ulm university universität
uulm

Fakultät für Ingenieurwissenschaften und Informatik
Institut für Neuroinformatik

Lern-basierte multimodale Registrierung

Masterthesis

vorgelegt von
Thomas Probst

Beginn:	15. Mai 2014
Abgabe:	15. November 2014

Betreuer:	Dr. Christian Wojek
1. Prüfer:	Prof. Dr. Heiko Neumann
2. Prüfer:	Prof. Dr. Günther Palm

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe und die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe.

Ulm, den 15.11.2014

Thomas Probst

Danksagung

An dieser Stelle möchte ich meinen Dank an alle Personen richten, die bei der Erstellung dieser Arbeit unterstützend mitgewirkt haben.

Mein besonderer Dank gilt Herrn Dr. Christian Wojek für die engagierte Betreuung, die ich vor Ort bei der Carl Zeiss AG in Form vieler Ideen, Anregungen und konstruktiver Kritik erfahren durfte.

Neben der universitären Betreuung und den vielen kritischen Nachfragen möchte ich Herrn Prof. Dr. Heiko Neumann für das informative und verständliche Heranführen an die vielseitige *Computer Vision* Thematik in den Vorlesungen danken.

Ich bedanke mich beim BMBF, dem Fördergeber des Projekts NanoCombine (FKZ: 13N11402), im Rahmen dessen diese Masterarbeit durchgeführt und gefördert wurde.

Außerdem möchte ich mich bei der Carl Zeiss Meditec AG für die Bereitstellung des Funduskopie-Bilddatensatzes bedanken.

Abstract

Die Registrierung von Bildern - also das örtlich korrekte Ausrichten und Zusammenfügen von Aufnahmen derselben Szene unter verschiedenen Aufnahmebedingungen - liegt einer Vielzahl von Anwendungen in der Bildverarbeitung zu Grunde, beispielweise dem *Stitching* von Panoramabildern oder der HDR[1]-Bildgebung. Registrieralgorithmen suchen dazu eine geometrische Transformation (bspw. Translation, Rotation oder Skalierung), welche die Ähnlichkeit überlappender Bildbereiche maximiert. Im multimodalen Fall stammen die Bildpaare aus unterschiedlichen Bildgebungsverfahren und ermöglichen daher die Extraktion und Fusion von unterschiedlichster Arten von Information über eine Probe. Da dieselbe Szene bei verschiedenen Aufnahmemodalitäten stark voneinander abweichende Erscheinungsbilder aufweisen kann, ist ein modalitätsübergreifendes Ähnlichkeitsmaß über konventionelle Metriken wie die euklidische Distanz hierbei nicht mehr gegeben. Unimodale Registrierverfahren können auf diesem Anwendungsgebiet daher oft nicht die gewünschte Robustheit und Genauigkeit erreichen. In dieser Arbeit werden verschiedene Ansätze untersucht, welche durch das Erlernen von applikationsspezifischer Ähnlichkeit aus Trainingsdaten den keypoint-basierten Registrierprozess im Bezug auf multimodale Daten unterstützen können.

Es werden zu diesem Zweck zum einen merkmalsbasierte Ansätze, darunter ein Wörterbuchverfahren sowie zwei *metric learning*-Verfahren, und zum anderen eine bildbasierte Methode (Image Analogies) verfolgt, welche eine filterartige Transformation zwischen Modalitäten erlernt. Nach einer Beschreibung der Methoden sowie der Auswahl verwendeter Trainingsdaten erfolgt die experimentelle Untersuchung auf verschiedenen multimodalen Datensätzen aus der Funduskopie sowie korrelativer Mikroskopie.

Die quantitative Evaluation und die anschließende Analyse der Ergebnisse zeigt auf, dass der Wörtbuchansatz für diese Anwendung eine ungenügende Skalierbarkeit und Generalisierungsfähigkeit aufweist. Die *metric learning*-Verfahren legen in dieser Hinsicht ein besseres Verhalten an den Tag, jedoch schränkt die zu geringe Diskriminanz der gelernten Metriken die Tauglichkeit für keypoint-basierte Registrierverfahren stark ein.

Die vorgeschaltete Transformation eines der Bilder in die jeweils andere Modalität mittels Image Analogies kann ein unimodales Registrierverfahren gerade bei stark abweichenden Erscheinungsbildern unterstützen, jedoch kann das Verfahren nur schlecht generalisieren und ist auf eine große Menge repräsentativer Trainingsbilder angewiesen. Für Anwendungsfälle mit vielen Variationen in Beleuchtung, Rauschen, Skalierung und Rotation von Strukturen skaliert das Verfahren ungenügend und stellt sich daher als ungeeignet heraus.

[1] High Dynamic Range

Inhalt

1 Einführung und Motivation **1**
1.1 Aufbau und Ziel der Arbeit 4

2 Stand der Technik **5**
2.1 Registrierung 5
2.2 Multimodale Registrierung 7
2.3 Verfahren zur multimodalen Registierung 8
2.4 Prozesspipeline und Eingriffsmöglichkeiten 10

3 Lern-basierte Ansätze **15**
3.1 Wörterbuch-basierter Ansatz 15
3.1.1 Beschreibung 15
3.1.2 Matcher 15
3.1.3 Einordnung 16
3.1.3.1 Parameter und Komplexität 16
3.1.3.2 Bewertung 16
3.1.3.3 Erweiterungen 17
3.2 Similarity-Sensitive Hashing 19
3.2.1 Beschreibung 19
3.2.2 Matcher 21
3.2.3 Einordnung 22
3.2.3.1 Parameter und Komplexität 22
3.2.3.2 Bewertung 24
3.3 Random Forests for Metric Learning 26
3.3.1 Beschreibung 26
3.3.2 Matcher 27
3.3.3 Einordnung 28
3.3.3.1 Parameter und Komplexität 28
3.3.3.2 Bewertung 28

3.4 Image Analogies 29
3.4.1 Beschreibung 29
3.4.2 Einordnung 30
3.4.2.1 Parameter und Komplexität 30
3.4.2.2 Bewertung 32

4 Auswahl der Trainingsdaten 35
4.1 Wörterbuch und diskriminative Verfahren 35
4.1.1 Positive Trainingspaare 35
4.1.2 Negative Trainingspaare 36
4.2 Image Analogies 39

