Evaluierungsbericht

Titel	Evaluierung der Erkennungsgenauigkeit am Markt erhältlicher Gesichtsidentifizierungssysteme für den Einsatz in der Kriminalistik
Geheimhaltungsgrad	OFFEN
Berichtsnummer	IGD-EB-2019-01
Versionsnummer	1.0
Status	Final
Erstellungsdatum	20. Dezember 2019
Prüfzeitraum	20. Juli bis 4. November 2019
Installation der Prüfgegenstände	29. April bis 6. Mai 2019
Auftraggeber	Bundeskriminalamt KT32 Äppelallee 45 65203 Wiesbaden
Erstellt von	[redacted]
Genehmigt durch	[redacted] Abteilungsleiter »Smart Living & Biometric Technology« Fraunhofer IGD

Fraunhofer-Institut für Graphische Datenverarbeitung IGD
Biometrie-Evaluierungslabor
Fraunhoferstraße 5
64283 Darmstadt

Kurzfassung

Vier am Markt erhältliche Gesichtsidentifizierungssysteme für den Einsatz in der Kriminalistik wurden hinsichtlich ihrer Erkennungsleistung evaluiert. Gesichtsidentifizierungssysteme für den Einsatz in der Kriminalistik liefern bei Recherchen Kandidatenlisten mit wählbarer Länge (hier: 100 Kandidaten), die von Experten für forensische Gesichtserkennung überprüft werden.

Die verwendeten Gesichtsbilder wurden vom BKA für die Evaluierung zur Verfügung gestellt. Als Referenzbilder dienten in jedem getesteten System ca. 4,8 Millionen frontale Gesichtsbilder von ca. 3 Millionen Personen. Die Referenzdatenbanken blieben für alle Recherchen unverändert. Wenn vorhanden, wurden mehrere Gesichtsbilder pro Testperson enrollt, ohne sie miteinander zu verknüpfen.

Als Probebilder dienten

- 10 000 Frontalbilder von 10 000 zufälligen Personen mit passendem Gegenstück in der Referenzdatenbank,
- 10 000 Frontalbilder von 10 000 zufälligen Personen ohne passendes Gegenstück in der Referenzdatenbank,
- 10 000 Frontalbilder von 10 000 Personen mit Brille und mit passendem Gegenstück in der Referenzdatenbank,
- 10 000 Frontalbilder von 10 000 Bartträgern und mit passendem Gegenstück in der Referenzdatenbank,
- 600 Frontalbilder von 147 Personen mit bekanntem zeitlichen Abstand zum passenden Gegenstück in der Referenzdatenbank (bis zu ca. 9 Jahre),
- 10 000 Halbprofilbilder von 10 000 zufälligen Personen mit passendem Gegenstück in der Referenzdatenbank,
- 10 000 Halbprofilbilder von 10 000 zufälligen Personen ohne passendes Gegenstück in der Referenzdatenbank,
- bis zu 257 Gesichtsbilder von 181 Personen aus verschiedenen Aufnahmewinkeln:
  - Bilder, auf denen der Kopf um 10°, 20°, 30°, 45°, 60°, 70°, 80° bzw. 90° in eine Richtung nur um die Hochachse gedreht ist (»Yaw Angle«),
  - Bilder, auf denen der Kopf um -45°, -30°, -20°, -10°, 10°, 20°, 30° bzw. 45° nur um die Querachse gesenkt bzw. gehoben ist (»Pitch Angle«),
  - Bilder, auf denen der Kopf um 10°, 20°, 30° bzw. 45° in eine Richtung nur um die Längsachse geneigt ist (»Roll Angle«).

Die für den Einsatz in der Kriminalistik interessanteste Kenngröße ist die Falschnegatividentifizierungsrate beim Rang 100 (abgekürzt Rang-100-FNIR). Bei den Recherchen anhand der 10 000 Frontalbilder von zufälligen Personen mit passendem Gegenstück in der Referenzdatenbank erreichten die besten getesteten Systeme eine Rang-100-FNIR von 0,3 ± 0,1 %. Die Rang-100-FNIR-Werte für Recherchen anhand der Frontalbilder von Personen mit Brille sind nicht signifikant höher als die für Recherchen anhand der Frontalbilder von zufälligen Personen. Für das in dieser Kategorie beste getestete System ist auch der Rang-100-FNIR-Wert für Recherchen anhand der Frontalbilder von Bartträgern (0,4 ± 0,1 %) nicht signifikant höher als der für Recherchen anhand der Frontalbilder von zufälligen Personen.

Mit den verfügbaren Daten konnte für die getesteten Systeme keine Abhängigkeit der FNIR von der seit der Referenzaufnahme verstrichenen Zeit festgestellt werden.

Um den Einfluß der Bildqualität auf die Erkennungsgenauigkeit zu evaluieren, wurde die Qualität von Kopien der 10 000 Frontalbilder von zufälligen Personen mit passendem Gegenstück in der Referenz-

datenbank auf verschiedene Weisen und in unterschiedlichem Grad verschlechtert. Die in dieser Kategorie besten getesteten Systeme zeigen bei Qualitätsminderungen, die ein Spitzen-Signal-Rausch-Verhältnis von mindestens 20 dB ergeben, keine signifikante Erhöhung der Rang-100-FNIR.

Bei Recherchen anhand von Halbprofilbildern erreicht das beste getestete System eine Rang-100-FNIR von 3,0 ± 0,3 %. Recherchen anhand der qualitativ hochwertigen Probebilder, auf denen der Kopf um bis zu 30° um die Hochachse gedreht ist, führten zu ähnlichen Rang-100-FNIR-Werten wie Recherchen anhand der Frontalbilder von zufälligen Personen. Auch Recherchen anhand der qualitativ hochwertigen Probebilder, auf denen der Kopf um bis zu 20° um die Querachse gesenkt bzw. gehoben ist, führten zu ähnlichen Rang-100-FNIR-Werten wie Recherchen anhand der Frontalbilder von zufälligen Personen.

Um die Schwächen der einzelnen Gesichtsidentifizierungssysteme zu umgehen, wurden die Kandidatenlisten von jeweils zwei Systemen auf einfache Weise (mittels Borda-Wahl) auf Rangebene zu einer Kandidatenliste fusioniert. Die Rang-100-FNIR-Werte der besten Systemkombinationen sind nur etwa halb so groß wie die Rang-100-FNIR-Werte der besten Einzelsysteme.

Die Evaluierungsergebnisse beziehen sich nur auf die Evaluierungsgegenstände in der jeweils getesteten Konfiguration.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CMC Cumulative Match Characteristic
DET Detection Error Tradeoff
FNIR False Negative Identification Rate, Falschnegatividentifizierungsrate
FPIR False Positive Identification Rate, Falschpositividentifizierungsrate
FRVT Face Recognition Vendor Test
FTER Failure-To-Enrol Rate
FTXR Failure-To-Extract Rate
GPU Graphics Processing Unit
JPEG Joint Photographic Experts Group
PSNR Peak-Signal-to-Noise Ratio
SSD Solid-State Drive
SOAP Simple Object Access Protocol

Glossar

95%-Vertrauensbereich Bereich, der mit einer Wahrscheinlichkeit von 95 % den tatsächlichen Wert einer Zufallsgröße überdeckt, [1]

ANMERKUNG Die untere Grenze π_u und die obere Grenze π_o des zweiseitigen Vertrauensbereichs für eine Wahrscheinlichkeit π liegen bei

$\pi_{u,o} = \frac{2 \cdot n \cdot h + z^2_{(1-\frac{\alpha}{2})} \mp z_{(1-\frac{\alpha}{2})} \cdot \sqrt{z^2_{(1-\frac{\alpha}{2})} + 4 \cdot n \cdot h \cdot (1-h)}}{2 \cdot (n + z^2_{(1-\frac{\alpha}{2})})}$, wobei *n* die Stichprobengröße, *h* die empirisch ermittelte relative Häufigkeit und $z_{(1-\frac{\alpha}{2})}$ das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung ist. Bei einer Irrtumswahrscheinlichkeit α von 5 % (also 95 % Vertrauen) gilt $z_{(1-\frac{\alpha}{2})} = 1{,}96$.

Ähnlichkeitsmaß (Similarity Score) numerischer Wert, der die Ähnlichkeit zwischen einer biometrischen Probe und einer biometrischen Referenz ausdrückt und mit steigender Ähnlichkeit zunimmt, [2]

biometrische Probe biometrische Daten, die mit (einer) biometrischen Referenz(en) zu vergleichen sind, [2]

biometrische Referenz gespeicherte biometrische Daten, die jeweils einer Person zugeordnet sind und als Vergleichsgrundlage dienen, [2]

biometrische Verifizierung Bestätigung einer behaupteten Identität durch Vergleich biometrischer Daten mit einer biometrischen Referenz [2]

CMC (Cumulative Match Characteristic) grafische Darstellung der geordneten Paare aus einem Rang k und der zugehörigen Rang-k-Richtigpositividentifizierungsrate [3]

ANMERKUNG 1 Rang-k-Richtigpositividentifizierungsrate = 1 – Rang-k-FNIR

ANMERKUNG 2 Die CMC hängt von der Größe der Referenzdatenbank ab[1].

DET (Detection Error Tradeoff) grafische Darstellung der geordneten Paare aus einer Falschpositivrate und der bei gleicher Schwellwerteinstellung zugehörigen Falschnegativrate [3]

ANMERKUNG Der DET-Graph hängt von der Größe der Referenzdatenbank ab[1].

Enrolment Erzeugung und Speicherung einer biometrischen Referenz in der Referenzdatenbank [2]

FTER (Failure-To-Enrol Rate) Anteil der Enrolment-Versuche, bei denen das biometrische System keine für den Vergleich geeigneten Merkmale aus dem Referenzbild extrahieren kann, [2]

FTXR (Failure-To-Extract Rate) Anteil der Rechercheversuche, bei denen das biometrische System keine für den Vergleich geeigneten Merkmale aus dem Probebild extrahieren kann, [4]

ANMERKUNG Wenn für die Merkmalsextraktion aus den Referenzbildern und für die Merkmalsextraktion aus den Probebildern derselbe Algorithmus mit denselben Parameterwerten verwendet wird und die Referenzbilder und Probebilder von gleicher Qualität sind, ist kein signifikanter Unterschied zwischen FTER- und FTXR-Werten zu erwarten. Dieser Evaluierungsbericht enthält

1 Je größer die Referenzdatenbank, desto höher ist die Wahrscheinlichkeit, dass der Identifikator der passenden Referenz aus den führenden Rängen verdrängt wird.

FTXR-Werte für Bilder von verminderter Qualität und mit starken Abweichungen von der Frontalen. Solche Bilder wurden nur als Probebilder, nicht jedoch als Referenzbilder verwendet.

Kandidat biometrische Referenz, die der biometrischen Probe mehr oder weniger ähnelt, [2]

Kandidatenliste Menge der Referenzidentifikatoren der Kandidaten, die der biometrischen Probe am ähnlichsten sind, [2]

ANMERKUNG Biometrische Identifizierungssysteme können so konfiguriert sein, dass

- die Kandidatenliste nur die Referenzidentifikatoren solcher Kandidaten enthält, für die das Maß der Ähnlichkeit zur biometrischen Probe einen festgelegten Schwellwert übersteigt oder
- die Kandidatenliste eine festgelegte Anzahl von Kandidaten umfasst.

Beim Einsatz in der Kriminalistik wird der Schwellwert meist auf 0 (oder einen sehr niedrigen Wert) gesetzt, so dass die Kandidatenliste immer eine festgelegte Anzahl von Kandidaten umfasst, um Falschnegatividentifizierungen möglichst zu vermeiden.

PSNR (Peak-Signal-to-Noise Ratio, Spitzen-Signal-Rausch-Verhältnis) dekadischer Logarithmus des Verhältnisses des Quadrats des maximal möglichen Wertes der Pixelintensität I_{max} eines Schwarz-Weiß-Bildes I der Größe $m \cdot n$ zur mittleren quadratischen Abweichung eines gestörten Bildes K,

$$\mathrm{PSNR} = 10 \cdot \lg \frac{I_{\max}^2}{\frac{1}{m \cdot n} \sum_{i}^{m-1} \sum_{j}^{n-1} (I_{ij} - K_{ij})^2} \mathrm{dB}$$

ANMERKUNG 1 Der maximal mögliche Wert der Pixelintensität I_{max} ist 255 bei Verwendung von 8 Bit.

ANMERKUNG 2 Für Farbbilder mit drei Farbkanälen wird das arithmetische Mittel der mittleren quadratischen Abweichung aller Farbkanäle verwendet.

Rang Position eines Kandidaten in der nach fallendem Ähnlichkeitsmaß geordneten Kandidatenliste [3]

Rang-k-FNIR (Rang-k-Falschnegatividentifizierungsrate) Anteil der Recherchen anhand von Probebildern von Personen, für die mindestens ein Enrolment-Versuch ausgeführt wurde, bei denen kein zugehöriger Referenzidentifikator unter den ähnlichsten *k* Kandidaten ist, [3]

ANMERKUNG 1 Alle Fehler, die dazu führen, dass für Personen, für die Enrolment-Versuche ausgeführt wurden, kein Referenzidentifikator in der Kandidatenliste der Länge *k* zu finden ist, seien es Enrolment-, Merkmalsextraktions- oder Erkennungsfehler, tragen zur (generalisierten) FNIR bei. Ein Enrolment-Fehler wird so behandelt, als ob das Enrolment abgeschlossen worden wäre, der zugehörige Referenzidentifikator taucht jedoch in keiner Kandidatenliste auf. Ein Merkmalsextraktionsfehler wird so behandelt, als ob die Recherche abgeschlossen worden wäre und dabei kein zugehöriger Referenzidentifikator gefunden wurde.

Rang-k-FPIR (Rang-k-Falschpositividentifizierungsrate) Anteil der Rechechen anhand von Probebildern von nicht enrollten Personen, bei denen das Ähnlichkeitsmaß der ähnlichsten *k* Kandidaten einen Schwellwert übersteigt, [3]

Recherche (biometrische Identifizierung) Suche in einer biometrischen Referenzdatenbank nach Kandidaten, die einer biometrischen Probe ähneln,

1 Einführung

1.1 Anwendungsbereich

Zur erkennungsdienstlichen Behandlung von Personen, die von polizeilichen Ermittlungen betroffen sind, gehört u.a. auch das Aufnehmen von Gesichtsbildern (Frontalansicht, Halbprofil, Profil) und deren Speicherung in einer Polizeidatenbank. Die Referenzbilder werden meist mit hoher Qualität aufgenommen. Für jede erkennungsdienstliche Behandlung wird ein Datensatz in der Referenzdatenbank angelegt. Das Zentralsystem der deutschen Polizeidatenbank (INPOL-Z) enthält ca. 5,5 Millionen Gesichtsbilddatensätze. Nur frontale Gesichtsbilder werden bisher zur automatisierten Gesichtserkennung verwendet.

Die gespeicherten Gesichtsbilder stehen Polizeidienststellen zur Identitätsfeststellung unbekannter Personen zur Verfügung. Die Probebilder können von minderer Qualität und aus verschiedenen Aufnahmewinkeln aufgenommen worden sein. Gesichtsidentifizierungssysteme für den Einsatz in der Kriminalistik liefern Kandidatenlisten mit wählbarer Länge (meist 100 Kandidaten), die von Experten für forensische Gesichtserkennung überprüft werden.

