Bundesamt
für Sicherheit in der
Informationstechnik

Studie:

„Untersuchung der Leistungsfähigkeit von Gesichtserkennungssystemen zum geplanten Einsatz in Lichtbilddokumenten – BioP I“

Öffentlicher Abschlussbericht

Bundesamt
für Sicherheit in der
Informationstechnik

Bundeskriminalamt

secunet

Version 1.1

07.04.2004

Bundesamt für Sicherheit in der Informationstechnik
Godesberger Allee 185-189, 53175 Bonn • Postfach 20 03 63, 53133 Bonn
Tel.: + 49 (0) 1888 9582-0 • Fax: + 49 (0) 1888 9582-400 • Internet: www.bsi.bund.de

Inhalt

Abkürzungsverzeichnis

BKA	Bundeskriminalamt
BSI	Bundesamt für Sicherheit in der Informationstechnik
GE	Gesichtserkennung
FAR	False Acceptance Rate
FER	Failed Enrolment Rate
FRR	False Rejection Rate
ICAO	International Civil Aviation Organization
ME	Matching Engine
MRZ	Machine Readable Zone
NTP	Network Time Protocol
OCR	Optical Character Recognition
ODBC	Open Database Connectivity
RefID 1	Bilddatei einer Frontalaufnahme als Referenzbasis
RefID 2	Foto auf Musterpersonalausweis als Referenzbasis (neuer Scan bei jeder Betätigung)
RefID 3	Foto auf EU-Visum als Referenzbasis
RefID 4	Komprimierte Bilddatei einer Frontalaufnahme als Referenzbasis
RefID 5	Bilddatei einer Halbprofilaufnahme als Referenzbasis
RefID 6	Foto auf aktuellem Personalausweis als Referenzbasis
RefID 7	System-Template aus Live-Enrolment als Referenzbasis
RefID 8	Foto auf Musterpersonalausweis als Referenzbasis
SQL	Structured Query Language
SSH	Secure Shell
User50	Teilmenge der Gesamtpopulation
USV	Unterbrechungsfreie Stromversorgung
WU	Weiterführende Untersuchungen
VNC	Virtual Network Computing
VPN	Virtual Private Network

1 Vorwort

Im Zuge der Maßnahmen zur Bekämpfung des internationalen Terrorismus wird eine Verbesserung der Identitätsprüfung in den verschiedenen Stufen der Einreise- und Aufenthaltskontrolle durch den Einsatz biometrischer Verfahren angestrebt. Als wesentliche Grundlage für eine Umsetzung entsprechender Überlegungen hat der Bundestag das am 9. Januar 2002 in Kraft getretene Terrorismusbekämpfungsgesetz beschlossen, das die Anpassung zahlreicher Sicherheitsgesetze an die neue Bedrohungslage beinhaltet. Zu den geänderten Bestimmungen zählen unter anderem das Passgesetz, das Gesetz über Personalausweise, das Ausländergesetz sowie das Asylverfahrensgesetz. Dabei werden insbesondere verschiedene Aspekte der Personenidentifizierung neu geregelt. So dürfen nun Pässe und Personalausweise neben dem Lichtbild und der Unterschrift weitere biometrische Merkmale beinhalten, die sich auf den Fingerabdruck, die Handgeometrie oder das Gesicht des Inhabers beziehen.

Weitere Motivation für die Befassung mit diesem Thema stellen die internationalen Aktivitäten und Rahmenbedingungen in diesem Umfeld dar. So hat der Kongress der USA ein Gesetzespaket zur Terrorismusbekämpfung beschlossen, das unter anderem wesentliche Änderungen des visafreien Einreiseprogramms („Visa Waiver Program“) beinhaltet. Hierbei wird von den teilnehmenden Staaten und damit auch Deutschland verlangt, dass bis zum 26. Oktober 2004 biometrische Merkmale in deren Reisedokumente eingeführt werden oder aber zumindest ein diesbezügliches Programm besteht. Entsprechend der Zielsetzung des US-amerikanischen „Enhanced Border Security and Visa Entry Reform Act“ vom 14.05.2002 auf Grundlage des „US Patriot Act“ für die Einführung von Biometrie auf Reisedokumenten der Visa-Waiver-Staaten befasst sich auch die ICAO (International Civil Aviation Organization) mit Empfehlungen zur Erweiterung von Reisedokumenten um biometrische Merkmale. Die ICAO gibt grundsätzlich den Einsatz von Gesichtserkennung als das biometrische Merkmal für globale Interoperabilität vor, lässt jedoch optional weitere Merkmale wie Fingerabdruck oder Iris zu.

In diesem Zusammenhang war es Ziel der Studie BioP I, die Leistungsfähigkeit von derzeit auf dem Markt verfügbaren Gesichtserkennungssystemen für eine Verwendung bei Personaldokumenten mit Lichtbildern zu untersuchen. Dabei wurde ein Testsieger ermittelt, der sich in der zweiten Projektphase BioP II einem verfahrensvergleichenden Systemtest zu Finger- und Iris-erkennungssystemen unterziehen wird.

Dieser Bericht stellt die Testkonzeption und die wichtigsten Ergebnisse von BioP I vor. Die Studie wurde unter der Gesamtprojektleitung des BSI gemeinsam mit dem BKA durchgeführt und durch die Firma secunet Security Networks AG als Auftragnehmer realisiert.

Bonn, Wiesbaden, Essen, im Januar 2004

2 Management Summary

Die Untersuchung biometrischer Gesichtserkennungsverfahren im Rahmen der Studie BioP I diente dazu, Aussagen zur Leistungsfähigkeit der zum gegenwärtigen Zeitpunkt auf dem Markt verfügbaren Gesichtserkennungssysteme bezüglich verschiedener Aspekte zu treffen und daraus Erkenntnisse für eine Verwendung von Gesichtserkennung im Zusammenhang mit Personaldokumenten zu gewinnen. Grundlage sind hierfür unter anderem die zahlreichen Gesetzesänderungen, die im Rahmen des Terrorismusbekämpfungsgesetzes vom 09.01.2002 im Bereich der Personaldokumente erfolgt sind und neben dem bisher traditionell eingesetzten Lichtbild und der Unterschrift eine Verwendung weiterer biometrischer Merkmale wie des Gesichts eröffnen. Die Betrachtungen in BioP I beinhalten zum einen eine vergleichende Untersuchung zweier, in einer Voruntersuchung ausgewählter Systeme in einem nach wissenschaftlichen Kriterien angelegten Systemtest, zum anderen den Vergleich mehrerer Algorithmen in unterschiedliche Ausprägung. Darüber hinaus wurden auf Grundlage der vielfältigen, in Betracht kommenden Personaldokumente mit Lichtbildern mehrere Referenzbasen untersucht, um Aussagen darüber zu treffen, ob und mit welchen Ergebnissen die getesteten Systeme Lichtbilder unterschiedlichster Art und Qualität auf Personaldokumenten in einem Verifikationsprozess verarbeiten können. Schließlich wurden die Testteilnehmer nach ihrer Einschätzung zu Biometrie allgemein, zur Gesichtserkennung im Besonderen und zu den konkret verwendeten Systemen in der Benutzung befragt.

Während somit in BioP I ausschließlich die Gesichtserkennung (GE) untersucht wurde, hat BioP II im unmittelbaren Anschluss daran eine vergleichende Betrachtung der Verfahren Gesichts-, Fingerabdruck- und Iriserkennung zum Ziel.

Der Gedanke an einen Einsatz von Gesichtserkennung bei Personaldokumenten liegt deswegen nahe, da etwa der deutsche Personalausweis sowie der Reisepass in seiner aktuellen Form bereits Lichtbildinformationen enthält und die Verwendung von Gesichtsbildern zur Identitätsüberprüfung bereits gängige Praxis ist. Zur Bereitstellung des biometrischen Merkmals Gesicht auf dem Ausweis sind als grundsätzliche Alternativen die Verwendung des vorhandenen Lichtbildes sowie die Speicherung einer Bilddatei und eines Templates in digitaler Form auf einem in das Ausweisdokument integrierten Chip denkbar.

Projektziele

Inhalt des Projekts BioP I ist demzufolge die Untersuchung der in diesem Zusammenhang auftretenden Fragestellungen zur Machbarkeit und technischen Realisierung:

- Ist die Gesichtserkennung für die Verwendung mit Lichtbildausweisen geeignet?
- In welcher Form und Qualität muss das biometrische Merkmal bereitgestellt werden?
- Was sind wesentliche Einflussfaktoren für die Gesichtserkennung?
- Wie ist es um die Überwindungssicherheit von Gesichtserkennungssystemen bestellt?
- Welches der getesteten Gesichtserkennungssysteme erzielt die besten Erkennungsleistungen?

Bei der Bearbeitung dieser Fragestellungen werden die internationalen Rahmenbedingungen – insbesondere die Richtlinien der ICAO für biometrisch verwertbare Gesichtsbilder – berücksichtigt.

Als in die Prüfung einzubeziehende Lichtbildausweise wurden der aktuelle Bundespersonalausweis, der Reisepass der Bundesrepublik Deutschland, das EU-Visum, längerfristige Aufenthaltstitel nach

EU-Modell sowie die neuen vorläufigen Reisepässe und Personalausweise der Bundesrepublik Deutschland ausgewählt.[1]

Für die genannten Zielvorgaben werden im Rahmen von BioP I Gesichtserkennungssysteme zweier verschiedener Hersteller im Verifikationsmodus (1:1) getestet, wobei eines dieser Systeme mehrere Gesichtserkennungsalgorithmen verschiedener Anbieter parallel bereitstellt. Die Entscheidung für diese Systeme wurde auf Basis eines im Vorfeld durchgeführten Auswahltests getroffen.

Mit der getroffenen Systemauswahl sind folgende Vergleiche möglich:

- Vergleich zweier Komplettsysteme
- Vergleich verschiedener Algorithmen innerhalb eines Komplettsystems
- Vergleich eines Algorithmus innerhalb zweier Komplettsysteme

Vorgehen

Bei der Gesichtserkennung wird eine aktuelle Aufnahme des Gesichts mit einer im Vorfeld gespeicherten Referenzaufnahme der entsprechenden Person, der so genannten Referenzbasis, verglichen. Bezüglich der Bereitstellung dieser Referenzbasis im Hinblick auf Personaldokumente gibt es mehrere Alternativen. Zunächst kann das bereits auf dem Ausweis vorhandene Lichtbild als Referenz verwendet werden. Dies bedeutet, dass bei der Identitätsprüfung einer Person das Lichtbild auf dem Ausweis gescannt und mit der aktuellen Aufnahme verglichen wird. Je nach Ausweistyp ergibt sich eine unterschiedliche Charakteristik des Lichtbildes. So ist beispielsweise das Lichtbild auf dem EU-Visum kleiner und hat mehr optische Störungen als ein Lichtbild auf dem Personalausweis.

Eine Alternative zur Verwendung des Lichtbildes stellt die Bereitstellung der Referenzbasis in elektronischer Form dar. Dies kann entweder eine Bilddatei des Gesichts oder eine spezielle, in der Regel proprietäre Kodierung des Gesichts, das Template, sein. Der Ausweis müsste dann entsprechend um einen Speicherplatz erweitert werden. Aus diesem würde bei der Identitätsprüfung in der Folge die Referenzbasis gelesen.

Zur Bewertung dieser Alternativen bezüglich ihrer Eignung für Gesichtserkennung wurden repräsentativ verschiedene Referenzbasen parallel getestet. Dies sind im Einzelnen das Lichtbild auf dem aktuellen Bundespersonalausweis, das Lichtbild auf einem speziell für das Projekt angefertigten Musterpersonalausweis mit frontaler Bildaufnahme[2], das Lichtbild des EU-Visums, die Bilddatei einer frontalen Bildaufnahme, die komprimierte Bilddatei einer frontalen Bildaufnahme gemäß ICAO, die Bilddatei einer Halbprofilaufnahme sowie ein proprietäres Template des jeweiligen Systemanbieters. Die Auswahl dieser Referenzbasen deckt alle oben genannten Ausweistypen ab. Sie erlaubt zudem weitere Aussagen zum Einfluss der Kompression des Bildmaterials auf die Gesichtserkennung, die Unterscheidung der Erkennungsleistung für Frontal- und Halbprofilaufnahmen sowie den Vergleich des aktuellen Bundespersonalausweises mit einem für die Gesichtserkennung optimierten Musterausweis.

[1] Das Lichtbild auf dem Reisepass entspricht dem des Bundespersonalausweises, daher wird dieser im Folgenden stellvertretend für beide Dokumente betrachtet. Ebenso steht das Lichtbild des EU-Visums im Folgenden stellvertretend für die Lichtbilder der längerfristigen Aufenthaltstitel sowie der vorläufigen Reisepässe und Personalausweise.

[2] Das Lichtbild des Musterpersonalausweises entspricht den Richtlinien der ICAO für die Erstellung von Passbildern zum Einsatz für biometrische Anwendungen.

Die Tests für die unterschiedlichen Referenzbasen wurden für Gesichtserkennungssysteme von zwei verschiedenen Herstellern durchgeführt, die auf Basis einer Voruntersuchung ausgewählt wurden. Eines dieser Systeme erlaubt die Integration und den parallelen Betrieb mehrerer Gesichtserkennungsalgorithmen unabhängig voneinander. In diesem System kamen im Rahmen von BioP I Algorithmen von drei verschiedenen Anbietern zum Einsatz, die jeweils zusätzlich um einen alternativen Gesichtsfinder modifiziert wurden (Plus-Version des jeweiligen Algorithmus).

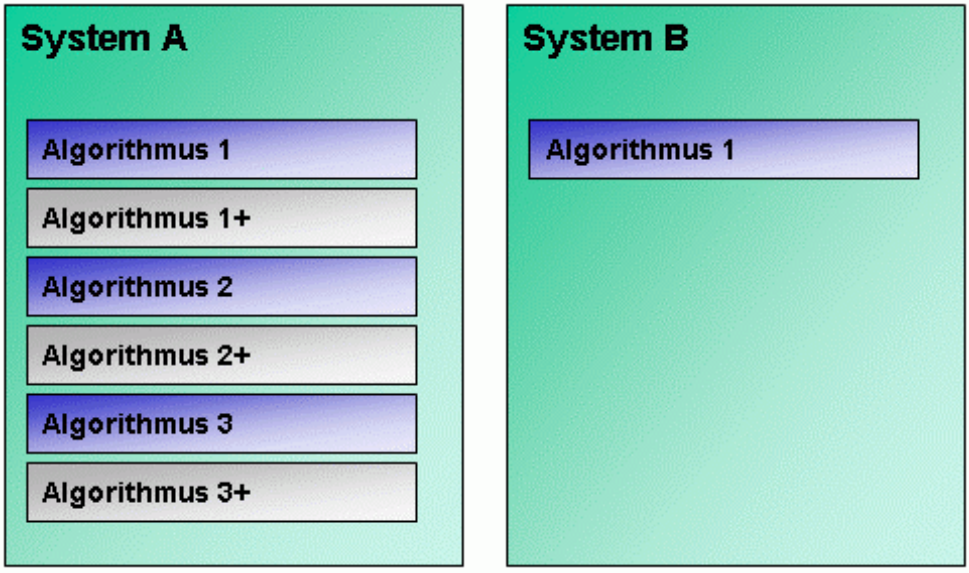

Abbildung 1: Übersicht Systeme und Algorithmen

Abbildung 1 zeigt, dass in System A drei Algorithmen zuzüglich der jeweiligen Plus-Version integriert sind und in System B ein Algorithmus.

Diese Konstellation erlaubt zum einen den Vergleich von Komplettsystemen auf Basis eines identischen Algorithmus, der in beiden Systemen integriert ist. Zum anderen können Algorithmen verschiedener Anbieter innerhalb eines identischen Komplettsystems verglichen werden. Entsprechend wird in BioP I zwischen einem **Systemvergleich** und einem **Algorithmenvergleich** unterschieden.

Um eine Repräsentativität der Testergebnisse zu gewährleisten, wurden die Untersuchungen auf Basis eines umfangreichen Feldtests in einer Liegenschaft des Bundeskriminalamts in Wiesbaden durchgeführt. Dabei wurden die Systeme über einen Zeitraum von 7 Wochen von annähernd 250 Personen benutzt, von denen 152 mehr als 50 Betätigungen wie im Folgenden beschrieben durchführten und damit in die Bewertung eingingen.

Eine Besonderheit bei der Durchführung des Tests stellte der Einsatz eines Ausweislesers der Bundesdruckerei (Verifier) dar. Die Testteilnehmer legten bei jeder Bedienung ihren Musterpersonalausweis auf, der die Personalausweisnummer und das gescannte Lichtbild für die Gesichtserkennungssysteme bereitstellte. Dieses Vorgehen lehnt sich an das oben beschriebene Zielszenario an.

Der Ablauf für die Bedienung eines Systems durch den Testteilnehmer stellte sich also folgendermaßen dar:

Nachdem der Ausweisleser die erforderlichen Informationen des Dokuments erfasst hatte, wurde die Gesichtserkennung angestoßen. Dabei wurden durch eine Kamera fortlaufend Bilder des Teilnehmers aufgenommen und einem Vergleich mit einer gespeicherten Referenzbasis unterzogen. Die Aufnahme

endete, wenn entweder ein Vergleich erfolgreich war oder wenn ein vorab festgelegtes Zeitlimit erreicht wurde. Das Aufnehmen der Bilder wurde dem Teilnehmer mittels eines gelben Lichtsignals angezeigt. Je nachdem, ob der Versuch erfolgreich war oder nicht, wurde dies der Person über ein grünes bzw. rotes Lichtsignal mitgeteilt. Im Hintergrund und für den Teilnehmer nicht wahrnehmbar erfolgten weitere Vergleiche des aufgenommenen Bildes mit allen im System integrierten Algorithmen und allen Referenzbasen. Die hierbei erzielten Ergebnisse wurden für die spätere Auswertung in einer Datenbank protokolliert.

Neben dem beschriebenen Feldtest wurden eine Reihe von weiterführenden Untersuchungen im Labor der Firma secunet Security Networks AG in Essen durchgeführt. Neben der Untersuchung von Einflussfaktoren auf die Gesichtserkennung wurden Möglichkeiten zur Reduzierung des Speicherplatzbedarfs bei der Bereitstellung der Referenzbasen geprüft. Weitere wesentliche Bestandteile waren die Prüfung der Überwindungssicherheit der beteiligten Systeme sowie die Durchführung einiger Offline-Tests.

Zur Darstellung und zum Vergleich der gewonnenen Ergebnisse wurden diese klassifiziert, anhand eines Notensystems bewertet und gemäß ihrer Bedeutung für die geplanten Anwendungen gewichtet in Bewertungsschemen abgebildet. Die Zusammenfassung der gewichteten Teilnoten resultierte in Gesamtnoten, die eine vergleichende Bewertung ermöglichen. Dieses Verfahren kam als geeignetes Werkzeug sowohl zum Vergleich der Algorithmen als auch zum Vergleich der Komplettsysteme zum Einsatz. Die ermittelten Ergebnisse sind in den nachfolgenden Abschnitten dargestellt.

Algorithmenvergleich

Innerhalb des Systems, in dem verschiedene Algorithmen integriert waren (System A), hat Algorithmus 1 bei nahezu allen wesentlichen Bewertungskriterien am besten abgeschnitten. Alle anderen folgten mit zum Teil erheblichen Abstand.

Systemvergleich

Während sich bezüglich der biometrischen Erkennungsleistung ein leichter Vorteil für den Anbieter A ergibt, lässt sich für andere wesentliche Bewertungskriterien ein zum Teil großer Vorsprung für den Hersteller B erkennen. Insbesondere bei Aspekten der Zuverlässigkeit, der Systemfehler, des Administrationsaufwands und des Supports konnte der Hersteller B ein deutlich positiveres Ergebnis erzielen. Diesen Kriterien kommt, neben der Erkennungsleistung, insbesondere bei Berücksichtigung eines breiteren Einsatzes und schließlich bei der Auswahl für BioP II eine große Bedeutung zu.

Referenzbasenvergleich

Eine besonders interessante Fragestellung in BioP I ist die Festlegung geeigneter Referenzbasen, insbesondere im Hinblick auf erforderliche Anpassungen deutscher Personaldokumente.

Es hat sich gezeigt, dass der **Bundespersonalausweis** in der gegenwärtigen Form nicht im Zusammenhang mit biometrischer Gesichtserkennung verwendbar ist. Dies begründet sich im Wesentlichen durch die Gesichtsdarstellung im Halbprofil sowie die im Einzelfall sehr schlechte Lichtbildcharakteristik bezüglich Kontrast und Helligkeit.

Der für das Projekt hergestellte **Musterpersonalausweis** mit einem Lichtbild gemäß den ICAO-Empfehlungen zeigt, dass Gesichterkennung grundsätzlich auf Basis eines vom Dokument zu scannenden Lichtbildes möglich ist. Die erzielten Ergebnisse sind zwar noch nicht zufrieden stellend, lassen aber ein gewisses Potenzial zur Verbesserung der Erkennungsleistungen erkennen.

Das getestete Lichtbild des **EU-Visums** fällt dagegen schon signifikant bezüglich der Erkennungsleistung ab. Ursache sind im Wesentlichen die Störungen innerhalb des Gesichtsbildes, die durch die optischen Sicherheitsmerkmale (zum Beispiel Melierfasern, Untergrunddruck) des Visums hervorgerufen werden.

Als Alternative zur direkten Verwendung des Lichtbilds wurde bereits die Nutzung einer **Bilddatei** aufgezeigt. Dies entspricht den ICAO-Empfehlungen und ermöglicht internationale Interoperabilität. Die mit dieser Alternative erzielbaren Erkennungsleistungen sind viel versprechend. Auch bei der durch ICAO empfohlenen Kompression fällt die Erkennungsrate nur in einem vertretbaren Rahmen ab. Bei Ausnutzung vorhandener Optimierungsspielräume ist hier durchaus ein erfolgreicher Einsatz denkbar. Dazu zählen beispielsweise die Verwendung spezieller Kamerasysteme, die Optimierung der Algorithmen auf die Verarbeitung von Bilddateien und eine geeignete Vorverarbeitung des Bildmaterials.

Der Test der **Bilddatei auf Basis einer Halbprofilaufnahme** zeigt deutlich, dass dieser Fototyp für die Gesichtserkennung ungeeignet ist. Dies unterstreicht die Ergebnisse zum aktuellen Personalausweis (siehe oben).

Erwartungsgemäß werden mit der Darstellung des Gesichts als **herstellerspezifisches Template** die mit Abstand besten Erkennungsleistungen bei allen Systemen erzielt.

Weiterführende Untersuchungen

Im Rahmen von weiterführenden Untersuchungen wurden für BioP I zusätzliche Detailfragen untersucht. Zunächst interessierten hier wesentliche Einflussfaktoren für die Gesichtserkennung. Die **Abhängigkeit von der Beleuchtung** bezüglich Intensität und Richtung ist hinlänglich als wichtigster Faktor für die Gesichtserkennung bekannt. Dies konnte durch die Untersuchungen in BioP I erneut bestätigt werden. Interessant ist dabei, dass für verschiedene Algorithmen und verschiedene Systeme die Beeinflussung unterschiedlich stark ist. Die Leistungsfähigkeit der Erfassungseinheit, also des Kamerasystems, spielt dabei eine wesentliche Rolle.

Der größte Abfall der Erkennungsleistung ist für alle Algorithmen und Systeme bei Lichteinfall von der Seite zu verzeichnen. Der Lichteinfall aus dem Hintergrund kann bei Einsatz eines geeigneten Kamerasystems annähernd vernachlässigt werden. Bei starkem Frontallicht fiel ein recht überraschender Effekt auf. Im Wesentlichen ergab sich hier eine Verschlechterung der Erkennungsleistung, während jedoch im Einzelfall eine deutliche Verbesserung auftrat. Von den beteiligten Algorithmen erwies sich in fast allen Fällen Algorithmus 1 als sehr robust.

Im Hinblick darauf, dass die Speicherkapazität eines möglichen Chips auf dem Personaldokument beschränkt ist und gemäß ICAO- und EU-Empfehlung ggf. mehrere Merkmale abzuspeichern sind, ist es weiterhin von Vorteil, die zu speichernden Informationen möglichst stark zu komprimieren. Deshalb wurde die Beeinflussung der Erkennungsleistung durch unterschiedliche **Kompressionsstufen** für die als Referenzbasis verwendeten Bilddateien geprüft. Dabei fiel auf, dass die Erkennungsleistung in der Regel mit zunehmender Kompression abnimmt. Während bei schwacher Kompression (Bildgröße ca. 75kB) ein zu vernachlässigender Rückgang zu verzeichnen ist, ergibt sich bei sehr starker Kompression (Bildgröße ca. 11kB) ein deutlicher Abfall. Eine Kompression, die sich in der von ICAO vorgeschlagenen Größenordnung bewegt (Bildgröße ca. 14kB), erzielt im Vergleich mit schwach komprimierten Referenzbasen noch akzeptable Erkennungsleistungen.

Als weitere Maßnahme zur Reduzierung des Speicherbedarfs wurden **Bilddateien mit geringerer Auflösung** getestet. Durch diese Modifikation ergeben sich bei allen Systemen leicht schlechtere Erkennungsraten. Für die Plus-Versionen der Algorithmen liegen in diesem Fall allerdings keine Ergebnisse vor, da entsprechende Bilddateien aufgrund nicht ausreichender Auflösung nicht verarbeitet werden konnten.

Ein weiterer wesentlicher Aspekt zur Beurteilung der Eignung von Gesichtserkennungssystemen im Zusammenhang mit Personaldokumenten ist der **Einfluss des Ausweisalters** und damit des darauf enthaltenen Referenzbildes auf die Erkennungsleistung. Eine entsprechende Untersuchung wurde auf Basis der aktuellen Personalausweise der Testteilnehmer vorgenommen. Da die Erkennungsleistung auf Basis dieser Ausweise jedoch generell sehr schlecht ist, können hier keine fundierten

Schlussfolgerungen gezogen werden. Trotzdem ist als Trend erkennbar, dass die Erkennungsleistung mit zunehmendem Ausweisalter abnimmt. Generell ist der Einfluss von Alterungseffekten auf Gesichtserkennungssysteme noch nicht ausreichend untersucht, wie eine im Rahmen von BioP I aktuell durchgeführte Sichtung diesbezüglicher Forschungsaktivitäten bestätigt.

Ein weiterer Einflussfaktor auf die Anwendung von Gesichterkennung bei Personalausweisen ist die **Qualität des Dokuments**. Dazu wurden die aktuellen Personalausweise der Testteilnehmer bezüglich Kratzern, Knicken, Rissen etc. klassifiziert. Dabei wurden kaum Ausweise mit mittlerer oder schlechter Qualität der Ausweisoberfläche im Bereich des Bildes identifiziert. Dies lässt den Schluss zu, dass der Bundespersonalausweis insbesondere im Bereich des Lichtbildes sehr robust ist. Aufgrund der sehr kleinen Stichprobe von Ausweisen niedriger Qualität können keine belastbaren Aussagen bzgl. der Beeinflussung der Erkennungsleistung getroffen werden.

Ein wesentliches Bewertungskriterium für biometrische Systeme insbesondere vor dem Hintergrund der Erhöhung der Sicherheit für das Einsatzszenario ist die **Überwindungssicherheit**. Die im Rahmen von BioP I durchgeführten Tests haben gezeigt, dass sich die beiden beteiligten biometrischen Systeme mit geringem Aufwand durch Kopien des biometrischen Merkmals Gesicht in Form von Fotos überwinden lassen. Allerdings war die Bereitstellung einer geeigneten Lebenderkennung kein Pflichtkriterium für die Systeme. Als ein sehr kritischer Aspekt ist dennoch herauszustellen, dass bei beiden Systemen in einem Einzelfall Verwechslungen von Personen auftraten, die sich nach visueller Beurteilung nur ansatzweise ähnlich sind. Dies kann dazu führen, dass sich eine Person ohne weiteren Aufwand mit dem Ausweis einer anderen Person identifiziert und als berechtigter Dokumenteninhaber vom System akzeptiert wird.

Benutzerakzeptanz

Im Rahmen des Projekts BioP I erfolgten statistische Untersuchungen über die Akzeptanz der getesteten biometrischen Systeme. Basis für die Akzeptanzuntersuchungen bildeten drei Befragungen der Testteilnehmer. Die erste Befragung fand vor Beginn der Testphase, die zweite nach etwa der Hälfte der Testphase und die dritte nach deren Ende statt. Das System des Herstellers B hat im Anwenderurteil deutlich besser abgeschnitten als das des Herstellers A. In jeder der fünf Kategorien Einfachheit der Bedienung, Erkennungsgenauigkeit, Schnelligkeit, Störanfälligkeit und Flexibilität sowie in der Gesamteinschätzung lag das System B vorn. Trotz des eindeutigen Rückstands ist aber auch das Ergebnis des Systems A auch noch als gut zu bewerten. Insgesamt kann die Beurteilung der Systeme damit als erfreulich gut bezeichnet werden. Die Bedienerfreundlichkeit der Systeme bereitet gemäß den erhobenen Ergebnissen keine Probleme, wogegen die Störanfälligkeit noch verbessert werden sollte.

Neben der Bewertung der konkret im Test eingesetzten Systeme waren die Testteilnehmer aufgefordert, auch Einschätzungen zur Gesichtserkennung und zu Biometrie im Allgemeinen vorzunehmen. Die Auswertung der Befragung lässt in erster Linie zwei Tendenzen erkennen. Zum einen haben die Teilnehmer im Verlauf des Tests eine zunehmend positive Einstellung zu zahlreichen Detailfragen der Gesichtserkennung entwickelt. So hält nur eine Minderheit die Gesichtserkennung für gesundheitsgefährdend, während die Praxisreife und die Zuverlässigkeit von einer großen Mehrheit der Teilnehmer als gegeben angesehen werden. Trotz dieser zustimmenden Haltung im Detail ist eine Skepsis der Teilnehmer in der Gesamtheit zu verzeichnen. So befürwortet eine Mehrheit die Forderung, dass Gesichtserkennung nicht unbetreut eingesetzt werden darf. Auch eine generelle Nützlichkeit wird nur von einem Drittel der Teilnehmer gesehen. Zu vermuten ist daher, dass ein Aufzeigen konkreter Nutzungsszenarien zur Steigerung der Akzeptanz der Biometrie in der Öffentlichkeit beitragen kann, während die Positionierung als Technologie für unterschiedlichste Anwendungen eher im Hintergrund stehen sollte.

Fazit

BioP I hat gezeigt, dass die Gesichtserkennung grundsätzlich für die Verwendung mit Personaldokumenten im Rahmen einer biometrischen Verifikation geeignet ist. Dies gilt jedoch nur unter Einhaltung folgender Randbedingungen und Erfüllung der dargestellten Grundvoraussetzungen:

- Die Referenzbasis ist auf dem Personaldokument bereitzustellen. Die besten Ergebnisse werden bei der Verwendung eines Templates erreicht. Realistischer bezüglich internationaler Einsetzbarkeit ist jedoch die Bereitstellung einer Bilddatei gemäß ICAO. Hier muss das vorhandene Optimierungspotential jedoch noch stärker ausgenutzt werden, um bessere Ergebnisse zu erzielen. Die Verwendung eines auf dem Ausweis vorhandenen Lichtbildes gemäß ICAO erscheint zwar möglich, hierfür müssen jedoch große Anstrengungen seitens der Algorithmenhersteller aufgewendet werden, um befriedigende Erkennungsleistungen zu erreichen.
- Eine wichtige Rahmenbedingung für den erfolgreichen Einsatz von Gesichterkennung ist die Schaffung einer kontrollierten Umgebung bezüglich des Lichteinflusses.
- Die Verbesserung der Überwindungssicherheit ist eine wesentliche Grundvoraussetzung für den Einsatz von Gesichtserkennungssystemen. Während die Überwindung mittels Fotos aufgrund einer voraussichtlich kontrollierten Identitätsprüfung nur eingeschränkt kritisch erscheint, ist die Verwechslung ähnlicher Personen nicht akzeptabel.
- Bezüglich der Eignung der Gesichtserkennung für Personaldokumente gilt auch der Vorbehalt, dass Alterungseffekte noch nicht ausreichend untersucht sind. Dies ist insbesondere vor dem Hintergrund der Gültigkeit dieser Dokumente relevant.
- Die genannten Randbedingungen implizieren einige notwendige Änderungen an deutschen Pässen und Personalausweisen, um eine qualitativ hochwertige Gesichtserkennung zu erreichen. Für eine geeignete Bereitstellung der Referenzbasen sollte der Personalausweis um ein Speichermedium erweitert werden. Als Rückfalllösung oder auch den parallelen bzw. Übergangseinsatz kommt durchaus das Lichtbild auf dem Ausweis infrage, sofern die aktuellen Aufnahmerichtlinien zur Lichtbilderstellung angepasst werden. Hier stellen die Richtlinien der ICAO für die Erstellung von Passbildern zum Einsatz für biometrische Anwendungen die geeignete Vorlage dar. Die gleichen Richtlinien sollten für die mittels Speichermedium des Ausweises bereitgestellten Bilddateien bindend sein.
- Die im Rahmen von BioP I ermittelten Ergebnisse werden innerhalb des Projekts BioP II auf Basis einer deutlich größeren Testgruppe überprüft und den biometrischen Verfahren Iris- und Fingerabdruckerkennung gegenübergestellt. Der in BioP I erarbeitete Algorithmenvergleich zeigt eine eindeutige Präferenz für Algorithmus 1, der für diese Untersuchungen ausgewählt wurde. Auch im Systemtest konnte eine klare Empfehlung ausgesprochen werden, da System B im BioP-I-Szenario bezüglich der neben der Erkennungsleistung für relevant erachteten Kriterien wie Fehlerverhalten, Zuverlässigkeit, Herstellerunterstützung sowie Akzeptanz durch die Testteilnehmer bessere Ergebnisse erzielen konnte.

3 Einleitung

Im Zuge der Terrorismus- und Kriminalitätsbekämpfung werden derzeit verschiedene Maßnahmen zur Erhöhung der inneren Sicherheit angestrebt. Daher erfolgten im Rahmen des Terrorismusbekämpfungsgesetzes im Januar 2002 Änderungen des Pass- und Personalausweisgesetzes. Diese ermöglichen die Erweiterung der Ausweisdokumente um biometrische Merkmale wie Gesicht, Fingerabdruck oder Handgeometrie zur Verbesserung des Prozesses der Identitätsfeststellung. Weiterhin gilt, dass keine zentrale Speicherung dieser biometrischen Merkmale erfolgen darf. Dies priorisiert die Verwendung von biometrischen Verfahren im Verifikationsmodus (1:1-Vergleich eines im Ausweis gespeicherten Merkmals gegen das entsprechende Live-Merkmal der Person).

Nahe liegend ist in diesem Rahmen der Einsatz von Gesichtserkennung (GE), da der Ausweis in seiner aktuellen Form bereits Lichtbildinformationen enthält und die Verwendung von Gesichtsbildern zur Identitätsüberprüfung allgemein akzeptiert ist. Zur Bereitstellung des biometrischen Merkmals Gesicht auf dem Ausweis sind als grundsätzliche Alternativen die Verwendung des vorhandenen Lichtbildes sowie die Speicherung einer Bilddatei und die Speicherung eines Templates in digitaler Form auf einem in das Ausweisdokument integrierten Chip denkbar.

Inhalt des Projekts BioP I ist die Untersuchung der in diesem Zusammenhang auftretenden Fragestellungen zur Machbarkeit und technischen Realisierung:

- Ist die Gesichtserkennung für die Verwendung mit Lichtbildausweisen geeignet? In welcher Form und Qualität muss dann das biometrische Merkmal bereitgestellt werden?
- Was sind wesentliche Einflussfaktoren für die Gesichtserkennung? Wie hoch ist die entsprechende Beeinflussung auf die Erkennungsleistung?
- Wie ist es um die Überwindungssicherheit von Gesichtserkennungssystemen bestellt?
- Welches der getesteten Gesichtserkennungssysteme erzielt die besten Erkennungsleistungen?

Bei der Bearbeitung dieser Fragestellungen werden die internationalen Rahmenbedingungen – insbesondere die Richtlinien der ICAO für biometrisch verwertbare Gesichtsbilder – berücksichtigt.

Für die genannten Zielvorgaben werden im Rahmen von BioP I Gesichtserkennungssysteme zweier verschiedener Hersteller im Verifikationsmodus getestet, wobei eines dieser Systeme mehrere Gesichtserkennungsalgorithmen verschiedener Anbieter parallel bereitstellt. Die Entscheidung für diese Systeme wurde auf Basis eines im Vorfeld durchgeführten Auswahltests getroffen.

