DAIMLER

Heterogene kamerabasierte Personenidentifikation

Andreas Bauerfeld

29. Juli 2017

Version:

Universität Stuttgart

Institut für Visualisierung und interaktive Systeme
Mensch-Maschine Interaktion

Masterarbeit

Heterogene kamerabasierte Personenidentifikation

Andreas Bauerfeld

1. Reviewer	Dr. Wolfgang Stolzmann RD/FAV Daimler AG
2. Reviewer	Emin Tarayan RD/UIF Daimler AG
3. Reviewer	Prof. Dr. Albrecht Schmidt VIS-HCI Universität Stuttgart
Supervisors	Dr. Wolfgang Stolzmann und Emin Tarayan

29. Juli 2017

Andreas Bauerfeld

Heterogene kamerabasierte Personenidentifikation

Masterarbeit, 29. Juli 2017
Reviewers: Dr. Wolfgang Stolzmann, Emin Tarayan und Prof. Dr. Albrecht Schmidt
Supervisors: Dr. Wolfgang Stolzmann und Emin Tarayan

Universität Stuttgart
Mensch-Maschine Interaktion
Institut für Visualisierung und interaktive Systeme

Pfaffenwaldring 5a
70569 und Stuttgart

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Andreas Bauerfeld | Stuttgart, 29. Juli 2017

Declaration

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

Andreas Bauerfeld | Stuttgart, 29. Juli 2017

Zusammenfassung

Die maschinelle Gesichtserkennung hält nicht zuletzt durch ihr breites Anwendungsspektrum Einzug in den Alltag. In der näheren Vergangenheit konnten erstmals Verfahren vorgestellt werden, welche die Möglichkeiten der menschlichen Wahrnehmung innerhalb enger Grenzen von Umwelteinflüssen übertreffen können. Eine bisher wenig betrachtete Ausgangssituation ist die heterogene Gesichtserkennung, bei welcher Abbildungsgeräte verschiedener Größen, wie etwa Wellenlänge oder Tiefe wechselseitig verwendet werden können.
Im speziellen fokussiert sich die Arbeit auf das Vorstellen einer Kategorisierung von Lösungsansätzen für welche Repräsentanten entwickelt und vergleichend evaluiert werden. Zum Einsatz kommen dabei neben modernen Abbildungsgeräten wie des Fovio-Systems auch verschiedenste neuronale Netze sowie öffentliche Benchmark-Datenbanken.
Abschließend wird eine Einschätzung des technisch Möglichen zusammen mit einem Ausblick von vielversprechenden Forschungsrichtungen gegeben und in Bezug zu den erlangten Erkenntnissen gesetzt.

Abstract

Facial recognition has a broad impact in every day life and became more and more common in the last years. Recently, it was possible to introduce techniques that surpass human grade accuracy, but only in a strictly limited way in matters of environmental variance. Human cognition still outperforms machines when it comes to robustness in sense of head rotation or facial expression. However, the following tries to solve a different problem: heterogeneous face recognition!
Heterogeneous face recognition is the task of automatically match identities captured by different sensing techniques. To accurately describe this process, a categorization is introduced. For each category several methods are developed and subsequently evaluated. The methods introduced, range from state of the art sensing devices like the Fovio system to Convolutional Neuronal Network and Deep Learning. The Evaluation is done by offical benchmark databases or a self conducted user study.
The thesis ends with a short look in the future, based on the results and knowledge obtained in the evaluation part.

Inhaltsverzeichnis

1 Motivation

Gesichtserkennung wird von den meisten als gelöstes Problem angesehen. Ein näherer Blick offenbart jedoch, dass diese Aussage nur mit vielen Einschränkungen zutrifft. Zumal die Robustheit moderner Methoden immer noch hinter dem menschlichen Leistungsvermögen zurückliegt, gibt es weitere Subgebiete die weitgehend neu und wenig entwickelt sind. Zu diesen gehört die heterogene Gesichtserkennung bei welcher verschiedene Medien kombiniert wird. Ziel ist es eine gemessene Eigenschaft auf eine andere Umgebung zu übertragen.

Die Identifikation über Modalitätsgrenzen hinweg ist keine rein akademische Fragestellung. Anwendungsfälle umfassen den Abgleich von konventionellen Bildmaterial mit Phantomzeichnungen oder die maschinelle Identifikation von Karikaturen. Auch die Anpassung und Erweiterung eines Gesamtsystems durch verschiedene Abbildungsgeräte, ermöglicht die Identifikation unter schwierigen Bedingungen. Zum Beispiel ist es möglich ein konventionelles System durch Infrarotkameras zu erweitern, so dass eine Erkennung in der Dunkelheit realisierbar wäre.

Ein konkreter Anwendungsfall wäre ein IR basiertes Driver Monitoring System, welches zum Anlernen eines Individuums auf Smartphones zurückgreifen und sich mit anderen Ausführungen kombinieren lässt. Es ist in diesem Anwendungsfall nicht möglich auf herkömmliche Kameras zurückzugreifen, da diese bei Nachtfahrten eine Lichtquelle benötigen würden, welche den Fahrer nachhaltig stört. Auch von Vorteil ist diese Unabhängigkeit bei der Erweiterung eines bestehenden Gesamtsystems durch neuere Modelle um eine Abwärtskompatibilität zu gewährleisten.

Durch eine Kombination von verschiedenen Geräten lassen sich ebenfalls die Kosten minimieren indem Abbildungsgeräte für Spezialfälle eben nur für diese zum Einsatz kommen und anderweitig auf klassische Kameras im eigentlichen Sinne zurückgegriffen wird. Ein Überwachungssystem könnte somit zum Beispiel durch IR-Kameras erweitert werden, wenn die Lichtverhältnisse eine konventionelle Aufnahme nicht zulassen. Die Bilder der IR-Kameras können dann über das gesamte System hinweg verwendet und mit den Aufnahmen der anderen Geräte abgeglichen werden.

Stand der Technik 2

> *Auf den Schultern von Giganten.*
>
> **– Google**

2.1 Immernoch eine Herausforderung?

Gesichtserkennung ist in aller Munde. Meist geht es dabei um datenschutzrechtliche Aspekte und innere Sicherheit. Wenig wird aber über den Reifegrad dieser Systeme gesprochen. Die Vielzahl an kommerziell erhältlichen Algorithmen, angefangen bei der Authentifizierung von Smartphonenutzern bis hin zur Beweissicherung im öffentlichen Raum, lassen vermuten, dass die aktuelle Technik, ein Gesicht mindestens ebenso präzise differenzieren kann, wie der Mensch.

Diese Annahme ist, wenn überhaupt, jedoch nur sehr eingeschränkt gültig: Sobald Störfaktoren wie z. B. Verzerrungen, Verdeckungen, Gesichtsausdruck, Kopfposition, Lichteinflüsse und andere Aufnahmevarianzen hinzukommen, nimmt die Präzision stark ab. Die verbleibende Fähigkeit ist weit unter der eines durchschnittlichen Menschen.(Ouyang et al., 2014)

Diese störenden Umwelteinflüssen treten ständig auf und können nur durch aktives Handeln des zu identifizierenden abgemildert werden, was wiederum den Grad der Immersion drastisch verringert. Die Toleranz eines Systems gegenüber derartigen Störfaktoren nennt man Robustheit. Man kann also daraus schließen, dass die automatisierte Identifikation von Menschen anhand des Erscheinungsbilds ihres Gesichts bisher nur mit eingeschränkten Umweltparametern hinreichend genau, dass heißt mindestens genauso präzise wie ein Mensch, möglich ist.(Ouyang et al., 2014)

Vielfältige Methoden zur Robustifizierung von Gesichtserkennungsverfahren wurden bereits publiziert und ihre Zahl steigt jedes Jahr stetig. Meist können diese Ansätze lediglich einen Störfaktor abschwächen und sind nicht frei kombinierbar mit anderen Robustifikationen. Im letzen Jahrzehnt haben sich neuronale Netze bewährt, welche auf variantenreichen, extrem großen Datensätzen trainiert werden. Das führt zu einer automatisierten Auswahl von Merkmalen die gegenüber im Trainingsdatensatz enthaltene Störfaktoren invariant sind. Der Black-Box-Automatismus dieser

Systeme bringt es allerdings mit sich, dass dessen Funktion ausgiebig analysiert werden muss, um sich der Korrektheit zu vergewissern.(Ouyang et al., 2014)

2.2 Gesichtserkennung

Um die Besonderheiten der heterogenen Gesichtserkennung genauer zu untersuchen ist es ratsam zunächst einen kleinen Überblick über Methoden der klassischen Gesichtserkennung zu erhalten, weshalb nachfolgend eine Kategorisierung vorhandener Ansätze vorgestellt wird. Im Verlauf wird man feststellen, dass diese Art der Einordnung sich nur auf die klassische Gesichtserkennung bezieht und für die Klassifikation von heterogenen Verfahren ungeeignet ist.(Ouyang et al., 2014)

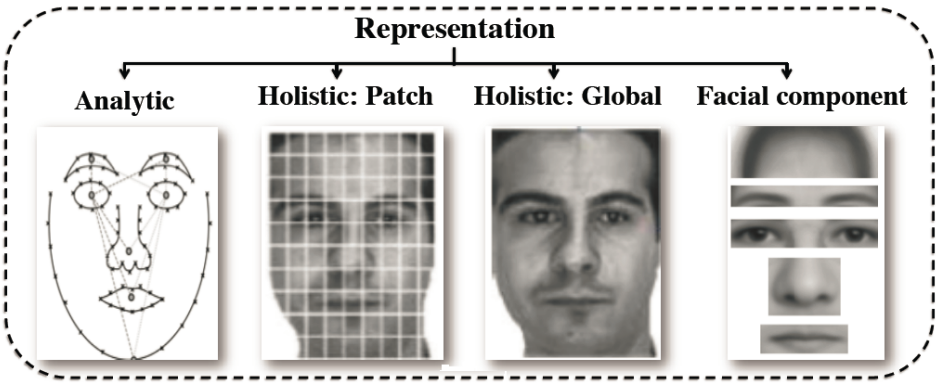

Abb. 2.1 Verschiedene Arten von Gesichtserkennungsrepräsentationen (Ouyang et al., 2014)

Homogene Klassifikatoren werden oft daran kategorisiert, wie Gesichtszüge innerhalb der Algorithmen repräsentiert werden. Da eine Repräsentation für den heterogenen Fall jedoch nicht gleich, sondern lediglich vergleichbar sein muss, ist eine derartige Aufteilung wenig hilfreich. Als Einstieg in die Grundlagen der Thematik aber recht nützlich.

2.2.1 Analytisch

Analytische Ansätze gehören zu den ältesten Methoden der Gesichtserkennung. Es wird versucht das aufgenommene Bild in eine mathematisch interpretierbare Repräsentation zu transformieren. Fourier basierte Techniken gehören genauso zu dieser Gruppe wie die Verwendung von geometrischen Masken und Punktbeziehungen.

Tatsächlich war der erste Algorithmus zur Gesichtserkennung ein analytisch geometrischer Ansatz. (Kanade, 1973)
Ist man in der Lage gewisse Gesichtsmerkmale innerhalb eines Bildes zu lokalisieren, so kann man eine Verteilung über eine Probandengruppe ermitteln und jedes neue Gesicht statistisch in diese Gruppe einordnen (siehe Abbildung 2.1 1. v. l.). Auch die Identifikation beruht bei derartigen Methoden auf einem statistischen Abstandsmaß. Jede Form von *Active Shape*(ASM) oder *Active Appearance Models*(AAM) gehört genauso zu dieser Kategorie wie *3D Morphable Models*(3DMM). Auch in dieser Arbeit werden verschiedene analytische Verfahren als einfacher ad-hoc Ansatz behandelt.
Der Vorteil dieser Repräsentation ist die geringe Datenmenge die für ein Gesicht verwaltet werden muss, auf der anderen Seite werden viele identifizierende Merkmale nicht betrachtet und ungenutzt verworfen. Im Allgemeinen kann man sagen, dass analytische Methoden nicht gut mit der Anzahl an Personen in der Datenbank skalieren. In den Anfängen der maschinellen Gesichtserkennung waren analytische Ansätze lange Zeit führend, mit Zunahme der Rechenkraft und dem Einzug von neuronalen Netzen haben sie jedoch viel an Bedeutung verloren. Praktisch kein modernes System beruht noch auf analytischen Repräsentation. Lediglich zur Frontalisierung des Gesichts, dass heißt zur geometrischen Vergleichbarkeit, werden sie noch als Vorstufe zu einer Identifikation verwendet. (Ouyang et al., 2014) Die global-hollistischen Ansätze bilden den Gegensatz zu der hier vorgestellten Kategorie.

2.2.2 Global-Hollistisch

Die global-hollistischen Methoden nutzen das Bild selbst als Repräsentation (siehe Abbildung 2.1 2. v. r.). Eine Aufteilung in einzelne Regionen findet nicht statt. Alle im Bild enthaltenen Informationen werden auch zur Identifikation herangezogen. Einerseits führt diese Verfahrensweise zu einer maximalen Informationsbasis, anderseits werden etwaige Störungen in die Identifikation eingebunden. Eine Vielzahl von Methoden ist verfügbar um derartige Bilder gegen Verdeckungen zu robustifizieren. Es muss also eine Gleichgewicht zwischen den in einem Bild verfügbaren, identifizierenden Informationen und der Varianz innerhalb eines Individuums gefunden werden.(Ouyang et al., 2014)
Diese Art von Verfahren sind in Kombination mit neuronalen Netzen weitverbreitet, obwohl, wie man später sehen wird, eine ganzheitliche Einordnung zu dieser Kategorie nicht ganz korrekt wäre. Ein neuronales Netz könnte seine identifizierenden

Merkmale durchaus nur aus Subregionen anstatt dem gesamten Bild beziehen, wie in Unterunterabschnitt 2.4.1.1 noch näher beleuchtet werden wird.
Die bekannteste Vertreter für global-hollistische Verfahren sind die *Eigenfaces*. (Belhumeur et al., 1997) Es handelt sich um dabei um eine *Principal Component Analysis*(PCA) die auf das gesamte Bild angewendet wird. Die Koeffizienten sind dabei der Identifikationsvektor. Jegliche Varianz von Kopfposition und Gesichtsausdruck wirkt sich direkt auf die Genauigkeit des Systems aus und muss durch geeignete Maßnahmen behandelt werden. Verdeckungen führen hierbei direkt zum Fehlschlag des Identifikationsversuchs, was die Anwendungsfälle dieses Ansatzes stark einschränkt. Dieses Problem wird von den lokal-hollistischen Ansätzen gelöst.

2.2.3 Lokal-Hollistisch

Die lokal-hollistischen Verfahren operieren auf kleinen Subregionen, auch *Patches* genannt (siehe Abbildung 2.1 2. v. l.). Diese können durchaus zu einem ganzheitlichen Identifikationsvektor zusammengesetzt werden um eine finale Aussage zum Vergleich zweier Gesichter zu erhalten, jedoch wird jedes Patch einzeln prozessiert. Das führt zu mehreren Vorteilen gegenüber den globalen Verfahren. Zum einen, können Verzerrungen, welche durch das Aufnahmegerät, Kopfrotation oder Gesichtsausdruck verursacht wurden, durch die separate Registrierung jedes Patches besser ausgeglichen werden, zum anderen sind Verdeckungen explizit modellierbar. Verdeckte Regionen werden aus dem Identifikationsvektor ausgeschlossen und nicht in den Abgleich miteinbezogen. (Ouyang et al., 2014)
Histogramm basierte Verfahren gehören zu dieser Gruppe. Hierbei werden gewisse Eigenschaften über eine eingeschränkte Nachbarschaft evaluiert. Bei den *Histogram of Oriented Gradients*(HOG) wird zum Beispiel die Verteilung von Gradientenrichtung und Stärke in einer gewissen Region berechnet und anschließend verglichen. HOG ist eine weitverbreitete und äußerst robuste Technik zur Gesichts und Mustererkennung.(Ouyang et al., 2014)
Bei den Subregionen handelt es sich um ein gleichmäßig angelegtes Raster. Eine Aufteilung nach Gesichtsregionen findet nur beim Komponenten basierten Ansatz Anwendung.

2.2.4 Komponenten basiert

Die Komponenten basierte Ansätze (siehe Abbildung 2.1 1. v. r.) teilen das Bild in verschiedene Gesichtskomponenten auf, die dann einzeln behandelt und anschließend zu einem Identifikationsvektor verbunden werden. Die Voraussetzung für die Aufteilung in Komponenten ist die genaue Lokalisation dieser, was nicht unbedingt trivial und ein eigenständiges Forschungsgebiet ist. Die Vorteile einer solchen Herangehensweise sind offensichtlich, verschiedene, charakteristische Regionen können unterschiedlich behandelt werden, was zu einen robusten Klassifikator führt. Der Mund unterliegt zum Beispiel starker Varianz (Sprechen, Lachen, usw.) und ist somit weniger charakteristisch als zum Beispiel die Augensymmetrie.
Allerdings ergeben sich auch die klassischen Nachteile der lokalen Ansätze: Erstens, kann die gesamtheitliche Form nicht für die Identifikation genutzt werden und zweitens ergeben sich durch die Zerteilung Randbedingungen die in manchen Verfahren mathematisch schlecht zu handhaben sind. (Ouyang et al., 2014)
Der Identifikationsvektor ergibt sich dann aus der Verbindung aller Komponenten. Gesichtsregionen welche in der prozessierten Aufnahme als verdeckt, verrauscht oder anderweitig als problematisch klassifiziert wurden, können explizit aus dem Vergleich genommen werden. Eine Wertung nach Zuverlässigkeit und Eignung der einzelnen Komponenten ist ebenso explizit gegeben.(Ouyang et al., 2014)
In der Praxis sind diese Verfahren eher selten, da die genaue Registrierung der Gesichter großen Einfluss auf die Qualität der Ergebnisse hat. Gerade diese Registrierung ist jedoch problematisch und fehleranfällig. (Ouyang et al., 2014)

2.3 Heterogene Gesichtserkennung

Nachdem verschiedene Gruppen von Methoden und Algorithmen für Gesichtserkennung erläutert wurden, widmet sich dieses Kapitel der *heterogenen* Gesichtserkennung. Um den Unterschied zur klassischen Problemstellung zu verdeutlichen und weitergehende Fragestellungen zu erarbeiten, wird der Begriff zu erst informell umrissen bevor eine Kategorisierung von Lösungsansätzen und deren Abgrenzung vorgestellt wird.

2.3.1 Definition

Wie bereits in Abschnitt 2.2 beschrieben ist die maschinengestützte Gesichtsanalyse eine der am weitesten verbreiteten Form der Mustererkennung. Nach über Vierzig Jahren intensiver Weiterentwicklung sind die meisten Verfahren deutlich besser als die menschliche Wahrnehmung. Allerdings nur innerhalb geringer Paramterintervalle bezüglich Beleuchtung, Kopfposition und Rotation sowie Gesichtsausdruck. Der „Big Data“ Ansatz, das heißt das stumpfe lernen aller genannten Varianzen, verspricht, wie bereits in Abschnitt 2.2 erwähnt, hierbei Besserung. Diese Art von Unabhängigkeit nennt man *Robustheit* und muss strengstens von der heterogenen Gesichtserkennung abgegrenzt werden.
Heterogene Gesichtserkennung beschreibt eben gerade nicht eine Robustifikation gegen eine Aufnahmeungenauigkeit, sondern die Verwendung verschiedener Aufnahmesysteme und die damit verbundenen Gesichtsrepräsentationen, welche im weiteren Verlauf der Arbeit, als *Modalitäten* bezeichnet werden.
Während die klassische Gesichtserkennung also gleiche Modalität verwendet und lediglich Robustheit gegen Veränderung der Szenerie ein Problem darstellt, befasst sich die heterogene Gesichtserkennung mit der Analyse verschiedenster Arten von Gesichtsrepräsentationen. Wobei eine Gesichtsrepräsentation von biometrischen Koeffizienten eine Hauptkomponentenanalyse bis hin zu einem Tiefenbild reichen kann. Dabei ist sie nicht auf bildgebende Sensoren beschränkt. Der Abgleich von forensischen Phantombildern mit einer Bilddatenbank aus dem VIS oder NIR Spektrum, zählt durchaus auch zu dieser Kategorie. Explizit ist damit die Übertragung der in einer Modalität erlangten Information in eine andere gemeint. Es ist zum Beispiel fast trivial einen Klassifikator für das VIS Spektrum derart auf das NIR Spektrum umzutrainieren. Die Übertragbarkeit der in Infrarot erlangten Informationen in das sichtbare Spektrum ist allerdings immer noch ein sehr forderndes Gebiet der Wissenschaft. Die weitreichenden Anwendungsgebiete, welche bereits in der Motivation besprochen wurden, macht es jedoch zu einem vielversprechenden Gebiet der Forschung.
Heterogene Gesichtserkennung umfasst nach (Ouyang et al., 2014), die in Abbildung 2.2 beschriebenen Modalitäten. Das aktivste Gebiet hierbei ist die bereits angesprochene Übertragbarkeit zwischen NIR und VIS, da beide Sensoren günstig und allgegenwärtig sind. Weiterhin ermöglicht das Infrarotmedium die Erkennung bei Nacht, ohne den Nutzer zu durch eine Lichtquelle zu beeinflussen. Die Übertragbarkeit zwischen äußerlichen Veränderungen wie z. B. Make-Up oder Frisur, zählt nach der Definition dieser Arbeit in den Bereich der Robustheit eines homogenen Klassi-

fikators. Es handelt sich nicht um verschiedene Modalitäten im oben dargestellten Sinne. Auch das verwenden gering aufgelöster und gestörter Abbildungen zählt zur Robustheit.

Wirft man einen Blick auf den de facto Benchmark für klassische, homogene Gesichtserkennung, der LFW Datenbank, so wird man viele Einträge mit einer Gesichtsbreite unter 50 Pixeln finden. In dieser Arbeit wird auf Grund der formalen Definition „Übertrag- und Wiederverwendbarkeit der erlangten Informationen über Modalitätsgrenzen hinweg“ das Schaubild Abbildung 2.2 von (Ouyang et al., 2014) zu Abbildung 2.2 abgewandelt.

Diese Abbildung kann auch als Referenz für die gesamte Ausarbeitung herangezo-

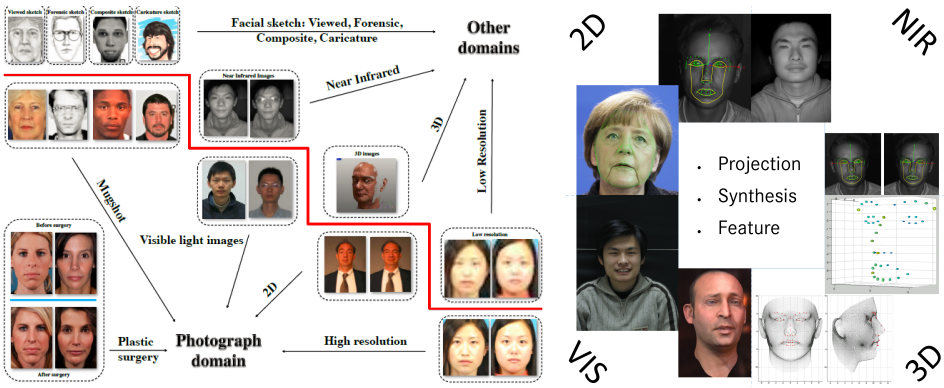

Abb. 2.2 Vergleich der Aufteilung von (Ouyang et al., 2014) und der hier verwendeten

gen werden, da alle dargestellten Modalitäten sukzessive abgehandelt werden. Weiterhin sind in diesem Schema drei Begriffe (Feature, Synthesis, Projection) zu finden, welche bislang nicht weiter erläutert wurden. In den nächsten Kapiteln werden Verfahren, welche diese Kategorien gruppiert werden können. Zunächst sollten die Begriffe jedoch im allgemeinen erläutert werden.

Feature bedeutet, dass alle vom Verfahren unterstützte Modalitäten die gleichen biometrischen Merkmale messen. Es ist dabei nicht zwingend notwendig, dass diese Merkmale direkt übertragbar sind. Eine statistische Umrechnung von 2D Gesichtslandmarken zu einer dreidimensionalen Repräsentation dergleichen, stellt ebenso ein Feature basiertes Verfahren dar, wie ein Abgleich zwischen nativen 3D-Gesichtslandmarken und einem Tiefenbild. Feature basierte Methoden sind aber nicht auf geometrische Merkmale beschränkt. Durchaus werden in Unterunterabschnitt 2.3.2.1 noch Systeme vorgestellt, welche auf einer photometrischen Invariante zwischen NIR und VIS basieren.

Synthesis ist der Oberbegriff für alle Vorgehensweisen die versuchen eine Modalität in eine andere zu transformieren. Beispielsweise wird in Unterunterabschnitt 2.3.3.1

beschrieben, wie aus Infrarotbildern einer Person ein korrespondierendes RGB Bild synthetisiert werden kann. Der besondere Reiz dieser Verfahrensgruppe ist die Verwendungsmöglichkeit eines homogenen Klassifikators. Ein Anwendungssystem kann so mit neuen Technologien wachsen ohne Kernfunktionalitäten weitreichend anpassen zu müssen. Durchaus könnte auch ein Aufbereiten von Bilddefekten, welche eine weitere Verarbeitung behindern, in diese Gruppe eingeordnet werden. Streng der hier eingeführten Definition von verschiedenen Modalitäten muss dieses jedoch der Robustheit und somit der homogenen Gesichtserkennung zugeordnet werden.
Als letztes bleibt nun die **Projection**, welche alle Systeme umschreibt, die versuchen identifizierende Merkmale aller unterstützten Modalitäten in einen gemeinsamen *Identifikationsraum* zu projizieren. Identifikationsraum bedeutet in diesem Zusammenhang derjenige Zahlenraum der als Basis zur Abstandsbestimmung zweier Messungen gebraucht wird (vgl. Unterunterabschnitt 2.4.1.2). Nicht zwangsweise müssen die einzelnen Modalitäten hierbei gesondert behandelt werden, wie es beispielsweise bei so genannten „Siamesichen Netzen“ (vgl. Absatz 3.4.0.0.1) der Fall ist. Das Trainieren eines Klassifikators mit NIR und VIS Daten (siehe Absatz 3.4.0.0.2) ist ebenso als Projektion anzusehen.
Nachfolgend werden nun Beispiele aus der aktuellen Literatur für alle genannten Kategorien vorgestellt. Im letzten Abschnitt 2.3.5 werden zu dem noch Kombinationsmöglichkeiten aller Kategorien behandelt.

2.3.2 Feature

Wie bereits zuvor ausführlich dargelegt handelt es sich bei Feature basierten Verfahren um Methoden welche ein gemeinsames Merkmal zur heterogenen Identifikation nutzen. Unterstrichen werden sollte hier noch einmal, dass dieses Merkmal auch abgeleitet, dass heißt nicht unmittelbar gemessen, sein kann.

2.3.2.1 Merkmal basierte NIR-VIS Verfahren

Durch aus gut funktionierende, weitverbreitete Verfahren zur Illuminationsvarianz, wie z. B. *Local Binary Pattern* (LBP) und *Histogram of oriented Gradients* (HOG) verlieren schnell an Güte, wenn anstatt Intensität und Richtung der Lichtquelle (homogene Varianz), das Spektrum (heterogene Varianz) verändert wird.
Zhu und Kollegen (Zhu et al., 2013) versuchen dieses Problem mit einer genauen

Analyse der zu Grunde liegenden physikalischen Vorgänge und einem darauf basierten invarianten *Deskriptor*, zu lösen. Das menschliche Gesicht wird hierbei als lambertsche Oberfläche, bei welcher alles einfallende Licht *diffus*, d. h. ungefähr proportional mit dem Winkel zur Oberflächennormale, reflektiert wird. Es ergibt sich die auch als *Cosine-Theta-Law* bekannte Formel: $\boldsymbol{I}(\Theta) \approx \boldsymbol{A} * \cos(\Theta)$. Ein grundlegendes Verständnis der Gegebenheiten von *Bidirectional Reflectance Distribution Functions* (BRDFs) vorausgesetzt, kann leicht festgestellt werden, dass die einzige, vom Spektrum der Lichtquelle abhängende, Variable $\boldsymbol{A}$, das Albedo ist. Das Albedo, im folgenden mit mit $\boldsymbol{A}(\lambda)$ abgekürzt, beschreibt materialabhängige Reflektanzeigenschaften und ist somit eine Funktion der Wellenlänge.
Da man sich in der Bildsynthese seit langem mit der realistischen Generierung von menschlichen Gesichtern beschäftigt, sind Analysen zur Annäherung von $\boldsymbol{A}(\lambda)$ zur Genüge vorhanden. Ein Beispiel für eine solche Näherung kann Abbildung 2.3 entnommen werden. Zhu und Kollegen machen sich diese Tatsache zur Nutze und igno-

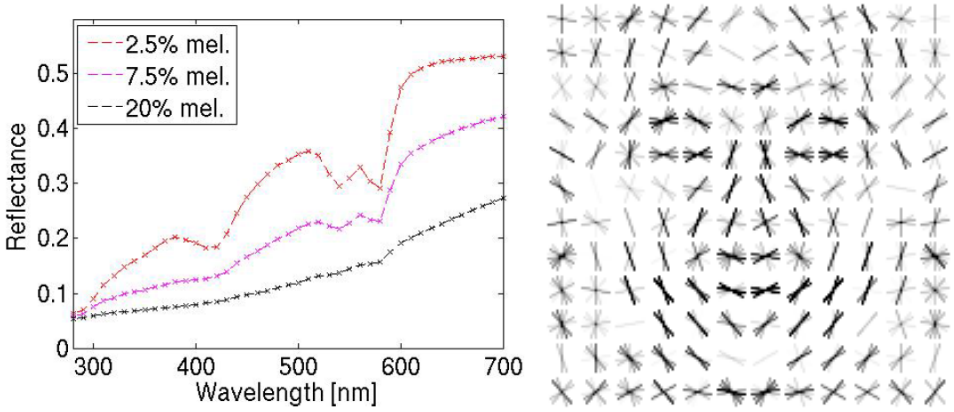

Abb. 2.3 Links Reflektanz menschlicher Haut abhängig von Wellenlänge und Melaninkonzentration (Hautpigment verantwortlich für Hautfarbe) (Zhu et al., 2013)
Rechts Exemplarische Visualisierung eines HOGs von einem menschlichen Gesicht (Zhu et al., 2013)

rieren die Abhängigkeit der Reflektanz von der Hautfarbe (vgl. Abbildung 2.3) und leisten darauf aufbauend folgende Beiträge: 1. den *Logarithmic Gradient Orientation* (LGO) und den *Logarithmic Gradient Magnitude* (LGM) 2. Eine robuste Kombination der beiden zuvor genannten Repräsentationen, sowie 3. eine tiefgehende physikalische Herleitung, welche als formale Systemverifikation herangezogen werden

kann. LGO und LGM basieren auf folgender Herleitung (Gleichung 2.3), bei welcher $\tilde{f}(x, y) = \log(f(x, y))$ gilt.

$$I(x, y) = R(x, y)L(x, y) \tag{2.1}$$
$$\tilde{I}(x, y) \approx \tilde{R}(x, y) + \tilde{L}(x, y) \tag{2.2}$$
$$\tilde{I}(x + \Delta x, y) = \tilde{R}(x + \Delta x, y) + \tilde{L}(x + \Delta x, y) \tag{2.3}$$

Die Intensität I ist das Produkt aus der Luminanz L und der Reflektanz R, jeweils an Position (x, y) innerhalb der Aufnahme. Weiterhin wird angenommen, dass man die logarithmische Intensität durch die Summe der logarithmischen Reflektanz und Luminanz annähern kann und das die Intensität stetig differenzierbar ist. Außerdem wird angenommen, dass die Differenz der Luminanz gegenüber der der Reflektanz verschwindend gering ist. Deshalb kann man Folgendes formulieren

$$\partial_x \tilde{I}(x, y) \approx \partial_x \tilde{R}(x, y) \tag{2.4}$$
$$\partial_y \tilde{I}(x, y) \approx \partial_y \tilde{R}(x, y) \tag{2.5}$$

Bedenkt man, dass die Reflektanz von Einfallsrichtung des Lichts abhängt, scheint die Annahme, dass die Änderung des ambienten Lichts gegenüber der der Reflektanz verschwindend gering ist, einleuchtend.
Mit dem Wissen, dass die Gradientorientierung(GO), sowie der Gradientbetrag(GM) wie folgt definiert ist:

$$GO(x, y) = \arctan\left(\frac{\partial_x I(x, y)}{\partial_y I(x, y)}\right) \tag{2.6}$$
$$GM(x, y) = \sqrt{(\partial_x I(x, y))^2 + (\partial_x I(x, y))^2} \tag{2.7}$$

Kann man durch die Approximation von I durch R lassen sich nun folgende, spektralinvariante Merkmale herleiten:

$$LGO(x, y) = \arctan\left(\frac{\frac{\partial_x R(x,y)}{R(x,y)}}{\frac{\partial_y R(x,y)}{R(x,y)}}\right) = \arctan\left(\frac{\partial_x \tilde{R}(x, y)}{\partial_y \tilde{R}(x, y)}\right) \tag{2.8}$$
$$LGM(x, y) = \frac{\sqrt{(\partial_x R(x, y))^2 + (\partial_x R(x, y))^2}}{R(x, y)} = \sqrt{(\partial_x \tilde{R}(x, y))^2 + (\partial_x \tilde{R}(x, y))^2} \tag{2.9}$$

Nimmt man nun das mit der Wellenlänge variierende Albedo wie folgt in die Gleichung $R_1(x, y) = A(\lambda)R_2(x, y)$ kann man durch einsetzen in Gleichung 2.9 zeigen

das eine Invarianz besteht.Ein Beispiel für eine derartig invariante Textur kann Abbildung 2.4 entnommen werden.
Es bleibt die Einbettung in einen gegen Rauschen robusten Rahmen, für welche die

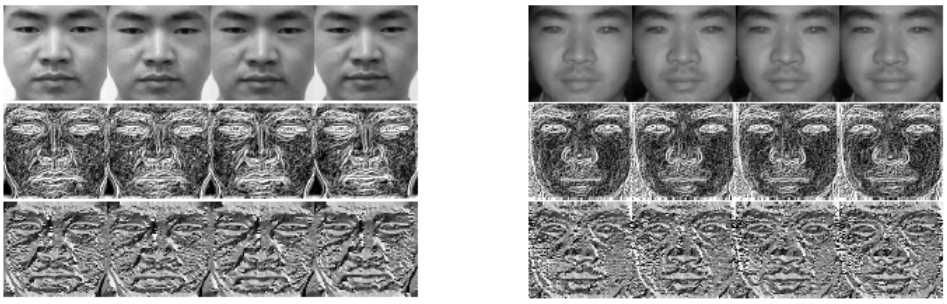

Abb. 2.4 Links VIS, rechts NIR, oben Textur, mitte LGO, unten LGM (Zhu et al., 2013)

Autoren (Zhu et al., 2013) auf einen Standard Histogram of oriented Gradients, oder in diesem Fall Histogram of oriented logarithmic Gradients zurückgreifen. Ein Beispiel dafür kann Abbildung 2.3 entnommen werden.
Zhu und Kollegen testen ihren Deskriptor auf einem viergeteilten HFB-Datensatz, was etwa 50 Personen pro Durchlauf zur Folge hat. Sie erreichen dabei Genauigkeiten über der 95%-Marke. Da es sich aber wie bereits erwähnt um sehr wenige Probanden handelt, sind die Ergebnisse fraglich. Auch die Vierteilung des Datensatzes ist unverständlich, da kein Lernen stattfindet dessen Generalisierungsverhalten analysiert werden müsste.

