Adrian Schwaninger

Objekterkennung und Signaldetektion: Anwendungen in der Praxis

1. Zusammenfassung

Beim Menschen ist der Sehsinn der differenzierteste aller Sinne. Über zwei Drittel des Cortex dienen der Verarbeitung visueller Information. Die schnelle und zuverlässige Erkennung von Objekten und Gesichtern spielt dabei eine zentrale Rolle. Durch eine interdisziplinäre Vorgehensweise wurde es in den letzten Jahrzenten möglich, die Erkennung und Detektion von Objekten besser zu verstehen und psychophysisch plausible Modelle zu entwickeln. Im Folgenden werden zunächst die wichtigsten Prozesse und Repräsentationen dargestellt, welche von Mensch und Maschine für die Erkennung von Objekten unter verschiedenen Wahrnehmungsbedingungen eingesetzt werden können (Kapitel 2). Anschliessend wird in Kapitel 3 die Signaldetektionstheorie (SDT) besprochen, welche interdisziplinär angewandte Methoden zur Messung von Detektions- und Erkennungsprozessen zur Verfügung stellt. In Kapitel 4 wird anhand ausgewählter Beispiele illustriert, wie theoretische Ansätze der Objekterkennung und psychophysische Methoden der SDT angewendet werden können. Dabei wird am Beispiel der Gepäckkontrolle an Flughäfen veranschaulicht, wie die SDT zur Messung der Erkennungsleistung verbotener Gegenstände in Röntgenbildern angewendet werden kann. Am Beispiel der Gesichtserkennung wird gezeigt, wie mittels SDT verschiedene Computeralgorithmen mit der menschlichen Erkennungsleistung verglichen werden können.

2. Theorien der Objekterkennung

Obwohl wir im Alltag das Gefühl haben, die Objekte der Umwelt innerhalb von Sekundenbruchteilen problemlos zu erkennen, handelt es sich dabei um eine komplexe Leistung unseres Gehirns. Dies lässt sich bereits daran zeigen, dass ein direkter Vergleich zwischen einem gespeicherten Bild in einem Gedächtnisspeicher und der Abbildung eines Objektes im Auge oder in einer Kamera selten zu einer hohen Übereinstimmung führen würde. Objekte können unterschiedlich weit vom Betrachter entfernt sein und sich an verschiedenen horizontalen und vertikalen Positionen im Gesichtsfeld befinden. Prinzipiell können Objekte um drei Achsen rotiert sein. Oft wird ein Objekt durch ein anderes verdeckt und manchmal sind die Wahrnehmungsbedin-

gungen suboptimal, so dass die Konturen eines Objektes nur unterbrochen sichtbar sind (z.B. bei Schneefall). Erschwerend kommt hinzu, dass jede Objektklasse verschiedene Exemplare enthält, welche unterschiedliche Formmerkmale aufweisen. All dies sind Beispiele dafür, dass robuste Objekterkennung viel mehr beinhaltet als der einfache bildhafte Vergleich von Stimulusbild und Gedächtnisbild.
Prinzipiell können vier verschiedene Ansätze unterschieden werden, welche zur Lösung dieser Probleme bei der Objekterkennung vorgeschlagen worden sind. Ein erster Ansatz besteht darin, relativ invariante Eigenschaften wie z.B. Farbe und lokale Textur oder Parallelität, Gekrümmtheit und gemeinsames Enden von Linien zu erkennen (Ansatz invarianter Eigenschaften). Ein zweiter Ansatz besteht darin, ein Objekt anhand der Teile und ihrer räumlichen Anordnung zu erkennen (Ansatz struktureller Beschreibung). Bei der Erkennung durch Ausrichtung und Transformation wird versucht, die bildhafte Repräsentation des Stimulus und das Gedächtnisbild möglichst in Übereinstimmung zu bringen, um sie dann zu vergleichen. Beim Ansatz multipler Gedächtnisrepräsentationen wird eine robuste Objekterkennung erzielt, indem viele Ansichten des Objektes im Gedächtnis abgespeichert werden. Theorien der Objekterkennung kombinieren in der Regel zwei oder mehr dieser Grundansätze. Im Folgenden wird eine Auswahl der wichtigsten Objekterkennungstheorien möglichst anschaulich dargestellt[29].

2.1 Traditioneller Ansatz nach Marr

David Marr wird als einer der wichtigsten Pioniere im Bereich der Objekterkennung angesehen. Wichtige Stufen visueller Informationsverarbeitung, wie sie Marr (1982) für die Objekterkennung durch Mensch und Maschine postuliert hat, sind in Abbildung 1 dargestellt. Das Input für das visuelle

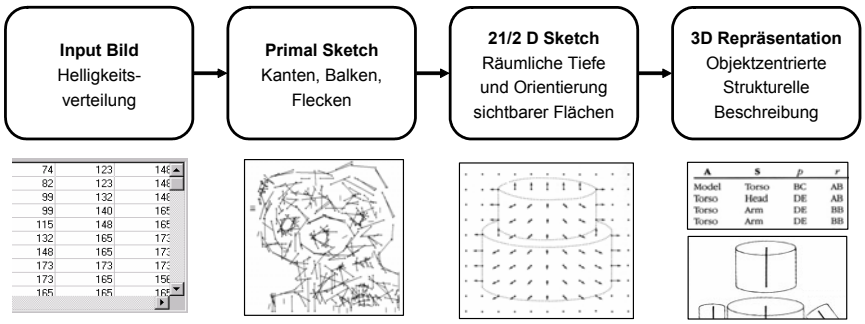

Abbildung 1: Stufen visueller Informationsverarbeitung bei der Objekterkennung im traditionellen Ansatz nach Marr (1982).

System kann vereinfachend als eine Vielzahl von Punkten mit unterschiedlichen Helligkeitswerten beschrieben werden. Nach Marr besteht nun der erste Schritt darin, in dieser Vielzahl von Punkten Kanten (edges) zu detektieren. Parallele Kanten werden zu Balken zusammengefasst (bars) und Balken einer definierten Endung werden zu Flecken (blobs) gruppiert. Durch diese Verarbeitungsschritte gelangt man vom Inputbild zur sogenannten Primärskizze (primal sketch). Für die 2.5 D Skizze muss nun als nächstes die räumliche Tiefe und Orientierung sichtbarer Flächen berechnet werden. Dazu werden von Marr verschiedene Informationsquellen vorgeschlagen, wie z.B. Querdisparation, Bewegungsparallaxe oder Helligkeits- und Texturgradienten. Als nächster Schritt wird eine 3D Repräsentation erstellt, welche ein Objekt anhand seiner Teile und ihrer räumlichen Anordnung definiert (strukturelle Beschreibung). Diese Repräsentation weist drei wichtige Merkmale auf: Sie ist objektzentriert, hierarchisch und modular. Im Gegensatz zu einer Beobachter zentrierten (viewer-centred) Repräsentation, welche das Bild in Bezug auf eine bestimmte Ansicht vom Betrachter beschreibt, ist eine objektzentrierte Repräsentation unabhängig vom Ort des Betrachters. Dabei werden alle Eigenschaften und Teile des Objektes in Relation zur Hauptachse des Objektes beschrieben. Dies ist in Abbildung 2 am Beispiel eines menschlichen Körpers dargestellt. Auf der ersten Beschreibungsstufe wird die Hauptachse des Körpers definiert. Der Körper besteht aus Rumpf, Armen und Beinen, deren mögliche Positionen im Bezug auf die Hauptachse definiert werden. Diese Teile bestehen wiederum aus Teilen. Beispielsweise unterscheidet man beim Arm den Ober- und Unterarm. Der Unterarm besteht

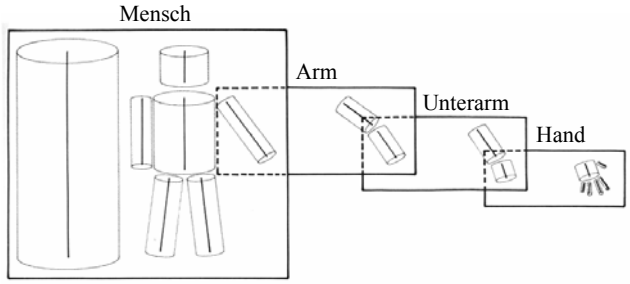

Abbildung 2: Objektzentrierte, modulare und hierarchische 3D Repräsentation (nach Marr, 1982).

aus dem eigentlichen Unterarm und der Hand. Diese wiederum enthält die fünf Finger, welche selbst wieder aus Teilen bestehen. Die 3 D Repräsentation nach Marr ist modular und hierarchisch organisiert, indem die Positionen

[29] Für eine ausführlichere Darstellung und Diskussion siehe Bülthoff, Edelman und Tarr (1995); Edelman (1999); Jolicoeur und Humphrey (1998); Kosslyn (1994); Tarr und Bülthoff (1998)

der Teile immer im Bezug auf die übergeordnete Hauptachse beschrieben werden.
Das Erkennen von Objekten im Rahmen der Theorie von Marr kann man sich folgendermassen vorstellen. Im Gedächtnis ist für jede Objektklasse gespeichert, aus welchen Teilen sie bestehen und wie die Teile angeordnet sind. Objekte in der Aussenwelt werden erkannt, indem die verschiedenen Verarbeitungsstufen von der Extraktion der Kanten bis zur Berechnung der 3D Repräsentation durchlaufen werden (Abbildung 1). Sobald eine strukturelle Beschreibung des Objektes der Aussenwelt vorliegt, welche die Teile und ihre räumliche Anordnung spezifiziert, wird im Gedächtnis nachgeschaut, welche Objektklasse die gleiche strukturelle Beschreibung aufweist. Das Ausmass an Übereinstimmung bestimmt dann, ob ein Objekt erkannt wird. Wie man in der Abbildung 2 sieht, genügen einfache Zylinder, um Arme, Beine, ja sogar den Kopf darzustellen. Aufgrund dieser Beobachtung ist Marr zum Schluss gekommen, dass eine begrenzte Anzahl einfacher volumetrischer Primitive ausreicht, um die verschiedenen Objektklassen anhand der Teile und ihrer räumlichen Relationen zu beschreiben. Diese Idee wurde nach dem Tod von Marr in der Theorie von Biederman umgesetzt, welche als nächstes dargestellt wird.

