Kwaadwillige intenties beter gedetecteerd met gecombineerde observaties*

Remco Wijn & Jack Vogels

Dit artikel laat zien dat het combineren van observaties van afwijkend gedrag door meerdere toezichthouders leidt tot betere oordelen over mogelijke kwaadwillige intenties dan oordelen van individuele toezichthouders. In de hier beschreven studie bekeken ervaren toezichthouders CCTV-beelden van een winkelcentrum waarin zakkenrollers actief waren. Zij konden mensen op de beelden aanwijzen als ze vonden dat die afwijkend gedrag vertoonden. Met signaaldetectiestatistiek berekenden we de accuratesse waarmee toezichthouders mensen op de beelden met kwade intenties (onze assistenten) konden onderscheiden van mensen zonder kwade intenties (winkelend publiek). We vergeleken de gemiddelde accuratesse van individuele toezichthouders met de accuratesse van combinaties van twee, drie of vier toezichthouders. Het blijkt dat het combineren van observaties van twee toezichthouders leidt tot een hogere sensitiviteit dan observaties van individuele toezichthouders.

1 Inleiding

Een trend voor beveiligingsorganisaties en -diensten is proactief beveiligen. Proactief beveiligen omvat zowel het gebruik van gedragsanalysemethoden om afwijkend gedrag te herkennen voordat een misdaad plaatsvindt, als de bereidheid en de vaardigheid om vroegtijdig risico's te mitigeren, bijvoorbeeld door individuen te benaderen op basis van hun afwijkende gedrag (Elias 2009; Van Pel, Verhagen & Wijn 2012; Van Rest, Roelofs & Van Nuenen 2014). Toezichthouders hoeven daarbij niet altijd een beroep te doen op hun bevoegdheden, maar kunnen vanuit hun rol als gastheer van een gebied of locatie iemand aanspreken (Van der Kleij, Roelofs & Van Hemert 2013). De trend naar proactief beveiligen wordt gevoed door de motivatie om veiligheidsrisico's te vermijden, beperken of voorkomen (Adam, Beck & Van Loon 2000) en is ten minste ten dele toe te schrijven aan risico's met grote, onomkeerbare gevolgen, zoals terroristische aanslagen (Minister van Justitie 2003).

Een belangrijke component van proactief beveiligen is de aandacht van toezichthouders voor afwijkend gedrag. Afwijkend gedrag is gedrag dat voorafgaat en gerelateerd is aan incidenten (Wijn, Van Rest, Burghouts & Lousberg 2012). Een goede, vroegtijdige beoordeling kan bijdragen aan het voorkomen van incidenten,

* Dit onderzoek is gefinancierd door de rijksoverheid en uitgevoerd binnen het TNO vraaggestuurd programma Veilige maatschappij; Topic 1: Afwijkend gedrag. De auteurs zijn Rick van der Kleij, Jeroen van Rest, en Maaike Lousberg dankbaar voor suggesties op eerdere versies van deze bijdrage.

doi: 10.5553/TvV/187279482015014001003

vergrijpen of delicten. Het inschatten van afwijkend gedrag is echter een lastige taak. Een van de redenen daarvoor is dat afwijkende gedragingen die voorafgaan en gerelateerd zijn aan incidenten zich vaak over langere tijd uitspreiden. Toezichthouders kunnen hierdoor stukjes informatie zien die mogelijk opvallend of afwijkend zijn, maar op zichzelf onvoldoende reden geven om de persoon in kwestie te benaderen of staande te houden. Een goed oordeel over kwaadwillige intenties kan vaak pas gegeven worden wanneer er meerdere relevante aanwijzingen of afwijkende gedragingen zijn.
Soms zijn toezichthouders in de gelegenheid om iemand bij wie ze afwijkend gedrag menen waar te nemen gedurende enige tijd te volgen. Zo zouden ze meerdere afwijkende gedragingen kunnen waarnemen. Vaak is deze gelegenheid slechts zeer beperkt of niet aanwezig. Bovendien verplaatst een potentiële dader zich vaak binnen of tussen verschillende toezichtgebieden, waardoor meerdere toezichthouders deze potentiële dader en eventueel afwijkend gedrag zouden kunnen waarnemen. In dergelijke gevallen zou het combineren van observaties van afwijkend gedrag (die volgens de afzonderlijke toezichthouders te weinig aanleiding vormen om in te grijpen of een interactie aan te gaan) een manier kunnen zijn om betrouwbaardere inschattingen te maken van een kwaadwillige intentie. Iemand die volgens meerdere toezichthouders afwijkend gedrag vertoont, zou dan aangesproken kunnen worden.
In dit artikel onderzoeken we of de inschatting van mogelijke kwaadwillige intentie beter wordt wanneer observaties van afwijkend gedrag van meerdere toezichthouders worden gecombineerd. Een goede inschatting betekent zowel correcte herkenning van een kwaadwillige, ofwel een *hit*, als een correcte herkenning van een niet-kwaadwillige, ofwel een *correct rejection*. In een eerder onderzoek waarin de vraag naar het combineren van tags centraal stond (Bouma, Vogels, Aarts, Kruszynski, Wijn & Burghouts, 2013) werden deze metrieken zelfstandig gebruikt om de effectiviteit van de combinaties aan te tonen.

In dit artikel willen we met behulp van signaaldetectiestatistiek een beter en completer antwoord geven op de vraag hoe observaties (van afwijkend gedrag) van meerdere afzonderlijke operators te combineren zijn en of dat leidt tot een beter onderscheid tussen kwaadwillige en niet-kwaadwillige individuen.