2.3.2.2 Merkmal basierte 2D-3D Verfahren

3D Verfahren haben den Reiz, deutlich akkurater als ihre 2D Pendants zu sein. Dieser Zuwachs an Genauigkeit kommt leider mit enormen Mehrkosten für Tiefensensoren und 3D Kameras. Eine Kombinationsmöglichkeit beider Modalitäten wäre sinnvoll um Genauigkeit und Kosten auf den jeweiligen Anwendungsfall zuschneiden zu können. In der Literatur wird dieser Fall *asymmetrische Gesichtserkennung* genannt, da meistens verschiedene Verfahren für den homogenen und heterogenen Abgleich verwendet werden.(Ouyang et al., 2014)
Ein bekanntes Merkmal für diese Kategorie ist die so genannte *Oriented Gradient Map* (OGM) von Huang und Kollegen.(Huang et al., 2012) Zuerst wird ein Deskriptor entwickelt, welcher sowohl auf das Tiefenbild als auch auf der Textur angewandt werden kann. Dabei orientiert man sich formal an der Arbeitsweise von Neuronen,

welche man für jede Pixelposition (x, y) wie in Gleichung 2.10 dargestellt, beschreibt. Dabei bezeichnet ∂o die partielle Ableitung in Richtung o (Orientierung) und $\boldsymbol{I}$ steht für ein Tiefen- oder Texturbild. Um nichtlineare Zusammenhänge zu modellieren, werden lediglich positive Neuronenantworten weiterverarbeitet. Negative Ergebnisse werden auf 0 gesetzt (vgl. RELU in Abschnitt 2.4.3) was in Gleichung 2.10 durch das $^+$ dargestellt wird.

$$\boldsymbol{G}_o = \left(\frac{\partial \boldsymbol{I}}{\partial o}\right)^+ \tag{2.10}$$

$$\rho_o^R = \boldsymbol{G}_R * \boldsymbol{G}_o \tag{2.11}$$

$$\rho^R(x, y) = \left[\rho_1^R(x, y), ..., \rho_o^R(x, y)\right]^t \tag{2.12}$$

$$\boldsymbol{J}_o(x, y) = \underline{\rho}_o^R(x, y) \tag{2.13}$$

Die Antwort ρ_o^R eines *komplexen Neurons*, dass heißt die Endreaktion einer Neuronenkette, in Richtung o ergibt sich aus der Faltung mit einem Gauss-Kernel $\boldsymbol{G}$, dessen Breite proportional zum Radius $\boldsymbol{R}$ ist (siehe Gleichung 2.11). Diese Faltung ist unbedingt notwendig um eine Gradientenglattheit über die einzelnen Eingangsregionen zu erzeugen. *Glattheit* in der englischen Optimierungsliteratur auch als (piecewise) *Smoothness* bezeichnet, beschreibt die kontinuierliche Änderung einer Variable. Insbesondere ist damit 1. Kontinuität und 2. eine geringe Norm der Hessematrix, also der 2. Ableitung, gemeint. Diese zwei Eigenschaften sind existentiell um eine korrekte mathematische Problemformulierung zu erreichen.
Die Resultate der einzelnen Richtungen $1 \ldots o$ lässt sich wie in Gleichung 2.12 zu einem Vektor ρ_p^R zusammenfassen. Die normalisierte Form dieses Deskriptors wird durch $\underline{\rho}_p^R$ bezeichnet. Die OGM an Position (x, y) wird dann mit $\boldsymbol{J}_o(x, y)$ (siehe Gleichung 2.13), als Anlehnung an die Jacobimatrix, bezeichnet. Ein Beispielbild für Textur und Tiefenbild, sowie deren acht OGMs, kann Abbildung 2.5 entnommen werden.
Auf dieser Basis wird nun die Erkennung durchgeführt. Wie bereits erwähnt nennt

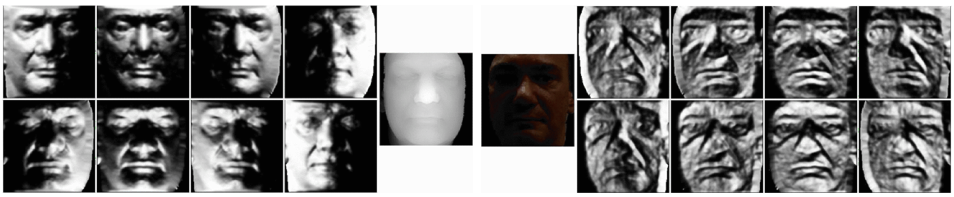

Abb. 2.5 Links Tiefenbild und daraus die daraus errechneten acht OGMs (Huang et al., 2012)
Rechts RGB-Bild und die daraus errechneten acht OGMs (Huang et al., 2012)

man diese Form der Gesichtserkennung auch asymmetrische Gesichtserkennung, da für den heterogenen (asymmetrischen) 3D-2D / 2D-3D Fall ein anderes Verfahren zum Zuge kommt, als im homogenen (symmetrischen), 2D-2D, Anwendungsfall. Im homogenen Fall wird ein *Local Binary Pattern*(LBP) (siehe Abschnitt 2.2.3) verwendet. Dafür wird das Gesicht wie bereits zu vor in (vgl. Unterunterabschnitt 2.3.3.1) erläutert, in mehrere Subregionen aufgeteilt. Die so errechneten LBPs der OGM werden dann paarweise mittels *Sparse Representation Classification*(SRC) verglichen. SRC basiert auf einem Näherungsverfahren, welches die Eingaben gegen regional begrenzte Verdeckungen robustifiziert. Es handelt sich um einen weitverbreiteten Ansatz für Mustererkennung, welcher in Abschnitt 2.2.2 bereits näher erläutert wurde. Das Residuum dieser Annährung wird dann als Abstandsmaß verwendet, wobei sich die Ähnlichkeit antiproportional zur L_2 Norm verhält. (Wright et al., 2009)
Interessanter ist jedoch der 2D-3D Fall, für welchen eine *Canocial Correlation Analysis*(CCA) durchgeführt wird. CCA maximiert die Korrelation zweier Variablenmengen entlang ihrer Hauptrichtung. Sie beruht wie auch die *Hauptkomponentenanalyse*(PCA) auf der Kovarianzmatrix. Das übertragbare Merkmal wird also aus zwei unabhängigen Messmethoden mit Hilfe der CCA erlernt, bevor es angewandt werden kann. Dieser Sachverhalt macht es wichtig das Generalisierungsverhalten, das heißt die Anwendbarkeit auf während der Trainingsphase nicht verfügbaren Daten, zu analysieren. Zunächst aber zur formalen Funktionsweise der CCA.
Nimmt man N Paare (x_i, y_i) aus (X, Y), $i = 1, \ldots, N$ an, wobei $X \in J_T$ (Textur) und $Y \in J_R$ (Range) gilt. Die maximale Korrelation in der jeweils anderen Modalität wird dann mit $x = w_x^T X$ sowie $y = w_y^T Y$ ermittelt. Deshalb nennt man diese auch *Canonische Varianten*. Sie ergeben sich aus dem Maximum der folgenden Funktion Gleichung 2.14. E steht dabei für für die statistische Erwartung.

$$\rho = \frac{E\left[w_x^T X Y^T w_y\right]}{\sqrt{E\left[w_x^T X X^T w_x\right] E\left[w_y^T Y Y^T w_y\right]}} \tag{2.14}$$

$$S(x', y') = \frac{x' y'}{||x'||\,||y'||} \tag{2.15}$$

Die Ähnlichkeit wird dann über Gleichung 2.15 ermittelt, wobei $x' = w_x^T X$ sowie $y' = w_y^T Y$ gilt. Ein höherer Wert bezeichnet eine höhere Ähnlichkeit.
Die Autoren geben eine Genauigkeit von 94.04 % im heterogenen Fall an, was leider nicht sehr aussagekräftig ist. Wie in Unterunterabschnitt 2.4.1.1 noch genauer beschrieben wird ist die Erkennungsrate immer in Abhängigkeit der Falscherkennungsrate zu sehen. Ein einzelner Wert macht es schwierig verschiedene Verfahren

in einer Evaluation gegenüber zu stellen. Weiterhin wird nicht auf das Generalisierungsverhalten eingegangen, d. h. die Evaluation hat auf den gleichen Daten wie das Training stattgefunden. Mit diesem Verfahren wäre es ein Einfaches gute Ergebnisse zu erzielen in dem man sich die Textur-Tiefenbildpaare pixelgenau speichert. Gerade lernbasierte Verfahren wie die hier besprochene CCA müssen daraufhin evaluiert werden. Die Autoren geben auch nicht an welche Datenbank verwendet wurde, oder wie viel Probanden und Bilder verfügbar waren. Eine Reproduktion und Validierung der Ergebnisse durch einen Dritten sind deshalb nicht möglich. Die erreichten 94.04% Genauigkeit sollten deshalb skeptisch betrachtet werden.
Hingegen genauer untersucht wurde die Auswirkung der Granularität der Quantizierung des Gradienten. Wie oben beschrieben wurden acht Orientierungen genutzt um die OGMs zu berechnen. Diese Quantizierung der Gradientenrichtung ist eine Robustifizierung gegen Rauschen und andere Störeinflüsse wie sie in vielen Methoden zu finden ist. Die bekanntesten sind die bereits vorgestellten SIFT und HOG Deskriptoren. Natürlich kommt diese Robustheit mit dem Preis des Präzisionsverlusts, weshalb eine Abwägung zwischen Präzision und Robustheit nur experimentell und anwendungsfallbezogen ermittelt werden kann. Die Autoren wenden deshalb verschiedene Konfigurationen an und stellen fest, dass mehr als acht Richtungen nicht nur zu keiner Verbesserung führen, sondern zu einem Verlust an Genauigkeit, proportional zur Anzahl der Richtungen. Ähnliches wurde für den Radius der Gaussglättung ermittelt.
Abschließend lässt sich zusammenfassen, dass die Autoren einen vielversprechenden Ansatz gefunden haben. Freie Parameter der Methode wurden umfangreich analysiert und optimiert. Einzig eine Validierung ihres Ansatzes bleiben sie schuldig. (Huang et al., 2012)

2.3.3 Synthesis

Wie bereits erwähnt beruht ein Synthese-basierter Ansatz auf der Rekonstruktion der jeweils anderen Modalität. Die meisten Methoden sind in ihrem Anwendungsfall symmetrisch, dass heißt sie können sowohl aus Modalität eins Modalität zwei generieren, als auch andersherum. In den meisten Fällen von NIR-VIS-Transformationen sogar ähnlich präzise. Natürlich ist es rein mathematisch und logisch gesehen deutlich einfacher aus 3D Punkten 2D Feature zu erzeugen, dann würde es sich allerdings um einen Feature basierten Ansatz handeln. Betrachtet man den Tiefenbild-Textur-Fall stellt sich die Sache deutlich anders dar. Während bei den 3D/2D Punk-

ten das gleiche Merkmal mit unterschiedlicher Sensorik gemessen wird, werden im Tiefen/Textur-Beispiel verschiedene Modalitäten synthetisiert. Synthese wird in diesem Abschnitt behandelt.
Das Verstehen dieses Unterschieds ist existentiell für die Sinnhaftigkeit der Kategorisierung. Sollte dies bis hierhin nicht klar gewesen sein, sollte man Abschnitt 2.3.1 erneut konsultieren.

2.3.3.1 Synthese basierte NIR-VIS Verfahren

Für heterogene Gesichtserkennung in den Modalitäten NIR und VIS ist der Synthese-basierte Ansatz der am weitesten verbreitete. Zunächst sollte man sich der Problemstellung genau bewusst werden und sie von der Colorierung eines Schwarzweiß Bildes unterscheiden. Im Falle der der Einfärbung sind nur zwei Parameter zu bestimmen, da die Luminanz eines Farbbildes der Intensität eines SW-Bildes entspricht. Dass heißt durch einfache Transformation in einen anderen Farbraum, hat man bereits eine Konstante und nur noch zwei Variablen. Im NIR-VIS-Fall handelt es sich jedoch um unterschiedliche Spektren und wie bereits in Unterunterabschnitt 2.3.2.1 erklärt wurde, ändert sich die Reflektanz eines Materials (Albedo) mit der Wellenlänge des einfallenden Lichts. Die Intensität eines NIR Bildes korrespondiert also nicht mit der Luminanz eines VIS Bildes. Der zweite Punkt ist die Masse und Qualität der Trainingsdaten. Während man jedes beliebige Farbbild in Schwarzweiß transformieren kann und somit über beliebig viele Bildpaare verfügt muss NIR-Bilder mit einer gesonderten Sensorik aufgenommen werden, was zu einem weiteren Problem führt. Im Falle der Einfärbung von SW-Bildern korrespondiert nicht nur der Luminanzkanal sondern auch Pixel. In dieser Problemstellung herrscht also pixelweise, im Fachjargon auch *spatial* genannte, Überlappung. Im NIR-VIS Fall müssen aufnahmebedingte Verzerrungen, sowie Unterschiede im Gesichtsausdruck und Kopfposition, berücksichtigt werden (siehe Abbildung 2.6). Die Trainingsdaten müssen also vorverarbeitet und optimiert werden. Man kann sie nicht ungefiltert als *Ground truth* verwenden. Zusätzlich muss bedacht werden, dass Datenbanken nur Bilder einer bestimmten Wellenlänge enthalten und nicht uneingeschränkt zum Training geeignet sind, falls die Wellenlänge der Aufnahmeeinrichtung des Anwendungssystem von der der Datenbank abweicht. Der Nahe-Infrarot Bereich ist ein Spektrum und keine exakte Wellenlänge.
Mathematisch gesehen kann man sagen dass bei einer Einfärbung lediglich die Funktion $\boldsymbol{SW \rightarrow RGB}$ approximiert werden muss währenddessen $\boldsymbol{RGB \rightarrow SW}$

trivial ist. Weiterhin gilt $\frac{R+G+B}{3} = SW$, wohingegen $NIR \rightarrow VIS$ und $VIS \rightarrow NIR$ beide unbekannt sind und es keinen trivialen mathematischen Zusammenhang gibt.

Aus diesem Grund sind die meisten wissenschaftlichen Arbeiten zum Thema NIR-

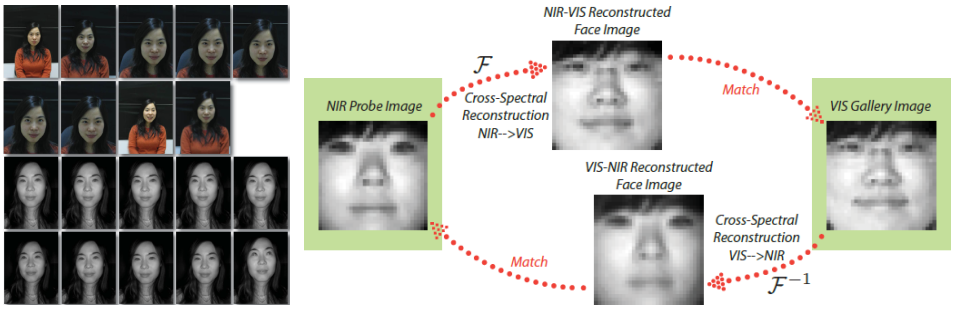

Abb. 2.6 Links Varianz innerhalb eines Subjekts (Li et al., 2013)
Rechts Mathematische Beschreibung von NIR-VIS Synthese (Juefei-Xu et al., 2015)

VIS Transformation und NIR-VIS Gesichtserkennung zweigeteilt. Zum einen ein Verfahren um aus vorhandenen Datenbanken verwertbare Bildpaare zu extrahieren die als Ground truth dienen und zum zweiten aus einer Methode um von einer Modalität in die andere zu synthetisieren.

Ein naheliegender Ansatz von (Juefei-Xu et al., 2015), basiert auf dem Erlernen eines NIR-VIS Wörterbuchs. Dabei werden die Bildpaare der CASIA NIR-VIS 2.0 Datenbank von (Li et al., 2013) in Subregionen, auch *Patches* genannt, zerschnitten. Das Farbbild wird dann in den *YCbCr-Farbraum* transformiert. Y steht hierbei für Luminanz (Helligkeit), Cb für den Blau-Gelb Kanal und Cr für den Rot-Grün Anteil. Der Unterschied zum klassischen RGB besteht in der Entkopplung von Farbe und Helligkeit.

Die so erhaltenen Patches werden anhand ihrer Helligkeit und mit Hilfe der Kreuzkorrelation affin aufeinander registriert. Ist die Korrelation der Patches und der zugehörige Gradient (erste Ableitung des Patches) über einen gewissen Schwellwert, wird das Bildpaar in das Wörterbuch aufgenommen. Das so generierte Wörterbuch kann sowohl für die NIR-VIS-Transformation als auch für die VIS-NIR-Richtung herangezogen werden. Das gewünschte Bild wird in Subregionen zerteilt und die daraus gewonnen Patches anhand des Wörterbuchs ersetzt. Man könnte diese Art von Transformation auch Bildsubstitution nennen. Selbstverständlich ist kann nicht für jedes Eingangspatch oder in unserer Metapher gesprochen *Wort* ein exakt passendes Pendant im Wörterbuch gefunden werden. Ist die Trainingsdatenbank, aus welcher das Wörterbuch generiert wurde, jedoch hinreichend groß, sollte zu jedem Patch ein hinreichend ähnlicher Eintrag im Wörterbuch gefunden werden. Eine Histogram-

normalisierung und eine Bereinigung von Umwelteinflüssen wie Rauschen oder Beleuchtung verbessert das Ergebnis selbstverständlich weiter. Ein Restfehler in der Rekonstruktion bleibt.

Der darauf aufbauende Gesichtsklassifikator muss deswegen hinreichend robust gegen diese Art von Ungenauigkeit sein. (Juefei-Xu et al., 2015) errechnen zu erst die durchschnittliche Abweichung ihrer Rekonstruktion von Ground-Truth um ein quantitatives Maß der benötigten Robustheit zu erlangen. Auf dieser Erkenntnis werden dann verschiedene Verfahren angewandt. Zu den in der Evaluation von (Juefei-Xu et al., 2015) erfolgreichsten Methoden gehören das bereits mehrfach erwähnt und erläuterte LBP sowie dessen Erweiterung *Discrete Cosine Transform Local Binary Pattern*(DLPB). Als kurze Erinnerung reicht es aus zu wissen, dass LBPs und die dazugehörigen Erweiterungen auf dem Skalenraum *scale space* basieren und ein Histogramm ermittelt wird, welches durch geeignete Normalisierung und Quantisierung robust gegen Rauschen ist.

DLBP, also ein LBP welches auf der Discrete Cosine Transform aufbaut ist deshalb robuster, da die DCT, welche auch in der ursprünglichen Form der JPEG-Kompression benutzt wurde, da das Bild zuerst approximiert wird. Rauschen, in diesem Fall durch Wörterbuchtransformation verursacht, wird damit, wie in allen spärlichen Approximationen (Sparsity-Ansatz), unterdrückt. Die *Rauschunterdrückung* lässt sich hierbei dynamisch, ähnlich der Kompressionsrate im JPEG Verfahren, anpassen. Man sollte allerdings anmerken, dass die so erreichte Robustheit mit dem Preis der geringeren Unterscheidbarkeit erkauft wird. Eine Abwägung beider Seiten muss für jeden Anwendungsfall erfolgen.

(Juefei-Xu et al., 2015) können mit diesem Verfahren dem Anschein nach sehr gute Resultate erzeugen. In ihrer Versuchsbeschreibung findet sich jedoch der Hinweis, dass *alle Bilder sowohl zum Training als auch zum Testen verwendet wurden*. (Juefei-Xu et al., 2015) Es wird der Eindruck erweckt, dass das alle Testbilder auch im Wörterbuch vorhanden sind, was eine Transformation trivial werden lässt und die exzellenten Ergebnisse erklärt. Weiterhin sollte bedacht werden, dass ein hinreichend großes Wörterbuch, dass heißt ein Wörterbuch welches in der Lage ist menschliche Gesichter aller Ethnien zu übertragen, äußerst groß wäre. Eine derart große Datenbasis ist schwer zu ermitteln und auch die Transformation eines Bildes würde durch die Verlängerung der Suchzeiten innerhalb des Wörterbuchs exorbitant viel Zeit in Anspruch nehmen.

Ein anderer Ansatz, der durchaus auch auf der Registrierung von Bildpatches basiert ist die NIR-VIS Halluzination von (Lezama et al., 2016). (Lezama et al., 2016) teilt die Datenbank von (Li et al., 2013) in vier gleiche Teile auf und evaluiert seine Methode

anhand der weitverbreiteten Technik der K-Cross-Validation, wobei K in diesem Fall 4 ist. Im speziellen heißt das, dass drei Teile zum Trainieren und ein Teil zum Validieren verwendet wird. Das Verfahren wird vier mal angewandt. Das hat den Vorteil, dass man keine Bilder der Datenbank *verschwendet* und die Varianz der Ergebnisse eine Aussage über die Generalisierungsgüte zulässt.
Zunächst jedoch zur Methode selbst: In einem ersten Schritt wird das gesamte Bild auf den Mittelwert aller in der Datenbasis enthaltenen Gesichtslandmarken transformiert. Die Autoren argumentieren, dass Transformationen mit mehr Freiheitsgraden als affin, der Registrierungsgüte nicht zuträglich sind. Die entstehenden Bildverzerrungen seien zu groß. (Lezama et al., 2016)

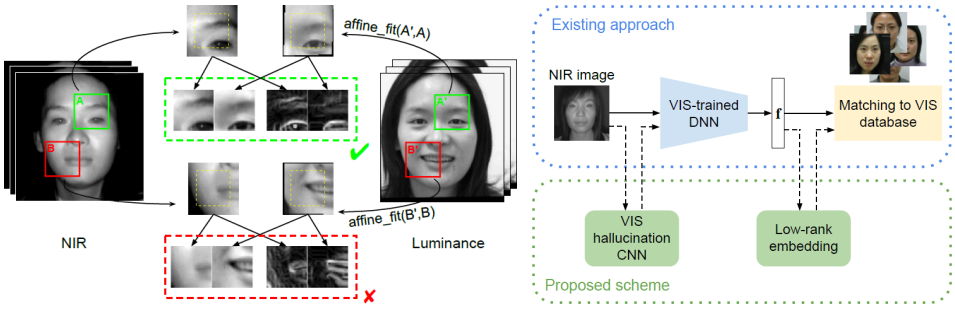

Abb. 2.7 Links Ablauf der Registrierung (Lezama et al., 2016)
Rechts Schematische Idee von (Lezama et al., 2016)
labelfig:patchesSchema

In einem zweiten Schritt wird dann das Bild in 64×64 Pixeln zerteilt. Diese Patches werden derart auf einander registriert, dass ein Maximum der Korrelation der Intensität, der Gradientenrichtung und der Gradientenmagnitude erreicht wird. Ist einer der drei Faktoren unter 0.5 wird das Bildpaar verworfen. (Lezama et al., 2016) konnten damit etwa 600.000 Patches erhalten. Diese werden dann auf eine eine Größe von 40×40 geschnitten um durch die Registrierung entstandene, nicht definierte Bildbereiche zu entfernen (siehe Abbildung 2.7 obere Reihe rechts).
Diese werden dann in einem dritten Schritt normalisiert bevor sie zum Training eines Encoder-Decoder Netzes (siehe Unterunterabschnitt 2.4.2.1 zum Einsatz kommen. Für jeden Bildkanal wird ein separates Netz verwendet, so dass die Luminanz durch ein Netz mit 11 Layern approximiert wird, wohingegen für die Farbinformation Cb und Cr, 5 Layer ausreichen.
Leider hat die Arbeit von (Lezama et al., 2016) auch einige Schwächen. Zum Beispiel werden zwei Klassifikatoren verwendet um zu zeigen, wie die Halluzination die Erkennung gegenüber dem direkten NIR-VIS Abgleich verbessert. Einer der Klassifikatoren, VGG-Face von (Parkhi et al., 2015), ist frei zugänglich was eine Reproduktion der Ergebnisse ermöglicht. Der zweite Klassifikator, welcher in der Evaluation signi-

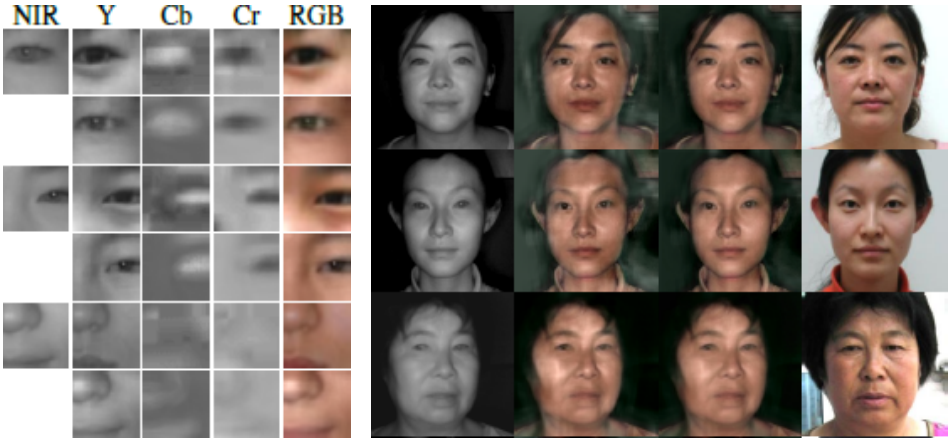

Abb. 2.8 Links YCbCr-Farbraum und RGB Rekonstruktion (Lezama et al., 2016) Rechts Von links nach rechts: NIR Eingangsbild, VIS-Hallzuination, VIS-Halluzination und Post-Processing, VIS-Bild (Lezama et al., 2016)

fikant besser als VGG-Face abschneidet wird jedoch nicht einmal benannt, sondern mit COTS (Commercial Of The Shelf) abgekürzt. (Lezama et al., 2016) Ein Nachprüfen dieser Ergebnisse ist somit nicht möglich. Weiterhin ist die Beschreibung des verwendeten CNNs dürftig. Die beschriebene Lernrate und Trainingszeit sind trotz höherer Rechenkapazität nicht ein mal annähernd nachzuvollziehen. Die angegebene Genauigkeit von 83% erscheint daher fragwürdig. Auch ROC-Kurven bleiben die Autoren schuldig. Eine weitergehende Analyse der Ergebnisse, kann Unterunterabschnitt 3.3.2.2 entnommen werden. (Lezama et al., 2016)
Abschließend kann gesagt werden, dass Halluzination ein vielversprechendes Gebiet der Synthese-basierten heterogenen Gesichtserkennung ist. Um einem Overfitting auf Trainingsdaten entgegenzuwirken ist es unablässig die Verfahren zu validieren. Die Aussagekraft von in der Literatur gefundenen Genauigkeiten muss in Relation zur Validierung gesehen werden. Ein Thema, welches von vielen Autoren vernachlässigt wird. Stellt man sich vor, dass man für das einmalige Trainieren eine Woche Rechen- und Arbeitszeit benötigt ist es verständlich, dass dieses ungern mehrmals durchgeführt wird. Es handelt sich hierbei jedoch um eine unabdingbare Notwendigkeit für verlässliche Ergebnisse.

2.3.3.2 Synthese basierte 2D-3D Verfahren

Die Synthese wurde noch nicht auf 2D-3D Verfahren angewandt, was jedoch nicht heißt, dass dieses nicht möglich wäre. (Ouyang et al., 2014) Zum Beispiel könnte man, ähnlich der NIR-VIS Halluzination, ein CNN mit Tiefenbild-Textur (RGB→Depth) trai-

nieren und der Annahme verfallen, dass aus einem Tiefenbild Texturdaten abgleitet werden können. Schaut man sich jüngste Ergebnisse des *Deep Learnings* an, mit welchem es möglich sein soll, aus einem Bild von 64×64 Pixeln, ein vollständiges Gesicht zu rekonstruieren. Aber wie in Abschnitt 2.4 berichtet, kann dieses rein mathematisch schon nicht möglich sein. Die Antwort eines CNNs ist immer die wahrscheinlichste gemessen an seinen Trainingsdaten.
Es läuft also darauf hinaus, ob genug statistische Korrelation zwischen einem Tiefen und einem Texturbild besteht. Gibt es statistische Korrelationen zwischen Kieferform und Hautfarbe? Ja, aber reicht diese aus um zwei Menschen zu unterscheiden? Ist die Augenfarbe abhängig von deren Form? Vermutlich nicht, was auch der Grund sein könnte warum es bisher noch keine Verfahren in dieser Kategorie gibt.
Man kann natürlich abseits von dem Tiefenbild-Textur Szenario argumentieren, dass eine Aufbereitung von 2D Gesichtslandmarken zu 3D Masken, wie es bei den in Abschnitt 2.2 vorgestellten Active Shape und Morphable Models getan wird, zu den Synthese-Verfahren gehört. Es handelt sich bei dieser Art jedoch um ein Anpassen von 3D Informationen, so dass diese den gemessenen 2D Informationen entsprechen. Weiterhin stellen sowohl 2D als auch 3D-Gesichtslandmarken die gleichen Merkmale dar, so dass sie den Feature-Verfahren zugeordnet werden müssen.
Auch eine Extraktion von Oberflächennormalen im Texturbild durch *Shape-From-Shading* und Vergleich mit der Ableitung(Oberflächennormale) des Tiefenbilds stellt ein Feature-Verfahren dar, da es sich um die gleiche Eigenschaft handelt.
Abschließend lässt sich sagen, dass die Kategorie der Synthese-2D/3D-Verfahren so gut wie nicht erforscht ist, dieses aber auch kein gravierendes Problem darstellt. Heterogene Gesichtserkennung im 2D/3D-Bereich ist bereits gut durch die Feature-Verfahren abgedeckt und funktionsfähig.