Mit der getroffenen Systemauswahl sind folgende Vergleiche möglich:

- Vergleich zweier Komplettsysteme
- Vergleich verschiedener Algorithmen innerhalb eines Komplettsystems
- Vergleich eines Algorithmus innerhalb zweier Komplettsysteme

Die Studie BioP I wurde unter Federführung von BSI (Gesamtprojektleitung) und BKA durchgeführt. Die secunet Security Networks AG ist Auftragnehmer für diese Studie. Das vorliegende Dokument stellt den öffentlichen Abschlussbericht der Studie BioP I dar.

4 Testüberblick

4.1 Referenzbasen

Bei der Gesichtserkennung wird eine aktuelle Aufnahme des Gesichts mit einer im Vorfeld gespeicherten Referenzaufnahme der entsprechenden Person, der so genannten Referenzbasis, verglichen. Bezüglich der Bereitstellung dieser Referenzbasis im Hinblick auf Personaldokumente gibt es mehrere Alternativen. Abbildung 2 gibt einen entsprechenden Überblick. Zunächst kann das bereits auf dem Ausweis vorhandene Lichtbild als Referenz verwendet werden. Dies bedeutet, dass bei der Identitätsprüfung einer Person das Lichtbild auf dem Ausweis gescannt und mit der aktuellen Aufnahme verglichen wird. Je nach Ausweistyp ergibt sich eine unterschiedliche Charakteristik des Lichtbildes. So ist beispielsweise das Lichtbild auf dem EU-Visum kleiner und hat mehr optische Störungen als ein Lichtbild auf dem Personalausweis.

Eine Alternative zur Verwendung des Lichtbildes stellt die Bereitstellung der Referenzbasis in elektronischer Form dar. Dies kann entweder eine Bilddatei des Gesichts oder eine spezielle, in der Regel proprietäre Kodierung des Gesichts, das Template, sein. Der Ausweis müsste dann entsprechend um einen Speicherplatz erweitert werden. Aus diesem würde bei der Identitätsprüfung in der Folge die Referenzbasis gelesen.

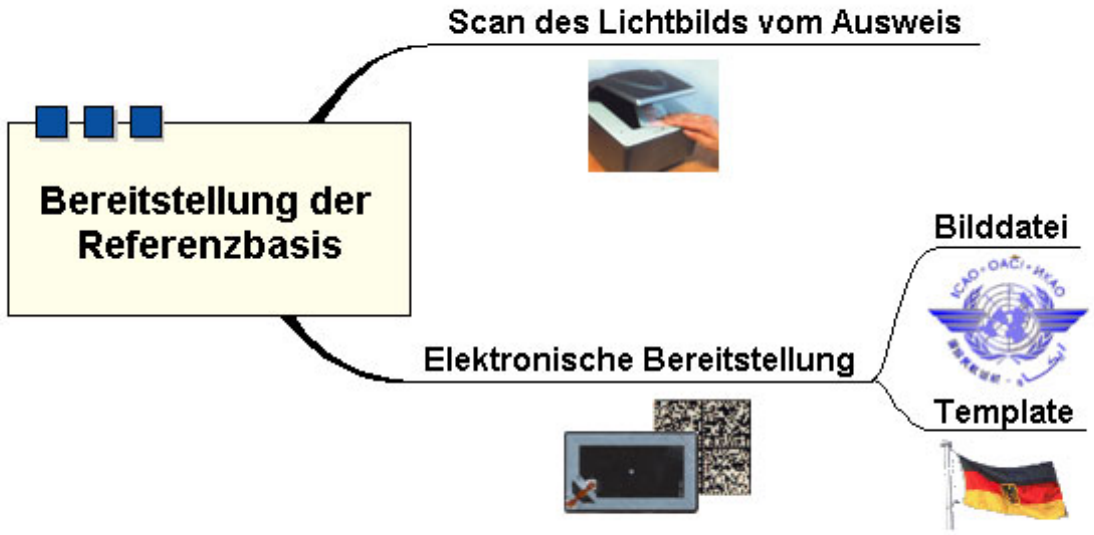

Abbildung 2: Bereitstellung der Referenzaufnahmen für die Gesichtserkennung

Abbildung 2 zeigt die Bereitstellung der Referenzbasen, zum einen als Scan vom Ausweis, zum anderen in elektronischer Form als Bilddatei und Template.

Zur Bewertung dieser Alternativen bezüglich ihrer Eignung für Gesichtserkennung wurden repräsentativ verschiedene Referenzbasen parallel getestet. Dies sind im Einzelnen:

- das Lichtbild auf dem aktuellen Bundespersonalausweis,

- das Lichtbild auf einem speziell für das Projekt angefertigten Musterpersonalausweis mit frontaler Bildaufnahme[3],
- das Lichtbild des EU-Visums,
- die Bilddatei einer frontalen Bildaufnahme,
- die komprimierte Bilddatei einer frontalen Bildaufnahme gemäß ICAO,
- die Bilddatei einer Halbprofilaufnahme sowie
- ein proprietäres Template.

RefID	Beschreibung	Bereitstellung im Zielszenario		
		Lichtbild	**Bilddatei**	**Template**
1	Foto als Frontalaufnahme		X	
2	Foto (1) auf Musterpersonalausweis (Test im Hinblick auf Bundespersonalausweis und Reisepass)	X		
3	Foto (1) auf Visumaufkleber (Test im Hinblick auf EU-Visum, längerfristige Aufenthaltstitel nach EG-Modell sowie vorläufige Reisepässe und Personalausweise)	X		
4	Komprimierte Bilddatei von (1)		X	
5	Foto als Halbprofilaufnahme		X	
6	Foto auf aktuellem Personalausweis	X		
7	System-Template auf Basis Live-Enrolment			X
8	Entspricht RefID 2, aber modif. Einbringen in den Test	X		

Tabelle 1: In BioP I untersuchte Referenzbasen

Die Auswahl dieser Referenzbasen deckt alle oben genannten Ausweistypen ab und erlaubt zudem sehr interessante Vergleiche. Dazu zählen der Einfluss der Kompression des Bildmaterials auf die Gesichtserkennung, die Unterscheidung der Erkennungsleistung für Frontal- und Halbprofilaufnahmen sowie der Vergleich des aktuellen Bundespersonalausweises mit einem für die Gesichtserkennung optimierten Musterausweis.

Den Gesichtserkennungssystemen wurden für jeden Testteilnehmer die in Tabelle 2 dargestellten Bilddateien übergeben. Auf Basis derer wurde ein File-Enrolment für alle beteiligten GE-Algorithmen im Vorfeld des Feldtests durchgeführt. Dagegen erfolgte für Referenzbasis 7 ein Live-Enrolment an den Systemen zur Generierung der System-Templates. Referenzbasis 2 (Musterpersonalausweis) wurde bei jeder Betätigung der Testteilnehmer erneut gescannt und anschließend sofort enroled. Für das Scannen des Ausweisdokuments wurde ein Ausweisleser der Bundesdruckerei – der Verifier – verwendet.

[3] Das Lichtbild des Musterpersonalausweise entspricht den Richtlinien der ICAO für die Erstellung von Passbildern zum Einsatz für biometrische Anwendungen

RefID	Foto	Format	Qualität (Photoshop)	Typische Dateigröße
1	Frontal	JPEG	10	75 kB
2	Scan des Fotos durch Verifier bei jeder Betätigung und Bereitstellung als JPEG für die GE			
3	Frontal	JPEG	10	142 kB
4	Frontal	JPEG	2	14 kB
5	Halbprofil	JPEG	10	75 kB
6	Bundespersonalausweis	JPEG Graustufen	10	65 kB
7	Live-Aufnahme mittels Kamera der GE-Systeme vor Start des Feldtests			
8	Musterpersonalausweis	JPEG Graustufen	10	65 kB

Tabelle 2: Übergebene Bilddateien

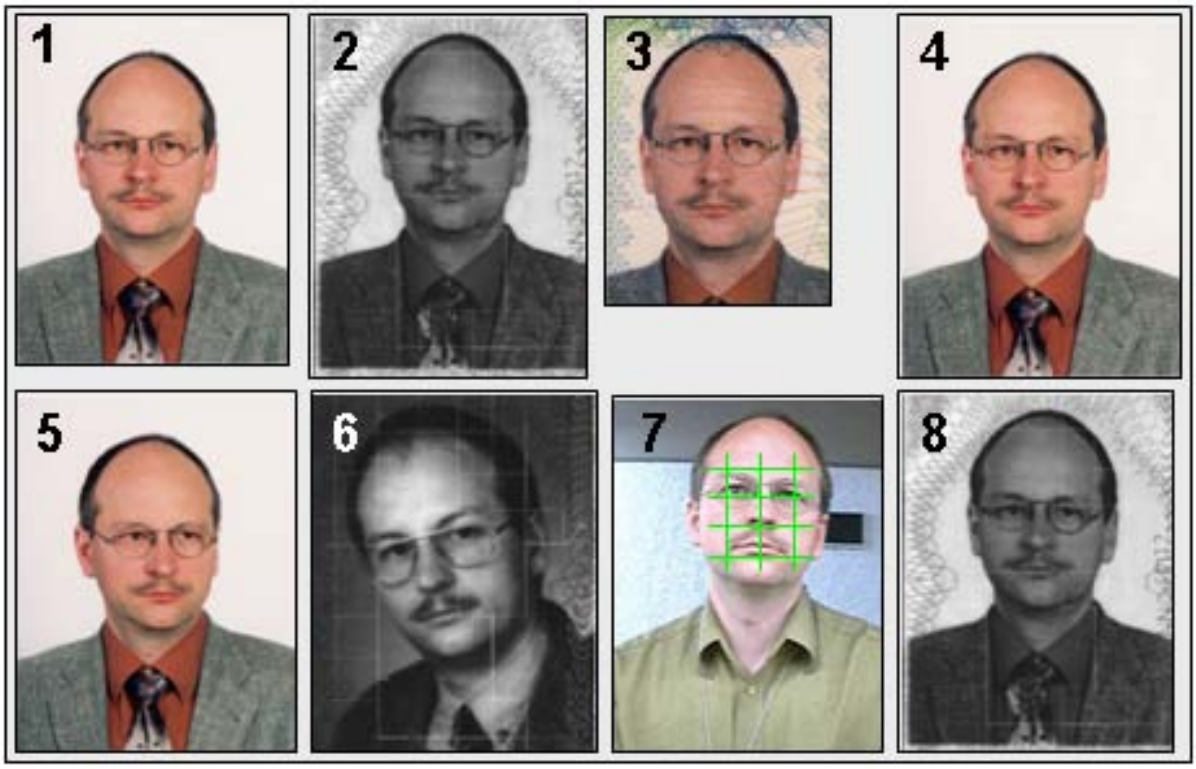

Abbildung 3: Beispielsatz für verschiedene Referenzbasen[4]

Abbildung 3 zeigt einen Beispielsatz für die verschiedenen involvierten Referenzbasen einer Person.

[4] Die Veröffentlichung der Bilder erfolgt mit Zustimmung der dargestellten Person.

4.2 Systeme und Algorithmen

Die Tests wurden für Gesichtserkennungssysteme zweier verschiedener Hersteller durchgeführt, die auf Basis einer Voruntersuchung ausgewählt wurden. Aufgrund der speziellen Zielsetzung des Projekts BioP I konnten keine Standardsysteme der beteiligten Hersteller zum Einsatz kommen. Stattdessen wurden speziell gemäß einem Pflichtenheft angepasste Systeme bereitgestellt, die daher den Status eines Prototyps besitzen.

Das System A erlaubt die Integration und den parallelen Betrieb mehrerer Gesichtserkennungs-algorithmen unabhängig voneinander. In diesem System kamen im Rahmen von BioP I Algorithmen von drei verschiedenen Anbietern zum Einsatz, die jeweils zusätzlich um einen alternativen Gesichtsfinder modifiziert wurden (Plus-Version des jeweiligen Algorithmus).

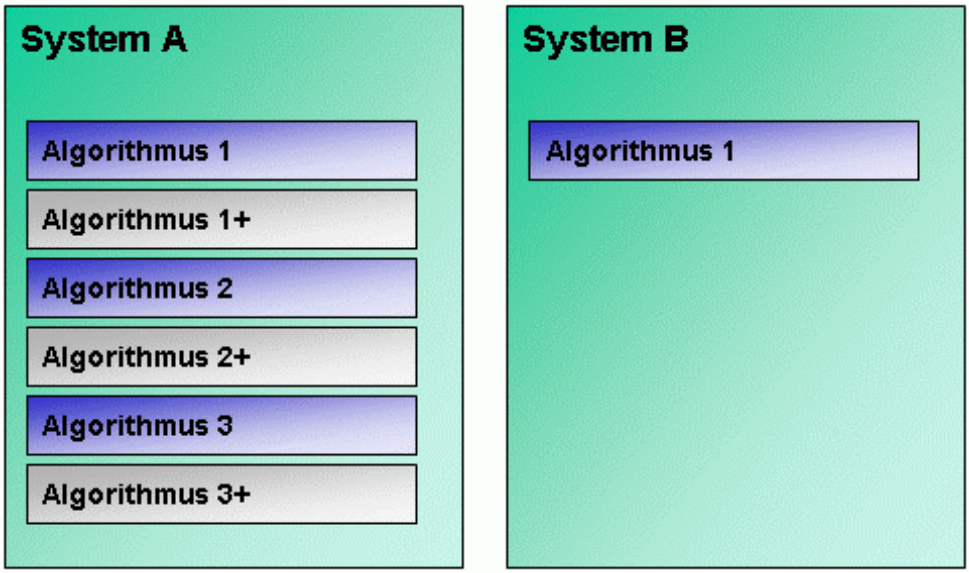

Abbildung 4: Übersicht Systeme und Algorithmen

Abbildung 4 zeigt, dass in System A drei Algorithmen zuzüglich der jeweiligen Plus-Version integriert sind und in System B ein Algorithmus.

Diese Konstellation erlaubt zum einen den Vergleich von Komplettsystemen auf Basis eines identischen Algorithmus, der in beiden Systemen zum Einsatz kommt. Zum anderen können verschiedene Algorithmen innerhalb eines identischen Komplettsystems verglichen werden.

4.3 Funktionsweise

Eine Besonderheit bei der Durchführung des Tests stellte der Einsatz eines Ausweislesers der Bundesdruckerei (Verifier) dar. Die Testteilnehmer legten bei jeder Bedienung ihren Muster-personalausweis auf, der die Personalausweisnummer und das gescannte Lichtbild für die Gesichts-erkennungssysteme bereitstellte.

Der Ablauf für die Bedienung eines Testteilnehmers stellte sich also folgendermaßen dar: Nachdem der Ausweisleser die erforderlichen Informationen des Dokuments erfasst hatte, wurde die Gesichts-erkennung angestoßen. Dabei wurden durch eine Kamera fortlaufend Bilder des Teilnehmers aufgenommen und einem Vergleich mit einer gespeicherten Referenzbasis unterzogen. Die Aufnahme endete, wenn entweder ein Vergleich erfolgreich war oder wenn ein vorab festgelegtes Zeitlimit erreicht wurde. Das Aufnehmen der Bilder wurde dem Teilnehmer mittels eines gelben Lichtsignals angezeigt. Je nachdem, ob der Versuch erfolgreich war oder nicht, wurde dies der Person über ein

grünes bzw. rotes Lichtsignal mitgeteilt. Im Hintergrund und für den Teilnehmer nicht wahrnehmbar erfolgten weitere Vergleiche des aufgenommenen Bildes mit allen im System integrierten Algorithmen und allen Referenzbasen. Die hierbei erzielten Ergebnisse wurden für die spätere Auswertung in einer Datenbank protokolliert. Der Ablauf einer Betätigung aus Sicht des Testteilnehmers ist in Abbildung 5 dargestellt.

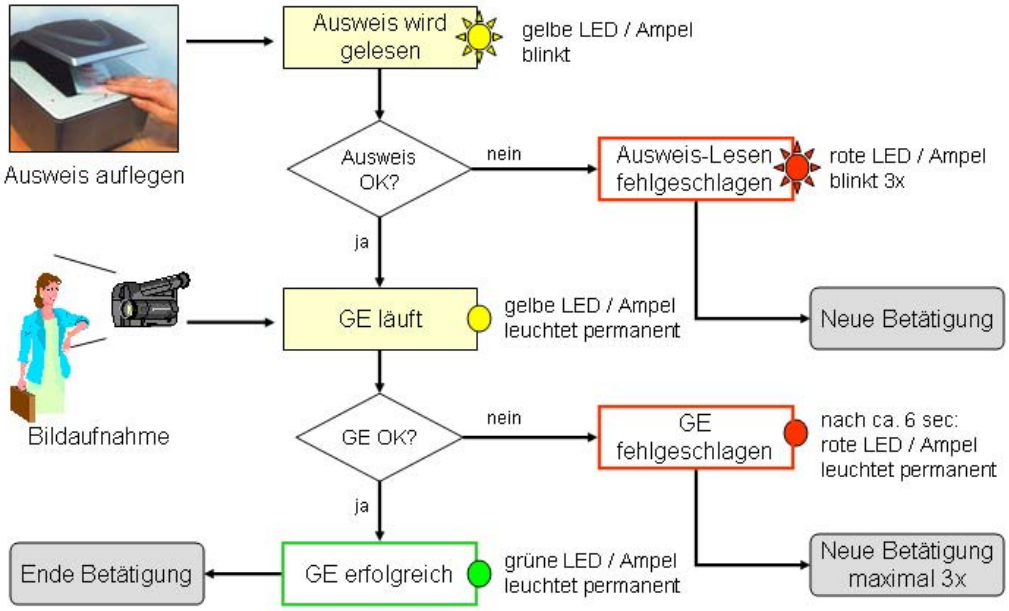

Abbildung 5: Ablauf einer Betätigung

Abbildung 5 zeigt das Ablaufschema einer Betätigung vom Auflegen des Ausweises über die Aufnahme des Gesichts bis zur Ergebnisanzeige.

Aufgrund der parallelen Untersuchung der Erkennungsleistung für unterschiedliche Referenzbasen und außerdem verschiedener GE-Algorithmen wurden durch die Betätigung einer Person eine Vielzahl biometrischer Verifikationen ausgelöst.

Ausgangspunkt ist dabei immer ein auf Basis einer so genannten Masterreferenz sowie eines Masteralgorithmus vorausgewähltes Live-Bild, das bei der Betätigung durch eine Person aufgenommen wird. Aus diesem wird durch die jeweils integrierten Algorithmen ein Template generiert (in den folgenden Abbildungen wird die Templatebildung als Funktion f_{tpl} dargestellt). Für jeden Algorithmus erfolgt dann ein Vergleich mit den bezüglich Person und Algorithmus zugehörigen Templates der verschiedenen Referenzbasen (RefID 1 bis 8). Der erzielte Matchscore wird in einer zentralen Datenbank protokolliert. Die entsprechenden Abläufe sind in Abbildung 6 und Abbildung 7 beschrieben. Die gelb markierten Elemente in der Abbildung zeigen die für den Feldtest festgelegte Konfiguration für Masterreferenz und Masteralgorithmus.

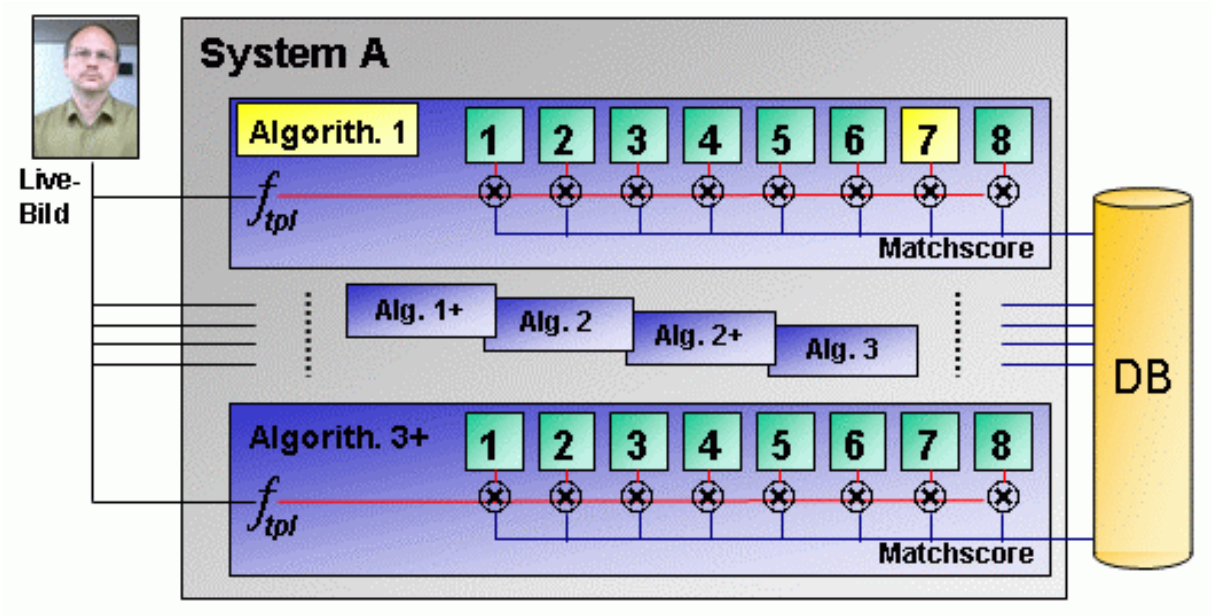

Abbildung 6: Verifikationsprozess bei System A

Abbildung 6 zeigt schematisch wie im System A das Live-Bild gegen alle integrierten Algorithmen und Referenzbasen verglichen wird.

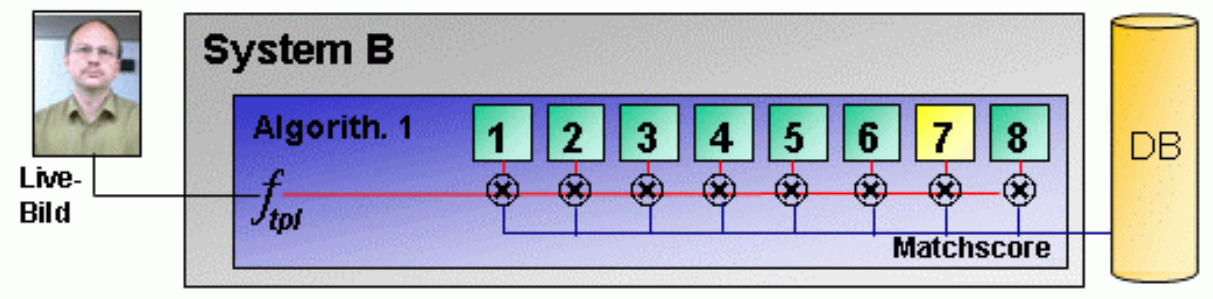

Abbildung 7: Verifikationsprozess bei System B

Abbildung 7 zeigt schematisch wie im System B das Live-Bild mit dem integrierten Algorithmus gegen alle Referenzbasen verglichen wird.

Diese Mehrfachverifikationen bleiben dem Testteilnehmer vollständig verborgen. Es wird lediglich das Verifikationsergebnis für die Masterreferenz und den Masteralgorithmus angezeigt.

4.4 Testbedingungen

4.4.1 Population

Die Teilnehmer des BioP-I-Feldtests rekrutierten sich aus Mitarbeitern des BKA, die sich freiwillig zur Teilnahme bereiterklärt haben. Diese Gruppe umfasst 241 Personen. Mitarbeiter, welche mit der Administration der Systeme und der Betreuung des Feldtests betraut waren, gehörten nicht der Feldtestgruppe an. Die statistischen Merkmale dieser Testgruppe können den folgenden Tabellen und

Diagrammen entnommen werden. Dabei erfolgt ein Vergleich der Struktur der Testpopulation mit der Struktur der Gesamtbevölkerung. Die Auswahl der erfassten statistischen Merkmale Geschlecht, Alter, Bildungsabschluss und ethnische Herkunft beruht auf [TechEval].

Die Population User50 ist eine Teilmenge der Gesamtpopulation. Sie umfasst die Testpersonen, die an beiden Systemgruppen während der Feldtestphase jeweils mindestens 50 unabhängige Versuche vollzogen haben, bei denen für die Gesichtserkennung geeignete Bilder aufgenommen wurden[5]. Die Population User50 umfasst 152 Personen.

Die Struktur der Testgruppe User50 im Vergleich zur Struktur der Gesamtpopulation des Feldtests sowie der Gesamtbevölkerung der Bundesrepublik Deutschland kann den folgenden Tabellen und Diagrammen entnommen werden.

		Männlich	Weiblich	K. Angabe	Gesamt
Testpopulation (gesamt)	Absolut	146	95	0	241
	Relativ [%]	60,58	39,42	0,00	100,00
User50	Absolut	99	53	0	152
	Relativ [%]	65,13	34,87	0,00	100,00
Gesamtbevölkerung	Relativ [%]	48,85	51,15	0,00	100,00

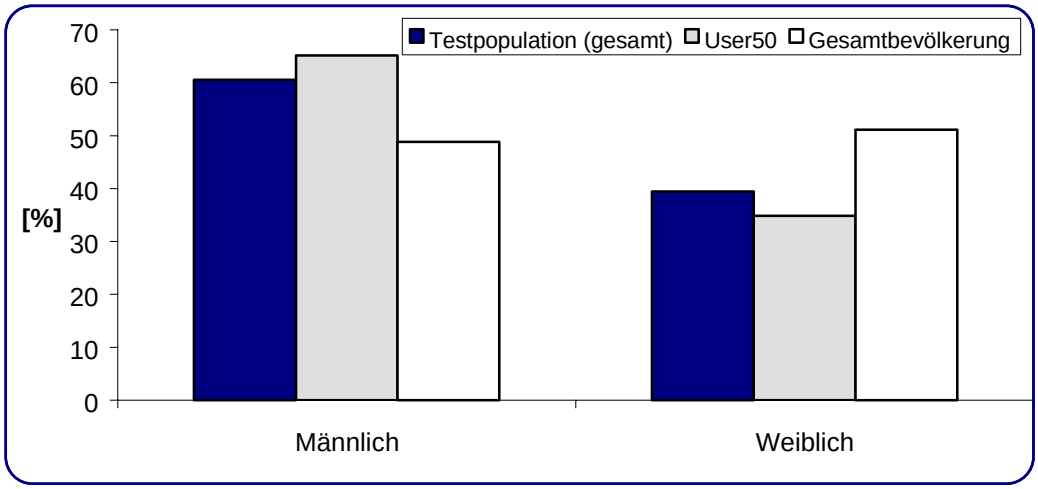

Abbildung 8: Geschlechtsstruktur der Testpopulationen im Vergleich zur Gesamtbevölkerung

Abbildung 8 stellt in einem Diagramm die Verteilung der Testpopulation anhand des Geschlechts dar.

[5] Betätigungen mit ungeeigneten Bildaufnahmen wurden markiert und für einzelne Bewertungen nicht berücksichtigt.

		<18	18-24	25-44	45-59	60-64	≥65	K. A.	Ges.
Testpopulation (gesamt)	Absolut	0	8	140	88	5	0	0	241
	Relativ [%]	0,00	3,32	58,09	36,51	2,07	0,00	0,00	100
User50	Absolut	0	6	92	51	3	0	0	152
	Relativ [%]	0,00	3,95	60,53	33,55	1,97	0,00	0,00	100
Gesamtbevölkerung	Relativ [%]	18,85	7,94	30,70	18,91	6,95	16,65	0,00	100

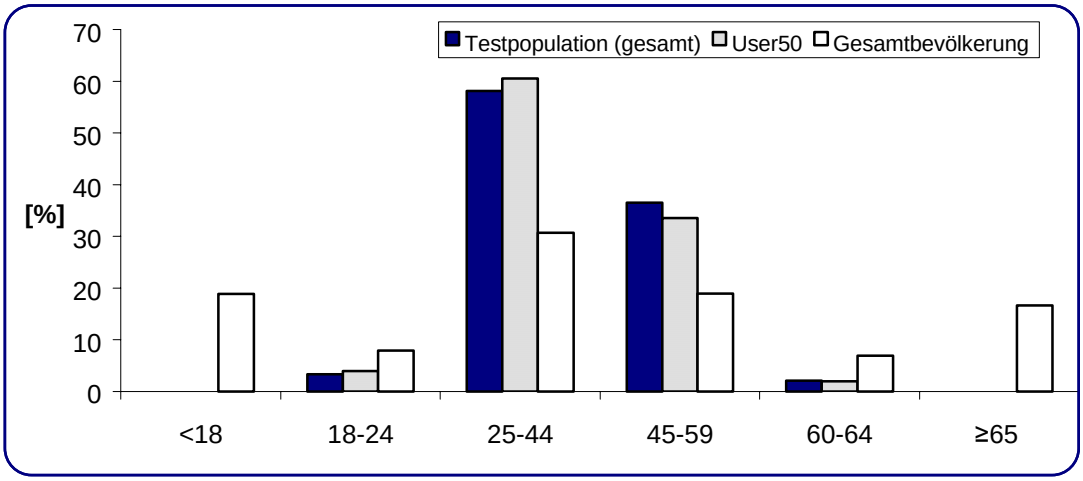

Abbildung 9: Altersstruktur der Testpopulationen im Vergleich zur Gesamtbevölkerung

Abbildung 9 stellt in einem Diagramm die Verteilung der Testpopulation anhand des Alters dar.

		Lehr-ausb.	Fach-sch.	Fach-sch. DDR	FH	Uni	Pro-mo-tion	K. A.	Ges.
Testpopulation (gesamt)	Absolut	74	33	2	66	18	34	14	241
	Relativ [%]	30,71	13,69	0,83	27,39	7,47	14,11	5,81	100
User50	Absolut	40	25	2	42	8	25	10	152
	Relativ [%]	26,32	16,45	1,32	27,63	5,26	16,45	6,58	100
Gesamtbevölkerung	Relativ [%]	52,16	6,59	1,60	3,72	5,94	0,89	29,09	100

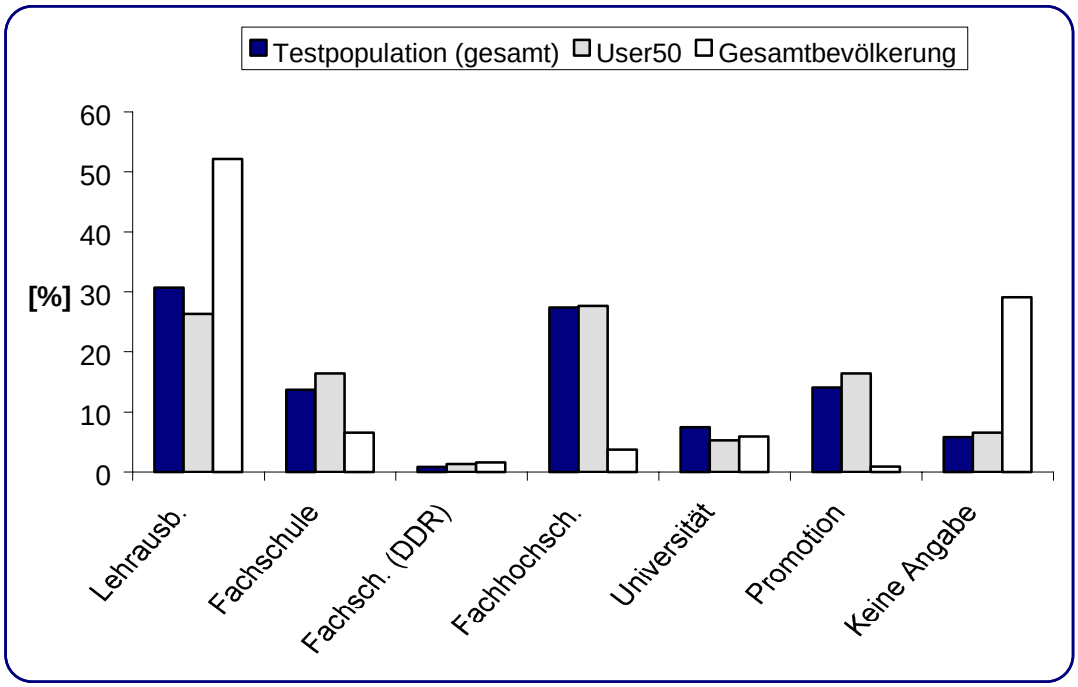

Abbildung 10: Ausbildungsstruktur der Testpopulationen im Vergleich zur Gesamtbevölkerung[6]

Abbildung 10 stellt in einem Diagramm die Verteilung der Testpopulation anhand der Ausbildung dar.

[6] Bei Gesamtbevölkerung entspricht der Wert bei „Keine Angabe“ dem Anteil der Personen ohne Ausbildungsabschluss

		Mittel-europa	Arab., Nordaf., Nahost	Schw.-afrika	Ost-asien	Andere	Keine Angabe	Gesamt
Testpo-pulation	Absolut	231	2	0	1	0	7	241
	Relativ [%]	95,85	0,83	0,00	0,41	0,00	2,90	100,00
User50	Absolut	149	1	0	0	0	2	152
	Relativ [%]	98,03	0,66	0,00	0,00	0,00	1,32	100,00

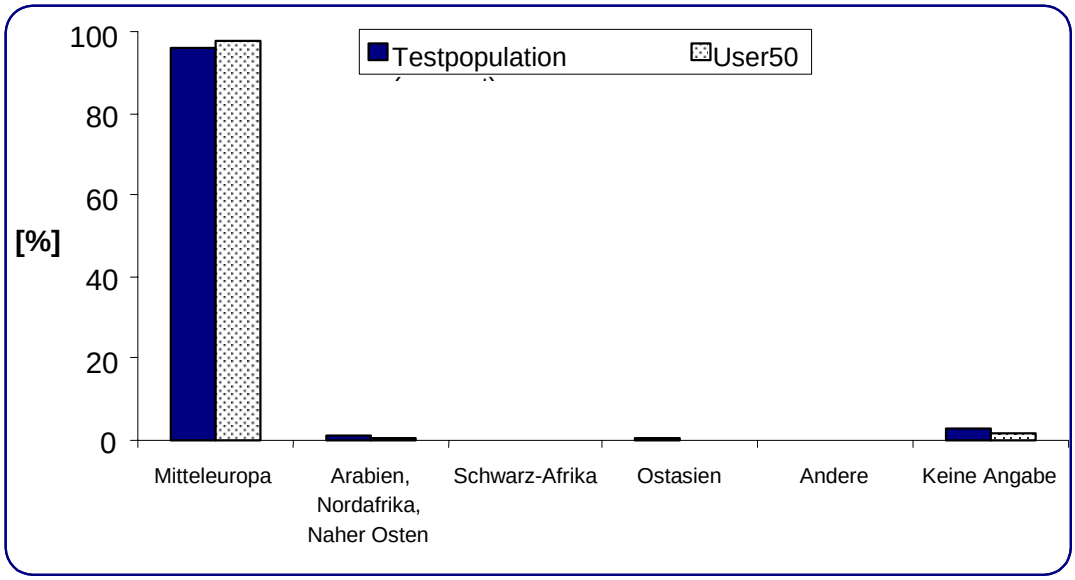

Abbildung 11: Ethnische Herkunft der Testpopulationen[7]

Abbildung 11 stellt in einem Diagramm die Verteilung der Testpopulation anhand der ethnischen Herkunft dar.

Aufgrund der Rekrutierung der Testteilnehmer aus Mitarbeitern des BKA repräsentiert die Testpopulation erwartungsgemäß nicht die Struktur der Gesamtbevölkerung der Bundesrepublik Deutschland.

4.4.2 Testumgebung

Die Gesichtserkennungssysteme wurden in einer Liegenschaft des Bundeskriminalamts in Wiesbaden aufgebaut. Zur Gewährleistung geeigneter Abstände zwischen der zu erfassenden Person und der Kameraeinheit wurden bei System A die Schränke für die Verifier entsprechend positioniert und bei System B Markierungen am Boden angebracht.

Um für die Gesichtserkennung geeignete und für alle Systeme einheitliche Beleuchtungsbedingungen zu schaffen, wurden folgende Maßnahmen durchgeführt:

- Abdeckung der Fenster durch weitgehend lichtundurchlässige Vorhänge

[7] Für das Merkmal ethnische Herkunft kann kein Vergleich mit der Gesamtbevölkerung erfolgen, da seitens des statistischen Bundesamtes lediglich die Staatsangehörigkeit der in Deutschland lebenden Bevölkerung, nicht aber deren ethnische Herkunft erfasst wird.

- Anbringen von Decken- und Wandverkleidungen mit sehr geringer Lichtabsorption
- Gewährleistung einer schlagschatten- und blendfreien Beleuchtung des Gesichts durch Montage indirekt abstrahlender Leuchten über dem Aufnahmebereich jedes Systems
- Gewährleistung von konstanten Beleuchtungsstärken (ca. 130 Lux) durch nicht veränderbare Dauerbeleuchtung (Licht konnte weder an- und ausgeschaltet noch gedimmt werden). Die Beleuchtungsstärke entspricht der Empfehlung gemäß DIN 5035 Teil 2 für Empfangsräume und Räume mit Publikumsverkehr.