1.2 Evaluierungsziele

Die Ziele waren die Evaluierung von am Markt erhältlichen Gesichtsidentifizierungssystemen für den Einsatz in der Kriminalistik hinsichtlich ihrer Erkennungsleistung und die Untersuchung, ob durch die Kombination mehrerer Gesichtsidentifizierungssysteme die Erkennungsleistung gesteigert werden kann.

1.3 Evaluierungsgegenstände

Nach Einholung von Informationen über Produkte von Herstellern von Gesichtserkennungssystemen, die im europäischen Wirtschaftsraum aktiv sind und deren Produkte erfolgreich an großen Vergleichstests (z.B. FRVT 2018 [4]) teilnahmen, wurden anhand von Ausschlusskriterien und Präferenzen des BKA vier Gesichtsidentifizierungssysteme für die Evaluierung ausgewählt. In diesem Evaluierungsbericht werden die getesteten Gesichtsidentifizierungssysteme als A, B, C und D bezeichnet.

Die ausgewählten Hersteller wurden gebeten, jeweils ihre aktuellsten und genauesten markterhältlichen Gesichtsidentifizierungssysteme für die Evaluierung zur Verfügung zu stellen. Zwei Hersteller stellten ihre Systeme mit grafischer Benutzeroberfläche zur Verfügung (B und D). Zwei Hersteller stellten für die Evaluierung ihre Algorithmen mit Befehlszeilen-Benutzeroberfläche zur Verfügung (A und C).

Die getesteten Systeme waren so konfiguriert, dass die ausgegebenen Kandidatenlisten jeweils 100 Kandidaten umfassten. Für andere Parameter (wie Schwellwerte für die Zuverlässigkeit der Gesichtsdetektion, die zulässige Winkelabweichung von der Frontalansicht und den zulässigen Augenabstand) wurden die Voreinstellungen der Hersteller verwendet. Die Evaluierungsergebnisse beziehen sich nur auf die Evaluierungsgegenstände in der jeweils getesteten Konfiguration. Andere Parametereinstellungen können zu anderen Ergebnissen führen.

Die grafischen Benutzeroberflächen der getesteten Systeme umfassen auch Werkzeuge zur Bildverbesserung und zur manuellen Korrektur der Augenposition, die im Praxiseinsatz helfen, die Ergebnisse zu verbessern. Solche Werkzeuge waren nicht Gegenstand dieser Evaluierung.

1.4 Verwendete Verfahrensanweisung

Die Evaluierungsgegenstände wurden auf der Grundlage einer Verfahrensanweisung [5], die ausgehend von den bewährten Methoden zur Technologieevaluierung biometrischer Systeme [3] ausgearbeitet wurde, evaluiert.

1.5 Inhalt dieses Berichts

Dieser Evaluierungsbericht ist im Weiteren wie folgt gegliedert: Abschnitt 2 beschreibt die Schritte zur Vorbereitung der Evaluierung. Abschnitt 3 beschreibt das Vorgehen bei der Durchführung und Auswertung sowie die Ergebnisse der Evaluierung.

2 Vorbereitung der Evaluierung

2.1 Hardware-Plattform

Um einen fairen Vergleich zu ermöglichen, verwendeten alle Gesichtsidentifizierungssysteme für den Test die gleiche Hardware-Plattform. Die Hardware-Plattform bestand aus einer Workstation pro getestetes Gesichtsidentifizierungssystem, einem Dateiserver für den Testdatenbestand (2 × 1 TB Festplattenkapazität auf SSD-Festplatten, welche kürzere Zugriffszeiten als herkömmliche magnetische Festplatten bieten) und einem Gigabit-Netzwerk-Switch. Die Konfiguration der Workstations wurde anhand der Mindestanforderungen der ausgewählten Gesichtsidentifizierungssysteme bestimmt. Die Workstations waren mit einer Grafikkarte mit GPU und einer SSD-Festplatte ausgestattet:

- Prozessor: Intel Xeon E5-2680 v4 @ 2,40 GHz
- Grafikkarte: NVIDIA GeForce GTX 1070
- Arbeitsspeicher: 64 GB
- Festplatte: SSD-Festplatte, 512 GB

Den Herstellern blieb überlassen, ob die zur Verfügung gestellte GPU genutzt wurde. Die Hersteller der Systeme A, C und D gaben an, dass GPU-Unterstützung die für das Enrolment erforderliche Zeit verkürzen kann, für Recherchen jedoch nicht erforderlich ist. System B nutzt die GPU nicht.

2.2 Installation, Funktionstests und Nachjustierung der getesteten Systeme

Die Hersteller installierten ihre Gesichtsidentifizierungssysteme jeweils auf einem Computersystem mit CentOS Linux als Betriebssystem. Die einzelnen Computersysteme wurden zu einer lokalen Netzwerkinsel verbunden.

Nach der Installation wurden Funktionstests aller Komponenten einschließlich der Auswertewerkzeuge vorgenommen. Nach den Funktionstests erhielten die Hersteller die Möglichkeit zur Nachjustierung. Nach der Nachjustierung durch die Hersteller wurden die zu testenden Systeme in einem verschließbaren Raum in den gesicherten Räumlichkeiten des BKA in der Thaerstraße 11 in Wiesbaden aufgestellt. Nach Beginn der eigentlichen Tests erfolgte kein Update durch die Hersteller der getesteten Systeme.

2.3 Evaluierungsskripte

Für den Test jedes Systems im Stapelbetrieb wurden Skripte erstellt. Für die Systeme A und C mit Befehlszeilen-Benutzeroberfläche stellten die Hersteller Evaluierungsskripte zur Verfügung, die versuchen, aus jedem Gesichtsbild im Referenzbildverzeichnis und jedem Gesichtsbild im Probebildverzeichnis Merkmale zu extrahieren, und die anhand der Probemerkmalsvektoren Recherchen gegen die Referenzmerkmalsvektoren ausführen. Für die Systeme B und D mit grafischer Benutzeroberfläche erstellte Fraunhofer IGD mit Hilfe der Hersteller Rechercheskripte, die auf die SOAP-Schnittstellen der Systeme zugriffen.

Die Skripte protokollierten für erfolgreiche Enrolment-Versuch die Bildnummer und die benötigte Zeit. Wenn ein Enrolment-Versuch fehlschlug, wurde dies sowie die Fehlerursache protokolliert. Bei jeder Recherche wurde die nach fallendem Ähnlichkeitsmaß geordnete Kandidatenliste und die benötigte Zeit aufgezeichnet. Wenn aus dem Probebild keine geeigneten Merkmale extrahiert werden konnten, wurde dies sowie die Fehlerursache protokolliert.

2.4 Bereitstellung des Datenbestands

Das BKA stellte für die Evaluierung die folgenden Datenbasen zur Verfügung:

- Kopien von ca. 5 Millionen digitalen Bildern, die in INPOL-Z als frontale Gesichtsbilder von ca. 3 Millionen Personen markiert sind,
- Kopien von ca. 3 Millionen digitalen Bildern, die in INPOL-Z als Halbprofilbilder markiert sind (d.h. Kopf um 45° um die Hochachse gedreht),
- von 147 freiwilligen Testpersonen mindestens zwei digitale frontale Gesichtsbilder, die unter idealen Bedingungen über einen Zeitraum von etwas mehr als neun Jahren aufgenommen wurden (insgesamt 747 Gesichtsbilder),
- digitale Gesichtsbilder von 181 freiwilligen Testpersonen aus bis zu 21 verschiedenen Aufnahmewinkeln:

  - jeweils ein Bild, bei dem der Kopf um 10°, 20°, 30°, 45°, 60°, 70°, 80° bzw. 90° in eine Richtung nur um die Hochachse gedreht ist (»Yaw Angle«),
  - ein Bild, bei dem der Kopf um -45°, -30°, -20°, -10°, 10°, 20°, 30° bzw. 45° nur um die Querachse gesenkt bzw. gehoben ist (»Pitch Angle«), wenn die Testperson dazu in der Lage war,
  - ein Bild, bei dem der Kopf um 10°, 20°, 30° bzw. 45° in eine Richtung nur um die Längsachse geneigt ist (»Roll Angle«), wenn die Testperson dazu in der Lage war, sowie
  - ein frontales Gesichtsbild (alle Winkel gleich 0°).

  Von 76 der 181 Testpersonen wurden zwei Serien von Gesichtsbildern aus verschiedenen Aufnahmewinkeln zur Verfügung gestellt: eine ohne und eine mit Brille (insgesamt 5358 Gesichtsbilder).

Für jede Datenbasis stellte das BKA ein Verzeichnis ohne Unterverzeichnisse mit allen Gesichtsbildern aller Testpersonen bereit. Jedes Bild war mit einer eindeutigen Teilnehmernummer und einer eindeutigen Bildnummer benannt. Damit die Gesichtsbilder nicht ohne Hinzuziehung zusätzlicher Informationen einer spezifischen Person zugeordnet werden konnten, wurden als Teilnehmernummern nicht die in INPOL-Z verwendeten P-Nummern und als Bildnummern nicht die in INPOL-Z verwendeten E-Nummern verwendet, sondern Pseudonyme.

Alle Gesichtsbilder lagen im JPEG-Format vor. Die durchschnittliche Dateigröße der Gesichtsbilder aus INPOL-Z betrug ca. 70 kByte.

Zu den frontalen Gesichtsbildern aus INPOL-Z stellte das BKA eine Liste von ca. 56 500 Bartträgern und ca. 19 500 Brillenträgern zur Verfügung, für die INPOL-Z jeweils mindestens zwei frontale Gesichtsbilder enthält. Ca. 6500 Personen auf der Liste waren sowohl Brillen- als auch Bartträger. Informationen zum Geschlecht der aufgenommenen Personen (männlich / weiblich) und zum Aufnahmedatum aller Gesichtsbilder aus INPOL-Z lagen für die Evaluierung nicht vor. Darum erfolgte keine Auswertung nach Geschlecht und zeitlichem Abstand der Aufnahmen aus INPOL-Z.

Für die Gesichtsbilder aus den eigens über mehrere Jahre aufgenommenen Serien war das Aufnahmedatum bekannt.

3 Durchführung und Ergebnisse der Evaluierung

3.1 Bereinigung des Datenbestands

Vor der Partitionierung des Datenbestands wurde mit Hilfe herstellerneutraler Algorithmen zur Gesichtsdetektion überprüft, ob alle in INPOL-Z als frontale Gesichtsbilder markierten digitalen Bilder tatsächlich Frontalansichten enthielten. Aus dem Datenbestand wurden 1621 Bilder entfernt, die zwar als frontale Gesichtsbilder markiert waren, aber keine Frontalansichten zeigten, sondern

- Halbprofil- oder Profilansichten,
- Tattoos an verschiedenen Körperteilen,
- Ganzkörperansichten von hinten oder von der Seite oder
- die Information, dass kein Gesichtsbild verfügbar ist.

Die Parameter für die Gesichtsdetektion waren so eingestellt, dass kein Bild, das tatsächlich ein Gesicht von vorn zeigte, aus dem Datenbestand entfernt wurde. Es konnte jedoch nicht ausgeschlossen werden, dass der Datenbestand weitere, unzutreffend markierte Bilder enthielt.

Von 42 Testpersonen lagen sowohl über mehrere Jahre aufgenommene Serien frontaler Gesichtsbilder als auch Gesichtsbilder aus verschiedenen Aufnahmewinkeln vor. Allen Aufnahmen dieser Testpersonen wurde die gleiche Teilnehmernummer zugeordnet.

3.2 Partitionierung des Datenbestands

3.2.1 Erforderliche Anzahl an Recherchen

Je kleiner die nachzuweisende Fehlerrate, desto größer ist die erforderliche Anzahl an Recherchen. FRVT 2018 [4] hat gezeigt, dass die Rang-1-FNIR bei der Suche nach frontalen Gesichtsbildern überwiegend guter Qualität in einer Referenzdatenbank mit 1,6 Millionen Gesichtsbildern überwiegend guter Qualität auf Werte von 0,3 % heruntergehen kann. Nach der »Rule of 30«[2] sind 10 000 Recherchen erforderlich, um mit 90%iger Sicherheit eine FNIR von ungefähr 0,3 % (zwischen 0,21 % und 0,39 %) nachzuweisen.

3.2.2 Frontalbilder aus INPOL-Z

Die frontalen Gesichtsbilder aus INPOL-Z wurden wie folgt in Referenzbilder und Probebilder partitioniert:

- Probebilder mit passenden Gegenstücken in der Referenzdatenbank waren jeweils ein Frontalbild von
  - 10 000 zufällig ausgewählten Personen (darunter können gemäß ihrem Anteil an INPOL-Z auch Brillen- und Bartträger sein),
  - 10 000 zufällig ausgewählten Brillenträgern sowie
  - 10 000 zufällig ausgewählten Bartträgern,

  für die jeweils mindestens ein weiteres frontales Gesichtsbild vorlag.

2 Rule of 30: Wenn mindestens M ≥ 30 Fehler beobachtet werden, dann liegt die tatsächliche Fehlerrate mit 90%iger Sicherheit innerhalb von ±30 % der beobachteten Fehlerrate M/N [3].

- Probebilder ohne passendes Gegenstück in der Referenzdatenbank waren die frontalen Gesichtsbilder von 10 000 zufällig ausgewählten Personen, für die nur ein einziges frontales Gesichtsbild vorlag.
- Referenzbilder waren alle übrigen frontalen Gesichtsbilder aus INPOL-Z, insgesamt 4 784 738 Bilder.

Nicht auf allen Aufnahmen trugen die als Brillen- bzw. Bartträger markierten Personen Brille bzw. Bart. Mit Hilfe herstellerneutraler Algorithmen zur Detektion von Brillen wurde sichergestellt, dass zumindest auf allen als Brillenträger-Probebild gewählten Bildern eine Brille zu sehen war.

3.2.3 Über mehrere Jahre aufgenommene Serien frontaler Gesichtsbilder

Die über mehrere Jahre aufgenommenen Serien frontaler Gesichtsbilder wurden wie folgt in Referenzbilder und Probebilder partitioniert:

- Referenzbild war jeweils das zuerst aufgenommene Gesichtsbild aus einer Serie, insgesamt 147 Bilder.
- Probebilder waren alle übrigen Gesichtsbilder aus diesen Serien, insgesamt 600 Bilder.

3.2.4 Halbprofilbilder aus INPOL-Z

Aus den Halbprofilbildern aus INPOL-Z wurden wie folgt Probebilder ausgewählt:

- Probebilder mit passenden Gegenstücken in der Referenzdatenbank waren jeweils ein Halbprofilbild von 10 000 zufällig ausgewählten Personen, für die mindestens ein frontales Gesichtsbild als Referenzbild ausgewählt wurde.
- Probebilder ohne passendes Gegenstück in der Referenzdatenbank waren die Halbprofilbilder von 10 000 zufällig ausgewählten Personen, für die kein frontales Gesichtsbild als Referenz gewählt wurde.

Halbprofilbilder wurden nicht als Referenzbilder verwendet.