2.3.4 Projection

Die Kategorie der Projektions-Verfahren reicht von der Anwendung eines homogenen Klassifikators auf heterogene Daten (umgs. *einfach mal Deep Learning*) über *(pseudo)siamesische Netze*(separate Netze, welche sich ihren Definitionsraum teilen) bis hin zu eigens entwickelten Verfahren, die darauf aufbauen, dass eine mehr oder minder große *Heterokomponente* zwischen den Modalitäten existiert. Diese Verfahren, nehmen wir der Einfachheit halber eine PCA an, projizieren zunächst beide Modalitäten separat, bevor sie in einem zweiten Schritt versuchen die Differenz bei-

der Modalitäten (Heterokomponente) zu minimieren. Nachfolgend werden mehrere dieser Verfahren vorgestellt.

2.3.4.1 Projektions basierte NIR-VIS Verfahren

Zunächst sollte vertieft werden, warum ein einfacher Deep Learning Ansatz nicht möglich ist, bevor mehrere Verfahren der Mustererkennung erläutert werden: Selbstverständlich wäre es möglich, ein Neuronales Netz zu trainieren, welches Unterscheidungskriterien vollständig selbst lernt, anstatt diese vorzugeben. Hierzu wäre allerdings, ein enorm großer, qualitativ guter Datensatz von Nöten. Dieser fehlt bislang. Die Schwächen der derzeit größten Datenbasis, der Casia NIR-VIS 2.0 wurden bereits ausführlich erläutert. (Li et al., 2013) Eine triviale Anwendung eines klassischen Klassifikators, ist deswegen vermutlich nicht von Erfolg gekrönt.
Ein Ansatz, welcher für heterogene Gesichtserkennung geeignet ist und nicht auf einem neuronalen Netz basiert, ist die sogenannte *Common Discrimant Feature Extraction*(CDFE) von (Lin und Tang, 2006). (Lin und Tang, 2006) Versuchen eine allgemeine Problemlösung in dem sie zwei Modalitäten in einen gemeinsamen Identifikationsraum projiziert. Die Komplexität der erlernten Funktion hat unmittelbaren Einfluss auf das Generalisierungsrisiko. Ein komplexes Modell neigt zur Überparametrisierung (engl. Overfitting) der Trainingsdaten. Das führt dazu, dass nicht im Trainingsdatensatz enthaltene Bilder nicht korrekt abgebildet werden. (Lin und Tang, 2006)
(Lin und Tang, 2006) definieren deshalb die Einschränkung, dass die Transformation stetig und glatt sein muss. Es wird argumentiert, dass damit das Überparametrisierungsproblem entschärft und lokale Strukturen erhalten werden. Das formulierte Problem, eine *Query-Modality* und eine *Reference-Modality* in einen gemeinsamen Identifikationsraum zu transformieren wird mittels Eigenwertzerlegung optimiert. Zwei Bedingungen werden hierbei für die Optimierung herangezogen:

1. Die *Intra-Class-Compactness*, dass heißt der Raum den Instanzen der gleichen Identität aufspannen soll möglichst gering sein (Abbildung aller Instanzen einer Identität auf einen Punkt).

$$id(f_1) = id(f_2) \rightarrow \min |f_1 - f_2| \tag{2.16}$$

2. Die *Inter-Class-Dispersion*, dass heißt der Abstand der Instanzen verschiedener Identitäten sollte möglichst groß sein.

$$id(f_1) = id(f_2) \rightarrow \max |f_1 - f_2| \tag{2.17}$$

Diese Anforderungen können nur von einer nicht linearen Transformation erfüllt werden, da eine lineare Transformation nicht ausreicht komplexe Überlappungen einzelner Instanzen zu trennen. Um dies zu ermöglichen wird eine Kernelization, ähnlich der einer Support-Vektor-Maschine, auf die einzelnen Modalitäten angewandt. (Lin und Tang, 2006)
Verschiedene mathematische Verbesserungen dieses Ansatzes sind bekannt. Die hier vorgestellte Basis erreicht Genauigkeiten um die 90% in der Evaluation von (Lin und Tang, 2006). Angesichts der Tatsache, dass die vorgestellte Probandenstudie lediglich 16 Personen mit insgesamt 64 Bildern beinhaltet sind die Ergebnisse kritisch zu betrachten.
Die Funktionsweise dieses Ansatzes ist jedoch bezeichnet und findet sich in im Grundsatz in vielen Methoden wieder. Zum Beispiel werden bei siamesischen Netzen zwei separate neuronale Netze (eins für jede Modalität) auf einen gemeinsamen Identifikationsraum verknüpft. Der Lernprozess wird dahingehend angepasst, dass je nachdem ob das Netz gerade die gleiche Person in unterschiedlichen Modalitäten sieht, der resultierende Fehler mit Lernfaktor 1 oder -1 multipliziert wird. So erreicht man, dass beide Netze, unterschiedliche Modalitäten einer Person auf den gleichen Punkt im Identifikationsraum projizieren.
Im Falle zweier unterschiedlicher Personen wird der Abstand maximiert (Faktor -1) im Falle der der gleichen Person minimiert (Faktor 1). Genauere Beschreibungen siamesischer Netze sind in Absatz 3.4.0.0.1 zu finden.

2.3.4.2 Projektions basierte 2D-3D Verfahren

Projektions basierte Ansätze für die 2D-3D Gesichtserkennung sind bisher nicht bekannt. Methoden, welche sowohl aus aus dem Tiefenbild, als auch aus der 2D Aufnahme, Oberflächennormalen extrahieren und statistisch analysieren, wie zum Beispiel (Rama et al., 2006), sind eher den Merkmal basierten Verfahren zuzuschreiben. Die eigentliche Identifikation über einen gemeinsame Eigenschaft erfolgt, welche in beiden Modalitäten messbar ist. (Ouyang et al., 2014) Die statistische Aufarbeitung der

gewonnenen Information reicht nicht aus um zu den Projektionsverfahren gezählt zu werden.
Weiterhin wird die *Canonical Correlation Analysis* gerne zu den Projektionsverfahren gezählt. Da es sich jedoch um die Rekonstruktion der gewünschten Modalität mit Hilfe von korrelierenden Merkmalen handelt, ist CCA, wie bereits zuvor vorgestellt, der Synthese zuzuordnen.

2.3.5 Hybride Verfahren

Nach der zuvor erläuterten Einteilung sind natürlich viele Grenzfälle denkbar, bei welchen ein Verfahren zu zwei Kategorien gehören könnte. Eine genauere Analyse der Methode gibt meist Aufschluss in welchem Schritt und mit welcher Idee das Verfahren arbeitet und wie die Identifikation zu bewerten ist. Dieser Abschnitt beschäftigt sich mit der Idee, Methoden verschiedener Kategorien zu kombinieren. Zum Beispiel ist es denkbar eine NIR-Aufnahme durch ein Syntheseverfahren in ein VIS-Bild zu transformieren und anschließend ein Projektionsverfahren zur Identifikation zu nutzen, welches lediglich auf VIS und Synthese-VIS Bildern trainiert wurde. Ein derartiger Ansatz würde äußerst robust sein und auch optimal mit den Eigenheiten von synthetisierten Bildern funktionieren, das Kernproblem der heterogenen Gesichtserkennung aber nicht abmildern. Das Kernproblem ist die geringe Datenverfügbarkeit. Die größte NIR-VIS Datenbank umfasst gerade einmal 735 Personen ohne jegliche ethnische Vielfalt. Legt man zu Grunde, dass es sich in beiden Fällen um lernende Methoden handelt, welche validiert werden sollten, folgt, dass nicht genug Bilder verfügbar sind. Eine etwaige Steigerung der Präzision wäre lediglich die Folge von Überparametrisierung. (Li et al., 2013) (Ouyang et al., 2014)
Gänzlich falsch wäre die Annahme, man könnte die Genauigkeit der kombinierten Verfahren in etwa multiplizieren, da diese voneinander unabhängig sind. Der Vorteil dieser Art ist, wie man auch in Abschnitt 4.2 sehen wird, das die automatisch ausgewählten, identifizierenden Merkmale, exakt denen entsprechen welche auch von der Synthese erzeugt werden. In der Synthese wird im Grunde eine Abbildung generiert, welche pixelbasiert den geringsten Abstand gemessenen an einer bestimmten mathematischen Norm (meist $\boldsymbol{L}_2$ oder $\boldsymbol{L}_1$) aufweist. Wie man jedoch in Unterunterabschnitt 3.3.3.1 sehen wird, ist der mathematisch geringste Abstand zweier Bilder nicht gleichbedeutend mit dem geringsten Abstand aus Identifikationssicht. Es ist zum Beispiel möglich, dass von einem rein visuellen Standpunkt aus, eine Synthese deutlich erfolgreicher aussieht als die Andere, in der Klassifikation jedoch

signifikant schlechter abschneidet. Das liegt daran, dass in der Trainingsphase des Syntheseverfahrens nicht der Fehler der Klassifikation zugrunde gelegt wird. Ein vielversprechender Ansatz für ein hybrides Verfahren wäre zum Beispiel eine Kombination des visuellen Fehlers (pixelweise $\boldsymbol{L}_2$-Norm) und des Identifikationsfehlers. Diesem Ansatz ist lediglich entgegenzusetzen, dass für jede Trainingsiteration eine vollständige Identifikation durchgeführt werden müsste, was die Trainingszeit extrem verlängert. Nimmt man an, dass eine Identifikation durch eine neuronales Netz an die 100ms in Anspruch nehmen kann und das ein Training aus mehrere Millionen Iterationen besteht, merkt man schnell, dass diese Idee nicht umsetzbar ist. (Ouyang et al., 2014)

2.4 Convolutional Neural Networks

Convolutional Neural Networks - zu deutsch also etwa „faltende neuronale Netze“, haben im letzten Jahrzehnt zunehmend an Bedeutung gewonnen. Die weitreichenden Anwendungsbereiche von Sprachanalyse bis hin zur Bildklassifikation - für welche diese Art Netz im übrigen entwickelt wurde, werden lediglich durch die enorme Rechenkraft, die benötigt wird um ein solches Netz zu trainieren, beschränkt. Zweifelsohne ist das auch der Grund, weshalb es fast ein halbes Jahrhundert gedauert hat, bis sich die initiale Idee eines „receptive Fields“ [Wisel 68] zu funktionieren Bildklassifikatoren mit Genauigkeiten jenseits der 99% weiterentwickelt hat.
Die ursprüngliche Idee von Wisel et al. ist ebenso interessant wie die hohen Genauigkeiten und vielfältigen Anwendungsgebiete. Wie häufig, hat man sich von biologischen Prozessen, in diesem Falle, dem menschlichen, visuellen Cortex, inspirieren lassen. Es ist weitläufig bekannt, dass die visuelle Wahrnehmung ein preattentiver Vorgang ist und somit keinerlei kognitiven Aufwand voraussetzt. Die menschliche Bildauswertung basiert also zu einem großen Teil auf massiv parallelen und hinreichend einfach mathematisch zu beschreibenden Neuronenvorgängen. Um ein Bild zu verstehen oder in unserem Falle ein Gesicht zu analysieren ist es folglich nicht notwendig komplexe Folgerungen durchzuführen. Weiterhin zeigt sich, dass eine verlustfreie Beschreibung von Gesichtern möglich ist ohne diskriminative Eigenschaften zu verlieren. Diese Tatsache ist für eine Umsetzung existentiell. Man bedenke nur es wäre nicht möglich und ein klassisches neuronales Netz, was sagen wir aus ein dutzend „Hiddenlayer“, d. h. Schichten zwischen den Ein- und Ausgängen besteht würde ein Farbbild mit der Auflösung 256×256×3, elementweise, also eine Neuronenkette per Pixel verarbeiten: Es wäre notwendig fast 2 mio. Parameter

zu approximieren.
Diese Gegebenheit sowie der Erfolg von CNNs auf dem Gebiet der Gesichtsanalyse, welche in den letzten Jahren sogar die menschliche Wahrnehmung übertroffen hat, macht es notwendig sich tiefer gehend über die Funktionsweise, dieser oft als „Black Box“ verwendeten Systeme zu informieren, um eine detailliertere Sicht auf deren Möglichkeiten und Einschränkungen zu erlangen. Die Tatsache, dass Begriffe wie „Big Data“ und „Deep Learning“ als heiliger Gral und Allheilmittel für jedes offene Problem der Wissenschaft gesehen werden, macht ein genauere Beleuchtung geradezu unabdingbar um die Problematik der heterogenen Gesichtserkennung akkurat zu vermitteln.

2.4.1 Schematische Funktionsweise

Die Kernoperation eines jeden Convolutional Neural Nets ist wie der Name bereits vermuten lässt: eine Faltung. Die mathematische Ergründung des Faltungstheorem sowie das Konzept der Fourier-Analyse wird an dieser Stelle ausgespart um ein abstrakteres Verständnis der Vorgänge in den Vordergrund zu rücken.
Wir definieren eine Bildfaltung als „schrittweise“ Abtastung eines Bildes mit einer Matrix, die sich Kernel nennt. Die Schrittweite bezeichnet man hierbei als Stride. Weiterhin sollte an dieser Stelle erwähnt werden, dass eine Faltung prinzipiell eine Linearkombination und somit eine affine Transformation darstellt. Das Abbilden von nichtlinearen Zusammenhängen wird durch verschiedene Gewichtungen erreicht, welche in Abschnitt 2.4.3 näher behandelt werden.
Betrachten wir nachfolgend Tabelle 2.1 zunächst verschiedene Konfigurationen von Stride und Kernel um einen besseren Eindruck von der elementaren Operation der Netze zu erlangen:

Es ist einfach zu erkennen, dass je nach Art des Kernels z. B. Kanten verstärkt oder abgeschwächt werden. Leser welche mit der Fourieranalyse vertraut sind werden unweigerlich feststellen, dass es sich hier um verschiedene Frequenzfilterungen handelt. Der Stride bestimmt Höhe und Breite des Ergebnisses, da bei einer Schrittweite von 2 nur jeder zweite Pixel verwendet wird, ist das Ergebnis auch nur ein Viertel so groß. Weiterhin ist zu entdecken, dass das Ergebnis einer Faltung wiederum ein „Bild“ im Sinne der Darstellung Breite×Höhe×Tiefe ist. Tiefe ist hierbei die Anzahl der Kanäle, das heißt für ein Farbbild drei und für ein Graubild eins. Es ist also vielmehr ein Volumen als ein zweidimensionales Bild im gebräuchlichen Sinne, wie

Name, Stride	Kernel	Ergebnis
Identität, st: 2	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	
Binomial, st: 1	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	
Gauss, st: 1	$\frac{1}{256}\begin{pmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{pmatrix}$	
Unsharp Masking, st: 1	$\frac{-1}{256}\begin{pmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & -476 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{pmatrix}$	

Tab. 2.1 Verschiedene Konfigurationen von Kernel und Stride

Abbildung 2.9 auch noch einmal veranschaulicht. Die Tiefe des Volumens kann sich während der Verarbeitung ändern, dann nämlich, wenn ein Filter ein Bild mehrmals abtastet und in diesem Zuge neue Kanäle entstehen. Dabei handelt es sich keineswegs um reine Redundanz, da die Faltungsresultate noch einmal gewichtet werden, was in Abschnitt 2.4.3 näher erläutert wird.

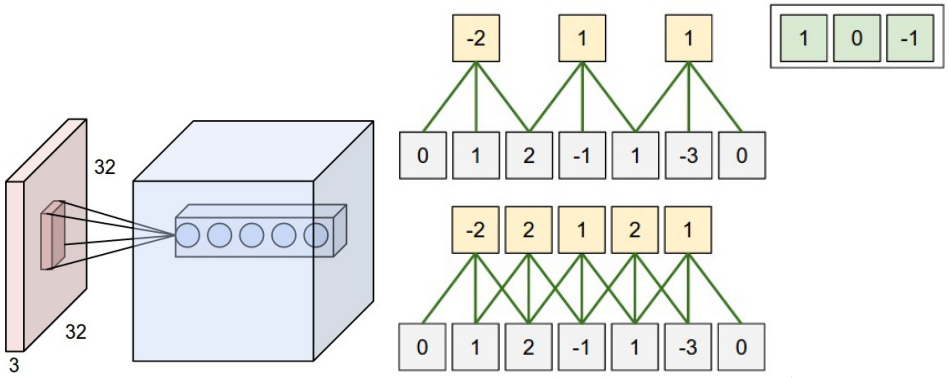

Abb. 2.9 Links Faltungsoperation (*Oxford VGG Pratical*)
Rechts Visualisierung der Schrittweite (*Oxford VGG Pratical*)

An dieser Stelle sollte auch noch ein mal auf die Analogie mit klassischen neuronalen Netzen eingegangen werden: Während in einem klassischen neuronalen Netz jeder Pixel ein eigenes Neuron hätte, werden in einem CNN ganze Bildregionen in ein einzelnes Neuron (Kreis in Abbildung 2.9) integriert. Selbstverständlich reduziert sich damit auch ganz erheblich der während der Trainingsphase zu approximierenden Parameter. Im obigen Beispiel würde sich für eine, auch Rezeptivfeldgröße gennant, von 5 und einer Schrittweite von 1 (5×5×3), 75 Parameter ergeben. Ein vergleichbares Neuronales Netz würde allein für diesen kleinen Teil der Verarbeitungspipeline 32×32×3, also 3072 Gewichte, benötigen.
Um die Veränderung der Abmessungen des „Volumens“ auch als Blobsize bekannt, für jedes Layer zu berechnen muss noch ein weiterer Parameter besprochen werden: Das Padding. Der Rand des Volumens ist eine Besonderheit, da der Kernel an dieser Stelle einen undefinierten Bereich außerhalb des Bildes, abtasten und akkumulieren müsste. Deshalb füllt man diesen „Überstand ins Leere“ mit Nullen auf, um eine vollständig definierte Faltung gewährleisten zu können. Raffiniertere Randbedingungen, wie z. B. der so genannten von-Neumann-Boundary, konnten ihren rechnerischen Mehraufwand nicht mit einer Verbesserung des Endergebnisses rechtfertigen und sind aus diesem Grund äußerst unüblich.
Nun sind alle Parameter bekannt um die Blobsize zu berechnen: Die Eingangsvolumengröße W, die Rezeptivfeldgröße F, der Stride S und das zuvor erwähnte Padding P. Die Dimensionen des Outputs ergeben sich aus folgender Formel:

$$\frac{(W - F + 2P)}{S} + 1 \tag{2.18}$$

Um eine mathematische Herleitung auszusparen betrachte man das praktische Beispiel aus Abbildung 2.9. Zum Verständnis reicht die Erklärung des eindimensionale Falles: Oben rechts sind die Gewichte der Beispielneuronen, von denen es 3 gibt (F = 3), zu sehen. Es wurde für beide Beispiele ein Stride S von 1 verwendet, welcher unter Berücksichtigung der Eingangsgröße W = 5 zu einer Ausgangsgröße von 3 für das obere und 5 für das untere führt. Ein Ausgangspixel wird durch ein gelbes Quadrat, ein Inputpixel durch ein graues, dargestellt. Die Formel würde mit diesen Faktoren für oben und unten wie folgt aussehen $\frac{5-3+2}{2} + 1 = 3$, $\frac{5-3+2}{1} + 1 = 5$. Das Padding von 2 ist wie bereits erwähnt notwendig, da das Inputvolumen in diesem Beispiel kein Vielfaches des Rezeptivfelds ist und ein CNN Pixel diskret behandelt werden.
Die Tatsache, dass alle Neuronen die gleichen Gewichte verwenden nennt man Weightsharing und kommt oft zum Einsatz um die Anzahl von Freiheitsgraden in einem

Netz zu begrenzen. Meist wird derartiges während des Trimmens angewandt. Trimmen ist hierbei als manuelles Verbessern der Genauig- oder Geschwindigkeit zu verstehen.
Der beschriebene Ablauf stellt die Kernoperation innerhalb eines Netzes da und kann als Schablone für alle Layerarten verwendet werden. Bevor man sich diesen in Abschnitt 2.4.3 zuwendet lohnt es sich zuerst einen zweiten Blick auf die Hauptarten der Verwendungszwecke zu werfen.

2.4.1.1 Klassifikation - vom Bild zum semantischen Etikett

Klassifikation auch Typifikation genannt ist per Definition die „systematische Einteilung oder Einordnung von Ausdrücken, Gegenständen, oder Ähnlichem in Klassen/Gruppen oder Unterklassen/Untergruppen“. Die Klassifikation im technischen Sinne und insbesondere in Verbindung mit CNNs ist tatsächlich ein sehr weit gefasster Begriff, welcher von der Textanalyse über Sprachverarbeitung (Natural Language Processing) bis hin zur semantischen Segmentierung eines Bildes (Scene Labeling). Formale Definitionen sind ebenso vielfältig und reichen von Binärklassifikatoren (entweder/oder) bis hin zu hierarchischen Multilabelcategorization bei der es auch sich einschließende Kategorien, wie z. B. Flüssigkeit ⊃ Getränk ⊃ Kaffee ⊃ Expresso, geben kann. Das angeführte Beispiel ist, eben durch diese Verkettung, auch Namensgeber des CNN Frameworks der ‚California University at Berkeley' „Caffe“ geworden.
Im Falle der Gesichtserkennung definiert sich ein Klassifikator, als dasjenige System/Methode/Algorithmus welches eine Testgesichtsrepräsentation (nicht zwangsläufig ein Bild) auch „Probe“ genannt, dem ähnlichstem Label zuordnet und ein Sicherheitsmaß berechnet, was ermöglicht die Zuordnung unterhalb eines Grenzwertes abzulehnen.

	Klassifikator entscheidet	
Wahrheit	zugehörig	nicht zugehörig
zugehörig	True Acceptance	False Rejection
nicht zugehörig	False Acceptance	True Rejection

Tab. 2.2 Konfusionsmatrix

Das Maß ist für die Gesichtserkennung existentiell, um Authentifizierungsversuche

abzulehnen. Insbesondere für Anwendungsfälle in denen z. B. Zahlungen autorisiert werden ist es ratsam einen höheren Grenzwert zu setzen, um die so genannte *False Acceptance Rate* (FAR) gering zu halten. Allerdings steigt mit dem Grenzwert, oder je nach dem antiproportional mit der FAR, die *False Rejection Rate* (FRR) also die inkorrekte Ablehnung eines Exemplars. Es handelt sich folglich um Eigenschaften, welche auf gegenseitige Kosten optimiert werden können. Ein System bei welchem FRR und FAR exakt gleich sind, nennt man *neutral*. Die Fehlerquote an dieser Schwelle nennt man *Equal Error Rate* (ERR) und dient als zuverlässiges Gütekriterium eines Klassifikators. Die Fälle in welchen der Klassifikator korrekt entscheidet, misst man mit der *True Rejection Rate* (TRR) und der *True Acceptance Rate* (TAR), wobei man unter TRR die Ablehnung einer Stichprobe welche auch tatsächlich nicht einer der bekannten Klasse zugeordnet werden kann, versteht. TAR ist demnach die korrekte Zuweisung einer bekannten Gesichtsrepräsentation.

Für Gesichtsklassifikation mit CNNs gibt es mehrere bekannte Beispiele, von denen *OpenFace*, die Gesichtserkennung von Facebook, sowie *VGG-Face* der Gewinner der „Labeled Faces in the Wild Challenge“, die am weitesten verbreiteten sind. Beide Netze werden zu einem späteren Zeitpunkt, genauer gesagt in Absatz 3.4.0.0.2 und Absatz 3.3.4.0.1, noch zum Einsatz kommen. Zuerst sollte der Blick jedoch noch einmal auf die generelle Funktion eines Klassifikators gerichtet werden.

Wie in Abschnitt 2.2 bereits erwähnt, „wählt“ ein CNN basierter Klassifikator seine Features selbstständig aus und bildet damit, neben den manuell ausgewählten Merkmalen und den statistisch-biometrisch basierten Methoden, eine weitere Kategorie von Verfahren zum Bildverstehen oder in unserem Falle der Gesichtserkennung. Der aufmerksame Leser wird sich nun fragen, wie aus den zuvor beschriebenen Faltungen eine numerische Zuordnung - Bild→ Klassenetikett - erfolgen kann. Durchaus denkbar wäre eine rekursive Faltung, welche das Volumen des Bildes solange verringert, bis es nur noch ein einzelner Pixel ist, dessen Wert dann wiederum als Etikett interpretierbar wäre. Dieses Vorgehen wäre allerdings nicht nur unnötig rechenintensiv, sondern auch äußerst anfällig für Störfaktoren wie partielle Verdeckung (Gegenstand/Gesicht nicht vollständig zu sehen), Bildrauschen (technisch bedinge Ungenauigkeit des Bildsensors), oder auch eine Änderung der Beleuchtungsbedingungen.

Ein robuster Klassifikator wird deshalb mit Hilfe eines *Pooling*(Pool)- und *Fully Connected*(FC)-Layers (vgl. Abschnitt 2.4.3), Bildregionen auswählen, welche von solchen Störfaktoren frei sind und die anderen Bildteile, so genannte *Patches*, verwerfen. Damit lässt sich nun auch einfach erklären, warum ein hierarchischer Bildklassifikator mit zunehmender Spezialisierung and Genauigkeit verliert: Ruft man sich das

namensgebende Kaffeebeispiel in Erinnerung, so kann man sich leicht vorstellen, dass eine Tasse mit dunkler Flüssigkeit das entsprechende Etikett Kaffee und Getränk erhält. Es handelt sich hierbei jedoch nicht um die „Erkenntnis“ des CNNs, dass jeder Kaffee ein Getränk ist, sondern um die Tatsache, dass die durch das Fully Connected Layer ausgewählten Bildregionen nicht nur das Etikett Kaffee sondern auch Getränk „auslösen“.
Welche Bildregionen zu welchem Etikett führen „lernt“ man durch die Vorgabe von Bild-Etikett Paaren während der Trainingsphase und der darauf basierenden iterativen Gewichtsanpassung aller Layer, welche in Abschnitt 2.4.4 noch ausführlicher dargestellt wird.
Allerdings ist es in den meisten Fällen nicht ausreichend nur FC-Layer zu kaskadieren. Durch vorherige Faltungen werden manche charakterisierende Merkmale erst sichtbar (vgl. Kantendetektion Tabelle 2.1). Auch ist es wichtig zu erwähnen, dass die Wahrscheinlichkeit der Zuordnungen als ein Sicherheitsmaß zu interpretieren sind und keinesfalls eine Ähnlichkeitsabbildung (vgl. Unterunterabschnitt 2.4.1.2) darstellen. Man betrachte nur folgendes Beispiel: Angenommen ein Bild führt zu den Etiketten: Getränk 97%, Kaffee 54% und Lastkraftwagen 72%, dann handelt es sich vermutlich um ein Bild welches alle genannten Objekte enthält, welcher aber je nach Sicherheit unterschiedlich gut zu erkennen sind. Keinesfalls handelt es sich um ein Objekt was einem LKW zu 72 und einem Getränk zu 97% ähnelt.
Um ein Ähnlichkeitsmaß zu erhalten verwendet man einen Embedder (Unterunterabschnitt 2.4.1.2) auf dessen Ergebnisse ein Klassifikator aufsetzen kann. Dieses kann, je nach Güte des Systems, von einem einfachen Korrelationsmaß, über eine *Support Vector Machine*(SVM) bis hin zu einem weiteren CNN reichen. Embedder können folglich als als merkmalserhaltende Dimensionsreduktionsverfahren beschrieben werden. Diese Eigenschaft ist äußerst nützlich für Gesichtserkennung die auch für Personen funktionieren sollte, welche nicht in der Trainingsphase gelernt wurde. Deshalb werden Sie im nächsten Abschnitt näher betrachtet.