Der Lichteinfluss durch die geöffnete Tür wurde von den Herstellern als nicht relevant erachtet. Die zunächst angedachten Maßnahmen zur Gewährleistung eines einfarbigen Hintergrundes, zum Ausschluss von Bewegungen im Hintergrund sowie zur seitlichen Eingrenzung des Aufnahmebereichs wurden nicht ergriffen, da dies nach Aussagen der Hersteller keinen relevanten Einfluss auf die Aufnahme- und Erkennungsleistung der Systeme gehabt hätte.

4.4.3 Architektur des Testaufbaus

Die Rechnersysteme und zugehörige Netzwerkinfrastruktur bestehen im Wesentlichen aus zwei Gruppen:

- Biometrische Systeme der beiden Hersteller
- Hintergrundsysteme für Datensammlung, Auswertung, Administration und Bereitstellung zentraler Funktionen

Das zentrale Datenbanksystem zur Erfassung sämtlicher Ergebnisdaten basiert auf folgender Konfiguration:

- RedHat Linux 8.0
- PostgreSQL 7.3.2
- Zeit-Serverdienst zur Gewährleistung, dass alle Log-Daten und Ergebnisdaten synchron auf einer Zeitbasis geschrieben werden
- RAID5-Conroller, Nutzdatengröße von ca. 600 GB zzgl. 250 GB externem Backup
- 2 GB Hauptspeicher
- Pentium 4 2,533 GHz
- 1 GBit/s Netzwerkanbindung

Alle Systeme wurden in einem autarken Testnetz betrieben. Remote-Zugriffe waren nur von speziellen Stand-alone-Systemen per VPN möglich.

4.4.4 Systemkonfiguration

4.4.4.1 Verifier-Einstellungen

In BioP I wurden die biometrischen Systeme ausschließlich im Verifikationsmodus betrieben. Entsprechend musste bei einer Betätigung immer die UserID der betroffenen Person mitgeteilt werden. Während des Feldtests kam hierzu ausschließlich der Dokumentenleser Verifier der Bundesdruckerei zum Einsatz.

Für den Test wurden sechs identische Verifier eingesetzt. Bezüglich der für die Gesichtserkennung bereitgestellten Bilddaten wurde für alle Systeme identisches Verhalten geprüft und sichergestellt[8]. Die letztlich bereitgestellte Bilddatei hat eine Auflösung von 472 x 620 Pixel (entspricht bei Lichtbildgröße ca. 300 dpi) und eine Farbtiefe von 8 Bit (Graustufen).

4.4.4.2 GE-System-Parameter

Folgende Parameter wurden vor Beginn des Feldtests für beide Systeme festgelegt und während des Tests nicht verändert.

- **Masterreferenz**: Als Referenzbasis, gegen deren Template das Live-Bild bei der interaktiven Betätigung einer Verifikation unterzogen wird, wurde Referenz 7 (beim Live-Enrolment erzeugtes Systemtemplate) gewählt.
- **Masteralgorithmus**: Als Matching Engine, mit der das Live-Bild bei der interaktiven Betätigung gegen die Masterreferenz geprüft wird, wurde Algorithmus 1 gewählt, der in beiden Systemen betrieben wurde.
- **Toleranzschwelle**: Für das Benutzerfeedback wurde die Grenze für den Matchscore, ab welcher eine Verifikation als erfolgreich gewertet wird festgelegt. Die Wahl der Toleranzschwelle erfolgte zunächst gemäß den Empfehlungen der Hersteller. In den Vortests zeichnete sich ab, dass bei Verwendung gleicher Toleranzschwellen System B eine bessere Erkennungsleistung liefert. Um eine diesbezügliche Beeinflussung für die parallel zum Feldtest durchgeführten Benutzerbefragungen auszuschließen, sollten die Systeme jedoch ähnliche Erkennungsleistungen für das Benutzerfeedback aufweisen. Daher wurde die Toleranzschwelle für das System B höher gesetzt. Trotz Nutzung des gleichen Algorithmus in beiden Systemen wurden somit unterschiedliche Toleranzschwellen im Feldtest verwendet.
- **Timeout bei Aufnahme des Live-Bildes**: Die maximale Zeitdauer, während der die Erfassungseinheit fortlaufend Live-Bilder aufzeichnet, wurde auf sechs Sekunden gesetzt.

4.4.4.3 Für GE-Systeme verfügbare Informationen

Die GE-Systeme verfügten zu einer UserID lediglich über die zugehörigen Templates der einzelnen Referenzbasen. Weitere Informationen, wie zum Beispiel die Körpergröße oder andere signifikante Merkmale, wurden nicht zur Unterstützung des Verifikationsprozesses herangezogen.

[8] Durch jeden Verifier wurde ein Ausweisbild der gleichen Person generiert und bezüglich Auflösung, Farbtiefe, Kompression sowie Bildcharakteristik bzgl. Helligkeit, Kontrast und Schärfe verglichen.

5 Testdurchführung

5.1 Feldtest

Nachfolgend werden die Vorbereitung und der Verlauf des Feldtests beim BKA in Wiesbaden beschrieben.

Als erste Aktion für die Testteilnehmer erfolgte die Erstellung der Fotos durch die BKA-Fotostelle.

Der zweite Vorgang war der Start der Herstellerinstallation. Voraussetzung hierfür war die Vorbereitung der Testräume und die Inbetriebnahme der technischen Infrastruktur für den Feldtest. Dies umfasste die Bereitstellung der erforderlichen Stromversorgung und Netzwerkumgebung und den Aufbau der Hintergrundsysteme. Die Hersteller begannen am 09.04.2003 bzw. am 15.04.2003 mit der Inbetriebnahme.

Gemeinsam mit den Herstellern wurden ab Bereitstellung der Funktionalität bis zur TeachIn-Phase einige PreTests zur optimalen Konfiguration der Systeme und Kalibierung auf die Umgebungsbedingungen durchgeführt. In dieser Zeit war es den Herstellern unter Informationspflicht gestattet, Updates einzuspielen und wesentliche Funktionsparameter zu modifizieren.

Am 16.04. bzw. am 22.04.03 wurden die BKA-Administratoren in die Bedienung der Systeme eingewiesen. Die Einweisung konzentrierte sich im Wesentlichen auf die Durchführung des Enrolments.

Am 24.04.03 fand eine Informationsveranstaltung für die Testteilnehmer statt. Hier wurden wesentliche Ziele des Projekts, die Aufgaben und der zu erwartende Aufwand für die Testteilnehmer im Rahmen des Projekts sowie die weitere Zeitplanung erläutert. Des Weiteren wurde über die datenschutzrechtlichen Belange informiert. Abschließend stand das Projektteam für Fragen zur Verfügung.

Direkt im Anschluss an die Veranstaltung begann die schriftliche Befragung zur Benutzerakzeptanz (Erstbefragung). Darüber hinaus wurden eine Mittelbefragung während des Feldtests und eine Abschlussbefragung nach dem Feldtest durchgeführt.

Am 28.04.03 wurde mit dem Live-Enrolment gestartet. Detaillierte Informationen zur Durchführung des Enrolments sind im nachfolgenden Abschnitt 5.1.1 zusammengefasst.

Dem eigentlichen Feldtest wurde noch eine dreitägige TeachIn-Phase vorangestellt. Wesentliche Unterscheidung zum Feldtest war, dass während dieser Phase immer Mitarbeiter des Projektteams in der Testumgebung zugegen waren. Dabei wurden zwei Ziele verfolgt. Zum einen sollten die Testteilnehmer im Umgang mit den Systemen trainiert werden. Zum anderen sollten bestehende Probleme und Fehlerquellen identifiziert und wenn möglich vor Beginn des Feldtests beseitigt werden.

Der Feldtest selbst begann am 15.05.03. Dieser wurde unüberwacht durchgeführt, das heißt Mitarbeiter des Projektteams waren in der Regel nicht zugegen. Im Testraum wurden Formblätter für Teilnehmernotizen ausgelegt. In diesen konnten die Testteilnehmer zum einen wesentliche Änderungen ihres Merkmals angeben (zum Beispiel Änderung bei Brille oder Frisur) und zum anderen Störungsmeldungen bezüglich der GE-Systeme protokollieren.

Der ursprüngliche Ansatz, während des Feldtests keinerlei Updates seitens der Hersteller zuzulassen, konnte nicht verwirklicht werden. Während bei Hersteller B ausschließlich Updates bzgl. der Kommunikation zwischen dem Dokumentenleser Verifier und dem Terminal-PC sowie eines für die Protokollierungs-Funktionalität erfolgten, waren beim System des Herstellers A grundlegende Modifikationen erforderlich. Hier gab es zwar keine Änderungen an den eingebetteten GE-Algorithmen. Die durch den Hersteller bereitgestellte Software zur Erfüllung der für BioP I

geforderten Funktionen wies jedoch einige gravierende Mängel auf. So wurden zum Beispiel fehlerhafte Verifikationsergebnisse geliefert.[9]

Der Feldtest endete am 02.07.2003.

5.1.1 Enrolment

In BioP I kamen zwei verschiedene Arten von Enrolments zum Einsatz. Zur Generierung der Templates auf Basis des Live-Bildes einer Person vor der Erfassungseinheit (RefID 7) wurde ein so genanntes Live-Enrolment durchgeführt. Die Templates für alle anderen Referenzbasen wurden auf Basis von Bilddateien generiert – also mittels File-Enrolment.

5.1.1.1 Live-Enrolment

Das Live-Enrolment erfolgte parallel für beide beteiligten Komplettsysteme. Die Testteilnehmer wurden jeweils in kleineren Gruppen zu festen Terminen eingeladen, um größere Wartezeiten zu vermeiden. Der Großteil der Testteilnehmer wurde zwischen dem 28. und 30.04.03 eingelernt. Für einige wenige Personen fand das Enrolment während der TeachIn-Phase statt.

Das Live-Enrolment wurde durch Mitarbeiter des BKA durchgeführt. Zur Unterstützung wurde den Mitarbeitern ein systemspezifisches Hinweisblatt zum Enrolment sowie ein Protokoll zur Erfassung der Ergebnisse an die Hand gegeben. Einige wesentliche Punkte sind nachfolgend dargestellt:

- Hinweise an die Testperson
  - Positionierung im festgelegten Bereich
  - Richtung Kamera blicken
  - Normaler Gesichtsausdruck (nicht betont fröhlich oder unfreundlich)
  - Hinweis bei Brillenträgern: Enrolment in der gleichen Weise wie Bild auf Musterpersonalausweis
- Berücksichtigung der Qualitätskontrolle der Systeme
- Ausschließen offensichtlich schlechter Aufnahmen
- Durchführen einer Testverifikation direkt im Anschluss an Enrolment (Funktionalität wurde durch einen Hersteller A nicht zur Verfügung gestellt)
- Enrolment gilt nach vier erfolglosen Versuchen als gescheitert

Insbesondere während der TeachIn-Phase und während der ersten Tage des Feldtests wurde darauf geachtet, ob einzelne Personen häufiger als andere zurückgewiesen wurden. Sofern kein fehlerhaftes Verhalten seitens der Teilnehmer festgestellt werden konnte, wurde ein Re-Enrolment durchgeführt. Bei jedem System wurden für drei Personen Re-Enrolments durchgeführt. Alle durchgeführten Re-Enrolments haben anschließend zu besseren Erkennungsergebnissen geführt.

5.1.1.2 File-Enrolment

Beim File-Enrolment werden den GE-Systemen Bilddateien übergeben, aus denen anschließend die Templates generiert werden. Die UserID des zum Foto gehörenden Testteilnehmers wird dabei aus dem Dateinamen abgeleitet. Eine Übersicht der bereitgestellten Bilddateien ist in Tabelle 2 dargestellt.

[9] Durch das Speichern aller Live-Bilder aus den Betätigungen der Testteilnehmer konnten die Verifikationen wiederholt und somit die korrekten Ergebnisse erzielt werden.

5.2 Weiterführende Untersuchungen

Im Labor der Firma secunet Security Networks AG in Essen wurden weitere Tests durchgeführt. Insbesondere umfasste dies folgende Schwerpunkte

- Untersuchung von Einflussfaktoren auf die Gesichtserkennung (Lichteinfluss)
- Reduzierung des Speicherplatzbedarfs bei der Bereitstellung der Referenzbasen
- Prüfung der Überwindungssicherheit der beteiligten Systeme
- Offline-Tests zur Bestimmung von Falschakzeptanzen

Parallel zum Feldtest wurde die Entwicklung der Benutzerakzeptanz untersucht.

6 Auswertung der Feldtestergebnisse

6.1 Auswertungskonzept

Nachfolgend wird das Auswertungskonzept von BioP I dargestellt, das die verschiedenen Projektziele berücksichtigt. Die Auswertung orientiert sich bezüglich Begriffen und Methoden grundsätzlich an [BestPrac], sofern dort entsprechende Themen behandelt werden. Abweichungen erfolgen nur in begründeten Fällen und es erfolgt ein entsprechender Hinweis.

6.1.1 Vergleichstypen

Die Zielsetzung von BioP I beinhaltet sowohl einen Vergleich verschiedener Gesichtserkennungssysteme als auch einen Vergleich bezüglich der zugrunde liegenden Referenzbasen. Durch die Systemauswahl für BioP I wurde außerdem noch der Vergleich verschiedener GE-Algorithmen innerhalb eines identischen Systems möglich.

Insgesamt können also folgende Vergleiche auf Basis der in BioP I ermittelten Ergebnisse durchgeführt werden (siehe Abbildung 12):

1. Vergleich von biometrischen Komplettsystemen unter vergleichbaren Bedingungen (GE-Algorithmus, Umgebungsbedingungen, Testzeitraum) = Systemvergleich (Horizontaler Vergleich 1 in Abbildung 12)
2. Vergleich von GE-Algorithmen unter vergleichbaren Bedingungen (identisches System, identische Enrolment-Bilder, identische Live-Bilder) = Algorithmenvergleich (Horizontaler Vergleich 2 in Abbildung 12)
3. Vergleich von verschiedenen Referenzbasen innerhalb eines Systems (identische Live-Bilder) = Referenzbasenvergleich (Vertikaler Vergleich in Abbildung 12)

Gemäß [BestPrac] ist der Systemvergleich und somit der Feldtest wie er sich für die Teilnehmer darstellt eine „Szenario-Evaluation“. Der Algorithmenvergleich ist dagegen vom Typ „Technische Evaluation“. Für den Algorithmenvergleich ist deshalb eine Einschränkung der durchgeführten Verifikationen erforderlich, um Fehler auszuschließen, die nicht den Algorithmen anzulasten sind. Für Gesichtserkennung ungeeignete Live-Bilder aufgrund von system- und benutzerspezifischem Fehlverhalten müssen ausgeschlossen werden. Dies wurde in BioP I sichergestellt (vergleiche Abschnitt 6.2.1).

Für den Referenzbasenvergleich stehen ebenfalls Fehler des Systems und der Benutzer im Hintergrund. Deshalb wurden die betreffenden Ergebnisse auf der gleichen Datenbasis wie der Algorithmenvergleich ermittelt.

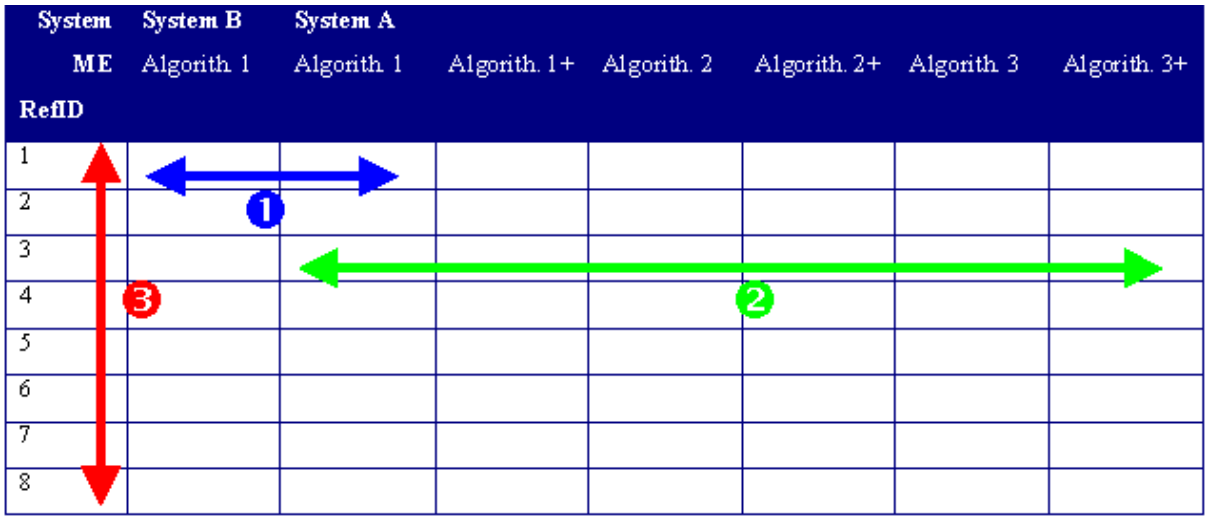

Abbildung 12: Vertikale und horizontale Vergleiche

Abbildung 12 gibt einen grafischen Überblick über die drei Vergleichstypen.

6.1.2 Bewertung der Erkennungsleistungen

Wesentliches Kriterium zur Bewertung von biometrischen Systemen ist deren Erkennungsleistung. Diese ergibt sich aus den Wahrscheinlichkeiten, dass eine berechtigte Person vom System zurückgewiesen wird und dass eine unberechtigte Person vom System akzeptiert wird. Diese Wahrscheinlichkeiten werden als False Rejection Rate (FRR) und False Acceptance Rate (FAR) bezeichnet. Die Werte für FRR und FAR sind nicht theoretisch herleitbar, sondern müssen immer statistisch auf Basis aufwendiger Tests ermittelt werden.

Da die Werte für FRR und FAR in einem Testszenario immer korrelieren, müssen zur Angabe der Erkennungsleistung eines Systems immer beide Werte angegeben werden. Dies erfolgt für so genannte Arbeitspunkte, an denen für einen festen Threshold die zugehörigen Werte für FRR und FAR berechnet werden. Singuläre Angaben von FAR und FRR sind wenig sinnvoll, da dann immer jeweils gute Arbeitspunkte herausgesucht werden können. So ist die Angabe einer niedrigen FRR nicht aussagekräftig, wenn der zugehörige Wert der FAR (der dann in der Regel hoch ist) nicht bekannt ist.

Per Definition berechnet sich eine FRR immer aus Verifikationen berechtigter Personen, die FAR dagegen aus Verifikationen Unberechtigter. Dies bedeutet für BioP I, dass zur Angabe von Erkennungsleistungen beide Vorgänge in hoher Anzahl durchzuführen sind, um statistisch signifikante Aussagen treffen zu können.

Zur Bestimmung der FRR finden die Verifikationen der Feldtestteilnehmer über den Feldtestzeitraum Verwendung. Die Ermittlung der FAR erfolgt dagegen auf Basis von im Feldtest gewonnenen Live-Bildern, die zur Verifikation gegenüber den Referenzen anderer Personen genutzt werden. Damit wird eine große Anzahl von Verifikationen Unberechtigter simuliert.

Für die Darstellung der Erkennungsleistung von biometrischen Systemen gibt es mehrere Verfahren. Für BioP I finden drei verschiedene Methoden Verwendung, die nachfolgend erläutert werden.

Eine einfache und interessante Methode ist die **Darstellung der relativen Häufigkeiten von Matchscores** (Genuine-Impostor-Frequency-Diagram). Ein Matchscore ist der erzielte Trefferwert beim Vergleich des aktuell generierten Templates der zu authentifizierenden Person mit dem gespeicherten Template aus dem Enrolment. Berechtigte Personen erzielen typischerweise hohe Matschscores, Unberechtigte in der Regel niedrige. Bei dieser Darstellung werden auf der Abszisse

alle im Test angefallenen Werte von Matchscores aufgetragen. Die Ordinate enthält die jeweils zugehörigen relativen Häufigkeiten des Auftretens (also die absolute Anzahl des Auftretens eines Matchscore-Wertes normiert bezüglich der Gesamtzahl aller Matchscores). Diese Werte werden jeweils für Berechtigte und Unberechtigte aufgetragen. Im Idealfall zeigen beide Verteilungskurven keine Überlappung.

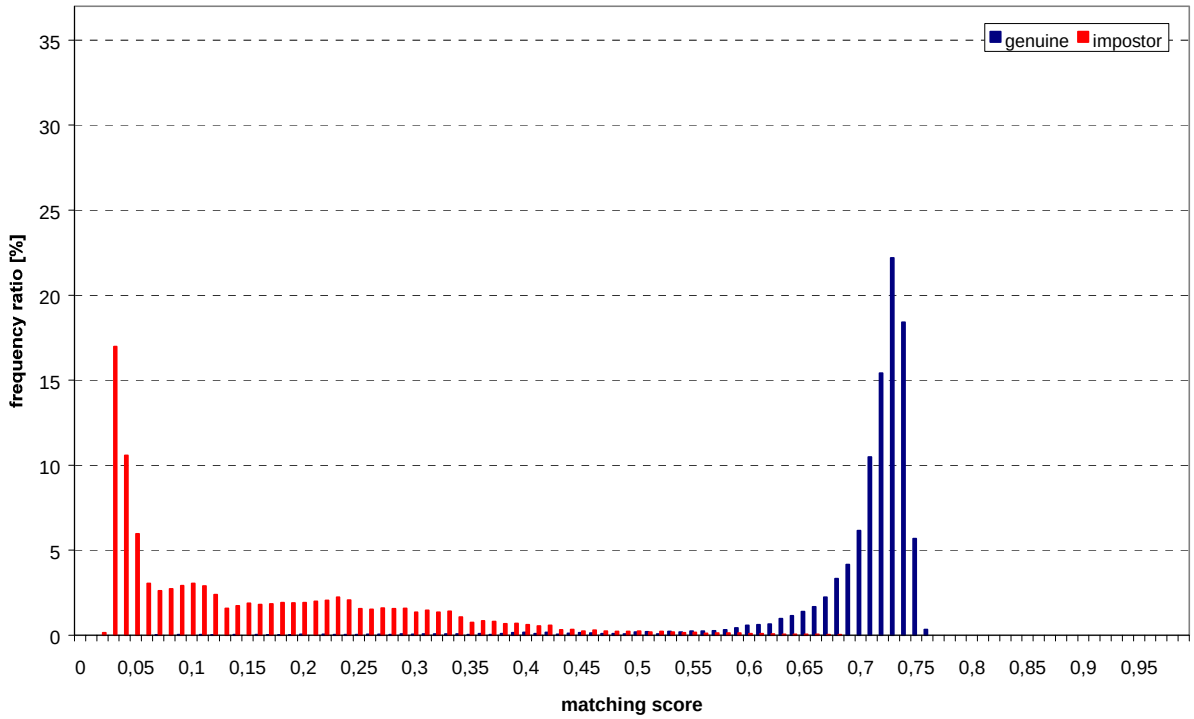

Abbildung 13: Beispiel eines Genuine-Impostor-Frequency-Diagram

Abbildung 13 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten dar.

Aus diesem Histogramm lassen sich recht einfach FAR und FRR ableiten. Legt man einen Matchscore als Toleranzschwelle (Threshold) zur Unterscheidung von Berechtigten und Unberechtigten fest, dann ergibt sich die FAR aus der Anzahl der oberhalb dieser Schwelle liegenden Matchscores von Unberechtigten im Verhältnis zur Gesamtanzahl der Versuche bzw. angefallenen Matchscores. Umgekehrt ergibt sich die FRR aus der unterhalb der Schwelle liegenden Anzahl von Matchscores von Berechtigten im Verhältnis zur zugehörigen Gesamtanzahl. Somit können sukzessive aus diesen Verteilungskurven FAR-FRR-Kurven ermittelt werden. Diese stellen die Fehlerraten in Abhängigkeit des Thresholds dar.

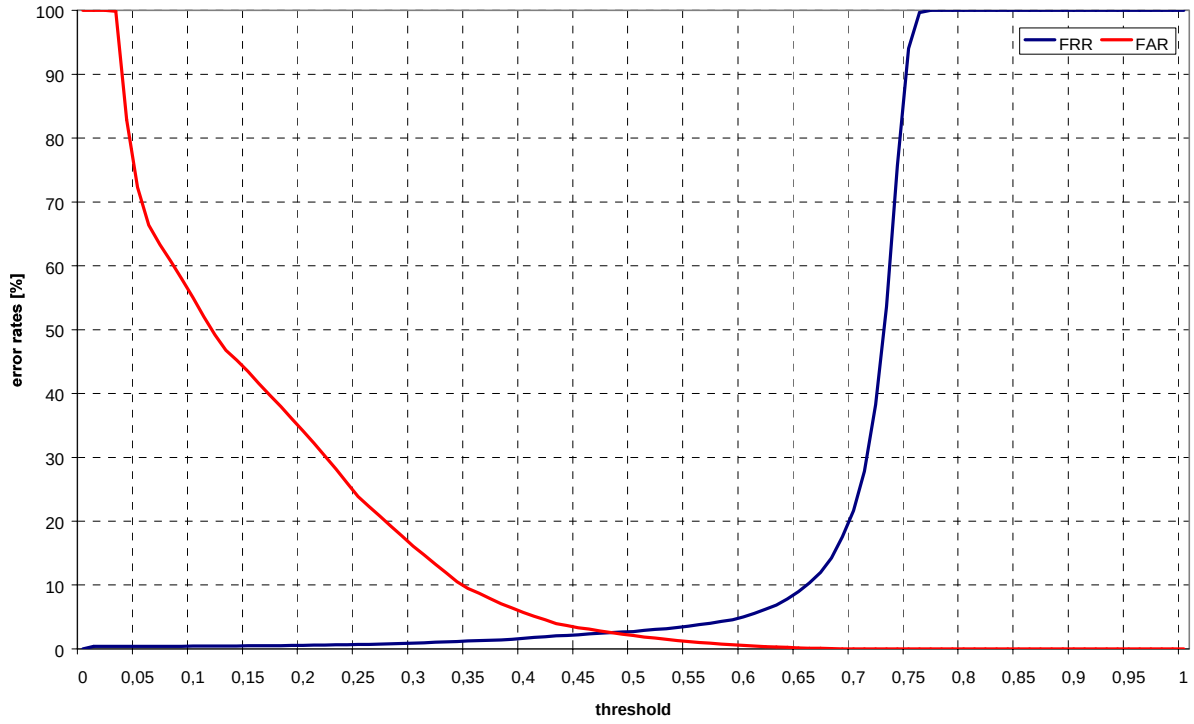

Abbildung 14: Beispiel eines FAR-FRR-Diagramm

Abbildung 14 stellt in einem Diagramm die FAR und FRR in Abhängigkeit des Schwellwertes gegenüber.

FAR-FRR-Diagramme sind die verbreitetste Darstellung der Erkennungsleistungen eines biometrischen Systems. Da sie leicht verständlich sind, werden sie auch für BioP I verwendet. Diese Darstellung ist insbesondere für die Festlegung eines Thresholds eines Systems für einen bestimmten Einsatzzweck geeignet. Absolute Aussagen über die tatsächliche Leistungsfähigkeit des Systems und insbesondere der Vergleich verschiedener biometrischer Systeme sind mit diesem Diagrammtyp nur beschränkt möglich. Das liegt vor allem daran, dass auftretende Matchscores für verschiedene Algorithmen sehr unterschiedlich realisiert sind. Damit sind die Matchscores und entsprechend die resultierenden Thresholds für unterschiedliche Systeme nicht vergleichbar. Für die Verteilung der Matchscores sind beliebige Skalierungen und Transformationen realisierbar, die das Aussehen von FAR-FRR-Kurven beeinflussen. Häufig werden beispielsweise einzelne Arbeitsbereiche der Kurve gedehnt, um das System robuster gegen Schwellwertänderungen erscheinen zu lassen. Diese Methoden beeinflussen jedoch nicht das Verhältnis der Werte FAR und FRR zueinander. Dies motiviert die direkte Darstellung der FRR in Abhängigkeit der FAR. Damit wird der Parameter Matchscore eliminiert und es wird eine von Schwellwertskalierungen unabhängige Darstellung erreicht, die eine echte Vergleichbarkeit unterschiedlicher biometrischer Systeme oder auch Systemkonfigurationen ermöglicht.

Diese Darstellung wird **ROC-Kurve** (Receiver Operating Characteristic) genannt. Die FRR wird als Funktion der FAR dargestellt. Die ideale ROC-Kurve nimmt nur Werte auf den Koordinatenachsen an ($FAR \neq 0 \Rightarrow FRR = 0$ und umgekehrt). Der oberste Punkt ist für alle Systeme durch FAR=0% und FRR=100% gegeben. ROC-Kurven können per Definition nicht ansteigen. Generell gilt, je enger die Kurve eines Systems an den Achsen liegt, umso besser ist die Erkennungsleistung. Die Equal Error

Rate (EER), bei der FAR=FRR, leitet sich aus dem Schnittpunkt der ROC-Kurve mit der Diagonalen des Koordinatensystems ab.

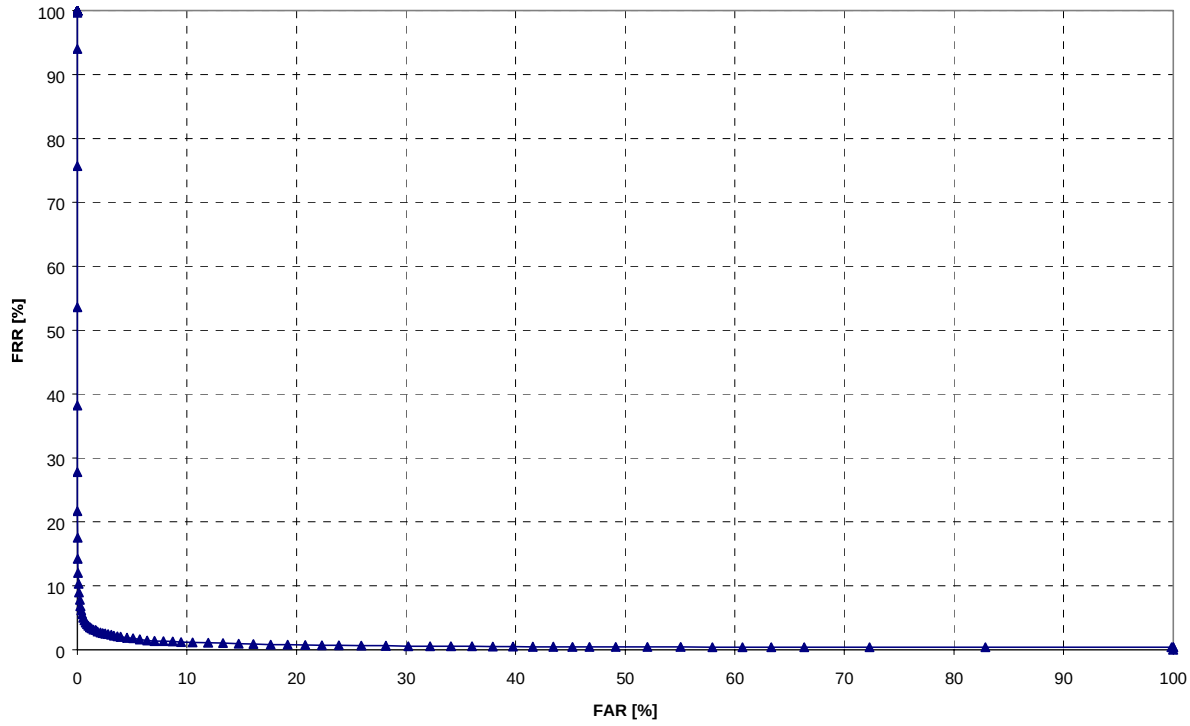

Abbildung 15: Beispiel eines ROC-Diagramm

Abbildung 15 stellt in einem Diagramm eine ROC-Kurve dar.

In BioP I werden ROC-Kurven ermittelt, indem für den gesamten Bereich möglicher Thresholds die jeweils zugehörige FAR und FRR berechnet wird. Damit wird die vollständige Leistungsfähigkeit dargestellt. Ein Vergleich verschiedener Systeme ist damit einfach möglich, indem verschiedene Kurven innerhalb eines Diagramms dargestellt werden. Da sich die resultierenden Kurven nicht notwendig hintereinander anordnen, so dass ein einfaches Ranking möglich ist, kann die Erkennungsleistung anhand einzelner Arbeitspunkte bewertet werden. Die Festlegung der Arbeitspunkte ist in Abschnitt 6.1.2.1 dargestellt.

Für ROC-Kurven zum Vergleich von biometrischen Systemen existieren in der Fachliteratur unterschiedliche Definitionen. Dort findet man beispielsweise die Auftragung von (1-FRR) statt FRR auf der Ordinate oder es werden die Fehlerraten FMR, FNMR (False Match Rate, False Non Match Rate) verwendet. [BestPrac] empfiehlt für den in BioP I angestrebten Systemvergleich die oben erläuterte Darstellung. Diese wird dort als ROC-Variante Detection-Error-Trade-off-Kurve (DET-Kurve) bezeichnet.

In BioP I werden keine kumulierten FRRs[10] bestimmt (zum Beispiel die Rückweisungsrate nach der zweiten Betätigung), sondern nur die Rückweisungsrate bei der ersten Betätigung. Die Durchführung mehrerer Erkennungsversuche ist bereits durch die Funktionsweise der beteiligten Systeme bedingt. Bei Nichterkennung werden automatisch neue Versuche bis zur Erreichung eines Timeout angestoßen.

[10] Kumuliert im Sinne von Zusammenfassung mehrer Betätigungen zu einem Versuch

6.1.2.1 Algorithmenvergleich und Referenzbasenvergleich

Die Erkennungsleistung eines biometrischen Systems muss immer in der Kombination FRR und zugehörige FAR angegeben werden. Um die beteiligten Algorithmen und Referenzbasen vergleichen zu können, werden hierzu Arbeitspunkte festgelegt. Diese Festlegung kann sich an fixen Werten für die FAR oder für die FRR orientieren. Für BioP I sind beide Richtungen von Interesse. Der Schwerpunkt der Betrachtung liegt jedoch auf dem Aspekt der Sicherheit, also einer niedrigen FAR. Folgende Punkte wurden innerhalb der Projektgruppe als betrachtenswert identifiziert:

- FAR: 0,01% 0,1% 1%
- FRR: 1% 2% 5%

Die Einordnung und Bewertung dieser Arbeitspunkte gemäß den Technischen Evaluierungskriterien des BSI [TechEval] ist Tabelle 3 zu entnehmen.

Fehlerrate	Wertebereich	Bewertung gemäß Kriterienkatalog
FAR	< 0,3%	Sehr stark
	0,3% - 1%	Stark
	1% - 5%	Mittel
	> 5%	Schwach
FRR	< 1%	Sehr stark
	1% - 3%	Stark
	3% - 7%	Mittel
	> 7%	Schwach

Tabelle 3: Einordnung von Fehlerraten

Bei einem Vergleich muss jeweils ein FAR-Wert oder FRR-Wert ausgewählt werden und für das betreffende System bzw. den betreffenden Algorithmus die korrespondierende Fehlerrate berechnet werden. Das Ergebnis kann dann als Vergleichskriterium verwendet werden.