5 Auswertung 41
5.1 Datensätze 41
5.2 Evaluationsmaße 42
5.3 Ergebnisse 43
5.3.1 Wörterbuch-basiert 44
5.3.2 Similarity-Sensitive Hashing 53
5.3.3 Random Forest Distance 62
5.3.4 Image Analogies 65

6 Ausblick 75

Literatur 77

Abbildungsverzeichnis

1.1 Anwendungsbeispiel für multimodale Registrierung aus der korrelativen Mikroskopie . . . 2
1.2 Funduskopie Beispiel . . . 3

2.1 Ungenaues Registrierergebnis . . . 7
2.2 Visualisierung des keypoint-basierten Registrierverfahrens. . . . 11

3.1 Similarity-Sensitive Hashing: Verlauf der Bitgewichte . . . 21
3.2 Similarity-Sensitive Hashing: Approximation der Bitgewichte . . . 23
3.3 Visualisierung Image Analogy Prozess . . . 30
3.4 Image Analogies auf Daten der korrelativen Mikroskopie . . . 31

4.1 Visualisierung von Beispielkorrespondenzen bei der Fundusbildgebung . . 37
4.2 Visualisierung gefundener positiver Korrespondenzen . . . 38
4.3 Trainingsdatenbeispiel für Image Analogies . . . 39

5.1 Visualisierung einer korrekten Korrespondenz . . . 42
5.2 Euklid-Baseline: Häufigkeitsverteilung der Matching-Distanzen . . . 43
5.3 Euklid-Baseline: Visualisierung false positives . . . 44
5.4 Euklid-Baseline: ROC-Plot . . . 45
5.5 Einfluss der zulässigen Nachbarschaftsgröße beim Wörterbuch . . . 46
5.6 Einfluss der zulässigen Nachbarschaftsgröße beim Wörterbuch mit linearer Approximation . . . 47
5.7 Einfluss der Anzahl an Einträgen beim Wörterbuch . . . 47
5.8 Einfluss der Anzahl an Einträgen beim Wörterbuch mit linearer Approximation 48
5.9 Einfluss des verwendeten Suchverfahrens bei der Wörterbuchsuche . . . 49
5.10 Einfluss verschiedener Trainingsdaten beim Wörterbuch mit linearer Approximation . . . 49
5.11 Vergleich der besten Wörterbuchverfahren zur Euklid-Baseline . . . 50
5.12 Visualisierung der geschätzten 512×512 Transformationsmatrix aus dem Least-Square-Verfahren . . . 51

5.13 Vergleich der linearen Approximation zur Euklid-Baseline mit *matching ratio* 1.035 52
5.14 Verteilung der Distanzen positiver und negativer Paare auf synthetisch generierten bimodalen Daten. 53
5.15 Einfluss der approximierten Kovarianzmatrix beim Similarity-Sensitive Hashing 54
5.16 Einfluss der Anzahl an zufälligen Linearkombinationen beim Similarity-Sensitive Hashing 56
5.17 Einfluss der Länge des Bitstrings beim Similarity-Sensitive Hashing 57
5.18 Similarity-Sensitive Hashing: ROC-Kurven verschiedener Trainingsdaten . 58
5.19 Verteilung der Distanzen positiver und negativer Paare auf „Fundus I"-Trainingsdaten. 59
5.20 Die Abbildung zeigt den Verlauf der Schnittfläche der Distanzverteilungen positiver und negativer Trainingspaare (links) sowie eine Visualisierung der Entwicklung der Distanzen (rechts) mit steigender Anzahl an Trennfunktionen (Bits). 60
5.21 Similarity-sensitive Hashing: Visualisierung false positives 61
5.22 Visualisierung Distanzverhältnisse zum besten Treffer 61
5.23 Distanzverhältnisse zum besten Treffer 63
5.24 Skizze der gelernten Trennfunktionen 63
5.25 Random Forest Distance: ROC-Kurven verschiedener Trainingsdaten . . . 64
5.26 Random Forest Distance: Visualisierung false positives 65
5.27 Image Analogies: Filterungsrichtung 66
5.28 Image Analogies: Filterergebnisse Funduskopie 67
5.29 Image Analogies: ROC-Plot „Fundus II"-Datensatz 68
5.30 Image Analogies: Hautaufnahme mittels korrelativer Mikroskopie 69
5.31 Image Analogies: korrelative Mikroskopie (Haut) ROC-Plot 70
5.32 Image Analogies: korrelative Mikroskopie (Zellgewebe) ROC-Plot 71
5.33 Image Analogies Ergebnisse bei der korrelativen Mikroskopie 73

Tabellenverzeichnis

3.1 Parameter beim Similarity-Sensitive Hashing 24
3.2 Parameter beim Random Forest Training 28
3.3 Parameter bei Image Analogies 32

5.1 Rechenzeiten beim Similarity-Sensitive Hashing 55

Kapitel 1

Einführung und Motivation

Bildregistrierung ist eines der klassischen Bildverarbeitungsprobleme und hat viele Anwendungen wie zum Beispiel Stitching [58], Image Fusion [57] oder HDR [4]. Viele der Arbeiten zu diesem Thema entwickeln Methoden zum Registrieren von Bildern der gleichen Modalität und erreichen dabei sehr gute Ergebnisse [58]. Aber auch die Registrierung von Bildern unterschiedlicher Bildmodalitäten ist ein wichtiges und weitaus anspruchsvolleres Problem, welches in dieser Arbeit näher behandelt werden soll. Als schwierig stellen sich für mikroskopische Anwendungen dabei insbesondere der starke Wechsel des Bildkontrasts in unterschiedlichen Modalitäten sowie große Unterschiede in der Auflösung dar. Aber gerade für biomedizinische Anwendungen ist die Registrierung von multimodalen Daten von Belang, da aus der Kombination unterschiedlicher Bildgebungsverfahren oftmals verschiedene Arten von Informationen gewonnen werden können.