2.2 Recognition by components (RBC)

Grundlegend für die Theorie von Biederman ist die Beobachtung, dass wenige elementare Teilkörper ausreichen, um viele Objektklassen des Alltags zu beschreiben. Biederman (1987) nennt solche elementaren Teilkörper Geone (geometrical ions). Wie man der Abbildung 3 entnehmen kann, sind verschiedene Alltagsobjekte durch zwei bis drei Geone beschreibbar. Was sich

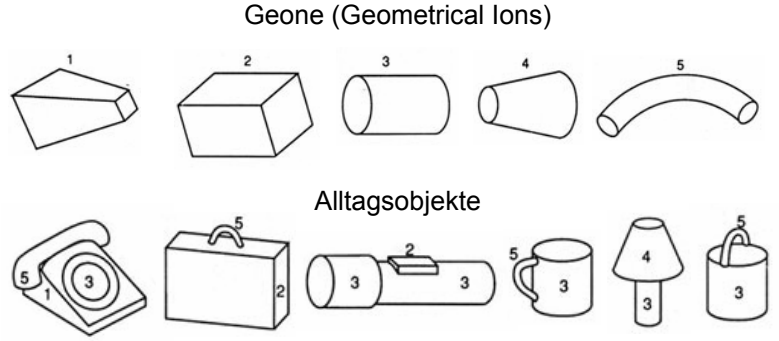

Abbildung 3: Wenige elementare Teilkörper genügen um die meisten Alltagsobjekte zu beschreiben (nach Biederman, 1995).

pro Objekt verändert ist die räumliche Anordnung der Geone und ihre Attribute wie z.B. die Orientierung oder das Verhältnis zwischen der Länge der

Hauptachse und dem Querschnitt. Um eine Erkennung unabhängig von der Grösse, Position und Orientierung zu erreichen, greift Biederman auf relativ invariante Eigenschaften zurück, welche von Lowe (1985) als nicht zufällige Merkmale (nonaccidental properties, NAPs) beschrieben worden sind. Solche Merkmale sind z.B. Parallelität, Gekrümmtheit oder die Art wie Konturen in einem Punkt enden (vertices). Sie bleiben weitgehend erhalten auch wenn sich die Grösse, Position oder Orientierung eines Objektes verändert. Die Geone von Biederman werden definiert durch das Vorhandensein und die Kombination von solchen invarianten Eigenschaften. In der RBC Theorie werden wie bei Marr als erstes Kanten und Linien extrahiert. Aus der Linienrepräsentation wird anschliessend versucht, die oben erwähnten invarianten Eigenschaften (NAPs) zu extrahieren, welche die Geone definieren. Danach wird die räumliche Anordnung der Geone bestimmt. Diese strukturelle Beschreibung der Teile (Geone) und ihrer räumlichen Relationen wird ähnlich wie bei Marr mit den gespeicherten strukturellen Beschreibungen im Gedächtnis verglichen. Findet sich eine genügend grosse Übereinstimmung, dann wird das Objekt erkannt.
Bei der RBC Theorie wird Erkennung mittels invarianter Eigenschaften und Erkennung durch strukturelle Beschreibung kombiniert. Eine leicht modifizierte Version wurde in einem konnektionistischen neuronalen Netz von Hummel und Biederman (1992) implementiert. Dieses Computerprogramm kann einfache Objekte erkennen, wenn sie als Linienzeichnung dargeboten werden, welche aus zwei Geonen bestehen.

2.3 Erkennung durch Ausrichtung und Transformation

Das SCERPO Vision System von Lowe (1985, 1987) ist eines der ersten Computermodelle, welche Objekte in Fotos erkennen kann. Es eignet sich

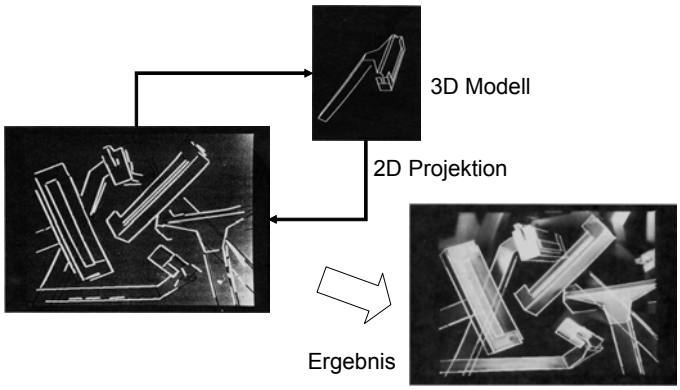

Abbildung 4 Illustration zum wissensbasierten Erkennungssystem nach Lowe (1987).

gut, um das Prinzip der Erkennung durch Ausrichtung und Transformation von 3D Repräsentationen zu erklären (Abbildung 4). Zunächst werden Linien extrahiert und gruppiert nach Gesetzmässigkeiten, welche Gestaltgesetzen ähneln. Dabei spielen Nähe, Parallelität, gemeinsames Enden an einem Punkt (vertices), oder Gekrümmtheit eine wichtige Rolle. Diese Eigenschaften sind relativ invariant (NAPs) und werden von Lowe verwendet, um eine bestimmte 3D Repräsentation im Gedächtnis auszuwählen. Letztere wird dann rotiert und verschoben, bis ihre 2D Projektion mit dem Inputbild hinreichend übereinstimmt. Die projizierten Linien können dann verwendet werden, um Konturen zu ergänzen (Abbildung 4). Das Erkennungssystem von Lowe (1985, 1987) kombiniert also zwei Grundansätze: Erkennung mittels Transformation und Ausrichtung, sowie Erkennung mittels invarianter Eigenschaften. Das Verfahren wurde von Lowe auch als wissensbasiert bezeichnet, weil das Erkennungssystem Information über die Dreidimensionalität sowie der erlaubten Transformationen enthalten muss. Anstelle eines 3D Modells, können aber auch mehrere 2D Repräsentationen angenommen werden, welche mit dem Inputbild in Übereinstimmung gebracht werden (siehe z.B. Huttenlocher & Ullman, 1990; für eine Übersicht siehe Graf, 2002). Die massiv höhere Rechenleistung von Computersystemen in den Neunziger Jahren ermöglichte die Verarbeitung der gesamten Bildinformation, ohne ein Inputbild auf die wichtigsten Kanten und Linien zu reduzieren (Ullman, 1996). Dies trifft auch auf das System zu, welches als Beispiel für den nächsten ansichtenbasierten Ansatz dient.

2.4 Erkennung durch Linearkombination von 2D Repräsentationen

Beim Ansatz von Ullman & Basri (1991) werden mehrere Ansichten eines Objektes als detaillierte ganze Bilder im Gedächtnis gespeichert. Diese kön-

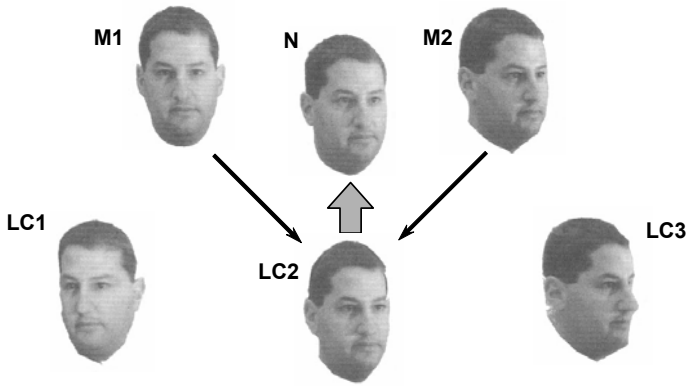

Abbildung 5 Erkennung mittels Linearkombination (nach Ullman, 1996).

nen mittels Linearkombination verrechnet werden, um neuen Ansichten eines Objektes zu bilden. Damit kann ein Objekt auch erkannt werden, wenn es in einer noch nie gesehenen Ansicht erscheint. Ohne auf die mathematischen Details näher einzugehen, sei dies am Beispiel von Gesichtern in Abbildung 5 veranschaulicht. Die Bilder M1 und M2 sind gespeicherte Ansichten. Das Bild N ist eine neue, dem Computersystem unbekannte Ansicht. Aus den Bildern M1 und M2 wurden mittels Linearkombination die Ansichten LC1, LC2 und LC3 berechnet. Wie man sieht, stimmt LC2 ziemlich gut mit dem realen Foto N überein, wodurch die in N abgebildete Person identifiziert werden kann. Mit diesem Verfahren können photorealistische Abbildungen von Objekten zuverlässig erkannt werden, was ein bedeutender Forschritt zu den linienbasierten Ansätzen von Marr und Biederman darstellt. Allerdings stellt sich die Frage, wie viele Ansichten von einem Objekt gespeichert werden müssen, damit alle anderen nicht gespeicherten Ansichten mittels Linearkombination zuverlässig erkannt werden können. Ullman und Basri (1991) konnten mathematisch beweisen, dass unter der Annahme orthographischer Projektion 2-5 gespeicherte Ansichten ausreichen, um ein Objekt in allen möglichen Rotationen und Positionen zu erkennen. Dabei muss das Objekt aber vollständig sichtbar sein. Bei teilweiser Verdeckung zeigten Computersimulationen mit ca. 10 gespeicherten Ansichten relativ gute Ergebnisse (Ullman, 1996).

