.

4. Interpretatie zaakscore

Met deze twee distributies kan bovendien de zaakscore geïnterpreteerd worden. Er kan nu namelijk gekeken worden of de zaakscore meer voorkomt bij zelfde-persoon vergelijkingen of juist meer voorkomt bij verschillende-personen-vergelijkingen. In Figuur 4 is hiervan een voorbeeld gegeven. De zaakscore van ongeveer 0.6 (aangegeven met de blauwe lijn) kruist de groene distributie op een veel hoger punt dan de rode distributie. Dat betekent dat een score van 0.6 veel vaker voorkomt bij zelfde-persoon vergelijkingen dan bij verschillende-personen vergelijkingen. Dit zou zich dus vertalen in een bewijskracht die naar de zelfde-persoon hypothese wijst. Als de zaakscore op 0.3 zou liggen, is dat juist andersom. Die zaakscore komt veel vaker voor bij verschillende-personen vergelijkingen dan bij zelfde-persoon vergelijkingen en het resultaat zou juist meer naar de verschillende-personen-hypothese wijzen.

5. Controle robuustheid bruikbaarheidsbepaling

In het onderzoek is een speciale testdatabase met een grote hoeveelheid verschillende kwaliteiten gezichtsbeelden (bijvoorbeeld lage resolutie, niet frontaal, onscherp, verschillende belichting, et cetera, zie referentie 4) gebruikt om een inschatting te maken van de bruikbaarheid van het betwiste beeldmateriaal voor vergelijking met het referentiemateriaal. Voor een goede bepaling van de bruikbaarheid van het betwiste materiaal is het nodig dat beeldmateriaal met vergelijkbare verstorende eigenschappen (bijvoorbeeld resolutie, pose en belichting) in deze testdatabase aanwezig is. Om te testen of de methode voor het bepalen van de bruikbaarheid robuust is, worden ook de uitkomsten van de vergelijking berekend onder de aanname dat de bruikbaarheid hoger of lager is.
Als de uitkomsten met als aanname een andere bruikbaarheid voor vergelijking in de buurt liggen van de oorspronkelijke uitkomst, is de methode kennelijk

robuust voor deze bruikbaarheidsbepaling. In dat geval wordt de ordegrootte van de uitkomst gebruikt in het bepalen van de eindconclusie.
Als de uitkomsten onderling sterk verschillen, dan blijkt dat de methode erg gevoelig is voor de bruikbaarheidsbepaling. In dat geval wordt de uitkomst van het onderzoek met meer voorzichtigheid behandeld.

6. De eindconclusie

Retrospectief zaakonderzoek heeft aangetoond dat er een goede correlatie is tussen de automatische gezichtsbeeldvergelijking en de vergelijkingen uitgevoerd door de NFI-onderzoekers. De uitkomsten van deze twee methoden kunnen worden samengenomen tot één eindconclusie resulterend in een betere berouwbaarheid van de uitspraak dan de onafhankelijke oordelen (1,2,3). Afhankelijk van de omstandigheden in de zaak kan de ene methode zwaarder worden meegewogen dan de andere. Per zaak wordt bekeken hoe deze uitkomsten zich vertalen naar één eindconclusie.
Daarbij wordt rekening gehouden met:
- hoe bruikbaar het zaakmateriaal is voor elk van de methodes
- hoe goed de gezichtsbeelden van de testpersonen overeenkomen met de beeldkwaliteit in het zaakmateriaal
- hoe robuust de methode bleek voor de verschillende inschattingen van bruikbaarheid (zie paragraaf 5)
- hoe goed elk van de methoden getest en gevalideerd is met het aangeboden type beelden.

De eindconclusie wordt uitgedrukt in een verbale schaal, waarin wordt aangegeven hoeveel waarschijnlijker de bevindingen zijn onder de ene hypothese dan onder de andere. Zie voor meer uitleg over de conclusieschaal de vakbijlagen “Algemene onderzoeksmethode vergelijking van gezichtsbeelden” en “De reeks waarschijnlijkheidstermen van het NFI”.

Referenties:

1) P.J. Phillips, A.N. Yates, Y.Hu, C.A. Hahn, A. Noyes, K. Jackson, et al. (2018). Face recognition accuracy of forensic examiners, superrecognisers,, and face recognition algorithms, PNAS, 201721355. https://doi.org/10.1073/pnas.1721355115.
2) R. Moreton (2021). Expertise in Applied Face Matching: Training, Forensic Examiners, Super Matchers and Algorithms. PhD thesis The Open University.
3) A. Towler, J. Dunn, S. Martínez, R. Moreton, F. Eklöf, A. Ruifrok, R. Kemp, et al. (2021). “Diverse Routes to Expertise in Facial Recognition.” PsyArXiv. November 25. doi:10.31234/osf.io/fmznh.
4) A.C.C. Ruifrok, P. Vergeer, A. Macarulla Rodrigues (2022). From facial images of different quality to score based LR. Forensic Science International 332,111.

.

Voor algemene vragen kunt u contact opnemen met de Frontdesk, telefoon (070) 888 68 88. Voor inhoudelijke vragen kunt u contact opnemen met het onderzoeksgebied Beeldonderzoek en Biometrie van de afdeling Digitale en Biometrische Sporen telefoon (070) 888 6425.

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid
Postbus 24044 | 2490 AA Den Haag

Telefoon (070) 888 66 66
www.forensischinstituut.nl

5 september 2022