3.2.5 Gesichtsbilder aus verschiedenen Aufnahmewinkeln

Die Gesichtsbilder aus verschiedenen Aufnahmewinkeln wurden wie folgt in Referenzbilder und Probebilder partitioniert:

- Referenzbilder waren alle frontalen Aufnahmen, insgesamt 257 Bilder.
- Probebilder waren alle Aufnahmen, die nicht als Referenzbilder verwendet wurden, pro Aufnahmewinkel bis zu 257 Bilder.

3.3 Enrolment frontaler Gesichtsbilder

Als erster Leistungstest wurde versucht, die als Referenzbilder gewählten 4 785 142 frontalen Gesichtsbilder (siehe Abschnitt 3.2) im Stapelbetrieb in die Referenzdatenbank jedes getesteten Systems zu überführen. Alle Referenzbilder, seien sie aus INPOL-Z, den über mehrere Jahre aufgenommenen Serien frontaler Gesichtsbilder oder den Serien von Gesichtsbildern aus verschiedenen Aufnahmewinkeln ausgewählt, wurden jeweils in die gleiche Referenzdatenbank überführt. Die Referenzdatenbanken blieben für alle Recherchen unverändert. Wenn vorhanden, wurden mehrere Gesichtsbilder pro Testperson getrennt voneinander enrollt, wie bei Gesichtsidentifizierungssystemen für den Einsatz in der Kriminalistik üblich. Die Referenzdatenbanken können also mehrere, nicht miteinander verknüpfte Gesichtsbilder der gleichen Testperson enthalten.

Aus den Protokolldaten wurden für jedes getestete System die FTER und ihr 95%-Vertrauensbereich für die frontalen Gesichtsbilder aus INPOL-Z ermittelt, siehe Tabelle 1. Da alle Aufnahmen in den über mehrere Jahre aufgenommenen Serien frontaler Gesichtsbilder sowie die frontalen Aufnahmen in den Serien von Gesichtsbildern aus verschiedenen Aufnahmewinkeln unter idealen Bedingungen aufgenommen wurden, traten für die Referenzbilder aus diesen Serien keine Enrolment-Fehler (Failure to Enrol) auf.

Tabelle 1: FTER für Frontalbilder aus INPOL-Z

System	FTER
A	0,013 % [0,012 %; 0,014 %]
B	0,066 % [0,063 %; 0,068 %]
C	0,056 % [0,054 %; 0,058 %]
D	0,156 % [0,152 %; 0,159 %]

Jedes getestete System nutzte die vom Hersteller voreingestellten Schwellwerte für das Enrolment. Funktionen zur manuellen Korrektur der maschinell bestimmten Augenposition, um das Enrolment doch abzuschließen, wurden nicht benutzt. Als Ursache von Enrolment-Fehlern wurde z.B. „failed to load image file“, „no face found“ und „image size lower than minimum allowed (96x96)“ protokolliert.

Insgesamt 324 der als Referenzbilder gewählten Bilder konnten in keinem der getesteten Systeme enrollt werden, 23 davon, weil die Datei sich nicht öffnen ließ. Die anderen Bilder, die nirgends enrollt werden konnten, zeigten

- vermummtes oder voll verschleiertes Gesicht,
- zu geringe Bildauflösung,
- zu wenig Kontrast,
- tief gesenkter Kopf,
- Kopf von hinten,
- andere Körperteile als der Kopf,
- Tattoos an verschiedenen Körperteilen,
- Ganzkörperansicht von hinten,
- Schuhsohlen oder
- die Information, dass kein Gesichtsbild verfügbar ist.

Tabelle 2 zeigt für jedes getestete System die mittlere Dauer der Enrolment-Versuche in Millisekunden. Die Ergebnisse lassen die Größenordnung der Dauer erfolgreicher Enrolment-Versuche erkennen, sind jedoch nicht miteinander vergleichbar, weil das Enrolment in den getesteten Systemen unterschiedliche Schritte umfasste:

- Bei System A (das für die Evaluierung mit Befehlszeilen-Benutzeroberfläche zur Verfügung gestellt wurde) umfasste das Enrolment nur die Gesichtsdetektion und die Merkmalsextraktion aus dem Referenzbild und das Schreiben des Referenzmerkmalsvektors in eine Datei.
- Bei System C (das für die Evaluierung ebenfalls mit Befehlszeilen-Benutzeroberfläche zur Verfügung gestellt wurde) kam zur Gesichtsdetektion, Merkmalsextraktion und dem Schreiben des Referenzmerkmalsvektors noch die Erzeugung einer Miniaturansicht des Referenzbildes hinzu.
- Bei den Systemen B und D (die mit grafischer Benutzeroberfläche zur Verfügung gestellt wurden) umfasste das Enrolment auch das Speichern des Referenzmerkmalsvektors und einer Miniaturansicht in einer relationalen Datenbank.

System D war so konfiguriert, dass nach dem Enrolment aller Referenzbilder ein Datenbankindex erzeugt wurde. Die dafür benötigte Zeit ist nicht in Tabelle 2 berücksichtigt. Beim Enrolment kleinerer Datenmengen kann das System auch so konfiguriert werden, dass der Datenbankindex während der Verarbeitung der Referenzbilder erzeugt wird.

Tabelle 2: Dauer von Enrolment-Versuchen

System	Mittelwert
A	0,4 ms
B	64,4 ms
C	5,7 ms
D	79,1 ms

3.4 Recherchen anhand frontaler Gesichtsbilder

3.4.1 Beliebige Frontalbilder aus INPOL-Z

Die 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder mit passenden Gegenstücken in der Referenzdatenbank und die 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder ohne passendes Gegenstück in der Referenzdatenbank (siehe Abschnitt 3.2.2) wurden im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Aus den Protokolldaten wurden für jedes getestete System die FTXR für frontale Gesichtsbilder aus INPOL-Z und ihr 95%-Vertrauensbereich ermittelt, siehe Tabelle 3.

Tabelle 3: FTXR für beliebige Frontalbilder aus INPOL-Z

System	FTXR
A	0,01 % [0,00 %; 0,03 %]
B	0,13 % [0,08 %; 0,18 %]
C	0,02 % [0,01 %; 0,04 %]
D	0,34 % [0,27 %; 0,43 %]

Aus den Protokolldaten für die Recherchen anhand der 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder mit passenden Gegenstücken in der Referenzdatenbank wurden für jedes getestete System die CMC sowie die Rang-k-FNIR über dem Rang ermittelt, siehe Abbildungen 1 und 2. Die Referenzdatenbank enthielt jeweils (1 – FTER)·4 785 142 (also ca. 4,8 Millionen) Referenzen. Enrolment-Fehler sind in der Rang-k-FNIR mit berücksichtigt: Bei Recherechen anhand von Probebildern von Personen, die enrollt sein sollten, tauchen die Referenzidentifikatoren der Bilder mit Enrolment-Fehler nicht in der Kandidatenliste auf. Wenn das Enrolment für kein Bild dieser Person gelang (es konnten mehrere Bilder der gleichen Person enrollt sein), kann die Person nicht identifiziert werden. Ein hoher FTER-Wert führt also nicht zu einer unfairen Verringerung der Rang-k-FNIR. Auch Merkmalsextraktionsfehler sind in der Rang-k-FNIR mit berücksichtigt: Bei Recherechen anhand von Probebildern mit Merkmalsextraktionsfehler kann die Person nicht identifiziert werden. Ein hoher FTXR-Wert führt also nicht zu einer unfairen Verringerung der Rang-k-FNIR. Eine Recherche zählt als erfolgreich, wenn die Kandidatenliste der Länge *k* mindestens einen zugehörigen Referenzidentifikator enthält.

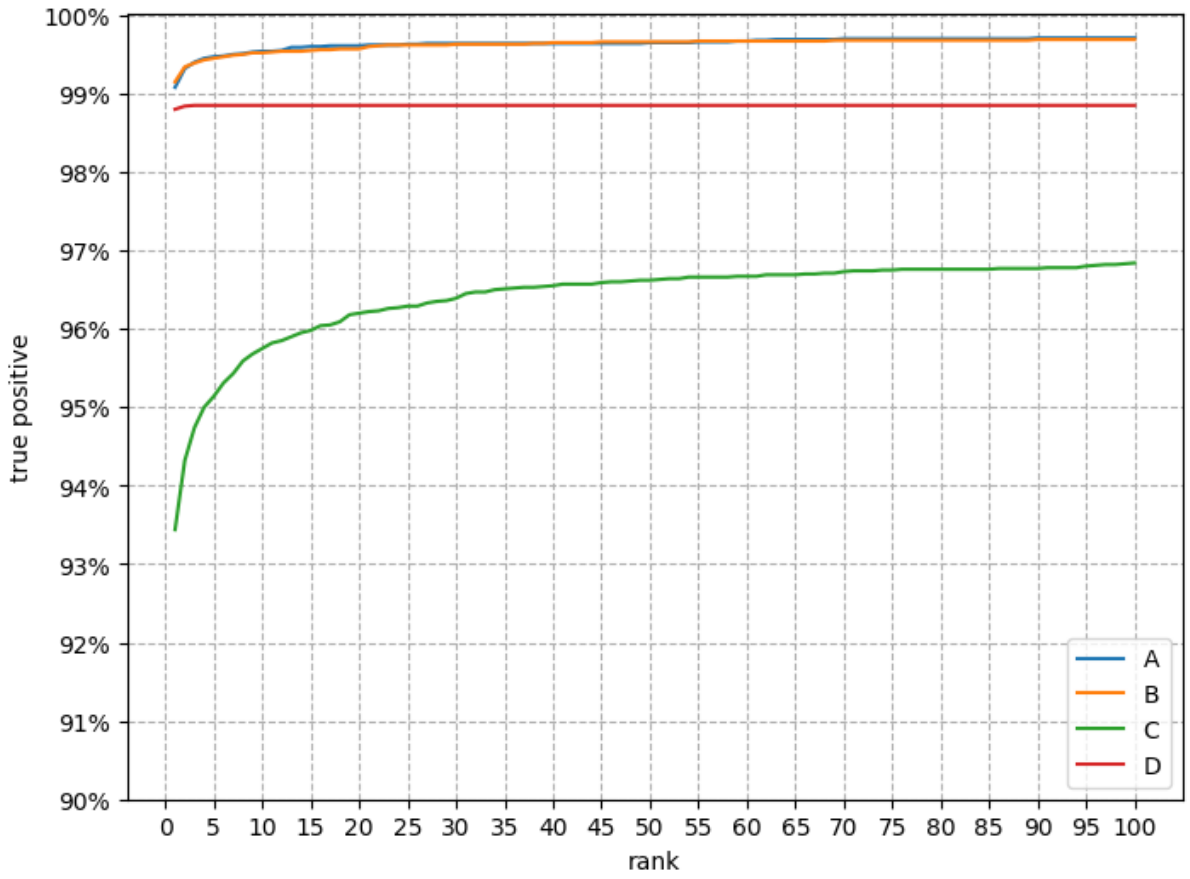

Abbildung 1: CMC für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

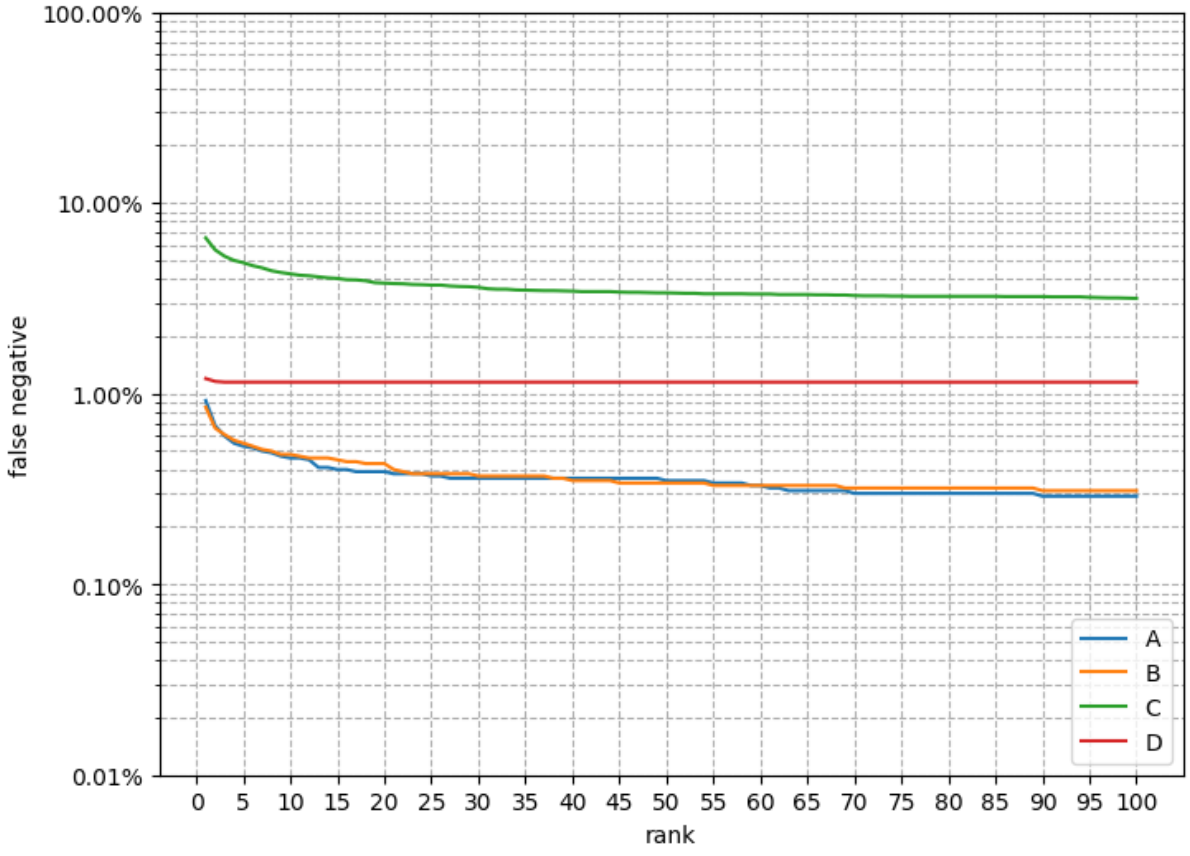

Abbildung 2: Rang-k-FNIR über dem Rang für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

Der Rang-k-FNIR-Graph über dem Rang von System D ist dank des nach dem Enrolment erstellten Datenbankindex besonders flach: Wenn System D einen zur Probe zugehörigen Referenzidentifikator findet, dann befindet sich dieser unter den ähnlichsten 5 Kandidaten.

Tabelle 4 fasst für jedes getestete System die sich bei Recherchen anhand von beliebigen frontalen Gesichtsbildern aus INPOL-Z ergebenden Rang-1- und Rang-100-FNIR-Werte (die auch aus Abbildung 2 ersichtlich sind) und deren 95%-Vertrauensbereiche zusammen. Die Überlappung der Vertrauensbereiche zeigt, dass der Unterschied zwischen den Rang-100-FNIR-Werten der Systeme A und B statistisch nicht signifikant ist.