2.5 Erkennung durch Interpolation von 2D Ansichten

Bei diesem Ansatz wird diejenige Bildübertragungsfunktion gesucht, welche verschiedene Inputbilder eines bestimmten Objektes auf den gleichen Wert überträgt (z.B. 1) und alle anderen Bilder auf einen anderen Wert (z.B. 0). Gesucht wird nun diejenige Funktion, welche neue Ansichten als Interpolation gespeicherter Ansichten darstellen kann. Poggio und Edelman (1990)

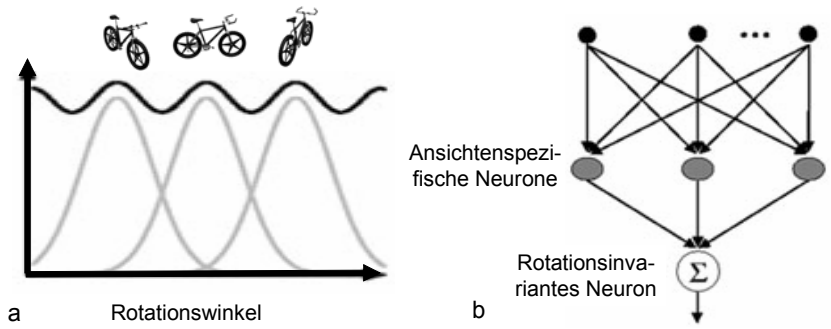

Abbildung 6 Veranschaulichung der Erkennung durch Interpolation von 2D Ansichten (nach Riesenhuber & Poggio, 1999).

konnten zeigen, dass solche Interpolationsfunktionen mit neuronalen Netzwerken gelernt werden können. Das Prinzip ist in Abbildung 6 veranschaulicht. Das neuronale Netz wurde mit mehreren verschiedenen Ansichten trainiert und hat für jede Ansicht eine radiale Basisfunktion (RBF) ausgebildet. Diese Basisfunktionen werden als radial bezeichnet, weil ihre Antwortstärke radial mit zunehmender Rotation (sowie anderen Transformationen) des Objektes abnimmt (in Abbildung 6a sind aus Gründen der Vereinfachung nur drei RBF als Gaussverteilungen zweidimensional dargestellt). Solche RBF kann man sich auch als einzelne Neurone vorstellen, welche auf eine bestimmte Ansicht spezialisiert sind („ansichtenspezifische Neurone"). Soll nun eine neue Ansicht des Objektes erkannt werden, so wird diese mit allen gespeicherten Ansichten verglichen. Die Ergebnisse werden gewichtet aufsummiert (Abbildung 6b), was der Gesamtantwort des neuronalen Netzes entspricht („rotationsinvariantes Neuron"). Die Analogie mit Neuronen ist dabei durchaus gerechtfertigt. Einzelzellableitungen im Inferotemporalcortex von Makaken haben nämlich ergeben, dass zahlreiche Nervenzellen ein Antwortverhalten zeigen, welches solchen ansichtsabhängigen RBF ähnelt (Logothetis, Pauls, & Poggio, 1995). Wie neurophysiologische Erkenntnisse mit Computeralgorithmen verbunden werden können, zeigt eine neueres RBF Modell von Riesenhuber und Poggio (1999). Für eine vertiefende Diskussion des Interpolationsmodells siehe aber auch Edelman (1999).
Im Vergleich zu den anderen ansichtenbasierten Ansätzen ist wichtig zu betonen, dass beim Interpolationsmodell die Erkennung mit dem Abstand von der jeweiligen Basisfunktion radial abnimmt. Bei der Erkennung mittels Linearkombination ist auch ausserhalb der gespeicherten Ansichten eine Erkennung relativ gut möglicht, nur orthogonal liegende Ansichten sind schwierig zu erkennen. Bei der Erkennung mittels Ausrichtung und Transformation müsste dagegen die Erkennung praktisch invariant sein, wenn angenommen wird, dass ein detailliertes 3D Modell gespeichert ist und die Transformation fehlerfrei funktioniert. Im Kapitel 4.2 werden diese Unterschiede wieder aufgegriffen, um die drei ansichtenbasierten Ansätze bezüglich ihrer Plausibilität für die menschliche Gesichtserkennung zu vergleichen.

3. Signaldetektionstheorie

Die Signaldetektionstheorie (SDT) gehört zum Methodeninventar der Psychophysik und wurde von Green und Swets (1966) begründet. Sie wurde in den folgenden Jahrzehnten erheblich weiterentwickelt (siehe MacMillan & Creelman, 1991) und wird sowohl in der Grundlagenforschung als auch in der angewandten Forschung in verschiedensten Bereichen eingesetzt (für eine Übersicht siehe z.B. Swets, 1996). Da die Grundannahmen und das statis-

tische Modell nicht ganz einfach sind, wird der Detektionsprozess zunächst anhand eines Beispiels aus der Praxis veranschaulicht.

3.1 Einführungsbeispiel

Bei Sicherheitskontrollen an Flughäfen müssen Passagiere ihr Handgepäck röntgen lassen, bevor das Flugzeug betreten werden darf. In *Abbildung 7* sind zwei Röntgenbilder abgebildet. Augrund solcher Bilder muss ein Si-

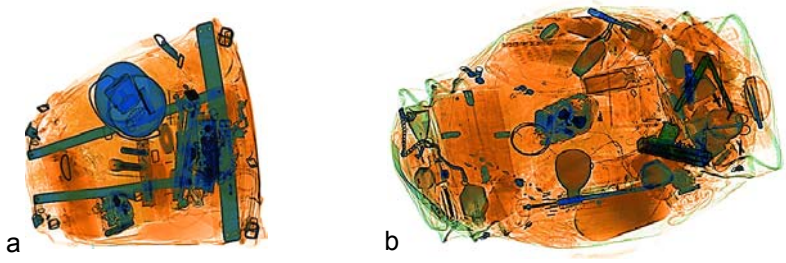

Abbildung 7 Röntgenbilder von zwei Gepäckstücken. a Das Gepäck auf der linken Seite ist ungefährlich. b Im Gepäck rechts ist eine Pistole und ein Messer enthalten.

cherheitsbeauftragter entscheiden, ob das Gepäckstück ungefährlich ist, oder ob Verdacht auf verbotene Gegenstände wie z.B. Schusswaffen oder Messer besteht und deshalb das Gepäck manuell nachkontrolliert werden muss. Intuitiv wird oft angenommen, dass die Erkennungsleistung umso besser ist, je mehr verbotene Gegenstände erkannt werden. Leider ist diese Annahme jedoch nicht unbedingt korrekt, weil jemand eine hohe Trefferrate auch dadurch erzielen kann, indem die meisten Gepäckstücke als „NICHT OK“ eingestuft werden. Dazu ein konkretes Beispiel. Nehmen wir an, zwei Sicherheitsbeauftragte A und B nehmen an einem Test teil, an welchem 200 Röntgenbilder gezeigt werden, wovon genau die Hälfte verbotene Gegenstände (z.B. Schusswaffen und Messer) enthält. Beide Probanden erkennen 90% aller verbotenen Gegenstände (Trefferrate in *Abbildung 8*). Die Trefferrate bezieht sich jedoch nur auf die Hälfte aller im Test gezeigten Röntgenbilder, nämlich diejenigen 100 Bilder, welche tatsächlich verbotene Gegenstände enthielten. Um die Erkennungsleistung valide beurteilen zu können, muss auch das Antwortverhalten für die anderen 100 Testbilder berücksichtigt werden, bei denen kein verbotener Gegenstand enthalten war. Dabei betrachtet man die Anzahl „Fehlalarme“, d.h. diejenigen Fälle, bei denen ein ungefährliches Gepäck als „NICHT OK“ beurteilt worden ist.