2.4.1.2 Embedding - CNN basierte Dimensionsreduktion

Ein Embedding kann als niedrig dimensionale Repräsentation angesehen werden. Ein Embedder ist demnach ein System zur Dimensionsreduktion. Wie in Abbildung 2.10 zu sehen ist, werden Bilder(Personen) anhand ihrer Ähnlichkeit sortiert. Der Abstand im Identifikationsraum ist also gleichbedeutend der Unähnlichkeit zweier Instanzen. Der schwierige Teil eines Embedders ist dessen Training. Um eine Ähnlich-

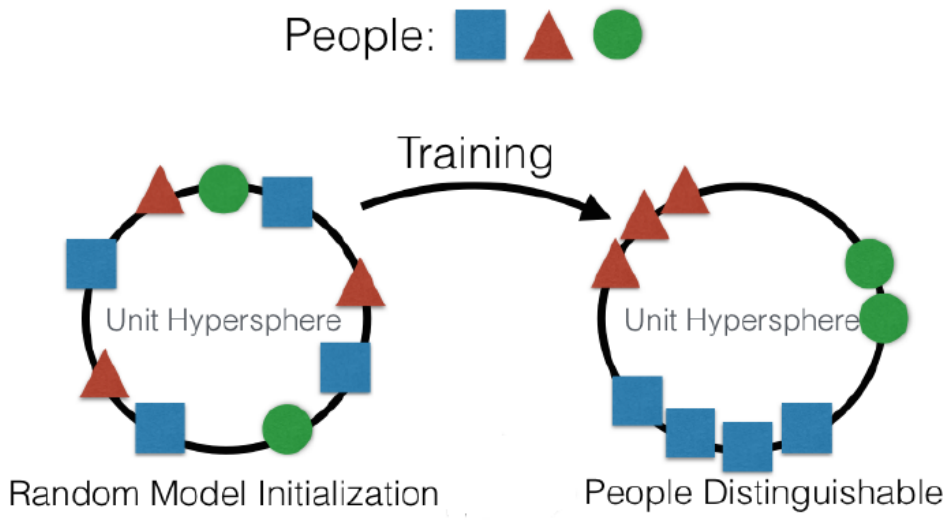

Abb. 2.10 Openface Trainingsprozess (Amos et al., 2016)

keit zu finden ist ein sogenanntes *Triplet Loss* und eine äußerst große Datenmenge notwendig. Openface zum Beispiel wurde auf der Datenbasis von 4.4 Millionen Bildern und etwa 200 Tausend Personen trainiert. (Amos et al., 2016)
Der Trainingsdatensatz wird in *Mini-Batches* aufgeteilt, welcher lediglich Q Personen mit P Bildern beinhaltet. Diese Mini-Batch wird dann weiter in Triplets geteilt. Ein Triplet besteht aus einen Anchor A, einer positiven P und einer negativen N Instanz. Positiv heißt in diesem Kontext, dass die Instanz das gleiche Individuum aufweist wie A. Negativ heißt, dass die Identität der Instanz ungleich der von A ist. Aus der Mini-Batch werden alle möglichen Triplets gebildet und prozessiert. Der Fehler (engl. Loss) wird dann wie folgt berechnet:

$$L(A, P, N) = ||f(x^A) - f(x^P)|| - ||f(x^A) - f(x^N)|| \tag{2.19}$$

Hierbei ist $f(x)$ als Repräsentation von Bild x zu verstehen. Diese Funktion wird kleiner je näher die zwei gleichen Instanzen zueinander sind und je weiter das ungleiche Paar auseinander ist. (Amos et al., 2016)
Das Prinzip der Mini-Batches und der Triplets beruht auf der Annahme, das das kontinuierliche, lokale Optimieren der Dimensionsreduktion zu einem globalen Minimum führt. Das finden des globalen Minimums erfordert viele Iterationen, durch das Aufteilen auf Mini-Batches lässt sich der Prozess jedoch gut parallelisieren. (Amos et al., 2016) argumentieren, dass sie für das Trainieren aller 4.4 Millionen Bilder nicht mehr als einen Tag auf handelsüblicher Hardware brauchen.
Das Embedding hat einen weiteren Vorteil: Die Repräsentation kann durch eine Support Vektor Machine zu einem Klassifikator umgewandelt werden und im Allgemei-

nen generalisiert ein so trainiertes neuronales Netz sehr gut. Die meisten der heute eingesetzten Systeme verwenden einen Embedder zur Gesichtserkennung. Problematisch wird es erst, wenn das Bild zu verzerrt ist. Ein Embedder ist darauf angewiesen, dass die einzelnen Gesichter präzise aufeinander abgebildet (normalisiert) sind, da das Netz keine Pooling-Layer verwendet, welche Bildregionen positionsunabhängig werden lässt. (Amos et al., 2016)

2.4.1.3 Regression - gelernte Filter und Bildtranslation

Nachdem die Unterschiede zwischen einem Klassifikator Unterunterabschnitt 2.4.1.1 und einem Embedder Unterunterabschnitt 2.4.1.2 herausgearbeitet wurden, wird mit der Regression die letzte Hauptart von CNNs erläutert. Bisher wurde beschrieben, wie semantische Informationen aus einem Bild extrahiert, verglichen und gruppiert werden können. Eine Regression kann jedoch durchaus als Regression im ursprünglichen, mathematischen Sinne verstanden werden: Eine Funktion (im Falle eines CNNs eine nichtlineare mit sehr vielen Parameters) approximiert einen Datensatz (Bildpaare während der Trainingsphase) unter Minimierung eines Abstandsmaßes (siehe Abschnitt 2.4.4).
Es ist also möglich mit Hilfe eines faltenden Neuronalen Netzes nichtlineare Bildfilter zu erlernen, was auch den Vorteil hat, dass man einer exakten mathematischen Beschreibung eines Filters schuldig bleiben kann und trotzdem ein funktionierendes, absicherbares System erhält. Ein prominentes Beispiel für das oben Erwähnte ist zum Beispiel das CNN basierte *Supersampling*. Während der Trainingsphase wird die Auflösung eines Bildes verringert und durch gleichverteiltes Rauschen gestört. Das Netz soll dabei die ungestörte, scharfe Urversion des Eingangsbildes wiederherstellen. In Abbildung 2.11 ist zu sehen wie ein derartiges Verfahren zum Nachschärfen von gescannten Texten verwendet wird. Die begrenzte Anzahl an Buchstabenformen sind für ein CNN leicht zu erlernen. Dieses „Wissen“, in der Fachwelt auch *Prior* genannt, macht diese Art von neuronalen Netzen herkömmlichen Methoden wie zum Beispiel der anisotropischen Diffusion, welche lediglich auf dem Bildgradienten beruht, überlegen.
Durchaus kann man diese Art Verfahren auch als Klassifikator bezeichnen, bei welchem die Anzahl der vorher zusagenden Etiketten gleich der Anzahl der gewünschten Bildpixel ist. Dieser Ansatz wäre allerdings äußerst unüblich da bei einer derartigen Bildtranslation eine „patchbasierter“ Ansatz Kontextinformationen verlieren würde, welche für eine korrekte Abbildung wichtig wären. Die Übergänge zwischen

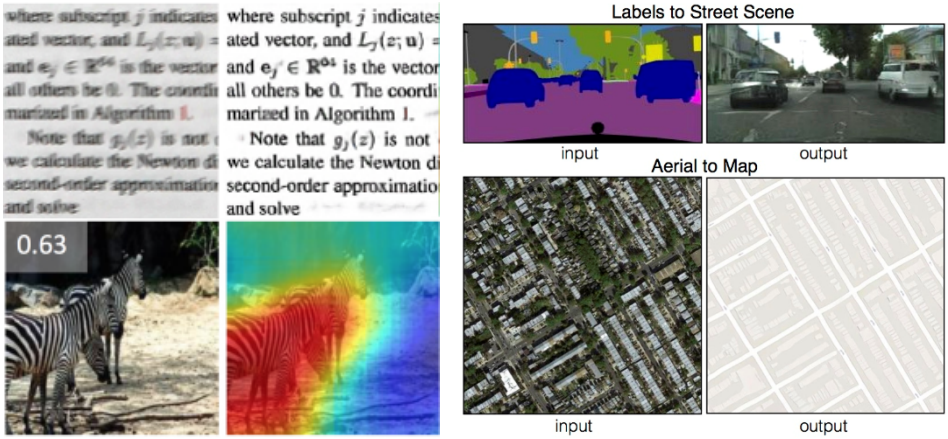

Abb. 2.11 Links Superresolution durch CNN (*Oxford VGG Pratical*)
Rechts Segmentierung durch CNN (*Oxford VGG Pratical*)

beiden Verfahren sind aber durchaus fließender als es zunächst den Eindruck macht: Die Segmentierung eines Bildes Abbildung 2.11 zum Bepsiel durchaus als Klassifikation als auch als Bildtranslation implementiert werden. Die Bildsegmentierung befasst sich mit dem Unterteilen und semantischen Einordnung von Bildregionen, wobei die Anzahl der Etiketten gegeünber der klassischen Klassifikation stark beschränkt ist. Im Beispiel von Abbildung 2.11 wird lediglich zwischen Straße, Fahrzeug und anderes unterschieden. Bei einer Klassifikationsvariante würde man den Ort der Etiketten über eine sogenannte *Neuron Activation Map* Abbildung 2.11 (unten) ermitteln. Man kann sich das als eine Rückverfolgung vorstellen, in dem herausgefunden wird welche Bildregion welches Etikett ausgelöst hat. Das Netz wird also sozusagen rückwärts durchlaufen. In der Bildtranslationsvariante würde man wie in Abbildung 2.11 jedes semantische Etikett in eine Farbe kodieren um eine Klassifikation zu erreichen.

Es kann aber auch bereits der erste signifikante Unterschied erkannt werden: In der Bildtranslationsvariante wäre es möglich von der Segmentierung aus ein passendes Bild zu halluzinieren. Dieses wäre mit dem Klassifikationsansatz nicht möglich. Bei aller Verblüffung über die Güte der Ergebnisse Abbildung 2.11, sollten die Fähigkeiten dieser Systeme nicht überschätzt werden. Im Grunde wendet ein solches CNN gelernte Gegebenheiten, z. B. Formen und Simplexe, an. Das ist auch der Anlass, weshalb man diesen Vorgang *Halluzination* nennt.

Erst kürzlich konnte man in der Fachwelt vernehmen, dass es mit „Deep Learning“, also ein neuronales Netz mit „vielen“ Ebenen, möglich sei aus einem 8×8 - 64 Pixel Bild - ein 256×256 Gesicht zu rekonstruieren, was selbstverständlich nicht die volle Wahrheit sein kann. Vielmehr handelt es sich dabei um das Wahrscheinlichste. Was das Wahrscheinlichste ist hängt vom Trainingsdatensatz ab. Ein Mensch kann zum

Beispiel auch, wenn er einen Teil eines Gesichtes sieht, die verdeckten Regionen anhand seiner Erfahrungswerte erahnen. Keinesfalls ist es jedoch so, dass Narben oder andere Irregularitäten, welche aber durchaus charakterisierend sind, rekonstruiert werden. Es handelt sich folglich um eine vage Möglichkeit, weshalb der Ausdruck *Halluzination* eine passende Umschreibung dieser Art von Anwendungsfällen ist.

2.4.2 Architekturen

Nach dem die grundlegenden Anwendungsfälle und Formen von CNNs besprochen wurden, werden nachfolgend verschiedene Architekturen vorgestellt. Eine Architektur beschreibt die Struktur des Netzes ohne auf einzelne Ebenen oder Besonderheiten einzugehen. Aus der Architektur lassen sich Eigenschaften des resultierenden Netzes ableiten und eignet sich außerdem zur Einordnung und Analyse verschiedener Ansätze und Lösungsideen.

2.4.2.1 Encoder-Decoder

Das Encoder-Decoder, auch Autoencoder genannt ist die einfachste Architektur eines neuronalen Netzwerks. Am einfachsten kann man sich das Verfahren unter zur Hilfenahme der Fourier-Transformation erklären. Die Kernelgröße der Faltungen wird zur Mitte hin immer geringer (siehe Abbildung 2.12) was zu einem Flaschenhals führt. Das Bild wird folglich zunächst herunterskaliert und dann schrittweise wieder auf die Ursprungsgröße hochskaliert. Im Frequenzraum betrachtet werden die niedrigen Frequenzen im Flaschenhals rekonstruiert. Die höheren Frequenzen in den Filtern mit größerer Kernelsize. (Isola et al., 2016)

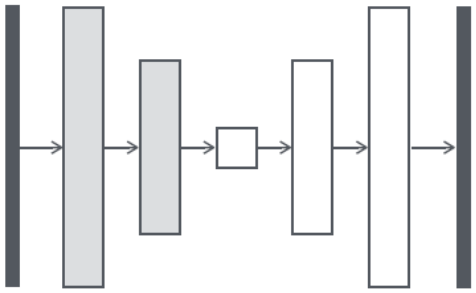

Abb. 2.12 Encoder-Decoder Architektur (Isola et al., 2016)

Das herunter und heraufskalieren führt zu einem Informationsverlust, sodass ge-

rade Regressionsmodelle, welche auf Encoder-Decoder Architekturen beruhen nur sehr unscharfe Ergebnisse liefern. Um dem Informationsverlust entgegenzuwirken wird im folgenden Abschnitt die *Skip-Connection* eingeführt.

2.4.2.2 U-Netze

Ein U-Netz ist eine Encoder-Decoder Architektur, bei welcher jede äquivalenten Ebene als Dimension kopiert wird. Die Informationen die in der ersten Ebene des Encoders enthalten sind werden als zusätzliche Dimension der letzten Ebene des Decoders verwendet (siehe Abbildung 2.13).

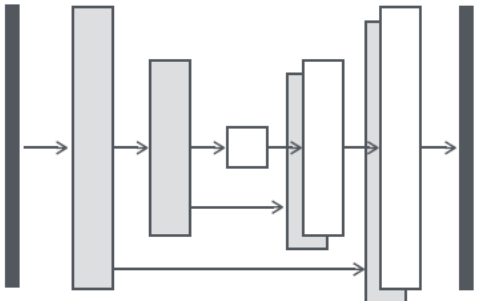

Abb. 2.13 U-Netz Architektur (Isola et al., 2016)

Die Faltungen, welche, wie in Unterunterabschnitt 2.4.1.1 beschrieben, 4- Dimensional sind, verfügen so mehr Informationen aus den ursprünglichen Eingangsdaten. Es handelt sich folglich um eine reine Erweiterung der zuvor vorgestellten Encoder-Decoder Architektur. Der Nachteil der Skip-Connections ist der damit einhergehende Mehraufwand an Speicher, weshalb es durchaus üblich ist nur von der ersten zur letzten Ebene zu kopieren. (Isola et al., 2016)

2.4.2.3 Adversarial-Netze

Die Adversarial Architektur besteht aus zwei Netzen: Einem generativen und einem dezisiven. Das generative Netz (G) synthetisiert ein Bild, zum Beispiel ein VIS aus einem NIR-Bild, welches dann vom dezisiven Netz (D) prozessiert wird. Das dezisive Netz versucht zu unterscheiden, ob es sich um ein generiertes oder ein echtes Bild handelt (siehe Abbildung 2.14). (Isola et al., 2016)
Spieltheoretisch ergibt sich zwischen G und D ein Nullsummenspiel. Da beide Netze

mit zunehmender Iteration genauer werden, können sie sich gegenseitig trainieren, so dass keine Lernfaktorkurve ermittelt werden muss. Weiterhin kann man dem Problem der Überparametrisierung entgegenwirken indem man das dezisive Netz mit mehr Parametern starten lässt als das Generative. So wird erreicht, dass mit weniger Parametern ein fast ebenso gutes Ergebnis erzielt wird, als mit ausgeglichen großen Netzen. Eine Überparametrisierung des generativen Modells wird somit explizit verhindert. (Isola et al., 2016)

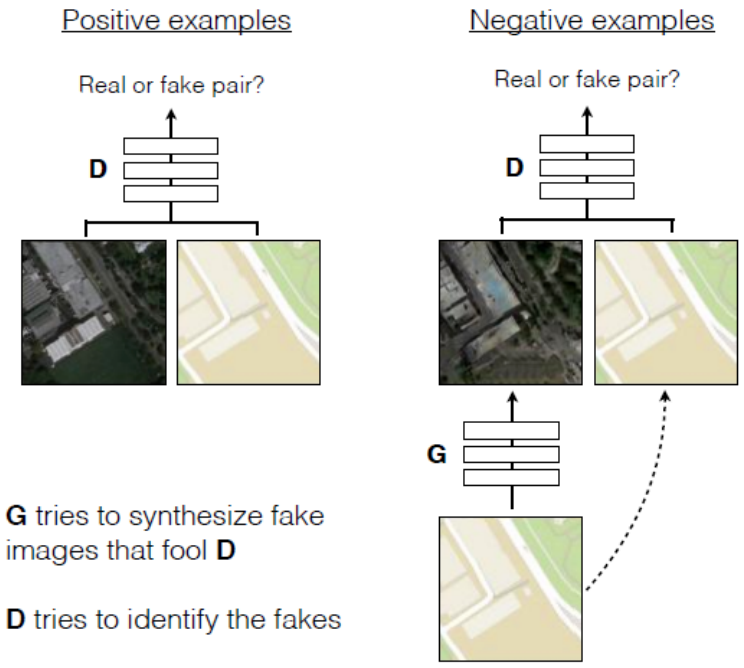

Abb. 2.14 Adversarial Architektur (Isola et al., 2016)

Da das generative Netz versucht das dezisive Netz zu täuschen ergibt sich ein weitere Eigenschaft: bei der Eingabe eines leeren (schwarzen) Bilds, wird das generative Netz das Mittelbild aller in der Trainingsphase gesehenen Daten ausgeben, da so die Wahrscheinlichkeit am höchsten ist, vom dezisiven Netz nicht erkannt zu werden. (Isola et al., 2016)
Abschließend lässt sich sagen, dass Adversarial-Netze vielversprechende Ergebnisse liefern und zu den komplexeren Architekturen von CNNs gehören. (Isola et al., 2016)

2.4.3 Bausteine eines CNNs

Einzelne Bausteine von CNNs wurden bereits erläutert. Im Folgenden wird nun noch einmal systematisch auf die Einzelheiten und Notwendigkeiten verschiedener Ebe-

nen eingegangen. Für neuronale Netze gibt es eine Vielzahl von Erweiterungen, die sich teilweise lediglich in der Implementierung unterscheiden. Grundsätzlich lassen sich jedoch alle Filter in die folgenden Kategorien aufteilen:

2.4.3.0.1 Faltung Jede Faltung kann als Verkettung von Funktionen dargestellt werden, bei welcher in jedem Schritt f_l ein Datum x_l und ein Parametervektor w_l. Die Sequenz von Funktionen ist meist manuell gewählt, wobei der Parametervektor während der Trainingsphase ermittelt wird. Bei den Daten handelt es sich wie bereits zuvor beschrieben um Bilder, oder allgemeiner Rasterdaten (da man CNNs auch für Ton und andere Informationen einsetzen kann).

$$f(x) = f_L(\dots f_2(f_1(x, w_1), w_2)\dots), w_L) \tag{2.20}$$

$$f : \mathbb{R}^{M\times N\times K} \to \mathbb{R}^{M'\times N'\times K'} \tag{2.21}$$

Eine Faltung f kann wie in Gleichung 2.21 mathematisch beschrieben werden, wobei $\boldsymbol{M}$ und $\boldsymbol{N}$ die Dimensionen der Ebene (Abmessung des Bilds) und $\boldsymbol{K}$ die Anzahl der Kanäle ist. Zum näheren Verständnis und weitergehenden Beispielen kann auch Abschnitt 2.4.1 herangezogen werden. (*Oxford VGG Pratical*)

2.4.3.0.2 Aktivierungsfunktion Die Faltung, wie sie zuvor beschrieben und in Tabelle 2.1 zu finden ist, ist eine lineare Funktion. Um nichtlineare Abbildungen modellieren zu können sind Aktivierungsfunktionen notwendig. Die meist verwendete ist hierbei die so genannte *Rectified Linear Unit*(ReLU). Es handelt sich hierbei um einen Grenzwert für Daten, welche bei nicht Erfüllung durch eine Konstante, meist 0, ersetzt werden.
Am einfachsten lässt sich das durch den folgenden mathematischen Zusammenhang beschreiben:

$$y_{ijk} = \max\{0, x_{ijk}\} \tag{2.22}$$

Zu Erinnerung: ijk beschreibt die Pixelposition(ij) und die Ebene(k) Da es sich um eine dreidimensionale Funktion handelt. Viele Erweiterungen, wie etwa die *Parametric Rectified Linear Unit*(PReLU) existieren sind vom Prinzip her jedoch ähnlich zu der hier vorgestellten Funktionsweise. (*Oxford VGG Pratical*)

2.4.3.0.3 Pooling Das Pooling vergleicht verschiedene Bildteile und liefert diejenigen Bildregionen welche bestimmte Kriterien erfüllen. Meist handelt es sich bei diesen Kriterien um relative Grenzwerte, zum Beispiel dem Maximum aller Pixel / Patches / Kanäle einer neuronalen Ebene. Auch die Summe mehrerer Einzelteile (Kanäle / Patches / Pixel) kann durch einen Poolingoperator für die Berechnung verfügbar gemacht werden. (*Oxford VGG Pratical*)
Das Max-Pooling, also das Maximum aller am *Teilnehmer*, kann mathematisch wie folgt beschrieben werden:

$$y_{ijk} = \max\{y_{i'j'k} : i \leq i' < i + p, j \leq j' < j + p\} \tag{2.23}$$

2.4.3.0.4 Normalisierung Die Nomalisierung ist wichtig um Vergleichbarkeit zwischen verschiedenen Trainingspaaren zu schaffen, da eine allgemeine Lösung gewünscht ist und der Fehler über die Gesamtheit aller Trainingsdaten minimiert wird. Eine bekannte Form der Normalisierung ist die Mittelwertsbefreiung, bei welcher der Mittelwert der Datenbasis errechnet und von allen Eingangsbildern abgezogen wird.
Reicht diese einfache Form nicht aus, so wird meist die Varianz der Datenbasis normiert, indem man die Varianz der Datenbasis errechnet und durch diese dividiert. Diese Form nennt man im Englischen *Zero Mean Unit Variance*(ZMUV). (*Oxford VGG Pratical*)
Viele weitere Formen wie zum Beispiel die Batch oder Channelnormalization sind in der einschlägigen Fachliteratur zu finden. Welche Methode zielführend für einen bestimmten Ansatz ist, hängt von zugrundeliegenden Datensatz und verarbeitenden Netz ab. Meist lässt sich diese Frage nur empirisch beantworten. (*Oxford VGG Pratical*)

2.4.4 Trainieren und iterativ lösen

Bisher wurde ausführlich über Architekturen, Ebenen und verschiedenen Funktionsarten berichtet. Um eine gewünschte Aufgabe zu erfüllen muss die zuvor eingeführte Parametermenge w gefunden werden. Betrachtet man den Fall der Regression so muss w derart gewählt werden, dass die Summe des Abstands der Funktion $f(x, w) = z$ (die Abbildung des CNNs) für den gesamten Trainingsdatensatz minimal ist. In anderen Worten: welche Gewichte müssen gewählt werden damit das CNNs bei gegeben Input x möglichst genau das gewünschte Ergebnis z erreicht.

Während der Trainingsphase lässt sich, dadurch das x und z bekannt sind, für jedes w ein Fehler $l(z, \tilde{z})$ für die Ausgabe des Netzes $\tilde{z}$ und der gewünschten Ausgabe z errechnen. Die einfachste Form der Fehlerberechnung ist die L_2-Norm. (*Oxford VGG Pratical*)
Der empirische Fehler wird durch den Mittelwert aller Ebenenfehler beschrieben und kann mathematisch wie folgt dargestellt werden:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, f(x_i, w)) \tag{2.24}$$

Die einfachste Methode das Minimum dieser Funktion zu finden ist das Gradientenverfahren, auch wenn dies in der Praxis durch komplexere, schneller konvergierende Ansätze weitestgehend abgelöst wurde. Das Gradientenverfahren würde in diesem Falle wie folgt aussehen:

$$w^{t+1} = w^t - \eta_t \frac{\partial f}{\partial w}(w^t) \tag{2.25}$$

Wobei $\eta_t \in \mathbb{R}_+$ die Lernrate ist. Im Anwendungsfall ist das Trainieren meist nur durch empirische Optimierung aller beteiligten Parameter, wie Fehlerberechnung, Optimierungsart und Lernrate durchzuführen. (*Oxford VGG Pratical*)
Abschließend kann gesagt werden, dass ein Netz in zwei Arten verwendet werden kann: im *forward-mode*, dass heißt $z = f(x, w)$ und im *backward-mode*, folglich $\langle p, f(x, w) \rangle$. (*Oxford VGG Pratical*)

3 Systementwurf

> *Was vorstellbar ist, ist auch machbar.*
>
> **– Albert Einstein**

Nachdem das Problem der heterogenen Gesichtserkennung ausführlich definiert und bestehende Lösungsansätze kategorisiert wurden, widmet sich dieses Kapitel dem Entwurf eigener Ansätze. Zunächst wird eine Anforderungsanalyse durchgeführt, die einen gemeinsamen Rahmen für alle Systeme beschreibt und zudem wünschenswerte Funktionalitäten und Möglichkeiten darlegt.
Danach wird jede im vorigen Kapitel vorgestellte Kategorie in einem Abschnitt abgehandelt. Als Feature-Ansatz wird die Verwendbarkeit von Gesichtlandmarken, auch als Fiducial Points bezeichnet, analysiert. Für die Syntheseverfahren werden mehrere Verfahren der NIR-VIS-Halluzination reproduziert sowie durch eigene Ideen verbessert und erweitert. Als Letztes werden – aus Gründen der Vollständigkeit – verschiedene homogene Klassifikatoren auf heterogene Datensätze angewandt, bevor auf spezifische Formen von siamesischen Netzen eingegangen wird.

3.1 Anforderungsanalyse

Sicherlich sind die Anwendungsfelder der heterogenen Gesichtserkennung vielfältig, in vorliegender Arbeit jedoch wird ein ganz spezifischer Anwendungsfall betrachtet. Es wurde bereits bei der Begriffsdefinition eine Einschränkung der Modalitäten auf 2D, 3D, NIR und VIS vorgenommen. Das Verfahren soll verschiedene *Driver-Monitoring- Systems* integrieren, es soll also ermöglichen, dass der Fahrer eines Autos über verschiedenste Sensorik anhand von Merkmalen seines Gesichtes erkannt wird.
Sowohl die Erkennung als auch die dazugehörige Datenbank bekannter Nutzer sollen erweiterbar sein. Konkret bedeutet dies, dass Personalisierungsdaten von Fahrzeug zu Fahrzeug über diverse Baureihen und Technologiegenerationen hinweg portiert werden, während die einmal extrahierten Merkmale eines Systems übertragbar sein sollen. Als Fahrerkamera können sowohl Tiefenkameras als auch Mono- oder Stereosysteme zum Einsatz kommen. Dadurch, dass Fahrerbeobachtungssysteme in

aller Regel innerhalb des Infrarotspektrums operieren, ist eine Übertragbarkeit der Erkennung in den sichtbaren Bereich wünschenswert, um ein Anlernen per Smartphone zu ermöglichen. Als erstrebenswert wird hierzu ein Klassifikator angesehen, der hinreichend genau ist, um Zahlungen, z. B. an Mautstationen, zu authentifizieren.

Drei Sekunden Zeit zwischen dem Einsteigen und dem Starten des Fahrzeugs werden als realistisch angesehen, weshalb bei der Systementwicklung davon ausgegangen werden darf, dass das zu identifizierende Gesicht für mindestens drei Sekunden vollständig sichtbar ist. Weiterhin darf angenommen werden, dass das Gesicht hinreichend gut aufgelöst werden kann, das heißt, mit mindestens 256 Pixel Augenabstand. Zusätzlich darf angenommen werden, dass innerhalb dieser drei Sekunden die Kopfrotation, relativ zum aufnehmenden Sensor, geringer als fünfzehn Grad ist. Die drei Sekunden sind als Mindestdauer zu verstehen, während der alle genannten Bedingungen gleichzeitig erfüllt sein sollten. Die Aufnahme muss dabei aber nicht zwingend konsekutiv, das heißt an einem Stück erfolgen. Eine Aufnahme von 10 Sekunden, bei der am Anfang zwei Sekunden und am Ende noch einmal eine Sekunde lang alle Bedingungen erfüllt sind, gilt als valide Eingabe für die Gesichtserkennung. Störfaktoren, wie Sensorrauschen, und schwierige Lichtverhältnisse, wie direkte Sonneneinstrahlung, sind im üblichen Maß zu berücksichtigen. Formeller kann gefordert werden, dass die heterogene Erkennung mindestens genauso robust sein sollte wie die homogene Variante.

Es ist wünschenswert, aber nicht notwendig, dass ein Verfahren für alle Modalitäten verwendet werden kann. Eine vorgeschaltete eigenständige Schnittstelle für jede Modalität ist durchaus denkbar. Weiterhin ist die heterogene Gesichtserkennung immer nur als Initial- oder Umstellungsprozess anzusehen, das heißt, es muss möglich sein, einen neuen Nutzer per Smartphone erstmalig anzulernen oder eine einmalige Portierung durchzuführen. Danach werden die spezifischen Merkmale der jeweiligen Modalität dem Nutzerprofil hinzugefügt, um eine homogene Erkennung zu ermöglichen. Die Anzahl von Nutzern, die in der Datenbank enthalten sind, sollte mit mehr als einer halben Million geplant werden.

Betrachtet wird außerdem lediglich die Erkennung mittels Gesichtsmerkmalen. Eine Identifikation mittels mehrerer weiterer Faktoren, wie zum Beispiel Smartphone, Fingerabdruck oder Stimme, ist nicht Gegenstand des zu entwerfenden Systems.

3.2 Feature

Nachdem in Unterunterabschnitt 2.3.2.1 verschiedenste Formen von Feature- Verfahren für 2D-, 3D und NIR-VIS-Gesichtserkennung vorgestellt wurden, widmet sich das folgende Kapitel dem Entwurf eines Merkmals, dass in allen Modalitäten, das heißt in 2D, 3D, NIR und VIS, gemessen werden kann. Die Idee dahinter ist, die Gesichtsform zu ermitteln, die in allen untersuchten Modalitäten (vgl. Abschnitt 3.1) messbar ist.
Zunächst werden Gesichtslandmarken sowohl in 2D als auch in 3D ermittelt, bevor mithilfe eines *3D Morphable Models*(3DMM) eine engmaschige Rekonstruktion des menschlichen Gesichts angestrebt wird. Zum Abschluss werden verschiedene Abstandsmaße zum Vergleich der Messdaten vorgestellt.

3.2.1 Tracking von Gesichtslandmarken

Zum Tracking von Gesichtslandmarken werden zwei bestehende Driver- Monitoring-Systems der Firma Seeing Machines Ltd. verwendet. Sowohl bei dem Mono-System als auch bei dem Stereo-System kommt die vierte Generation der Fovio-Software zum Einsatz. Nachfolgend werden die einzelnen Systeme vorgestellt und es wird kurz auf die zugrunde liegenden Algorithmen eingegangen. Beide Systeme geben die gleichen Gesichtslandmarken relativ zum *Nasion*(Mittenauge) aus. Eine Normalisierung der Kopfrotation findet bereits innerhalb der Software statt.

3.2.1.1 Mono-System

Das Fovio-Mono-System besteht aus einer NIR-Kamera der Marke Ximea, die durch zwei IR-LED unterstützt wird. Die Helligkeit dieser LEDs wird dynamisch gesteuert, wodurch eine konstante Bildhelligkeit ermöglicht wird. Durch die Beleuchtung wird jedoch nicht nur eine Unabhängigkeit von Umwelteinflüssen erreicht: Die LEDs blitzen und erzeugen so einen Hornhaut-Pupillen-Reflex, mit dessen Hilfe die Blickrichtung des Nutzers bestimmt werden kann. Zusätzlich ergeben sich durch die Änderung von Schattierungen, welche von der Position der momentanen Beleuchtung abhängig ist, Forminformationen über das Gesicht. Diese Informationen wiederum können im Prozess der Lokalisation der Gesichtslandmarken eingearbeitet werden.

Beim Fovio-System handelt es sich um eine proprietäre Software, weshalb es schwierig ist, genaue Informationen über den Headtracking-Algorithmus in Erfahrung zu bringen. Um dieses Problem der lückenhaften Dokumentation der Funktionsweise zu umgehen, wird stellvertretend das am weitesten verbreitete Verfahren und der Gewinner der *300-Faces-in-the-Wild-Challenge* von (Zhu und Ramanan, 2012) vorgestellt.

Bei diesem Verfahren werden, relativ zur aufnehmenden Kamera, gleichzeitig sowohl die Kopfposition und die Kopfrotation als auch der Abstand zur Kamera geschätzt. Die Basis bilden dabei die auch im Fovio-System zur Anwendung kommenden 68 I-BUG Featurepoints, welche auch in Abbildung 3.1 zu sehen sind.

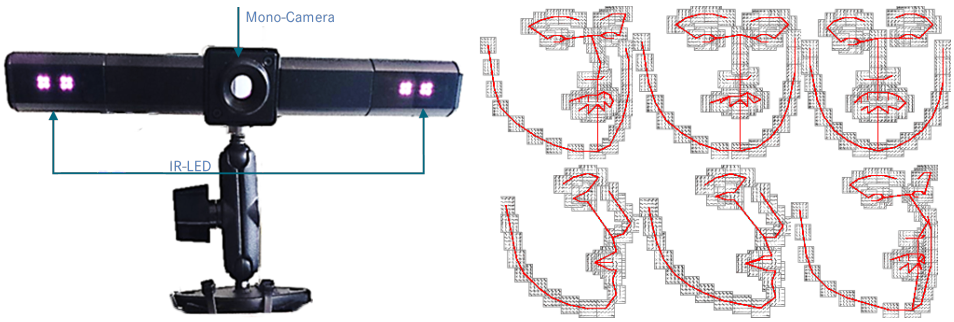

Abb. 3.1 Links Beschreibung Fovio-Mono-System
Rechts Topologie(Federn) in Rot, HOG der einzelnen Punkte in Grau (Zhu und Ramanan, 2012)

Einer der vielen Vorteile bei diesem ganzheitlichen Ansatz ist die Robustheit gegenüber Verdeckungen. Modelliert wird das Ganze durch eine Baumstruktur, das heißt, es gibt eine Topologie von Gesichtslandmarken, deren Zusammenhang als Feder interpretiert wird. Über das Elastizitätsgesetz lässt sich dann eine Spannung für jede *Feder* finden. Die Summe der Spannungen aller Federn (68 Gesichtspunkte in einer Baumstruktur ergeben 67 Federn, siehe Abbildung 3.1) wird dann minimiert, um die wahrscheinlichste Konstellation von Gesichtspunkten, Kopfposition und Rotation zu finden. Weiterhin können durch diesen topologischen Ansatz elastische Deformationen sehr gut modelliert und somit Abweichungen aufgrund von verschiedenen Gesichtsausdrücken beschrieben werden.

Die Methode lässt sich wie jedes Optimierungsproblem in einen *Daten- und einen Modellpart*(Data Term and Smoothness Term) teilen. Der Datenpart, das heißt derjenige Teil der Optimierung, der sich lediglich am Bild orientiert, ist hierbei die Lokalisation der 68 Gesichtslandmarken via einen HOG-Deskriptor. Die Gesichtslandmarken sind somit in 2D bekannt und entsprechen dem gesehenen Bild. Verdeckte Landmarken sind folglich nicht detektiert.