Tabelle 4: Rang-1- und Rang-100-FNIR für beliebige Frontalbilder aus INPOL-Z

System	Rang-1-FNIR	Rang-100-FNIR
A	0,92 % [0,75 %; 1,13 %]	0,29 % [0,20 %; 0,42 %]
B	0,85 % [0,69 %; 1,05 %]	0,31 % [0,22 %; 0,44 %]
C	6,56 % [6,09 %; 7,06 %]	3,16 % [2,83 %; 3,52 %]
D	1,20 % [1,00 %; 1,43 %]	1,15 % [0,96 %; 1,38 %]

Wenn man von Merkmalsexraktionsfehlern (nach denen gar keine Kandidatenliste erstellt wurde) absieht und nur solche Fälle berücksichtigt, bei denen eine Kandidatenliste erstellt wurde, die aber keinen zugehörigen Referenzidentifikator enthielt, sind die Werte der Falschnegativraten um den FTXR-Wert aus Tabelle 3 kleiner als die Rang-k-FNIR-Werte aus Abbildung 2, siehe Abbildung 3.

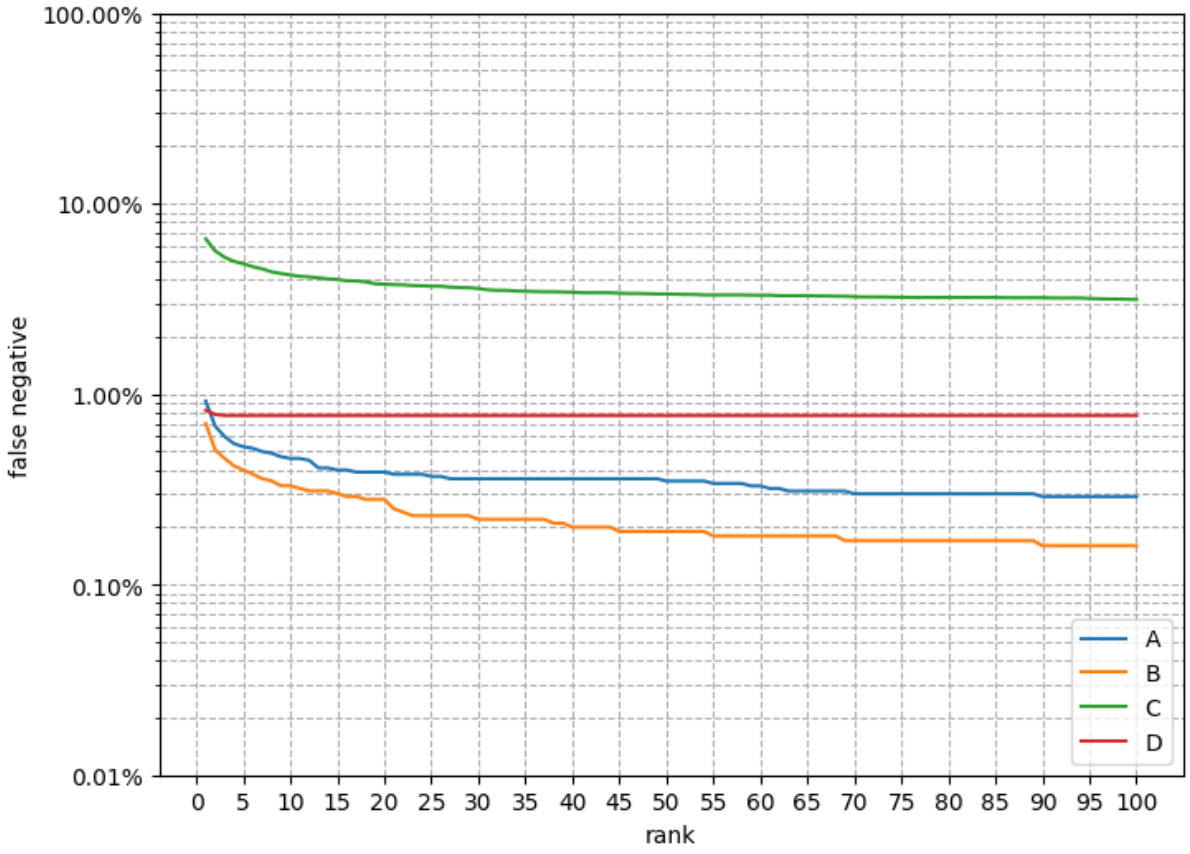

Abbildung 3: Rang-k-FNIR – FTXR über dem Rang für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

Um einen Überblick über die mit frontalen Gesichtsbildern erreichbaren Werte der Fehlerraten zu erlangen, wurde für jedes getestete System ein Rang-1-DET-Graph[3] für frontale Gesichtsbilder berechnet, siehe Abbildung 4. Rang-1-DET bedeutet, dass die Kandidatenliste nur einen Kandidaten enthält, sofern die Ähnlichkeit zwischen der Probe und der ähnlichsten Referenz den Schwellwert übersteigt. Ein DET-Graph ist nützlich, wenn das Gesichtsidentifizierungssystem so konfiguriert werden muss, dass nur eine begrenzte Anzahl von Kandidaten, für die das Ähnlichkeitsmaß einen Schwellwert übersteigt, in die Kandidatenliste eingehen, z.B. für den Einsatz in der Videoüberwachung oder Fotofahndung. Wenn das Gesichtsidentifizierungssystem z.B. so konfiguriert werden muss, dass die Rang-1-FPIR 0,1 % beträgt, würde System D die geringste Rang-1-FNIR liefern.

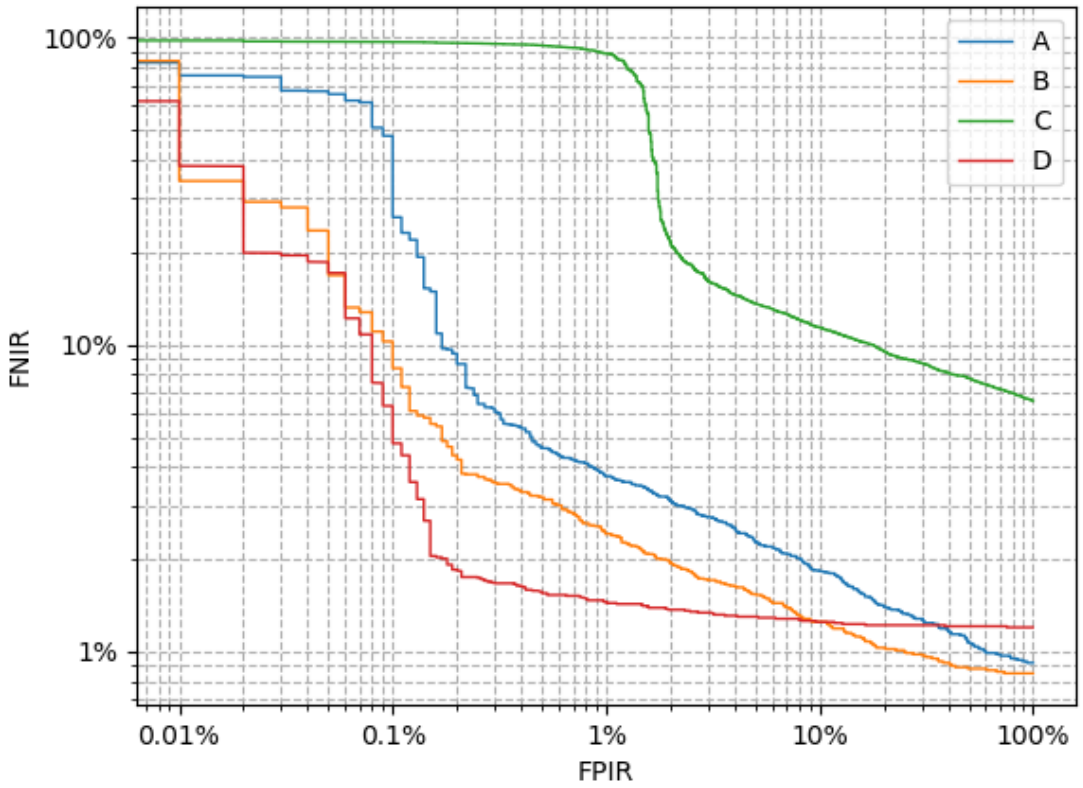

Abbildung 4: Rang-1-DET-Graph für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

Aus den Protokolldaten für die 20 000 Recherchen wurde für die getesteten Systeme jeweils die mittlere Dauer der ausgeführten Recherchen in Millisekunden ermittelt, siehe Tabelle 5. Gemessen wurde die Zeit von Anfang bis Ende der Recherche einschließlich der Zeit für die Merkmalsextraktion aus dem Probebild. Bei System A (das für die Evaluierung mit Befehlszeilen-Benutzeroberfläche zur Verfügung gestellt wurde) wurden jeweils die Dauer der Merkmalsextraktion aus dem Probebild und die Dauer der Recherche anhand des Probemerkmalsvektors, die in unterschiedlichen Dateien protokolliert wurden, addiert. Bei System C (das ebenfalls mit Befehlszeilen-Benutzeroberfläche zur Verfügung gestellt wurde) fehlt jeweils die Dauer der Recherche anhand des Probemerkmalsvektors. Dank des von System D erzeugten Datenbankindex (siehe Abschnitt 3.3) gingen die eigentlichen Recherchen in System D deutlich schneller als in den anderen getesteten Systemen.

3 Für den Einsatz in der Kriminalistik sind die DET-Graphen nur von geringem Interesse, da der Schwellwert auf 0 (oder einen sehr niedrigen Wert) gesetzt wird, um Falschnegatividentifizierungen zu vermeiden.

Tabelle 5: Dauer von Recherchen anhand von Frontalbildern

System	Mittelwert
A	668,3 ms
B	653,8 ms
C	–[4]
D	241,9 ms

3.4.2 Frontalbilder aus INPOL-Z von Brillenträgern

Die 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder von Brillenträgern (siehe Abschnitt 3.2.2) wurden im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Aus den Protokolldaten für diese 10 000 Recherchen wurden für jedes getestete System die FTXR und ihr 95%-Vertrauensbereich ermittelt, siehe Tabelle 6.

Tabelle 6: FTXR für Frontalbilder aus INPOL-Z von Brillenträgern

System	FTXR
A	0,04 % [0,02 %; 0,10 %]
B	0,20 % [0,13 %; 0,31 %]
C	0,11 % [0,06 %; 0,20 %]
D	0,55 % [0,42 %; 0,72 %]

Aus den Protokolldaten wurden für jedes getestete System die CMC sowie die Rang-k-FNIR über dem Rang für frontale Gesichtsbilder aus INPOL-Z von Brillenträgern ermittelt, siehe Abbildungen 5 und 6.

4 System C hat nur die Dauer der Merkmalsextraktion aus jedem Probebild protokolliert, nicht jedoch die Gesamtdauer der Recherchen gegen die Referenzdatenbank. Die durchschnittliche Dauer der Merkmalsextraktion aus einem der 20 000 zufällig aus INPOL-Z ausgewählten frontalen Gesichtsbildern betrug 6,1 ms, war also ähnlich der mittleren Dauer von Enrolment-Versuchen im System C (siehe Tabelle 2).

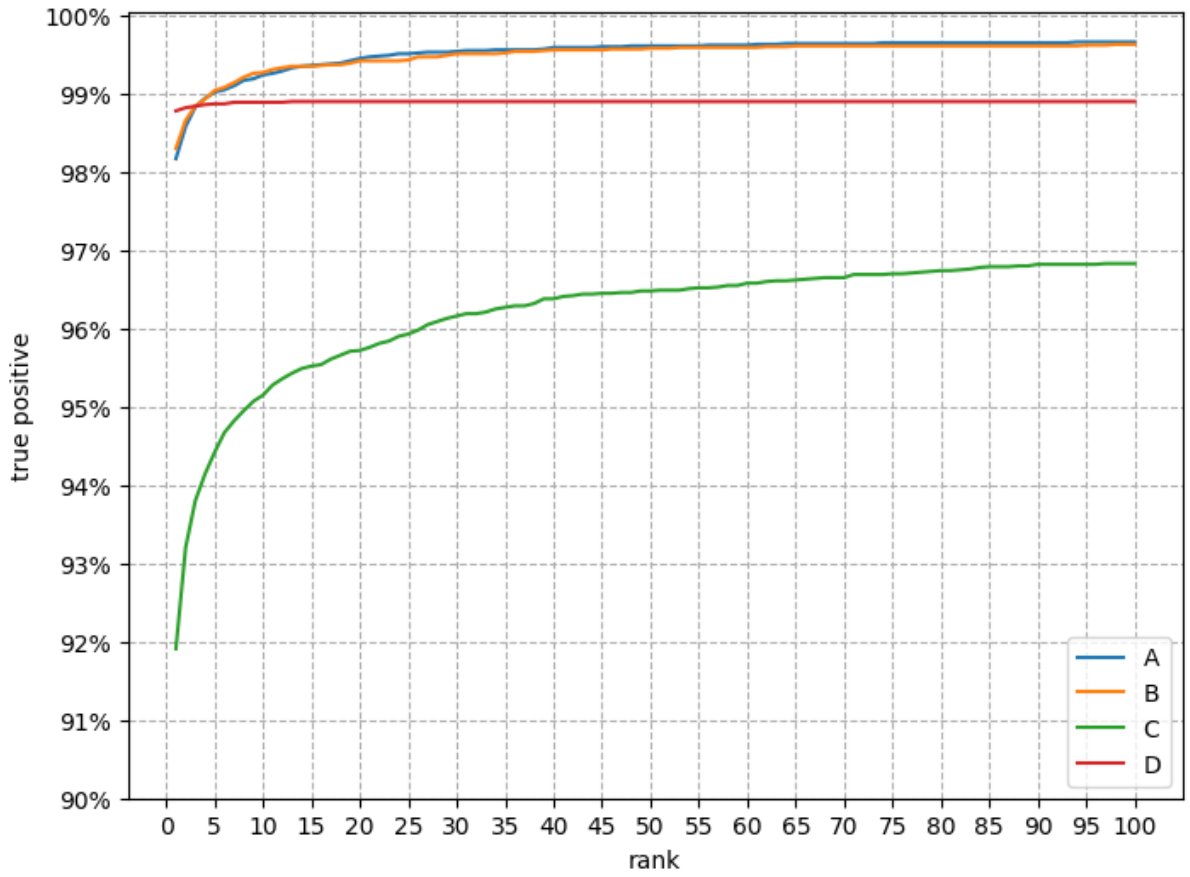

Abbildung 5: CMC für Frontalbilder von Brillenträgern bei ca. $4{,}8 \cdot 10^6$ Referenzen