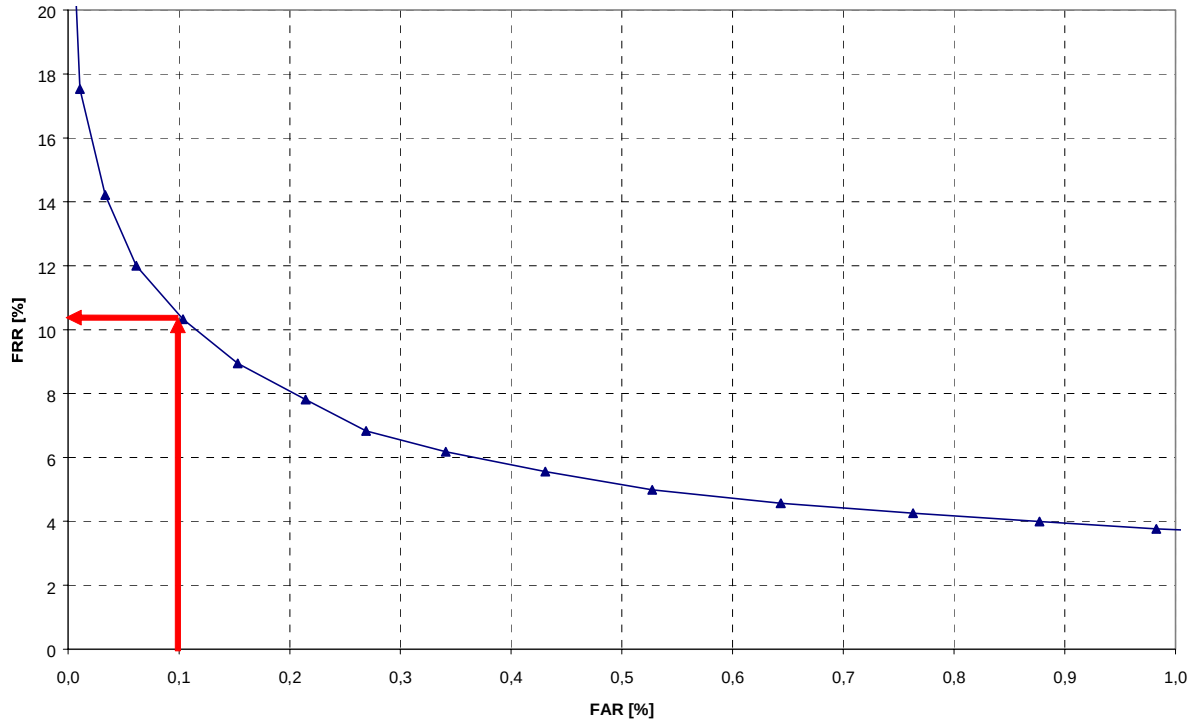

Abbildung 16: Arbeitspunkt FAR = 0,1%

Abbildung 16 stellt in einem Diagramm den Arbeitspunkt bei FAR = 0,1% auf einer ROC-Kurve dar.

In BioP I konzentrieren sich die meisten Untersuchungen auf den Arbeitspunkt FAR = 0,1%.

6.1.2.2 Systemvergleich

Für die Bewertung der Erkennungsleistung von biometrischen Komplettsystemen wird in BioP I ein identischer GE-Algorithmus in verschiedenen Komplettsystemen eingesetzt. Die Ermittlung der Erkennungsleistung erfolgt für eine in beiden Systemen identisch festgelegte Toleranzschwelle (Threshold). Für diesen Threshold werden die Fehlerraten FRR und FAR berechnet. Die Ergebnisse können dann als Vergleichskriterium verwendet werden.

6.2 Testergebnisse

6.2.1 Definition der Basisdatenmengen

In BioP I wurden sämtliche Betätigungen in der Datenbank protokolliert, sofern die UserID der betreffenden Person bekannt war. Nachfolgend wird die Einschränkung der Gesamtmenge von Betätigungen auf die für die jeweilige Untersuchung erforderlich Teilmenge beschrieben.

In Abhängigkeit der zu ermittelnden Ergebnisse ist der Ausschluss einzelner Betätigungen erforderlich. Im Folgenden sind entsprechende Kriterien aufgezählt:

1. Betätigungen außerhalb des definierten Feldtestzeitraums
2. Betätigungen von Projektmitarbeitern / Administratoren

3. Durch variierte Umgebungsbedingungen signifikant beeinflusste Ergebnisse (während des Feldtests war ein Beleuchtungsausfall im Testraum zu verzeichnen)
4. Durchführung einer Folge von Betätigungen an einem System durch eine Person (bewirkt nicht zu vernachlässigende statistische Abhängigkeiten nach der ersten Betätigung)
5. Systeme befinden sich in undefiniertem Zustand (beim System A trat mehrfach der Fall auf, dass Systemstatus und Benutzerfeedback nicht konsistent waren)
6. Fehlfunktion der Kamera
7. Vertauschung von Ausweisen durch Feldtestteilnehmer (Spieltrieb)
8. Nicht kooperatives Verhalten der Testteilnehmer (bewusst variiertes Aussehen durch Grimassen, Sonnenbrille etc.)

Während die Kriterien 1 bis 4 auf einfache Weise technisch ausgeschlossen werden können, ist dies bei den Kriterien 5 bis 8 nur durch manuelle Unterstützung möglich. Zur Markierung entsprechender Fälle wurde ein spezieller Datenbankreport implementiert, der eine auf Basis von Erkennungsleistungen parametrisierte Auswahl aus der Gesamtmenge trifft. Mit Hilfe der durch diesen Report dargestellten Informationen kann zum einen eine Klassifikation des entsprechenden Bildes erfolgen. Zum anderen können Ausweisvertauschungen durch Gegenüberstellung von zugehörigen Enrolment-Bildern identifiziert werden. Kriterien zur Klassifizierung der durch diesen Report dargestellten Live-Bilder sind in Tabelle 4 dargestellt. Die Anwendung des Reports führte zu den in der Tabelle angegebenen Häufigkeiten.

Auswahlcode	Klassifikation	Häufigkeit System A	Häufigkeit System B
0	keine Person	115	12
1	kein Gesicht	69	0
2	Teilgesicht	306	40
3	Kopfhaltung	19	1
4	zu dunkel	18	1
5	zu hell	0	1
6	Tausch	12	9
7	Sonnenbrille / Grimasse	8	6
8	mehr Personen	10	0
9	Kamera unscharf	57	0

Tabelle 4: Kriterien zur Klassifizierung ungeeigneter Live-Bilder

Beim Systemvergleich kommt nur ein Teil der beschriebenen Ausschlusskriterien zum Einsatz, da es sich um eine „Szenario-Evaluation“ gemäß [BestPrac] handelt. Ausgeschlossen werden Betätigungen mit Ausweistausch sowie Live-Bilder, bei denen keine Person im Bild erfasst ist. Letzteres trat im Wesentlichen durch einen Fehler auf, bei dem Systemstatus und Benutzerfeedback inkonsistent waren. Vereinzelt traten Bilder ohne Personen durch beabsichtigte Fehlbedienung auf.

Gemäß dem Auswertungskonzept in Abschnitt 6.1 müssen beim Algorithmen- und Referenzbasenvergleich für Gesichtserkennung ungeeignete Live-Bilder ausgeschlossen werden, die aufgrund von

system- und benutzerspezifischem Fehlverhalten entstanden sind. Somit kommen alle genannten Kriterien zur Anwendung.

Die relevanten Einschränkungskriterien in Abhängigkeit des durchzuführenden Vergleichs sind in Tabelle 5 zusammengefasst.

Betätigungen	Kürzel	Einschränkungskriterien	Anzahl System A	Anzahl System B
Gesamtmenge	Over_All_Set	Feldtestzeitraum Testteilnehmer (keine Mitarbeiter des Projektteams) Nur erste Betätigung eines Versuchs	14532	14176
Menge für Systemvergleich (Szenario-Evaluation)	Scenario_Attempt_Set	Einschränkungen entsprechend Over_All_Set Testteilnehmer mit mehr als 50 Betätigungen an beiden Systemen (User50) Nur Verifikationen mit Matchscore ≥ 0 Bereinigung der Live-Bilder gemäß Auswahlcode 0 und 6	11028	10886
Menge für Algorithmenvergleich und Referenzbasen-vergleich (Technische Evaluation)	Tech_Attempt_Set	Einschränkungen entsprechend Scenario_Attempt_Set Bereinigung der Live-Bilder gemäß Auswahlcode 0 bis 9	10680	10865
Menge für FAR-Bestimmung	FAR_Set	Verifikationen mit Live-Bildern Unberechtigter gegen die Referenzbasen der jeweils anderen Teilnehmer	57120	57120

Tabelle 5: Betätigungsmengen für die Auswertung

6.2.2 Failed Enrolment Rate (FER)

Vor dem Hintergrund des Zielszenarios ist es von eminenter Bedeutung, dass alle Personen enrolt werden können. Bei biometrischen Systemen ist dies nicht immer gegeben. Einige Algorithmen nehmen bereits im Vorfeld der Template-Generierung eine Bewertung der Bildqualität vor. Die Template-Generierung erfolgt nicht nur beim Enrolment. Auch bei jeder Betätigung wird aus dem aufgenommenen Live-Bild ein Template gebildet. Schlägt die Template-Bildung fehl, wird dies in einer realen Einsatzumgebung als Falschrückweisung interpretiert. Im Feldtest von BioP I wurden diese Effekte separat betrachtet und gingen nicht in die FRR ein. Daher ist die FER bei der Bewertung der Erkennungsleistung mit zu berücksichtigen.

Beim Live-Enrolment war für alle Algorithmen die FER=0. Eine FER>0 trat nur bei Algorithmus 1 beim File-Enrolment auf. Insbesondere bei den Bildern vom aktuellen Bundespersonalausweis der Testpersonen schlug aufgrund der schlechten Bildqualität das Enrolment in 7% der Fälle fehl.

6.2.3 Erkennungsleistungen

6.2.3.1 Verifikationen Unberechtigter

Zur Ermittlung von statistisch aussagekräftigen Falschakzeptanzraten ist eine große Anzahl von Verifikationen unberechtigter Personen erforderlich.

Ein Vergleich aller gespeicherten Live-Bilder gegen alle gespeicherten Referenztemplates war aus Kapazitätsgründen (Zeit- und Speicherbedarf) nicht durchführbar. Deshalb wurde für alle Testteilnehmer, die mindestens eine Betätigung durchgeführt haben (238 Personen), je ein repräsentatives Live-Bild[11] von System A und System B ausgewählt. Diese wurden in einem Batch-Lauf in dem jeweiligen System gegen alle Referenztemplates aller Testteilnehmer verglichen. Die Ergebnisse wurden für die spätere Ermittlung der FAR in der zentralen Ergebnisdatenbank protokolliert.

Hierbei ist zu beachten, dass jeweils nur ein Input-Bild für die Verifikation eines Unberechtigten übergeben wurde und dieses aus der Betätigung eines Berechtigten stammt. Dieses Vorgehen führt zu geringeren Matchscores, als wenn das System die Möglichkeit hat, aus einer Sequenz das Bild mit der höchsten Übereinstimmung auszuwählen. Dafür bietet dieses Verfahren die Möglichkeit, eine große Anzahl von Verifikationen Unberechtigter durchzuführen und somit FARs auf Basis einer statistisch aussagekräftigen Datenmenge zu ermitteln.

Neben der Ermittlung von FARs ermöglicht dieses Verfahren auch die Erstellung von Similarity-Matrizen. In einer solchen Matrix werden die erreichten Matchscores von Betätigungen eines Teilnehmers bei Vergleichen gegen die Referenztemplates aller anderen Teilnehmer dargestellt. Die Diagonale enthält somit die Matchscores, die der jeweilige Teilnehmer bei Betätigungen mit Vergleichen seines eigenen Referenztemplates erzielt.

	4699000019	4699000020	4699000031	4699000042	4699000053	4699000064	4699000075	4699000086
4699000019	0,7072	0,0734	0,0381	0,0945	0,1371	0,0981	0,1941	0,2434
4699000020	0,0937	0,7225	0,0927	0,1989	0,0442	0,0622	0,2173	0,2429
4699000031	0,0566	0,0572	0,711	0,0919	0,0317	0,0326	0,0984	0,0301
4699000042	0,3919	0,2697	0,1052	0,7212	0,1881	0,0911	0,5688	0,3001
4699000053	0,0715	0,3784	0,1955	0,4943	0,6186	0,3903	0,0987	0,1227
4699000064	0,0373	0,1624	0,1124	0,0405	0,0574	0,6727	0,1659	0,2671
4699000075	0,3483	0,1982	0,1819	0,0572	0,0556	0,0579	0,7314	0,3537
4699000086	0,1269	0,0643	0,0372	0,1326	0,0379	0,032	0,0829	0,7089

Abbildung 17: Auszug einer Similarity-Matrix

In Abbildung 17 ist exemplarisch ein Auszug einer Similarity-Matrix dargestellt.

6.2.3.2 Algorithmenvergleich

Ein Ziel von BioP I ist die Bestimmung des besten Algorithmus innerhalb eines Systems. Hierfür ergibt sich ein eindeutiges Ergebnis. Für alle untersuchten Referenzbasen liefert der Algorithmus 1 die besten Erkennungsleistungen. Beispielhaft wird dies an den ROC-Diagrammen für die Referenzbasis 4 (Abbildung 18), Referenzbasis 7 (Abbildung 19) und Referenzbasis 8 (Abbildung 20) verdeutlicht.

[11] Zusätzlich wurde getestet, ob Bilder aus dem Live-Enrolment zu höheren Matchscores führen als Live-Bilder aus regulären Betätigungen. Dies kann verneint werden.

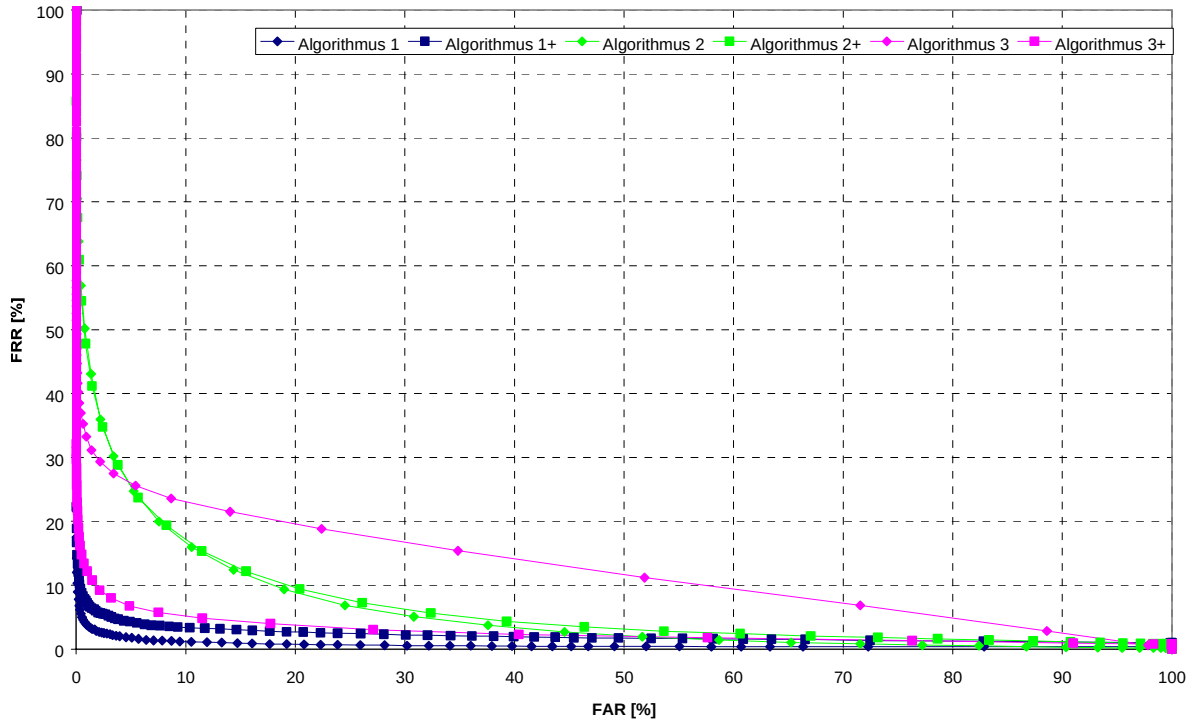

Abbildung 18: ROC-Kurven für Referenzbasis 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 18 stellt in einem Diagramm die ROC-Kurven der Algorithmen für die komprimierte Bilddatei gemäß ICAO dar.

Anhand dieses Diagramms lässt sich sehr gut ein Ranking der Algorithmen ablesen (je näher die Kurve an den Achsen verläuft, desto besser ist die Erkennungsleistung des Algorithmus):

1. Algorithmus 1
2. Algorithmus 1+
3. Algorithmus 3+
4. Algorithmus 2 und Algorithmus 2+ nahezu gleichauf
5. Algorithmus 3

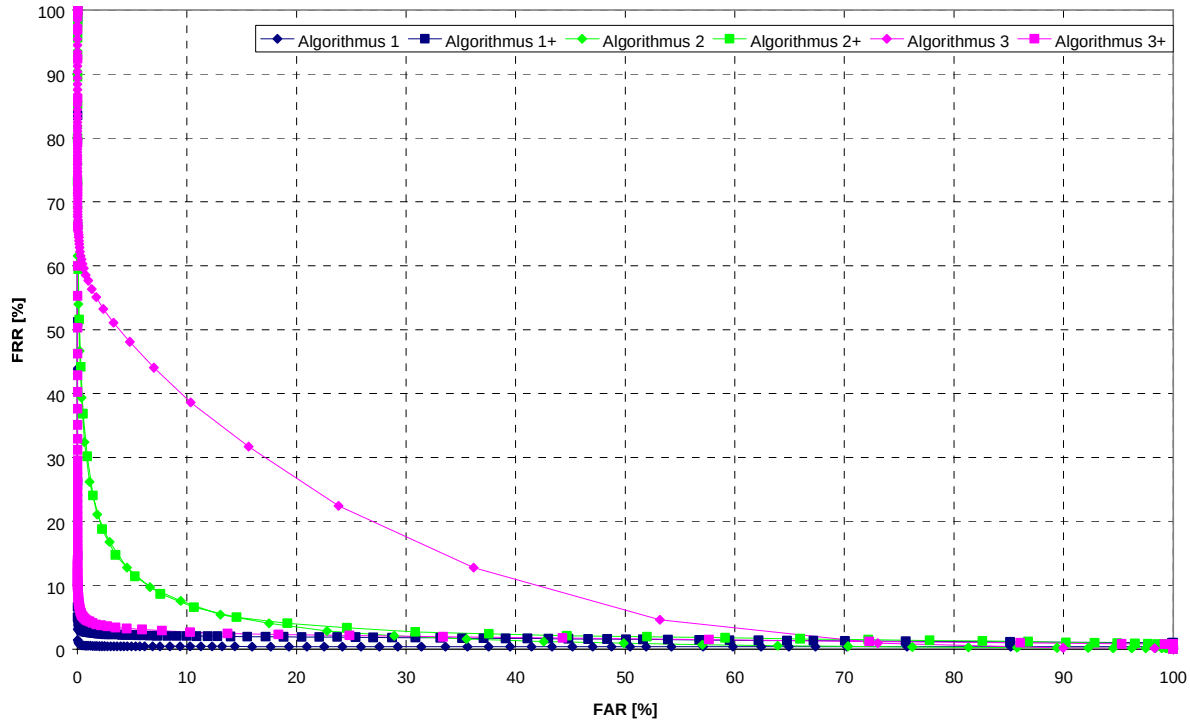

Abbildung 19: ROC-Kurven für Referenzbasis 7 (Systemtemplate aus Live-Enrolment)

Abbildung 19 stellt in einem Diagramm die ROC-Kurven der Algorithmen für das Systemtemplate aus dem Live-Enrolment dar.

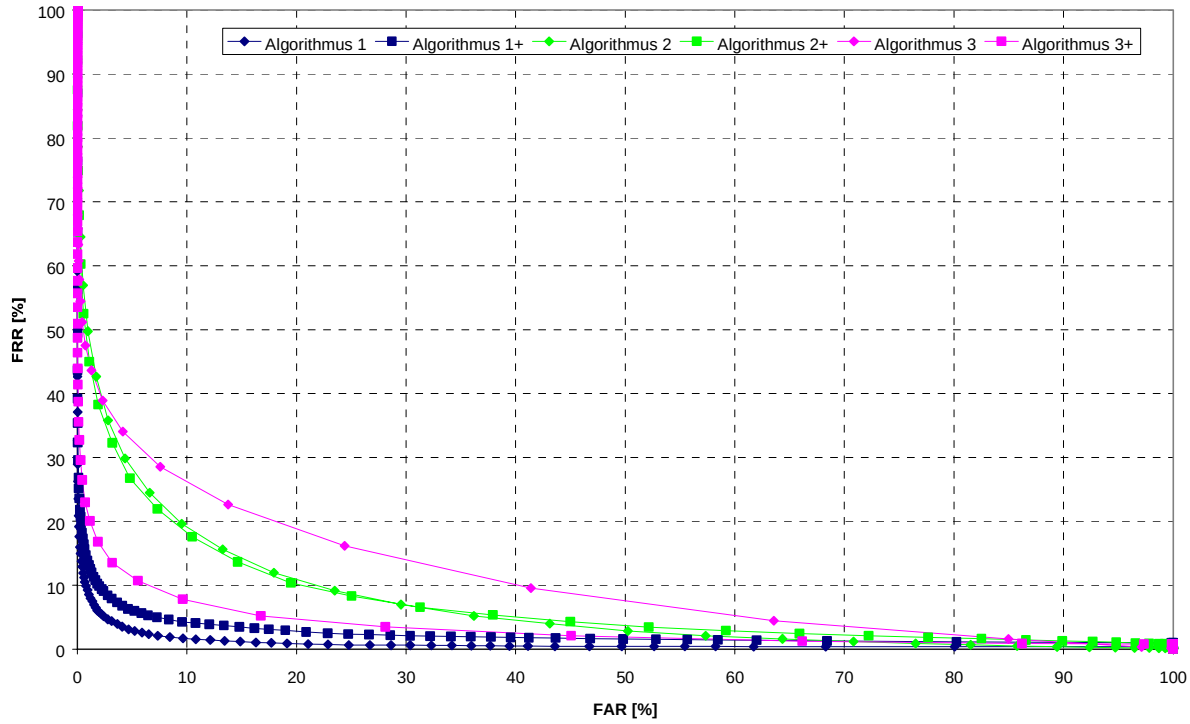

Abbildung 20: ROC-Kurven für Referenzbasis 8 (Lichtbild vom Musterpersonalausweis)

Abbildung 20 stellt in einem Diagramm die ROC-Kurven der Algorithmen für das Lichtbild vom Musterpersonalausweis dar.

Das aufgestellte Ranking bestätigt sich für alle Referenzbasen.

Zur weiteren Veranschaulichung der Ergebnisse der Algorithmen sind im Folgenden einige FAR-FRR-Kurven für die Referenzbasis 4 dargestellt. Aus FAR-FRR-Kurven kann als Gütemerkmal die Trennschärfe zwischen Akzeptanzen Unberechtigter und Rückweisungen Berechtigter abgelesen werden. Im Idealfall schneiden sich beide Kurven auf der Abszisse. Aus dem Bereich, in dem beide Kurven gemeinsam auf der Abszisse verlaufen, kann dann ein Threshold für den Algorithmus ausgewählt werden. Real ist dies jedoch nicht zu erreichen. Damit ein geeigneter Threshold gewählt werden kann, sollte es einen Bereich geben, in dem beide Kurven sehr dicht an der Abszisse verlaufen.

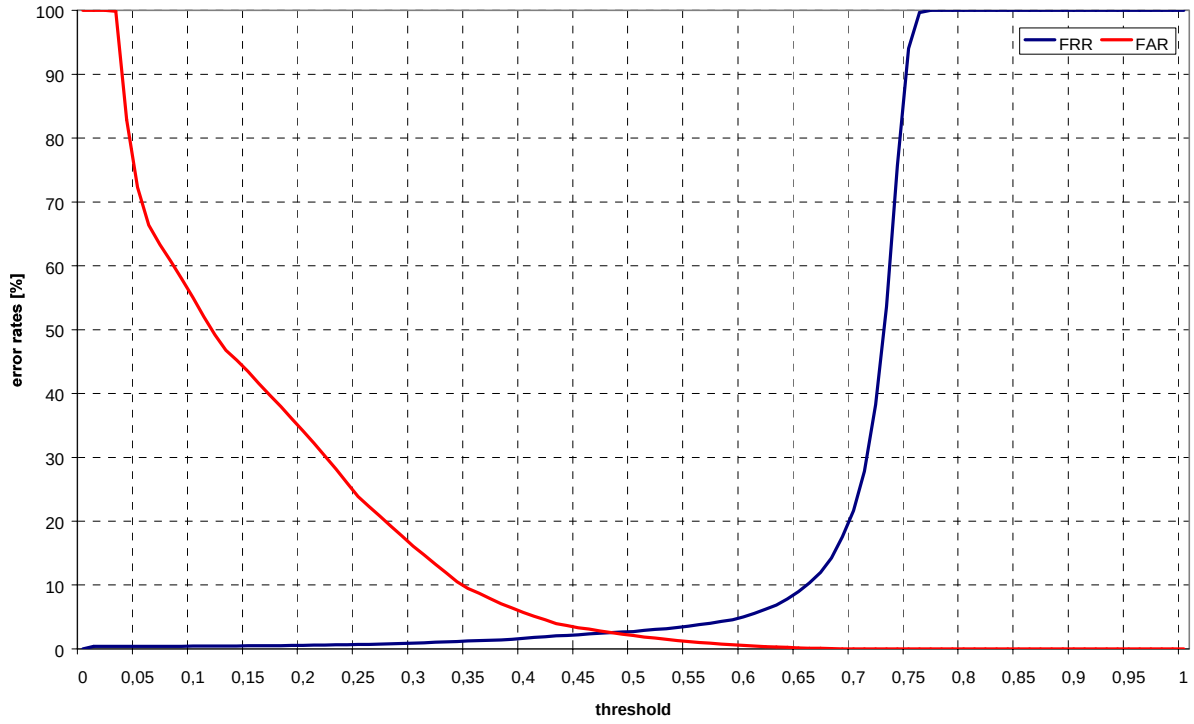

Abbildung 21: FAR-FRR-Curve für Algorithmus 1 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 21 stellt in einem Diagramm die FAR und FRR von Algorithmus 1 für RefID 4 in Abhängigkeit des Schwellwertes gegenüber.

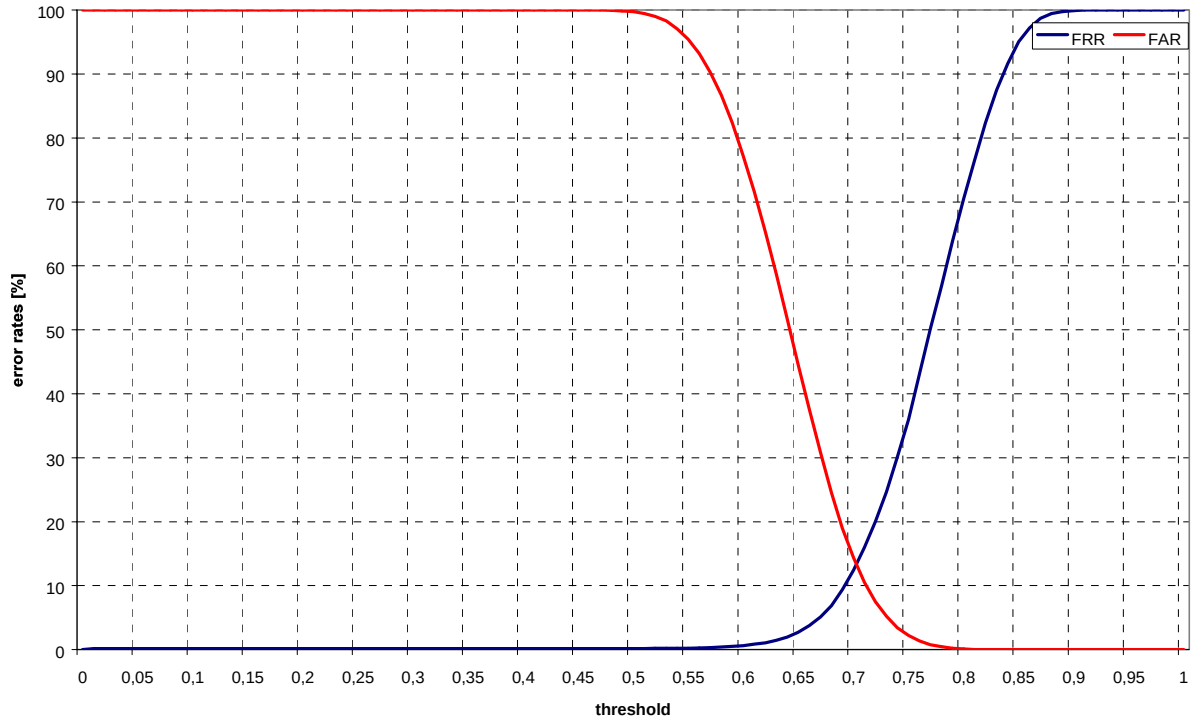

Abbildung 22: FAR-FRR-Curve für Algorithmus 2 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 22 stellt in einem Diagramm die FAR und FRR von Algorithmus 2 für RefID 4 in Abhängigkeit des Schwellwertes gegenüber.

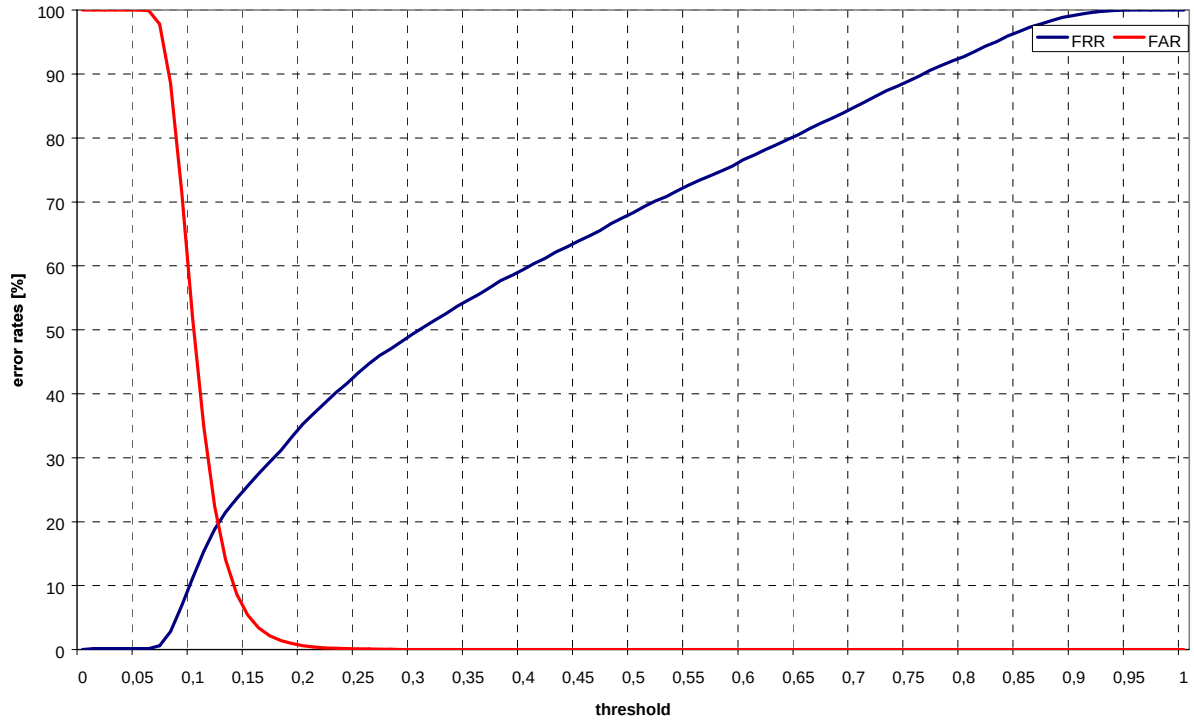

Abbildung 23: FAR-FRR-Curve für Algorithmus 3 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 23 stellt in einem Diagramm die FAR und FRR von Algorithmus 1 für RefID 4 in Abhängigkeit des Schwellwertes gegenüber.

Auch anhand dieser Übersichten wird deutlich, dass bei den Algorithmen 2 und 3 kein Bereich existiert, in dem die Fehlerraten in einem akzeptablen Verhältnis zueinander stehen.

Dies wird noch deutlicher bei der Betrachtung der relativen Häufigkeiten von Matchscores für Betätigungen von Berechtigten und Unberechtigten. Die Darstellung erfolgt (wie in Kapitel 6.1.2 erläutert) mit Hilfe von Genuine-Impostor-Frequency-Diagrammen. Das wesentliche Gütemerkmal, welches aus Genuine-Impostor-Frequency-Diagrammen abgeleitet werden kann, ist die Trennschärfe zwischen dem Auftreten von Matchscores Berechtigter und Unberechtigter. Im Idealfall zeigen beide Kurven keine Überlappung. Aus dem Bereich zwischen den beiden Kurven kann dann ein Threshold für den Algorithmus ausgewählt werden. Real ist eine Überlappung jedoch nicht auszuschließen. Damit ein geeigneter Threshold gewählt werden kann, sollte diese Überlappung jedoch möglichst gering sein.

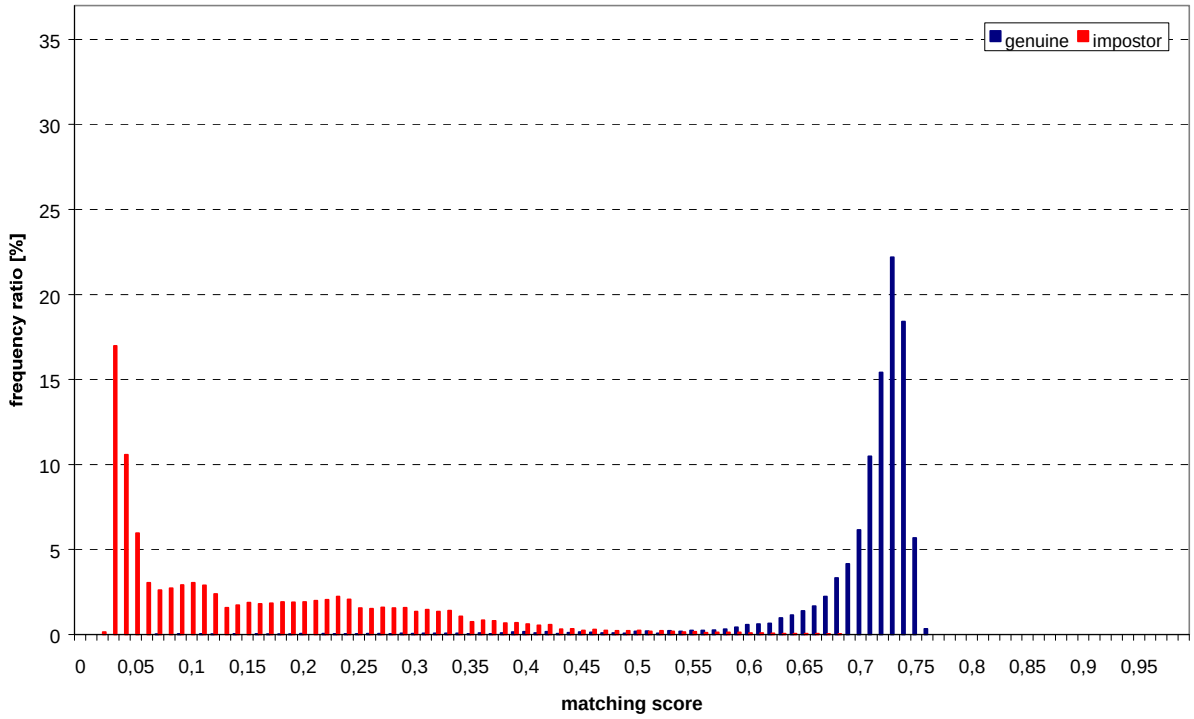

Abbildung 24: Matchscore-Verteilung für Algorithmus 1 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 24 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten für Algorithmus 1 und RefID 4 dar.

Abbildung 24 zeigt die besten im Rahmen von BioP I erreichten Verteilungen für RefID 4. Diese wurden beim Algorithmus 1 erzielt. Während sich die Matchscores Unberechtigter in einem Bereich deutlich unter 0,7 bewegen, ordnet sich eine zwar nicht zu vernachlässigende, aber noch relativ kleine Menge von Matchscores Berechtigter ebenfalls in diesem Bereich an.