Verdeckte Landmarken sowie Kopfposition und Rotation sind also nur zu berech-

nen, wenn weitergehende Hintergrundinformationen über das menschliche Gesicht eingebracht werden. Das geschieht mithilfe des Modellterms. Dieser hat eine durchschnittliche 3D-Gesichtsform *Mean Shape*, die von einem Trainingsdatensatz mit bekannten Werten für alle Landmarken ermittelt wurde. Die zuvor bereits erwähnten Federn unterliegen bei der durchschnittlichen Gesichtsform keinerlei Spannungen. Um beide Informationsquellen zu verbinden, lassen sich der Datenterm und der Modellterm summieren und danach minimieren. Je nachdem ist es dann möglich, einen Parameter hinzufügen, der beide Summanden gegeneinander gewichtet und sich dadurch anwendungsspezifisch entweder eher auf Daten oder eher auf gelernte Modelle stützt.
Eine genaue mathematische Definition der zugrunde liegenden alternierenden Optimierung beider Terme kann (Zhu und Ramanan, 2012) entnommen werden. Eine tiefer gehende Beschreibung ist zum Verständnis des des darauf aufbauenden Systems nicht nötig. Es reicht zu wissen, dass es 3D-Modellinformationen, messbare 2D-Positionen und ein gemeinsames Minimum gibt.

$$\min \left\| \underbrace{\text{HOG-Deskriptor}}_{\text{Datenterm}} - \alpha \overbrace{\text{Gesichtsform}}^{\text{Modellterm}} \right\| \tag{3.1}$$

Das Fovio-System kennt drei Qualitätszustände von Gesichtslandmarken: 0 - nicht gefunden; 1 - modelliert; 2 - gemessen. Mit obiger Information ist es nun ein Leichtes, diese drei Zustände zu interpretieren. 0 bedeutet, dass kein Gesicht gefunden wurde. Das ist zum Beispiel der dann Fall, wenn die Verdrehung relativ zur Kamera derart groß ist, dass nur ein sehr kleiner Teil des Gesichts zu erkennen ist, welcher nicht ausreicht, um alle anderen Punkte daraus abzuleiten. Der Zustand 1 – modelliert – heißt, dass ein Gesichtspunkt zwar für die Kamera nicht sichtbar ist, jedoch anhand der anderen Gesichtspunkte ermittelt werden kann. Wenn somit nur eine Gesichtshälfte sichtbar ist, so ist die Annahme naheliegend, dass durch die Symmetrie die nicht sichtbare Seite ähnlich aussehen wird. Bei (Zhu und Ramanan, 2012) wird dies durch die Topologie der Punkte erreicht, bei welcher nicht sichtbare Punkte die Deformation ihrer sichtbaren, übergeordneten Landmarken *erben*. Gemessen ist dann der beste Zustand. Punkte mit diesem Gütemaß konnten im Bild gefunden werden und deren Position ist somit für das aufgenommene Individuum spezifisch – genau das, was für eine Identifikation benötigt wird.
Es wird angenommen, dass während der in der Anforderungsanalyse definierten *drei Sekunden Aufnahme mit wenig Kopfrotation*, der Zustand 2 –gemessen – herrscht.

Unter dieser Annahme gibt es insgesamt für drei Sekunden Bildmaterial, in welchen das Fovio-System tatsächlich alle gewünschten Gesichtslandmarken messen kann. Wie bereits erwähnt, werden die Gesichtslandmarken normalisiert, sodass das Mittenauge der Ursprung des Gesichtskoordinatensystems ist. Weiterhin findet ein Rotationsausgleich statt. Alle verbleibenden Asymmetrien sind somit individuellen Gegebenheiten geschuldet und nicht durch Kopfposition oder Rotation verursacht.

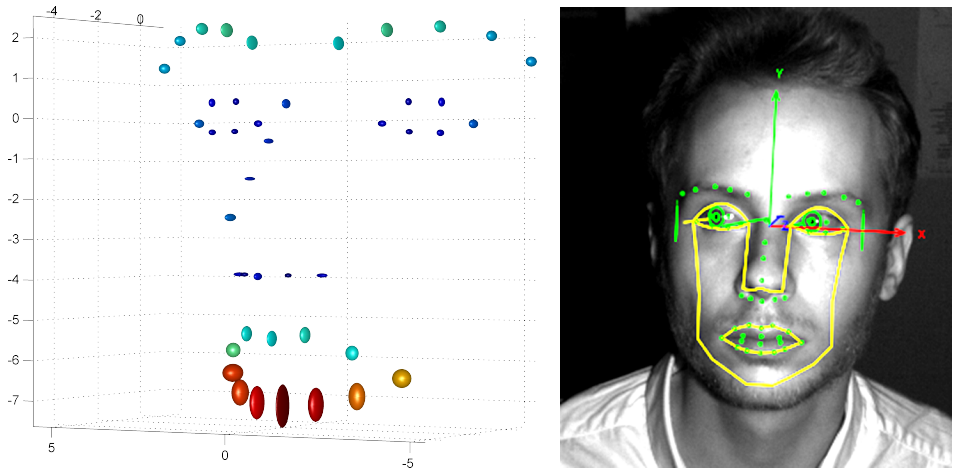

Abb. 3.2 Links Intraindividuumstandardabweichung des Fovio-Mono-Systems. Angaben in cm.
Rechts Augmentierte Ausgabe des Fovio-Mono-Systems mit Kopfkoordinatensystem

In Abbildung 3.2 ist die Standardabweichung des Verfahrens für jede Gesichtslandmarke zu sehen. Der Radius der Ellipsoide beschreibt hierbei die Standardabweichung in Richtung der jeweiligen Dimension. Die Farbe codiert die Summe aller Standardabweichungen. Die Position entspricht dem Durchschnitt aller gemessenen Positionen. Entstanden ist dieses Ergebnis durch eine in Abschnitt 4.1.1 näher beschriebene Nutzerstudie. Es handelt sich dabei um die durchschnittliche Standardabweichung pro Individuum, also genau um den Anteil, der nicht zur Identifikation genutzt werden kann, da es sich um Messungenauigkeiten handelt.
Im nächsten Abschnitt wird der Unterschied des Fovio-Stereo-Systems zu dem vorstehend beschriebenen Fovio-Mono-System aufgezeigt. Es wird insbesondere eine deutliche Differenz zu Abbildung 3.2 gezeigt und erläutert.

3.2.1.2 Stereo-System

Das Fovio-Stereo-System besteht aus zwei NIR-Kameras der Marke Ximea, die durch zwei IR-LEDs unterstützt werden (siehe Abbildung 3.3). Die Helligkeit dieser LEDs wird dynamisch gesteuert und passt sich somit der Umgebung an. Weiterhin blitzen diese LEDs abwechselnd, um einen Hornhaut-Pupillen-Reflex zu erzeugen. Mithilfe dieses Reflexes kann die Blickrichtung des Nutzers bestimmt werden.
Das Stereo-Verfahren unterscheidet sich leicht von dem des Mono-Systems. Es gibt eine Master- und eine Slave-Kamera. Das Bild der Master-Kamera wird genau wie im Mono-System verarbeitet, mit dem Unterschied, dass die resultierenden 3D-Punkte in das Koordinatensystem der Slave-Kamera projiziert werden. Dies geschieht mithilfe des *Epipolarzusammenhangs*(engl. Epipolar Constraint). Ein zugrunde liegender Term ist nicht notwendig. Es reicht aus zu wissen, dass beide Kameras mithilfe eines Kalibrationsmusters bekannter Abmessung aufeinander kalibriert sind und dass eine Matrix existiert, welche 3D-Punkte von einem Kamerasystem in das jeweils andere übertragen kann.

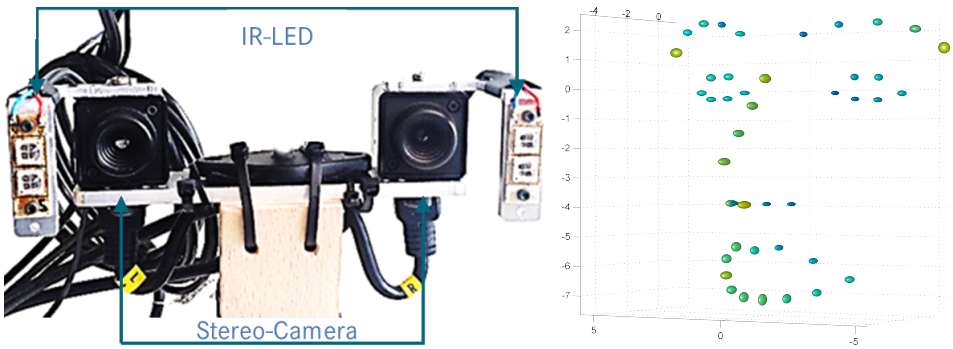

Abb. 3.3 Links Beschreibung Fovio-Stereo-System
Rechts Intraindividuumstandardabweichung des Stereo-Systems. Angaben in cm.

Die ermittelten 3D-Punkte werden also relativ zur Slave-Kamera dargestellt. Das Bild dieser Kamera wird nun als zusätzlicher Datenterm benutzt. Dabei liegt die Annahme zugrunde, dass die ermittelten 3D-Punkte der Master-Kamera, projiziert in das Koordinatensystem der Slave-Kamera, dem von dieser aufgenommenen 2D-Bild entsprechen.
Es ist durchaus denkbar, für beide Bilder mittels der der zuvor erwähnten HOG-Deskriptoren 2D-Punkte zu ermitteln und diese auf eine sogenannte *Epipolarlinie* im jeweils anderen Koordinatensystem zu projizieren. Der Schnittpunkt beider Linien würde dann den 3D-Punkt ergeben. Dieser Ansatz ist zwar der übliche für vollständige 3D-Rekonstruktionen, in dem hier beschriebenen Fall ist jedoch statistisch-

biometrisches Wissen verfügbar, das sich so nicht ganzheitlich einarbeiten lassen würde.
Außerdem ist die Gesichtsdetektion ein sehr aufwendiges Verfahren, das notwendig ist, um überhaupt einen Bildausschnitt zu finden, der ein Gesicht enthält. Dieses Verfahren ist notwendig, um die HOG-Deskriptoren für die eigentlichen Landmarken zu initialisieren. Wird das Wissen der Master-Kamera ausgenutzt, dann wird das aufwendige Suchen des Gesichts unnötig, da bereits präzise Informationen über die Position der einzelnen Landmarken verfügbar sind.
Sicherlich ist es auch möglich, das System noch weiter zu verbessern, da zwei Kameras mehr aufzeichnen als nur eine. Doch auch in dem Fall, dass nur eine Kamera einen Gesichtspunkt sieht, lässt sich dieser noch besser lokalisieren als über ein reines Mono-System. Da die Fragestellung sich aber auf identifizierende, übertragbare Merkmale richtet, sind derartige Finessen erst einmal zweitrangig. Es kann aber davon ausgegangen werden, dass das kommerzielle Fovio-Stereo-System über solche Verfahren und Verbesserungen verfügt und dass diese auch zum Einsatz kommen.
Bei einer Betrachtung der aus diesem System entstehenden intraindividuellen Standardabweichungen (Abbildung 3.3) und einem Vergleich dieser mit denen des Mono-Systems (Abbildung 3.2) lässt sich feststellen, dass insbesondere die Abweichung auf der Z-Achse im Stereo-System deutlich höher ist. Allgemein kann festgehalten werden, dass die Standardabweichung im Stereo-System größer als im Mono-System ist. Dies mag paradox erscheinen, da im Stereo-System doch deutlich mehr Informationen verfügbar sind und es somit auch präziser sein sollte. Der entscheidende Aspekt ist jedoch, dass es sich um tatsächlich gemessene Punkte handelt, die durch zwei Datenterme und einen Modellterm entstanden sind. Das Verfahren kann also viel mehr Gewicht auf das Gemessene legen, als es beim Mono-System möglich wäre. Das Mono-System muss sich stark am internen Gesichtsmodell orientieren, um konstante 3D-Punkte zu erzeugen, denn es liegen eigentlich zu wenig Informationen vor, um einen 3D-Punkt zu ermitteln. Vereinfacht kann gesagt werden, dass das Mono-System seine Tiefeninformationen vollständig aus dem Modell ableitet.
Für den Anwendungsfall der heterogenen Gesichtserkennung lässt sich bemerken, dass es dabei zu sehr unterschiedlichen Ergebnissen kommt, obwohl ähnliche Software genutzt wird, bei der das gleiche Gesichtsmodell zugrunde liegt. Es ist also nicht so, dass die hier ermittelten übertragbaren Features auf exakt vergleichbare Art messbar sind und dass die zwei Systeme derart unterschiedlich wären, dass sie alle Definitionsteile der heterogenen Gesichtserkennung erfüllen.
Bevor jedoch verschiedene Distanzermittlungen erläutert werden, die tatsächlich in der Lage sind, mit derart verschiedenen und doch gleichen Punktwolken umzuge-

hen, sollte das zugrunde liegende Gesichtsmodell detaillierter erläutert werden. Die hier besprochene Repräsentation durch charakteristische Punkte ist durchaus nicht die einzige Methode, um die digitale Form eines Gesichts zu erhalten.

3.2.2 3D Morphable Model

Da es nicht ohne Weiteres möglich ist 3D-Daten aus einem 2D-Bild zu extrahieren wird ein statistisches Gesichtsmodell als Prior genutzt, so dass sich das Problem auf die Anpassung des Modells an das gesehene 2D-Bild beschränkt. Das statistische Modell ist durch die Koeffizienten der Hauptkomponenten parametrisiert. Es wäre also denkbar zufällige Startwerte zu wählen und dann, durch eine orthographische Projektion, den Abstand zum Eingangsbild minimieren. Diese Art von Algorithmen nennt man *Analysis-by-Synthesis* Verfahren, da lediglich von 3D auf 2D projiziert wird ohne das 2D-Bild zu Analysieren. Leider haben diese Methoden den Nachteil, dass sie dazu neigen, nicht das globale Minimum zu erreichen und dadurch abhängig von den zufällig gewählten Startwerten sind. Die Optimierung ist folglich nicht deterministisch und kann nicht zur Identifikation herangezogen werden.(Bas et al., 2016)
Auf der anderen Seite gibt es die Feature basierten Verfahren, welche ihre initialen Parameter passend zu gefundenen Gesichtslandmarken wählen, so dass keine Abhängigkeit von zufällig gewählten Startparametern besteht. Die korrespondierenden Positionen der Gesichtslandmarken welche im 2D-Bild gefunden werden können, sind innerhalb des 3DMMs bekannt, weshalb es möglich ist, detektierte Kanten im Eingangsbild auf der 3D-Oberfläche des 3DMMs nachzuvollziehen. Das geschieht unter der Annahme, das 2D-Bildkanten mit einer Änderung der Form und dadurch einer Veränderung der Lichtreflexion einhergehen.(Bas et al., 2016) Im einfachsten

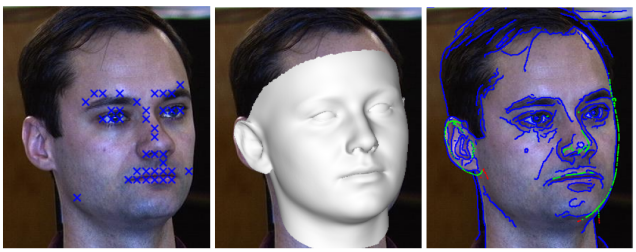

Abb. 3.4 3DMM-Optimierung von (Bas et al., 2016)

Fall kann nach der Initialisierung mit Hilfe der Gesichtlandmarken und der anschließenden Projektion der Gesichtsform durch eine skalierte orthographische Projektion eine iterative Lösung durch eine Singulärwertzerlegung gefunden werden. Die meisten Methoden benutzen jedoch zusätzlich eine *Iterative Closest Edge* Optimierung,

dass heißt die gefunden, korrespondierenden Kanten zwischen dem Eingangsbild und der Gesichtsform, werden sukzessive zur weiteren Verfeinerung eingebunden. Dieser Vorgang ist nichtlinear und mathematisch schwer zu formulieren und optimieren.
3DMMs werden hauptsächlich eingesetzt um Gesichtsbewegungen und Rotationen zu analysieren und zu verändern. Ob es die zugrundeliegenden Koeffizienten der Hauptkomponenten ausreichen um heterogene Gesichtsaufnahmen zu identifizieren bleibt fraglich, da sie lediglich die am besten passende Form zum gegebenen Eingangsbild widerspiegeln.(Bas et al., 2016)

3.2.3 Distanzmetriken

Nachdem verschiedene Aufnahmesysteme vorgestellt wurden, die in der Lage sind, Gesichtslandmarken dreidimensional zu lokalisieren, stellt sich die Frage, wie daraus ein identifizierendes Abstandsmaß abgeleitet werden kann. Zuerst werden deshalb geometrische Verfahren, wie die Procrustes- oder Hausdorff-Distanz vorgestellt. Bei dieser Kategorie von Verfahren hat der ermittelte Abstand eine geometrische Bedeutung. Genauer gesagt handelt es sich also um einen metrischen Abstand, der im engeren Sinne keine Rückschlüsse hinsichtlich der Unterscheidung von Individuen zulässt. Beispielsweise wäre ein halber Zentimeter Unterschied bei der Ohrgröße eine durchaus signifikante Differenz, aber die Biologie lehrt, dass Ohren nicht aufhören zu wachsen und dass somit bei einem Individuum eine derartige Abweichung durchaus nicht ausgeschlossen werden kann.
Aus diesem Grund ist es von Vorteil, ein biometrisches Abstandsmaß zu errechnen, also ein Abstandsmaß, das in eine Referenzgruppe eingeordnet werden kann und dessen Differenz ein statistisches Maß für Unterscheidbarkeit darstellt.

3.2.3.1 Geometrisch

Es gibt verschiedenste geometrische Verfahren. Teilweise finden diese Anwendung in der Registrierung von Punktwolken und sollen verhindern, dass schlechte Aufnahmen das Gesamtergebnis zu sehr verfälschen. Meist sind derartige Methoden jedoch, wie z. B. die zuvor erwähnte *Procrustes-Distanz*, an ein Verfahren gekoppelt, welches die Aufnahmen zunächst derart verformt, dass sich die zu vergleichenden Punktmengen maximal überdecken. Es gibt aber auch Verfahren, wie zum Beispiel

die *Hausdorff-Distanz*, welche ein reines Abstandsmaß darstellen und je nach Anwendungsfall um eine Registrierung erweitert werden müssen. Auch diese Registrierungen können in globale und lokale unterschieden werden. Wird zum Beispiel eine einzige Transformation für alle Punkte gesucht, wie zum Beispiel bei der Procrustes-Distanz, handelt es sich um eine globale Registrierung. Wird hingegen jeder Punkt für sich genommen verarbeitet, wie es zum Beispiel beim *Iterative Closest Point*(ICP)-Verfahren der Fall ist, handelt es sich um eine globale Methode.

3.2.3.1.1 Procrustes-Distanz Zunächst werden die Procrustes-Distanz und die dazugehörige Iterative-Procrustes-Distanz näher beleuchtet. Unschwer ist zu erkennen, dass der Name aus dem Griechischen stammt und einen Banditen der Mythologie bezeichnet, der seine Opfer an sein Bett anpasste. Ganz in diesem Sinne wird nun versucht, zwei Punktmengen derart aufeinander zu transformieren, dass die Summe der euklidischen Abstände minimal ist. Die Punktwolken aus Abbildung 3.5 dienen hierbei exemplarisch als Ausgangssituation. Es handelt sich um tatsächlich gemessene 3D-Punkte, die aus Gründen der Übersichtlichkeit in 2D dargestellt sind.

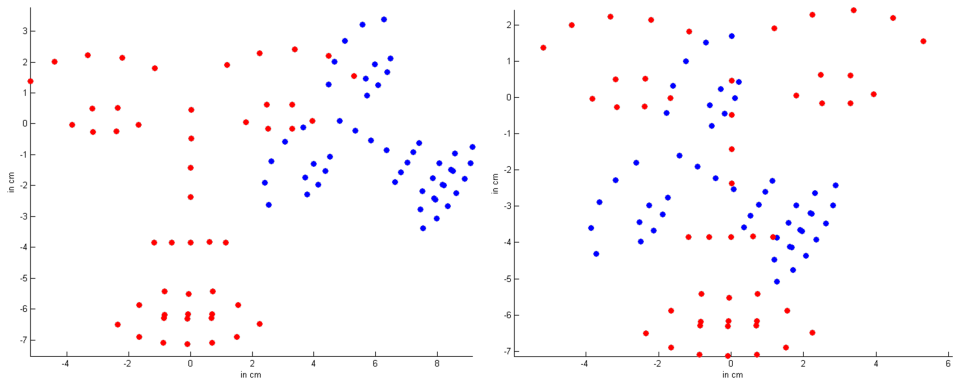

Abb. 3.5 Links Ausgangssituation für die Procrustes-Distanz, $\sum L_2 = 1,17$
Rechts Nach Translationm $\sum L_2 = 1,00$

In einem ersten Schritt wird die Translation gesucht, welche die Summe der euklidischen Abstände minimiert. Diese findet sich, indem von beiden Punktmengen der Durchschnitt(Zentroid) bestimmt wird. Die Differenz beider Zentroide ist der ge-

wünschte Vektor (siehe Gleichung 3.5). Wird Abbildung 3.5 links wie beschrieben transliiert, entsteht rechts Abbildung 3.5.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_k}{k} \tag{3.2}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_k}{k} \tag{3.3}$$

$$\bar{z} = \frac{z_1 + z_2 + \cdots + z_k}{k} \tag{3.4}$$

$$(x, y, z) = (x - \bar{x}, y - \bar{y}, z - \bar{z}) \tag{3.5}$$

Nach der Translation folgt die Skalierung, sodass die Varianz beider Punktmengen 1 ist. Dies ist mathematisch einfach auszudrücken, wie Gleichung 3.7 zeigt. Das Ergebnis dieser Transformation ist in Abbildung 3.6 links zu sehen. Das Skalieren auf Varianz 1 hat den Vorteil, dass das Residuum (Procrustes-Distanz) unabhängig von dem Volumen der Punktwolken ist. Anders ausgedrückt ist die Procrustes-Distanz sowohl einheitslos als auch skaleninvariant und stellt somit ein einheitliches Abstandsmaß dar.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (y_1 - \bar{y})^2 + \cdots}{k}} \tag{3.6}$$

$$(x_1, y_1, z_1) = \left(\frac{(x_1 - \bar{x})}{s}, \frac{(y_1 - \bar{y})}{s}, \frac{(z_1 - \bar{z})}{s} \right) \tag{3.7}$$

Danach fehlt nur noch der Rotationsausgleich. Die Rotation berechnet sich etwas komplizierter. Wenn – wie in diesem Fall – die Daten mehr als zwei Dimensionen haben, muss eine *Singular Value Decomposition*(SVD) zu Hilfe genommen werden. Es ist für ein ausreichendes Verständnis nicht erforderlich, die SVD im Detail zu erklären, weshalb an dieser Stelle darauf verzichtet wird. Das Endergebnis wird in Abbildung 3.6 rechts dargestellt.

Es gibt auch eine Erweiterung, die sogenannte Iterative-Procrustes-Distanz . Hierbei geht es um die Aufarbeitung von Messdaten. Zum Beispiel handelt es sich bei Gesichtslandmarken um Daten, die in einer Frequenz von 50 Hz errechnet werden. In Rücksicht auf die drei Sekunden aus der Anforderungsanalyse ergeben sich 150 Punktmengen. Bei der Ermittlung des Durchschnitts ist es natürlich möglich, dass Ausreißer diesen verzerren. Iterative Procrustes löst dieses Problem, indem es alle Punktmengen aufeinander registriert. Messungen, bei denen das Residuum einen Grenzwert überschreitet, werden verworfen. Danach wird ohne diese Ausreißer erneut die Procrustes-Distanz errechnet. Dies geschieht so lange, bis es keine Ausreißer mehr gibt.

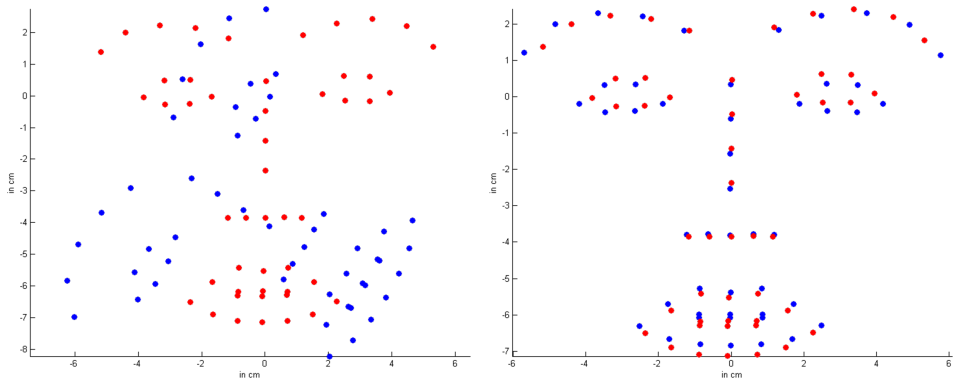

Abb. 3.6 Links Nach Skalierung, $\sum L_2 = 0,50$
Rechts Nach Rotation, Procrustes-Distanz $= \sum L_2 = 0,01$

3.2.3.1.2 Hausdorff-Distanz Die Hausdorff-Distanz, benannt nach ihrem Erfinder, dem Mathematiker Felix Hausdorff, hat keinen integrierten Registrierungsprozess. Sicherlich ließe sich Procrustes durchführen und danach die Hausdorff-Distanz messen oder zuvor das bereits erwähnte Verfahren Iterative Closest Point anwenden. Es sollte aber bedacht werden, dass mit jeder Anpassung des Datensatzes nicht nur Verzerrungen minimiert, sondern unter Umständen auch identifizierende Abstände ausgewaschen werden. Da die Fovio-Systeme bereits die Kopfrotation normalisieren und jede Gesichtslandmarke relativ zur Kopfposition angeben, ist es möglich, die durchschnittliche Position jeder Landmarke pro Proband heranzuziehen und die daraus resultierenden Features direkt mittels der Hausdorff-Distanz zu vergleichen.

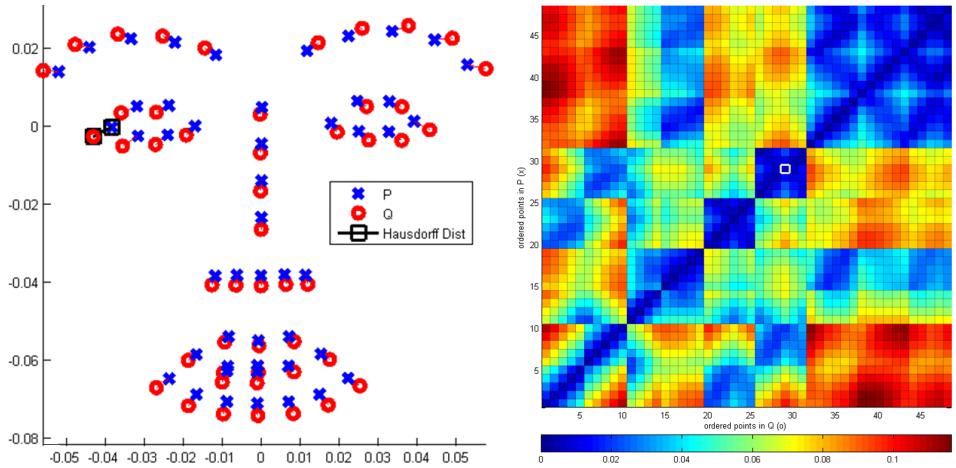

Abb. 3.7 Links Hausdorff-Plot, schwarze Kästen markieren Distanz.
Rechts Distanz-Matrix. Weißer Rahmen zeigt Hausdorff-Distanz

Mathematisch gesehen handelt es sich um das Maximum der minimalen Abstände. Es wird also der euklidische Abstand von jedem Punkt aus der ersten Punktmen-

ge zu jedem Punkt aus der zweiten Punktmenge errechnet. Jedem Punkt aus der ersten Punktmenge werden genauso viele Distanzen zugeordnet, wie es Punkte in der zweiten Punktwolke gibt. Aus dieser Menge wird nun die minimale Distanz für jeden Punkt ausgewählt und das Maximum dieser Menge ist die Hausdorff-Distanz. Eine mathematische Beschreibung ist Gleichung 3.8 zu entnehmen.

$$\delta(\boldsymbol{P}, \boldsymbol{Q}) = \max_{p \in P} \left\{ \min_{q \in Q} \left\{ (p - q)^2 \right\} \right\} \tag{3.8}$$

Ein Blick auf Abbildung 3.7 zeigt links die in 2D projizierten zwei Punktmengen P und Q. Die zwei Punkte am rechten Auge, innerhalb des schwarzen Rahmens, bilden die Hausdorff-Distanz. Auf der rechten Seite sind alle Distanzen erkennbar. Die Hausdorff-Distanz ist hierbei weiss umrandet.

3.2.3.2 Biometrisch

Bisher wurden verschiedene geometrische Abstände behandelt, die nur eine bedingte Aussagekraft für die Unterscheidung zweier Individuen haben. Es fehlt eine statistische Einordnung. Würde zum Beispiel mit dem Fovio-System eine große Probandenstudie durchgeführt werden, sodass für jede Gesichtslandmarke eine *Point Distribution Function*(PDF) ermittelt werden kann, dann könnte eine Messung statistisch eingeordnet werden. Im engeren Sinne ließe sich eine Wahrscheinlichkeit errechnen, wie oft bei zwei Menschen die gleiche Konstellation von Abweichungen auftritt. Weiterhin ist es geboten, wie bereits in Abschnitt 3.2.2 erwähnt, Hauptrichtungen für die Varianz herauszuarbeiten. Es wäre zum Beispiel möglich, weitergehendes Wissen zu ermitteln, wie z. B. das Alter oder das Geschlecht. In Abschnitt 3.2.2 wurde gezeigt, mit welchen Verfahren ein 3D-Morphable-Model an ein Bild angepasst werden kann. Da die Aufnahme von großen Datenmengen mit dem Fovio-System, welche zudem ethnisch repräsentativ sind, aus Zeitgründen nicht durchführbar ist, wird im Folgenden aufgezeigt wie die Informationen eines beliebigen Modells für die Gesichtserkennung mit den Fovio-Gesichtslandmarken genutzt werden können. Das hat zudem den Vorteil der erweiterten Übertragbarkeit, wie es bereits ausführlich in Abschnitt 3.2.2 dargelegt wurde.
Als Beispiel dient das Basel-Face-Model von (Paysan et al., 2009). Grundsätzlich lässt sich sagen, dass im Folgenden versucht wird, mithilfe der hoch präzisen, aus mindestens 150 Bildern entstandenen Gesichtslandmarken das gesamte Gesicht zu rekonstruieren. Natürlich ist es nicht möglich, mit 68 Punkten alle 199 Koeffizienten

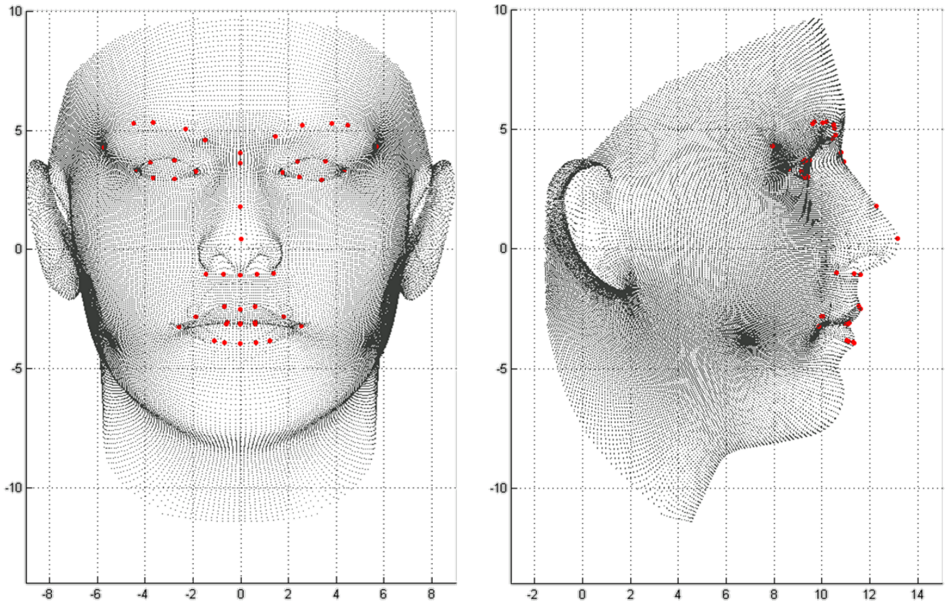

Abb. 3.8 Links Basel-Face-Modell, frontal, mit Fovio-Markern in rot.
Rechts Basel-Face-Modell, seitlich, mit Fovio-Markern in rot.

des Basel-Face-Modells zu finden. Es ist aber sehr wohl möglich, genauso viele Koeffizienten zu finden, wie Punkte ermittelt wurden. Dadurch dass die Eigenwerte der Hauptkomponentenbasis mit dem Index des Koeffizienten abnehmen, ist es möglich, das Residuum zu bestimmen. Dieses Residuum entspricht dann auch exakt der Ungenauigkeit dieses Ansatzes. Es wird an dieser Stelle angenommen, dass dieses Residuum klein genug ist und eine Identifikation nicht verhindert. Weiterhin hat dieser Ansatz den Vorteil, dass er Konstellationen von Abweichungen codieren und nutzen kann und nicht auf einzelnen Landmarken basiert, wie zum Beispiel die Hausdorff-Distanz.