Om deze vraag te beantwoorden hebben we een onderzoek uitgevoerd waarbij we acteurs in een winkelcentrum andere acteurs lieten zakkenrollen. We toonden opgenomen beelden aan toezichthouders die geïnstrueerd waren om afwijkend gedrag onder bezoekers van het winkelcentrum te signaleren. Deze meldingen van afwijkend gedrag combineerden we vervolgens om de vraag te beantwoorden of meer paar ogen meer zien dan één, of dat meer paar ogen meer fouten maken dan één. Voordat we ingaan op het onderzoek en de resultaten, gaan we nader in op de signaaldetectietheorie (Swets, Tanner & Birdsall 1961) die de basis vormt voor onze definitie en berekening van effectiviteit.

De signaaldetectietheorie (SDT) wordt gebruikt bij het onderzoeken van beslissingen die zijn omgeven door onzekerheid, zoals de vraag of geobserveerd afwijkend gedrag een voorbode is van een incident. De onzekerheid bestaat dan uit het

Tabel 1 Reacties op signalen

	Signaal aanwezig	**Signaal afwezig**
Signaal wel gezien	Hit	False alarm
Signaal niet gezien	Miss	Correct rejection

onderscheiden van relevante *signalen* uit een veelheid aan *ruis*. Gedragingen die voorafgaan en gerelateerd zijn aan een incident zijn in SDT-termen 'signalen'. Een zakkenroller communiceert bijvoorbeeld op afstand met zijn handlanger over het benaderen van een slachtoffer, gaat dichtbij potentiële slachtoffers staan en probeert ze af te leiden. Deze gedragingen zijn signalen; ze gaan vooraf en zijn gerelateerd aan een zakkenrolincident. Een toezichthouder die dit ziet gebeuren, kan besluiten dat een signaal aanwezig is (dan heeft hij of zij het goed gezien: een *hit* – een incident kan worden voorkomen) of dat het signaal afwezig is (dan heeft hij of zij het niet goed gezien: een *miss* – het delict kan plaatsvinden, omdat de toezichthouder de signalen niet heeft onderkend). Dezelfde beslissingen kan een toezichthouder nemen wanneer er in werkelijkheid geen signaal is. Een toezichthouder kan ook dan besluiten dat een signaal aanwezig is (dan heeft hij of zij het verkeerd gezien: *false alarm*) of dat het signaal afwezig is (dan heeft hij of zij het goed gezien: een *correct rejection*).

Bij het herkennen van afwijkend gedrag ontstaat onzekerheid (en dus moeilijkheid), omdat normaal gedrag en afwijkend gedrag overlappen. Iemand die dicht bij een ander gaat staan, kan de intentie hebben om deze te zakkenrollen, maar kan dit ook om tal van andere, niet kwaadwillige redenen doen. Het punt waar een toezichthouder vindt dat gedrag van normaal overgaat in afwijkend is een inschatting die per toezichthouder kan verschillen:

– liberaal criterium (in SDT-termen): toezichthouder vindt gedrag al snel, op basis van weinig informatie, afwijkend . Dit leidt tot een grotere kans op hits, maar ook tot een grotere kans op *false alarms*;
– conservatief criterium (in SDT-termen): toezichthouder bestempelt gedrag minder snel als afwijkend en grijpt dus pas bij veel aanwijzingen in. Dit leidt tot meer *misses* (mogelijk leidend tot een incident), maar ook tot meer *correct rejections* (niet aanhouden van onschuldigen).

Het vorenstaande laat zien dat de kansen op *hits* en *false alarms* beïnvloed worden door de algemene geneigdheid van een toezichthouder om een liberaal of conservatief criterium te hanteren. Die geneigdheid wordt de *respons bias* genoemd. Daarnaast worden kansen op *hits* en *false alarms* beïnvloed door de mate waarin een toezichthouder in staat is signaal van ruis, of in dit geval daders van winkelend publiek, te onderscheiden. Dit vermogen wordt de sensitiviteit genoemd. Het belang van SDT ligt in haar theoretische en statistische vermogen om *respons bias* en sensitiviteit van elkaar te onderscheiden en invariant voor elkaar te maken (Stanislaw & Todorov 1999).

doi: 10.5553/TvV/187279482015014001003

SDT wordt doorgaans gebruikt om individuele prestaties in kaart te brengen. In het huidige artikel gebruiken we SDT om oordelen van toezichthouders te combineren en de sensitiviteit van gecombineerde prestaties van twee, drie of vier toezichthouders in kaart te brengen en af te zetten tegen individuele prestaties. Het combineren van oordelen met signaaldetectiestatistiek werd al eerder theoretisch onderzocht (o.a. door Batchelder & Romney 1986; Sorkin, West & Robinson 1998). Sorkin, Hays en West (2001) leverden een empirische onderbouwing voor het combineren door deelnemers aan hun studies te laten schatten of de gemiddelde hoogte van negen naast elkaar afgebeelde staafjes hoger (signaal) of lager (ruis) was dan een gestelde grens. Vervolgens onderzochten ze hoe ze de waarnemingen van de deelnemers konden combineren en welke gevolgen dat had voor de sensitiviteit op groepsniveau.
Hoewel een dergelijke taak het mogelijk maakt om heel nauwkeurig individuele met groepsprestaties te vergelijken, geeft het onvoldoende inzicht of complexere stimuli succesvol te combineren zijn en tot hogere groepsprestaties leiden. Volgens ons is afwijkend gedrag een dergelijke categorie van complexe stimuli. Bovendien wordt met afwijkend gedrag vaak een heel patroon aan gedragingen bedoeld die in samenhang en afhankelijk van de context afwijkend kunnen zijn (Van Rest, Roelofs & Van Nunen 2014). Vanwege deze complexiteit lijken toezichthouders dan ook vooral op intuïtie of impliciete kennis te vertrouwen als het gaat om het herkennen van afwijkend gedrag.