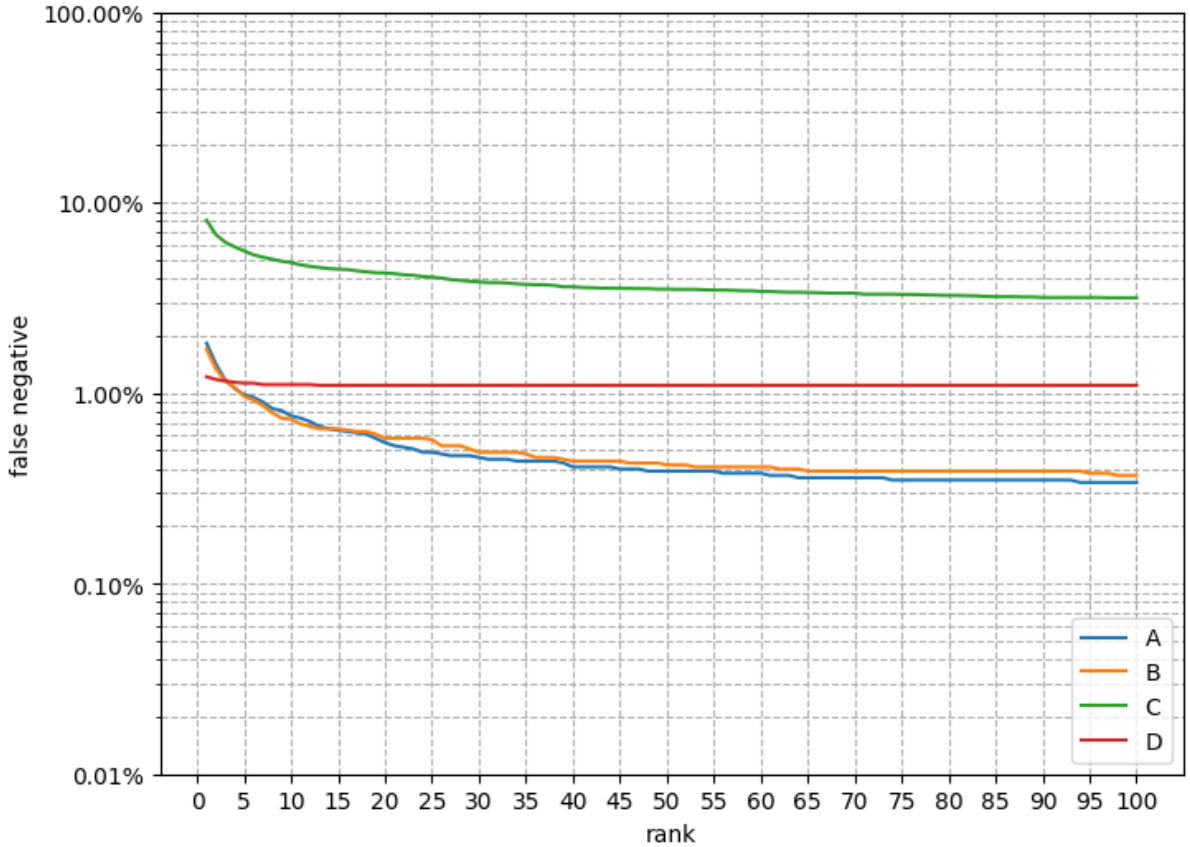

Abbildung 6: Rang-k-FNIR über dem Rang für Frontalbilder von Brillenträgern bei ca. $4{,}8 \cdot 10^6$ Referenzen

Tabelle 7 fasst für jedes getestete System die sich bei Recherchen anhand von frontalen Gesichtsbildern mit Brille ergebenden Rang-1- und Rang-100-FNIR-Werte (die auch aus Abbildung 6 ersichtlich sind) und deren 95%-Vertrauensbereich zusammen. System D ist das einzige getestete System, bei dem sich der Rang-1-FNIR-Wert für Frontalbilder mit Brille nicht signifikant von dem für beliebige Frontalbilder unterscheidet (vgl. Tabelle 4). Die anderen getesteten Systeme finden einen zugehörigen Referenzidentifikator für Brillenträger auf hinteren Rängen. Für jedes getestete System unterscheidet sich der Rang-100-FNIR-Wert für Frontalbilder mit Brille nicht signifikant von dem für beliebige Frontalbilder (vgl. Tabelle 4), wie die Überlappung der Vertrauensbereiche zeigt.

Tabelle 7: Rang-1- und Rang-100-FNIR für Frontalbilder aus INPOL-Z von Brillenträgern

System	Rang-1-FNIR	Rang-100-FNIR
A	1,83 % [1,59 %; 2,11 %]	0,34 % [0,24 %; 0,47 %]
B	1,70 % [1,46 %; 1,97 %]	0,37 % [0,27 %; 0,51 %]
C	8,09 % [7,57 %; 8,64 %]	3,17 % [2,84 %; 3,53 %]
D	1,22 % [1,02 %; 1,45 %]	1,10 % [0,91 %; 1,32 %]

3.4.3 Frontalbilder aus INPOL-Z von Bartträgern

Die 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder von Bartträgern (siehe Abschnitt 3.2.2) wurden im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Aus den Protokolldaten für diese 10 000 Recherchen wurden für jedes getestete System die FTXR und ihr 95%-Vertrauensbereich ermittelt, siehe Tabelle 8.

Tabelle 8: FTXR für Frontalbilder aus INPOL-Z von Bartträgern

System	FTXR
A	0,01 %[0,00 %, 0,06 %]
B	0,09 % [0,05 %, 0,17 %]
C	0,04 % [0,02 %, 0,10 %]
D	0,32 % [0,23 %, 0,45 %]

Aus den Protokolldaten wurden für jedes getestete System die CMC sowie die Rang-k-FNIR über dem Rang für frontale Gesichtsbilder aus INPOL-Z von Bartträgern ermittelt, siehe Abbildungen 7 und 8.

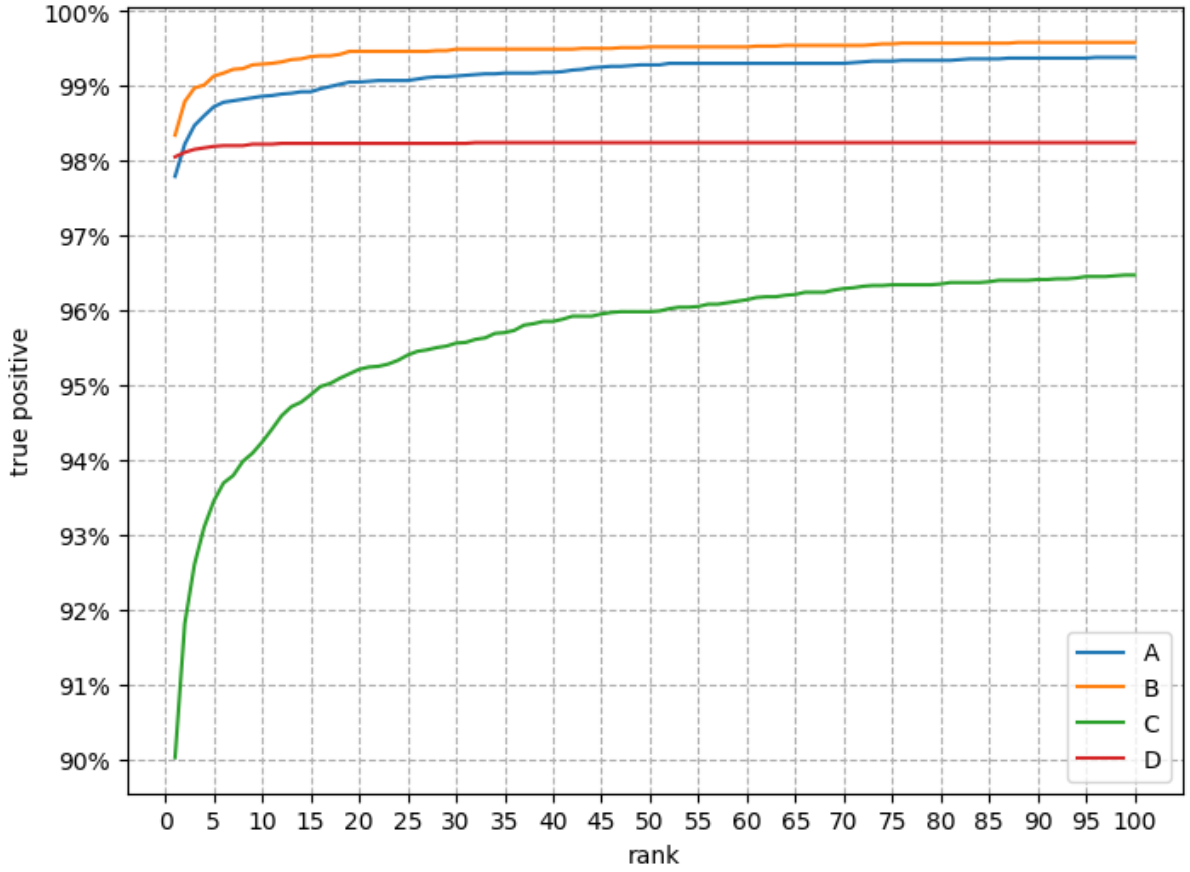

Abbildung 7: CMC für Frontalbilder von Bartträgern bei ca. $4{,}8 \cdot 10^6$ Referenzen

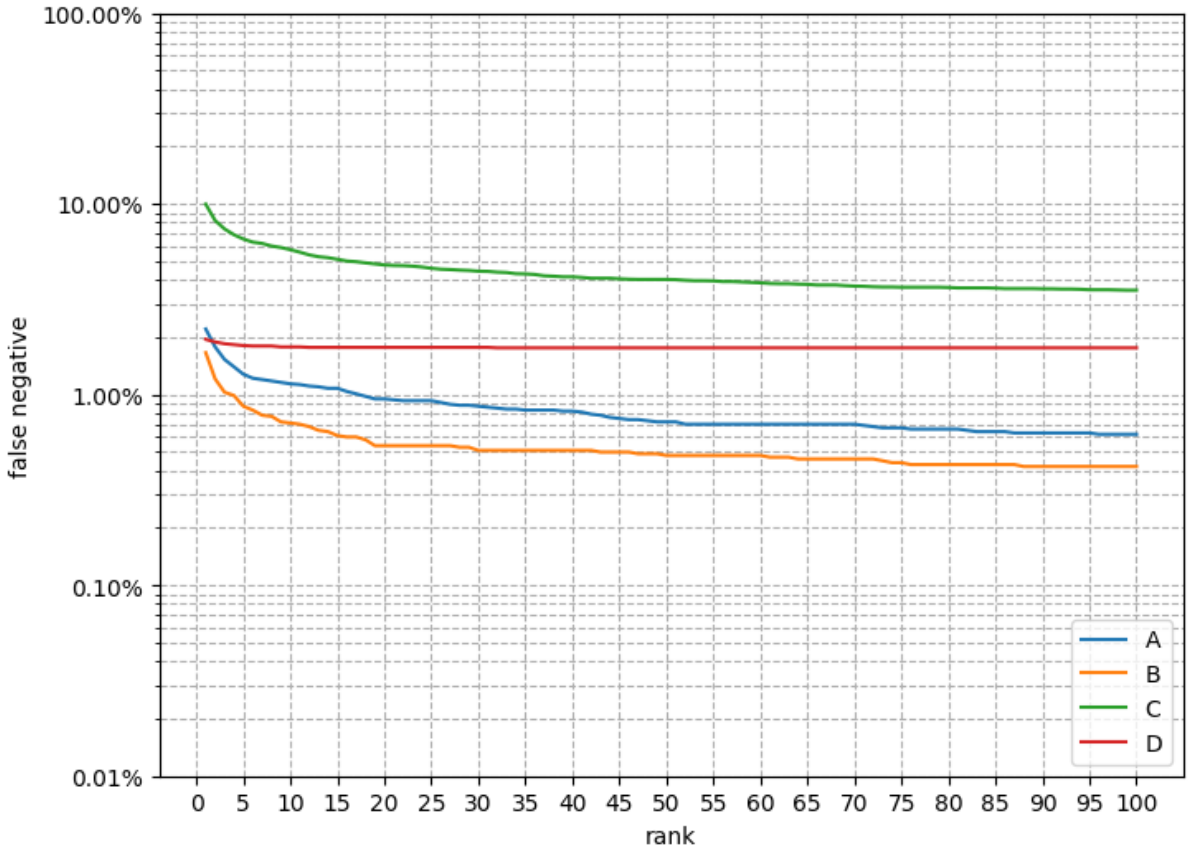

Abbildung 8: Rang-k-FNIR über dem Rang für Frontalbilder von Bartträgern bei ca. $4{,}8 \cdot 10^6$ Referenzen

Tabelle 9 fasst für jedes getestete System die sich bei Recherchen anhand von frontalen Gesichtsbildern von Bartträgern ergebenden Rang-1- und Rang-100-FNIR-Werte (die auch aus Abbildung 8 ersichtlich sind) und deren 95%-Vertrauensbereich zusammen. System B ist das beste getestete System, bei dem sich der Rang-100-FNIR-Wert für Frontalbilder von Bartträgern nicht signifikant von dem für beliebige Frontalbilder unterscheidet (vgl. Tabelle 4). Für die anderen getesteten Systeme sind sowohl die Rang-1- als auch die Rang-100-FNIR-Werte für Frontalbilder von Bartträgern signifikant höher als die für beliebige Frontalbilder (vgl. Tabelle 4).

Tabelle 9: Rang-1- und Rang-100-FNIR für Frontalbilder aus INPOL-Z von Bartträgern

System	Rang-1-FNIR	Rang-100-FNIR
A	2,21 % [1,94 %; 2,52 %]	0,62 % [0,48 %; 0,79 %]
B	1,66 % [1,43 %; 1,93 %]	0,42 % [0,31 %; 0,57 %]
C	9,98 % [9,41 %; 10,58 %]	3,53 % [3,19 %; 3,91 %]
D	1,95 % [1,70 %; 2,24 %]	1,76 % [1,52 %; 2,04 %]

3.4.4 Über mehrere Jahre aufgenommene Serien frontaler Gesichtsbilder

Alle 600 Probebilder aus den über mehrere Jahre aufgenommenen Serien frontaler Gesichtsbilder wurden im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Da alle Aufnahmen in diesen Serien unter idealen Bedingungen aufgenommen wurden, traten für diese Probebilder keine Extraktionsfehler (Failure to Extract) auf.

Wegen der geringen Anzahl dieser Probebilder werden nur zwei Klassen unterschieden, die ungefähr gleich groß sind:

- Probebilder, die innerhalb von weniger als 3 Jahren und 4 Monaten nach der Referenzaufnahme aufgenommen wurden, insgesamt 307 Gesichtsbilder, und
- Probebilder, die 3 Jahre und 4 Monate bis 9 Jahre und 3 Monate nach der Referenzaufnahme aufgenommen wurden, insgesamt 293 Gesichtsbilder.

Unabhängig vom zeitlichen Abstand zwischen der Aufnahme der Referenz- und Probebilder geht die Rang-100-FNIR für alle getesteten Systeme gegen 0. Darum wurde die Rang-1-FNIR und ihr 95%-Vertrauensbereich in Abhängigkeit vom zeitlichen Abstand zwischen der Aufnahme der Referenz- und Probebilder betrachtet, siehe Tabelle 10. Die Überlappung der Vertrauensbereiche zeigt, dass die Unterschiede zwischen den beobachteten Fehlerraten statistisch nicht signifikant sind. Mit den gegebenen Daten konnte für die getesteten Systeme keine Abhängigkeit der FNIR von der seit der Referenzaufnahme verstrichenen Zeit festgestellt werden.