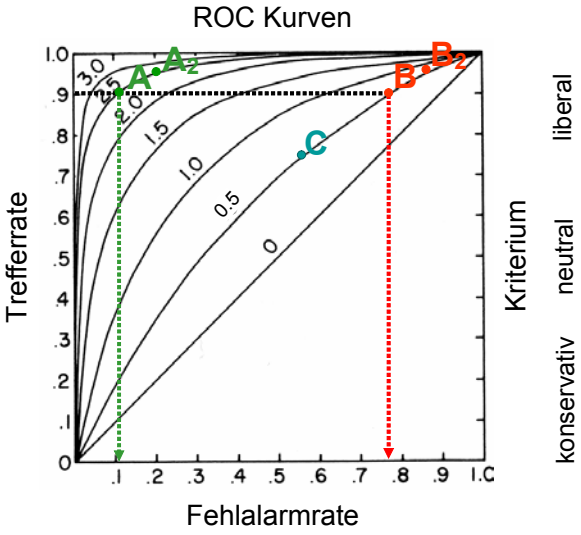

Abbildung 8 ROC Kurven, Sensitivität und Kriterium (nach Schwaninger, 2003a).

Wie man der *Abbildung 8* entnehmen kann, unterscheiden sich dabei die beiden Probanden erheblich. Person B weist eine viel höhere Anzahl Fehlalarme auf als Person A. Dies bedeutet, dass B die hohe Trefferrate von 90% vor allem dadurch erzielt hat, indem die meisten Gepäckstücke – ob gefährlich oder ungefährlich – prinzipiell als „NICHT OK" beurteilt worden sind. Im Arbeitsalltag hätte dies unnötig lange Warteschlangen bei der Sicherheitskontrolle zur Folge. Anders verhält es sich bei Person A. Dieser Sicherheitsbeauftragte ist sehr effizient. Er hat die gleich hohe Trefferrate von 90%, aber eine viel kleinere Anzahl Fehlalarme (11%). Mit anderen Worten: A erkennt sowohl treffsicher wenn ein verbotener Gegenstand enthalten war (hohe Trefferrate), als auch wenn ein Gepäckstück ungefährlich ist (tiefe Fehlalarmrate). Die Funktion, welche die Trefferrate in Abhängigkeit der Fehlalarme darstellt, wird receiver operating characteristic, oder abgekürzt ROC Kurve genannt. Die Diagonale entspricht Rateverhalten, d.h. Anzahl Treffer = Anzahl Fehlalarme. Ein valides Mass der Erkennungsleistung bei Detektionsaufgaben ist die sog. Sensitivität d'. Sie entspricht dem Abstand der Diagonalen zum Wendepunkt der jeweiligen ROC Kurve. Beispielsweise befindet sich Person A in *Abbildung 8* auf der ROC Kurve mit d' = 2.5. Person B hat wegen der hohen Fehlalarmrate eine viel geringere Erkennungsleistung. Diese Person befindet sich auf der ROC Kurve, welche einer Sensitivität von d' = 0.5 entspricht. Der Ort auf einer bestimmten ROC Kurve wird durch das Kriterium bestimmt (Antwort Bias). Es kann sich sehr schnell ändern, weil es von subjektiven Kosten/Nutzen Einschätzungen, Ar-

beitsmotivation und der erwarteten Auftretenswahrscheinlichkeit von Signalen abhängig ist. Beispielsweise hat sich die subjektive Auftretenswahrscheinlichkeit von verbotenen Gegenständen in Gepäck kurz nach dem 11. September 2001 massiv verändert. Natürlich konnte sich die tatsächliche Erkennungsleistung nicht von einem Tag auf den anderen verbessern. Die Sensitivität d' von Person A war nach wie vor viel besser als diejenige von Person B. Was sich sofort änderte war das Kriterium. Die meisten Sicherheitsbeauftragten beurteilten sofort viel häufiger Gepäckstücke als „NICHT OK", wodurch sich plötzlich viel längere Warteschlangen bei den Sicherheitskontrollen ergaben. Diese Verschiebung des Kriteriums ist in *Abbildung 8* veranschaulicht. Sowohl Person A als auch Person B haben sich auf ihrer jeweiligen ROC Kurve in Richtung liberaleres Kriterium verschoben (A2 und B2). Das Kriterium kann auch von Person zu Person erheblich variieren. Beispielsweise hat Person C in *Abbildung 8* ein viel konservativeres Kriterium als Person B und produziert daher viel weniger Fehlalarme. Da die Trefferrate aber auch entsprechend abnimmt, ist die Erkennungsleistung von Person C gleich schlecht wie bei Person B. Beide Personen sind auf der gleichen ROC Kurve, welche einer Sensitivität von d' = 0.5 entspricht.
Wichtig ist festzuhalten, dass die Erkennungsleistung einer Person ein stabileres Merkmal ist, welches sich durch gezieltes Training verändern lässt (siehe z.B. Hofer & Schwaninger, in press; Schwaninger, 2004; Schwaninger & Hofer, 2004). Im Gegensatz dazu kann sich das Kriterium sehr schnell verändern, weil es abhängig ist von subjektiven Kosten / Nutzen Abwägungen, Arbeitsmotivation und der eingeschätzten Auftretenswahrscheinlichkeit.

3.2 Statistisches Modell

Das Erkennen verbotener Gegenstände in Röntgenbildern kann als eine Detektionsaufgabe interpretiert werden. Dabei werden die Begriffe Signal und Rauschen unterschieden. Was detektiert werden soll, wird als Signal bezeichnet. Im vorliegenden Beispiel also verbotene Gegenstände im Röntgen-

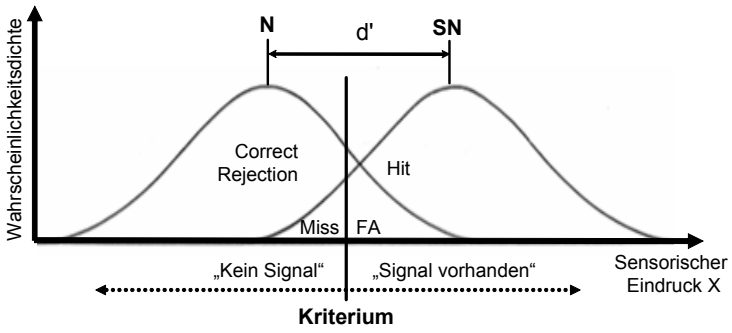

Abbildung 9 Statistisches Modell der Signaldetektionstheorie.

bild des Gepäcks wie z.B. eine Pistole oder ein Messer. Für die Information, welche das Signal umgibt, wird der Begriff Rauschen (noise) verwendet. Im Einführungsbeispiel wäre das also diejenige Information im Röntgenbild, welche von ungefährlichen Gegenständen stammt, d.h. der ungefährliche Gepäckinhalt. Die Aufgabe des Beobachters ist es zu entscheiden, ob es sich beim Stimulus nur um Rauschen (Gepäck ungefährlich) oder um Signal plus Rauschen handelt (z.B. Pistole im Gepäck). Diese Entscheidung basiert auf der Intensität des sensorischen Eindruckes, welcher bei der Verarbeitung des Stimulus entsteht, im vorliegenden Beispiel also beim Betrachten des Röntgenbildes. Erreicht der sensorische Eindruck eine bestimmte Intensität, welche höher als das subjektive Kriterium ist, so entscheidet sich der Betrachter dafür dass ein Signal vorhanden ist (Gepäck „NICHT OK“). Wird das Kriterium unterschritten, so nimmt der Betrachter an, dass kein Signal enthalten ist (Gepäck „OK“). Dabei lassen sich prinzipiell vier Fälle unterscheiden. Ist nur Rauschen vorhanden (Gepäck ungefährlich) und der Betrachter entscheidet sich für „Gepäck OK“, dann spricht man von korrekter Zurückweisung (correct rejection). Lautet die Antwort bei Rauschen jedoch „Gepäck NICHT OK“ so handelt es sich um einen Fehlalarm (false alarm). Ist ein verbotener Gegenstand im Gepäck enthalten (Signal plus Rauschen) und wird das Signal nicht detektiert („Gepäck OK“), so handelt es sich um einen Verpasser (miss). Im anderen Fall („Gepäck NICHT OK“) spricht man von einem Treffer (hit).

Wie bereits erwähnt wurde, kann das Kriterium je nach subjektiven Kosten / Nutzen Einschätzungen, Arbeitsmotivation und vermuteten Auftretenswahrscheinlichkeiten variieren. Ein neutrales Kriterium liegt zwischen der Verteilung von Rauschen und Signal plus Rauschen, ähnlich wie es in der *Abbildung 9* veranschaulicht ist. Bei einem sehr vorsichtigen Betrachter würde das Kriterium weiter links liegen, es sollen ja möglichst keine Signale verpasst werden (liberales Kriterium, vgl. auch *Abbildung 8*). Die höhere Trefferrate bei Signal plus Rauschen wird aber erkauft mit einer höheren Anzahl Fehlalarme wenn nur Rauschen vorhanden war. Ein konservatives Kriterium liegt rechts von der Mitte der beiden Verteilungen. Dies tritt z.B. ein, wenn möglichst keine Fehlalarme gemacht werden sollen, was aber mit einer kleineren Trefferrate einhergeht. Die Sensitivität d’ entspricht dem Abstand zwischen den Mittelwerten der Verteilungen für Rauschen und Signal plus Rauschen. Je grösser dieser Abstand, umso besser kann ein Betrachter zwischen Rauschen und Signal plus Rauschen unterscheiden. Wie man der *Abbildung 9* entnehmen kann, ist die Sensitivität d’ statistisch unabhängig vom Kriterium.