In der Literatur finden sich zur Registrierung multimodaler Bilder vorwiegend Arbeiten aus dem medizinischen Anwendungsbereich mit Daten aus CT und MRT-Aufnahmen [64, 41, 17, 56, 67]. Der Unterschied zwischen den Modalitäten ist hier beispielweise in den Kontrasten sicher vorhanden, aber nicht sehr ausgeprägt, so dass oftmals Ähnlichkeitsmaße wie *mutual information* zu guten Ergebnissen führen. Insbesondere im Bereich der Mikroskopie gibt es jedoch Anwendungen, bei denen aufgrund von Kontrastinversionen, fehlenden Detailstrukturen oder Bildstörungen Unterschiede in den Erscheinungsbildern der Modalitäten sehr gravierend sind, so dass einfache Metriken wie *mutual information* oftmals kein geeignetes Ähnlichkeitsmaß darstellen. Zusätzlich spielen unterschiedliche Auflösungen der Aufnahmegeräte eine wichtige Rolle.

Diese Arbeit beschäftigt sich unter anderem mit Bilddaten aus der *korrelativen Mikroskopie* (siehe Abbildung 1.1). Ziel dabei ist es, die Informationen aus einem funktionellen Bildgebungsverfahren (Fluoreszenzmikroskopie) mit der hochauflösenden Information eines Elektronenmikroskops zu kombinieren [22]. Dabei kann die Komplementarität der Modalitäten genutzt werden - Fluoreszenzlichtmikroskopie erlaubt das Markieren von Molekülen mit Farbstoffen, die mittels Fluoreszenz sichtbar werden, ist allerdings in der maximalen Auflösung begrenzt. In der Elektronenmikroskopie hingegen kann Fluoreszenz nicht spezifisch beobachtet werden, allerdings ist die verfügbare Auflösung um eine Vielfaches besser als im Lichtmikroskop, sodass betroffene Strukturen im Detail untersucht werden können. Neben

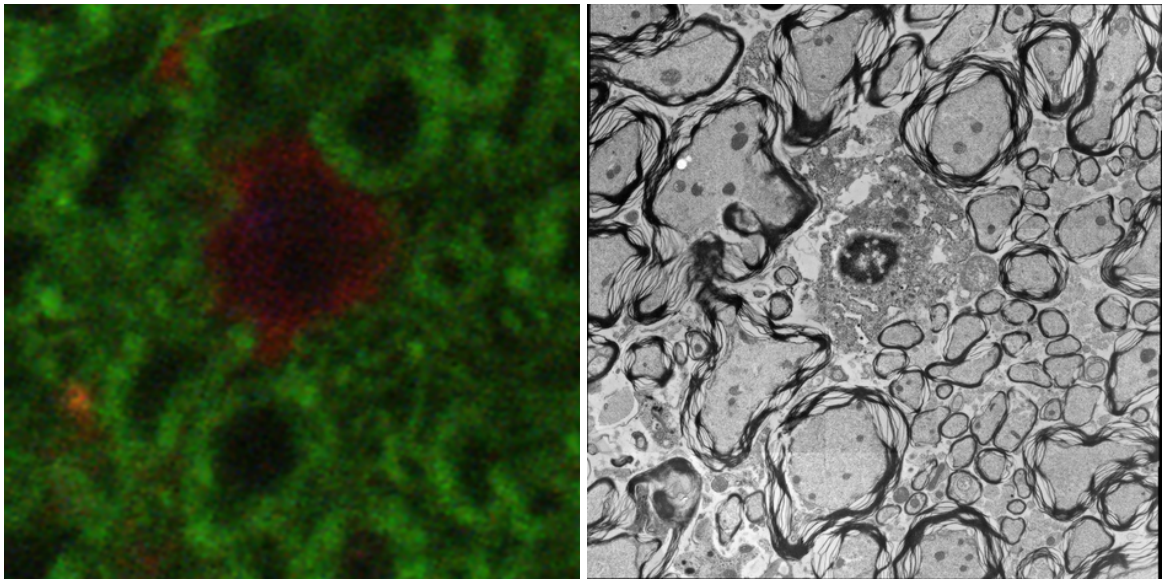


Abbildung 1.1: Anwendungsbeispiel für multimodale Registrierung aus der korrelativen Mikroskopie. Das linke Bild zeigt eine Aufnahme von Zellgewebe eines Lichtmikroskops. Rechts ist derselbe Bereich dargestellt, aufgenommen durch ein Transmissions-Elektronenmikroskop (TEM). Obwohl dieselben Strukturen abgebildet sind, ist eine Ähnlichkeit ausschließlich auf Basis von Pixelintensitäten nicht gegeben. Bildmaterial aus [15].

der Erschließung neuer Ansätze im Bereich der Zellbiologie hat das BMBF-geförderte Projekt *NanoCombine* zum Ziel, Instrumente zur Erforschung von Krankheiten auf molekularer Ebene auf Basis einer solchen Technologie realisierbar zu machen.

Ähnlich können in der Funduskopie (siehe Abbildung 1.2) in der Augenheilkunde einige Krankheitsmerkmale auf der Oberfläche der Retina besser im Farbfundusbild erkannt werden, wohingegen andere Merkmale in der Autofluoreszenzbildgebung besser zu sehen sind. Zusätzlich ist es mit Hilfe des OCT-Verfahrens[1] möglich, die tieferen Schichten der Retina zu erfassen, so dass eine registrierte Kombination von Fundusfarbbild, Fundusautofluoreszenzbild und OCT-Volumen dem Augenarzt ein umfassendes Bild gibt.

Wie aus den Abbildungsbeispielen ersichtlich wird, sind die Unterschiede im Erscheinungsbild zwischen Modalitäten sehr erheblich. Diese Arbeit befasst sich daher mit Methoden des maschinellen Lernens, die es erlauben sollen, die Kontrastwechsel und Darstellungsunterschiede zu erlernen. Insbesondere sollen dabei mehrere in der Literatur für direkte intensitätsbasierte Registrierung vorgeschlagene Lernverfahren [13, 65, 33] in einem Framework für keypoint-basierte Registrierung eingebettet und evaluiert werden.