In deze studie onderzoeken we of het combineren van observaties van afwijkende gedragingen (die een indicatie kunnen zijn van een zakkenrolincident) door verschillende toezichthouders leidt tot een hogere sensitiviteit dan de observaties van individuele toezichthouders. We verwachten dat meerdere toezichthouders meer (en systematisch beter) afwijkend gedrag zien dan individuele toezichthouders en dat dit leidt tot een hogere groepssensitiviteit. Om dit te onderzoeken lieten we ervaren beveiligers kijken naar beelden van een winkelcentrum waarin zakkenrollers te zien waren en lieten we ze passanten omkaderen (taggen) die volgens hen afwijkend gedrag vertoonden. We gebruikten hiervoor V-TAG, een softwareprogramma dat is ontwikkeld voor dit onderzoek. Daarmee kunnen de beveiligers op CCTV-beelden een kader om mensen trekken, waarna de tag wordt opgeslagen en kan worden vergeleken met andere opgeslagen tags.

2 Methode

2.1 Deelnemers

Tien ervaren beveiligers (acht mannen en twee vrouwen) namen tijdens werktijd deel aan het onderzoek, waarvoor ze geen extra compensatie ontvingen.

2.2 Stimulusmateriaal

We trainden acht acteurs in zakkenrollerstechnieken. De training duurde een halve dag, waarin deelnemers op basis van theorie en beveiligingsbeelden, en met hulp van experts uit de praktijk leerden om te zakkenrollen. Vervolgens kregen

acteurs opdracht om in telkens wisselende uiterlijke verschijningen en in wisselende samenstellingen van één, twee of drie in een winkelcentrum tussen winkelend publiek andere acteurs te zakkenrollen. Daarvoor droegen we in totaal tien andere acteurs op om individueel op een naïeve manier door het winkelcentrum te lopen en daar op aangewezen plekken boodschappen te doen. De zakkenrollende individuen of teams rolden zo gedurende drie uren bij achttien mensen hun portemonnee. Een individu of team was telkens een uur actief in het winkelcentrum en beroofde in dat uur drie mensen. Op elk moment waren er twee teams tegelijkertijd actief. Per uur werden zo zes zakkenrollersincidenten gecreëerd die verdeeld over het winkelcentrum plaatsvonden.
Statische CCTV-apparatuur nam alle incidenten op. Uit het opgenomen beeldmateriaal creëerden we zes sets van camerabeelden. Drie sets besloegen de ene helft van het winkelcentrum en de andere drie sets besloegen de andere helft. Elke set bestond uit vier camerabeelden en toonde ongeveer één uur synchroon afgespeelde camerabeelden.

2.3 Procedure

Na binnenkomst namen de deelnemers plaats achter een computer, uitgerust met een 21 inch-scherm en een muis. We vertelden hen dat ze naar CCTV-beelden zouden gaan kijken van een winkelcentrum in Utrecht. We vertelden dat op die beelden normaal winkelend publiek te zien was, maar dat er ook criminelen doorheen liepen die zakkenrolden van individuen in het publiek. We vroegen deelnemers om mensen een tag te geven wanneer ze dachten afwijkend gedrag te zien (i.e. gedrag dat gerelateerd is en voorafgaat aan een incident). Dit konden ze doen door het beeld stil te zetten en een kader (tag) om de verdachte te trekken. Dit beeld werd vervolgens opgeslagen om de getagde persoon te kunnen koppelen aan tags van dezelfde persoon, gemaakt door dezelfde of andere deelnemers.
Iedere deelnemer bekeek twee sets van ongeveer een uur met daartussen een pauze. We registreerden per deelnemer de specifieke set die ze bekeken, de tijd en het cameranummer waarin een tag was gegeven en de uitsnede van de tag waarop de verdachte te zien was. Na het verzamelen van deze data sorteerden we handmatig de foto's. We gaven alle personen die waren getagd een uniek identificatienummer en zorgden ervoor dat tags van dezelfde personen hetzelfde identificatienummer kregen. Zo konden we zien hoe vaak één individu getagd was, door wie, waar en op welke tijd. Vervolgens kenden we aan de identificatienummers een waarde toe die aangaf of het betreffende individu een van onze zakkenrollende acteurs was, een slachtofferacteur of een onbekende, onschuldige omstander.[1]

Voor SDT moeten we weten hoe de zichtbaarheid van signalen en ruis zich tot elkaar verhouden. In dit onderzoek zijn signalen en ruis de aanwezigheid van respectievelijk zakkenrollers en omstanders in de videobeelden. Deze waren echter

1 Navraag bij de beveiliging van het winkelcentrum en bij de politie leerde dat er geen meldingen van echte incidenten waren gedaan op de dag van de opnames. Onze eigen zakkenrollersincidenten waren dus de enige echte incidenten op deze dag.

Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003

niet allemaal even lang in beeld. Onze zakkenrollers liepen een uur lang door het winkelcentrum en waren dus per cameraset van vier camera's die de helft van het winkelcentrum besloeg bij benadering dertig minuten te zien. Overig winkelend publiek was doorgaans korter te zien. Op basis van trackinggegevens is af te leiden dat het aantal personen dat per minuut een willekeurig punt in het winkelcentrum passeert ongeveer op 25 ligt. Per uur zouden dus bij benadering 1500 mensen een willekeurig punt in het winkelcentrum passeren. Observaties van publiek laten zien dat winkelend publiek gemiddeld vijftien minuten (0,25 uur) zichtbaar is op de camerasets. Dit leidt tot de schatting dat er 1500*0,25 = 375 bezoekers per uur in het winkelcentrum zichtbaar waren. Per cameraset die de helft van het winkelcentrum besloeg, zouden dus in een uur 375/2 = 187,5 bezoekers te zien geweest zijn. Voor de berekening van de proportie *false alarms* gebruiken we deze waarde.
We berekenden voor elke deelnemer de proportie *hits* door het aantal correct getagde zakkenrollers te delen door de aanwezigheid van zakkenrollers in het tijdsblok.

Formule 1:

Hits:

$$H = \frac{Z_{tagged}}{Z_{totaal}}$$

De proportie *false alarms* berekenden we door voor elke deelnemer het aantal getagde onschuldige omstanders te delen door het (geschatte) totale aantal omstanders in het tijdsblok.

Formule 2:

False alarms:

$$F = \frac{O_{tagged}}{O_{totaal}}$$

Mensen op de CCTV-beelden laten een continue stroom gedragingen zien waarvan er één of meer door deelnemers als afwijkend zouden kunnen worden gezien. Deelnemers konden daarom zo vaak ze wilden een individu taggen. Het was echter ook mogelijk voor deelnemers om in korte tijd meerdere tags te geven zonder dat daar nieuwe observaties van afwijkend gedrag aan ten grondslag lagen. Een deelnemer kan bijvoorbeeld onzeker zijn geweest of zijn tag wel goed was, kan nog een tag hebben gegeven, omdat de mogelijke dader een moment later beter herkenbaar in beeld was, of heeft met snelle opeenvolgende tags willen aangeven erg zeker te zijn van de verdenking. Om de data te schonen van dit type doublures die niet op basis van nieuwe afwijkende gedragingen ontstonden, gebruikten we voor onze analyses een minimaal interval van dertig seconden tussen twee tags

op eenzelfde individu. Daarbij redeneerden we dat binnen dat interval gemaakte herhaalde tags niet door afwijkend gedrag waren ontstaan.
Omdat de standaarddeviaties van signaal en ruis in de meeste gevallen niet gelijk zijn, wordt meestal gekozen voor een niet-parametrische benadering van de sensitiviteit. Er zijn verschillende niet-parametrische methoden om sensitiviteit (het vermogen om signalen van ruis te onderscheiden) te berekenen waarvan A' de meest gebruikte is. Een A'-waarde van 0.5 geeft aan dat er geen onderscheid te maken is tussen signalen en ruis, terwijl een A'-waarde van 1 duidt op een perfect onderscheid tussen signalen en ruis.
In het huidige onderzoek transformeerden we voorafgaand aan het berekenen van A' alle proporties gelijk aan 0 naar 0.001 en proporties gelijk aan 1 naar 0.999 om oneindige z-scores en daardoor dataverlies te vermijden (Stanislav & Todorov 1999).

Formule 3:

Sensitiviteit:

$$A' = .5 + \left[sign\,(H - F) \frac{(H - F)^2 + |H - F|}{4\max(H, F) - 4HF} \right]$$

Sensitiviteit kan per individuele deelnemer worden berekend, zoals hiervoor. Sensitiviteit kan ook voor combinaties van deelnemers worden berekend. Door tags van verschillende deelnemers te combineren, kunnen we berekenen of de sensitiviteit toeneemt ten opzichte van de gemiddelde individuele sensitiviteitsuitkomsten. Hiervoor blijft formule 3 ongewijzigd en definiëren we collectieve *hits* en collectieve *false alarms* respectievelijk als:

Formule 4:

Collectieve hits:

$$Hc = \frac{\sum_2^n Z_{tagged}}{Z_{totaal}}$$

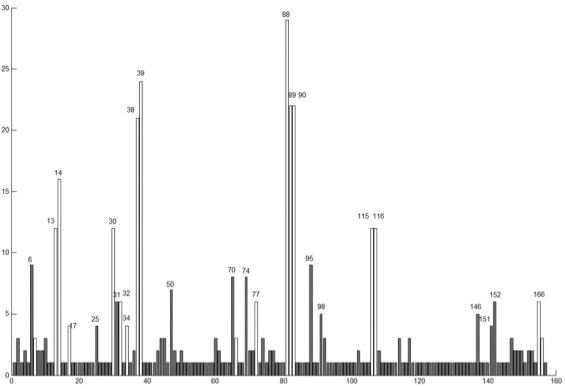
Formule 5:

Collectieve false alarms:

$$Fc = \frac{\sum_2^n O_{tagged}}{O_{totaal}}$$

In deze berekeningen worden dus de tags voor twee of meer deelnemers opgeteld en gedeeld door het totaalaantal zakkenrollers of onschuldige omstanders. Dit kunnen we doen voor elke mogelijke combinatie van deelnemers die hetzelfde tijdsblok hebben bekeken. Elk tijdsblok is door maximaal vier deelnemers beke-

Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003



Figuur 1 ***Totaal aantal tags per persoon gegeven door de tien operators. Witte kolommen betreffen de zakkenrollers, grijze kolommen onschuldige passanten. Nummers boven de kolommen zijn de nummers van personen met meer dan drie tags.***

ken, dus we kunnen per tijdsblok combinaties maken van alle varianten van twee, drie en vier deelnemers.