Es seien also die die ersten 68 Hauptkomponenten des Basel-Face-Models , die dazugehörigen Mittel- und Eigenwerte sowie die aus den Fovio-Systemen gesammelten 3D-Daten herangezogen. Außerdem seien die Grundlagen der Hauptkomponentenanalyse an dieser Stelle Voraussetzung, da sie bereits in Abschnitt 3.2.2 ausführlich abgehandelt wurden. Zunächst muss eine Korrespondenz zwischen den Punkten des Basel-Face-Models und denen des Fovio-Systems gefunden werden. Eine derartige Zuordnung wurde manuell durchgeführt und kann Abbildung 3.8 entnommen werden. Danach können durch das Lösen folgender Gleichung 3.9 die gewünschten

Koeffizienten gefunden werden.

$$-\underbrace{\begin{pmatrix} \mu_1^x & \mu_1^y & \mu_1^z \\ \vdots & \vdots & \vdots \\ \mu_k^x & \mu_k^y & \mu_k^z \end{pmatrix}}_{\text{Mittelwerte}} + \underbrace{\begin{pmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_k & y_k & z_k \end{pmatrix}}_{\text{Fovio-Messpunkte}} = \underbrace{\begin{pmatrix} c_1^x & c_1^y & c_1^z \\ \vdots & \vdots & \vdots \\ c_k^x & c_k^y & c_k^z \end{pmatrix}}_{\text{Koeffizienten}} \underbrace{\begin{pmatrix} E_1^x & \dots & E_k^x \\ E_1^y & \dots & E_k^y \\ E_1^z & \dots & E_k^z \end{pmatrix}}_{\text{Eigenvektoren}} \tag{3.9}$$

Es ist hierbei wichtig, dass die Eigenvektoren mit den Eigenwerten multipliziert werden, was oft unterlassen wird. In diesem Fall ist dies aber enorm wichtig, um eine Gewichtung mit der Varianz zu erreichen. Ein weiterer Vorteil ist die Interpretierbarkeit der Koeffizienten. Da (Paysan et al., 2009) Alter, Gewicht, Geschlecht und Herkunft der in der Basis enthaltenen Individuen kannten, war es ihnen möglich, Richtungen anzugeben, die diesen Eigenschaften entsprechen. Es ist nun also nicht nur möglich, zu identifizieren, sondern auch zu analysieren. Erweiterungen, die diese Eigenschaften mit einbeziehen und somit für eine robustere Identifikation sorgen, sollten jedoch vorerst zurückgestellt werden, da es sich hierbei um Feinheiten handelt.

3.3 Synthesis

Im vorigen Abschnitt war der NIR/VIS-Anwendungsfall nicht sonderlich fordernd, da davon ausgegangen werden darf, dass Gesichtslandmarken sowohl im sichtbaren als auch im Infrarotspektrum gefunden werden können. Als problematisch stellte sich hingegen der 2D/3D-Anwendungsfall heraus. Im Folgenden wird nun das Hauptaugenmerk auf die NIR/VIS-Gesichtserkennung gelegt.
In der homogenen Gesichtserkennung sind die Texturverfahren, welche auf ein CNN aufbauen, unangefochten führend. Sie sind deutlich robuster und genauer als vergleichbare Verfahren, die auf manuell ausgewählten Merkmalen beruhen. Ziel dieses Abschnitts ist es, verschiedene Systeme zu entwickeln, die in der Lage sind, Bilder zwischen beiden Modalitäten zu transformieren. Insbesondere wird darauf Wert gelegt, dass der homogene Klassifikator nicht angepasst wird. Es handelt sich folglich um einen – in der Anforderungsanalyse bereits behandelten – Modalitätswrapper, der vor der eigentlichen Erkennung angewandt wird.
Zuerst wird ein kurzer Überblick über die wenigen verfügbaren Datenbasen für dieses Vorhaben gegeben, bevor verschiedene Arten der Datenaufbereitung dargestellt

werden. Im Anschluss werden zwei verschiedene Netze mit den zuvor aufbereiteten Daten trainiert, bevor eine Klassifikation durch Opensource-Gesichtserkennungsverfahren durchgeführt wird. Es handelt sich also um eine reine Halluzination.

3.3.1 Verfügbare Datenbasis

Wie bereits erwähnt, ist die Datenbasis für ein solches Vorhaben sehr klein, da auch der Anwendungsfall noch eine Forschungsnische mit Wachstumspotenzial ist. Der Aufbau einer eigenen Datenbank ist zeitaufwendig und im Rahmen dieser Arbeit nicht machbar. Das einzige Forschungszentrum, das mehrere Datenbasen zur Verfügung stellt, ist das *Center for Biometrics and Security Research*(CBSR) der *Chinese Academy of Sciences*(CASIA). Drei der dort veröffentlichten Sammlungen sind hierbei besonders interessant:

CASIA-NIR-VIS umfasst 80 Personen (73,2% männlich) mit jeweils 10 Bildern in NIR und VIS, welche auf Kommando verschiedene Gesichtsausdrücke(Angst, Ekel usw.) imitieren. Zur Datenbank existiert keine wissenschaftliche Arbeit.

CASIA-NIR-VIS-2.0 enthält 725 Personen, verteilt auf etwa 5000 VIS- und 12000 NIR-Bilder. Es existiert eine wissenschaftliche Arbeit, die den Aufnahmeprozess beschreibt (Li et al., 2013). Zusätzlich steht ein Evaluationsprotokoll zur Verfügung, um Ergebnisse verschiedener Verfahren vergleichen zu können.

CASIA-HFB umfasst 100 Personen (57% männlich) mit jeweils 4 Bildern pro Modalität. Es sind sowohl NIR-Bilder als auch VIS-, TIR(Thermal Infrared)- und Tiefenbilder enthalten. Die Datenbasis wird von (Li et al., 2009) näher beschrieben.

Im Folgenden wird nur die Datenbasis von (Li et al., 2013) genutzt. Sie umfasst einerseits die meisten Bilder und Individuen, andererseits wird sie auch in der aktuelleren Literatur als Standard angesehen, sodass die hier erhaltenen Resultate vergleichbar werden.

3.3.2 Datenaufbereitung

Da die Datenbank von (Li et al., 2013) darauf ausgelegt ist, heterogene Verfahren zu evaluieren, eignen sich ihre Bilder wenig, um eine Halluzination zu trainieren. Da aber, wie im vorigen Abschnitt aufgezeigt wurde, keine anderen Datensätze existieren, werden die Bilder aufbereitet.

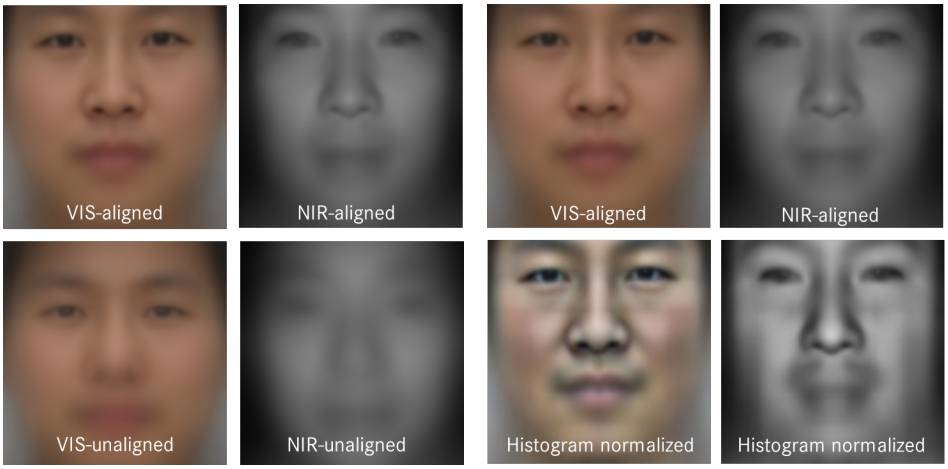

Abb. 3.9 Links Mittelbild der Datenbank mit verschiedenen Registrierungen. Rechts Mittelbild mit verschiedenen Normalisationen.

Das Problem hierbei ist vielfältig. Einerseits sind die Modalitäten nicht deckungsgleich, andererseits sind auch die Bildhelligkeit und der Gesichtsausdruck der Probanden nicht eingeschränkt. Es gibt folglich drei Probleme, die gelöst werden müssen, um aus der NIR-VIS-2.0-Datenbank eine Trainingsbildmenge zu extrahieren.
Das Erste ist die Korrespondenz. Alle Bilder müssen pixelweise übereinstimmen. Das bedeutet, dass jegliche Form von Verzerrung, Kopfposition und Gesichtsausdruck vereinheitlicht werden muss. Dieser Prozess wird Registrierung genannt (siehe Abbildung 3.11).
Zweitens muss das Histogramm aller Bilder derart ausgeglichen sein, dass die Intensität eines Pixels übertragbar ist. Dieses Problem ist unter dem Begriff Histogrammnormalisation bekannt (siehe Abbildung 3.9). Es darf keine zu dunklen Aufnahmen geben, auch keine Gesichter, die durch verschiedene Lichtquellen unterschiedliche Schatten werfen.
Das letzte Problem, das behandelt werden muss, ist die Entwicklung eines Gütekriteriums. Es muss möglich sein, Bilder oder Datenpaare auszusortieren, nachdem sie den Extraktionsprozess durchlaufen haben und immer noch nicht die oben genannten Kriterien erfüllen. Diese Fragestellung wird im Folgenden unter dem Namen Ausreißerkontrolle geführt.

3.3.2.1 Geometrisch rektifiziert

Bei den meisten Gesichtsklassifikationsverfahren können nur Bilder verarbeitet werden, die zuvor geometrisch rektifiziert wurden. Genauer gesagt muss eine räumliche Übereinstimmung einzelner Gesichtsmerkmale vorliegen, um optimale Ergebnisse zu erhalten. Es ist also naheliegend, diese für das Problem der Trainingsdatensatzerzeugung zu verwenden.

Abb. 3.10 Links Erfolgreiche Registrierung.
Rechts Fehlgeschlagene Registrierung.

Der von Openface verwendete HOG-Deskriptor wird dazu benutzt, die Mundwinkel sowie die äußersten Punkte der Augen zu finden und diese auf eine einheitliche Position zu transformieren. Wie in Abbildung 3.10 zu sehen, kann es dabei zu unschönen und unnatürlichen Verzerrungen kommen, die vom Einzelfall abhängen und deshalb nicht mit der Modalität korrelieren. Mit anderen Worten: Eine pixelweise Korrespondenz, wie sie in der Problemstellung formuliert wurde, kann nur für die zuvor genannten Stützstellen garantiert werden. Weiterhin muss mit einem durch Interpolation hervorgerufenen Fehler gerechnet werden. Die eigentlichen Gesichtsklassifikatoren sind gegen solche Verzerrungen robust, da sie diese bereits im Training gesehen haben, eine Halluzination, bei der lediglich die Farbe gelernt werden soll, die jedoch nicht die Geometrie verändern soll, kann derartiges schwerlich handhaben.
Die Einfachheit dieses Ansatzes lässt es interessant erscheinen, diesen weiterzuverfolgen und zu analysieren. Eine klassische Halluzination mittels eines Autoencoders (siehe Unterunterabschnitt 3.3.3.1) kann nicht erfolgreich sein, ein Adversarial-Netz (siehe Unterunterabschnitt 3.3.3.2), das keine Pixel transformiert, sondern versucht, ein möglichst echtes Gesamtbild zu erzeugen, unter Umständen schon.

3.3.2.2 Patch basiert

In diesem Abschnitt wird die Arbeit von (Lezama et al., 2016) reproduziert und an mehreren Stellen verbessert. Wie zuvor bereits erwähnt, registriert (Lezama et al., 2016) die Bilder zunächst paarweise durch eine affine Transformation, die anhand von 2D-Gesichtslandmarken errechnet werden. Hier wird die Erkenntnis aus Abschnitt 3.2 verwendet, um diesen Ansatz zu verbessern. Anstatt eine zweidimensionale, affine Transformation aus 2D-Punkten zu errechnen, wird das zuvor erarbeitete 3D Modell verwendet, um eine dreidimensionale, projektive Transformation zwischen beiden Bildern zu finden. Weiterhin werden von (Lezama et al., 2016) lediglich 6 Punkte verwendet, um die Transformation zu errechnen, wohingegen in dieser Arbeit 40 Punkte zum Einsatz kommen. Das Gleichungssystem ist in diesem Fall über bestimmt und kann somit leichter falsch detektierte Landmarken handhaben.

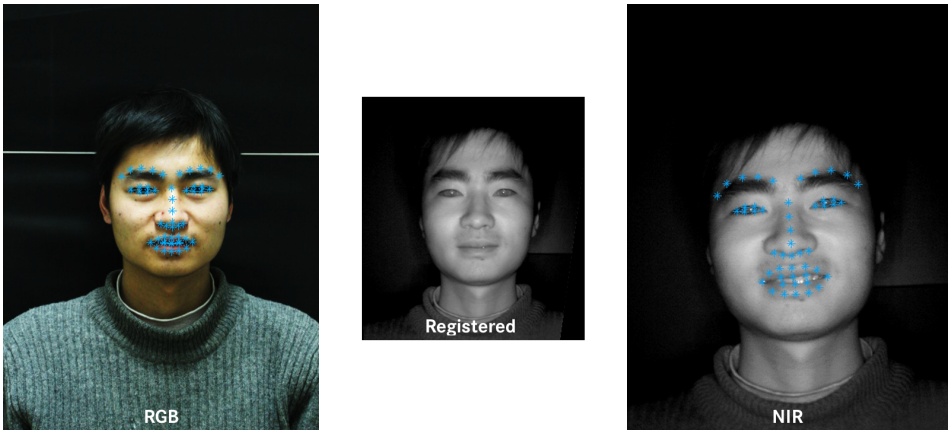

Abb. 3.11 Links Texturextraktion in VIS, VP07
Rechts Texturextraktion in NIR, VP07

Zunächst wird also sowohl im NIR-Bild als auch im VIS-Bild die Position des Gesichts detektiert, bevor die Gesichtslandmarken auf die Gegebenheiten optimiert werden. Zusehen ist dies in Abbildung 3.11 links für VIS und in der Mitte für NIR. Die 3D-Gesichtslandmarken sind in blau eingezeichnet. Danach wird mithilfe dieser Landmarken eine projektive Transformation errechnet, die das NIR-Bild auf das VIS-Bild transformiert. Das Ergebnis ist in Abbildung 3.11 rechts zu sehen. Es hat sich als vorteilhaft erwiesen, das NIR-Bild auf das VIS-Bild zu transformieren, da auf den VIS-Bildern deutlich weniger Varianz im Hinblick auf den Gesichtsausdruck und die Kopfrotation herrscht.
Im Anschluss wird die Bildhelligkeit beider Modalitäten durch eine Histogrammnor-

malisation ausgeglichen, bevor Patches, also Bildsubregionen, aus dem Bildpaar extrahiert werden. Hierbei kommt auch der YCbCr-Farbraum zum Einsatz, in welchen das VIS-Bild überführt wird. Dies hat den Hintergrund, dass in diesem Farbraum die Farb- und die Helligkeitsinformationen voneinander entkoppelt sind, wie es bereits in Unterunterabschnitt 2.3.2.1 ausführlich dargestellt wurde.

Abb. 3.12 Verkleinerung des Ausschnitts um nicht definierte Bereiche (schwarz) zu vermeiden.

Auf dieser Basis werden nun mit einer Schrittweite von 32 Pixeln, 64×64 große Patches erzeugt. Diese werden wiederum paarweise affin derart aufeinander registriert, dass die Kreuzkorrelation zwischen dem Luminanzkanal des VIS-Bildes und des NIR-Bildes maximal ist. Danach wird das Patchpaar auf 40×40 Pixel zurechtgeschnitten. Der Rand wird folglich verworfen. Es findet keine Interpolation statt. Das Verwerfen der Randregion ist notwendig, um eventuell durch die affine Transformation entstandene, nicht definierte Bildbereiche zu entfernen (siehe Abbildung 3.12). Abschließend wird ein Gütekriterium angewandt, um zu entscheiden, ob das Verfahren erfolgreich war und ob mithin das Patchpaar in den Trainingsdatensatz aufgenommen werden kann. Hier dient wieder die Kreuzkorrelation als Entscheidungsmaß. (Lezama et al., 2016) verwenden nur Patchpaare, bei welchen die Summe der Kreuzkorrelation zwischen VIS-Y und NIR sowie deren Gradienten mehr als 1 betragen. Liegt hingegen einer der beiden Faktoren unter 0.5, dann wird das Paar ebenfalls verworfen.
(Lezama et al., 2016) konnte mit seinem Verfahren und unter Anwendung der eben genannten Schwelle etwa 600.000 Patchpaare extrahieren. Das hier beschriebene Verfahren kann mit dem gleichen Schwellwert etwa 900.000 Patchpaare, 1.5 Mal so viel wie (Lezama et al., 2016), generieren. Es ist aber nicht zu erwarten, dass durch diese Verbesserung der Datenaufbereitung auch die Genauigkeit um den Faktor 1.5 steigt.

3.3.2.3 Texturunwrapping

In diesem Teil werden nicht wie in (Lezama et al., 2016) und dem vorigen Abschnitt quadratische Bildpatches extrahiert, sondern es kommt eine vollständige Gesichtsmaske zum Einsatz. Das Surrey-Face-Mesh besteht aus 3443 3D-Punkten und den von diesen aufgespannten Oberflächen, auch *Triangles* genannt.
Dieser Ansatz lässt sich auf verschiedene Arten mit den vorangegangen Verfahren vergleichen. Einerseits könnte er als Weiterentwicklung der in Unterunterabschnitt 3.3.2.1 vorgestellten Methode interpretiert werden, wobei die Anzahl der Gesichtslandmarken der Anzahl der Punkte der Gesichtsmaske entspricht. Andererseits ist es denkbar, die Triangles als Patches und damit als Erweiterung zu Unterunterabschnitt 3.3.2.2 zu sehen.

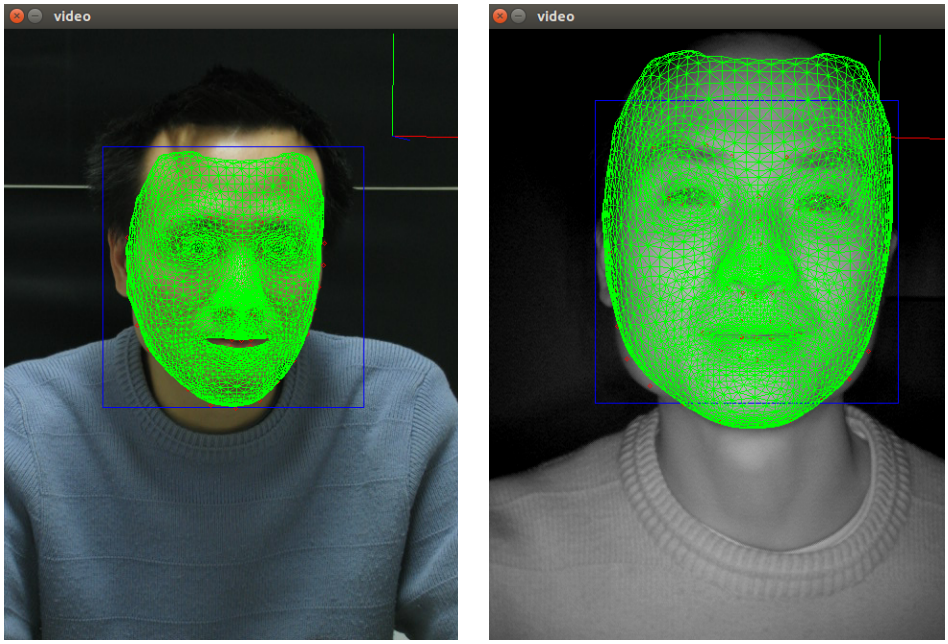

Abb. 3.13 Links Texturextraktion in VIS, VP07
Rechts Texturextraktion in NIR, VP07

Das Surrey-Face-Mesh ist aus Abschnitt 3.2.2 bereits bekannt. Dieses Mal wird es jedoch nicht dazu genutzt, Forminformationen zu extrahieren, sondern um genau diese zu entfernen. Die Textur des Individuums soll nicht von 3D zu 2D projiziert, sondern vielmehr abgerollt werden. Dieser Vorstellung zuträglich ist der Gedanke der Häutung des Gesichts, da genau dieses mathematisch passiert. Dazu wird zuerst das Modell wie in Abschnitt 3.2.2 auf alle Bilder der Datenbank angewandt, und es wird eine sogenannte UV-Map für jedes Bild generiert. Jeder 3D-Punkt der Gesichtsmaske hat eine UV-Koordinate. Die Bildregion, welche mit den benachbarten Faces

korrespondiert, wird zwischen den eingrenzenden UV-Koordinaten in die Isomap eingetragen (siehe Abbildung 3.13).
Dieses Vorgehen hat weitere Vorteile gegenüber der Projektion. Zum einen ist es möglich, diejenigen Regionen, die im Bild nicht sichtbar sind – das heißt Faces, deren Oberflächennormale von der Kamera weg zeigen – auszulassen. Zum anderen kann das Surrey-Face-Model verschiedene Gesichtsausdrücke erkennen und diese neutralisieren. Weiterhin könnten in einem erweiterten Anwendungsfall die Mehrinformationen, die ein Video mit sich bringt, in einem Bild vereint werden, indem die Textur durch die Gesichtsmaske aufaddiert wird. Das hat insbesondere im Hinblick auf das Oberflächenwinkelkriterium den Vorteil, dass immer nur die gerade gut sichtbaren Gesichtsbereiche in die Textur kombiniert werden würden. Das kann lässt sich zum Beispiel an den schwarzen Bereichen in Abbildung 3.14 gut erkennen.

Abb. 3.14 Links VIS-3D-View und Isomap
Rechts NIR-3D-View und Isomap

Es kann jedoch durchaus auch passieren, dass das Verfahren fehlschlägt und dass somit ein falsches Trainingspaar erzeugt wird. Wie bereits in der Problemstellung formuliert, müssen hier nach einem tauglichen Gütekriterium eventuell erzeugte Fehler erkannt und aussortiert werden, weshalb ein Gesichtsdetektor auf die Isomap angewendet wird. Findet dieser Detektor kein Gesicht, gilt es als Fehlschlag und wird nicht in den Trainingsdatensatz aufgenommen.
Es muss noch hinzugefügt werden, dass eine derartige Normalisierung vermutlich schlecht mit Standardklassifikatoren harmoniert, da dabei sämtliche, womöglich identifizierende, geometrischen Informationen verloren gehen. Entweder muss der Klassifikator daraufhin angepasst werden, oder es werden diese Paare nur als Training für die Halluzination benutzt und es finden im echten Anwendungsfall normale, rektifizierte Bilder Verwendung.

3.3.3 Halluzination

Nachdem verschiedene Verfahren zur Datenaufbereitung beschrieben wurden, widmet sich dieser Abschnitt der Halluzination. In Abschnitt 2.4 wurden die folgenden Architekturen bereits vorgestellt. Die extrahierten Trainingsdatenpaare werden hier benutzt, um passende Werte für alle Parameter des jeweiligen Netzes zu errechnen. Es wird angenommen, dass die trainierten Modelle in der Lage sind, ein NIR-Bild in das sichtbare Spektrum zu transformieren. Eine Evaluation hinsichtlich der visuellen Genauigkeit der Rekonstruktion findet nicht statt. Das Ziel ist die Synthese von identifizierenden Merkmalen in der jeweils anderen Modalität. Da diese aber nicht zwingend auf sichtbaren Merkmalen beruhen, sondern durchaus komplexe Zusammenhänge modellieren, ist ein derartiges Abstandsmaß nicht geeignet, um eine Aussage bezüglich der Eignung für heterogene Gesichtserkennung zu treffen.

3.3.3.1 Encoder-Decoder Netze

Dieser Abschnitt beschäftigt sich mit dem Ansatz von (Lezama et al., 2016) und zeigt auf, dass es keine triviale Aufgabe darstellt, das Netz erfolgreich zu trainieren, selbst wenn Architektur und Rahmenbedingungen des Netzes bekannt sind. Unglücklicherweise sind die Angaben von (Lezama et al., 2016) lückenhaft, sodass wichtige Metaparameter, wie z. B. die Lernrate oder die zugrunde liegende Bilddatenbank, geschätzt werden müssen. Eine Rekonstruktion der Ergebnisse ist somit zeitaufwendig und nicht abschließend bewertbar. Dennoch wird versucht, passende Metaparameter zu finden, was angesichts der Tatsache, dass das Trainieren auf einer TITAN-X Pascal-GPU volle drei Tage in Anspruch nimmt, kein schnelles Unterfangen darstellt. Das Caffe-Framework bietet leider auch keine automatische Anpassung der Lernrate, wie es bei Torch der Fall ist. Auch die Normalisierung der Bilder ist im Paper lediglich umschrieben. Die Aussage, dass alle Bildhistogramme auf ein geeignetes Bild angepasst werden, ist nicht ausreichend, um die Ergebnisse zu reproduzieren. Die Fülle an Metaparametern und der enorme Zeitaufwand machen es fast unmöglich.
In Abbildung 3.15 sind die Netzarchitekturen zu sehen. Da die mittleren Schichten in ihrer Architektur identisch sind, wurden sie aus Platzgründen entfernt. Wie im Paper beschrieben, wird für die Luminanz ein größeres Netz als für die Farbkanäle verwendet. Es sollte an dieser Stelle daraufhin hingewiesen werden, dass es sich um eine Reproduktion von den in (Lezama et al., 2016) gefundenen Netzen handelt. Es ist jedoch durchaus möglich, dass das Netz von (Lezama et al., 2016) von diesem

divergiert, da die Angaben unvollständig waren und teilweise nicht der gängigen Fachsprache entsprachen.

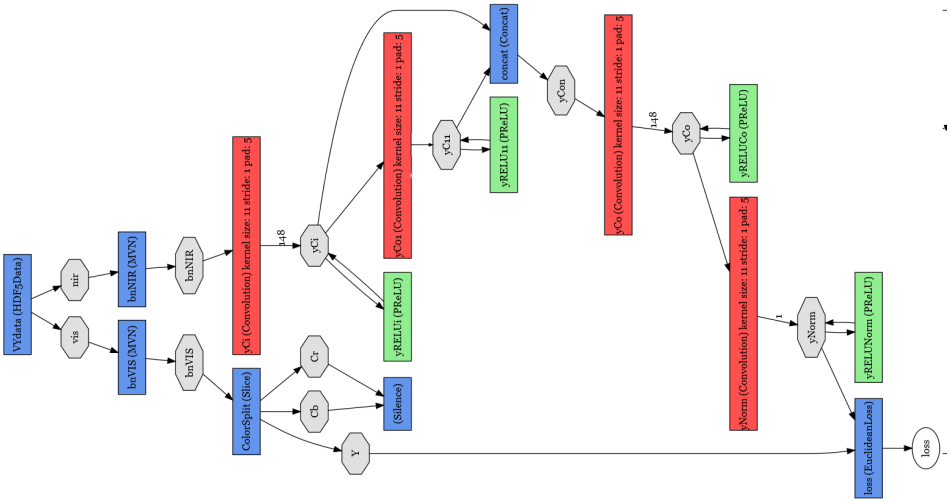

Abb. 3.15 Architektur von (Lezama et al., 2016)

Zum Trainieren wurden jeweils 50 Patches in eine *Batch* zusammengefasst, sodass Ausreißer durch das Mitteln über alle Ergebnisse in einer Batch keinen zu großen Einfluss auf den durch den Adam-Solver errechneten Gewichtsgradienten nehmen können. Diese Auswahl erfolgte zufällig, da eine monotone Reihenfolge negativen Einfluss auf das Lernen des Netzes hat. Die Patches werden aus einer *HDF5*(Hierachical Data Format 5), welche auch von Matlab benutzt wird, geladen. Dies ist notwendig, um zu garantieren, dass die Patches paarweise (NIR-VIS) geladen werden. Würden zwei getrennte Datenbanken für NIR und für VIS verwendet werden, dann ließe sich bei einer zufälligen Auswahl der Daten nicht garantieren, dass die NIR-VIS-Zuordnung intakt bleibt. Je nachdem, welcher Farbkanal des YCbCr-Farbraumes trainiert wird, wird durch ein Splitlayer der jeweilige Kanal herausgeschnitten und weitergeleitet. Die übrigen Daten der VIS-Patches werden dann in ein Silencelayer geführt, da alle Daten, die nicht Input eines anderen Layers sind, auf die Konsole ausgegeben werden, was bei einem Bild äußerst störend ist.
Zuerst wurden die Bilder weder normalisiert noch durchschnittsbefreit. Das Minimieren des Fehlers führte nicht zu einer Verbesserung, ein Lernen konnte nicht festgestellt werden. Nach drastischer Erhöhung der Lernrate konnte ein sprunghafter signifikanter Abschwung festgestellt werden. Ein Prüfen mit Testdaten ergab ein konstantes Bild, völlig unabhängig gegenüber dem Eingangsbild. Eine statistische Analyse des Trainingsdatensatzes ergab, dass die Helligkeit des konstanten Ausgangsbildes (0.38) exakt dem Mittelwert aller Bilder entspricht. Im Hinblick darauf,

dass es sich hier im weiteren Sinne um eine Regression handelt, ist die Mittelwertbefreiung tatsächlich die beste Art, die gesehenen Daten zu approximieren. Ein weiteres Lernen wird durch die hohe Lernrate, die notwendig ist, um den Mittelwert zu finden, verhindert. Das lokale Minimum kann danach nicht mehr verlassen werden. Dies ist ein sehr treffendes Beispiel, dass die Normalisierung der Datenbank existenziell ist und keinesfalls aus einer wissenschaftlichen Ausarbeitung herausgekürzt werden sollte.

Die Datenbank wurde nach dieser Erkenntnis vom Mittelwert befreit. Ein erneutes Training brachte dann die in Abbildung 3.16 erkennbaren Ergebnisse hervor. Es lässt sich feststellen, dass das Ergebnis mit zunehmender Anzahl der Iterationen immer schärfer wird. Das Endergebnis ist jedoch immer noch deutlich unschärfer als das Eingangsbild. Dies ist ebenfalls auf die Regression zurückzuführen, denn da Kanten nicht übereinanderliegen, werden diese aufsummiert, sodass bei der Anwendung des Netzes ein Verlauf anstatt einer klaren Kante zu sehen ist.

Weitere Versuche der Normalisierung, zum Beispiel auf eine einheitliche Varianz, brachten keine Verbesserung in den Ergebnissen. Auch eine Batchnormalisierung, das heißt eine Normalisierung, die den Mittelwert der aktuellen Batch berechnet und diesen dann subtrahiert, konnte das Ergebnis nicht weiter verbessern. Zusätzlich sollte erwähnt werden, dass das Netz auf den Patches trainiert, jedoch auf den kompletten Bildern angewandt wird. Es ist also nicht nötig, ein Eingangsbild in Patches der Trainingsgröße zu zerschneiden. Sie werden im Ganzen prozessiert. Das schwächt auch das Schärfeproblem ab, da die Bilder vergrößert in das Netz geleitet und danach wieder verkleinert werden können.

Nach einer visuellen Bewertung von Abbildung 3.16 kann festgestellt werden, dass diese nach menschlichem Empfinden kein reales Bild abgeben und im Allgemeinen nicht sehr ansprechend sind. Da das Ziel die heterogene Gesichtserkennung ist, ist es aber ein ausschlaggebendes Kriterium, dass eine Verbesserung der Erkennung entsteht. Dies wird in Abschnitt 4.2 untersucht.