[1] Optische Kohärenztomographie

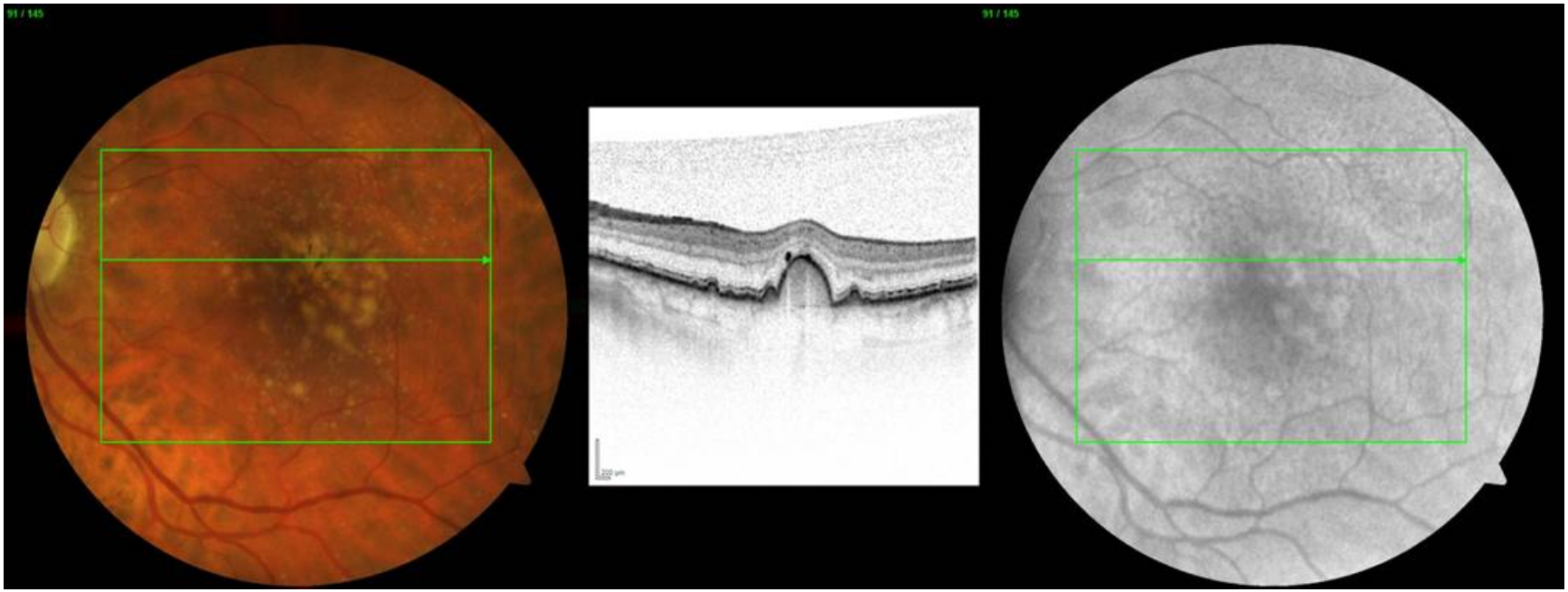


Abbildung 1.2: In der Funduskopie werden verschiedene Bildgebungsverfahren zur Gewinnung von diagnoserelevanten Merkmalen der Netzhaut eingesetzt. Links ist ein Farbfundusbild abgebildet, rechts das zugehörige Autofluoreszenzbild. Zusätzlich können über das OTC-Verfahren Querschnittsbilder über tieferliegende Netzhautschichten gewonnen werden (mitte).

1.1 Aufbau und Ziel der Arbeit

Nach einer generellen Erörterung der Eingriffsmöglichkeiten und einem Überblick über geeignete Lernverfahren in Kapitel 2 werden die ausgewählten Ansätze in Kapitel 3 im Detail beschrieben und in Bezug auf theoretische Vor- und Nachteile bewertet. Nach einer Diskussion über Möglichkeiten zur Auswahl von Trainingsdaten in Kapitel 4 folgt die Beschreibung der experimentellen Untersuchung der Verfahren und die zugehörige Auswertung in Kapitel 5. Abschließend soll im letzten Kapitel ein Ausblick über weitere Vorgehensweisen zum Thema gegeben werden.

Kapitel 2

Stand der Technik

Dieses Kapitel soll nach einer kurzen Beschreibung gängiger Algorithmen zur Bildregistrierung die Schwierigkeiten multimodaler Bilddaten darlegen. Im Anschluss werden verschiedene Ansätze für das multimodale Registrierungsproblem aus der Literatur vorgestellt und im Hinblick auf Bilddaten und Registrierverfahren diskutiert. Daraufhin wird ein Überblick über das verwendete Registrierframework gegeben um sinnvolle Eingriffsmöglichkeiten aufzuzeigen.

2.1 Registrierung

(Bild-)Registrierung (*image alignment*) bezeichnet allgemein die Aufgabe bzw. das Vorgehen, die verschiedenen Koordinatensysteme eines vorliegenden Bildpaares zueinander in Bezug zu setzen und diese somit in einem gemeinsamen, globalen Koordinatensystem zu vereinen. Das Ausrichten erfolgt dabei über das Wiederfinden von Strukturen oder Bildausschnitten im Überlappungsbereich der beiden Eingabebilder, woraus eine geometrische Beziehung zwischen den Koordinatensystemen abgeleitet werden kann.

Ein Registrierverfahren lässt sich allgemein anhand der folgenden drei Aspekte charakterisieren [58, 31]:

(i) Transformationsmodell
(ii) Ähnlichkeitsmaß
(iii) Optimierungsstrategie

Zunächst muss das zugrundeliegende Transformationsmodell (i) ausgewählt werden. Die Wahl hängt dabei stark vom Anwendungsfall ab und hat einen großen Einfluss auf Rechenaufwand und Robustheit des Verfahrens. So stehen Modelle mit variabler Anzahl und Art von Freiheitsgraden zur Verfügung [58, 1], beispielsweise für Translation, Rotation, Scherung und Skalierung von Bildern (affines Modell, starre Transformation). Für Anwendungen mit starker räumlicher Perspektive, optischen Verzerrungen oder anderen nichtlinearen Effekten reicht das affine Modell allerdings nicht aus und es muss auf projektive Transformationen oder allgemein deformierbare Modelle zurückgegriffen werden. Je mehr Parameter

bzw. Stützstellen ein solches Modell aufweist, umso höherdimensional ist der resultierende Suchraum und darauf aufbauende Optimierung wird entsprechend aufwändiger.