3.3 Berechnung von Sensitivität und Kriterium

Ausgehend von den empirisch ermittelten Treffern und Fehlalarmen lässt sich die Sensitivität d' und das Kriterium einfach mit Hilfe einer Z-Tabelle berechnen. Nehmen wir an, die Fehlalarmrate beträgt 2% (*Abbildung 10* o-

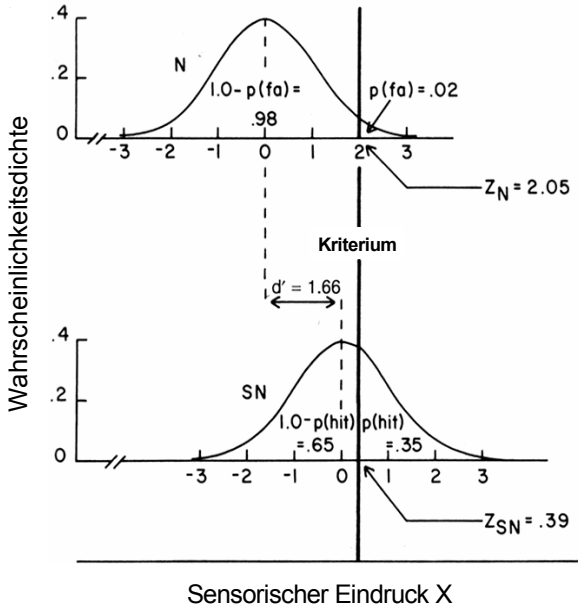

Abbildung 10 Berechnung der Sensitivität d' (nach Gescheider, 1997).

ben). Weil in der Z-Tabelle jeweils die Unterschreitungswahrscheinlickeit angegeben wird, schauen wir den Z-Wert für 1-0.02 nach und erhalten Z_N = 2.05 für den Ort des Kriteriums im Bezug auf die Verteilung von Rauschen (N Verteilung). Ähnlich verfahren wir für die Verteilung von Signal plus Rauschen (SN Verteilung). Dazu brauchen wir die Anzahl Treffer, im vorliegenden Beispiel 35%. Den Z-Wert schaut man wiederum in der Z-Tabelle nach für 1-0.35 und erhält Z_{SN} = 0.39. Die Sensitivität entspricht dem Abstand der beiden Verteilungen und ergibt d' = 2.05-0.39 = 1.66. Zum gleichen Ergebnis kommt man, wenn man Z(Trefferrate) – Z(Fehlalarmrate) berechnet[30]. Für das Kriterium werden verschiedene Masse verwendet. Das be-

[30] Diese Berechnung lässt sich übrigens einfacher und exakter mit dem Tabellenkalkulationsprogramm Microsoft Excel (deutsche Version) mit folgender Formel berechnen: = STANDNORMINV(Trefferrate)-STANDNORMINV(Fehlalarmrate). In unserem Fall ergibt dies d' = STANDNORMINV(0.35)-STANDNORMINV(0.02) =

kannteste Mass ist β; es entspricht der Steigung auf der ROC Kurve (*Abbildung 8*). Da β aber nicht unabhängig von der Sensitivität d' ist (Macmillan & Creelman, 1991), wird oft das Kriteriumsmass C angegeben. Es berechnet sich als $C=0.5(Z_{SN}+Z_N)$. Ein neutrales Kriterium entspricht C=0 und es befindet sich exakt in der Mitte des Abstandes zwischen der Rauschen und Signal plus Rauschen Verteilungen. Bei Werten C>0 liegt das Kriterium weiter rechts (konservatives Kriterium), bei Werten C<0 spricht man von einem eher liberalen Kriterium, welches weiter links liegt (vgl. dazu auch *Abbildung 8*).

3.4 „Nicht parametrische" Masse A' und B''

Die Berechnungen von d' und C setzen voraus, dass Rauschen und Signal plus Rauschen normalverteilt sind und gleiche Varianzen aufweisen. Diese Voraussetzungen lassen sich mittels verschiedener Verfahren überprüfen (Green & Swets, 1966; MacMillan & Creelman, 1991). Ein einfaches Verfahren basiert auf Z-transformierten ROC Werten: Die ROC Kurven werden zu Geraden, welche parallel zur Diagonalen von d' = 0 verlaufen, sofern die Voraussetzungen erfüllt sind. Ist die Annahme der Varianzhomogenität verletzt, so haben die Geraden eine Steigung ungleich 1. Ist dagegen die Annahme der Normalverteilung verletzt, so liegen die Z-transformierten ROC Punkte nicht auf einer Geraden (siehe z.B. Hofer & Schwaninger, in press).
Gibt es Belege für die Verletzung der statistischen Voraussetzungen oder ist die Durchführung einer ROC Kurven Analyse nicht möglich, so wird häufig auch A' verwendent (siehe z.B. Schwaninger, Hardmeier, & Hofer, in press). Dieses Mass wird oft als „nicht parametrisch" oder „verteilungsfrei" bezeichnet, weil die Berechnung keine a priori Annahmen über die zugrundeliegenden Verteilungen braucht. A' entspricht einem Schätzwert für das Integral unter der ROC Kurve und kann Werte zwischen 0.5 und 1 annehmen. Die Formel lautet folgendermassen (H steht für Trefferrate (hit rate) und F für Fehlalarmrate)[31]:

$$A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)]$$

Als „nicht parametrisches" Mass für das Kriterium wird häufig B'' verwendet[32]:

$$B'' = [H(1-H)-F(1-F)]/[H(1-H)+F(1-F)]$$

1.668. Für die englische Version von Microsoft Excel lautet die Formel = NORMSINV(Trefferrate)-NORMSINV(Fehlalarmrate)

[31] Die Gleichung muss allerdings angepasst werden, falls die Fehlalarmrate grösser als die Trefferrate ist (was allerdings nur selten vorkommt). In diesem Fall gilt: A' = 0.5 - [(F-H)(1+F-H)]/[4F(1-H)]

32 Im seltenen Fall F > H gilt jedoch: B'' = [F(1-F)-H(1-H)]/[H(1-H)+ F(1-F)]

Für eine vertieftere Diskussion dieser und anderer Detektionsmasse sowie eine detaillierte Methodenbeschreibung siehe MacMillan und Creelman (1991).

4. Anwendungsbeispiele

In zahlreichen Anwendungsbereichen wird visuelle Information verarbeitet, um bestimmte Objekte zu erkennen. Oft wird ein binärer Entscheid verlangt. Wie in vorangehenden Abschnitten erläutert wurde, ist ein anschauliches Beispiel dafür die Gepäckkontrolle bei Flughäfen. Der Sicherheitsbeauftragte muss für jedes Röntgenbild entscheiden, ob es ungefährlich ist, oder ob es manuell nachkontrolliert werden muss. Auch in der medizinischen Diagnostik bilden Röntgenbilder die Grundlage binärer Entscheide. Bei Verdacht auf Lungen- oder Brustkrebs beispielsweise muss der Arzt entscheiden, ob das Gewebe im Röntgenbild gesund ist, oder ob Gewebeproben genommen werden müssen. Auch bei Materialprüfungen müssen binäre Entscheide gefällt werden. Beispielsweise muss bei regelmässigen Kontrollen am Flugzeug entschieden werden, ob gefährliche Materialermüdungen vorliegen, oder ob mit einem aufwändigen Service noch gewartet werden soll. Dies sind typische Anwendungsbereiche der SDT (für eine Übersicht siehe Swets, 1996).
Im Folgenden wird am Beispiel der Gepäckkontrolle an Flughäfen gezeigt, wie die SDT als nützliches Instrument bei der Untersuchung der Erkennung verbotener Gegenstände eingesetzt werden kann. Anschliessend wird am Beispiel der Gesichtserkennung verdeutlicht, wie die SDT eingesetzt werden kann, um verschiedene Computererkennungsmodelle mit der menschlichen Erkennungsleistung zu vergleichen.

4.1 Sicherheitskontrollen am Flughafen

Wie gut werden verbotene Gegenstände erkannt, die in der Handtasche oder im Koffer über das Rollband des Röntgenprüfgeräts gleiten? Welche gefährlichen Gegenstände sind besonders schwierig zu erkennen? Ist die Erkennung von Waffen leichter als die von Gefahrengütern wie z.B. Gaskartuschen oder Taucherlampen? Die Beantwortung solcher Fragen ist bedeutsam für Qualitätskontrollen und das Abschätzen von Sicherheitsrisiken. Eine zuverlässige Messung der Erkennungsleistung bildet aber auch die Grundlage für eine faire Beurteilung und Auswahl von Sicherheitsbeauftragten. In enger Zusammenarbeit mit der Kantonspolizei Zürich, Flughafenpolizei und mit finanzieller Unterstützung der Flughafen Zürich AG Unique, wurden mehrere Untersuchungen durchgeführt, in denen verschiedene Aspekte der Objekterkennung und Methoden der SDT eine wichtige Rolle spielten (Hofer & Schwaninger, in press; Schwaninger, Hardmeier & Hofer, in press; Schwaninger, 2003a, 2003b, 2004). Im Folgenden werden die wichtigsten

Ergebnisse der Studien zur Erkennung verbotener Gegenstände in Röntgenbildern dargestellt.
Zur Messung der Erkennung verbotener Gegenstände wurden drei Tests eingesetzt. Dazu wurden Bilder verwendet, welche mit VIVID VIS Dual Energy System Röntgengeräten aufgenommen worden sind. Bei allen Tests mussten die Sicherheitsbeauftragten jeweils entscheiden, ob ein Gepäck ungefährlich ist, oder ob Verdacht auf verbotene Gegenstände wie z.B. Schusswaffen, Messer, Gefahrengüter etc. besteht und deshalb das Gepäck manuell nachkontrolliert werden müsste[33].