Zusätzlich lässt sich in Abbildung 3.16 zeigen, dass Verzerrungen von Gesichtsformen durch die Frontalisation zu ungewöhnlichen weißen Artefakten führen. Dies ist damit zu erklären, dass diese Art der Verzerrung nicht im Trainingsdatensatz zu finden ist, da dort die Gesichter paarweise registriert wurden. Das Beispiel zeigt sehr gut, dass diese Art von CNN komplexe Formen erlernt. Unbekannte Formen, die eine hohe Abweichung zu allem vorab Trainierten aufweisen, funktionieren deshalb nicht, was auch die Grenzen aller Deep-Learning-Ansätze aufzeigt.

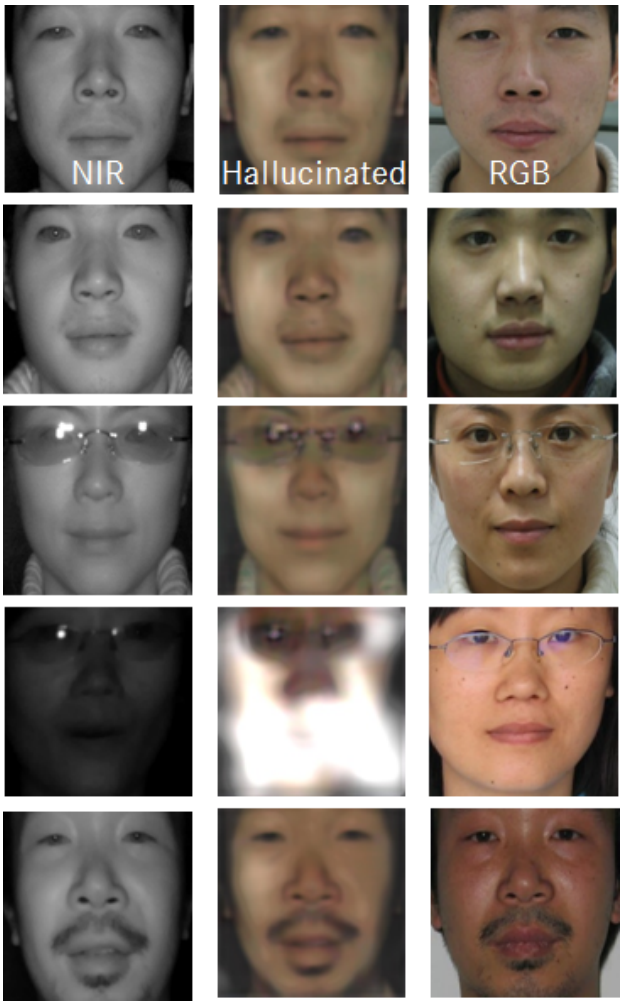

Abb. 3.16 Ergebnis des Encoder-Decoder Netzes mit gekachelten Daten

3.3.3.2 Adversarial Netze

Zuvor wurde das Encoder-Decoder-Netz von (Lezama et al., 2016) verwendet. Dieser Abschnitt beschäftigt sich mit dem Adversarial-Netz von (Isola et al., 2016). Wie bereits in Unterunterabschnitt 2.4.2.3 beschrieben, handelt es sich hierbei nicht um eine Regression im üblichen Sinne. Vielmehr wird das Problem als Nullsummenspiel interpretiert, bei dem zwei Netze gleichzeitig trainiert werden. Das eine versucht ein generiertes, eingefärbtes NIR-Bild von einem echten VIS-Bild zu unterscheiden, und das andere versucht ein NIR-Bild derart einzufärben, dass es nicht von einem echten unterscheidbar ist. Wie ausführlich in Unterunterabschnitt 2.4.2.3 dargestellt, wird dieses Verfahren in Anlehnung an den Turing-Test als Turing-Lernen bezeichnet und kann als Primal-Dual-Optimierung von CNNs gesehen werden.
Für das Einfärben wird Pix2Pix von (Isola et al., 2016) verwendet. Das trainierte Netz

beinhaltet 64 Diskriminatoren und 64 Generatoren. Nach kurzer Versuchsphase haben sich diese Parameter als sinnvoll herausgestellt. Es ist durchaus eine weitergehende Optimierung und Abwägung sinnvoll, aus Zeitgründen jedoch nicht im Rahmen dieser Arbeit machbar. Eine kurzer Vergleich, wie sich die Ergebnisse prinzipiell bei einer Veränderung der Generatoren und Diskriminatoren verhalten, kann Abbildung 3.17 entnommen werden.

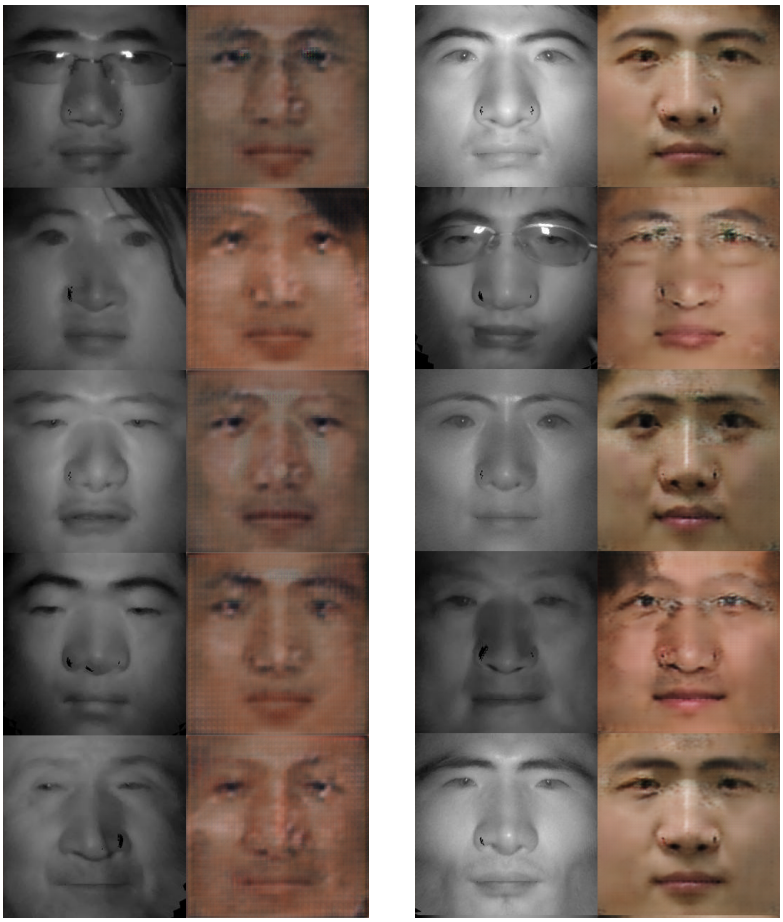

Abb. 3.17 Links 32 Generativeinheiten
Rechts 128 Generativeinheiten

Pix2Pix baut auf dem Torch Framework auf und ist in Lua-Script implementiert. Die Verbreitung dieser recht neuen Sprache ist gering und die Installation deutlich unhandlicher als bei Caffe. Dennoch ist die Implementierung des Netzes deutlich besser, sodass viele Dinge, wie zum Beispiel das Generieren von Zwischenergebnissen, automatisch geschehen, während sie in Caffe aufwendig selbst implementiert werden müssen. Die Organisation größerer Studien fällt hier deutlich leichter und auch das Ausnutzen sämtlicher Hardwareressourcen scheint hier effizienter zu sein. Weiterhin ist es nicht erforderlich, eine Datenbank zu erstellen, denn gewünschte Bild-

paare können einfach Seite an Seite abgespeichert und vom Framework geladen und zerteilt werden. Auch die Lernrate wird hier automatisch über den Gradienten bestimmt und muss nicht manuell optimiert werden, was – wie in Unterunterabschnitt 2.4.2.3 beschrieben – auch gar nicht möglich ist, da es sich hier um ein Nullsummenspiel handelt, bei welchem beide Spieler stetig besser werden.

3.3.3.2.1 Patch basiert Zuerst werden die in Unterunterabschnitt 3.3.2.2 generierten Patches verwendet, um das oben beschriebene Netz zu trainieren. Durch die sehr kleinen Bildausschnitte ist der Vorteil dieses generativen Ansatzes begrenzt. Da der Abgleich nicht mit vollständigen Bildern, sondern lediglich mit Patches erfolgt, kann kein visuell ansprechendes Bild erzeugt werden. Wie in Unterunterabschnitt 3.3.2.2 bereits beschrieben, wird das Netz mit den Patches trainiert, im tatsächlichen Anwendungsfall werden jedoch vollständige Bilder verwendet. Das Ergebnis im Verlauf des Trainings kann in Abbildung 3.18 gefunden werden.

Abb. 3.18 Ergebnis des Adversarial-Patch-Ansatzes

Es lässt sich erkennen, dass die Resultate sehr schwach bleiben und allgemein eher einem unscharfen Schatten als einem tatsächlichen VIS-Bild gleichen.

3.3.3.2.2 Patch-Kollage Im vorigen Absatz wurden die Schwächen des Patch basierten Ansatzes aufgezeigt, welche nun verbessert werden sollen. Dazu wird die Patchdatenbank zu Kollagen zusammengefasst, wie in Abbildung 3.19 zu sehen ist. Sinn dahinter ist, dass die recht geringe Patchgröße besser auf die netzinternen Prozesse abgebildet wird. Auch dieser Ansatz wird für 4 Epochen trainiert. Die Ergebnisse sind in Abbildung 3.19 zu finden.
Die Bilder sind deutlich markanter als der direkte Patch-Ansatz, leider werden die Patch-Grenzen auch in der Anwendung des Netzes reproduziert. Zweifellos ist das durch die Architektur des Adversarial Netzes bedingt, welches versucht ein möglichst *echtes* Bild zu reproduzieren. Da in diesem Fall die echten Bilder alle solche

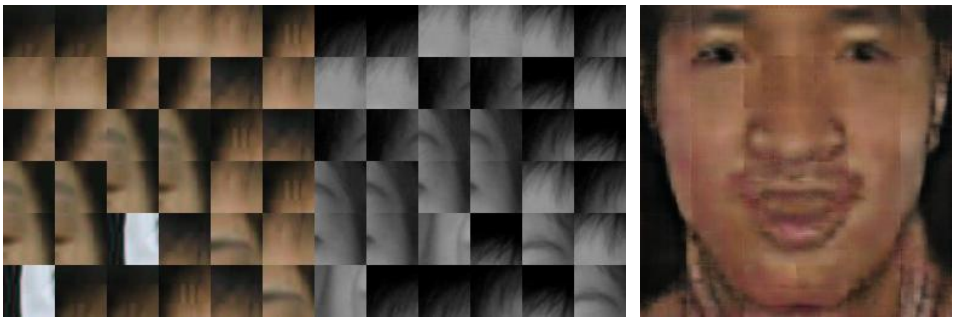

Abb. 3.19 Patchkollage als Input und Ergebnis nach Adversarialprozess

Kanten, sogar an exakt der gleichen Stelle, aufweisen wird das Produkt der Halluzination auch derartige Kachelfugen aufweisen.
Abschließend lässt sich zusammenfassen, dass aufgrund der speziellen Art der Adversarial Netze, welche versuchen ein täuschend echtes Ergebnis (im Vergleich zur Trainingsdatenbank) zu erzeugen, der Patch basierte Ansatz grundlegend ungeeignet erscheint. Gesamtheitliche Erscheinungsinformationen, welche für dieses Verfahren existenziell sind, gehen durch ihn verloren und führen zu unschönen Ergebnissen.

3.3.3.2.3 Geometrisch rektifiziert Der geometrisch rektifizierte Ansatz ist im Vergleich zu den Patch basierten Ansätzen viel mehr darauf bedacht Bildinformationen und deren Zusammenhang zu erhalten. Wie bereits bei der Datenaufbereitung besprochen wird hier lediglich eine affine Transformation für das gesamte Bild angewandt um weitergehende Verzerrungen zu verhindern. Dafür ist die Registrierungsgüte die schlechteste unter allen hier besprochenen Ansätzen.

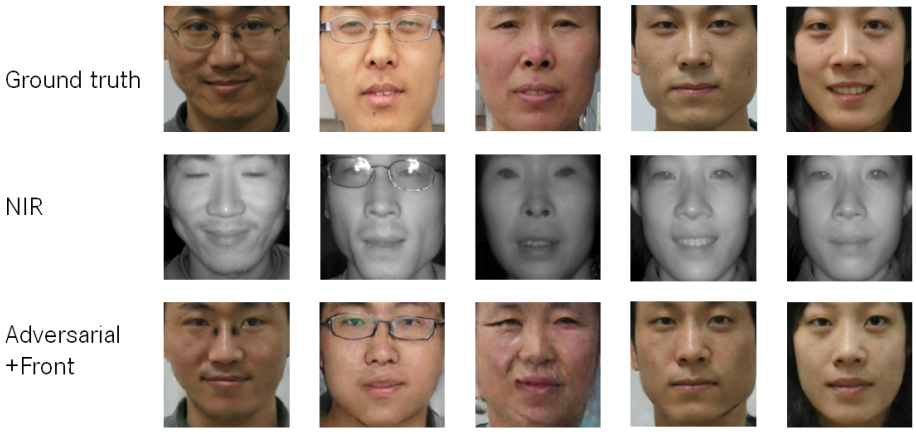

Abb. 3.20 Ergebnis des rektifizierten Datensatzes nach Adversarialprozess

Das Ergebnis nach vier Epochen kann in Abbildung 3.20 gefunden werden. Es ist

zu erkennen, dass Textur und Hautton sehr gut reproduziert wurden, aber auch, dass die mangelnde Übereinstimmung zu Diskontinuitäten der Form geführt hat. Ein Schärfeproblem wie bei den Patch basierten Versuchen ist nicht zu erkennen. Kanten sind markant und die allgemeine Textur ist detailliert und kontrastreich. Brillen stellen ein kleineres Problem dar, da diese nicht akkurat reproduziert werden. Vermutlich ist das dem geringen Anteil an Brillenträgern in der Datenbank zu zuschreiben.
Abschließend lässt sich zusammenfassen, dass die geometrische Rektifizierung die Vorteile des Adversarial Netzes vollständig ausnutzt, es aber durch die geringe Registrierungsgüte, zu visuell wenig ansprechenden, Artefakten in der Rekonstruktion kommt und die Gesichtsform nachhaltig verändert wird.

3.3.3.2.4 Texturunwrap Der Texturunwrap vereint die lokale Registrierung mit einem ganzheitlichen Ansatz und sollte somit in der Lage sein die Vorteile eines Adversarial Netzes auszunutzen und geometrische Artefakte zu verhindern. Während der Texturzusammenhang gewahrt wird, werden alle Forminformationen auf ein Standardgesicht abgebildet und stehen somit nicht mehr zur Identifikation von menschlichen Gesichtern zur Verfügung. Das Ergebnis, auch hier wieder nach vier Epochen Training, kann in Abbildung 3.21 gefunden werden.
Es ist ersichtlich, dass die Textur detailliert und kontrastreich reproduziert wurde. Geometrische Artefakte entstehen keine. Das Problem mit Brillen wurde etwas abgeschwächt ist jedoch immer noch präsent. Bei dieser Problematik kann folglich nur ein größerer Datensatz Abhilfe schaffen.

3.3.4 Klassifikation

In den vorigen Abschnitten wurden verschiedene Trainingsdatensätze aus der CASIA-NIR-VIS-2.0-Datenbank extrahiert, die danach zum Trainieren verschiedener Netze verwendet wurden. Dieser Abschnitt zieht nun den Kreis zur heterogenen Gesichtserkennung. Es werden zwei verschiedene Opensource-Klassifikatoren verwendet, die nur mit VIS-Bildern initialisiert wurden. Genauer gesagt werden die bereits trainierten homogenen Modelle in keinster Weise an den hiesigen Anwendungsfall angepasst. Das Resultat dieser Klassifikation ist ein Punkt innerhalb des Definitionsbereichs des jeweiligen Netzes. Der Abstand zwischen zwei Punkten ist nicht notwendigerweise proportional zur Unähnlichkeit der klassifizierten Probanden. Wie bereits in Abschnitt 2.4 dargelegt wurde, trifft diese Eigenschaft nur auf Openface

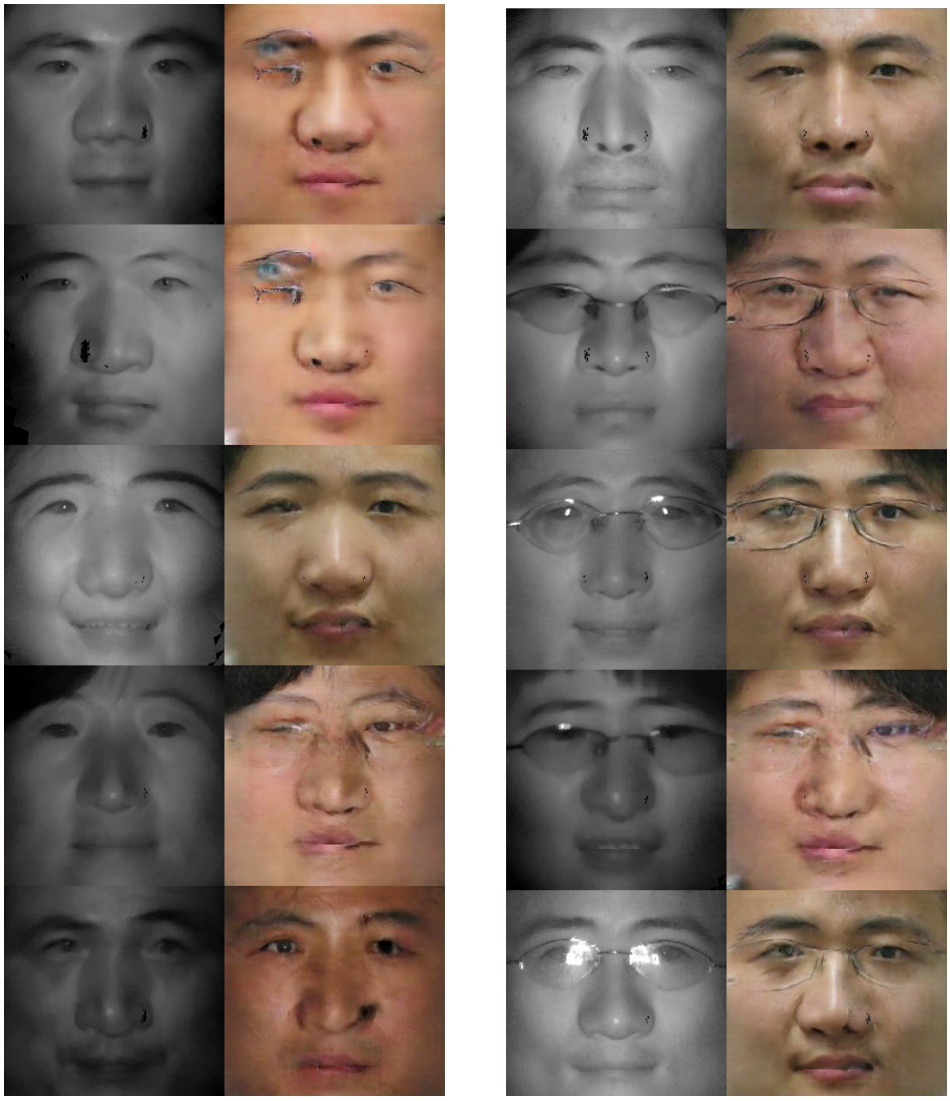

Abb. 3.21 Ergebnis des Texturdatensatzes nach Adversarialprozess

zu. VGG-Face ist ein Klassifikator, der eine Identität in einen Cluster fester Größe zu projizieren versucht. Wichtig ist folglich nur, ob der Abstand der zwei Punkte geringer ist als der Radius der Cluster-Hyper-Sphere.
Bevor die effiziente Berechnung der Abstandsmatrix untersucht wird, wird zunächst technisch dargelegt, wie die einzelnen Verfahren verwendet werden können. Eine abschließende Evaluation und Interpretation der hier ermittelten Ergebnisse ist in Abschnitt 4.2 zu finden.

3.3.4.0.1 VGG-Face VGG-Face (Parkhi et al., 2015) beruht auf dem bereits verwendeten Caffe-Framework (Jia et al., 2014) und war Sieger der LFW-Challenge. Er ist extrem robust gegenüber Verdeckungen und Verzerrungen. Durch die wiederholten

Fully-Connected-Layer kann angenommen werden, dass die von ihm ausgewählten, identifizierenden Merkmale, rein auf Textur basieren und nur sehr wenig bis keine geometrischen Informationen verwendet werden.
Es wird das von den Autoren trainierte Modell zur Identifikation genutzt, weshalb keine Notwendigkeit der Kreuzvalidierung besteht. Die Eingangsauflösung des Netzes beträgt 224×224 Pixel. Alle Bilder, dass heißt jedes VIS- und NIR-Bild, als auch jede Form der Halluzination wird durch den Klassifikator in den Identifikationsraum projiziert. Das Ergebnis ist also für jedes Bild ein 4096 Elemente großer Vektor. Dieser wird zusammen mit dem Label (Identifikationsnummer) in eine Matlabdatei abgespeichert.
Die L_2-Distanz zwischen zwei Identifikationsvektoren dient als Abstandsmaß zur Unterscheidung, ist aber in im Falle von VGG-Face kein Ähnlichkeitsmaß zweier Gesichter, sondern ein reines Unterscheidungskriterium. Eine Abstandsmatrix wird für jede Variante berechnet, sodass für jeden Versuch eine obere Dreiecksmatrix mit dem Abstands zwischen i und j entsteht.

3.3.4.0.2 Openface Openface (Amos et al., 2016) ist ein Embedder, das heißt der Abstand zwischen zwei Identifikationsvektoren kann als Maß der Unähnlichkeit zweier Gesichter gesehen werden. Der Identifikationsraum umfasst 128 Dimensionen. Durch die sich wiederholenden Faltungen ist das Verfahren nur geringfügig robust gegen Verzerrungen und Verdeckungen. Weiterhin spielt die Gesichtsform eine große Rolle in der Identifikation. Textur ist ein weniger ausschlaggebender Faktor. Die Eingangsauflösung des Netzes beträgt 96×96 Pixel. Es muss genausten darauf geachtet werden, dass alle Gesichter korrekt frontalisiert wurden. Weicht die Augenposition eines Probanden nur um wenige Pixel von der vom Netz erwarteten ab, kann es zu einem signifikanten, zweistelligen Präzisionsverlust kommen.
Auch hier dient die L_2-Distanz als Abstandsmaß mit welcher eine obere Dreiecksmatrix erzeugt werden kann. Je größer der Eintrag, desto unähnlicher sind sich i und j.

3.4 Projection

Nachdem die Feature- und Synthesis-Methoden beleuchtet wurden, widmet sich dieser Abschnitt der Projection. Wie bereits mehrfach erwähnt kann ein Projektionsverfahren sowohl eine aufwendige, für den heterogenen Anwendungsfall entwickelte, Methode, als auch die simple Anwendung eines homogenen Klassifikators sein.

Nachfolgend werden drei Beispiele behandelt, welche zu den Projektsmethoden gezählt werden können. Alle basieren auf einem neuronalen Netz, wobei eins, das siamesische Netz, eine Besonderheit darstellt, die nachfolgend näher beschrieben wird. Die verbleibenden Ansätze sind die bereits aus den vorigen Abschnitten bekannten homogenen Klassifikatoren, welche nun mit einem heterogenen Datensatz trainiert werden. Hierbei könnte man durchaus sagen, dass es sich um den klassischen Deep Learning Ansatz handelt, bei welchem man das Netz unüberwacht arbeiten lässt.

3.4.0.0.1 Siamesische Netze Bei siamesischen Netzen handelt es sich um zwei Subnetze mit der gleichen Architektur. Ein Unterschied zwischen der Funktionsweise beider Subsysteme kann deshalb nur auf den gelernten Gewichten beruhen. Netze welche gemeinsame Gewichte nutzen (*Parameter Sharing*) werden als *pseudo siamesisch* bezeichnet und meist zur Unterscheidung zweier Bilder verwendet. Pseudo siamesische Netze sind für die heterogene Gesichtserkennung weniger nützlich, weshalb man sich im Folgenden auf die Siamesischen Netze beschränkt.
Die Idee ist es zwei gleiche Netze zu verwenden von denen eins jeweils eine Modalität in der Trainings- sowie der Anwendungsphase verarbeitet. Beide Systeme sollen eine Identität auf den gleichen Punkt im gemeinsamen Identifikationsraum abbilden um eine heterogene Erkennung zu ermöglichen.
Dadurch, dass beide Netze ihre eigenen Parameter erlernen können ergeben sich mehr Freiheitsgrade mit welchen eine modalitätsspezfische Abbildung auf den Identifikationsraum ermittelt werden kann. Es handelt sich hierbei folglich um einen automatisierten Prozess der die bereits vorhandenen Möglichkeiten zur homogenen Gesichtserkennung verbindet. Mehr Parameter haben jedoch nicht nur Vorteile: Der Datensatz der zum erfolgreichen Training benötigt wird muss deutlich größer sein als der für ein einzelnes Netz. Ist er zu klein kann es zu Überanpassung kommen und das Generalisierungsverhalten einschränken. Gerade im Fall der heterogenen Gesichtserkennung, für welche die verfügbare Datenbasis oft bescheiden ist, ist dieser Ansatz weniger geeignet.
Versuche ein siamesisches Netz auf Basis des VGG-Klassifikators aufzubauen sind zum einen an der zu geringen Datenbasis zum anderen an der Resourcenintensität gescheitert. Eine effizientes Training wurde durch den zu hohen Speicherbedarf (ca 18 GB) behindert und konnte nicht erfolgreich abgeschlossen werden.

3.4.0.0.2 Openface Openface (Amos et al., 2016) ist eines der bekanntesten Gesichtserkennungssysteme. Er basiert auf einem Triplet-trainierten Embedder Ansatz. Einzelheiten zur Funktionsweise können den entsprechenden, vorangestellten Ka-

piteln entnommen werden. Für das Training wird auch hier die von (Li et al., 2013) vorgeschlagene Vierteilung verwendet.
Der Datensatz von (Li et al., 2013) wird dazu in die korrekte Größe von 96×96px gebracht. Zusätzlich werden die äußeren Augen- sowie Mundwinkel auf eine einheitliche Position transformiert (affin). Gelernt wird für 10 Epochen, was in etwa einen Tag in Anspruch nimmt pro Partition in Anspruch nimmt. Das Lernen einer Konfiguration von manuellen Parametern ist folglich mit einem zeitlichen Aufwand von 4 Tagen verbunden.
Für die manuell festgelegten Parameter werden die Standardwerte der Autoren (Amos et al., 2016) verwendet, da bei 4 Tagen Trainingszeit ein Feintuning der Konfiguration zu aufwendig für eine Machbarkeitsstudie ist. Für den hier angestellten kategorischen Vergleich ist die argumentative Aussagekraft trotzdem hinreichend.

4 Evaluation

> *Grau, teurer Freund, ist alle Theorie.*
>
> **– Johann Wolfgang von Goethe**

Nachdem verschiedene Ansätze beschrieben, implementiert und angewandt wurden, widmet sich dieses Kapitel der Analyse der Ergebnisse, für welche einerseits öffentlich verfügbare Datenbanken und andererseits Probandenstudien verwendet werden. Insbesondere wird auf die Vergleichbar- sowie auf die Belastbarkeit der ermittelten Ergebnisse eingegangen um eine Basis zur abschließenden Eignungsbewertung zu erhalten.

4.1 Feature

Die Merkmal basierten Verfahren zeichnen sich durch ihren analytischen Ansatz aus. Eine Eigenschaft welche in allen Modalitäten messbar ist, wird zur Identifikation herangezogen. Da die Evaluation für das 3DMM auf der CASIA NIR-VIS Li et al., 2013 beruht und das des Gesichtstrackings auf einer eigenen Probandenstudie, wird im folgenden klar zwischen beiden Systemen unterschieden. Die so erlangten Ergebnisse sind nicht uneingeschränkt vergleichbar.

4.1.1 Tracking von Gesichtsmerkmalen

Die Funktionsweise des Trackings von Gesichtsmerkmalen ist in Abschnitt 3.2 beschrieben. Da es sich um eine Kombination von Hard- und Software handelt ist es nicht möglich das Verfahren anhand von anerkannten, öffentlichen Datenbanken zu evaluieren. Das angewandte Evaluationsverfahren und die Zusammensetzung der Probanden wird im Folgenden beschrieben, bevor die Ergebnisse analysiert werden.

4.1.1.1 Evaluationsprotokoll

Um eine maximale Vergleichbarkeit zu erreichen und etwaigen Aufnahmevarianzen vorzubeugen, werden beide Kamerasysteme übereinander angeordnet und gleichzeitig verwendet. Das spart nicht nur Zeit bei der Durchführung, sondern garantiert auch, dass etwaige Unterschiede in Mimik, Kopfrotation und Position von beiden Systemen äquivalent wahrgenommen werden. In Abbildung 4.1 (links) ist der verwendete Kameraaufbau zu sehen.

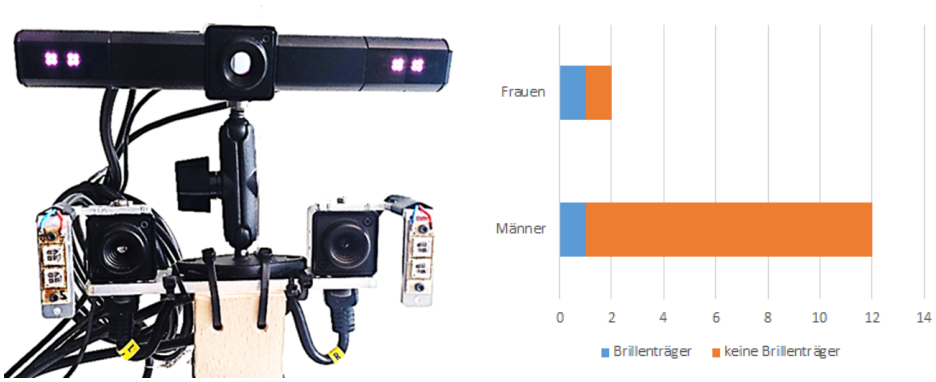

Abb. 4.1 Links Fovio-Stereo-Mono-Testaufbau
Rechts Probandenzensus der durchgeführten Studie

In Abbildung 4.1(rechts) ist die Zusammensetzung der Probanden visualisiert. 14 Personen haben an der Studie teilgenommen, darunter zwei Frauen und zwei Brillenträger. Jeder Proband wurde zwei mal aufgenommen, was zu 28 Datensätzen führt.
Die Aufnahmedauer beträgt etwa 30 Sekunden, welche jedoch nicht vollständig zur weiteren Identifikation herangezogen wird. Nach den zuvor definierten Anforderungen (siehe Abschnitt 3.1) kann lediglich davon ausgegangen werden, dass die Kameras einen Nutzer für 3 Sekunden mit eine Kopfrotation von weniger als 15 Grad und ohne Verdeckungen sehen können. Deshalb wird jeder Datensatz auf die ersten 3 Sekunden zusammengeschnitten in welchem die Kopfrotation geringer als 15 Grad gemessen an der optischen Achse beträgt. Diese drei Sekunden werden dann gemittelt und anhand der zuvor erklärten Abstandsmaße verglichen. Betrachtet wird nur der heterogene Fall, dass heißt Mono-Stereo (2D-3D). Da die hier verwendeten Abstandsmaße symmetrisch sind (siehe Abschnitt 3.2) ist es irrelevant ob die Richtung 2D-3D oder 3D-2D verglichen wird.
Da es sich um einen analytischen Ansatz handelt ist es nicht notwendig die Datenbasis zu teilen um eine Kreuzvalidierung zu ermöglichen. Weiterhin ist eine Datenbasis von 14 Probanden, welche auch nicht ethnisch repräsentativ sind, nicht geeignet

um eine vergleichende Aussage zu veröffentlichten, wissenschaftlichen Arbeiten zu geben. Sehr wohl jedoch ist es möglich, die verschiedenen Abstandsmaße qualitativ zu untersuchen, was auch das Ziel der folgenden Auswertung ist.

4.1.1.2 Auswertung

Nachfolgend kann ein ROC-Diagramm gefunden werden, welches sich hervorragend eignet die Eigenschaften eines Klassifikators darzustellen. Allerdings bedarf es einer Erklärung um ein solches lesen zu können.