Eine Registrierung lässt sich als Optimierungsprozess interpretieren, da eine Parametrierung für die Transformation (i) gesucht wird, die eine entsprechend gewählte Kostenfunktion minimiert. Dies setzt zum einen ein Ähnlichkeitsmaß (ii) voraus, welches die Güte einer Transformation und damit der Registrierung bewertet. Zum anderen kann die Registrierung auf verschiedene Arten, also über unterschiedliche Optimierungsstrategien (iii) erfolgen.

Man kann dabei prinzipiell zwei Arten unterscheiden [58]:

Bildbasierte (direkte) Registrierung Bei bildbasierten Verfahren wird in jedem Iterationsschritt die Kostenfunktion evaluiert, indem das eine Bild (*floating image*) gemäß der Parameter aus dem letzten Schritt transformiert (*warping*) wird und anschließend die Ähnlichkeit zum anderen Bild (*target image*) herangezogen wird. Dazu muss ein angemessenes Ähnlichkeitsmaß gewählt werden, beispielsweise pixelweise ℓ_2-Distanzen. Die Transformation mit dem geringsten Fehler wird schließlich als Lösung übernommen. Diese Methode weist aufgrund der lokalen Bewertung eine hohe Genauigkeit auf, allerdings besteht stets die Gefahr lediglich ein lokales Optimum der im Allgemeinen nichtkonvexen Kostenfunktion zu finden, was zu einer inkorrekten Registrierung führt. Ob das globale Optimum gefunden wird hängt stark von der Initialisierung der Optimierungsparameter ab. Zwar könnten diese Probleme zum Teil mit einer Grob-nach-Fein-Strategie (Auflösungspyramide) reduziert werden, was jedoch zusätzlichen Rechenaufwand mit sich bringt. Dabei steigt der Berechnungsaufwand bei der numerischen Auswertung der Kostenfunktion bzw. des Gradienten sowie beim anschließenden *warping* entsprechend mit der Bildgröße an.

Keypoint-basierte Registrierung Hierbei werden auf beiden Bildern markante Punkte (*keypoints*) gesucht und über die Ähnlichkeit der Beschreibungen ihrer jeweiligen Umgebung (Deskriptoren) in Korrespondenz gesetzt. Anschließend wird mit robusten Schätzverfahren die Transformation ermittelt, die alle relevanten Korrespondenzen global am besten erklärt. Dies führt in vielen Anwendungsfällen insgesamt zu einen höheren Robustheit bei einem geringerem Berechnungsaufwand. Dabei muss die Annahme erfüllt sein, dass an korrespondieren Stellen der beiden Bildern Keypoints vorliegen und die Deskriptoren über eine geeignete Ähnlichkeitsmetrik verglichen und zugeordnet werden können. Da im Rahmen dieser Arbeit ausschließlich Experimente mit diesem Verfahren durchgeführt werden, wird die gesamte Prozesskette in Kapitel 2.4 nochmals im Detail beschrieben.

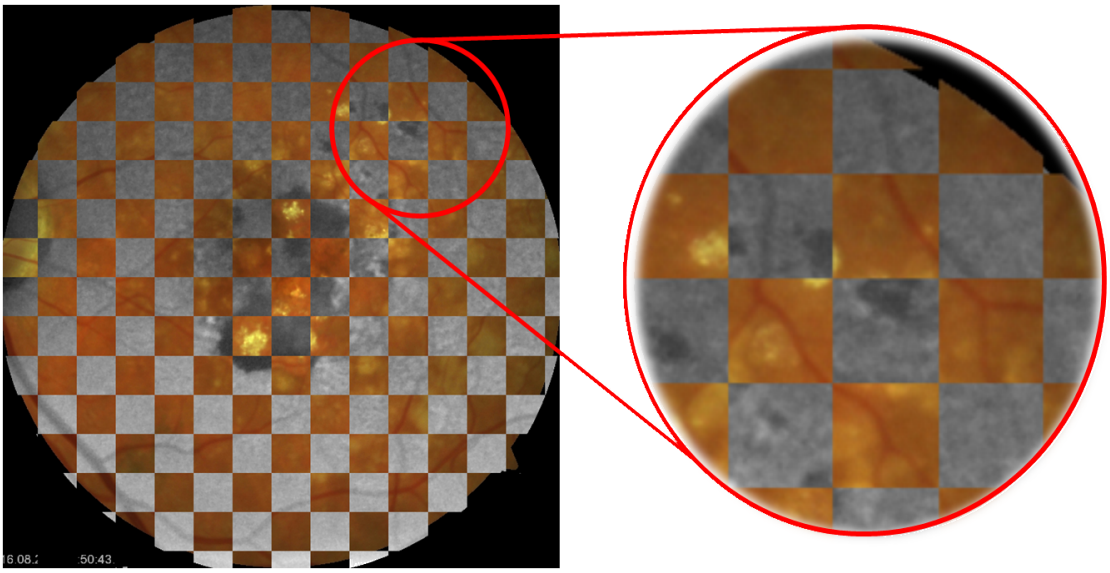


Abbildung 2.1: Die Abbildung zeigt am Beispiel eines multimodalen Bildpaares aus der Fundusbildgebung (siehe Kapitel 5.1) das Ergebnis der Registrierung, indem schachbrettartig jeweils abwechselnd ein Bildausschnitt der einen Modalität (Farbphotographie) und der anderen (Autofluoreszenzbild nach Anwendung der geometrischen Transformation aus dem Registrierergebnis) dargestellt wird. An der vergrößerten Stelle (rechts) sind anhand der zueinander verschobenen und nicht kontinuierlich ins benachbarte Feld übergehenden Adern Ungenauigkeiten zu erkennen, die auf fehlende oder falsch zugeordnete Korrespondenzen in diesem Bildbereich zurückzuführen sind. Auch Verzerrungseffekte oder ein ungenügend flexibles Transformationsmodell können an dieser Stelle problematisch sein.

2.2 Multimodale Registrierung

Unabhängig von der Methode liegt jedem Registrierverfahren demnach eine Ähnlichkeitsmetrik zugrunde, die maßgeblich die Qualität des Ergebnisses beeinflusst.