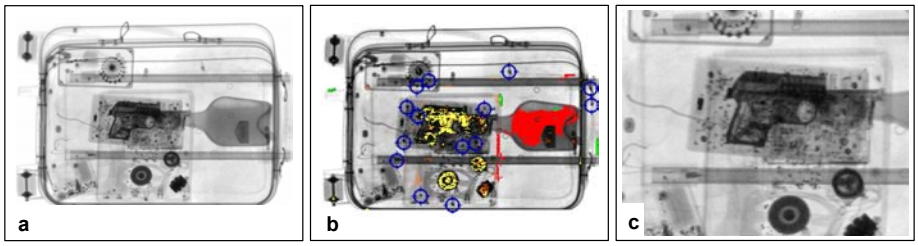

Abbildung 11 Beispielbilder des ersten Tests zur Messung der Erkennung verbotener Gegenstände. a Farbbild, b Schwarzweissbild, c vergrössertes Bild.

Im ersten Test ging es um die Messung der Erkennung verbotener Gegenstände wenn Grundfunktionen wie Vergrösserung und Farbe zur Verfügung stehen. Jedes Röntgenbild konnte maximal 10 Sekunden betrachtet werden. Es konnte jederzeit zwischen Farbbild und Schwarzweissbild gewechselt werden und die Auflösung konnte verdoppelt werden (*Abbildung 11*).

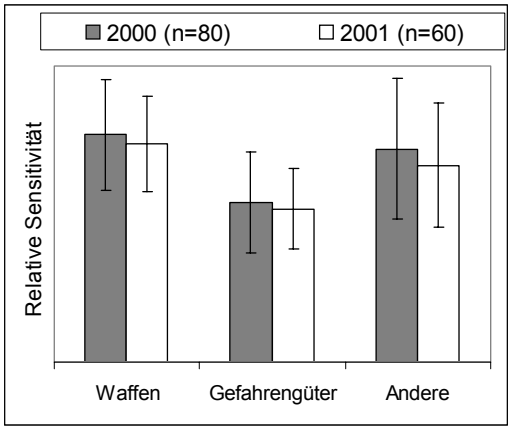

Abbildung 12 Relative Unterschiede in der Erkennungsleistung verschiedener verbotener Gegenstände. Fehlerbalken geben die Standardabweichung an.

[33] Bei allen Tests wurde auch die Erkennung von Bomben untersucht. Im vorliegenden Text sind dazu aber keine Angaben enthalten.

Die Röntgenbilder wurden in zufälliger Reihenfolge am Computer gezeigt und bei jedem Bild musste entschieden werden, ob das Gepäckstück „OK“ ist, oder ob es verbotene Gegenstände enthält und daher „NICHT OK“ ist. Der Test enthielt 154 Röntgenbilder. In einem fünftel aller Gepäckstücke waren Waffen versteckt, 14 Prozent enthielten Gefahrengüter wie z.B. Gaskartuschen, gefüllte Benzinkanister oder Feuerwerk und 10 Prozent enthielten andere verbotene Gegenstände wie z.B. Elfenbein oder ausgestopfte geschützte Tiere. Die verschiedenen relativen Häufigkeiten wurden gewählt, um der unterschiedlichen Auftretenswahrscheinlichkeit im Alltag Rechnung zu tragen. An diesem Test haben insgesamt 140 Sicherheitsbeauftragte der Kantonspolizei Zürich, Flughafenpolizei teilgenommen, 80 Personen im Jahre 2000 und 60 Personen ein Jahr später. Alle Mitarbeiter hatten mindestens 24 Monate Berufserfahrung sowie ein Mindestbeschäftigungsgrad von 70 %. Die Daten wurden mittels SDT ausgewertet (siehe Kapitel 3). Von Hauptinteresse war dabei die Sensitivität d’, als Mass dafür, wie gut die unterschiedlichen Kategorien von verbotenen Gegenständen erkannt worden sind. Wie man der *Abbildung 12* entnehmen kann, wurden Gefahrengüter schlechter erkannt als Waffen und andere verbotene Gegenstände. Dieser Befund zeigte sich in beiden Erhebungen (2000 und 2001), welche sehr ähnliche Daten ergaben. Interessant ist auch die beachtliche Streuung zwischen den Versuchspersonen, welche auf erhebliche interindividuelle Unterschiede in der Erkennungsleistung hinweist. Dies könnte auf unterschiedlicher Lernerfahrung oder auf Unterschiede in eher stabilen Eigenschaften visueller Informationsverarbeitung zurückzuführen sein (siehe auch Schwaninger, Hardmeier & Hofer, in press).

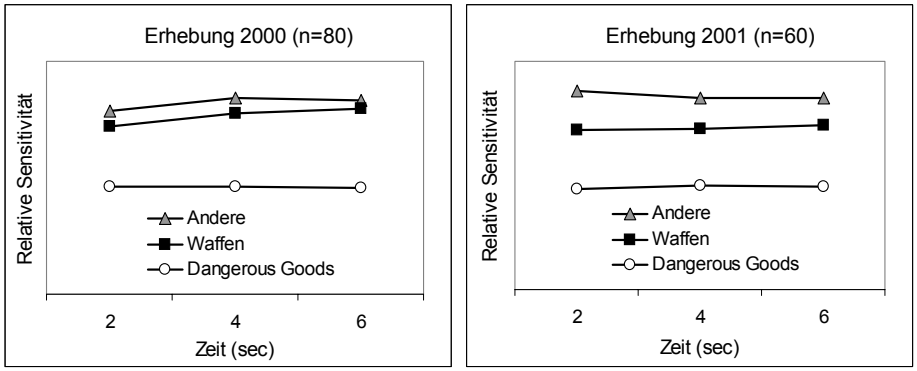

Abbildung 13 Relative Unterschiede bei der Erkennungsleistung verschiedener verbotener Gegenstände bei Präsentationsdauern von 2, 4 und 6 Sekunden.

Im Alltag wird von Sicherheitsmitarbeitern öfters berichtet, dass sie schon nach wenigen Sekunden ein Gefühl dafür hätten, ob beim Gepäck „etwas nicht stimmt“. Bei der Kontrolle von Handgepäck hat man oft nur 3-6 Se-

kunden Zeit, um das Röntgenbild zu beurteilen. Bei der Kontrolle von Gepäck, welches im Frachtraum mitgeführt wird, ist die Entscheidungszeit länger, d.h. oft zwischen 12 und 18 Sekunden. Im zweiten Test sollte untersucht werden, inwiefern die ersten Sekunden visueller Informationsverarbeitung die Erkennungsleistung bestimmen. Dazu wurden 90 Röntgenbilder verwendet, welche in zufälliger Reihenfolge je einmal 2, 4 und 6 Sekunden lang gezeigt wurden. Im Gegensatz zum ersten Test waren alle Bilder immer Schwarzweiss und die Auflösung konnte nicht verändert werden. Die Häufigkeit der verschiedenen Kategorien verbotener Gegenstände war jedoch gleich wie beim ersten Test. Wiederum musste für jedes Bild am Computer entschieden werden, ob es „OK“ oder „NICHT OK“ ist. Es nahmen die gleichen Sicherheitsbeauftragten teil, welche bereits am ersten Test teilgenommen hatten (80 Personen im Jahre 2000 und 60 Personen im Jahre 2001). Als Mass für die Erkennungsleistung wurde wiederum basierend auf der SDT die Sensitivität d’ berechnet. Wie man der *Abbildung 13* entnehmen kann, verändert sich die Erkennungsleistung nach 2 Sekunden nicht mehr stark. Dieser Befund zeigte sich in beiden Erhebungen (2000 und 2001). Er passt sehr gut zum subjektiven Eindruck vieler Sicherheitsbeauftragter, wonach

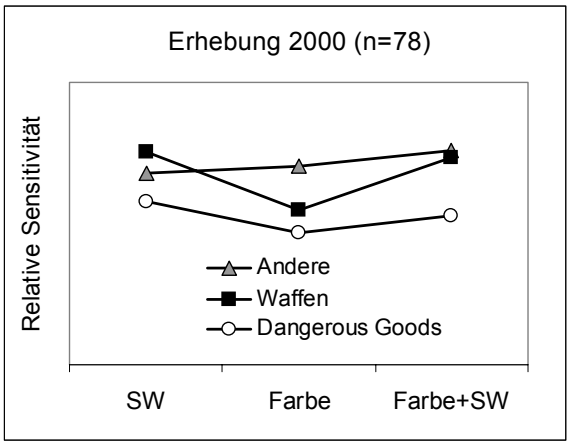

Abbildung 14 Erkennung verbotener Gegenstände bei Präsentation in Schwarzweiss (SW), Farbe sowie in Farbe plus Schwarzweiss.