Tabelle 10: Rang-1-FNIR in Abhängigkeit von der seit der Referenzaufnahme verstrichenen Zeit

System	weniger als 3,3 Jahre	3,3 Jahre bis 9,25 Jahre
A	2,0 % [0,9 %; 4,2 %]	1,4 % [0,5 %; 3,5 %]
B	2,0 % [0,9 %; 4,2 %]	1,7 % [0,7 %; 3,9 %]
C	2,3 % [1,1 %; 4,6 %]	1,4 % [0,5 %; 3,5 %]
D	2,6 % [1,3 %; 5,1 %]	1,4 % [0,5 %; 3,5 %]

3.5 Recherchen anhand frontaler Gesichtsbilder mit verminderter Bildqualität

3.5.1 Vorgehensweise

Um den Einfluß der Bildqualität auf die Erkennungsgenauigkeit zu evaluieren, wurde die Qualität jeder Kopie der 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder (siehe Abschnitt 3.2.2) auf 16 verschiedene Weisen in unterschiedlichem Grad verschlechtert:

(1) Verlustbehaftete Komprimierung: Es wurde JPEG-Kompression mit verschiedenen Werten für die JPEG-Qualität angewendet.
(2) Verringerung der Auflösung: Um kleinere Bilder zu simulieren, wurde die Auflösung mittels linearer Interpolation verkleinert.
(3) Mittelwertfilter: Zur Simulation von Unschärfeeffekten wurde das Bild mittels eines Mittelwertfilters weichgezeichnet.
(4) Medianfilter: Zur Simulation von Unschärfeeffekten wurde das Bild mittels eines Medianfilters weichgezeichnet.
(5) Gaußscher Filter: Zur Simulation von Unschärfeeffekten wurde das Bild mittels eines Gaußschen Filters weichgezeichnet.
(6) Gaußsches Rauschen: Es wurde jedem Pixel Gaußsches Rauschen hinzugefügt.
(7) Überlagerung mit Text: Zur Simulation von Text, der auf ein Bild aufgedruckt oder aufgestempelt wurde, wurde Text (bestehend aus Buchstaben, Ziffern und Sonderzeichen) hinzugefügt.
(8) Überlagerung mit Wasserzeichen: Zur Simulation von Gesichtsbildern, die aus Ausweisdokumenten eingescannt wurden, wurden die Bilder mit Grafiken überlagert.
(9) Überlagerung mit Guillochen: Das Gesichtsbild wurde mit unterschiedlichen Guillochen überlagert.
(10) Überlagerung mit »Salz und Pfeffer«: Es wurden zufällig Pixel im Bild ausgewählt und auf Schwarz oder Weiß gesetzt.
(11) Überlagerung mit Flecken: In einem Kreis mit zufällig ausgewähltem Radius und zufällig ausgewählter Position wurde die Pixelintensität um einen zufälligen Betrag verändert.
(12) Helligkeitsunterschiede: Das größte Problem für Gesichtserkennungsalgorithmen bilden Helligkeitsvariationen innerhalb eines Bildes. Zur Modellierung dieser Variationen wurde in den Bildern jeweils eine Helligkeitsänderung angewendet, die linear von einem Minimalwert an einer Kante zu einem Maximalwert an der gegenüberliegenden Kante ansteigt.
(13) Fischaugenabbildung: Mit Hilfe einer Fischaugenabbildung wurden kurze Brennweiten simuliert.
(14) Farbfehler: Es wurden unterschiedliche Farbstiche angewendet.
(15) Graustufen: Es wurden unterschiedliche Anzahlen an Graustufen angewendet.
(16) Graustich: Nach Konversion in ein Graustufenbild wurden unterschiedliche Graustiche angewendet.

Um sicherzustellen, dass zufällig variierendes Rauschen erzeugt wird, wurden für jedes Bild andere Werte für die Parameter zufällig aus einem Wertebereich ausgewählt. Die Wertebereiche der Parameter wurden jeweils anhand von Beispielbildern festgelegt.

Da die Qualität des Ausgabebilds nicht nur von den gewählten Werten der Transformationsparameter abhängt, sondern auch von der Qualität des Eingabebilds, kann anhand der Transformationsparameter die Qualität des Ausgabebilds nicht genau quantifiziert werden. Nur einzelne Merkmale wie JPEG-Qualität und Bildbreite können quantifiziert werden. PSNR dient als Maß für die Qualtitätsminderung im Vergleich zum Eingabebild. Bei früheren Evaluierungen von Gesichtserkennungssystemen verursachten Qualitätsminderungen, die PSNR-Werte über 30 dB ergeben, nur eine vernachlässigbare Abnahme der Erkennungsgenauigkeit [6].

Alle Gesichtsbilder mit gezielt verminderter Bildqualität waren Probebilder. Die Gesichtsbilder mit verminderter Bildqualität wurden nicht als Referenzbilder verwendet.

3.5.2 Ergebnisse in Abhängigkeit von der JPEG-Qualität

Für jedes getestete System wurde mit Hilfe von insgesamt 10 000 unterschiedlich komprimierten Gesichtsbildern die FTXR und die Rang-1-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit von der JPEG-Qualität ermittelt, siehe Tabellen 11 und 12. Auch die beste betrachtete JPEG-Qualität (31 % bis 40 %) führt noch zu erhöhten Rang-100-FNIR-Werten. JPEG-Kompression der Gesichtsbilder sollte vermieden werden.

Tabelle 11: FTXR in Abhängigkeit von der JPEG-Qualität

System	1 % bis 10 %	11 % bis 20 %	21 % bis 30 %	31 % bis 40 %
A	2,0 % [1,5 %; 2,6 %]	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,2 %]
B	6,0 % [5,1 %; 7,0 %]	6,3 % [5,4 %; 7,3 %]	6,5 % [5,6 %; 7,5 %]	5,3 % [4,4 %; 6,2 %]
C	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,2 %]
D	5,7 % [4,9 %; 6,7 %]	6,1 % [5,2 %; 7,1 %]	6,5 % [5,6 %; 7,6 %]	5,1 % [4,3 %; 6,1 %]

Tabelle 12: Rang-100-FNIR in Abhängigkeit von der JPEG-Qualität

System	1 % bis 10 %	11 % bis 20 %	21 % bis 30 %	31 % bis 40 %
A	13,3 % [12,0 %; 14,7 %]	0,9 % [0,6 %; 1,4 %]	1,1 % [0,8 %; 1,6 %]	1,2 % [0,8 %; 1,7 %]
B	14,9 % [13,6 %; 16,4 %]	6,9 % [6,0 %; 7,9 %]	7,2 % [6,2 %; 8,2 %]	6,0 % [5,1 %; 7,0 %]
C	22,3 % [20,7 %; 24,0 %]	6,3 % [5,4 %; 7,3 %]	6,7 % [5,8 %; 7,8 %]	6,2 % [5,3 %; 7,2 %]
D	13,7 % [12,4 %; 15,1 %]	7,5 % [6,6 %; 8,6 %]	7,5 % [6,5 %; 8,6 %]	6,5 % [5,6 %; 7,6 %]

3.5.3 Ergebnisse in Abhängigkeit von der Bildgröße

Für jedes getestete System wurde mit Hilfe von insgesamt 10 000 unterschiedlich großen Gesichtsbildern die FTXR und die Rang-1-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit von der Bildgröße ermittelt, siehe Tabellen 10 und 11. Auch die breitesten der betrachteten Bilder mit verringerter Auflösung führten noch zu erhöhten Rang-100-FNIR-Werten. Der überwiegende Teil der Bilder mit unverminderter Bildqualität hatte eine Bildgröße von 600 × 800 Pixeln, die nicht unterschritten werden sollte.

Tabelle 13: FTXR in Abhängigkeit von der Bildbreite

System	weniger als 125 Pixel	126 Pixel bis 250 Pixel	251 Pixel bis 600 Pixel
A	0,0 % [0,0 %; 0,2 %]	0,0 % [0,0 %; 0,1 %]	0,0 % [0,0 %; 0,2 %]
B	6,1 % [5,3 %; 7,2 %]	5,7 % [5,1 %; 6,4 %]	6,2 % [5,3 %; 7,2 %]
C	0,1 % [0,0 %; 0,4 %]	0,1 % [0,0 %; 0,2 %]	0,1 % [0,0 %; 0,3 %]
D	6,9 % [6,0 %; 8,0 %]	5,7 % [5,1 %; 6,4 %]	5,8 % [5,0 %; 6,8 %]

Tabelle 14: Rang-100-FNIR in Abhängigkeit von der Bildbreite

System	weniger als 125 Pixel	126 Pixel bis 250 Pixel	251 Pixel bis 600 Pixel
A	1,3 % [0,9 %; 1,8 %]	1,3 % [1,0 %; 1,7 %]	0,8 % [0,5 %; 1,2 %]
B	6,9 % [6,0 %; 8,0 %]	6,6 % [5,9 %; 7,3 %]	7,0 % [6,1 %; 8,1 %]
C	6,9 % [6,0 %; 8,0 %]	7,6 % [6,9 %; 8,4 %]	6,8 % [5,9 %; 7,8 %]
D	7,9 % [6,9 %; 9,0 %]	7,0 % [6,3 %; 7,7 %]	7,1 % [6,2 %; 8,2 %]

3.5.4 Ergebnisse in Abhängigkeit vom PSNR

Für jedes getestete System wurde mit Hilfe der anderen 140 000 Gesichtsbilder mit verminderter Qualität die FTXR und die Rang-100-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit vom PSNR ermittelt, siehe Tabellen 15 und 16. Die Systeme A und C zeigen bei Qualitätsminderungen, die PSNR-Werte über 20 dB ergeben, keine signifikante Erhöhung der Rang-100-FNIR. Die anderen beiden Systeme zeigen selbst bei geringen Qualitätsminderungen, die PSNR-Werte über 40 dB ergeben, eine signifikante Erhöhung der Rang-100-FNIR.

Tabelle 15: FTXR in Abhängigkeit vom PSNR

System	0 bis 10 dB	11 dB bis 20 dB	21 dB bis 30 dB	31 dB bis 40 dB	über 40 dB
A	0,4 % [0,3 %; 0,5 %]	0,1 % [0,0 %; 0,1 %]	0,0 % [0,0 %; 0,0 %]	0,0 % [0,0 %; 0,0 %]	0,0 % [0,0 %; 0,1 %]
B	9,2 % [8,8 %; 9,6 %]	6,4 % [6,1 %; 6,6 %]	5,7 % [5,5 %; 5,9 %]	6,5 % [6,2 %; 6,7 %	5,3 % [4,7 %; 5,9 %]
C	2,2 % [2,0 %; 2,4 %]	0,4 % [0,3 %; 0,4 %]	0,0 % [0,0 %; 0,1 %]	0,0 % [0,0 %; 0,0 %]	0,0 % [0,0 %; 0,1 %]
D	10,4 % [10,0 %; 10,9 %]	8,4 % [8,1 %; 8,7 %]	5,7 % [5,5 %; 5,9 %]	6,3 % [6,0 %; 6,5 %]	5,1 % [4,5 %; 5,7 %]

Tabelle 16: Rang-100-FNIR in Abhängigkeit vom PSNR

System	0 bis 10 dB	11 dB bis 20 dB	21 dB bis 30 dB	31 dB bis 40 dB	über 40 dB
A	2,0 % [1,8 %; 2,2 %]	0,7 % [0,6 %; 0,8 %]	0,4 % [0,3 %; 0,5 %]	0,3 % [0,3 %; 0,4 %]	0,3 % [0,2 %; 0,4 %]
B	9,6 % [9,2 %; 10,1 %]	6,6 % [6,4 %; 6,9 %]	6,0 % [5,8 %; 6,2 %]	6,6 % [6,4 %; 6,9 %]	5,5 % [4,9 %; 6,1 %]
C	6,6 % [6,2 %; 7,0 %]	3,6 % [3,4 %; 3,8 %]	3,2 % [3,1 %; 3,4 %]	3,5 % [3,3 %; 3,7 %]	3,1 % [2,7 %; 3,6 %]
D	12,0 % [11,6 %; 12,5 %]	9,4 % [9,1 %; 9,7 %]	7,1 % [6,8 %; 7,3 %]	7,4 % [7,2 %; 7,7 %]	6,3 % [5,7 %; 6,9 %]

3.6 Recherchen anhand nichtfrontaler Gesichtsbilder

3.6.1 Halbprofilbilder aus INPOL-Z

Die 10 000 zufällig aus INPOL-Z ausgewählten Halbprofilbilder mit passenden Gegenstücken in der Referenzdatenbank und die 10 000 zufällig aus INPOL-Z ausgewählten Halbprofilbilder ohne passendes Gegenstück in der Referenzdatenbank (siehe Abschnitt 3.2.4) wurden im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Aus den Protokolldaten für diese 20 000 Recherchen wurden für jedes getestete System die FTXR für Halbprofilbilder aus INPOL-Z und ihr 95%-Vertrauensbereich ermittelt, siehe Tabelle 14. Für jedes getestete System ist die FTXR für Halbprofilbilder signifikant höher als für Frontalbilder (vgl. Tabelle 3). Die Systeme B und D weisen mit den herstellerseitigen Voreinstellungen für Halbprofilbilder eine signifikant höhere FTXR auf als A und C. Andere Parametereinstellungen können zu anderen Ergebnissen führen.

Tabelle 17: FTXR für Halbprofilbilder aus INPOL-Z

System	FTXR
A	0,5 % [0,4 %, 0,6 %]
B	24,7 % [24,1 %, 25,3 %]
C	0,5 % [0,4 %, 0,6 %]
D	41,3% [40,7 %, 42,0 %]

Aus den Protokolldaten für die Recherchen anhand der 10 000 zufällig ausgewählten Halbprofilbilder mit passenden Gegenstücken in der Referenzdatenbank wurden für jedes getestete System die CMC sowie die Rang-k-FNIR über dem Rang ermittelt, siehe Abbildungen 9 und 10.