Liegt ein Bildpaar mit geringen Änderungen in Beleuchtung, Auflösung, Farbe sowie lediglich kleineren Störungen vor, so spricht man von Aufnahmen derselben *Modalität*. Hierbei sind ℓ_p-Distanzen zwischen Intensitätswerten oder abgeleiteten Merkmalen als Ähnlichkeits- bzw. Abstandsmaß sehr gut geeignet.

Sind dagegen die obigen Annahmen (teilweise) verletzt, oder stammen die Bilder aus verschiedenen Bildgebungsverfahren, liegt ein *multimodaler* Fall vor. Dies ist unter anderem der Fall bei Bestrahlung und Messung mit unterschiedlichen Wellenlängen, Variationen in der Ausleuchtung oder der Verwendung unterschiedlicher Sensorarten und -auflösungen.

Multimodale Registrierung kann in diesem Sinne als Ansatz zur Informationsfusion betrachtetet werden, der insbesondere in Anwendungsfällen von großem Nutzen ist, bei denen Bildgebungssysteme nicht oder nur ungenügend genau vorab zueinander kalibriert werden können. Abbildung 1.1 veranschaulicht dies am Beispiel der korrelativen Mikroskopie. Die korrekte örtliche Zuordnung ermöglicht somit die Extraktion einer Vielzahl von Eigenschaften einer Szene bzw. eines Objekts von physikalischer, optischer, medizinischer oder biologi-

scher Relevanz. Die Integration aller verfügbarer Informationen stellt oft eine essentielle Voraussetzung für eine darauf aufbauende Analyse oder Klassifikation dar und findet daher Anwendung in verschiedensten Bereichen wie in den Materialwissenschaften, in der Medizin (z.B. Funduskopie, MRT, CT) oder bei der korrelativen Mikroskopie.

Für ein multimodales Bildpaar bedeutet dies, dass dieselbe Szene auf verschiedenen Größenordnungen, in anderen Farbkanälen, variierten Intensitätsverteilungen, mit invertieren Kontrastorientierungen oder zusätzlichen bzw. fehlenden Details dargestellt sein kann. Die reine Pixelähnlichkeit einer Nachbarschaft ist somit kein zuverlässiges Maß mehr für Ähnlichkeit der abgebildeten Strukturen. Dies hat zur Folge, dass die Qualität der Registrierung im Sinne von Genauigkeit und Robustheit nachlässt (siehe Abbildung 2.1).

Um eine bessere Vergleichbarkeit zwischen Bildern zweier Modalitäten zu erreichen, können mittels einer geeigneten Vorverarbeitung (*preprocessing*) beispielsweise leichte Beleuchtungsschwankungen und Rauscheinflüsse kompensiert werden, oder auch Umrechnungen in andere Farbräume (z.B. Projektion auf Grauwerte) stattfinden. Bei schwierigeren Problemen ist dies für eine robuste Registrierung nicht ausreichend, so dass darüber hinaus auf andere Verfahren zurückgegriffen werden muss, welche im nächsten Abschnitt diskutiert werden.

2.3 Verfahren zur multimodalen Registierung

Aufgrund der in Kapitel 2.1 genannten Vorteile keypoint-basierter Registrierverfahren gegebenüber direkter Verfahren, insbesondere im Hinblick auf Robustheit und Berechnungsaufwand, sollen im Folgenden Methoden diskutiert werden, die keypoint-basierte Verfahren bei der Anwendung auf multimodalen Bilddaten unterstützen können.

Yang et al. [66] stellen ein iteratives keypoint-basiertes Registrierverfahren für multimodale Bilddaten vor, welches unter der Annahme funktioniert, dass einige charakteristische Keypoints über deren SIFT[1]-Deskriptoren [40] modalitätsübergreifend korrekt zugeordnet werden können. Ausgehend von der besten gefundenen Korrespondenz wird für die zugehörige lokale Region ein Transformationsmodell geschätzt. Anschließend wird für alle detektierten Ecken- und Kantenpunkte innerhalb dieser Region gemäß der geschätzten Transformation die Zuordnung auf dem anderen Bild gesucht. Mit den zusätzlichen Korrespondenzen kann nun das Transformationsmodell mittels robuster Schätzung verfeinert oder ggf. auf eines mit einer größeren Anzahl an Freiheitsgraden gewechselt werden. Abschließend wird in jedem Schritt die betrachtete Region gemäß des lokalen Transformationsfehlers vergrößert. Das vorgehen wird bis zur Konvergenz iterativ wiederholt. Schlägt das „Dual Bootstrap“ genannte Verfahren fehl, so wird mit der zweit- bzw. nächstbesten Korrespondenz neu initialisiert. Zwar verhält sich dieses Verfahren viel robuster als eine globale Korrespondenzfindung in einem Schritt, dennoch hat es neben dem hohen Berechnungsaufwand den Nachteil, dass im multimodalen Fall die SIFT-Deskriptoren im Allgemeinen keine Ähnlichkeit zwischen Modalitäten im Sinne der verwendeten euklidschen Distanz aufweisen. Insbesondere bei korrelativer Mikroskopie können Strukturen mit unterschiedlichen Kontrasten und

[1] Scale-invariant feature transform

-richtungen oder Texturen abgebildet sein, was diese Annahme verletzt. So ist das Verfahren beispielweise nicht in der Lage, die in Kapitel 5.3.4 verwendeten Haut- und Zellaufnahmen (Abbildung 5.30 und 5.33)zu registrieren.

Kelman et al. [37] schlagen daher eine Variante des SIFT-Deskripors vor, die invariant gegenüber Kontrastwechseln bzw. der Gradiantenrichtung agiert. Dazu wird unter anderem der Wertebereich des Gradientenhistogramms auf den Winkel π halbiert, Werte darüber werden gespiegelt aufgetragen (*gradient mirroring*). So erhalten Umgebungen mit antiparallelen Gradientenrichtungen einen ähnlichen Merkmalsvektor. Während diese Art von Deskriptor die Registrierung multimodaler Daten mit einem dominanten Anteil invertierter Kontraste verbessern kann, ist die Tauglichkeit für Bilddaten mit nichtlinearen Zusammenhängen zwischen Intensitäten nicht gegeben. Zudem geht durch die Halbierung des Wertebereichs Information über die Umgebungsgradienten verloren, die für die Korrespondenzfindung über andere Strukturen im Bild von Vorteil sein kann.