3 Resultaten

De tien deelnemers gaven in totaal 469 tags. Toepassing van de interval van dertig seconden na een tag, waarin we een volgende tag op dezelfde persoon negeerden voor onze analyses, leverde 44 doublures op die we verwijderden. De resterende 425 tags werden gegeven aan 157 verschillende personen, waaronder alle 18 zakkenrollers. De 139 passanten werden in totaal 196 maal getagd (M = 1.41 tags per passant). Onze 18 zakkenrollers werden in totaal 229 maal getagd (M = 12.72 tags per zakkenroller). Een overzicht van de tags die de tien operators gezamenlijk hebben gegeven, wordt gevisualiseerd in figuur 1. In deze figuur zijn de kolommen die corresponderen met zakkenrollers in wit weergegeven en de kolommen die corresponderen met getagd onschuldig publiek grijs.

Het centrale dilemma (bij hoeveel tags verkrijgen we de optimale verhouding *hits* en *false alarms*) wordt in deze figuur onmiddellijk duidelijk door de grote verscheidenheid aan tags die wordt gegeven aan zowel bekende zakkenrollers als onschuldige passanten. De meeste zakkenrollers (personen 13, 14, 30, 38, 39, 88, 89, 90, 115 en 116; N=10) ontvangen meer tags dan de meest getagde onschuldige omstanders. Andere zakkenrollers (personen 7, 17, 32, 34, 71, 77, 166 en 167; N=8) worden echter juist minder gezien dan de meest getagde onschuldige omstanders. Die meest getagde onschuldige omstanders vormt een aanzienlijke groep (personen 6, 31, 50, 70, 74, 95, 98, 146, 151 en 152; N=10) die evenveel of meer tags ontvangen dan de groep van weinig getagde zakkenrollers. Met andere

woorden, de groep weinig getagde zakkenrollers dreigt te ontkomen, terwijl een groep veel getagde onschuldigen onterecht als verdachte gezien wordt. De vraag die hieruit voortvloeit, is bij hoeveel tags het beste onderscheid (de hoogste sensitiviteit) tussen zakkenrollers en onschuldige omstanders wordt bereikt.

4 Sensitiviteit

In dit onderzoek staat de vraag centraal of het combineren van tags van verschillende operators leidt tot betere oordelen. Daarvoor vergelijken we de gemiddelde A' over alle individuele deelnemers met de A's van de verschillende groepsgroottes (twee, drie of vier deelnemers binnen één tijdsblok). Die vergelijking maken we per niveau van criterium (liberaal/conservatief), ofwel het aantal tags dat nodig is om een individu als verdachte aan te merken. In tabel 2 staan deze gemiddelde A's. Een 4 (groepsgrootte) x 4 (criteria)-variantieanalyse (ANOVA) op A' laat een hoofdeffect zien van groepsgrootte, $F(3, 170) = 12.81$, $p < .001$. Dit betekent dat de gemiddelde A' toeneemt naarmate de groep operators groter is. Met andere woorden, als de tags van meer operators gecombineerd worden, wordt de prestatie beter (een hogere A', veroorzaakt door een hoge *hit*-ratio en/of een lage *false alarm*-ratio). Een Tukey Multiple Comparison-test laat voorts zien dat de gemiddelde prestatie (A') van individuele deelnemers lager is dan de prestatie van een groep van twee operators ($t(170) = -3.00$, $p < .02$), van 3 ($t(170) = -5.57$, $p < .001$), of een groep van vier operators ($t(170) = -4.07$, $p < .001$). De gemiddelde prestatie van een groep van twee operators is lager dan die van een groep van drie ($t(170) = -3.71$, $p < .001$) en vier operators ($t(170) = -2.66$, $p < .05$), maar de gemiddelde prestatie van een groep van drie operators is niet lager dan van een groep van vier operators ($t(170) = -0.83$, *ns*). Dit patroon geldt voor alle criteria. Deze data laten dus zien dat de overall sensitiviteit toeneemt wanneer meerdere observaties worden gecombineerd.

De 4 (groepsgrootte) x 4 (criteria)-variantieanalyse (ANOVA) op A' laat ook een hoofdeffect zien van criterium $F(3, 170) = 10.26$, $p < .001$. Dit hoofdeffect betekent dat de gemiddelde A' afneemt naarmate het criterium conservatiever wordt. Met andere woorden, naarmate meer tags gebruikt worden om een individu als verdachte aan te merken (het criterium), wordt de prestatie minder (een lagere A', veroorzaakt door een lage *hit*-ratio en/of een hoge *false alarm*-ratio). Een Tukey Multiple Comparison-test laat hier zien dat de gemiddelde prestatie (A') bij een criterium van één niet hoger is dan de gemiddelde prestatie bij een criterium van twee ($t(170) = 1.33$, *ns*), maar wel hoger dan de gemiddelde prestatie bij een criterium van drie ($t(170) = 4.27$, $p < .001$) of een criterium van vier ($t(170) = 4.61$, $p < .001$). De gemiddelde prestatie bij een criterium van twee is hoger dan bij een criterium van drie ($t(170) = 2.95$ $p < .02$) en een criterium van vier ($t(170) = 3.33$, $p < .01$), maar de gemiddelde prestatie bij een criterium van drie is weer niet hoger dan bij een criterium van vier ($t(170) = 0.46$, *ns*). Deze data laten dus zien dat de prestatie (A') toeneemt wanneer een lager criterium gehanteerd wordt.