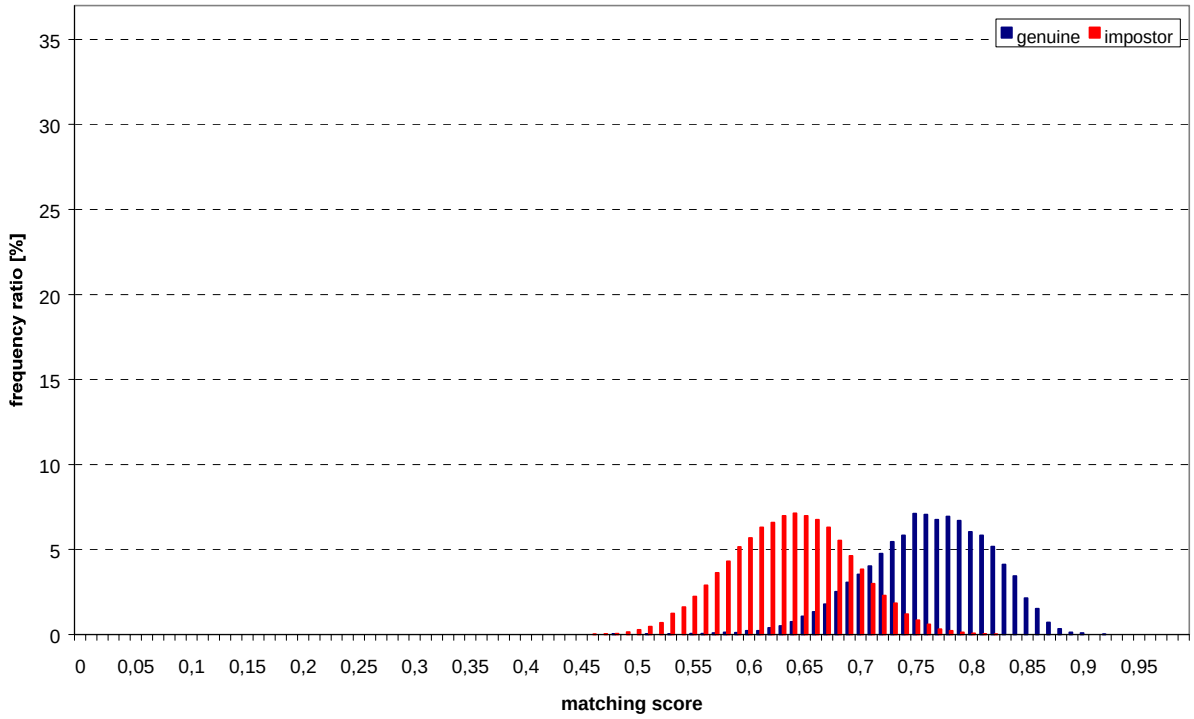

Abbildung 25: Matchscore-Verteilung für Algorithmus 2 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 25 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten für Algorithmus 2 und RefID 4 dar.

Die in Abbildung 25 dargestellte Verteilung der mit dem Algorithmus 2 erreichten Matchscores zeigt eine sehr starke Überlappung. In einem solchen Fall ist keine sinnvolle Konfiguration für einen Einsatz möglich.

Noch extremer als bei Algorithmus 2 stellt sich die Verteilung beim Algorithmus 3 dar. Hier liegt sogar eine Häufung der Matchscores Berechtigter im gleichen Bereich wie die Matchscores Unberechtigter (siehe Abbildung 26).

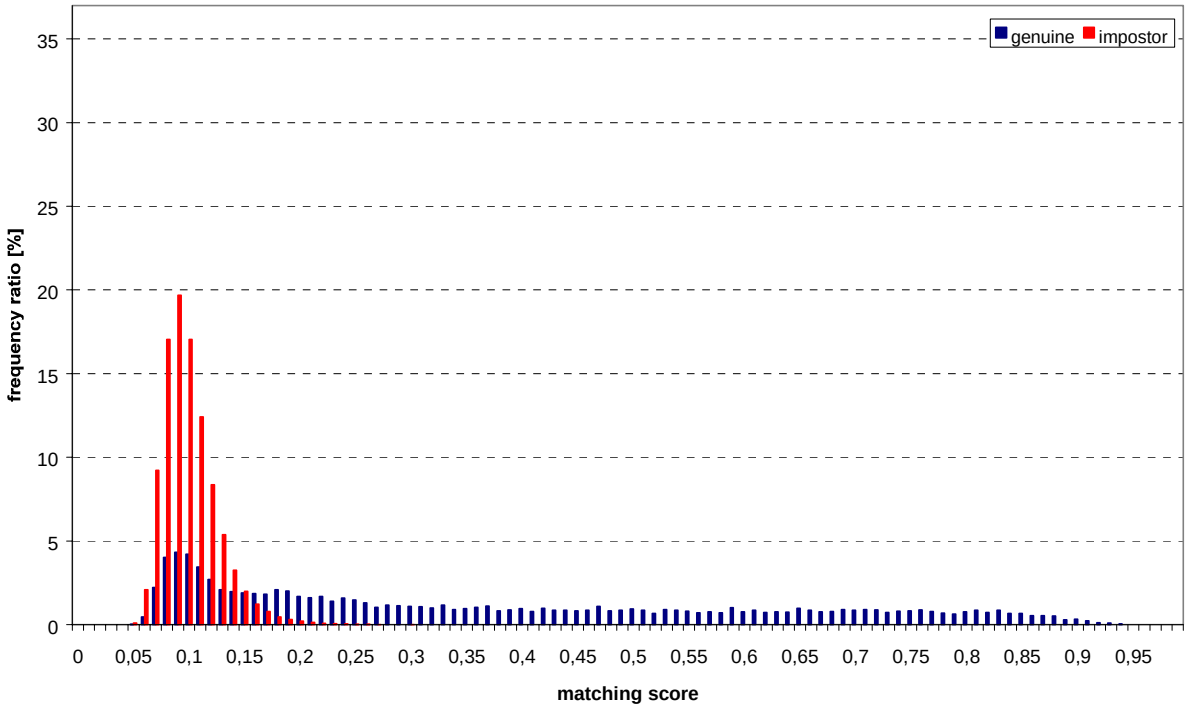

Abbildung 26: Matchscore-Verteilung für Algorithmus 3 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 26 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten für Algorithmus 3 und RefID 4 dar.

6.2.3.3 Referenzbasenvergleich

Anhand des Algorithmus 1, der im Algorithmenvergleich für alle Referenzbasen die besten Erkennungsleistungen erzielt hat, kann ein Vergleich der Referenzbasen untereinander erfolgen. Dies lässt sich wiederum am geeignetesten auf Basis der ROC-Kurven darstellen.

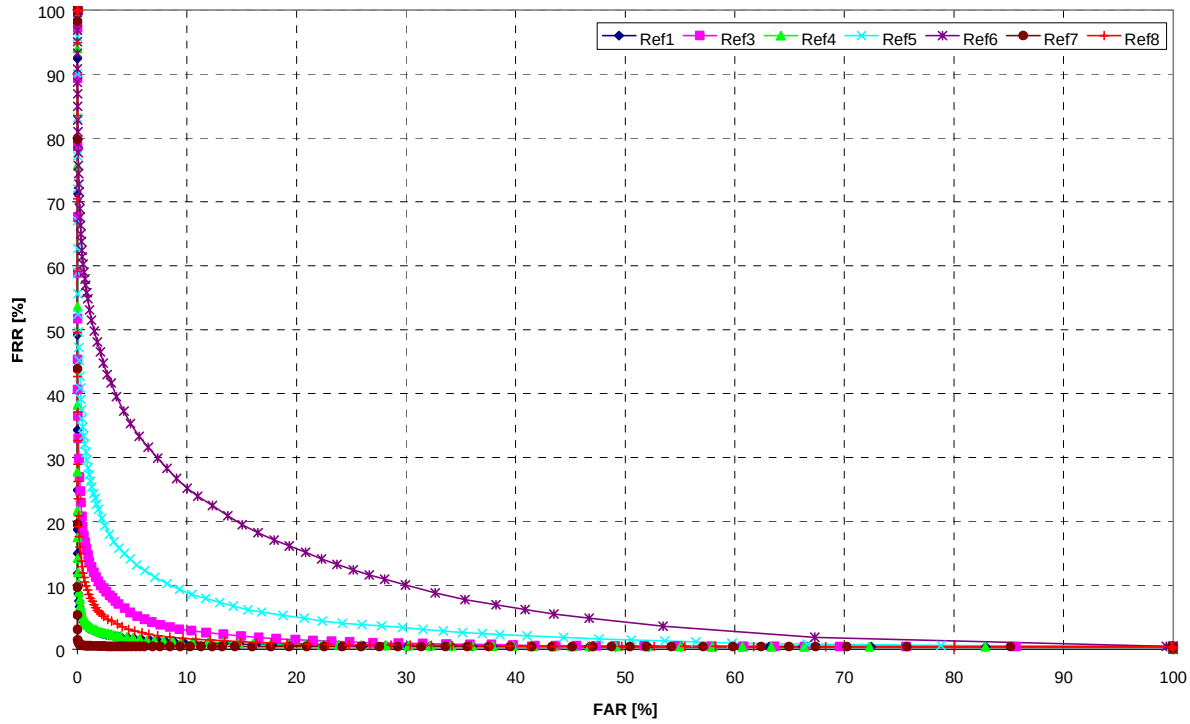

Abbildung 27: ROC-Kurven für Algorithmus 1

Abbildung 27 stellt in einem Diagramm die ROC-Kurven von Algorithmus 1 für alle Referenzbasen dar.

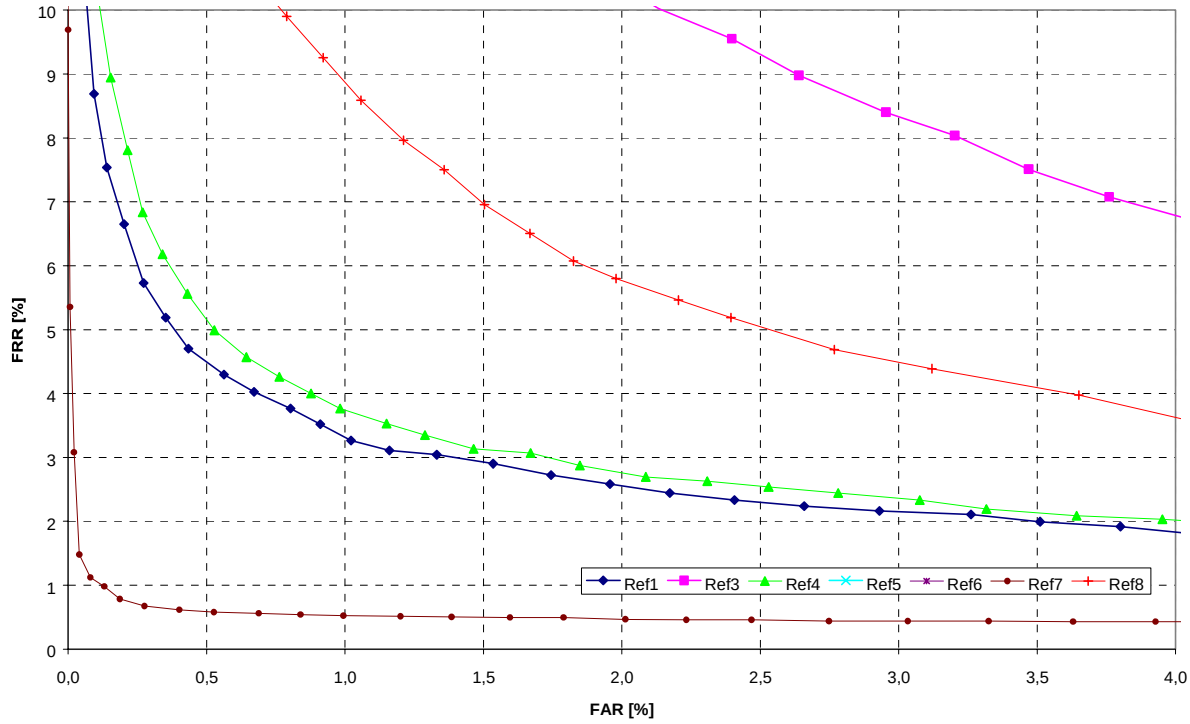

Abbildung 28: Detailausschnitt der ROC-Kurven für Algorithmus 1

Abbildung 28 stellt in einem Diagramm einen Detailausschnitt der ROC-Kurven von Algorithmus 1 für alle Referenzbasen dar.

Aus den Diagrammen in Abbildung 27 und Abbildung 28 ist eine klare Abstufung zwischen den einzelnen Referenzbasen ersichtlich. Die beste Erkennungsleistung bietet das systemspezifische Template des Live-Enrolments mit deutlichem Abstand vor den Bilddateien auf Basis des Frontalfotos. Die schwach komprimierte Bilddatei liefert dabei leicht bessere Ergebnisse als die stärker komprimierte. Mit wiederum erheblichem Abstand folgen die Ergebnisse der Lichtbilder des Musterpersonalausweises und des EU-Visums. Sehr schlechte Erkennungsleistungen treten bei der Bilddatei auf Basis der Halbprofilaufnahme sowie beim aktuellen Bundespersonalausweis auf.

Zusammenfassend stellt sich das Ranking wie folgt dar:

1. RefID 7: Systemtemplate aus Live-Enrolment
2. RefID 1: Bilddatei Frontalaufnahme unkomprimiert
3. RefID 4: Bilddatei Frontalaufnahme komprimiert
4. RefID 8: Lichtbild Musterpersonalausweis
5. RefID 3: Lichtbild EU-Visum
6. RefID 5: Bilddatei Halbprofilaufnahme
7. RefID 6: Lichtbild Bundespersonalausweis

Dieses Ergebnis wird durch die im Folgenden dargestellten Diagramme beispielhaft verdeutlicht.

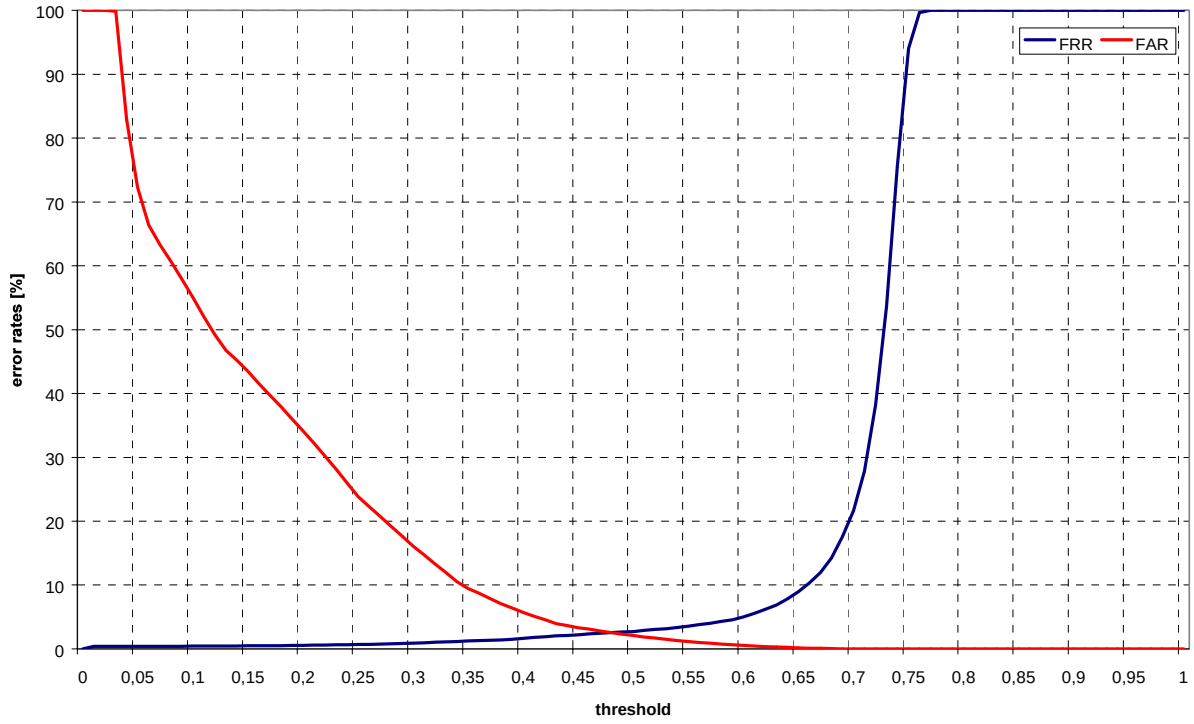

Abbildung 29: FAR-FRR-Curve für Algorithmus 1 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 29 stellt in einem Diagramm die FAR und FRR von Algorithmus 1 für die komprimierte Bilddatei gemäß ICAO in Abhängigkeit des Schwellwertes gegenüber.

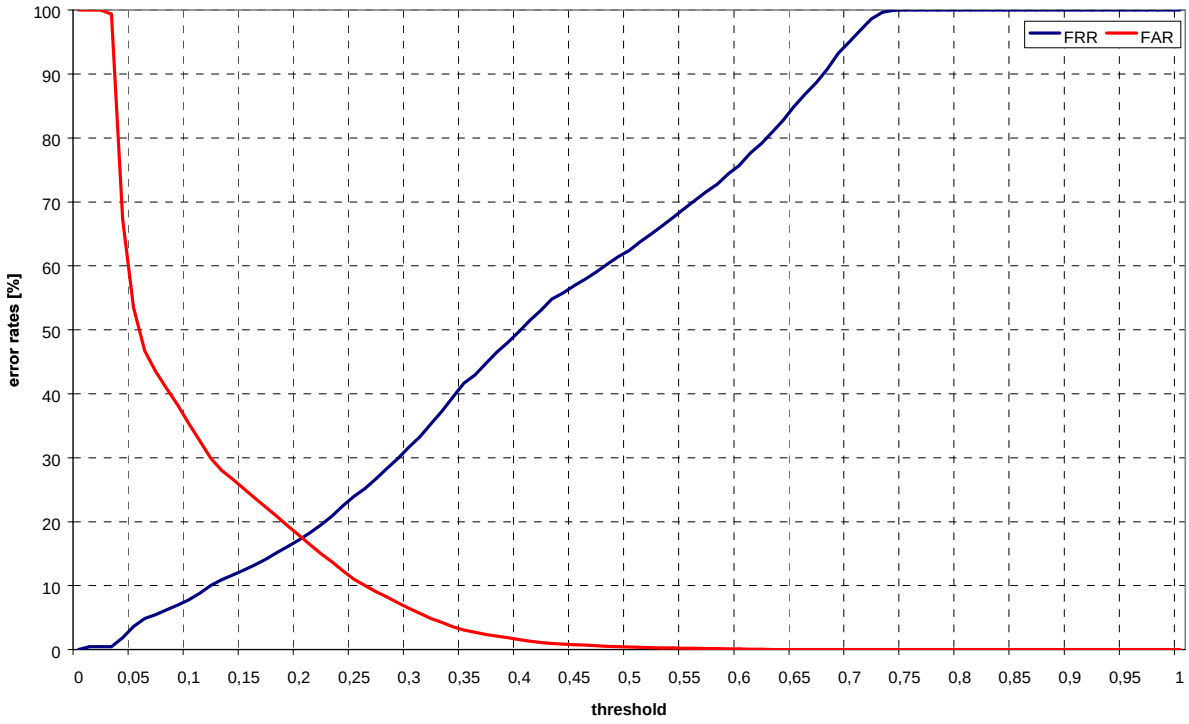

Abbildung 30: FAR-FRR-Curve für Algorithmus 1 und RefID 6 (aktueller Bundespersonalausweis)

Abbildung 30 stellt in einem Diagramm die FAR und FRR von Algorithmus 1 für den aktuellen Bundespersonalausweis in Abhängigkeit des Schwellwertes gegenüber.

Während Abbildung 29 einen typischen Kurvenverlauf für biometrische Systeme zeigt, ist ein System mit dem in Abbildung 30 dargestellten Verhalten völlig ungeeignet.

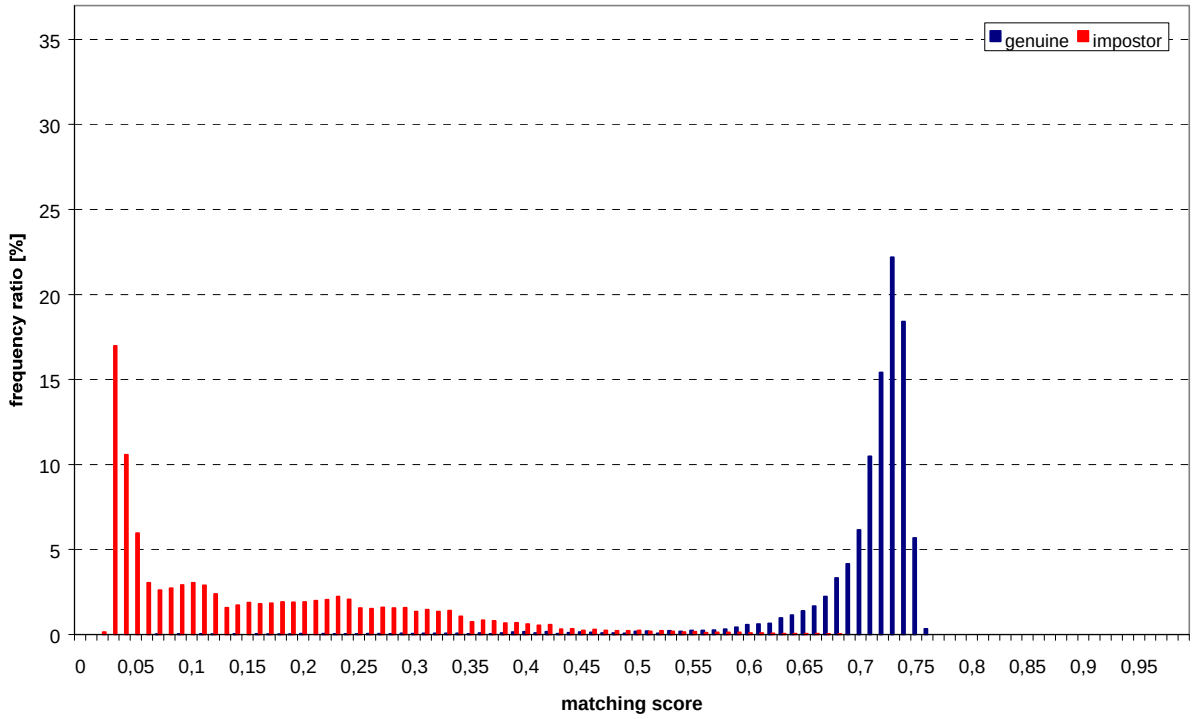

Abbildung 31: Matchscore-Verteilung für Algorithmus 1 und RefID 4 (komprimierte Bilddatei gemäß ICAO)

Abbildung 31 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten für Algorithmus 1 und die komprimierte Bilddatei gemäß ICAO dar.

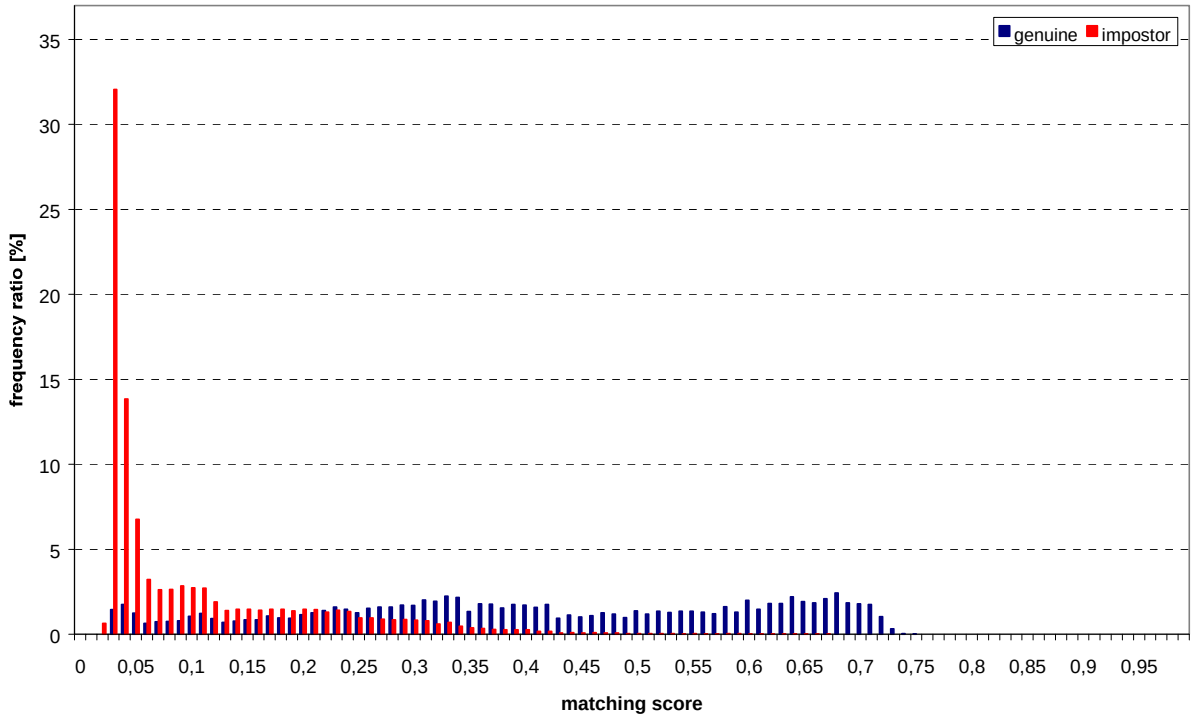

Abbildung 32: Matchscore-Verteilung für Algorithmus 1 und RefID 6 (aktueller Bundespersonalausweis)

Abbildung 32 stellt in einem Diagramm die Häufigkeitsverteilung der Matchscores von Berechtigten und Unberechtigten für Algorithmus 1 und den aktuellen Bundespersonalausweis dar.

Während das in Abbildung 31 dargestellte Ergebnis für die Referenzbasis 4 (Bilddatei gemäß ICAO) eine akzeptable Verteilung mit nicht allzu starker Überlappung zeigt, wird aus Abbildung 32 deutlich, dass Referenzbasis 6 (aktueller Bundespersonalausweis) absolut untauglich für den Einsatz von Gesichtserkennungsverfahren ist. Zwar liegen alle Matchscores Unberechtigter in einem niedrigen Bereich, jedoch verteilen sich die Matchscores der Berechtigten nahezu gleichmäßig über einen sehr breiten Bereich, der auch die niedrigen Matchscores umfasst.

6.2.3.4 Systemvergleich

Um die Erkennungsleistung der beiden Komplettsysteme vergleichen zu können, werden diese für eine einheitliche Konfiguration betrachtet. Dies bedeutet, dass bei beiden Systemen die Verifikationsergebnisse des integrierten identischen Algorithmus bei einer einheitlichen Toleranzschwelle von 0,7 betrachtet werden.

RefID	FRR [%]		FAR [%]	
	System A	System B	System A	System B
1	21,23	48,24	0,0070	0,0035
2	51,74	74,43	0	0
3	60,16	80,84	0,0053	0,0018
4	24,08	51,86	0,0053	0,0018
5	77,89	88,62	0	0
6	95,25	98,51	0	0
7	4,61	0,65	0,0407	0,1442
8	51,15	74,22	0	0

Tabelle 6: Fehlerraten bei Toleranzschwelle 0,7

Aus diesen Ergebnissen geht hervor, dass das System A für Bilddateien bessere Erkennungsleistungen liefert. Für das beim Live-Enrolment gewonnene Systemtemplate liefert jedoch das System B die deutlich besseren Erkennungsleistungen.

6.2.3.5 Erkennungsleistung im zeitlichen Verlauf

Bei der Betrachtung der Entwicklung der FRRs über den Feldtestzeitraum lassen sich die im Folgenden dargestellten Punkte beobachten. Dabei entwickeln sich die jeweiligen FRRs eines Systems für das Technical_Attempt_Set und das Scenario_Attempt_Set tendenziell ähnlich, jedoch auf verschiedenen Niveaus.

- **System A**: Abgesehen von den ersten zwei Tagen ist in den ersten drei Feldtestwochen im Wesentlichen ein Rückgang der FRR festzustellen. Nach einem kurzen Anstieg zu Beginn der vierten Woche, geht die Fehlerrate wieder zurück.
- **System B**: Nach einem deutlichen Rückgang in der ersten Woche des Feldtests stabilisiert sich die FRR auf einem relativ konstant verlaufenden niedrigen Niveau.

Bei beiden Systemen ist also nach einer Gewöhnungsphase der Benutzer, in der die Fehlerraten deutlich zurückgegangen sind, eine Stabilisierung zu beobachten. Bemerkenswert dabei ist, dass diese Gewöhnungsphase beim System A deutlich länger dauert als beim System B.

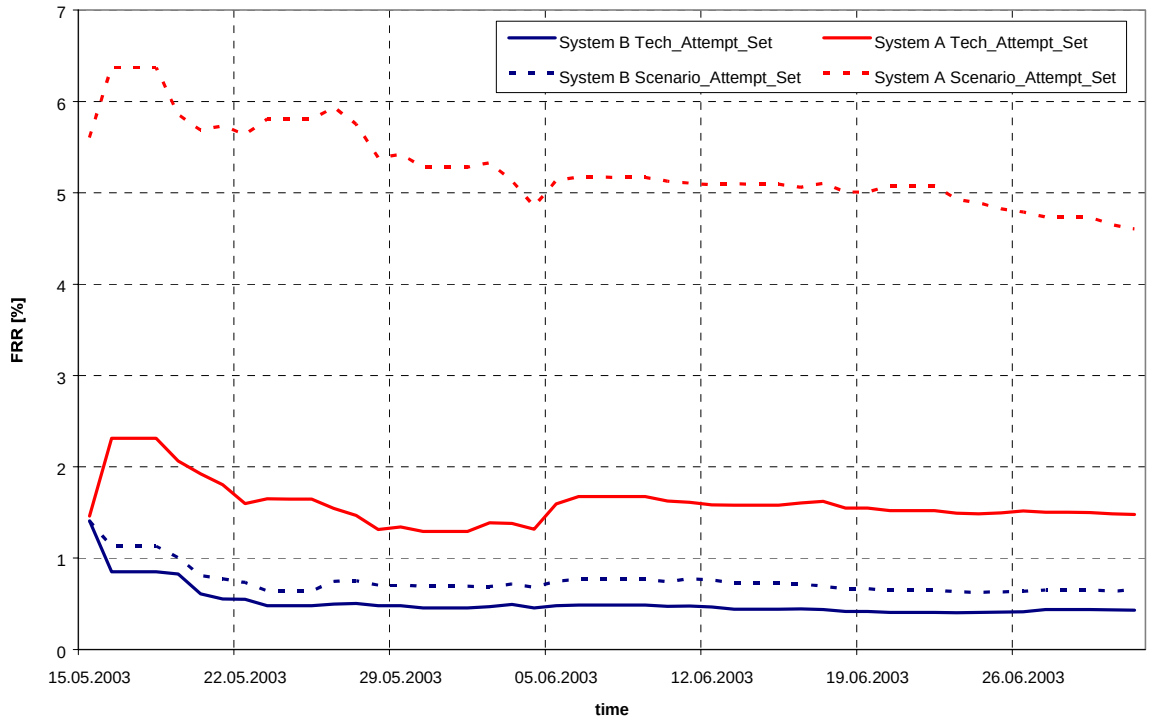

Abbildung 33: Entwicklung der FRRs im zeitlichen Verlauf (Threshold 0,7)

Abbildung 33 stellt in einem Diagramm die Entwicklung der FRR über den Feldtestzeitraum für beide Systeme sowohl anhand des Tech_Attempt_Set als auch anhand des Scenario_Attempt_set dar.

6.2.4 Einzelbenutzerstatistik

6.2.4.1 Tech_Attempt_Set

Die Einzelbenutzerstatistik des Tech_Attempt_Set stellt die FRR (bei FAR = 0,1%) spezifisch für jeden Testteilnehmer aus der Population User50 dar. Idealerweise sollten die FRRs der verschiedenen Teilnehmer alle möglichst niedrig sein. Hohe FRRs einzelner Personen deuten darauf hin, dass das Merkmal dieser Personen vom Algorithmus bzgl. der betrachteten Referenzbasis nicht hinreichend gut verwertet werden kann.

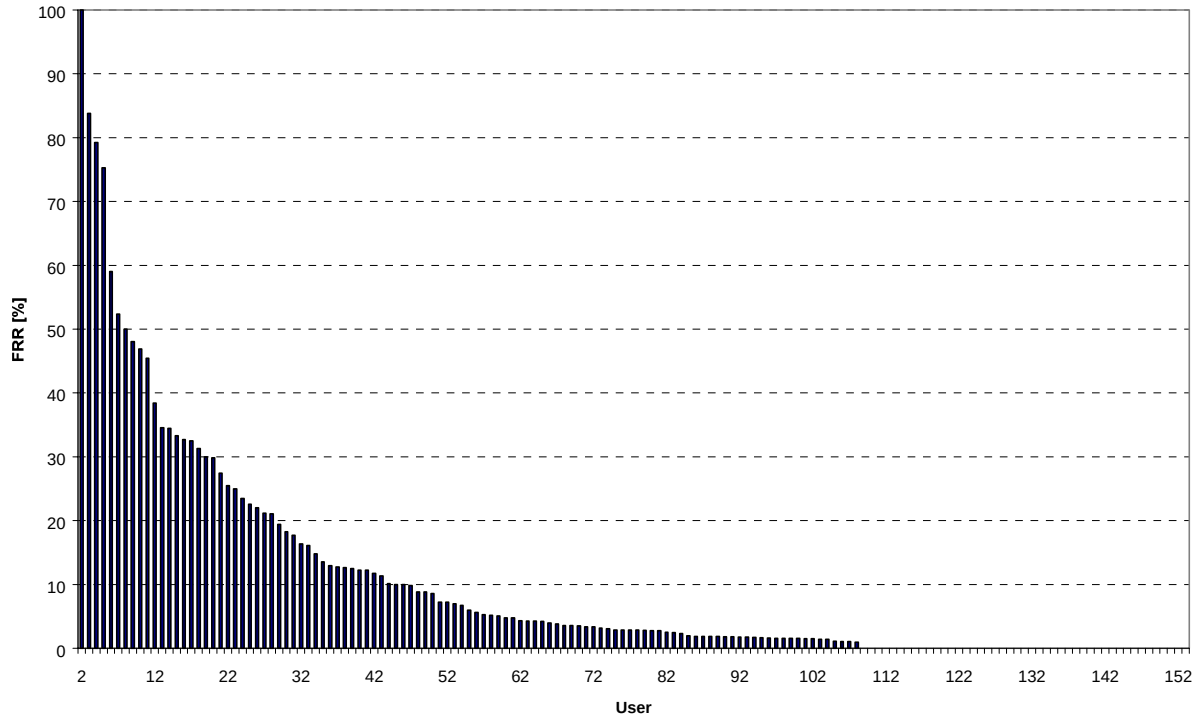

Abbildung 34: Einzelbenutzerstatistik für Algorithmus 1 und RefID 4 (komprimierte Bilddatei gemäß ICAO) (FAR=0,1%)

Abbildung 34 stellt in einem Diagramm die FRR der einzelnen Benutzer für Algorithmus 1 und die komprimierte Bilddatei gemäß ICAO dar.

Aus Abbildung 34 kann abgelesen werden, dass bei RefID 4 (komprimierte Bilddatei gemäß ICAO) ca. ein Zehntel der Testpopulation bei mehr als einem Drittel ihrer Betätigungen zurückgewiesen werden.

Für Referenzbasis 7 (Template aus Live-Enrolment) sieht die Einzelbenutzerstatistik bei gleicher FAR deutlich anders aus. Während lediglich zwei Testteilnehmer eine FRR von über 10% aufweisen, ist für den überwiegenden Teil der Population User50 die FRR gleich Null.

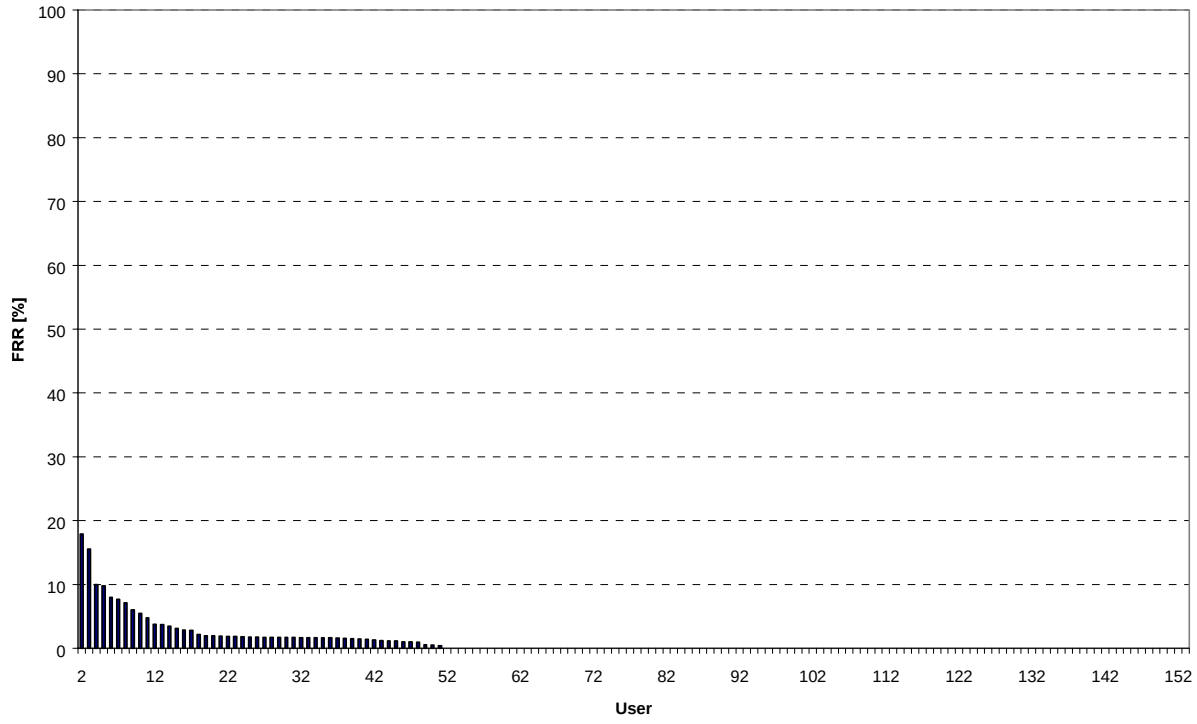

Abbildung 35: Einzelbenutzerstatistik für Algorithmus 1 und RefID 7 (Template aus Live-Enrolment) (FAR=0,1%)

Abbildung 35 stellt in einem Diagramm die FRR der einzelnen Benutzer für Algorithmus 1 und das Systemtemplate aus dem Live-Enrolment dar.