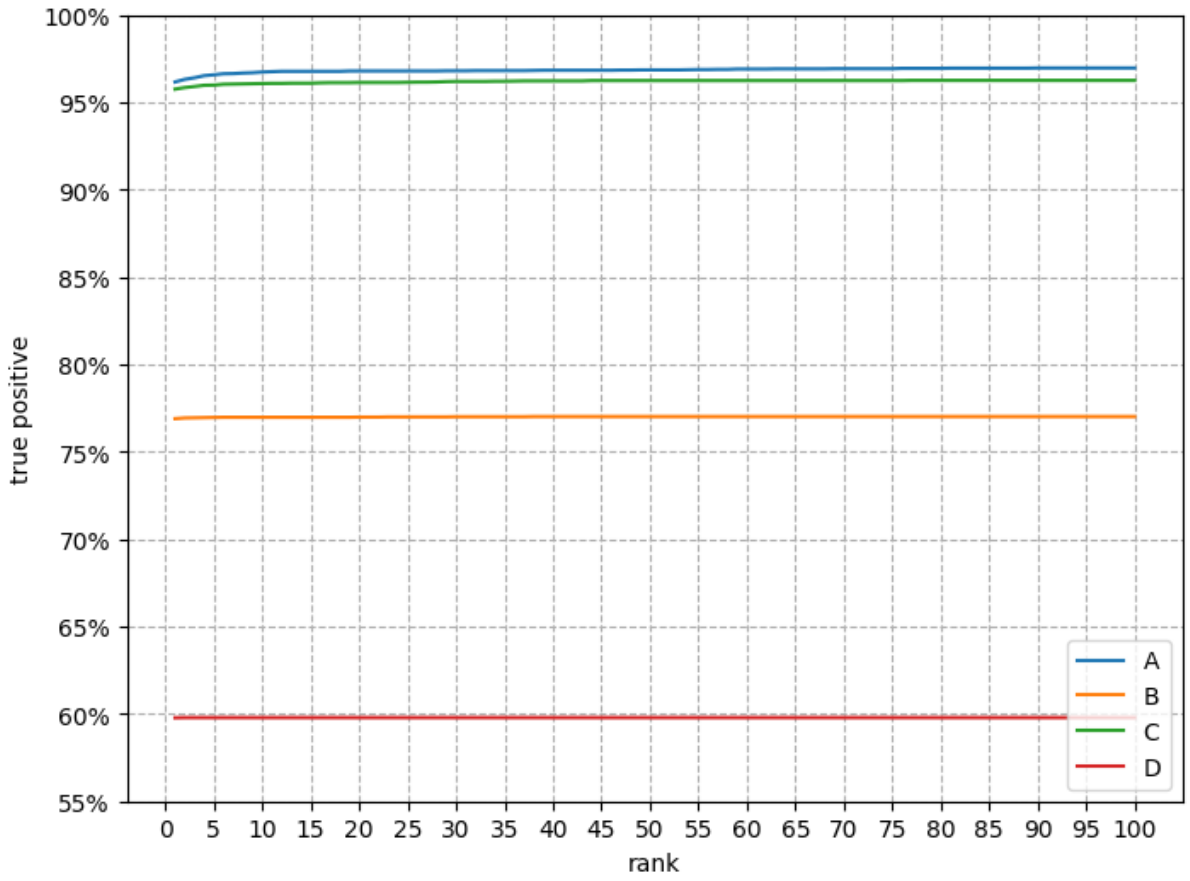

Abbildung 9: CMC für Halbprofilbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

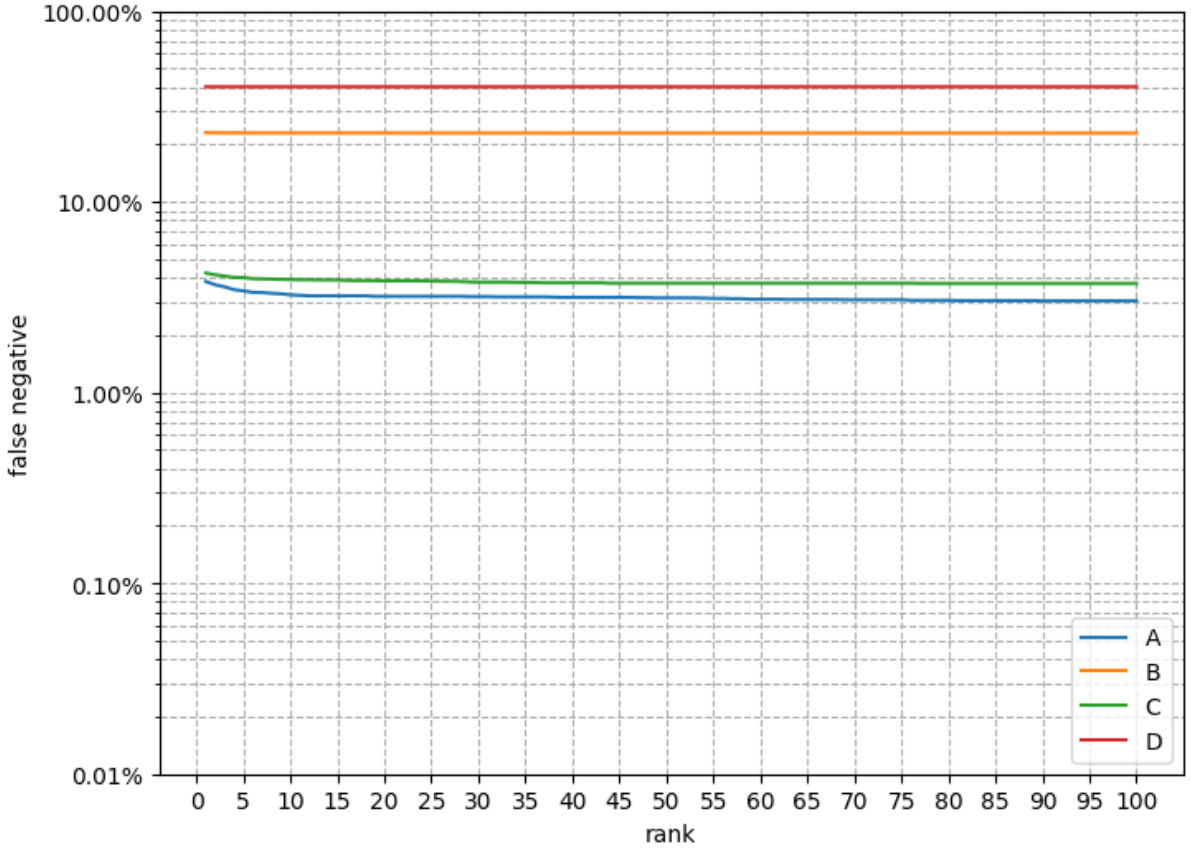

Abbildung 10: Rang-k-FNIR über dem Rang für beliebige Halbprofilbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

Tabelle 18 fasst für jedes getestete System die sich bei Recherchen anhand von Halbprofilbildern ergebenden Rang-1- und Rang-100-FNIR-Werte (die auch aus Abbildung 10 ersichtlich sind) und deren 95%-Vertrauensbereich zusammen. Für jedes getestete System sind die Rang-1- und Rang-100-FNIR-Werte für Halbprofilbilder signifikant höher als die für Frontalbilder (vgl. Tabelle 4).

Tabelle 18: Rang-1- und Rang-100-FNIR für Halbprofilbilder aus INPOL-Z

System	Rang-1-FNIR	Rang-100-FNIR
A	3,83 % [3,47 %; 4,22 %]	3,03 % [2,71 %; 3,38 %]
B	23,10 % [22,28 %; 23,94 %]	22,98 % [22,17 %; 23,81 %]
C	4,24 % [3,86 %; 4,65 %]	3,74 % [3,39 %; 4,13 %]
D	40,22 % [39,26 %; 41,18 %]	40,21 % [39,25 %; 41,17 %]

Wenn man von Merkmalsexraktionsfehlern (nach denen gar keine Kandidatenliste erstellt wurde) absieht und nur solche Fälle berücksichtigt, bei denen eine Kandidatenliste erstellt wurde, die aber keinen zugehörigen Referenzidentifikator enthielt, sind die Werte der Falschnegativraten um den FTXR-Wert aus Tabelle 17 kleiner als die Rang-k-FNIR-Werte aus Abbildung 10, siehe Abbildung 11. Abbildung 11 zeigt, dass die Systeme B und D für die Halbprofilprobebilder, aus denen sie jeweils fehlerlos Merkmale extrahieren konnten, deutlich geringere Falschnegativraten erreichen.

Um einen Überblick über die mit Halbprofilbildern erreichbaren Werte der Fehlerraten zu erlangen, wurde für jedes getestete System ein Rang-1-DET-Graph für Halbprofilbilder berechnet, siehe Abbildung 12.

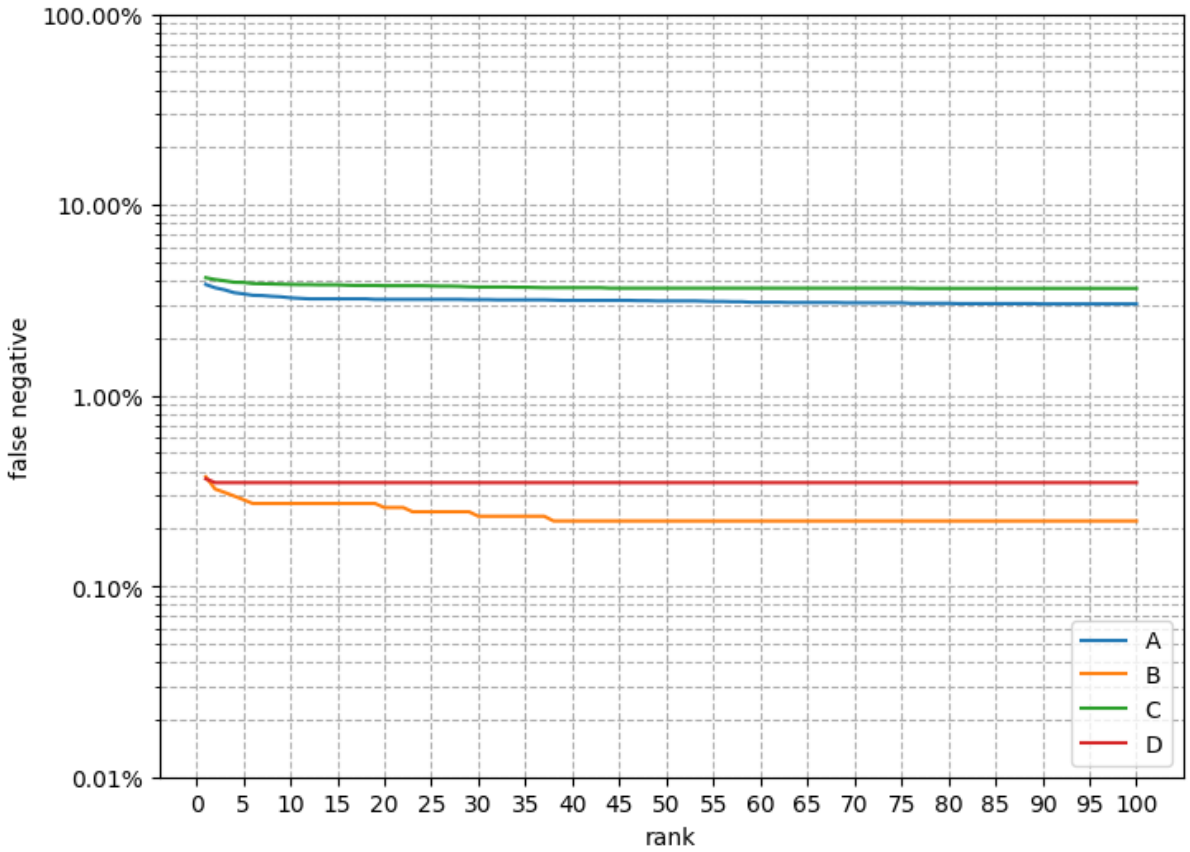

Abbildung 11: Rang-k-FNIR – FTXR über dem Rang für beliebige Halbprofilbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

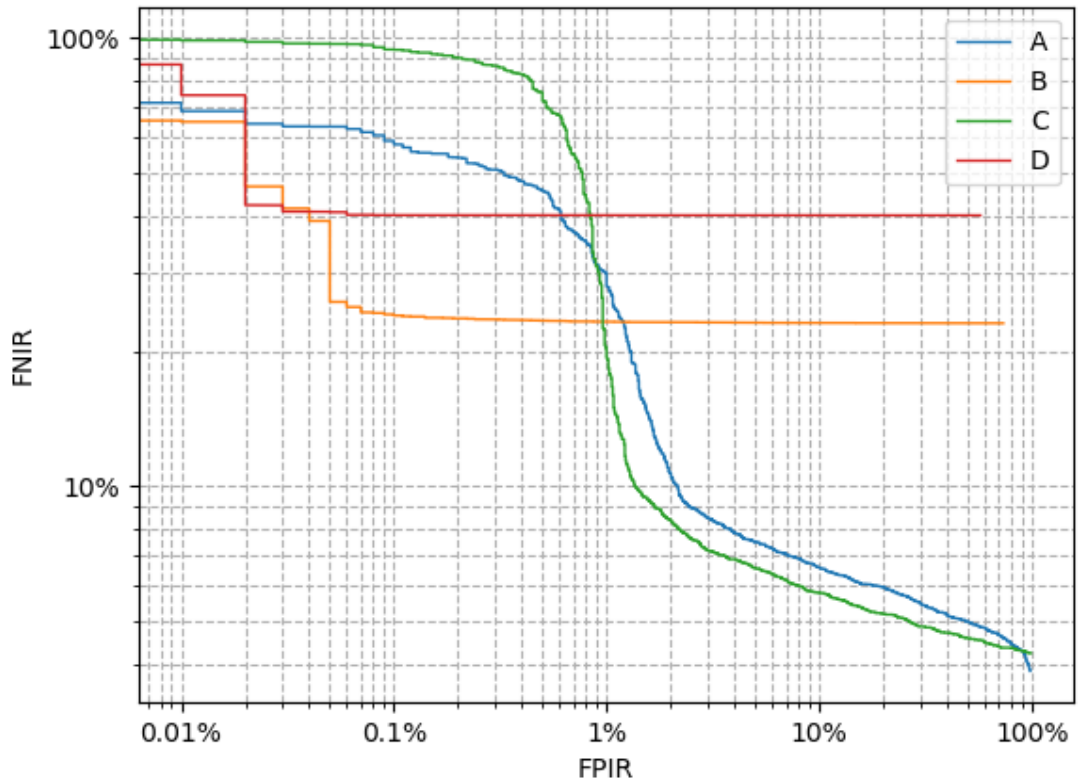

Abbildung 12: Rang-1-DET-Graph für Halbprofilbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

3.6.2 Gesichtsbilder aus verschiedenen Aufnahmewinkeln

Für jeden der verschiedenen Aufnahmewinkel wurden bis zu 257 Probebilder (siehe Abschnitt 3.2.5) im Stapelbetrieb gegen die ausschließlich aus frontalen Gesichtsbildern bestehende Referenzdatenbank recherchiert. Aus den Protokolldaten wurden für jedes getestete System die FTXR und die Rang-100-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit von Drehachse und -winkel ermittelt.

Die Tabellen 19 und 20 zeigen für jedes getestete System die FTXR und die Rang-100-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit vom Drehwinkel um die Hochachse. Qualitativ hochwertige Probebilder, auf denen der Kopf um bis zu 30° um die Hochachse gedreht ist, führten bei allen getesteten Systemen zu ähnlichen Rang-100-FNIR-Werten wie beliebige frontale Gesichtsbilder (vgl. Tabellen 20 und 4). System C erreichte sogar bei Recherchen anhand qualitativ hochwertiger Probebilder, auf denen der Kopf um bis zu 70° um die Hochachse gedreht ist, noch einen ähnlichen Rang-100-FNIR-Wert wie bei beliebigen frontalen Gesichtsbildern.