Andere Ansätze versuchen Korrespondenzen über Beschreibung von modalitätsspezifischer Selbstähnlichkeit (*self-similarity*) lokaler Keypointumgebungen zu finden [34, 10]. Dies setzt allerdings eine gewisse Konsistenz innerhalb der Umgebung voraus, die bei fehlenden Strukturen (beispielsweise aufgrund von Auflösungsunterschieden) und Bildstörungen verletzt ist. Des Weiteren können Deskriptoren dieser Art in einigen Fällen nur ungenügend diskriminative Beschreibungen extrahieren, da insbesondere Skalierungs- und Orientierungsinformation nur begrenzt ausgenutzt werden kann. Daher sind *self-similarity*-Ansätze nur für eine begrenzte Menge an multimodalen Daten geeignet.

Anstatt die Korrespondenzfindung oder die Art des Deskriptors an das jeweilige Problem anzupassen, ist es ebenso möglich beide Bilder in eine einheitliche Repräsentation zu bringen. Dazu können sich je nach Anwendungsfall Gradientenbilder [32] oder strukturelle Beschreibungen [61] (beispielweise über Entropiebilder [62]) anbieten, die allerdings als Datenbasis für SIFT-ähnliche Deskriptoren weniger geeignet sind. Alternativ können Verfahren eingesetzt werden, die direkt lernen, wie sich Strukturen und deren Umgebungen zwischen Modalitäten abbilden. Dies kann sowohl auf der Bildebene (siehe *Image Analogies* [33], Kapitel 3.4), als auch auf davon abgeleiteten Merkmalen erfolgen (siehe Wörterbuch-basierte Verfahren, Kapitel 3.1).

Eine andere Herangehensweise besteht darin, die zugrundeliegende Ähnlichkeitsmetrik und die damit verbundene Kostenfunktion an das multimodale Problem zu adaptieren, anstatt die Bilder selbst anzugleichen. Da dieses Vorgehen in der Literatur selten in Kombination mit keypoint-basierten Registrierverfahren diskutiert wird, sollen die Methoden zunächst anhand direkter Verfahren vorgestellt werden.

So ist es möglich von der Übereinstimmung jeweiliger Pixelintensitäten zu abstrahieren, indem ein Ähnlichkeitsmaß auf Basis der Entropie gewählt wird. Somit können Korrelationen zwischen Bildbereichen und damit die optimale Registrierung verschiedener Modalitäten gefunden werden, indem die Mutual Information maximiert wird [64, 41]. Hierbei wird ausgenutzt, dass Bilder verschiedener Modalitäten zwar mit anderen Intensitätswerten repräsentiert sein können, diese allerdings statistisch voneinander abhängig sind. Das bedeutet, dass das Bild der einen Modalität Information über das der anderen enthält, was sich

in der Mutual Information (MI) widerspiegelt. Neben bereits genannten Nachteilen direkter Registrierverfahren sind für derartige Methoden vor allem fehlende (Detail-)Strukturen und Rauschen problematisch, da hierbei Zusammenhänge zwischen korrespondierenden Intensitätswerten nicht global konsistent sind. Da zudem keinerlei örtlichen Korrelationen zwischen Pixelintensitäten ausgenutzt werden, kann die MI-Metrik um einen ortsvarianten Term (beispielsweise lokale Gradienteninformation [50]) erweitert werden, was Registriergenauigkeit und Konditionierung der Kostenfunktion zugute kommt.

Anhand bereits registrierter Trainingsbildpaare kann die Registrierung unterstützt werden, indem die Statistik der Häufigkeiten zusammen auftretender Pixelintensitäten (oder abgeleiteter Merkmale [67]) im einen Bild und der korrespondierenden Pixel im anderen Bild in Form von bivariaten Häufigkeitsverteilungen (*joint intensity distributions* [39]) erfasst werden. Wurde eine solches Verteilungsmodell gelernt, können somit auch statistische Maße zur Messung der Ähnlichkeit zweier Häufigkeitsverteilungen, wie die Bhattacharyya- [56] oder Kullback-Leibler-Distanz [17] und davon abgeleitete Varianten als zu optimierende Kostenfunktionen herangezogen werden, denn die resultierende *joint distribution* sollte sich bei korrekter Registrierung der gelernten Verteilung stark ähneln.

Da mit den genannten statistischen und entropiebasierten Methoden Korrelationen zwischen einer größeren Anzahl von Merkmalen nur mit hochdimensionalen Verteilungsräumen und daher nicht effizient ausgenutzt werden können (*curse of dimensionality*), begrenzt sich deren Anwendung auf bild-basierte Registrierverfahren, sofern örtlich unkorrelierte Pixelintensitäten angenommen werden. Dagegen ermöglichen *metric learning* Ansätze [13, 65, 38, 28, 27, 46] das effiziente Lernen applikationsspezifischer Ähnlichkeiten auch hoher Dimensionalität auf robuste Art. Somit können neben ganzer Bildpatches für direkte Verfahren auch Merkmalsdeskriptoren keypoint-basierter Registrierverfahren über Modalitäten hinweg in eine semantisch sinnvolle Ähnlichkeitsbeziehung gesetzt werden. Die Metrik wird dabei diskriminativ über ähnliche und unähnliche Trainingsbeispiele gelernt. Als vielversprechende Ansätze dieser Art sollen in dieser Arbeit das *Similarity-Sensitive Hashing* [13] sowie *Random Forests for Metric Learning* [65] (siehe Kapitel 3.2 und Kapitel 3.3) im Hinblick auf multimodale keypoint-basierte Registrierung untersucht werden.

2.4 Prozesspipeline und Eingriffsmöglichkeiten

Als Grundlage für die im Rahmen dieser Arbeit durchgeführten Experimente steht ein Keypoint-basiertes Registrierframework zur Verfügung, welches im Folgenden schrittweise beschrieben werden soll. Abbildung 2.2 gibt einen visuellen Überblick über den Verarbeitungsprozess.