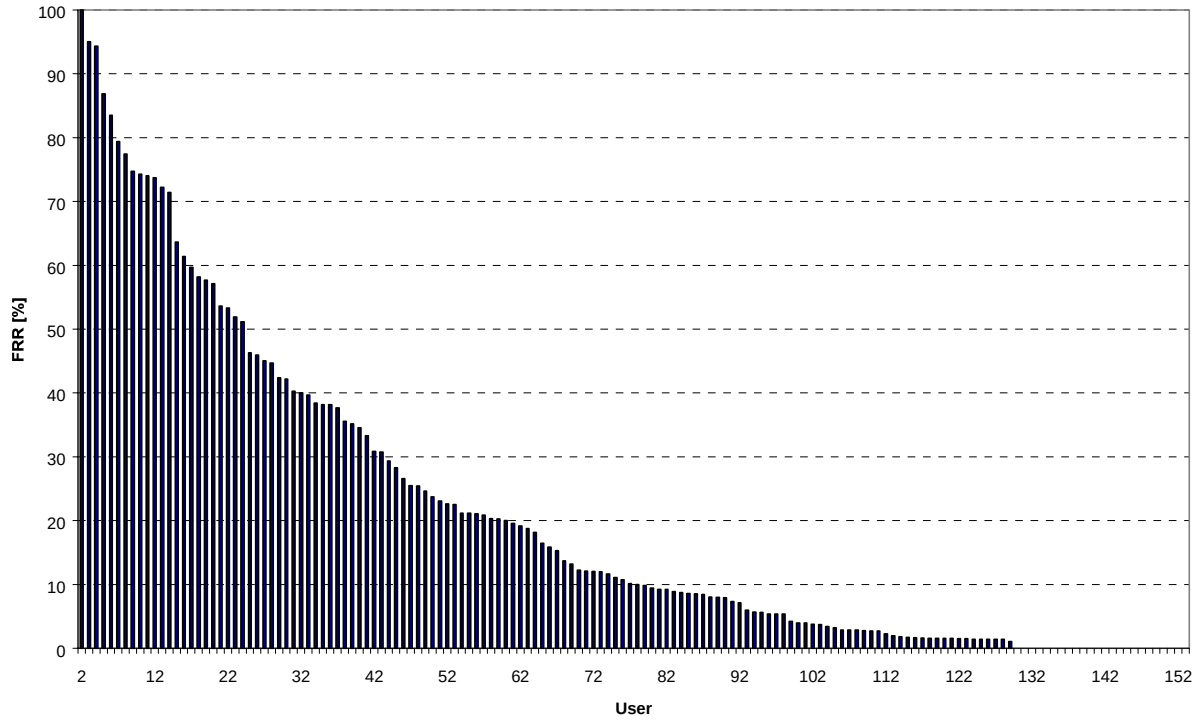

Abbildung 36: Einzelbenutzerstatistik für Algorithmus 1 und RefID 8 (Bild vom Musterpersonalausweis) (FAR=0,1%)

Abbildung 36 stellt in einem Diagramm die FRR der einzelnen Benutzer für Algorithmus 1 und das Bild vom Musterpersonalausweis dar.

Für Referenzbasis 8 (Bild vom Musterpersonalausweis) ergibt sich keine befriedigende Verteilung der Einzelbenutzer-FRRs. Für etwa ein Drittel der Testpopulation User50 liegen die FRRs zum Teil sehr weit über der Gesamt-FRR dieser Referenzbasis.

Für die anderen Algorithmen wurden signifikant schlechtere Einzelbenutzerstatistiken ermittelt.

Ein weiteres interessantes Ergebnis ist die Standardabweichung der Einzelbenutzer-FRRs. Eine hohe Standardabweichung bedeutet, dass für eine relativ große Anzahl der Testpersonen schlechtere Erkennungsleistungen als im Mittel erzielt werden. Hier schneidet Referenzbasis 7 (Template aus Live-Enrolment) deutlich am besten ab, gefolgt von Referenzbasis 1 (Bilddatei Frontalaufnahme) und Referenzbasis 4 (komprimierte Bilddatei gemäß ICAO). Die höchste Streuung tritt bei den Referenzbasen 5 (Bilddatei Halbprofilaufnahme) und 6 (aktueller Bundespersonalausweis) auf.

6.2.4.2 Scenario_Attempt_Set

Die Einzelbenutzerstatistik des Scenario_Attempt_Set stellt die FRR (bei Threshold 0,7) spezifisch für jeden Testteilnehmer aus der Population User50 dar. Idealerweise sollten die FRRs der verschiedenen Teilnehmer alle möglichst niedrig sein. Hohe FRRs einzelner Personen deuten darauf hin, dass entweder die Erfassung des Merkmals für diese Personen ein Problem bereitet oder das Merkmal dieser Personen vom Algorithmus bzgl. der betrachteten Referenzbasis nicht hinreichend gut verwertet werden kann.

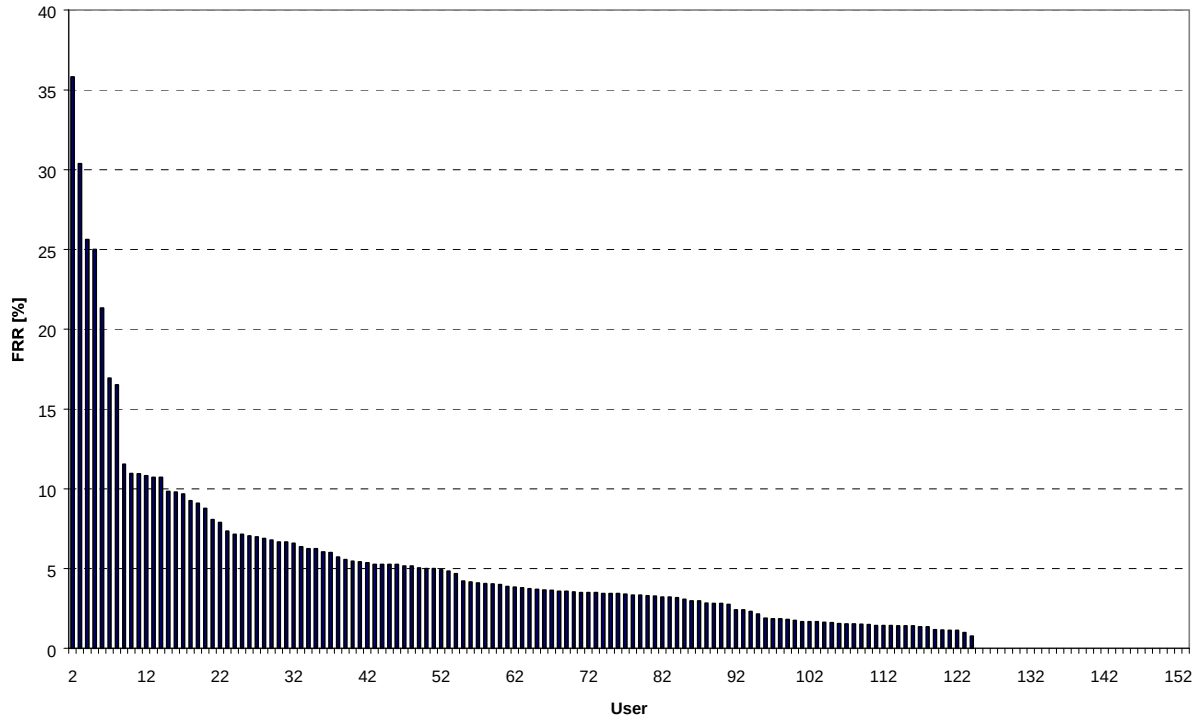

Abbildung 37: Einzelbenutzerstatistik für System A und RefID 7 (Threshold 0,7)

Abbildung 37 stellt in einem Diagramm die FRR der einzelnen Benutzer beim Threshold 0,7 für System A und das Template aus dem Live-Enrolment dar.

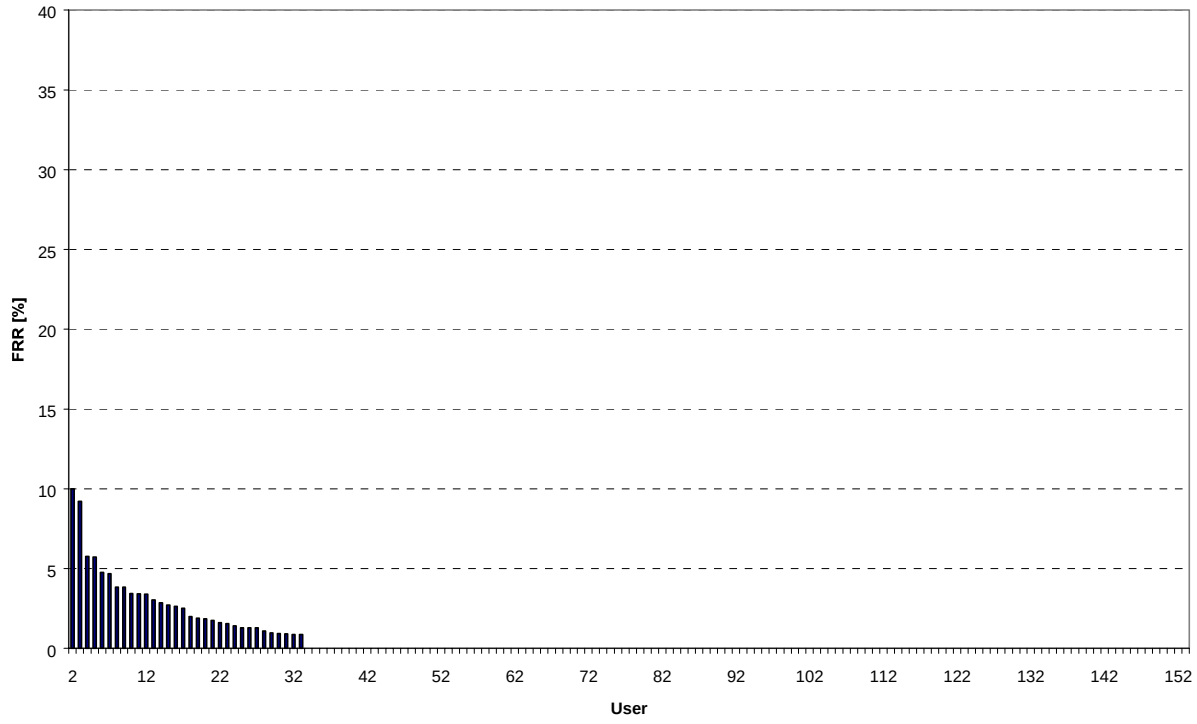

Abbildung 38: Einzelbenutzerstatistik für System B und RefID 7 (Threshold 0,7)

Abbildung 38 stellt in einem Diagramm die FRR der einzelnen Benutzer beim Threshold 0,7 für System B und das Template aus dem Live-Enrolment dar.

Aus dem Vergleich von Abbildung 37 und Abbildung 38 wird deutlich, dass für einen Großteil der betrachteten Testpopulation das System A deutlich schlechtere Erkennungsleistungen bietet. Untermauert wird dies durch den Vergleich der Standardabweichungen der Einzelbenutzer-FRRs (System A: 5,45%; System B: 1,56%).

Für dieses deutliche Ergebnis kann es zwei Ursachen geben:

- Das System B liefert für Referenzbasen, die aus Mehr-Bild-Enrolments resultieren wie sie beim Live-Enrolment durchgeführt werden, deutlich bessere Erkennungsleistungen als das System A.
- Die Erfassungseinheit des System A liefert für eine Reihe von Testpersonen schlechte Aufnahmen des Merkmals.

Auf eine Gegenüberstellung der Einzelbenutzerstatistiken der beiden Systeme für die anderen Referenzbasen wird an dieser Stelle verzichtet. Aufgrund der Ermittlung der Fehlerraten anhand des für Referenzbasis 7 eingestellten Thresholds ist die Betrachtung der anderen Referenzbasen nur bedingt aussagekräftig. Trotzdem bestätigt sich anhand dieser Ergebnisse die in 6.2.3.4 gewonnene Erkenntnis, dass das System A für Referenzbasen, die aus Ein-Bild-Enrolments resultieren (Bilddateien, Lichtbilder), Vorteile gegenüber dem System B besitzt.

6.2.5 Betrachtung der Gesichtsfindung

Der Prozess der Gesichtserkennung lässt sich grundsätzlich in die Gesichtsfindung und die eigentliche Gesichtserkennung unterscheiden. Die Gesichtsfindung (face detection) stellt dabei die Erkennung des

Gesichts als solches innerhalb der Umgebung dar. Das Gesicht wird entsprechend ausgeschnitten und dem Gesichtserkennungsalgorithmus als so genanntes kanonisches Bild übergeben. Entsprechend kann die Gesichtserkennung als Gesamtprozess nur so gut sein, wie die jeweilige Kombination aus Gesichtfindungs- und Gesichtserkennungsalgorithmus.

Die in BioP I beteiligten Komplettsysteme lassen sich nicht unter diesem Aspekt vergleichen. Während System B den klassischen Ansatz verwendet, der eine entsprechende Analyse erlaubt, geht System A aufgrund eines integrierten intelligenten Kamerasystems einen anderen Weg: Unabhängig von den integrierten Algorithmen nimmt bereits eine Vorverarbeitungseinheit (PreProcessor) eine Gesichtsfindung vor. Den integrierten Algorithmen wird dann bereits eine auf das Gesicht zugeschnittene Szene übergeben. Hierauf wird dann bei den integrierten Algorithmen der klassische Gesichtsfinder angewendet und auf dieses Ergebnis schließlich der eigentliche Gesichtserkennungs-algorithmus.

Aufgrund der zum Teil nicht Pflichtenheft-gemäßen Umsetzung durch die Hersteller lassen sich keine genauen Angaben zur Fehlerrate bezüglich Gesichtsfindung angeben. Das Verhältnis zwischen den beiden Komplettsystemen ist jedoch sehr deutlich. Während bei System B sehr selten Fehler in der Gesichtsfindung auftraten, versagte die Vorverarbeitungseinheit von System A vergleichsweise häufig. In mehr als fünf Prozent der Aufnahmen wurden komplette Bildausschnitte (nicht auf das Gesicht fokussiert) bereitgestellt.

6.2.6 Allgemeine Ergebnisse zu Systemen und Herstellern

Gemäß Pflichtenheft waren die beteiligten Hersteller aufgefordert, ausgereifte, für einen unbetreuten Feldtest geeignete Systeme termingerecht bereitzustellen. Im Hinblick auf das Zielszenario für die Gesichtserkennungssysteme besitzen diese nicht-biometriespezifischen Anforderungen eine nicht zu vernachlässigende Bedeutung.

Die beteiligten Systeme und deren Hersteller wurden bezüglich entsprechender Kriterien untersucht.

Dazu zählen:

- **Inbetriebnahme**: Im Rahmen der Inbetriebnahme sind besonders die termingerechte Bereitstellung der geforderten Funktionen sowie deren korrekte Umsetzung relevant.
- **Systemfehler**: Hier wird dargestellt, welche wesentlichen Fehler während des Feldtests aufgetreten sind.
- **Ausfallverhalten**: Bei unüberwachtem Betrieb von Systemen ist deren Zuverlässigkeit von hoher Bedeutung. Entsprechend wurden über den Feldtest die Systemausfälle protokolliert. Dies erfolgte über Informationen der Administratoren seitens des BKA, über das spezielle Tool EventSentry sowie die Teilnehmernotizen vorbehaltlich einer Prüfung auf Basis von EventSentry und Ergebniseinträgen.
- **Administrationsaufwand**: Der Einsatz von IT-Systemen erfordert immer einen gewissen Administrationsaufwand. Dieser sollte möglichst gering sein. Des Weiteren sollten geeignete Werkzeuge zur Unterstützung des Administrators bereitstehen, wie beispielsweise Reporting-Mechanismen. Nicht zuletzt ist eine intuitive Bedienung relevant.
- **Benutzerbedingte Probleme**: Bei biometrischen Systemen kommt es vor, dass beispiels-weise aufgrund von Fehlern im Systemdesign einzelne Personen wegen individueller Eigenschaften (zum Beispiel Körpergröße) schlechter erkannt werden als andere. Dieser Aspekt ist für die Bewertung benutzerbedingter Probleme relevant.
- **Service und Support durch die Hersteller**: Unter diesem Kriterium wird dargestellt, ob der Hersteller für eine reibungslose Testdurchführung ausreichenden Service und Support

geleistet hat. Insbesondere interessiert, ob aufgetretene Fehler jeweils kurzfristig gelöst werden konnten.

Eine Darstellung dieser Ergebnisse ist nicht Bestandteil dieses öffentlichen Abschlussberichts.

6.3 Statistische Aussagekraft der Ergebnisse und Fehlerbetrachtung

6.3.1 Bewertung der statistischen Aussagekraft der Ergebnisse

Die in [TechEval] angegebene Formel zur Messung der Signifikanz wird hier nicht angewendet. Statistische Signifikanz bezeichnet vereinfacht ausgedrückt die Wahrscheinlichkeit, dass der beobachtete empirische Befund eintritt, wenn eine diesbezügliche, vorab gegebene Hypothese richtig ist. Dies basiert auf der Annahme einer Zufallsstichprobe, in der jedes Element der Grundgesamtheit die gleiche Chance hat, in die Stichprobe zu gelangen. Dies war in der im Projekt BioP I entstandenen und untersuchten Gesamtheit nicht der Fall (siehe auch Kapitel 6.3.2.1.1).

Auch unter Berücksichtigung der in Abschnitt 6.3.2 dargestellten Fehlerquellen haben die in BioP I ermittelten Ergebnisse eine starke Aussagekraft. Diese Aussage stützt sich auf die Tatsache, dass ein ausgesprochen umfangreicher und gehaltvoller empirischer Datenbestand zur Verfügung stand. So wurden 241 Teilnehmer in die empirische Erhebung aufgenommen. Die ermittelten Ergebnisse basieren auf einer Population von 152 Teilnehmern, welche jeweils 50 oder mehr Betätigungen durchführten. Insgesamt wurden so von dieser schon relativ großen Testpopulation über 10.000 Betätigungen für jedes der beteiligten Systeme (und somit für jeden der integrierten Algorithmen und jede der zu untersuchenden Referenzbasen) realisiert. Der erhebliche Umfang der entstandenen Datenlage erlaubt einen weitreichenden und gewichtigen Blick auf die Leistungsfähigkeit und Beurteilung von Gesichtserkennungssystemen.

6.3.2 Fehlerbetrachtung

Bei der Erhebung der Ergebnisse ist zu berücksichtigen, dass bei der Durchführung technischer Untersuchungen Fehlerquellen nie komplett auszuschließen sind. Die auftretenden Fehler können gemäß ihrer Ursache klassifiziert werden:

- **Systematische Fehler**: Fehler, die aus grundsätzlichen Rahmenbedingungen der Testdurchführung resultieren
- **Implementationsfehler**: Fehler, die durch nicht korrekte Implementation bzw. Integration von (Teil-)Komponenten entstehen
- **Auswertungsfehler**: Fehler, die durch die verwendeten Auswertungsmethoden bedingt sind

6.3.2.1 Systematische Fehler

6.3.2.1.1 Auswahl der Testpopulation

Die im Feldtest untersuchten Fragestellungen beschränken sich zumeist nicht auf den in der Untersuchung betrachteten Personenkreis. Gewonnene Erkenntnisse sollen beispielsweise auf alle Personen in Deutschland verallgemeinert werden. Für dieses Problem der so genannten Aussagekraft oder Repräsentativität stehen in der empirischen Forschung zweierlei Lösungsmöglichkeiten zur Verfügung.

Die erste Lösungsmöglichkeit ist die so genannte bewusste Auswahl, in der bekannte Parameter der Grundgesamtheit (also etwa alle deutschen Staatsbürger) wie beispielsweise der Anteil von Männern und Frauen oder der Altersverteilung in der Teilerhebung nachgebildet werden. Dem liegt die

Vorstellung zugrunde, dass die so entstandene Stichprobe die zu beschreibende Grundgesamtheit ersetzt. Eine derartige Konstruktion der Stichprobe konnte im Rahmen des Projekts BioP I jedoch nicht durchgeführt werden, da sich die Teilnehmer ausschließlich aus dem Personalbestand des BKA rekrutierten.

Die zweite Möglichkeit zur Erstellung einer Stichprobe bildet die so genannte zufällige Auswahl. Hierbei hat jedes Element der Grundgesamtheit (also beispielsweise jeder deutsche Staatsbürger) die gleiche Chance, in die Stichprobe zu gelangen. Ist eine zufällige Auswahl zumindest annähernd gegeben, so ist es möglich, die bestehenden Ergebnisse über die Anwendung so genannter induktivstatistischer Verfahren mit einem quantifizierbaren Grad an Gewissheit, der statistischen Signifikanz oder Irrtumswahrscheinlichkeit, auf die jeweilige Grundgesamtheit zu übertragen. Da in BioP I die Stichprobe durch die Auswahl der BKA-Mitarbeiter vorgegeben wurde, konnte jedoch keine Zufallsstichprobe realisiert werden.

6.3.2.1.2 Auswahl Bildmaterial für Verifikation

Der parallele Vergleich verschiedener Referenzbasen und verschiedener Algorithmen erfordert, dass alle im Hintergrund ablaufenden Verifikationsprozesse mit demselben Bildmaterial arbeiten müssen. Dieses ist jedoch nur für die Kombination von Masterreferenz und Masteralgorithmus optimal ausgewählt. Eine Verschlechterung der erzielten Erkennungsleistungen für die Nicht-Masterreferenzbasen und Nicht-Masteralgorithmen ist somit nicht auszuschließen. Trotz dieses systematisch auftretenden Fehlers ist eine Vergleichbarkeit für die im Hintergrund ablaufenden Verifikationen gegeben.

6.3.2.1.3 Einflüsse auf Performance

Zur Erfassung der zur Auswertung erforderlichen Ergebnisdaten, mussten die GE-Systeme Mechanismen bereitstellen, welche entsprechende Datensätze in eine zentrale Datenbank protokollieren. Insbesondere bei der Übertragung von Bilddatensätzen ist somit eine Beeinflussung der Prozesszeit nicht zu vermeiden. Da jedoch für alle Systeme diesbezüglich gleiche Voraussetzungen galten, ist die Vergleichbarkeit der gewonnenen Ergebnisse durch diese Fehlerquelle nicht beeinträchtigt.

6.3.2.2 Implementierungsfehler

Bei der Integration von Komponenten der GE-Systeme existieren potentielle Fehlerquellen, die einen nicht unerheblichen Einfluss auf die Erkennungsleistung des Gesamtsystems haben. Beispielhaft seien an dieser Stelle folgende wesentliche Punkte aufgeführt:

- Nicht optimale Integration und Konfiguration eines Algorithmus
- Einsatz von Kamerasystemen, die kein optimales Bildmaterial liefern

6.3.2.3 Auswertungsfehler

Neben den durch die numerischen Berechnungen bedingten Fehlern (zum Beispiel Rundungsfehler, beschränkte Genauigkeiten) sind weitere wesentliche Fehlerquellen zu berücksichtigen.

- **Verwendetes Datenmaterial**: In die Auswertung gehen Betätigungen aus der gesamten Feldtestphase ein. Trotz einer vorangehenden TeachIn-Phase sind in den ersten Tagen des Feldtests vergleichsweise hohe Fehlerraten zu beobachten. Bezöge man diese Tage nicht in die Auswertung mit ein, würden bessere Gesamtergebnisse resultieren.
- **FAR-Ermittlung**: In BioP I wurde die FAR auf Basis von Einzelbild-Verifikationen ermittelt. In realen Einsatzszenarien werden jedoch Verifikationen Unberechtigter auf Basis von Bild-

sequenzen durchgeführt. Da somit höhere Matchscores erzielt werden können, ist es erforderlich, höhere Thresholds zu verwenden. Dies hat wiederum zur Folge, dass sich die FRR verschlechtert.

- **Statistische Abhängigkeiten bei FRR-Ermittlung**: Zur Ermittlung der FRR wurden die Betätigungen einer festen Testpopulation herangezogen. Da jeder der Testteilnehmer aus dieser Population eine Vielzahl von Betätigungen durchgeführt hat, die alle in die Bewertung einfließen, basiert die FRR-Ermittlung nicht auf komplett unabhängigen statistischen Ereignissen.
- **Statistische Abhängigkeiten bei FAR-Ermittlung**: Zur Ermittlung der FAR wurden Verifikationen von Live-Bildern aller Testteilnehmer gegen die Referenztemplates jeweils aller anderen Testteilnehmer durchgeführt. Da somit wechselseitige Vergleiche stattfanden, basiert die FAR-Ermittlung nicht auf komplett unabhängigen statistischen Ereignissen.

7 Auswertung der weiterführenden Untersuchungen

7.1 Technische Untersuchungen

7.1.1 Verifikationen Unberechtigter

Zur Ermittlung der Robustheit bei Versuchen Unberechtigter wurden Verifikationen durchgeführt, bei denen Vergleiche von Live-Aufnahmen von Testpersonen mit den Referenztemplates anderer Testpersonen stattfanden. Dabei handelt es sich gemäß [BestPrac] um Zero-effort attempts. Die Durchführung dieses Tests mit Bilddateien ist in Kapitel 6.2.3.1 beschrieben.

Um zu testen, wie stark bei Verifikationen Unberechtigter die Übergabe einer Bildsequenz anstelle eines einzelnen Bildes die erzielten Matchscores beeinflusst, wurde ein entsprechender Test mit Live-Betätigungen durchgeführt.

Hierfür wurde aus der Feldtestpopulation eine Teilgruppe von 21 Personen ausgewählt, deren Struktur bzgl. der Merkmale Geschlecht, Bartträger und Brillenträger mit der Struktur der gesamten Feldtestgruppe annähernd übereinstimmt.

In diesem Test verwendete jede Person der Testgruppe jeweils das Verifikationsmittel (Musterpersonalausweis) jedes anderen Teilnehmers. Dieser Test wurde in der gleichen Umgebung und unter identischen Bedingungen wie der reguläre Feldtest durchgeführt. Im Gegensatz zum Feldtest waren diese Betätigungen jedoch überwacht. Die Ergebnisse wurden in der zentralen Datenbank protokolliert.

Die Auswertung der Ergebnisse bestätigt die Vermutung, dass bei Verifikationen auf Basis von Bildsequenzen im Mittel deutlich bessere Matchscores von Unberechtigten erreicht werden. Dies muss bei der Auswahl der Toleranzschwellen für reale Einsatzumgebungen berücksichtigt werden.

7.1.2 Variation der Referenzdaten

In diesem Test wurde untersucht, inwieweit starke Kompressionen und eine reduzierte Auflösung der Referenzdaten die Erkennungsleistung beeinträchtigen. Basis war in allen Fällen das Foto der Testpersonen in Frontalaufnahme. Dieses ging in den in Tabelle 7 dargestellten Varianten als Referenzbasis in den erweiterten Test ein.

Hintergrund dieser Untersuchung ist, dass die Bilder – sofern sie als Datei auf dem Ausweisdokument abgelegt werden – einen möglichst geringen Speicherplatzbedarf haben sollen (deshalb Test verschiedener Kompressionen und Auflösungen).

RefID	Beschreibung	Format	Typische Dateigröße	Test Kompression	Test Auflösung
AltRef1	Identisch mit RefID 1 aus Feldtest (Photoshop-Qualität 10)	JPEG	75 kB	X	X
AltRef2	Identisch mit RefID 4 aus Feldtest (Photoshop-Qualität 2)	JPEG	14 kB	X	
AltRef3	Zweitstärkste Photoshop-Kompression (Qualität 1)	JPEG	12 kB	X	
AltRef4	Stärkste Photoshop-Kompression (Qualität 0)	JPEG	11 kB	X	
AltRef5	Reduzierte Auflösung (150 dpi) mit Photoshop-Qualität 10	JPEG	32 kB		X

Tabelle 7: Übergebene Bilddateien für zusätzliche Referenzbasen[12]

Dieser Test erfolgte ohne interaktive Betätigungen von Testpersonen. Es wurde von allen Teilnehmern des Feldtests, die mindestens eine Betätigung durchgeführt haben (238 Personen), je ein Live-Bild eines Systems ausgewählt. Diese wurden in einem Batch-Lauf in dem jeweiligen System gegen die zugehörigen – zusätzlich enrolten – alternativen Referenztemplates des jeweiligen Testteilnehmers verglichen. Die verschiedenen Testfälle basieren somit auf je 238 unabhängigen Verifikationen. Die Ergebnisse der einzelnen Vergleiche wurden in der zentralen Ergebnisdatenbank protokolliert.

Beim Enrolment der Bilddateien sind folgende Punkte zu bemerken:

- Die Referenzbasen AltRef1 - AltRef4 (verschiedene Kompressionsstufen) konnten problemlos von allen Algorithmen enrolt werden.
- Die Referenzbasis AltRef5 (niedrige Auflösung) konnte von den um einen alternativen Gesichtsfinder ergänzten Versionen der Algorithmen (Plus-Versionen) nicht enrolt werden.

Anhand der erfassten Verifikationsergebnisse (Matchscores) wurden algorithmenspezifisch die FRRs berechnet (siehe Tabelle 8). Als Threshold wurde jeweils der Wert von dem Arbeitspunkt gewählt, an dem die FAR für AltRef1 0,1% ist. Innerhalb eines Algorithmus wurde also für alle alternativen Referenzbasen derselbe Threshold zur Berechnung der FRR verwendet.

[12] Die Qualitätsstufen von Photoshop gehen von 0 bis 10, wobei 0 die schlechteste Qualität (stärkste Kompression) und 12 die beste Qualität (schwächste Kompression) repräsentiert.

MeID	AltRef1	AltRef2	AltRef3	AltRef4	AltRef5
Algorithmus 1	4,64	5,91	7,59	8,86	6,78
Algorithmus 1+	8,40	9,66	11,34	10,50	-
Algorithmus 2	65,97	65,97	64,71	64,71	66,24
Algorithmus 2+	61,76	63,03	60,92	62,18	-
Algorithmus 3	26,05	30,25	30,25	33,19	32,07
Algorithmus 3+	14,29	18,49	18,07	24,37	-

Tabelle 8: FRR in [%] für alternative Referenzbasen (FAR=0,1%)

Erwartungsgemäß ist bei steigender Kompressionsstärke eine Verschlechterung der Erkennungsleistung zu beobachten. Gleiches gilt bei einer Verringerung der Auflösung.

7.1.3 Variation der Umweltbedingungen

Da biometrische Systeme charakteristische Merkmale von Personen aus der Umgebung aufnehmen, ergibt sich ein signifikanter Einfluss durch die gegebenen Umweltbedingungen. Für Gesichtserkennung sind hier im Wesentlichen die vorherrschenden Lichtbedingungen relevant. Eine Beeinflussung ergibt sich insbesondere aus Beleuchtungsintensität und der Art des Lichteinfalls, wobei letztgenanntes den wesentlicheren Faktor darstellt.

Da diese Tests einen relativ hohen Aufwand erzeugen und der Fokus hier eher auf qualitativen als auf quantitativen Ergebnissen liegt, wurde die Testgruppe sehr klein gehalten. Die Testgruppe umfasste 13 Personen, die sich aus Mitarbeitern der secunet am Standort Essen rekrutierten. Die Auswahl der Testpersonen erfolgte in der Art, dass für die Merkmale Geschlecht, Bart, Brille, Frisur, Makeup und ethnische Herkunft verschiedene Ausprägungen vorhanden sind.

Diese Tests wurden in einem Labor der secunet am Standort Essen durchgeführt. Durch die vorhandene Deckenbeleuchtung und das Verdecken der Fenster mit einem weitgehend lichtundurchlässigen Vorhang wurde eine gleichmäßige Ausleuchtung gewährleistet.

Die Erhöhung der Lichtintensität sowie die Erzeugung verschiedener Arten des Lichteinfalls wurden wie folgt realisiert:

- **Lichteinfall frontal auf die Testperson**: Beleuchtung durch einen handelsüblichen Halogenstrahler (500 Watt) aus ca. 1,5 m Abstand zwecks Erzeugung von Blendstellen
- **Lichteinfall seitlich auf die Testperson**: Beleuchtung durch einen handelsüblichen Halogenstrahler (500 Watt) aus ca. 1 m Abstand zwecks Erzeugung von Schlagschatten
- **Lichteinfall aus dem Hintergrund auf der Testperson**: Tageslichteinfall durch Öffnung der Fenstervorhänge zwecks Erzeugung von Gegenlicht

Lichteinfall	System A	System B
Normalbedingungen	120	123
Frontal	310	290
Seitlich	1300	1300
Hintergrund	310	260

Tabelle 9: Beleuchtungsstärke [Lux] im Labor der secunet[13]

Jede Testperson hat für jedes der beiden Systeme über mehrere Tage verteilt Betätigungen für alle zu untersuchenden Bedingungen durchgeführt. Insgesamt wurden dabei folgende Betätigungszahlen erreicht:

- Normalbedingungen: 650
- Frontal: 325
- Seitlich: 325
- Hintergrund: 325

Aufgrund optimaler Laborbedingungen und einer Betreuung der Versuchsteilnehmer wurden an beiden Systemen sehr hohe Matchscores unter den normalen Lichtbedingungen erzielt.

Anhand der erfassten Verifikationsergebnisse (Matchscores) wurden für die Arbeitspunkte, an denen die FAR jeweils 0,1% beträgt, algorithmenspezifisch die FRRs berechnet.

Zusammenfassend lassen sich die folgenden Punkte feststellen:

- **Frontaler Lichteinfall** auf die Person bewirkt bei System A eine signifikante Verschlechterung der FRR. Lediglich für RefID 7 (Template aus Live-Enrolment) fällt die Verschlechterung moderat aus.
- **Frontaler Lichteinfall** auf die Person bewirkt für das System B eine Verbesserung der FRR für die Templates auf Basis von Bilddateien (RefID 1, 3, 4, 5, 8). Für das Template auf Basis des Live-Enrolments (RefID 7) verschlechtert sich die FRR sehr stark.
- **Seitlicher Lichteinfall** auf die Person bewirkt für beide Systeme extreme Verschlechterungen der FRRs.
- **Lichteinfall von hinten** auf die Person bewirkt für System A eine signifikante Verschlechterung der FRR.
- **Lichteinfall von hinten** auf die Person hat für das System B keine nennenswerten Auswirkungen auf die FRR.

7.1.4 Einfluss des Ausweisalters

In dieser Untersuchung wurde geprüft, wie weit die Erkennungsleistung gegenüber dem Lichtbildausweis von dessen Alter abhängt. Dies erfolgte anhand der aktuellen Bundespersonal-ausweise der Testteilnehmer – also für RefID 6 des Feldtests.

[13] Da die Tests mit Lichteinfall aus dem Hintergrund der Person mit Tageslicht durchgeführt wurden, variiert die Beleuchtungsstärke je nach Witterung. Die angegebenen Werte wurden bei leichter Bewölkung erfasst.

Aus der gesamten Feldtestpopulation stellte eine Teilgruppe von 228 Personen ihren aktuellen Bundespersonalausweis für diese Untersuchung freiwillig zur Verfügung[14]. Dabei wurde das Alter des Ausweises erfasst. In die Untersuchung gingen jedoch nur die Ausweise der Population User50 ein – insgesamt 144 Stück.