Tabelle 19: FTXR bei Drehung um die Hochachse (»Yaw Angle«)

System	10°	20°	30°	45°	60°	70°	80°	90°
A	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	1,2 % [0,4 %; 3,4 %]	14,8 % [11,0 %; 19,6 %]	37,7 % [32,0 %; 43,8 %]	40,9 % [35,0 %; 47,0 %]	43,2 % [37,3 %; 49,3 %]
B	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	0,4 % [0,1 %; 2,2 %]	28,0 % [22,9 %; 33,8 %]	83,7 % [78,6 %; 87,7 %]	97,7 % [95,0 %; 98,9 %]	99,6 % [97,8 %; 99,9 %]	100,0 % [98,5 %; 100,0 %]
C	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,8 % [0,2 %; 2,8 %]	0,0 % [0,0 %; 1,5 %]	4,7 % [2,7 %; 8,0 %]
D	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	62,6 % [56,6 %; 68,3 %]	98,8 % [96,6 %; 99,6 %]	98,1 % [95,5 %; 99,2 %]	100,0 % [98,5 %; 100,0 %]	98,8 % [96,6 %; 99,6 %]

Tabelle 20: Rang-100-FNIR bei Drehung um die Hochachse (»Yaw Angle«)

System	10°	20°	30°	45°	60°	70°	80°	90°
A	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	1,2 % [0,4 %; 3,4 %]	29,2 % [24,0 %; 35,0 %	78,6 % [73,2 %; 83,2 %]	100,0 % [98,5 %; 100,0 %]	100,0 % [98,5 %; 100,0 %]
B	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	0,4 % [0,1 %; 2,2 %]	28,0 % [22,9 %; 33,8 %]	84,0 % [79,1 %; 88,0 %	98,1 % [95,5 %; 99,2 %]	99,6 % [97,8 %; 99,9 %]	100,0 % [98,5 %; 100,0 %]
C	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %	0,8 % [0,2 %; 2,8 %]	3,1 % [1,6 %; 6,0 %]	23,0 % [18,2 %; 28,5 %]
D	0,4 % [0,1 %; 2,2 %]	0,4 % [0,1 %; 2,2 %]	0,8 % [0,2 %; 2,8 %]	62,6 % [56,6 %; 68,3 %]	100,0 % [98,5 %; 100,0 %]	100,0 % [98,5 %; 100,0 %]	100,0 % [98,5 %; 100,0 %]	99,2 % [97,2 %; 99,8 %]

Die Tabellen 21 und 22 zeigen für jedes getestete System die FTXR und die Rang-100-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit vom Drehwinkel um die Querachse. Qualitativ hochwertige Probebilder, auf denen der Kopf um bis zu 20° um die Querachse gesenkt bzw. gehoben ist, führten bei allen getesteten Systemen zu ähnlichen Rang-100-FNIR-Werten wie beliebige frontale Gesichtsbilder (vgl. Tabellen 22 und 4).

Tabelle 21: FTXR bei Drehung um die Querachse (»Pitch Angle«)

System	-45°	-30°	-20°	-10°	10°	20°	30°	45°
A	3,9 % [2,1 %; 7,0 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]
B	16,0 % [12,0 %; 21,0 %]	1,2 % [0,4 %; 3,4 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	9,4 % [6,4 %; 13,7 %]
C	0,8 % [0,2 %; 2,8 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	5,9 % [3,6 %; 9,5 %]
D	18,4 % [14,1 %; 23,6 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	9,7 % [6,7 %; 14,0 %]	74,4 % [68,7 %; 79,4 %]

Tabelle 22: Rang-100-FNIR bei Drehung um die Querachse (»Pitch Angle«)

System	-45°	-30°	-20°	-10°	10°	20°	30°	45°
A	6,3 % [3,9 %; 9,9 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	26,8 % [21,7 %; 32,5 %]
B	16,4 % [12,4 %; 21,4 %]	1,2 % [0,4 %; 3,4 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,2 %]	23,2 % [18,5 %; 28,8 %]
C	7,0 % [4,5 %; 10,8 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	1,2 % [0,4 %; 3,4 %]	36,6 % [30,9 %; 42,7 %]
D	19,5 % [15,1 %; 24,8 %]	1,2 % [0,4 %; 3,4 %]	0,4 % [0,1 %; 2,2 %]	0,8 % [0,2 %; 2,8 %]	0,0 % [0,0 %; 1,5 %]	1,6 % [0,6 %; 3,9 %]	10,1 % [7,0 %; 14,4 %]	75,2 % [69,5 %; 80,1 %]

Die Tabellen 23 und 24 zeigen für jedes getestete System die FTXR und die Rang-100-FNIR sowie deren 95%-Vertrauensbereiche in Abhängigkeit vom Drehwinkel um die um die Längsachse. In den getesteten Konfigurationen der Systeme B und D schlugen Recherchen anhand von Probebildern, auf denen der Kopf um 45° um die Längsachse gedreht ist, fehl. Im Einsatz in der Kriminalistik stellt dies kein Problem dar, da Drehungen um die Längsachse (»Roll Angle«) vor der Recherche leicht manuell korrigiert werden können. Es wurden wieder die herstellerseitigen Voreinstellungen verwendet. Andere Parametereinstellungen können zu anderen Ergebnissen führen.

Tabelle 23: FTXR bei Drehung um die Längsachse (»Roll Angle«)

System	10°	20°	30°	45°
A	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,6 %]
B	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	1,2 % [0,4 %; 3,4 %]	97,4 % [94,4 %; 98,8 %]
C	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,4 % [0,1 %; 2,4 %]
D	0,4 % [0,1 %; 2,2 %]	0,4 % [0,1 %; 2,2 %]	7,1 % [4,5 %; 10,9 %]	98,7 % [96,2 %; 99,6 %]

Tabelle 24: Rang-100-FNIR bei Drehung um die Längsachse (»Roll Angle«)

System	10°	20°	30°	45°
A	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,6 %]
B	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	1,2 % [0,4 %; 3,4 %]	97,8 % [95,0 %; 99,1 %]
C	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	0,0 % [0,0 %; 1,5 %]	3,1 % [1,5 %; 6,2 %]
D	0,4 % [0,1 %; 2,2 %]	0,4 % [0,1 %; 2,2 %]	7,5 % [4,8 %; 11,3 %]	100,0 % [98,4 %; 100,0 %]

3.7 Untersuchung möglicher Systemkombinationen

Um die Schwächen der einzelnen Gesichtsidentifizierungssysteme zu umgehen, wurden die Kandidatenlisten von jeweils zwei Systemen auf einfache Weise auf Rangebene [7] zu einer Kandidatenliste fusioniert (mittels Borda-Wahl). Den Ausgangspunkt bildeten die Kandidatenlisten der Recherchen anhand der 10 000 zufällig aus INPOL-Z ausgewählten frontalen Probebilder mit passenden Gegenstücken in der Referenzdatenbank (siehe Abschnitt 3.4.1). Aus den verschmolzenen Kandidatenlisten aller möglichen Kombinationen aus zwei Systemen wurden die CMC und die Rang-k-FNIR über dem Rang ermittelt, siehe Abbildungen 13 und 14.

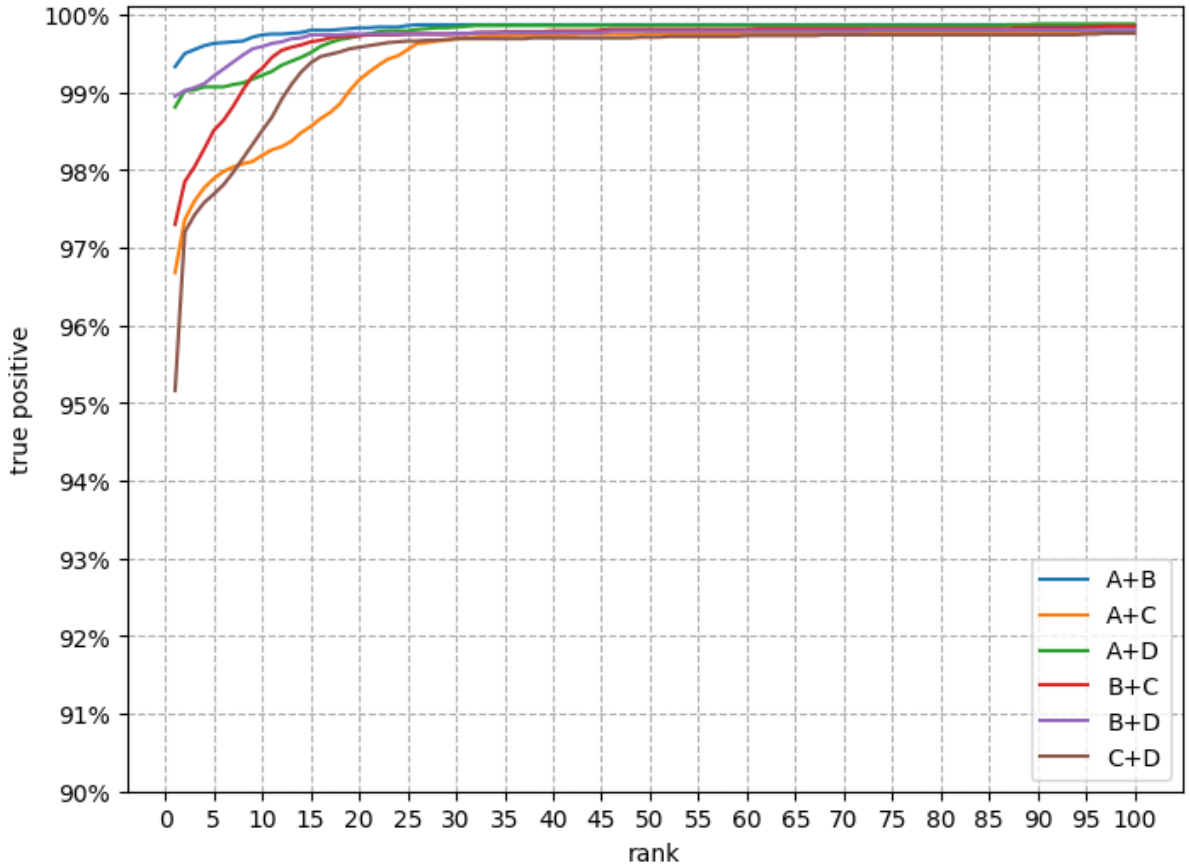

Abbildung 13: CMC von Systemkombinationen für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

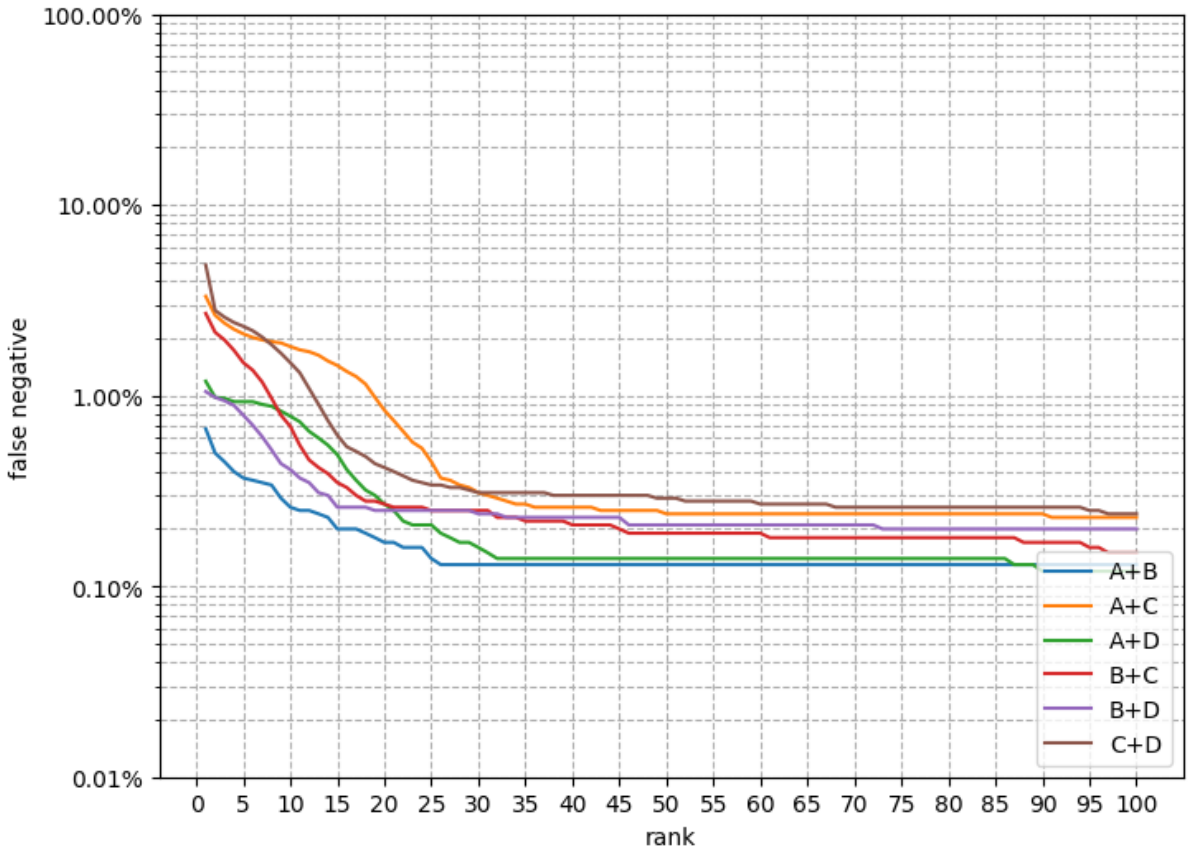

Abbildung 14: Rang-k-FNIR über dem Rang für Systemkombinationen für beliebige Frontalbilder bei ca. $4{,}8 \cdot 10^6$ Referenzen

Tabelle 25 fasst für jede mögliche Systemkombination die Rang-100-FNIR (die auch aus Abbildung 14 ersichtlich ist) und deren 95%-Vertrauensbereich zusammen. Die Werte sind in Tabelle 25 aufsteigend geordnet. Der Vergleich mit der Rang-100-FNIR der einzelnen Systeme (Tabelle 4) zeigt, dass die Rang-100-FNIR jeder Systemkombination geringer ist als die Rang-100-FNIR jedes einzelnen Systems. Die Rang-100-FNIR-Werte der besten Systemkombinationen sind nur etwa halb so groß wie die Rang-100-FNIR-Werte der besten Einzelsysteme.

Tabelle 25: Rang-100-FNIR für mögliche Systemkombinationen für beliebige Frontalbilder aus INPOL-Z

System	Rang-100-FNIR
A+D	0,12 % [0,07 %; 0,21 %]
A+B	0,13 % [0,08 %; 0,22 %]
B+C	0,15 % [0,09 %; 0,25 %]
B+D	0,20 % [0,13 %; 0,31 %]
A+C	0,23 % [0,15 %; 0,34 %]
C+D	0,24 % [0,16 %; 0,36 %]

Literaturverzeichnis

[1] Michael Eid, Mario Gollwitzer, Manfred Schmitt: Statistik und Forschungsmethoden. 5., korr. Aufl., Beltz Verlag, 2017

[2] International Standard ISO/IEC 2382-37:2017, Information technology – Vocabulary – Part 37: Biometrics

[3] International Standard ISO/IEC 19795-1:2006, Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework

[4] NIST: Ongoing Face Recognition Vendor Test (FRVT) – Part 2: Identification. NIST Interagency Report NISTIR 8238, November 2018

[5] Fraunhofer IGD: Evaluierung am Markt erhältlicher Gesichtserkennungssysteme für den Einsatz in der Kriminalistik hinsichtlich der Erkennungsgenauigkeit, Robustheit und der Einhaltung weiterer polizeilicher Anforderungen. Verfahrensanweisung IGD-VA-2018-03, Version 1.0, 2019

[6] Andreas Mascher-Kampfer, Herbert Stögner, Andreas Uhl: Comparison of compression algorithms' impact on fingerprint and face recognition accuracy. Proc. SPIE 6508, Visual Communications and Image Processing 2007, 650810 (29 January 2007)

[7] Tin Kam Ho, Jonathan J. Hull, Sargur N. Srihari: Decision Combination in Multiple Classifier Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 16, No. 1, S. 66–75 (1994)