bereits nach wenigen Sekunden ein Gefühl da ist, ob das Gepäck in Ordnung ist oder nicht. Dennoch muss betont werden, dass aufgrund dieser Daten keinesfalls geschlossen werden darf, dass für die Bildbeurteilung immer 2 Sekunden ausreichen. Erstens werden gewisse verbotene Gegenstände erst nach einer länger dauernden visuellen Suche erkannt. Zweitens werden bei vielen Gepäckstücken von den Sicherheitsbeauftragten verschiedene Darstellungsfunktionen eingesetzt, um das Bild besser beurteilen zu können. Dies

benötigt natürlich ein paar zusätzliche Sekunden. Ein dritter wichtiger Punkt ist, dass bei einer länger dauernden Bildbeurteilung von über 6 Sekunden mehr Zeit für perzeptuelle Lernprozesse zur Verfügung stehen, was vor allem bei neuen Mitarbeitern sehr wichtig ist. Viertens ist anzumerken, dass kurze Präsentationsdauern von weniger als 4 Sekunden in einem Test zwar vertretbar sind, bei der alltäglichen Arbeit jedoch zu erheblichen kognitiven Belastungen und Ermüdungen führen würden. Die meisten eingesetzten Röntgenprüfsysteme besitzen eine Vielzahl von Bilddarstellungsfunktionen wie z.B. Farbbild, Schwarzweissbild, Negativ, Kantenverstärkung, Kontrast- und Helligkeitsänderungen, usw. In Fachkreisen war man sich insbesondere über die Rolle der Farbe bei der Bildauswertung uneinig. Mit unterschiedlichen Farben werden aufgrund der Atomzahl verschiedene Materialien angezeigt. Im Röntgenbild, welches in *Abbildung 11* dargestellt ist, wird z.B. sprengstoffverdächtiges organisches Material mit rot markiert, schlecht durchdringbares Metall mit gelb. Bei diesen älteren Röntgenprüfgeräten besteht das Problem, dass die Farbe deckend ist. Dies könnte die Detektion verbotener Gegenstände sogar beeinträchtigen. Der Effekt der Farbe auf die Erkennungsleistung wurde im dritten Test untersucht. Dabei wurden 90 Röntgenbilder in zufälliger Reihenfolge je 4 Sekunden in Farbe, 4 Sekunden schwarzweiss oder aber 2 Sekunden in Farbe gefolgt von 2 Sekunden schwarzweiss gezeigt. Die prozentualen Anteile verbotener Gegenstände waren gleich wie im ersten und zweiten Test. Der dritte Test wurde nur in der Erhebung 2000 von 78 Mitarbeitern durchgeführt. In der *Abbildung 14* ist die relative Sensitivität d' als Mass für die Erkennungsleistung in den drei Bedingungen dargestellt. Tatsächlich zeigte sich eine Beeinträchtigung der Erkennungsleistung durch die deckende Farbe der älteren Röntgenanlagen. In der Bedingung, in welcher nur das Farbbild gezeigt worden ist, war die Erkennungsleistung schlechter als in der Schwarzweiss-Bedingung oder wenn zuerst 2 Sekunden das Farbbild und anschliessend 2 Sekunden das Schwarzweissbild gezeigt wurde. Dieser Effekt war bei Waffen besonders ausgeprägt und hängt wohl damit zusammen, dass bei Waffen in der Regel viel Metall enthalten ist, welches im Röntgenbild mit gelber Farbe überdeckt wird. Während einige Gefahrengüter auch Metall enthalten (z.B. Campingkocher, Taucherlampen, Motorsägen mit Benzin) ist bei anderen das gefährliche Material organisch (gefüllte Benzinkanister aus Plastik oder Feuerwerk). Dies erklärt, weshalb die beeinträchtigende Wirkung der deckenden Farbe bei Gefahrengütern weniger stark ausgefallen ist. Bei den anderen verbotenen Gegenständen wie z.B. Elfenbein oder ausgestopften geschützten Tieren ist viel seltener Metall oder sprengstoffähnliches organisches Material enthalten. Tatsächlich zeigte sich hier kein signifikanter Effekt der Darstellungsbedingung. Interessant ist auch, dass bei der Anzeige des Farb- plus Schwarzweissbildes die Erkennungsleistung nicht viel besser ausfiel. Offen-

bar ist für die Erkennung verbotener Gegenstände die Farbe im Mittel nicht sehr diagnostisch.
Heute wird am Flughafen Zürich die Reisegepäckprüfung mit moderneren Röntgenanlagen durchgeführt, welche eine viel bessere Bildqualität mit durchsichtiger Farbe besitzen. Die im dritten Test gefundene Beeinträchtigung der Erkennungsleistung durch die deckende Farbe konnte damit behoben werden.

4.2 Gesichtserkennung

Gesichter gehören zu den relevantesten visuellen Stimuli des Alltags. Obwohl sie sehr unterschiedlich aussehen, handelt es sich hierbei geometrisch gesehen um eine sehr homogene Stimuluskategorie. Jedes Gesicht besteht aus den gleichen Teilen wie Nase, Mund, Augen, Kinn etc. in einer ähnlichen Anordnung. Während aufrechte Gesichter ziemlich gut erkannt werden, ist die Erkennung bei Erwachsenen stark beeinträchtigt, wenn Gesichter auf den Kopf gedreht werden (für eine Übersicht siehe Valentine, 1988; Schwaninger, Carbon, & Leder, 2003). Dies ist darauf zurückzuführen, dass Gesichter im Alltag meist aufrecht gesehen werden und hängt auch damit zusammen, dass Gesichter zu komplexe Stimuli sind, um sie als ganzes mental rotieren zu können (Rock, 1973, 1988; Schwaninger et al., 2003).

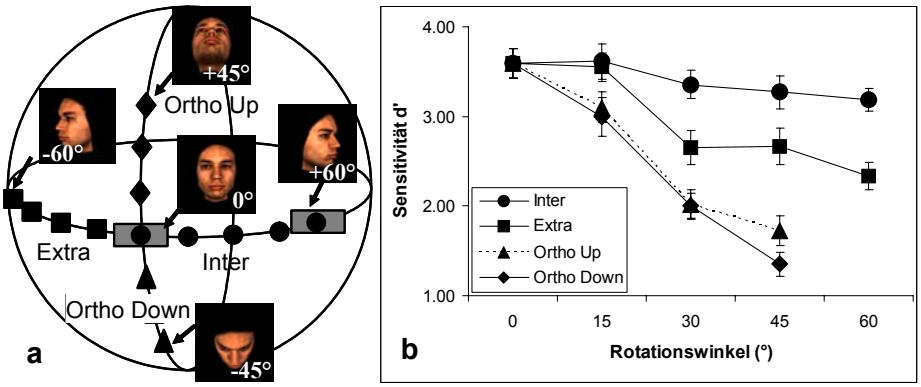

Abbildung 15 a Präsentationsbedingungen. Gelernt wurde die 0° und 60° Ansicht (graue Balken). Getestet wurden alle abgebildeten 15 Ansichten (Bedingungen Inter, Extra, Ortho Up und Ortho Down). b Erkennungsleistung gemessen durch die Sensitivität d' für die verschiedenen Bedingungen und Rotationswinkel (nach Wallraven, Schwaninger, Schumacher, & Bülthoff, 2002).

Wallraven, Schwaninger, Schumacher und Bülthoff (2002) haben die drei in Kapitel 2 vorgestellten ansichtenbasierten Ansätze im Bezug auf ihre Plausibilität für die menschliche Gesichtserkennung untersucht. Dabei wurde das Inter-Extra-Ortho Paradigma nach Bülthoff und Edelman (1992) verwendet.

Im Experiment lernten die Versuchspersonen 10 Gesichter, welche von vorne (0°) und von der Seite (60°) abgebildet waren. Danach wurden diese Gesichter und 10 Distraktoren in den 15 Winkeln präsentiert, welche in Abbildung 15 dargestellt sind (Bedingungen Inter, Extra, Ortho Up, Ortho Down). Die Versuchsperson musste jedes Mal entscheiden, ob es sich um ein gelerntes Gesicht oder um einen Distraktor handelte. Die Daten wurden

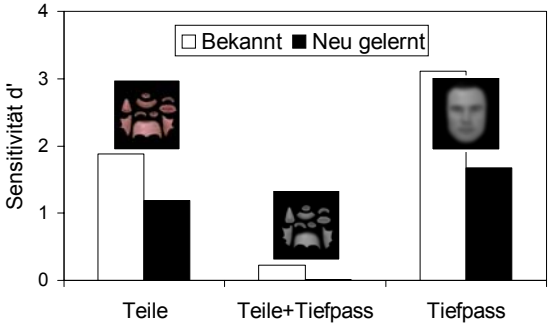

Abbildung 16 Studie von Schwaninger et al. (2002). Gesichter als neuartig angeordnete Teile (links), zusätzlich tiefpassgefiltert (Mitte), ganze Gesichter mit selbem Tiefpass gefiltert (rechts).