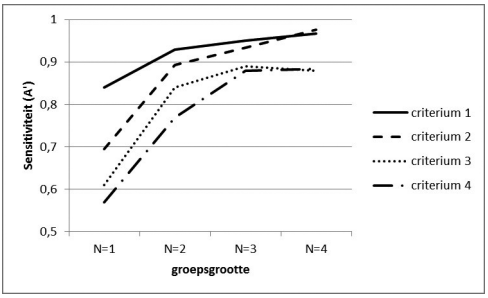
Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003

Tabel 2 Gemiddelde sensitiviteit (A') per criterium en groepsgrootte

Criterium	N=1	N=2	N=3	N=4
1	0,8396	0,9289	0,9497	0,9670
2	0,6944	0,8924	0,9334	0,9767
3	0,6093	0,8405	0,8898	0,8799
4	0,5685	0,7688	0,8789	0,8838

Met andere woorden, de prestatie wordt beter naarmate een lagere grens (minder tags) wordt gehanteerd om een individu als verdachte aan te merken.

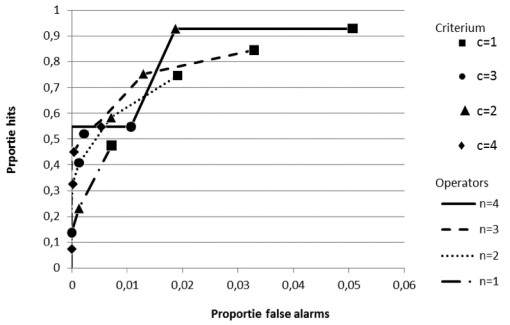
De ANOVA laat geen statistisch significant interactie-effect zien tussen criterium en groepsgrootte op A', (3, 170) = 1.49, $p < .16$. Dat betekent dat er geen wisselwerking aantoonbaar is tussen de factoren. Met andere woorden, in onze data is de invloed van het criterium op A' bij de vier groepsgroottes gelijk, en omgekeerd is de invloed van de groepsgrootte op A' bij de verschillende criteria gelijk.



Figuur 2 Sensitiviteit (A') per criterium en groepsgrootte

5 ROC-curve

Om meer inzicht te krijgen in de individuele en gecombineerde deelnemersprestaties analyseren we de samenhang tussen de proportie *hits* en de proportie *false alarms*, de bouwstenen van A'. We doen dit door middel van een ROC-curve waarin per groepsgrootte en bij verschillende criteria de *hits* worden afgezet tegen de *false alarms*. De ROC-curve wordt vaak gebruikt bij kosten-batenanalyses waarbij een optimaal model moet worden gekozen. Hiervoor wordt het punt in de grafiek bepaald dat het dichtst in de linkerbovenhoek ligt (waarvoor geldt: *hit* = 1, *false alarm* = 0), of dat volgens andere overwegingen een optimale verhouding laat zien. Figuur 2 laat de ROC-curve zien van de hits afgezet tegen de *false alarms*. Uit deze figuur is af te lezen dat met de huidige data en de aanname van gelijke kosten voor *hits* en *false alarms* het punt dat het dichtst bij de linkerbovenhoek van de grafiek ligt, bereikt wordt met een groepsgrootte van drie of vier operators en een criterium meer dan twee tags (c=2; individuen die meer dan twee tags krijgen,



Figuur 3 ***ROC-curve (proportie hits vs. proportie false alarms) voor alle operators***

worden als verdachte aangemerkt). Maar de hoge *hit*-ratio die deze groepsgrootte en dit criterium opleveren, gaat ook gepaard met een hogere *false alarm*-ratio dan wanneer een lagere groepsgrootte of een lager criterium gekozen zou worden. Wat het optimale criterium is, is daarom niet alleen het resultaat van het zoeken naar het optimale nettoresultaat, maar is ook een strategische keuze. Een belangrijk deel van die strategische keuze is het gewicht dat *hits* en *false alarms* krijgen om ze in balans te brengen.

6 Conclusie en discussie

De huidige studie laat zien dat het combineren van observaties van verschillende toezichthouders leidt tot een toename van sensitiviteit in het onderscheiden van zakkenrollers van onschuldige passanten. Hiertoe lieten we ervaren beveiligers videobeelden zien van toezichtcamera's op een winkelcentrum en vroegen we aan hen om mensen die zich afwijkend gedroegen te markeren. We analyseerden vervolgens individuele prestaties en prestaties van combinaties van tags van twee, drie of vier deelnemers. De analyses laten een hoofdeffect zien van groepsgrootte (aantal gecombineerde operators) op sensitiviteit. Concreet leidt een grotere groep tot een hogere sensitiviteit. Ook laten de analyses een hoofdeffect zien van criterium (aantal tags), waarbij een lager criterium geassocieerd wordt met betere prestaties. Er wordt dan bijvoorbeeld al bij twee tags ingegrepen in plaats van bij vier tags. Dit onderzoek laat dus zien dat naarmate de oordelen van meer toezichthouders gecombineerd werden, de gecombineerde oordelen betere inschattingen gaven van de schuld of onschuld van mensen in het gebied. Belangrijk is hier dat dit niet alleen betekent dat er meer zakkenrollers werden herkend (*hits*) of dat minder onschuldige passanten verdacht werden (*false alarms*), maar dat deze twee gezamenlijk in één maat een beter resultaat gaven.

Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003

Een ROC-curve laat zien dat voor de huidige data een optimaal punt ligt bij de combinatie van drie of vier operators, waarbij een criterium van twee tags wordt gehanteerd. Onder die omstandigheden kunnen zakkenrollers het meest accuraat van overig publiek worden onderscheiden. De ROC-curve is echter vooral een handig middel om inzichtelijk te maken hoe *hits* en *false alarms* samenhangen onder wisselende condities van groepsgrootte en criterium. Dit geeft beleidsmakers de mogelijkheid om een strategische keuze te maken om de kans op *hits* te vergroten, met een bijbehorende toenemende kans op *false alarms*, of juist om de kans op *false alarms* te verlagen met een bijbehorende afnemende kans op *hits*. De strategische keuze voor een criterium en groepsgrootte wordt daarbij voornamelijk bepaald door twee factoren. De eerste is het niveau van dreiging en de tweede is het belang van bescherming.

Het niveau van dreiging betreft de specifieke dreiging waartegen het publiek moet worden beschermd, zoals zakkenrollerij of een terroristische aanslag. In het eerste geval volstaat een hoger criterium (met als gevolg een kleinere kans op *hits* én *false alarms*) dan in het tweede geval.

Het belang van bescherming betreft het potentiële doel van een dreiging, zoals winkelend publiek of een bezoek van een hoogwaardigheidsbekleder. Het risico (het product van kans en effect) houdt sterk verband met deze twee overwegingen. Als bijvoorbeeld bekend is dat in een winkelcentrum veel zakkenrollers actief zijn, dan kan het verstandig zijn om het criterium (tijdelijk) te verlagen (dus sneller ingrijpen), zodat de kans op *hits* toeneemt, ook al betekent dat ook een toegenomen kans op *false alarms*.

Naast sensitiviteit als algemene prestatiemaat zal in de praktijk dus ook moeten worden gekeken naar de *hit*-ratio en *false alarm*-ratio, de bouwstenen van sensitiviteit. De dreiging die geldt en de bescherming die nodig is, zullen namelijk bepalen hoe sterk de invloed van de een ten opzichte van de ander moet zijn. Daarbij geldt dat wanneer één van de waarden (*hit* of *false alarm*) vastgezet wordt, de ander ook vast komt te staan. Beleidsmakers of professionals kunnen zo dus strategische keuzes maken over de kansen om zakkenrollers of andere vormen van criminaliteit te voorkomen tegen welke kosten (in casu onterechte staandehoudingen). Bovendien kunnen ze daarbij overwegen om meerdere operators in te zetten om de algemene prestatie (sensitiviteit) te verhogen of om een van de bouwstenen (*hit*- en *false alarm*-ratio's) daarvan te versterken.

Tegelijkertijd kan inspectie van de *false alarms* informatie geven over mogelijke vooringenomenheid van de deelnemers. Sommige gedragingen of uiterlijke kenmerken kunnen voor deelnemers onterecht een aanleiding zijn voor een verdenking. Door te onderzoeken op welke kenmerken veel getagde onschuldigen verschillen van andere onschuldigen (bijvoorbeeld geslacht of kledingstijl) kunnen deze gedragingen of uiterlijke kenmerken aan het licht komen en kunnen operators extra getraind worden om hen bewust te maken van die vooringenomenheid en zo discriminatie op uiterlijke kenmerken te voorkomen.

Een systeem zoals beschreven in het huidige artikel berust op een aantal aannames. Om te beginnen aannames met betrekking tot de manier van taggen. In onze instructies aan deelnemers hadden we gevraagd om reeds bij een geringe verdenking of afwijkend gedrag van een individu een tag te geven. We hebben niet onderzocht hoe deze instructie zich verhoudt tot prestaties na andere instructies, zoals een instructie om terughoudend te zijn met tags.
Voorts kan het zo zijn dat individuen verschillende afwijkende gedragingen verspreid over de tijd hebben laten zien die indicaties waren van kwaadwillige intenties. Deelnemers hebben mogelijk om die reden soms een individu vaker dan eens een tag gegeven. Daarbij spelen verschillende potentiële problemen. Het belangrijkste probleem is dat we in het huidige systeem uitgaan van onafhankelijkheid van observaties. Die gold wel tussen de deelnemers, maar niet binnen een deelnemer. Met andere woorden, een individu dat een tag heeft gekregen omdat hij of zij afwijkend gedrag vertoonde en om die reden langer werd gevolgd door de deelnemer, loopt een grotere kans om nog een tag te krijgen. Als gevolg daarvan krijgen die tags in werkelijkheid een verschillend gewicht (immers, de eerste tag verklaart deels waarom het individu een tweede tag kreeg), terwijl dat niet tot uitdrukking komt in de data. Daar komt bij dat er andere redenen kunnen zijn waarom deelnemers vaker dan eens eenzelfde individu hebben getagd waarvan we niet op de hoogte zijn. Misschien was het een uitdrukking van de zekerheid waarmee deelnemers dachten te weten dat het betreffende individu kwaadwillige intenties had of misschien dacht de deelnemer dat zijn of haar eerdere tag niet goed was geregistreerd. Dit probleem kan deels worden verholpen door duidelijke afspraken en deels vereist dit nader onderzoek naar de waarde van opeenvolgende tags.
Een vooralsnog onopgelost methodologisch probleem heeft te maken met de grote invloed van *correct rejections*, ofwel de terecht niet als verdacht aangemerkte onschuldige passanten. In onze studie vonden we over het algemeen erg hoge A's. Een belangrijke oorzaak hiervan is dat er veel onschuldige passanten zijn die terecht niet getagd zijn. De proportie *false alarms* is daardoor erg laag. Het grote aantal onschuldige passanten legt dus veel gewicht in de schaal en is verantwoordelijk voor de totstandkoming van de hoge A'. Bovendien zorgt de constantheid van de lage *false alarm*-ratio ervoor dat fluctuaties in A' voornamelijk toe te schrijven zijn aan de invloed van de proportie *hits*. Naarmate groepen groter worden en meer deelnemers in de groep tags kunnen geven, stijgt logischerwijs ook de proportie *hits* en dus A'. Weliswaar geeft de maat A' zoals gebruikt in dit artikel de meest objectieve inschatting van succes in het onderscheiden van zakkenrollers van winkelend publiek, maar het is onwenselijk dat door de constante, lage *false alarm*-ratio fluctuaties in A' bijna uitsluitend door veranderingen in de *hit*-ratio worden bepaald. Vervolgonderzoek zou in moeten gaan op alternatieve methoden om evenwicht te brengen in de invloed van *hit*- en *false alarm*-ratio's, terwijl een objectief beeld van sensitiviteit behouden blijft.
Daaraan gerelateerd is de vraag hoe het totaal van aanwezige zakkenrollers en totaal winkelend publiek moet worden gemeten. In dit onderzoek waren zakkenrollers en omstanders niet allemaal even lang in beeld. Onze zakkenrollers liepen een uur lang door het winkelcentrum en waren dus per cameraset die de helft van

Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003

het winkelcentrum besloeg bij benadering dertig minuten te zien. Overig winkelend publiek was doorgaans korter te zien. Dat verschil hebben we in dit onderzoek niet meegenomen, omdat we niet weten hoe lang exact winkelend publiek gemiddeld te zien was. Zou dat verschil wel meegenomen kunnen worden, bijvoorbeeld door te rekenen met het product van tijdseenheden maal aantal zakkenrollers of onschuldigen, dan wordt het relatieve verschil tussen tijd x aantal zakkenrollers en tijd x aantal onschuldigen kleiner en zal dat resulteren in een relatief kleinere *bias* als gevolg van de grotere aanwezigheid van onschuldige passanten.

In dit onderzoek stond de vraag centraal of onder geïsoleerde, optimale omstandigheden het combineren van tags leidt tot betere prestaties. Daarbij zijn de prestaties van de verschillende combinaties van twee, drie of vier operators versus individuele operators vergeleken. Als onderdeel van die geïsoleerde, optimale omstandigheden gebruikten we acteurs als daders en slachtoffers. Hoewel we deze zakkenrollers en slachtoffers zo natuurlijk mogelijk gedrag hebben laten vertonen, is het mogelijk dat ze niet volkomen levensecht gedrag hebben vertoond dat past bij hun respectievelijke rollen. Bovendien hebben we de werkelijkheid overdreven door in een halve dag tijd achttien incidenten te laten plaatsvinden, terwijl dat er normaal slechts enkele zijn. De belangrijkste reden voor de inzet van acteurs was dat alleen op die manier de *hit*-ratio (aangewezen zakkenrollers gedeeld door het totaalaantal zakkenrollers) te berekenen is. Zonder kennis over die *ground-truth* moet daarvan een schatting worden gemaakt. Nu we succesvol de fundamentele onderzoeksvraag hebben beantwoord of het combineren van tags van toezichthouders effectiever is dan het niet delen en combineren van tags, is de vervolgstap om te testen of de bevindingen standhouden in een operationele omgeving.

Literatuur

Adam, B., U. Beck & J. van Loon (2000) *The Risk Society and Beyond: Critical Issues for Social Theory*. London: Sage.

Batchelder, W.H. & A.K. Romney (1986) The statistical analysis of a general Condorcet model for dichotomous choice situations. In: B. Grofman & G. Owen (Eds.), *Decision Research, 2*, 103-112. Greenwich, CT: JAI Press.

Bouma, H., J. Vogels, O. Aarts, C. Kruszynski, R. Wijn & G.J. Burghouts (2013) Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators. *Proc. SPIE, 8745*.

Elias, B. (2009) *Airport passenger screening: Background and issues for Congress* (No. R40543), Congressional Research Service.

Kleij, R. van der, M. Roelofs & D.A. van Hemert (2014) Gaan veiligheidsmaatregelen ten koste van de servicebeleving?. *Tijdschrift voor Veiligheid, 4*, 3-19.

Minister van Justitie, Handelingen II 2003/04, nr. 33, p. 3338.

Pel, B. van, B. Verhagen & R. Wijn (2012) Predictive profiling of proactief beveiligen: Security questioning & prikkelen. *Security Management, 9*, 40-43.

Rest, J.H.C. van, M.L. Roelofs & A.M. van Nunen (2014) *Afwijkend Gedrag: Maatschappelijk verantwoord waarnemen van gedrag in context van veiligheid* (2de herziene druk). Delft: TNO.

Sorkin, R.D., C.J. Hays & R. West (2001). Signal-detection analysis of group decision making. *Psychological Review, 108*, 183-203.

Sorkin, R.D., R. West & D.E. Robinson (1998) Group performance depends on the majority rule. *Psychological Science, 9*, 456-463.

Stanislaw, H. & N. Todorov (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149.

Swets, J.A., W.P. Tanner & T.G. Birdsall (1961) Decision processes in perception. *Psychological Review, 68*, 301-340.

Wijn, R., J.H.C. van Rest, M. Lousberg & G.J. Burghouts (2012) Naar een beter begrip van afwijkend gedrag: Herkenning door mens en computer. In: E.R. Muller (Ed.), *Veiligheid: Veiligheid en Veiligheidsbeleid in Nederland* (565-587). Deventer: Kluwer.

Tijdschrift voor Veiligheid 2015 (14) 1
doi: 10.5553/TvV/187279482015014001003