Ausweisalter [Jahre]	Gesamtpopulation		Population User50	
	Anzahl absolut	Anteil relativ [%]	Anzahl absolut	Anteil relativ [%]
a ≤2	64	28,32	39	27,08
2 < a ≤4	76	33,63	44	30,56
4 < a ≤6	47	20,80	34	23,61
6 < a ≤8	25	11,06	16	11,11
8 < a ≤10	14	6,19	11	7,64
Gesamt	**226**	**100,00**	**144**	**100,00**

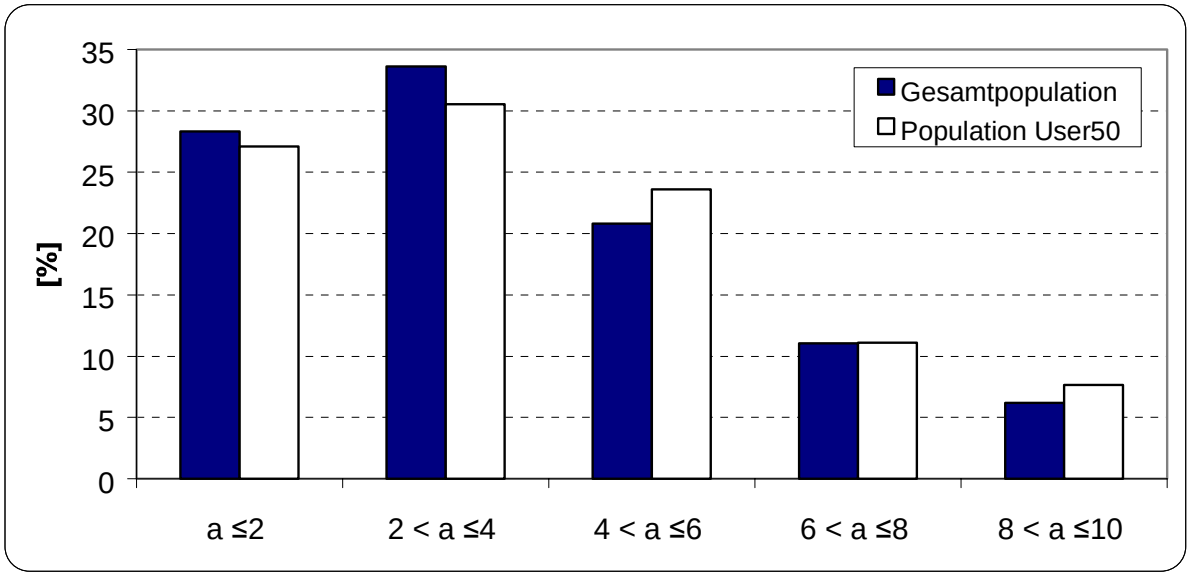

Abbildung 39: Verteilung Ausweisalter (nach Jahren)

Abbildung 38 stellt in einem Diagramm die Verteilung des Ausweisalters sowohl für die Gesamtpopulation als auch die Population User50 dar.

Anhand der im Feldtest erhobenen Verifikationsergebnisse (Matchscores) wurden für die Population User50 und die Arbeitspunkte, an denen die FAR für RefID 6 jeweils 0,1% beträgt, algorithmen-spezifisch die FRRs berechnet (siehe Abbildung 40).

[14] Von zwei Ausweisen lag das Ausweisalter außerhalb des Betrachtungszeitraums.

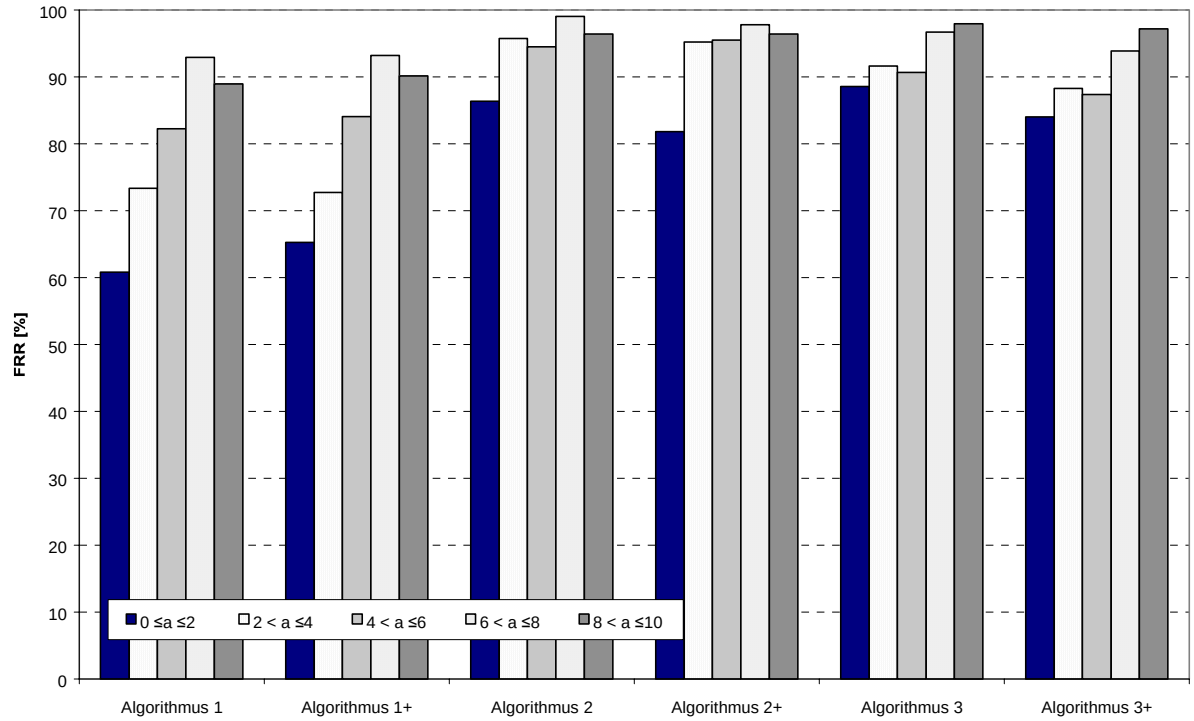

Abbildung 40: FRR in Abhängigkeit des Ausweisalters (FAR=0,1%)

Abbildung 40 stellt in einem Diagramm die FRR in Abhängigkeit des Ausweisalters für alle beteiligten Algorithmen dar.

Für alle Algorithmen lässt sich als Trend erkennen, dass die FRR mit dem Ausweisalter steigt. Dass für die höchste betrachtete Stufe (Ausweisalter 8 bis 10 Jahre) die FRR zum Teil wieder abfällt, kann auf die relativ niedrige Stichprobe (elf Ausweise in dieser Gruppe) zurückgeführt werden.

7.1.5 Einfluss der Ausweisqualität

In dieser Untersuchung wurde geprüft, wie weit die Erkennungsleistung gegenüber dem Lichtbildausweis von dessen Beschaffenheit abhängt. Da die verwendeten Musterpersonalausweise fabrikneu im Feldtest zum Einsatz kamen, wurde die Untersuchung bzgl. der Ausweisqualität anhand der aktuellen Bundespersonalausweise der Testteilnehmer – also für RefID 6 – durchgeführt.

Aus der gesamten Feldtestpopulation stellte eine Teilgruppe von 228 Personen ihren aktuellen Bundespersonalausweis für diese Untersuchung zur Verfügung.

Beim Scannen der aktuellen Bundespersonalausweise im Vorfeld des Feldtests wurden die Ausweise gemäß den folgenden Kriterien bewertet und nach Beeinträchtigungsstufen klassifiziert:

- Knicke
- Kratzer
- Risse
- Verschmutzung

Als Beeinträchtigungsstufen gab es:

- Keine Beeinträchtigung Wert 0

- Geringe Beeinträchtigung Wert 1
- Starke Beeinträchtigung Wert 2

Bei der Bewertung der Ausweisqualität wurde dabei lediglich der Bereich des Lichtbildes berücksichtigt.

Die Summe der Werte aller Kriterien für einen Ausweis stellt dann die Ausweisqualität dar. Diese liegt somit zwischen 0 und 8, wobei 0 die bestmögliche Ausweisqualität darstellt.

Dies erlaubt eine Klassifikation in folgende Qualitätsbereiche:

- Hohe Ausweisqualität Summe < 1
- Mittlere Ausweisqualität Summe zwischen 1 und 3
- Niedrige Ausweisqualität Summe > 3

Ausweis-qualität	Gesamtpopulation		Population User50	
	Anzahl absolut	Anteil relativ [%]	Anzahl absolut	Anteil relativ [%]
hoch	199	87,28	123	84,83
mittel	27	11,84	21	14,48
niedrig	2	0,88	1	0,69
Gesamt	**228**	**100,00**	**145**	**100,00**

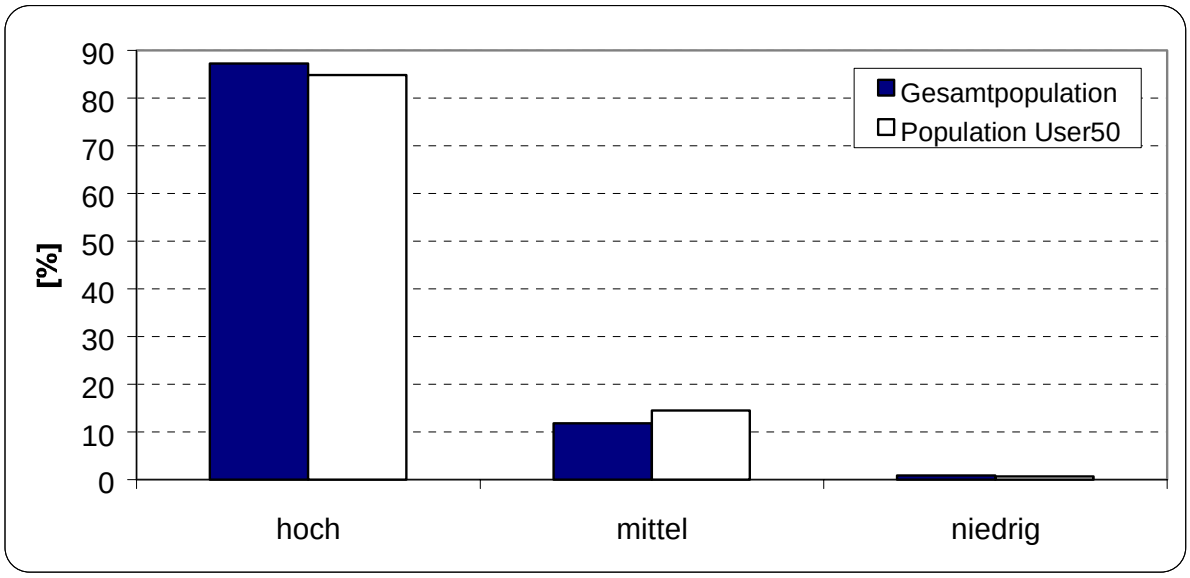

Abbildung 41: Verteilung Ausweisqualität

Abbildung 41 stellt in einem Diagramm die Verteilung der Ausweisqualität sowohl für die Gesamtpopulation als auch die Population User50 dar.

Der hohe Anteil von Ausweisen mit hoher Qualität legt den Schluss nahe, dass der Ausweis im Bereich des Bildes sehr robust ist.

Aufgrund der geringen Stichprobenmenge für Ausweise mittlerer und niedriger Qualität ist eine quantitative Gegenüberstellung der FRRs für die verschiedenen Qualitätsstufen nicht sinnvoll. Daher wird an dieser Stelle darauf verzichtet.

7.1.6 Überwindungssicherheit

Ein wichtiger Aspekt bei der Bewertung biometrischer Systeme ist der Test auf die Möglichkeit deren Überwindung. Die Überwindung ist als erfolgreich anzusehen, wenn es durch Nachstellen des biometrischen Merkmals gelingt, eine positive Verifikation zu erreichen, obwohl das präsentierte Merkmal nicht authentisch ist.

Da die Untersuchung der Überwindungssicherheit nicht das primäre Ziel des Projekts BioP I war, wurden von den Systemherstellern keine speziellen Sicherheitsmechanismen (zum Beispiel Lebenderkennung) gefordert.

Die Verifikation biometrischer Merkmale wird eine Überwindung nie hundertprozentig verhindern können. Dies liegt unter anderem darin begründet, dass das Ziel einer akzeptablen Falschabweisungsrate zur Einstellung eines Schwellwertes führt, der von einem Angreifer zur Überwindung ausgenutzt werden kann.

Bei einer Verifikation kommt für einen Angreifer erschwerend hinzu, dass er im Besitz eines Tokens (bei BioP I der Musterpersonalausweis) der Person sein muss, deren biometrisches Merkmal er überwinden möchte.

Die Überwindungsversuche mit Fakes wurden im Testlabor von secunet in Essen durchgeführt. Bei diesen Tests wurde untersucht, ob Verifikationen mit Fakes zu erfolgreichen Ergebnissen führen und ob Überwindungsversuche akzeptiert werden. Da die sicherheitstechnische Untersuchung nicht zu den primären Zielen von BioP I gehört, wurde auf Angriffe mit hohem Aufwand verzichtet (zur Klassifikation von Angriffen siehe Tabelle 10). Ebenso wurde auf Angriffe auf das jeweilige Betriebssystem sowie die Anwendungssoftware verzichtet, da davon auszugehen ist, dass diese in der angestrebten Einsatzumgebung entsprechend gegen externe Angreifer geschützt werden können. Die eingesetzten Rechnersysteme wurden lediglich mit einem Penetrationstool auf Schwachstellen untersucht.

	Niedriger Aufwand	Mittlerer Aufwand	Hoher Aufwand
Motivation des Angreifers	Unbeabsichtigtes Eindringen, Spieltrieb	Neugierde, Wettkampf	Kriminelle Energie, Geheimdienstaktivität, Spionage
Benötigte Informationen	Keine	Öffentlich zugängliche Informationen	Insider-Kenntnisse
Zeitaufwand für die Vorbereitung	Gering	Stunden bis Tage	Wochen
Zeitaufwand für die Durchführung	Gering	Stunden bis Tage	Wochen
Finanzieller Aufwand	Keiner	Gering	Kaum eingeschränkte finanzielle Mittel sind nötig
Hilfsmittel	Keine	Einfache Hilfsmittel	Spezialwerkzeug u. ä.

Tabelle 10: Kriterien zur Klassifikation von Angriffen (externe Täterklassen)

7.1.6.1 Überwindungsversuche mit Fakes

Die auf Fakes beruhenden Versuche wurden nur für Referenzbasis 7 (Systemtemplate) und Algorithmus 1 durchgeführt, da für diese Kombinationen im Feldtest relativ niedrige Streuungen bei den Matchscores auftraten und die Ergebnisse der Überwindungsversuche daher besser reproduzierbar sind.

Es wurden Überwindungsversuche mit Fotos (sowohl Schwarz/Weiß als auch Farbe) und Videos durchgeführt. Dabei wurde jeweils eine entsprechende Aufnahme einer berechtigten Person angefertigt und der Erfassungseinheit präsentiert. Bei beiden Systemen wurde das jeweilige Fake als der Berechtigte akzeptiert.

7.1.6.2 Überwindungsversuche mit biometrisch ähnlichen Personen

Grundsätzlich sind biometrisch ähnliche Personen nicht zwingend auch visuell ähnlich. Trotzdem wurden zunächst mit visuell ähnlichen Personen Überwindungsversuche durchgeführt. Die erzielten Matchscores lagen weitgehend in einem relativ niedrigen Bereich.

In einem Fall wurden jedoch die im Folgenden dargestellten Werte erreicht.

Vorgehensweise:

- Ein Unberechtigter führt mit dem Ausweis eines Berechtigten eine Reihe von zehn Betätigungen durch. Das Aussehen des Unberechtigten wurde hierfür nicht an das Aussehen des Berechtigten angepasst (zum Beispiel durch Ankleben eines Barts, Änderung der Frisur, Make-up).
- Unter gleichen Bedingungen führt der Berechtigte eine Reihe von zehn Betätigungen durch, damit eine Vergleichbarkeit der vom Unberechtigten erzielten Matchscores möglich ist.

Matchscore	Betätigungen des Unberechtigten	Betätigungen des Berechtigten
Minimum	0,719	0,763
Durchschnitt	0,742	0,768
Maximum	0,754	0,772

Tabelle 11: Vergleich Matchscores System A

Matchscore	Betätigungen des Unberechtigten	Betätigungen des Berechtigten
Minimum	0,635	0,750
Durchschnitt	0,690	0,766
Maximum	0,726	0,773

Tabelle 12: Vergleich Matchscores System B

An beiden Systemen erreicht der Unberechtigte reproduzierbar hohe Matchscores und wird somit als Berechtigter akzeptiert.

7.1.6.3 Überwindungsversuche durch Systemmanipulation

Die Überwindungsversuche durch Systemmanipulation sind in zwei Gruppen aufzuteilen:

- Manipulation von Systemkomponenten, zum Beispiel durch Installation eines Trojaners oder eines weiteren Gerätes zur Videoaufzeichnung
- Manipulation der Übertragungswege

Die Ergebnisse dieser Untersuchungen sind nicht Bestandteil des öffentlichen Abschlussberichts.

7.1.6.4 Fazit

Die dargestellten Tests haben gezeigt, dass sich die beiden biometrischen Systeme mit geringem Aufwand überwinden lassen. Ein weitaus kritischerer Aspekt ist jedoch, dass bei beiden Systemen selbst mit Zero-effort attempts (siehe Abschnitt 7.1.6.2) Überwindungen mit hohen Matchscores erzielt wurden.

Die Überwindung von biometrischen Systemen kann durch die drei folgenden Maßnahmen erschwert werden:

- durch eine Verbesserung der biometrischen Messmethode am Erfassungsgerät (Lebenderkennung),
- durch Übertragungstechniken, welche die Vertraulichkeit und Authentizität der Daten vom Erfassungsgerät zur Auswertungseinheit gewährleisten (unter anderem Verschlüsselung) und
- durch die Einschränkung des Einsatzbereichs (zum Beispiel nur in überwachten Umgebungen).

7.2 Untersuchung der Benutzerakzeptanz

Im Rahmen des Projekts BioP I erfolgten ausführliche statistische Untersuchungen über die Akzeptanz der getesteten biometrischen Systeme. Grundlage für die Akzeptanzuntersuchungen bildeten drei Befragungen der Testteilnehmer, die als Erstbefragung, Mittelbefragung und Abschlussbefragung bezeichnet wurden. Die Erstbefragung fand vor Beginn der Testphase, die Mittelbefragung nach etwa der Hälfte der Testphase und die Abschlussbefragung nach deren Ende statt.

7.2.1 Bewertung der Systeme

Für die Bewertung der im Feldtest eingesetzten Gesichtserkennungssysteme lassen sich die folgenden wesentlichen Aussagen zusammenfassen (Bewertung gemäß Schulnotensystem):

- Bei der Frage nach der **Einfachheit der Bedienung** hat das System B einen leichten Vorteil gegenüber dem System A, wobei die Beurteilung sich von der Mittel- zur Endbefragung nicht wesentlich änderte. Dass beide Systeme deutlich besser als Note 2 abgeschnitten haben, lässt darauf schließen, dass die Einfachheit der Bedienung der getesteten Systeme bereits ein hohes Niveau erreicht hat.
- Die Bewertungen der **Erkennungsgenauigkeit**, bei denen erneut System B vorne lag, wurden im Lauf des Versuchs etwas schlechter, obwohl erwartungsgemäß eine wachsende Vertrautheit mit den Systemen die Erkennungsgenauigkeit tatsächlich erhöhen sollte. Möglicherweise beruht diese Einschätzung auf zunehmenden Ansprüchen, die die Nutzer an die Systeme stellen.
- Bei der Beurteilung der **Schnelligkeit** ergibt sich ein Vorteil für System B, während das System A hier schlechter bewertet wurde. Die Diskrepanz zwischen den beiden Systemen ist in diesem Bereich größer als in allen anderen abgefragten Kategorien.
- Auch bei der **Störanfälligkeit** liegt der Vorteil bei System B. Die Bewertungen sind an dieser Stelle die schlechtesten unter den abgefragten Kategorien. Dies bedeutet, dass die Störanfälligkeit in der Einschätzung der Anwender der größte Mangel unter den vorgegebenen Möglichkeiten ist.
- Bei der Beurteilung der **Flexibilität** ist der Anteil der Personen, die keine Angabe machten, hoch. Möglicherweise beruht dies darauf, dass bei vielen der Teilnehmer während der Testphase keine wesentlichen Änderungen des äußeren Erscheinungsbildes bzgl. Brille, Frisur oder Bart auftraten. Die Bewertung verschlechterte sich im Verlauf des Tests. Dies könnte

zum einen daran liegen, dass der Unterschied zwischen dem Erscheinungsbild und den Referenzdaten im Laufe der Zeit größer wurde, zum anderen könnte sich aber auch das subjektive Empfinden der Teilnehmer geändert haben.

- Die positivere Bewertung von System B gegenüber System A bestätigt sich schließlich in der **zusammenfassenden Beurteilung der Systeme**, die sich im Testverlauf nur unwesentlich verändert hat. In der Endbewertung erreicht das System A eine Durchschnittsnote von 2,7 gegenüber System B mit 1,8.

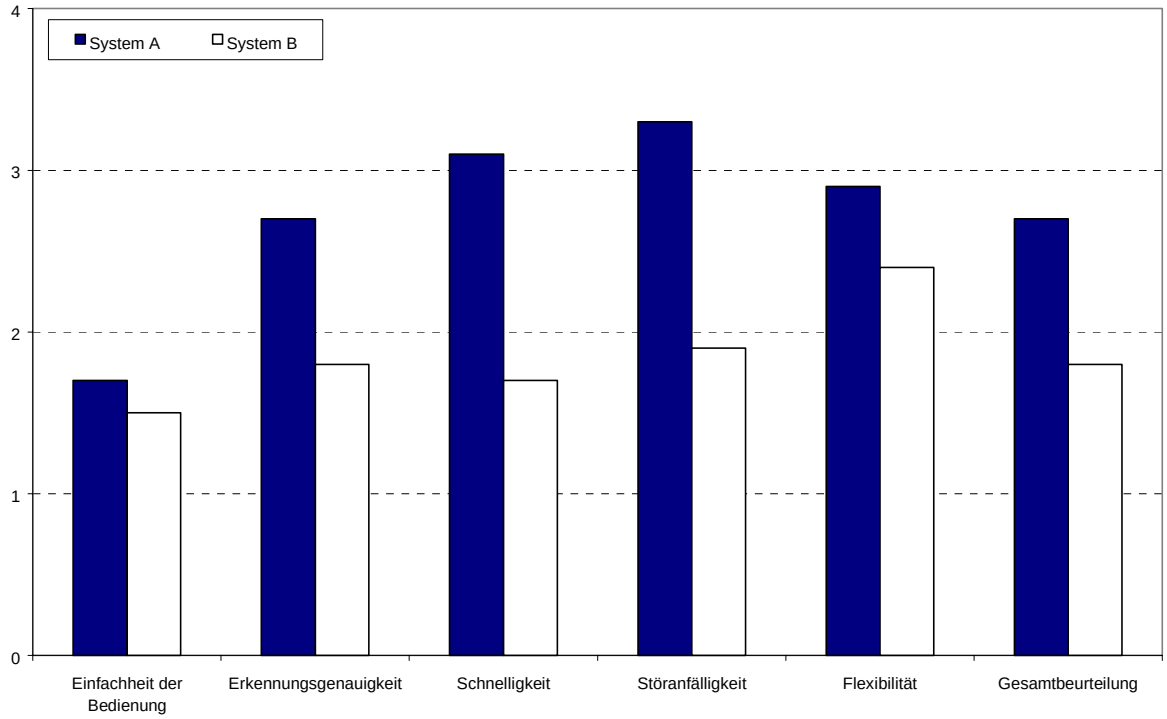

Abbildung 42: Systembewertung durch die Testteilnehmer nach Feldtestende

Abbildung 42 stellt in einem Diagramm die Systembewertung der Teilnehmer für beide Systeme anhand der Einzelkategorien sowie die Gesamtbeurteilung dar.

Zusammenfassend lässt sich feststellen, dass System B im Anwenderurteil besser abgeschnitten hat als das System A. In allen fünf Einzelgattungen sowie in der Gesamteinschätzung liegt System B um mehrere Zehntelnoten vorn. Trotz des Rückstands ist das Ergebnis von System A nicht negativ zu bewerten, da alle Notenmittelwerte stets zwischen 1,7 und 3,3 liegen.

Insgesamt kann die Beurteilung der Systeme damit als positiv bezeichnet werden. Zu berücksichtigen ist jedoch, dass die Teilnehmer sowohl die Erkennungsgenauigkeit als auch die Toleranz gegenüber Änderungen des Gesichts im Verlauf des Versuchs schlechter einstuften. Die Bedienerfreundlichkeit der Systeme bereitet gemäß den hier erhobenen Ergebnissen zwar keine Probleme, allerdings ist deren Störanfälligkeit noch verbesserungswürdig.

7.2.2 Akzeptanz biometrischer Verfahren

Neben der Bewertung der konkret im Test eingesetzten Systeme waren die Testteilnehmer aufgefordert, auch Einschätzungen zur Gesichtserkennung und zu Biometrie im Allgemeinen zu geben. Im Folgenden sind die aus den Antworten zu den jeweiligen Fragen ermittelten Ergebnisse zusammengefasst:

- **"Gesichtserkennungssysteme sind in der Lage, in einer für die Praxis hinreichenden Güte Personen zu identifizieren":** Hier zeigt sich, dass der Anteil der Testteilnehmer, die die Erkennungsleistungen von Gesichtserkennungssystemen für praxistauglich halten, im Verlauf des Versuchs gestiegen ist. Bei der Endbefragung machte der Anteil der Zustimmenden und Unentschlossenen zusammen über 95 Prozent aus. Lediglich 2,4 Prozent stimmen dieser Aussage nicht zu.
- **"Gesichtserkennungssysteme haben mittlerweile in ihrer Entwicklung ein Stadium erreicht, in dem sie für alltägliche Aufgaben genutzt werden können":** Auch bei dieser Frage hat die Zustimmung bereits in der Mittelbefragung zugenommen und sich in der Endbefragung noch einmal leicht erhöht. Mehr als die Hälfte der Teilnehmer (53,3 Prozent) hält Gesichtserkennungssysteme nach ersten eigenen Erfahrungen für praxistauglich. Lediglich 7,7 Prozent stimmen dieser Aussage nicht zu.
- **"Gesichtserkennungssysteme sollten zur Personenerkennung nicht eigenständig eingesetzt werden, sondern immer nur als Unterstützung für eine Person dienen, die die Kontrolle durchführt":** Die Antworten stehen hier in einem gewissen Widerspruch zu den zuvor beschriebenen Resultaten. So zeigt sich zwar die Mehrheit der Teilnehmer nach Abschluss des Tests davon überzeugt, dass Gesichtserkennungssysteme bereits alltagstauglich sind. Gleichzeitig ist aber eine noch deutlichere Mehrheit (68,6 Prozent), die im Verlauf des Tests gestiegen ist, dafür, derartige Technik nicht isoliert einzusetzen.
- **"Den Anteil Fehlerkennungen bei Gesichtserkennungssystemen schätze ich wie folgt: Anteil in Prozent: _____%":** Die Testteilnehmer ließen sich auf Grund der gemachten Erfahrungen offensichtlich davon überzeugen, dass Fehlerkennungen bei Gesichtserkennungs-systemen die Ausnahme sind. Daher sank der Mittelwert der geschätzten Fehlerkennungen von 32,4 Prozent im Verlauf des Versuchs auf 16,1 Prozent.
- **"Gesichtserkennungssysteme sind in ihrer Anwendung ausreichend flexibel gegenüber optischen Erscheinungsvariationen des Anwenders wie Brille, Bart, Frisur usw.":** Bei dieser Frage ist eine ansteigende Zustimmung zu verzeichnen, die jedoch auch in der Endbefragung insgesamt nicht höher als 22 Prozent ausfiel.
- **"Bei der Bedienung von Gesichtserkennungssystemen ist für den Anwender ersichtlich, wie die biometrische Identifikation im Detail abläuft":** Die Anzahl der Zustimmungen von der Mittel- zur Endbefragung ist nach einem vorherigen Anstieg wieder zurückgegangen. Die Mehrheit der Teilnehmer (59,1 Prozent) stimmt nicht zu.
- **"Die Bedienung von Gesichtserkennungssystemen ist in ihrem Ablauf einfach, schnell und bequem":** Nachdem sich die Teilnehmer mit den Testsystemen vertraut gemacht hatten, hielt eine Mehrheit (92,9 Prozent) die Bedienung von Gesichtserkennungssystemen für einfach, schnell und bequem. Vor Testbeginn war dies noch eine Minderheit (25,8 Prozent).
- **"Die Anwendung von Gesichtserkennungssystemen stellt keine Gefahr für die Gesundheit dar":** Nur eine kleine Minderheit, deren Anzahl im Verlauf des Versuchs von 3,3 Prozent auf 1,8 Prozent zurückging, hält Gesichtserkennungssysteme offenbar für möglicherweise gesundheitsschädlich.
- **"Die Nutzung von Merkmalen meines Körpers bei der Anwendung von Gesichtserkennungssystemen ist mit einem prinzipiell unangenehmen Gefühl verbunden":** Hier zeigt sich die Tendenz, dass die Teilnehmer kein unangenehmes Gefühl bei der Verwendung von Gesichtserkennungssystemen verspüren. Nur bei 6,5 Prozent trifft die Aussage zu.

- **"Bei der Nutzung von Gesichtserkennungssystemen habe ich das Gefühl, ‚erkennungsdienstlich behandelt' zu werden":** Die Mehrheit der Teilnehmer verneinte Assoziationen zu einer erkennungsdienstlichen Behandlung, bejaht wurde dies dagegen von 14,2 Prozent. Dabei zeigte sich ein leichter Anstieg, der sich von der Mittel- zur Endbefragung ergeben hat.
- **"Ich bin für die alltägliche Anwendung von Gesichtserkennungssystemen":** Hier überwiegt zwar die Zustimmung, sie geht jedoch von der Mittel- zur Endbefragung leicht zurück von 57,5 Prozent auf 54,5 Prozent.
- **"Wo sehen Sie zukünftige Vorteile des Einsatzes der Gesichtserkennungssysteme in zentralen Bereichen des öffentlichen Lebens?":** Höhere Sicherheit (68,5 Prozent) und einfacherer Zugang (65,7 Prozent) leuchten den Teilnehmern am meisten ein. Dagegen scheint der alltägliche Ärger mit Passwörtern nur wenig Einfluss auf die befürworteten Anwendungsfelder zu haben.
- **"Wo sehen Sie Probleme beim Einsatz von Gesichtserkennungssystemen?":** Technische Probleme (73,7 Prozent) sowie Fragestellungen des Datenschutzes (37,1 Prozent) stehen bei den Bedenken der Testteilnehmer vorne. Dagegen betragen mögliche gesundheitliche lediglich 3,3 Prozent.
- **"Den Anteil der Befürworter der alltäglichen Anwendung von Gesichtserkennungssystemen in meinem Bekannten- und Verwandtenkreis schätze ich auf_____%":** Die meisten Teilnehmer schätzen den Anteil der Befürworter in ihrem Bekanntenkreis auf einen Wert um die 50 Prozent. Auffällig dabei ist der hohe Prozentsatz von Personen, die keine Angabe machten (ca. 40 Prozent). Dies legt die Vermutung nahe, dass die meisten Befragten keine genaue Vorstellung von der Einstellung ihrer Mitmenschen zum Thema Gesichtserkennung haben.

Die folgenden Aussagen wurden nur während der Erstbefragung erhoben:

- **"Falls Gesichtserkennungssysteme im Rahmen der Maßnahmen zur Terrorismusbekämpfung eingesetzt werden, finde ich das...":** Das Thema Biometrie in der Terrorismusbekämpfung wird von den Testteilnehmern positiv gesehen (84,5 Prozent).
- **"Falls Gesichtserkennungssysteme von staatlichen Behörden im Zusammenhang mit dem Personalausweis eingesetzt werden, finde ich das....":** Die Testpersonen stehen einem solchen Einsatz von Gesichtserkennungssystemen positiv gegenüber (72,8 Prozent).
- **"Falls eine andere Form der biometrischen Personenerkennung von staatlichen Behörden eingesetzt wird (etwa Fingerabdruckerkennung oder Handerkennung), finde ich das...":** Die Antworten weichen hier nur unwesentlich von der vorhergehenden Frage ab.
- **"Wie schätzen Sie generell die Nützlichkeit von Systemen zur biometrischen Personenerkennung ein?":** Hier ist eine Polarisierung zwischen „sehr nützlich" (33,3 Prozent) und „teils teils" erkennbar, wobei die Mehrheit (60,6 Prozent) zu „teils teils" tendiert.
- **"Welches Körpermerkmal halten Sie für biometrische Personenerkennung als am besten geeignet?":** Der überwiegende Anteil der Testteilnehmer (43,7 Prozent) halten die Fingerabdruckerkennung für am besten geeignet. Iriserkennung wird mit 37,1 Prozent für am besten geeignet gehalten, Gesichtserkennung erreicht 14,1 Prozent.

Die Auswertung der Befragung lässt insgesamt zwei Tendenzen erkennen. Zum einen scheinen die Teilnehmer im Verlauf der Tests eine zunehmend positive Einstellung zu zahlreichen Detailfragen der Gesichtserkennung entwickelt zu haben. So hält nur eine Minderheit die Gesichtserkennung für

gesundheitsgefährdend, während die Praxisreife und die Zuverlässigkeit von einer großen Mehrheit der Teilnehmer als gegeben gesehen werden.

Auf der anderen Seite ist trotz dieser grundsätzlich zustimmenden Haltung Skepsis der Teilnehmer in der Gesamtheit zu verzeichnen. So befürwortet eine Mehrheit die Forderung, dass Gesichtserkennung nicht isoliert eingesetzt werden darf. Auch eine generelle Nützlichkeit wird nur von einem Drittel der Teilnehmer gesehen. Möglicherweise dient zur Steigerung der Akzeptanz der Biometrie in der Öffentlichkeit das Aufzeigen konkreter Nutzungsszenarien.

8 Bewertungsschema

Ein zentrales Ziel des Projekts BioP I ist eine Einschätzung der Qualität der getesteten Gesichtserkennungssysteme. Dadurch soll insbesondere ein Vergleich der beiden Komplettsysteme sowie der einzelnen Algorithmen (Matching Engines) möglich werden. Voraussetzung für einen Vergleich sind Kennzahlen, die mit einem Bewertungsschema berechnet werden. Das für BioP I entwickelte Bewertungsschema wird im Folgenden dargestellt.

8.1 Aufbau des Bewertungsschemas

Im Mittelpunkt des Bewertungsschemas steht die Berechnung zweier Indizes, die als Komplettsystem-Index (KSI) und als Matching-Engine-Index (MEI) bezeichnet werden. Ein Index ist in der Statistik ein gewichteter Mittelwert, der aus mehreren Parametern gewonnen wird. In diesem Fall sind die Parameter Bewertungskriterien aus dem Feldversuch, den weiterführenden Untersuchungen sowie herstellerspezifischem Verhalten. Der KSI liefert jeweils eine Kennzahl für die beiden Komplettsysteme, der MEI entsprechend für die einzelnen Matching Engines.

Die Ergebnisse, die ein Komplettsystem oder eine Matching Engine im Feldtest und in einigen der weiterführenden Untersuchungen erreicht, unterscheiden sich für die verschiedenen Referenzbasen zum Teil erheblich. Das Bewertungsschema ist daher auf drei Referenzbasen beschränkt, die mit unterschiedlicher Gewichtung betrachtet werden.

Nicht alle Bewertungskriterien sind für das Endergebnis gleich wichtig. Daher wird jedes Kriterium mit einem Gewicht versehen, welches den Anteil des Kriteriums an der Gesamtnote widerspiegelt.

Der Wertebereich eines jeden Bewertungskriteriums beruht auf dem gängigen Schulnotensystem mit Werten von 1 bis 6.

Für jedes Bewertungskriterium wird das Gewicht mit der Note multipliziert. Die Addition der gewichteten Notenwerte ergibt den Indexwert des jeweiligen Komplettsystems bzw. der Matching Engine. Es handelt sich dabei somit um eine Zahl zwischen 1 und 6, die die Gesamtnote wiedergibt.