mittels SDT ausgewertet. Wie in Kapitel 3.3 erklärt wurde, bestimmt sich die Sensitivität durch d' = Z(Hit) – Z(FA). Die Hit Rate entspricht im vorliegenden Paradigma der Anzahl korrekt erkannter Gesichter und die Fehlalarmrate (FA) der Anzahl fälschlich als vorher gelernt bezeichneter Gesichter. Die theoretischen Vorhersagen sind analog zur Studie von Bülthoff und Edelman (1992), welche mit Drahtobjekten und amoebenartigen Objekten durchgeführt worden ist. Wird angenommen, dass Gesichter erkannt werden, indem ein 3D Modell fehlerfrei rotiert werden kann, dann müsste die Erkennung in den Bedingungen Inter, Extra und Ortho etwa gleich gut sein. Unter der Annahme einer Erkennung durch Linearkombination ergibt sich eine andere Vorhersage. Die Linearkombination kann verschiedene Rotationen um die Hochachse nachbilden; sie wird aber fehlerhaft, wenn orthogonale Rotationen berechnet werden müssen. Die Vorhersage ist deshalb Inter = Extra > Ortho. Wird schliesslich angenommen, dass Gesichter mittels Interpolation gespeicherter 2D Ansichten erkannt werden, so müsste die Erkennung in der Inter Bedingung besser als in der Extra und in der Ortho Bedingung sein. Wie man der Abbildung 15 entnehmen kann, ist genau dies eingetreten, es zeigte sich Inter > Extra > Ortho Up = Ortho Down. Die Erkennung von Gesichtern beim Menschen scheint also weder durch die Transformation eines 3D Modells, noch durch Linearkombination zu geschehen. Die Annahme der Interpolation von 2D Ansichten passt dagegen sehr gut zu den psychophysi-

schen Daten und stimmt auch mit Befunden aus verschiedenen anderen psychophysischen und neurophysiologischen Studien überein (Tarr & Bülthoff, 1998).

In der Studie von Schwaninger, Collishaw und Lobmaier (2002) wurde untersucht, inwiefern beim Menschen die Information von Teilen und deren räumliche Anordnung (konfigurale Information) eine Rolle spielt. Dies ist von besonderem theoretischen Interesse weil in der Wahrnehmungspsychologie für Gesichter ein rein holistischer Verarbeitungsmechanismus postuliert worden ist (Farah, Tanaka, & Drain, 1995; Tanaka & Farah, 1993). Holistisch bedeutet dabei, dass nur das Ganze verarbeitet wird, ohne dass Teile und ihre Relationen explizit enkodiert werden. Im Experiment von Schwaninger et al. (2002) wurden zuerst 10 Gesichter gelernt. Im Test wurden die gelernten Gesichter und 10 Distraktoren in Teile auseinander geschnitten und neuartig angeordnet dargeboten. Die Daten wurden wiederum mit SDT ausgewertet. Wie man der Abbildung 16 (links) entnehmen kann, konnten die Gesichter klar überzufällig anhand der Teile erkannt werden. Die Sensitivität d' erreichte Werte, welche klar über der Ratewahrscheinlichkeit von d' = 0 waren. Dieser Befund zeigt eindeutig, dass bei der menschlichen Gesichtswahrnehmung lokale Information der Teile explizit enkodiert wird. Eine rein holistische Gesichtsverarbeitung, bei der nur das Gesicht als Ganzes enkodiert würde, kann diese Daten nicht erklären. In einem weiteren Experiment wurden die Gesichter so stark tiefpassgefiltert, bis sie nicht mehr anhand der Teile erkannt werden konnten. Der Tiefpassfilter hatte tatsächlich sämtliche lokale Detailinformation der Teile eliminiert, was daran ersichtlich ist, dass d' nicht mehr von der Ratewahrscheinlichkeit d'=0 verschieden war (Abbildung 16, Mitte). Im dritten Experiment wurden nun Gesichter als Ganzes gezeigt, welche mit dem gleichen Filter bearbeitet worden waren. Wegen der Tiefpassfilterung enthielten diese Gesichter per definitionem keine lokale Detailinformation der Teile sondern nur noch die konfigurale Information über die räumliche Anordnung der Teile. Tatsächlich konnten diese Gesichter wieder klar überzufällig erkannt werden, was für eine separate Enkodierung von konfiguraler Information spricht. Diese drei Experimente wurden mit anderen Versuchspersonen repliziert, welche die gelernten Gesichter schon kannten. Generell war die Erkennungsleistung besser, aber es zeigten sich keine qualitativen Unterschiede. Sowohl bei der Erkennung von bekannten als auch bei neu gelernten Gesichtern wird also die Information der Teile und die konfigurale Information enkodiert und kann für die Erkennung verwendet werden. Dabei ist mindestens bezüglich der konfiguralen Information wichtig, dass zwischen Wahrnehmungs- und Erkennungsprozessen unterschieden wird (Collishaw, Hole, & Schwaninger, in press; Schwaninger, Ryf, & Hofer, 2003).

5. Schlussbemerkung

Marr hat als einer der Pioniere der Objekterkennung betont, wie wichtig eine interdisziplinäre Vorgehensweise ist. Die Implementation von Theorien mit Computern ermöglicht einen wichtigen Plausibilitätstest. Eine Theorie sollte aber auch physiologisch plausibel sein und mit den Erkenntnissen aus der Neurophysiologie übereinstimmen. Betrachtet man die Entwicklung im Bereich der Objekterkennung während der letzten zwanzig Jahre, so kommt der interdisziplinären Zusammenarbeit zwischen Wahrnehmungspsychologen, Informatikern und Neurowissenschaftlern eine immer grössere Bedeutung zu. Durch diese Zusammenarbeit gelingt es, das menschliche Gehirn besser zu verstehen und nicht selten führt dies zu neuen Erkenntnissen, welche zur Lösung von Wahrnehmungsproblemen in der Praxis eingesetzt werden können.

6. Literaturliste

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115-147.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *An Invitation to Cognitive Science* (2nd ed., Vol. 2, pp. 121-165). Cambridge, Massachusetts: MIT Press.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, USA, 89*, 60-64.

Bülthoff, H. H., Edelman, S., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5*(3), 247-260.

Collishaw, S., Hole, G., & Schwaninger, A. (in press). Configural processing and perceptions of head tilt. *Perception*, in press.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, Massachusetts: MIT Press.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 628-634.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals*. Mahwah, NJ: Lawrence Erlbaum Associates.

Graf, M. (2002). Form, space and object. Geometrical transformations in object recognition and categorization. Berlin: Wissenschaftlicher Verlag.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, J.B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas, *Psychological Bulletin,* 75, 424-429.

Hofer, F. & Schwaninger, A. (in press). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings*, in press.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480-517.

Huttenlocher, D. P., & Ullman, S. (1990). Recognizing solid objects by alignment with an image. *International Journal of Computer Vision, 5*, 195-212.
Jolicoeur, P., & Humphrey, G. K. (1998). Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In V. Walsh & J. Kulikowski (Eds.), *Perceptual Constancy. Why Things Look as They Do* (pp. 69-123). Cambridge: Cambridge University Press.
Kosslyn, S. M. (1994). *Image and Brain. The resolution of the imagery debate*. Cambridge, Massachusetts: MIT Press.
Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552-563.
Lowe, D. G. (1985). *Perceptual organization and visual recognition*. Boston: Kluwer Academic Publishing.
Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence, 31*, 355-395.
Marr, D. (1982). *Vision*. San Francisco: Freeman.
McMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.
Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*(6255), 263-266.
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.
Rock, I. (1973). Orientation and form. New York: Academic Press.
Rock, I. (1988). On Thompson's inverted-face phenomenon (research note). *Perception, 17*(6), 815-817.
Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security (pp. 147-156).* Wessex: WIT Press.
Schwaninger, A. (2003a). Reliable measurements of threat detection. *AIRPORT, 1/2003*, 22-23.
Schwaninger, A. (2003b). Screener evaluation and selection. *AIRPORT, 2/2003*, 14-15.
Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation security international*, FEB/2004, 31-36.
Schwaninger, A., Carbon, C.C., & Leder, H. (2003). Expert face processing: Specialization and constraints. In G. Schwarzer & H. Leder, *Development of face processing* (pp. 81-97)*,* Göttingen: Hogrefe.
Schwaninger, A., Collishaw, S. M., & Lobmaier, J. (2002). Role of featural and configural information in familiar and unfamiliar face recognition. *Lecture Notes in Computer Science, 2525*, 643-650.
Schwaninger, A., Hardmeier, D., & Hofer, F. (in press). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings*, in press.
Schwaninger, A., Ryf, S., & Hofer, F. (2003). Configural information is processed differently in perception and recognition of faces. *Vision Research, 43,* 1501-1505.
Swets, J. A. (1996). *Signal detection theory and roc analysis in psychology and diagnostics*. Mahwah, New Jersey: Lawrence Erlbaum.
Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology A, 46*(2), 225-245.
Tarr, M. J., & Bülthoff, H. H. (1998). *Object recognition in man, monkey and machine.* Cambridge, Massachusetts: MIT Press.

Ullman, S. (1996). *High-level vision*. Cambridge, Massachusetts: MIT Press.
Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(10), 992-1006.
Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *British Journal of Psychology, 79*, 471-491.
Wallraven, C., Schwaninger, A., Schumacher, S., & Bülthoff, H. H. (2002). View-based recognition of faces in man and machine: re-visiting inter-extra-ortho. *Lecture Notes in Computer Science, 2525*, 651-660.