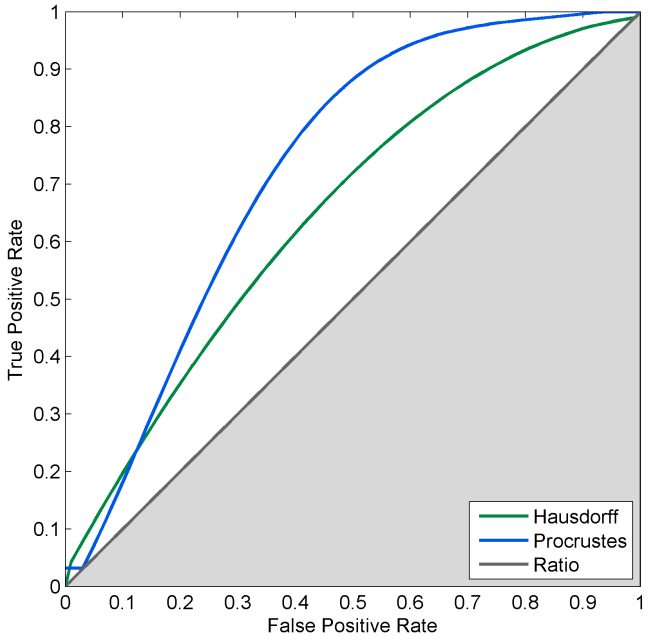

Abb. 4.2 ROC von Procrustes und Hausdorff Distanz im Vergleich

Auf der X-Achse wird die *False Positive Rate* abgebildet. Es handelt sich folglich um den Anteil an Bildpaaren deren Identität das System als gleich (positiv) beurteilt die es aber nicht sind. Auf der Y-Achse ist die *True Positive Rate* abgebildet. Sie beschreibt den Anteil von Bildpaaren deren Identität gleich ist und welche auch das System als gleich beurteilt. Es liegt auf der Hand, dass die beiden Quotienten gegensätzlich sind. Nimmt man an, dass das System immer negativ als Ergebnis ausgibt, so würde die True Positive Rate Null betragen und die False Positive Rate ebenfalls Null. Im Diagramm wäre dieser Punkt unten links, am Schnittpunkt von X- und Y-Achse zu finden, was auch die Bedeutung der grauen Linie erklärt. Die graue Linie

bezeichnet das Ergebnis, welches ein auf Zufall basierendes Verfahren erzielen würde. Unterhalb der grauen Linie sollte es keine Kurven geben, da ein System, welches signifikant schlechter als der Zufall sein kann, in gleicher Weise über der Zufallslinie liegen kann.
Sieht man sich die Kurven von Hausdorff und Procrustes genauer an, stellt man fest, dass keine allgemeingültige Aussage über die Überlegenheit eines der beiden Verfahren gestellt werden kann. Bei niedrigen False Positive Raten, im Bereich welcher insebsondere für Authentifizierungsanwendungen von Bedeutung ist, ist die Hausdorff-Distanz leicht besser. Es muss allerings klar sein, dass beide Kurven in diesem Bereich nicht signifikant über der Zufallskurve liegen und das es sich hierbei auch um einen Effekt handeln könnte, welcher bei einer größeren Probandenstudie nicht auftreten würde. Diese Hypothese wird durch den Sprung der Procrustes-Kurve am Anfang der X-Achse untermauert. Eine derartige Diskontinuität lässt sich lediglich durch einen starken Quantisierungseffekt erklären.
Im der zweiten Hälfte der X-Achse, dem Identifkationsbereich, schneidet Procrustes deutlich besser ab als Hausdorff. Es können folglich zwei Lehren aus dem Versuch gezogen werden: Erstens, sind beide geometrischen Verfahren nicht sehr vielversprechend, aber in Relation zu ihrem Implementierungsaufwand recht effektiv. Zweitens kann festgehalten werden, dass Procrustes signifikant besser funktioniert als Hausdorff und die anfängliche Unterlegenheit im Authentifizierungsbereich vermutlich auf einen Quantisierungseffekt zurückzuführen ist. Procrustes ist vermutlich durch seine gesamtheitlichen Datenanpassungsstrategie der reinen Hausdorff Distanzmessung überlegen.

4.1.2 3D Morphable Model

Das 3DMM wurde verwendet um eine heterogene Identifikation zwischen dem NIR- und VIS-Spektrum zu erhalten. Ein biometrisches Modell wird dabei auf Gesichtslandmarken angepasst. Im Gegensatz zu den geometrischen Abstandsmaßen wie Hausdorff und Procrustes, hat die Korrelation der Eigenwerte des 3DMM eine statistische Aussagekraft in Bezug auf die Datenbasis auf welches es trainiert wurde. Näheres kann den entsprechenden Kapiteln Unterunterabschnitt 2.3.2.1 und Abschnitt 3.2.2 entnommen werden.

4.1.2.1 Evaluationsprotokoll

Die biometrischen Informationen des 3DMMs sind nicht aus der CASIA NIR-VIS Datenbank weshalb auch in diesem Fall keine Kreuzvalidierung erfolgen muss. Betrachtet wird ebenfalls lediglich der heterogene Fall. Da der Abstand der Eigenwerte anhand der Kreuzkorrelation gemessen wird und diese symmetrisch ist, ist es nicht notwendig zwischen NIR-VIS und VIS-NIR zu unterscheiden.
Die CASIA NIR-VIS Datenbank beinhaltet 725 Personen und ist somit die größte öffentlich verfügbare Datenbank für die heterogene NIR-VIS Gesichtserkennung.(Li et al., 2013) Die Problematik der Datenbasis ist vielfältig und bei weitem nicht auf die geringe Größe (verglichen mit homogenen Datenbasen) oder die geringe ethnische Vielfalt, begrenzt.
Das heißt jedoch nicht, dass die mit ihr ermittelten Ergebnisse nicht aussagekräftig sind. Da das Evaluationsprotokoll für alle heterogenen NIR-VIS Verfahren verwendet wird, kann eine qualitative Vergleichbarkeit unterstellt werden, auch wenn die Datenbasis nicht zwangsläufig realen Bedingungen entspricht.
Keinesfalls jedoch darf qualitativ mit homogenen Klassifikatoren verglichen werden. Da Gesichtserkennungssyteme logischerweise dazu tendieren mit der Anzahl der bekannten Identitäten an Genauigkeit zu verlieren (Identifikationsraum wird dichter besiedelt, was zu einem geringeren Interindividuumsabstand führt), ist es nicht aussagekräftig, Ergebnisse, welche auf 725 Personen beruhen mit Evaluationen von mehreren hunderttausend Probanden zu vergleichen.

4.1.2.2 Auswertung

Nachfolgend findet man das bereits zuvor erläuterte ROC-Diagramm. Ein direkter Vergleich mit den zuvor vorgestellten geometrischen Verfahren sollte jedoch trotzdem nicht uneingeschränkt erfolgen. Wie im Evaluationsprotokoll beschrieben handelt es sich in diesem Versuch um eine deutlich größere Datenbank, welche zudem aus Bildern statt aus Bildsequenzen besteht.
Zunächst einmal ist zu bemerken, dass das Diagramm keinerlei Quantisierungseffekte aufweist, was mit der deutlich gestiegenen Anzahl an Testentitäten zu erklären ist. Das 3DMM kann sich deutlich von der Zufallslinie abgrenzen, der erzielte maximale Abstand zu dieser ist jedoch nicht zufriedenstellend. Es kann also festgestellt werden, dass 3D Informationen zwar prinzipiell übertragbar, jedoch keineswegs zuverlässig bestimmbar sind. Es ist nicht möglich stabile 3D Gesichtslandmarken aus

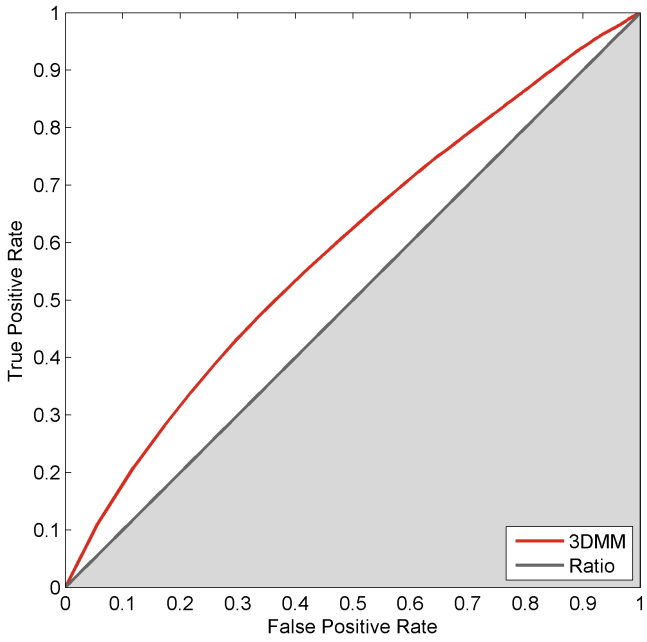

Abb. 4.3 ROC des entwickelten 3DMM Verfahrens

einem einzigen 2D Bild zu extrahieren. Dadurch ist die Übertragbarkeit von 2D-3D und NIR-VIS durch das 3DMM irrelevant, da eine präzise Messung nicht möglich ist.

4.2 Synthesis

Die hier vorgestellten Verfahren zu Synthese von VIS-Bildern aus NIR-Daten sind gelernte Ansätze und müssen im Gegensatz zu den analytischen Methoden, validiert werden. Da der Trainingsdatensatz mangels alternativer Quellen auch als Evaluationsdatenbank benutzt wird, wird er wie von (Li et al., 2013) beschrieben in vier Teile zerteilt. Jeweils drei davon werden zum Training verwandt, der vierte Teil dient der Evaluation. Führt man dies mit allen kombinatorisch möglichen Varianten durch, erhält man die volle Datenbank zu Evaluation ohne das ein trainiertes Modell während des Tests ein Bild verarbeitet, welches es zuvor in der Trainingsphase bereits gesehen hat. Es wird folglich das Generalisierungsverhalten überprüft, da man davon ausgehen darf, dass im realen Einsatz die allermeisten Gesichter nicht in der Trainingsphase enthalten sind.
Die üblichen und zuvor mehrfach ausführlich dargelegten Probleme der geringen

ethnischen Vielfalt der CASIA Datenbank, bleiben selbstverständlich bestehen und können nicht durch eine Kreuzvalidierung entkräftet werden.

4.2.1 Encoder-Decoder Netze

Im untenstehenden Diagramm bilden die Farben jeweils die Art der Methode und die Linienart den evaluierenden Klassifikator ab. Zunächst muss festgestellt werden, dass Openface deutlich besser mit dem heterogenen Anwendungsfall umgeht, als VGG. Das Interessantere ist jedoch der direkte Vergleich der Reihenfolge: Während Openface durch den vorangestellten Encoder-Decoder einen leichten, dennoch konstanten und signifikanten Zuwachs erhält, wirkt sich diese Form von Synthese äußerst negativ auf die Performanz des VGG Klassifikators aus.

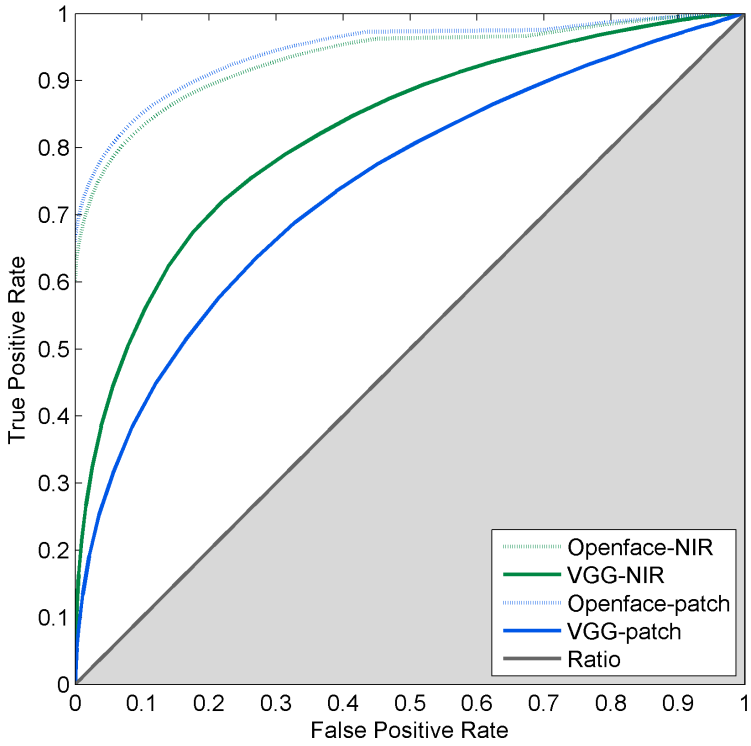

Abb. 4.4 ROC der Encoder-Decoder-Methoden

Abschließend kann festgestellt werden, dass das Encoder-Decoder Modell zu wenig Details für texturbasierte Klassifikatoren abbildet. Lediglich die vollständige Erhaltung der Gesichtsform ermöglicht die Erkennung durch einen Embedder, dessen Genauigkeit durch die Synthese leicht verbessert werden kann.

4.2.2 Adversarial Netze

Im nachfolgendem Diagramm sind die Adversarialmethoden aufgeschlüsselt. Auffällig ist zunächst das keine der Methoden präziser ist als der homogene Embedder Openface (gepunktet grün). Das untere Ende bildet VGG-Face ohne Synthese (in grün durchgezogen). Es untermauert folglich die Hypothese, dass sich Adversarialnetze besonders für die Rekonstruktion von Textur eignen, dabei jedoch Formen und Bildkanten nur sehr ungenau wiederspiegeln. Dieses führt im Umkehrschluss auch zu einem Grund für die hohe Ungenauigkeit von Openface.
Das Adversarialnetz zusammen mit unwrapped Gesichtsbildern führt, mit großem Abstand zu den halluzinierten Bildern, zu dem besten Ergebnis von VGG. Auch der Halluzinationsansatz führt zu einer Verbesserung von VGG und einer signifikanten Verschlechterung von Openface.
Um den Zusammenhang der Synthese und den eingesetzten Klassifikatoren genauer zu untersuchen, wird im nachstehenden Kapitel ein gegenüberstellender Vergleich durchgeführt.

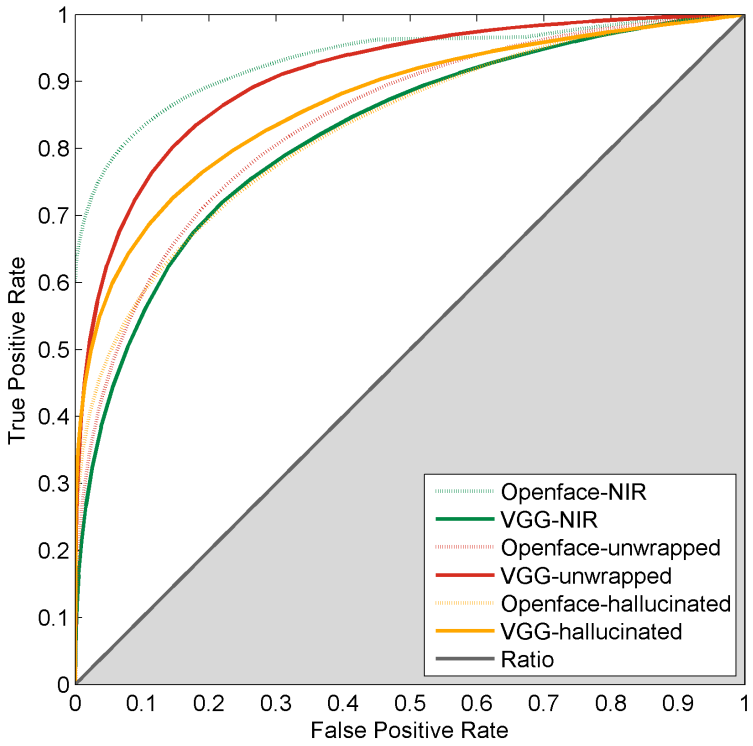

Abb. 4.5 ROC der Adversarial-Methoden

4.2.3 Vergleich nach Klassifikator

Nachdem die einzelnen Syntheseverfahren verglichen wurden lohnt es sich den Vergleich auf die zur Evaluation nachgestellten Klassifikatoren zu konzentrieren. Prüft man die untenstehenden Diagramme gegeneinander, so wird man feststellen, dass die qualitative Ordnung der Kurven zwischen beiden Diagramm invertiert zu sein scheint.

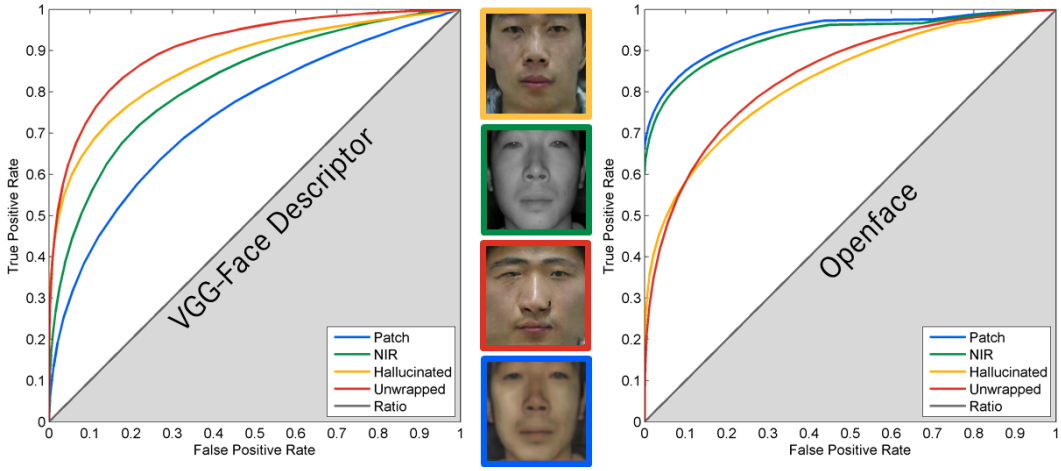

Abb. 4.6 ROC der Synthese Methoden, aufgeschlüsselt nach Klassifikator

Die unwrapped Bilder schneiden mit Openface am schlechtesten und mit VGG-Face am besten ab. Die Reihenfolge folgt für VGG-Face der Qualität der Registrierung. Openface erzielt die besten Ergebnisse mit den am wenigsten verzerrten Bildern. Es kann folglich festgestellt werden, dass Openface, als Embedder, eher die allgemeine Gesichtsform und deren Züge beschreibt, wohingehend VGG-Face sein Ergebnis auf Texturvergleiche stützt. Die Erkenntnis geht jedoch deutlich weiter als die Auswirkung unterschiedlicher Architekturen und deren Auswirkung zu kennen: Es kann gesagt werden, dass die Güte der Synthese vollständig abhängig von dem zum Einsatz kommenden Identifikationssystem ist.

4.3 Projection

Der letzte Teil der Evaluation widmet sich der dritten Lösungsmöglichkeit für heterogene Gesichtserkennung. Nachdem Feature verglichen und Bilder eingefärbt wurden, wird nun die Standartlösung für homogene Gesichtserkennung auf ihre Tauglichkeit zur heterogenen Identifikation hin untersucht.

Es wird der exakt gleiche Ansatz zu Validierung verwendet wie er bereits für die Synthese-Verfahren genutzt von (Li et al., 2013) beschrieben wurde.

4.3.1 Openface

Für das untenstehende ROC-Diagramm wurden zur besseren Einordnung nochmals die Kurven der performantesten Synthese-Systeme eingezeichnet. Da beide Verfahren sowohl auf der gleichen Datenbank beruhen, als auch gleichsam evaluiert wurden, ist eine direkte Vergleichbarkeit gegeben.

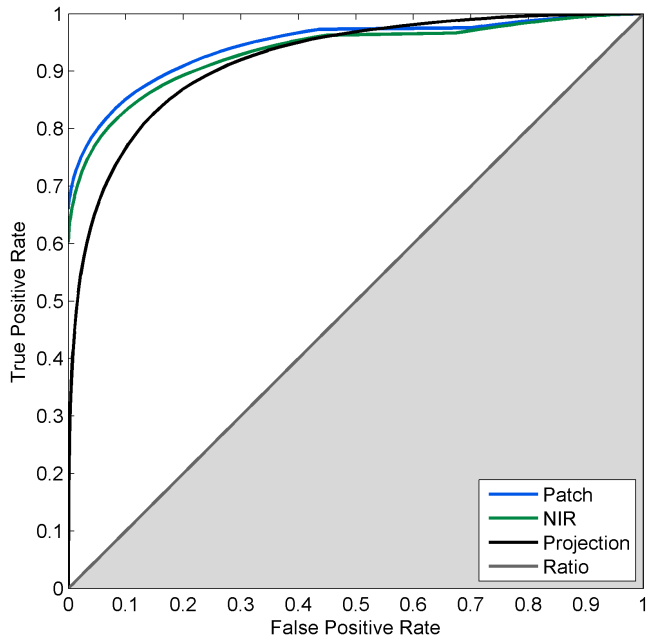

Abb. 4.7 ROC des selbst trainierten Openface Embedders

Es ist ersichtlich, dass das Projektionsverfahren deutlich über der Zufallskurve, jedoch unterhalb der Syntheseverfahren, liegt. Erst im Identifikationsbereich ist Openface performanter als die Synthese basierten Methoden. Dies kann durch die höhere Robustheit gegen Kopfverdrehungen erklärt werden. Wirft man einen Blick auf Zwischenergebnisse der Synthese so kann man feststellen, dass bei leichter Verzerrung des Gesichts, die Einfärbung fehlschlägt und eine nachfolgende Identifikation erschwert. In diesem Bereich schneidet das Projektionsverfahren besser ab, da es die Verzerrungen während der Trainingsphase explizit erlernen konnte.
Es bleibt die Frage ob Syntheseverfahren ausnahmslos besser geeignet sind, als Projektionsverfahren: Legt man zugrunde, dass moderne Projektionsverfahren auf

Millionen von Bildern trainiert werden und diesem im Schnitt lediglich 12000 zur Verfügung standen so ist davon auszugehen, dass bei einem hinreichend großen Datensatz der Perfomanzzuwachs bei Projektionsansätzen deutlich höher ausfällt als bei Syntheseverfahren.
Aber auch der Umkehrschluss ist eine Erkenntnis: Bei kleiner Datenbasis, kann ein Syntheseverfahren deutlich bessere Ergebnisse erzielen als ein reines Projektionsverfahren. Legt man zu Grunde, dass heterogene Datenbasen meisten klein und teuer zu beschaffen sind, so sind Syntheseverfahren der Projektion häufig vorzuziehen.

5 Ausblick

Verfahren der verschiedenen Kategorien wurden zuvor erläutert, implementiert und evaluiert. Der dadurch erreichte Erkenntnisgewinn eröffnet neue Richtungen für zukünftige Arbeiten. Die Abhängigkeit des Identifikationssystems von der vorangestellten Synthese und die visuelle Güte von Adversarialnetzen lassen vermuten, dass eine Verschmelzung beider Systeme einen beträchtlichen Genauigkeitszuwachs zur Folge hat.

Wie bereits mehrfach angesprochen besteht das Adversarialsystem in der Lernphase aus zwei Netzen: dem Dezisiven und dem Generativen. Die Idee ist, das dezisive Netz durch den später zum Einsatz kommenden Klassifikator zu ersetzen. Somit wäre garantiert, dass nicht die visuelle Abbildung optimiert wird, sondern die Erhaltung der Identität. Ein Embedder würde sich dafür besonders eignen, da der Abstand im Identifikationsraum als Maß der Unähnlichkeit zu verstehen ist, welches es in diesem Falle zu minimieren gilt. Der Nachteil dieses Verfahrens ist der hohe Rechenaufwand, welcher durch die wiederholte Identifikation im dezisiven Teil des Systems entstehen würde.

Fernab der rein technischen Zukunft, lässt sich auch eine Prognose der für die verschiedenen Anwendungsfälle treffen. In den letzten Jahren haben Kamerasysteme, welche zunächst für Spezialfälle entwickelt wurden, Einzug in Smartphones und handelsübliche Abbildungsgeräte gehalten. Die Dual-Kamera des IPhones ist ein Paradebeispiel für diesen Trend. Es erscheint naheliegend, dass in Zukunft auch NIR-Kameras integriert werden, um auch bei schwachen Lichtverhältnissen gute Ergebnisse zu erzielen. Die Intel RealSense R200, ein System mit zwei NIR und einer RGB-Kamera wird teilweise bereits in Tablets integriert und für Gesichtserkennung eingesetzt. Mit der Diversifizierung der Abbildungsgeräte gewinnt die heterogene Gesichtserkennung an Bedeutung. Durch die weitere Verbreitung der Anwendungssysteme wächst auch die verfügbare Datenbasis, welches die Möglichkeiten CNNs zu trainieren deutlich verbessert.

Abschließend lässt sich sagen, dass die heterogene Gesichtserkennung, trotz der Tatsache, dass sie eine Subkategorie der fortgeschrittenen Gesichtserkennung ist, relativ unentwickelt zu sein scheint. Während andere Teilgebiete deutlich präziser als die menschliche Wahrnehmung sein können, ist das für die heterogene Gesichtserkennung nicht der Fall. Von den drei Subkategorien: Feature, Synthesis und Projection, wird die Projection mit Zunahme der verfügbaren Datenbasis die erfolgreichste Verfahrensgruppe sein.

6 Fazit

Die Problematik der heterogenen Gesichtserkennung wurde zusammen mit den dazugehörigen Anwendungsfällen vorgestellt. Die Systematik von Lösungsansätzen wurde erläutert und bekannte Arbeiten der näheren Vergangenheit kategorisiert.
Es hat sich gezeigt, dass Merkmal-basierte Verfahren mit geringen Aufwand implementiert werden können, jedoch eher enttäuschende Ergebnisse liefern. Für NIR-VIS-Bilder ist der Klassifikator nur leicht besser als der Zufall. Für 2D-3D Videos sind die Merkmal-Klassifikatoren deutlich besser als der Zufall, aber immer noch zu schlecht um weiteren Entwicklungsaufwand zu investieren.
Darauffolgend wurden die Synthese-basierten Verfahren erläutert und mehrere Datenaufbereitungsverfahren entwickelt. Es konnte ermittelt werden, dass der Erfolg maßgeblich vom Zusammenspiel aller Komponenten (Datenregistrierung, Netzarchitektur, Klassifikator) abhängt. Textur-basierte Klassifikatoren funktionierten am besten mit einem vollständigen Gesichtsunwrapping wohingegen Erscheinung-basierte Verfahren besser mit Frontalisierungsverfahren funktionierten. Adversarialnetze, als neue Klasse von neuronalen Netzen spielten dabei eine große Rolle. Mit Hilfe eines Adversarialnetzes konnten Farbbilder halluziniert werden, welche sich nur lose an dem Inputbild orientieren. Es war also möglich nicht nur Textur, sondern auch Form zu variieren. Diese Netze funktionierten besonders gut mit Textur basierten Klassifikatoren jedoch sehr schlecht mit Form basierten Embeddern.
Zum Abschluss wurde die klassische Herangehensweise, die Projektion, auf ihre Tauglichkeit bezüglich des heterogenen Anwendungsfall hin überprüft. Mit deutlich weniger Aufwand in puncto Datenaufbereitung und Trainingsdauer, konnten ähnlich gute Ergebnisse wie in der Synthese erzielt werden. Die Synthese eignet sich gut für kleine Datenbasen, bei welchen das Training eines Klassifikators nicht möglich ist. Im jetzigen Stand der Technik sind sie folglich eine gute Alternative und Kombinationsmöglichkeit für herkömmliche Methoden.
Es kann davon ausgegangen werden, dass CNN basierte Projektionsverfahren auch in der heterogenen Gesichtserkennung das Mittel der Wahl werden, wenn die Datenbasis auf eine zum homogenen Problem vergleichbaren Größe ansteigt.

Literatur

Amos, Brandon, Bartosz Ludwiczuk und Mahadev Satyanarayanan (2016). *OpenFace: A general-purpose face recognition library with mobile applications*. Techn. Ber. (zitiert auf den Seiten 33, 34, 75–77).

Bas, Anil, William AP Smith, Timo Bolkart und Stefanie Wuhrer (2016). „Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences". In: *Asian Conference on Computer Vision*. Springer, S. 377–391 (zitiert auf den Seiten 51, 52).

Belhumeur, Peter N., João P Hespanha und David J. Kriegman (1997). „Eigenfaces vs. fisherfaces: Recognition using class specific linear projection". In: *IEEE Transactions on pattern analysis and machine intelligence* 19.7, S. 711–720 (zitiert auf Seite 6).

Huang, Di, Mohsen Ardabilian, Yunhong Wang und Liming Chen (2012). „Oriented gradient maps based automatic asymmetric 3D-2D face recognition". In: *Biometrics (ICB), 2012 5th IAPR International Conference on*. IEEE, S. 125–131 (zitiert auf den Seiten 13, 14, 16).

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou und Alexei A Efros (2016). „Image-to-Image Translation with Conditional Adversarial Networks". In: *arxiv* (zitiert auf den Seiten 36–38, 69).

Jia, Yangqing, Evan Shelhamer, Jeff Donahue et al. (2014). „Caffe: Convolutional architecture for fast feature embedding". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, S. 675–678 (zitiert auf Seite 74).

Juefei-Xu, Felix, Dipan K Pal und Marios Savvides (2015). „NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, S. 141–150 (zitiert auf den Seiten 18, 19).

Kanade, Takeo (1973). „Picture processing system by computer complex and recognition of human faces". In: *Doctoral dissertation, Kyoto University* 3952, S. 83–97 (zitiert auf Seite 5).

Lezama, José, Qiang Qiu und Guillermo Sapiro (2016). „Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-spectral Hallucination and Low-rank Embedding". In: *CoRR* abs/1611.06638 (zitiert auf den Seiten 19–21, 62–64, 66, 67, 69).

Li, Stan, Dong Yi, Zhen Lei und Shengcai Liao (2013). „The casia nir-vis 2.0 face database“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, S. 348–353 (zitiert auf den Seiten 18, 19, 23, 25, 59, 60, 77, 79, 83, 84, 88).

Li, Stan Z, Zhen Lei und Meng Ao (2009). „The HFB face database for heterogeneous face biometrics research“. In: *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, S. 1–8 (zitiert auf Seite 59).

Lin, Dahua und Xiaoou Tang (2006). „Inter-modality face recognition“. In: *Computer Vision–ECCV 2006*, S. 13–26 (zitiert auf den Seiten 23, 24).

Ouyang, Shuxin, Timothy M. Hospedales, Yi-Zhe Song und Xueming Li (2014). „A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-resolution“. In: *CoRR* abs/1409.5114 (zitiert auf den Seiten 3–9, 13, 21, 24–26).

Oxford VGG Pratical. `http://www.robots.ox.ac.uk/~vgg/practicals/cnn/`. Accessed: 2017-02-15 (zitiert auf den Seiten 28, 35, 39–41).

Parkhi, Omkar M, Andrea Vedaldi und Andrew Zisserman (2015). „Deep Face Recognition.“ In: *BMVC*. Bd. 1. 3, S. 6 (zitiert auf den Seiten 20, 74).

Paysan, Pascal, Reinhard Knothe, Brian Amberg, Sami Romdhani und Thomas Vetter (2009). „A 3D face model for pose and illumination invariant face recognition“. In: *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, S. 296–301 (zitiert auf den Seiten 56, 58).

Rama, Antonio, Francesc Tarres, Davide Onofrio und Stefano Tubaro (2006). „Mixed 2D-3D Information for pose estimation and face recognition“. In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. Bd. 2. IEEE, S. II–II (zitiert auf Seite 24).

Turk, Matthew A und Alex P Pentland (1991). „Face recognition using eigenfaces“. In: *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, S. 586–591.

Wright, John, Allen Y Yang, Arvind Ganesh, S Shankar Sastry und Yi Ma (2009). „Robust face recognition via sparse representation“. In: *IEEE transactions on pattern analysis and machine intelligence* 31.2, S. 210–227 (zitiert auf Seite 15).

Zhu, Jun-Yong, Wei-Shi Zheng und Jian-Huang Lai (2013). „Logarithm gradient histogram: A general illumination invariant descriptor for face recognition“. In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, S. 1–8 (zitiert auf den Seiten 10, 11, 13).

Zhu, Xiangxin und Deva Ramanan (2012). „Face detection, pose estimation, and landmark localization in the wild“. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, S. 2879–2886 (zitiert auf den Seiten 46, 47).

Abbildungsverzeichnis

Tabellenverzeichnis