8.2 Auswahl zu betrachtender Referenzbasen

In die Bewertung der Systeme und Algorithmen sollen nur Ergebnisse für die Referenzbasen eingehen, die für potentielle Anwendungsszenarien eine Rolle spielen. Dabei sind im Wesentlichen die drei folgenden Realisierungsalternativen für zukünftige Ausweisdokumente von Interesse:

- **Speicherung einer Bilddatei auf dem Ausweis**: Wahrscheinlichste Realisierungsvariante ist die Speicherung einer Bilddatei auf dem Ausweisdokument gemäß ICAO-Richtlinien. Wegen der Problematik des begrenzten Speicherplatzbedarfs wird die komprimierte Variante der Bilddatei präferiert. Dies entspricht in BioP I Referenzbasis 4, die daher mit höchstem Gewicht (60 Prozent) in die Bewertung eingeht.
- **Verwendung eines auf dem Ausweisdokument vorhandenen Lichtbildes**: Als Fallback-Variante im Falle eines Defekts des Speicherchips bzw. für Dokumente, die keinen entsprechenden Chip besitzen, sowie für einen parallelen Einsatz in einer Übergangsphase ist die Verwendung des Lichtbildes relevant. Das Lichtbild muss dabei ICAO-Richtlinien erfüllen. Dies entspricht in BioP I Referenzbasis 8. Diese geht mit einem Gewicht von 30 Prozent in die Bewertung ein. Der aktuelle Personalausweis (Referenzbasis 6) ist dagegen nach den hier erzielten Ergebnissen aufgrund des Halbprofils der Lichtbilder sowie zum Teil schlechter Bildqualität nicht verwendbar.

- **Speicherung eines Templates auf dem Ausweis**: Dies entspricht in BioP I Referenzbasis 7. Diese Alternative wird mitbetrachtet, da sie die besten Erkennungsleistungen liefert. Da diese jedoch nicht den ICAO-Empfehlungen entspricht, geht sie nur mit niedrigem Gewicht (10 Prozent) in die Bewertung ein.

8.3 Bewertungskriterien

Die Bewertung der Algorithmen und Systeme erfolgt anhand der folgenden Kriterien:

- **FER**: Die FER ist in Zusammenhang mit der FRR zu sehen. Algorithmen bzw. Systeme, die niedrige FRRs liefern, führen ggf. eine Vorsortierung von akzeptierten Aufnahmen durch. Beim realen Einsatz, zum Beispiel bei der Identitätsprüfung an der Grenzkontrolle, geht die FER direkt in die FRR ein. Daher darf sie bei einer separaten Betrachtung nicht vernachlässigt werden. Dieses Kriterium ist sowohl für den Algorithmen- als auch den Systemvergleich relevant.
- **FRR (FAR=X%)**: Für den Algorithmenvergleich ist die Erkennungsleistung das wesentliche Kriterium. Dabei wird jedoch differenziert nach Erkennungsleistungen für verschiedene Sicherheitsniveaus.
    - FAR=0,01% entspricht sehr starker Sicherheit und ist primär für High-Security-Anwendungen relevant.
    - FAR=0,1% entspricht am ehesten einem realistischem Einsatzszenario mit einem akzeptablen Sicherheitsniveau.
    - FAR=1% wird niedriger bewertet, da in diesem Fall das Sicherheitsniveau schon etwas aufgeweicht ist.
- **FRR(TH=0,7), FAR(TH=0,7)**: Die Erkennungsleistung hat beim Vergleich biometrischer Systeme das höchste Gewicht. Ob dabei FRR oder FAR stärker gewichtet werden, hängt davon ab, ob eher ein akzeptables Sicherheitsniveau oder ein akzeptabler Komfort zu gewährleisten ist. Im Falle der Betrachtung der Erkennungsleistung bei einem festen Threshold hängen FRR und FAR direkt voneinander ab und können daher gleich gewichtet werden.
- **Standardabweichung Einzelbenutzerstatistik**: Eine hohe Standardabweichung besagt, dass ein großer Anteil der Nutzer hohe Rückweisungsraten hat. Beim Algorithmenvergleich ist dies ein Kriterium, um Algorithmen abzuwerten, die bei vielen Personen Probleme bereiten. Beim Systemvergleich ist die Standardabweichung schwächer zu gewichten, da hier ein direkter Zusammenhang mit den benutzerbedingten Problemen besteht. Allgemein sollte die Standardabweichung in Relation zur Erkennungsleistung nicht zu stark gewichtet werden, da sonst ggf. Algorithmen bzw. Systeme eine bessere Bewertung erhalten, die zwar eine sehr niedrige Standardabweichung haben, dies aber bei hohen FRRs.
- **Mittlere Bedienzeit**: Dieses Kriterium besitzt für den Systemvergleich Relevanz, da dies die für den Benutzer spürbare Zeit ist. Da die biometrische Identitätsprüfung im realen Einsatz in der Regel in einen Gesamtprozess (etwa Einreisekontrolle) eingebettet ist, wird die Bedienzeit im Verhältnis zu den anderen Kriterien nicht stark gewichtet.

- **Einfluss Lichtbedingungen**: Da auch in den geplanten Einsatzszenarien keine absolut idealen Bedingungen geschaffen werden können, muss der Einfluss der Lichtbedingungen stark bei der Bewertung berücksichtigt werden. Dies gilt sowohl für den Algorithmenvergleich, bei dem die Robustheit gegenüber schlecht ausgesteuertem Bildmaterial bewertet wird, als auch für den Systemvergleich, bei dem eine gegenüber Störeinflüssen relativ resistente Erfassungseinheit wichtig ist.

Die weiteren Kriterien sind nur für den Systemvergleich, nicht aber für den Algorithmenvergleich relevant:

- **Benutzerakzeptanz**: Die Akzeptanz ist eine relevante Kennzahl, da biometrische Systeme in der Regel ein kooperatives Verhalten seitens der Benutzer erfordern.
- **Systemfehler, Ausfallverhalten, Administrationsaufwand**: Für Systeme, die in Umgebungen mit Publikumsverkehr und daraus resultierenden hohen Betätigungszahlen eingesetzt werden, ist die Gewährleistung eines stabilen Betriebs ein Aspekt, der nach der Erkennungsleistung am stärksten gewichtet ist.
- **Benutzerbedingte Probleme**: Durch Design-Fehler des Systems verursachte Probleme für Personen mit bestimmten Merkmalen (zum Beispiel geringe Körpergröße) sind ein negativ zu bewertender Faktor.
- **Überwindungssicherheit bzgl. Zero-effort Attempt**: Diese Angriffe stellen aufgrund der einfachen Durchführbarkeit das größte Problem für die Überwindungssicherheit dar und werden daher entsprechend gewichtet.
- **Support/Service**: Beim Einsatz komplexer technischer Systeme ist zuverlässiger Hersteller-Support ein wichtiges Kriterium und geht daher mit vergleichsweise hohem Gewicht ein.

Zu den hier dargestellten gibt es noch weitere Kriterien. Diese gehen aber nur mit einem Gewicht von jeweils weniger als zwei Prozent in die Bewertung ein und werden daher an dieser Stelle nicht weiter erläutert.

8.4 Klassifikation der Ergebnisse

In diesem Abschnitt erfolgt die Darstellung der Richtlinien für die Notenvergabe zu den einzelnen Bewertungskriterien.

FER	Note
≤ 0,0001%	1
≤ 0,001%	2
≤ 0,01%	3
≤ 0,1%	4
≤ 1%	5
> 1%	6

Tabelle 13: Notenklassifikation zur FER

FRR	Note
≤ 2%	1
≤ 4%	2
≤ 8%	3
≤ 16%	4
≤ 32%	5
> 32%	6

Tabelle 14: Notenklassifikation zur FRR

FAR	Note
≤ 0,01%	1
≤ 0,1%	2
≤ 1%	3
≤ 3%	4
≤ 5%	5
> 5%	6

Tabelle 15: Notenklassifikation zur FAR

Standardabweichung in der Einzelbenutzerstatistik	Note
≤ 1%	1
≤ 2%	2
≤ 4%	3
≤ 8%	4
≤ 16%	5
> 16%	6

Tabelle 16: Notenklassifikation zur Standardabweichung in der Einzelbenutzerstatistik

Mittlere Bedienzeit	Note
≤ 2 s	1
≤ 4 s	2
≤ 6 s	3
≤ 8 s	4
≤ 10 s	5
> 10 s	6

Tabelle 17: Notenklassifikation zur mittleren Bedienzeit

Systemfehler	Note
Keine oder nur marginale Fehler während des Feldtests	1
Kleine Fehler während des Feldtests, die durch einfache Maßnahmen behoben werden konnten	2
Fehler während des Feldtests, die durch Maßnahmen mittleren Aufwandes behoben werden konnten	3
Fehler während des Feldtests, die durch Maßnahmen hohen Aufwand behoben werden konnten	4
Fehler während des Feldtests, die nicht behoben werden konnten	5
Fehler während des Feldtests, die einen zweckgemäßen Einsatz verhinderten	6

Tabelle 18: Notenklassifikation zu Systemfehlern

Ausfallverhalten	Note
Keine Ausfälle während des Feldtests	1
Weniger als zwei Ausfälle während des Feldtests, Betriebsbereitschaft konnte dabei jeweils innerhalb einer Stunde wiederhergestellt werden	2
Weniger als fünf Ausfälle während des Feldtests, Betriebsbereitschaft konnte dabei jeweils innerhalb einer Stunde wiederhergestellt werden	3
Weniger als zehn Ausfälle während des Feldtests, Betriebsbereitschaft konnte dabei jeweils innerhalb 24 Stunden wiederhergestellt werden	4
Zehn oder mehr Ausfälle während des Feldtests, Betriebsbereitschaft konnte dabei jeweils innerhalb 24 Stunden wiederhergestellt werden	5
Zehn oder mehr Ausfälle während des Feldtests, Betriebsbereitschaft konnte dabei nicht immer innerhalb 24 Stunden wiederhergestellt werden	6

Tabelle 19: Notenklassifikation zum Ausfallverhalten

Kriterien für die Bewertung des Administrationsaufwands:

- Nach Bereitstellung des Betriebszustandes ist keine Administration erforderlich.
- Das Enrolment wird durch geeignete Werkzeuge wie Qualitätsbewertung der Enrolmentbilder und eine Testverifikation unterstützt.
- Die Administrationsoberfläche ist intuitiv bedienbar.
- Es stehen geeignete Reporting-Werkzeuge zur Verfügung.

Administrationsaufwand	Note
Einhaltung aller Kriterien	1
Ein Kriterium wird in zumutbarer Weise nicht erfüllt	2
Ein Kriterium wird nicht erfüllt	3
Zwei Kriterien werden nicht erfüllt	4
Drei Kriterien werden nicht erfüllt	5
Die Kriterien werden nicht erfüllt	6

Tabelle 20: Notenklassifikation zum Administrationsaufwand

Benutzerbedingte Probleme	Note
Keine Probleme festgestellt, die auf Eigenschaften oder Verhalten von Benutzern zurückzuführen sind	1
Marginale Probleme	2
Vereinzelt Probleme, die durch einfache Maßnahmen behoben werden konnten	3
Vereinzelt Probleme, die nicht durch einfache Maßnahmen behoben werden konnten	4
Gehäuft Probleme, die durch einfache Maßnahmen behoben werden konnten	5
Massive Probleme, die nicht durch einfache Maßnahmen behoben werden konnten	6

Tabelle 21: Notenklassifikation zu benutzerbedingten Problemen

Benutzerakzeptanz	Note
Bewertung des Systems mit „sehr gut“	1
Bewertung des Systems mit „gut“	2
Bewertung des Systems mit „befriedigend“	3
Bewertung des Systems mit „ausreichend“	4
Bewertung des Systems mit „mangelhaft“	5
Bewertung des Systems mit „ungenügend“	6

Tabelle 22: Notenklassifikation zur Benutzerakzeptanz

Zero-effort Attempts	Note
Keine Verwechslung herbeigeführt mit Stichprobe < 100.000	1
Verwechslung herbeigeführt mit Stichprobe > 10.000	2
Verwechslung herbeigeführt mit Stichprobe > 1.000	3
Verwechslung herbeigeführt mit Stichprobe > 100	4
Verwechslung herbeigeführt mit Stichprobe > 10	5
Verwechslung herbeigeführt mit Stichprobe ≤ 10	6

Tabelle 23: Notenklassifikation zu Zero-effort Attempts

Die Inbetriebnahme ist nicht als separates Bewertungskriterium im Bewertungsschema aufgeführt. Die ermittelte Note geht zu 1/3 in die Note für Support / Service ein.

Inbetriebnahme	Note
Termingerechte Bereitstellung eines voll funktionsfähigen Systems	1
Termingerechte Bereitstellung des Systems, aber nachträgliche Anpassungen erforderlich	2
Termingerechte Bereitstellung des Systems und wesentliche nachträgliche Anpassungen erforderlich oder verspätete Bereitstellung eines voll funktionsfähigen Systems	3
Verspätete Bereitstellung des Systems und nachträgliche Anpassungen erforderlich	4
Verspätete Bereitstellung des Systems und wesentliche nachträgliche Anpassungen erforderlich	5
Keine Bereitstellung eines geeigneten Systems	6

Tabelle 24: Notenklassifikation zur Inbetriebnahme

Kriterien für die Bewertung von Support / Service:

- Reaktion und Lösung während Feldtest immer innerhalb 24 h
- Reaktion und Lösung während weiterführender Untersuchungen immer innerhalb 48 h
- Benennung eines kompetenten Ansprechpartners
- Bereitstellung einer geeigneten Dokumentation

Support / Service	Note
Einhaltung aller Kriterien	1
Ein Kriterium wird in zumutbarer Weise nicht erfüllt	2
Ein Kriterium wird nicht erfüllt	3
Zwei Kriterien werden nicht erfüllt	4
Drei Kriterien werden nicht erfüllt	5
Die Kriterien werden nicht erfüllt	6

Tabelle 25: Notenklassifikation zu Support / Service

Die ermittelte Note geht zu 2/3 in die Bewertung von Support / Service ein, die Note der Inbetriebnahme zu 1/3.

9 Zusammenfassung und Interpretation der Ergebnisse

Ziel des Projekts BioP I war die Untersuchung der Leistungsfähigkeit von Gesichtserkennungssystemen für den geplanten Einsatz in Lichtbilddokumenten. Dabei steht die Frage im Vordergrund, ob Gesichtserkennung für diesen Einsatzzweck geeignet ist und in welcher Form und Qualität die Referenzbasis bereitgestellt werden muss.

Die Beantwortung dieser Fragestellung kann nur auf Basis konkreter Implementierungen von Gesichtserkennung in Form von Algorithmen und Komplettsystemen erfolgen. Weiterhin müssen verschiedene infrage kommende Referenzbasen parallel getestet werden.

Entsprechend basiert BioP I auf verschiedenen Vergleichstypen: einem Algorithmenvergleich, einem Systemvergleich und einem Referenzbasenvergleich. Die Ergebnisse dieser Gegenüberstellungen sind nachfolgend zusammengefasst.

Im Anschluss daran werden die wesentlichen Einflussfaktoren für Gesichtserkennungssysteme sowie Aspekte der Überwindungssicherheit zusammengefasst. Auf Basis dieser Ergebnisse wird abschließend die Frage zur grundsätzlichen Eignung von Gesichtserkennung für die Verwendung mit Lichtbildausweisen beantwortet.

9.1 Algorithmenvergleich

Die folgenden Übersichten stellen eine Einordnung der getesteten Algorithmen bezüglich ihrer Erkennungsleistung für die verschiedenen Referenzbasen dar. Tabelle 26 bezieht sich dabei auf ein Sicherheitsniveau, das gemäß den technischen Evaluierungskriterien des BSI [TechEval] der Stufe „stark“ entspricht. Das Tabelle 27 zugrunde liegende Sicherheitsniveau entspricht der Stufe „sehr stark“. Die Farbgebung bezieht sich jeweils auf die im Feldtest ermittelten Falschrückweisungsraten gemäß der folgenden Einteilung:

- Grün: $0\% \leq FRR \leq 2\%$ (Note 1 gemäß Bewertungsschema)
- Gelb: $2\% < FRR \leq 8\%$ (Note 2 oder 3 gemäß Bewertungsschema)
- Orange: $8\% < FRR \leq 16\%$ (Note 4 gemäß Bewertungsschema)
- Rot: $FRR > 16\%$ (Note 5 oder 6 gemäß Bewertungsschema)

RefID	Algorith. 1	Algorith. 1+	Algorith. 2	Algorith. 2+	Algorith. 3	Algorith. 3+
1						
2						
3						
4						
5						
6						
7						
8						

Tabelle 26: Einordnung der Kombinationen vom Algorithmus und Referenzbasis bei Sicherheitsniveau „stark“ (FAR=1%)

Tabelle 26 enthält die Einordnung der Kombination von Algorithmus und Referenzbasis gemäß der Farbklassifikation für das Sicherheitsniveau „stark“.

RefID	Algorith. 1	Algorith. 1+	Algorith. 2	Algorith. 2+	Algorith. 3	Algorith. 3+
1						
2						
3						
4						
5						
6						
7						
8						

Tabelle 27: Einordnung der Kombinationen vom Algorithmus und Referenzbasis bei Sicherheitsniveau „sehr stark“ (FAR =0,1%)

Tabelle 27 enthält die Einordnung der Kombination von Algorithmus und Referenzbasis gemäß der Farbklassifikation für das Sicherheitsniveau „sehr stark“.

Der Algorithmus 1 hat bei allen Referenzbasen am besten abgeschnitten.

9.2 Systemvergleich

Neben dem Vergleich der Algorithmen interessiert bei der Berücksichtigung des Einsatzszenarios der Vergleich der Komplettsysteme. Neben biometriespezifischen Kriterien sind hier auch Aspekte wie Ausfallverhalten, Systemfehler, Administrationsaufwand und Support relevant. Tabelle 28 gibt einen Überblick zur Gegenüberstellung der Systeme. Die Farbgebung orientiert sich dabei an der Notenklassifikation gemäß Systembewertungsschema wie folgt:

- Grün: Note 1
- Gelb: Note 2 oder Note 3
- Orange: Note 4
- Rot: Note 5 oder Note 6

Kriterium	System A	System B
Erkennungsleistung[15]	3,11	3,33
Systemverhalten[16]	4,50	2,23
Weiterführende Untersuchungen[17]	5,16	4,93
Herstellerbeurteilung[18]	5,00	2,30

Tabelle 28: Bewertung der Komplettsysteme bzgl. Kriteriengruppen

Während sich bezüglich der biometrischen Erkennungsleistung ein leichter Vorteil für das System A ergibt, lässt sich für die anderen Bewertungskriterien ein zum Teil großer Vorsprung für System B erkennen. Insbesondere bei Aspekten der Zuverlässigkeit, der Systemfehler, des Administrationsaufwands und des Supports konnte der Hersteller von System B ein deutlich positiveres Ergebnis erzielen. Entsprechenden Kriterien kommt bei Berücksichtigung eines breiten Einsatzes und auch bei einer Auswahl für BioP II große Bedeutung zu.

9.3 Referenzbasenvergleich

Für die Referenzbasis kommen grundsätzlich drei Alternativen in Betracht, die nachfolgend mit den entsprechenden Ergebnissen dargestellt sind. Eine Zusammenfassung der Ergebnisse mit Farbklassifizierung ist in Abbildung 43 dargestellt.

9.3.1 Bereitstellung des biometrischen Merkmals als Lichtbild

In jedem Fall hat sich gezeigt, dass der Bundespersonalausweis in der gegenwärtigen Form nicht im Zusammenhang mit biometrischer Gesichtserkennung verwendbar ist. Dies begründet sich im Wesentlichen durch die Gesichtsdarstellung im Halbprofil sowie die im Einzelfall sehr schlechte Lichtbildcharakteristik bezüglich Kontrast und Helligkeit.

Der für das Projekt hergestellte Musterpersonalausweis mit einem Lichtbild gemäß den Richtlinien der ICAO zeigt, dass Gesichterkennung grundsätzlich auf Basis eines zu scannenden Lichtbildes möglich ist. Die erzielten Ergebnisse sind zwar noch nicht zufrieden stellend, lassen aber ein gewisses Potential erkennen. In jedem Fall müssen hierfür jedoch große Anstrengungen seitens der Algorithmenhersteller aufgewendet werden, um befriedigende Erkennungsleistungen zu erreichen.

Das getestete Lichtbild des EU-Visums fällt dagegen schon signifikant bezüglich der Erkennungsleistung ab. Ursache sind im Wesentlichen die Störungen innerhalb des Lichtbildes, welche durch die optischen Sicherheitsmerkmale des Visums hervorgerufen werden.

[15] Besteht aus den Einzelkriterien FER, FAR, FRR, Standardabweichung Einzelbenutzerstatistik

[16] Besteht aus den Einzelkriterien Systemfehler, Ausfallverhalten, Administrationsaufwand, benutzerbedingte Probleme, mittlere Bedienzeit

[17] Besteht aus den Einzelkriterien Benutzerakzeptanz, Überwindungssicherheit, Einfluss Lichtbedingungen

[18] Besteht aus den Einzelkriterien Support / Service (inkl. Inbetriebnahme)

9.3.2 Bereitstellung des biometrischen Merkmals als Bilddatei

Als Alternative zur direkten Verwendung des Lichtbilds ist die Nutzung einer in elektronischer Form auf dem Ausweisdokument vorliegenden Bilddatei zu betrachten. Dies entspricht den ICAO-Empfehlungen und ermöglicht somit internationale Interoperabilität. Die mit dieser Alternative erzielbaren Erkennungsleistungen liegen zwar in einem nicht sehr zufrieden stellenden Bereich. Das vorhandene Optimierungspotential seitens der Algorithmen lässt jedoch eine signifikante Verbesserung erwarten, sodass bei Ausnutzung der vorhandenen Verbesserungsspielräume (zum Beispiel Verwendung spezieller Kamerasysteme, Optimierung der Algorithmen auf die Verarbeitung von Bilddateien, geeignete Vorverarbeitung des Bildmaterials) durchaus ein erfolgreicher Einsatz denkbar ist. Die durch ICAO empfohlene Kompression der zugrunde liegenden Bilddateien hat keinen wesentlichen Einfluss auf die Erkennungsleistung.

Der Test der Bilddatei auf Basis einer Halbprofilaufnahme zeigt deutlich, dass dieser Fototyp für die Gesichtserkennung ungeeignet ist. Dies unterstreicht die Ergebnisse zum aktuellen Personalausweis.

9.3.3 Bereitstellung des biometrischen Merkmals als Template

Eine weitere Alternative ist die Nutzung eines in elektronischer Form auf dem Ausweisdokument gespeicherten Templates. In den Tests von BioP I wurden erwartungsgemäß mit der Repräsentation des Gesichts als herstellerspezifisches Template die mit Abstand besten Erkennungsleistungen erzielt.

Bezüglich der Interoperabilität sind bei dieser Realisierungsalternative jedoch erhebliche Schwierigkeiten zu erwarten. Da ein solches Template immer spezifisch für das System eines Herstellers ist, müsste international eine Festlegung auf ein System bzw. ein Verfahren erfolgen.

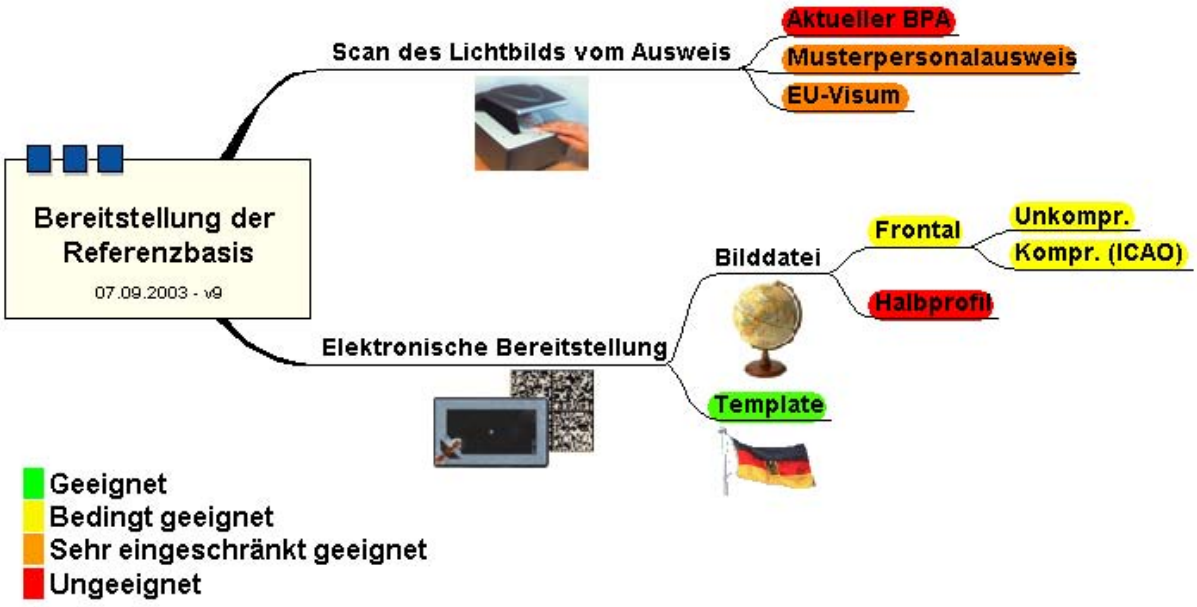

Abbildung 43: Eignung der Referenzbasen zur Gesichtserkennung

Abbildung 43 gibt einen grafischen Überblick über die Eignung der im Test untersuchten Referenzbasen anhand einer Farbklassifikation.

9.4 Einflussfaktoren für Gesichtserkennung

9.4.1 Lichtverhältnisse

Die Abhängigkeit von der Beleuchtung bezüglich Intensität und Richtung ist hinlänglich als hauptsächlicher Störeinfluss für die Gesichtserkennung bekannt. Dies konnte durch die Untersuchungen in BioP I bestätigt werden. Allerdings ist für verschiedene Algorithmen und verschiedene Systeme die Beeinflussung unterschiedlich stark. Die Leistungsfähigkeit der Erfassungseinheit, also des Kamerasystems, spielt dabei eine wesentliche Rolle.

Der größte Abfall der Erkennungsleistung ist für alle Algorithmen und Systeme bei Lichteinfall von der Seite zu verzeichnen. Der Lichteinfall aus dem Hintergrund kann bei Einsatz eines geeigneten Kamerasystems annähernd vernachlässigt werden. Bei starkem Frontallicht fiel ein recht überraschender Effekt auf. Im Wesentlichen ergab sich hier eine Verschlechterung der Erkennungsleistung, während jedoch im Einzelfall eine deutliche Verbesserung auftrat. Von den beteiligten Algorithmen erwies sich in fast allen Fällen Algorithmus 1 am robustesten.

Die Schaffung stabiler Beleuchtungsverhältnisse ist somit eine Grundvoraussetzung für den erfolgreichen Einsatz von Gesichtserkennungssystemen.

9.4.2 Qualität der Bilddatei

Im Hinblick auf die beschränkte Speicherkapazität eines möglichen Chips auf dem Personaldokument ist es von Vorteil, die zu speichernden Informationen möglichst stark zu komprimieren. Deshalb wurde die Beeinflussung der Erkennungsleistung durch unterschiedliche Kompressionsstufen für die als Referenzbasis verwendeten Bilddateien geprüft. Dabei fiel auf, dass die Erkennungsleistung mit zunehmender Kompression abnimmt. Während bei schwacher Kompression (Bildgröße ca. 75kB) ein zu vernachlässigender Rückgang zu verzeichnen ist, ergibt sich bei sehr starker Kompression (Bildgröße ca. 11kB) ein deutlicher Abfall. Eine Kompression, die sich in der von ICAO vorgeschlagenen Größenordnung bewegt (Bildgröße ca. 14kB), erzielt im Vergleich mit schwach komprimierten Referenzbasen noch akzeptable Erkennungsleistungen.

Als weitere Maßnahme zur Reduzierung des Speicherbedarfs wurden Bilddateien mit geringerer Auflösung getestet. Durch diese Modifikation ergeben sich leicht schlechtere Erkennungsraten.

9.4.3 Qualität des Lichtbildes auf dem Ausweis

Ein weiterer Einflussfaktor bezüglich Gesichterkennung und Personalausweis ist die Qualität des entsprechenden Dokuments. Dazu wurden die aktuellen Personalausweise der Testteilnehmer bezüglich Kratzern, Knicken, Rissen etc. klassifiziert. Dabei wurden kaum Ausweise mit mittlerer oder schlechter Qualität im Bereich des Bildes identifiziert. Dies lässt den Schluss zu, dass der Bundespersonalausweis insbesondere im Bereich des Lichtbildes sehr robust ist. Aufgrund der sehr kleinen Stichprobe von Ausweisen niedriger Qualität können keine belastbaren Aussagen bzgl. der Beeinflussung der Erkennungsleistung getroffen werden.

9.4.4 Alterungseffekte

Ein wesentlicher Aspekt zur Beurteilung der Eignung von Gesichtserkennungssystemen im Zusammenhang mit Personaldokumenten ist der Einfluss des Ausweisalters und damit des darauf enthaltenen Referenzbildes auf die Erkennungsleistung. Eine entsprechende Untersuchung wurde auf Basis der aktuellen Personalausweise der Testteilnehmer vorgenommen. Da die Erkennungsleistung auf Basis dieser Ausweise jedoch generell sehr schlecht ist, können hier keine fundierten Schlussfolgerungen gezogen werden. Trotzdem ist als Trend erkennbar, dass die Erkennungsleistung mit zunehmendem Ausweisalter abnimmt. Generell ist der Einfluss von Alterungseffekten auf

Gesichtserkennungssysteme noch nicht ausreichend untersucht, wie eine im Rahmen von BioP I durchgeführte Sichtung diesbezüglicher Forschungsaktivitäten bestätigt.

9.5 Überwindungssicherheit

Ein wesentliches Bewertungskriterium für biometrische Systeme insbesondere vor dem Hintergrund der Erhöhung der Sicherheit für das Einsatzszenario ist die Überwindungssicherheit. Die im Rahmen von BioP I durchgeführten Tests haben gezeigt, dass sich die beiden beteiligten biometrischen Systeme mit geringem Aufwand durch Kopien des biometrischen Merkmals Gesicht in Form von Fotos überwinden lassen. An dieser Stelle muss jedoch hinzugefügt werden, dass die Bereitstellung einer geeigneten Lebenderkennung kein Pflichtkriterium war. Ein sehr kritischer Aspekt ist jedoch, dass bei beiden Systemen in einem Einzelfall Verwechslungen von Personen auftraten, die sich nur ansatzweise ähnlich sind. Dies kann dazu führen, dass eine Person ohne weiteren Aufwand mit dem Ausweis einer anderen Person identifiziert wird.

9.6 Generelle Eignung der Gesichtserkennung

BioP I hat gezeigt, dass die Gesichterkennung grundsätzlich für die Verwendung mit Personaldokumenten geeignet ist. Dies gilt jedoch nur unter Einhaltung folgender Randbedingungen und Erfüllung der dargestellten Grundvoraussetzungen:

- Die Referenzbasis ist in geeigneter Weise auf dem Personaldokument bereitzustellen. Die besten Ergebnisse werden bei der Bereitstellung eines Templates erreicht. Realistischer bezüglich internationaler Einsetzbarkeit ist jedoch die Bereitstellung einer Bilddatei gemäß ICAO. Hier muss das vorhandene Optimierungspotential jedoch noch besser ausgenutzt werden, um zufrieden stellende Ergebnisse zu erzielen. Die Verwendung eines auf dem Ausweis vorhandenen Lichtbildes gemäß ICAO erscheint zwar möglich, hierfür müssen jedoch große Anstrengungen seitens der Algorithmenhersteller aufgewendet werden, um befriedigende Erkennungsleistungen zu erreichen.
- Eine wichtige Rahmenbedingung für den erfolgreichen Einsatz von Gesichterkennung ist die Schaffung einer kontrollierten Umgebung bezüglich des Lichteinflusses.
- Die Verbesserung der Überwindungssicherheit ist eine wesentliche Grundvoraussetzung für den Einsatz von Gesichtserkennungssystemen (insbesondere in unüberwachten Umgebungen). Während die Überwindung mittels Fotos aufgrund einer voraussichtlich betreuten Identitätsprüfung nur eingeschränkt kritisch erscheint, ist die Verwechslung ähnlicher Personen nicht akzeptabel.
- Bezüglich der Eignung der Gesichtserkennung für Personaldokumente gilt auch der Vorbehalt, dass Alterungseffekte noch nicht ausreichend untersucht sind. Dies ist insbesondere vor dem Hintergrund des langen Gültigkeitszeitraums dieser Dokumente zu betrachten.
- Die genannten Randbedingungen implizieren einige notwendige Änderungen an deutschen Pässen und Personalausweisen. Für eine geeignete Bereitstellung der Referenzbasen sollten Pass und Personalausweis um ein Speichermedium – vorzugsweise einen kontaktlosen Chip – erweitert werden. Als Rückfalllösung oder auch den parallelen bzw. Übergangseinsatz kommt durchaus das Lichtbild auf dem Ausweis infrage, sofern die aktuellen Aufnahmerichtlinien zur Lichtbilderstellung angepasst werden. Hier stellen die Richtlinien der ICAO für die Erstellung von Passbildern zum Einsatz für biometrische Anwendungen eine geeignete Vorlage dar. Die gleichen Richtlinien sollten für die mittels Speichermedium des Ausweises bereitgestellten Bilddateien bindend sein.

Die im Rahmen von BioP I ermittelten Ergebnisse werden innerhalb des Projekts BioP II auf Basis einer größeren Testgruppe überprüft und den biometrischen Verfahren Iris- und Fingerabdruck-erkennung gegenübergestellt. Der in BioP I erarbeitete Algorithmenvergleich zeigt eindeutig, dass der Algorithmus 1 für diese Untersuchungen präferiert werden sollte. Als Komplettsystem erscheint System B als geeignete Wahl, da dieses eindeutig bezüglich Kriterien wie Fehlerverhalten, Zuverlässigkeit, Herstellerunterstützung aber auch Akzeptanz durch die Testteilnehmer besser abgeschnitten hat.

Literatur

[BestPrac] Biometric Working Group: *Best Practices in Testing and Reporting Performance of Biometric Devices*; Version 2.10; 2002

[BioFace] BSI: *Studie BioFace I & II – Vergleichende Untersuchung von Gesichtserkennungssystemen, Öffentlicher Abschlussbericht*; 2003

[BioIS] BSI: *Studie BioIS – Vergleichende Untersuchung biometrischer Identifikationssysteme, Technische Untersuchung*; 2000

[Breite] Marco Breitenstein: *Biometrische Authentifizierung – Übersicht und Evaluation von Gesichtserkennungssystemen*; Diplomarbeit an der Technischen Universität Clausthal, Institut für Informatik; 2000

[CESG] Tony Mansfield, Gavin Kelly, David Chandler, Jan Kane: *Biometric Product Testing - Final Report*; 2001

[FRVT00] P. Jonathon Phillips, Patrick Grother, Duane M. Blackburn, Mike Bone: *Face Recognition Vendor Test 2000*

[FRVT02ER] P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi, Mike Bone: *Face Recognition Vendor Test 2002 – Evaluation Report*; März 2003

[FRVT02OS] P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi, Mike Bone: *Face Recognition Vendor Test 2002 – Overview and Summary*; März 2003

[OsbHar] Prepared by Australia (John Osborne & Terry Hartmann) for NTWG: *Guidelines for Maximising Interoperability of Facial Biometrics*; in: ICAO TECHNICAL REPORT, BIOMETRICS DEPLOYMENT, Development And Specification Of Globally Interoperable Biometric Standards For Machine Assisted Identity Confirmation Using Machine Readable Travel Documents; Outline Draft 5; 03.12.2002

[TechEval] BSI: *Technische Evaluierungskriterien zur Bewertung und Klassifizierung biometrischer Systeme*; Version 0.6; 2000

[TeTrKK] TeleTrusT AG 6: *Bewertungskriterien zur Vergleichbarkeit biometrischer Verfahren*; 2002

[VielStei] Claus Vielhauer, Ralf Steinmetz: *Sicherheitsaspekte biometrischer Verfahren: Klassifizierung von sicherheitsrelevanten Vorfällen und wesentlicher Größen zur Beurteilung der Funktionssicherheit*; 7. Deutscher IT-Sicherheitskongress des BSI; 2001

[WaAsMaMu] Wayman, Ashbourne, Mansfield, Munde: *Principles of Biometric Security System Vulnerabillity Assessment*; U.K. Biometric Working Group; 2001

Dateiname:	BioPI-Abschlussbericht_oeff-1-2.doc
Verzeichnis:	E:\internet
Vorlage:	C:\WINNT\Profiles\GerhardDieter\Anwendungsdaten\Microsoft\Vorlagen\Normal.dot
Titel:	BioP I - Öffentlicher Abschlussbericht
Thema:	
Autor:	Breitenstein, Niesing
Stichwörter:	
Kommentar:	
Erstelldatum:	09.01.2004 10:22
Änderung Nummer:	41
Letztes Speicherdatum:	08.04.2004 07:29
Zuletzt gespeichert von:	GerhardDieter
Letztes Druckdatum:	08.04.2004 07:30
Nach letztem vollständigen Druck	
Anzahl Seiten:	94
Anzahl Wörter:	23.983 (ca.)
Anzahl Zeichen:	136.708 (ca.)