(1) Vorverabeitung

Zu Beginn werden die zu registrierenden Bilder eingelesen und in einem gemeinsamen *preprocessing*-Schritt verarbeitet. Je nach Konfiguration und Bildformat findet dabei eine Reduktion auf ein Graustufenbild mittels Farbraumumrechnung oder Selektion eines Farbkanals statt. Nach einer Reskalierung auf eine einheitliche Bildgröße werden zur

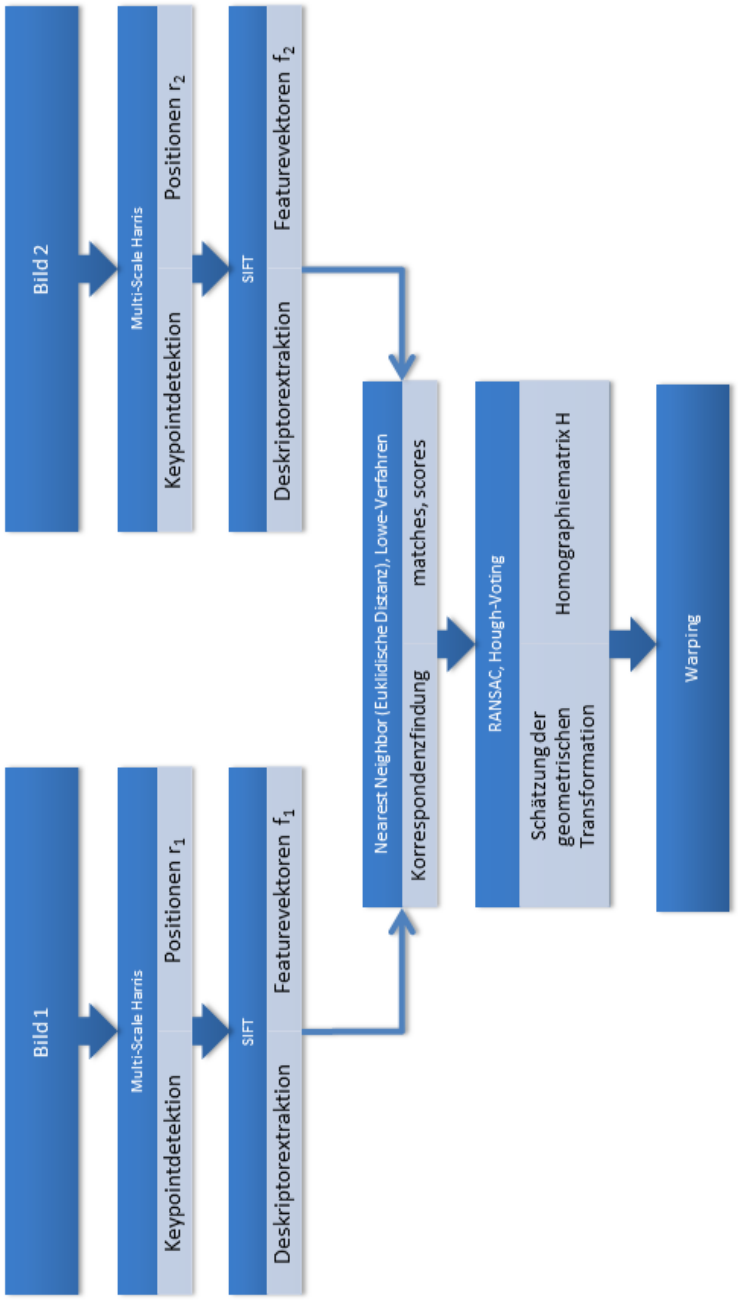


Abbildung 2.2: Visualisierung des keypoint-basierten Registrierverfahrens.

Rauschunterdrückung beide Bilder anschließend mit einen Gaußschen Tiefpass gefiltert. Zudem können weitere Filteroperationen, ein Helligkeitsausgleich (speziell bei Fundusbildern) und ein adaptives Histogrammausgleichsverfahren nachgeschaltet werden, um den Kontrast zu erhöhen und den verfügbaren Dynamikbereich möglichst gut auszunutzen. Das weitere Verfahren läuft ausschließlich auf den auf diese Weise vorverarbeiteten Bildern ab.

(2) Keypointdetektion

Im nächsten Schritt werden diejenigen Punkte in den Bildern gesucht, die einen hohen Wiedererkennungswert aufweisen (Keypoints) und daher für eine spätere Korrespondenzfindung geeignet sind. Hierfür sind vor allem Punkte in der Nachbarschaft von ausgeprägten 2D-Strukturen geeignet, welche mit geeigneten Eckendetektoren gefunden werden können. Für die durchgeführten Experimente wurde ein *Multi-Scale Harris*-Detektor [47] herangezogen, der über die Eigenwerte des Strukturtensores Informationen über die lokale Gradientenorientierung auf mehreren Skalen auswertet [59]. Eine grundlegende Annahme dieser Methode besteht darin, dass Umgebungen interessanter Punkte in beiden Modalitäten eine signifikante 2D-Struktur aufweisen und sich daher auch mit großer Wahrscheinlichkeit Keypoints an den korrespondierenden Orten beider Bilder finden lassen.

(3) Deskriptorextraktion

Um die auf diese Art gefundenen Keypoints bei der Korrespondenzfindung wiederzuerkennen, werden diese anschließend über eine Beschreibung ihrer jeweiligen Umgebungen im Bild in Form von Merkmalsdeskriptoren (Vektoren) charakterisiert. Als Merkmale dienen hierbei SIFT-ähnliche Histogramme von Gradientenorientierungen der Nachbarschaft auf der zugehörigen Keypoint-Skala auf Basis der *VLFeat*-Implementierung [60], welche eine Rotations- und Skalierungsinvarianz aufweisen und unabhängig von Rauschen, leichten Perspektivunterschieden und niederfrequenten Helligkeitsschwankungen sind [40]. Im Hinblick auf eine multimodale Registrierung kann optional auf die Berücksichtigung des Gradientenvorzeichens verzichtet werden [37], um den Fall einer Kontrastinversion abzudecken. Die extrahierten Deskriptoren haben eine Dimension von bis zu 512 Elementen (8×8 örtliche *bins* mit jeweils 8 Orientierungen).

(4) Korrespondenzfindung

Alle detektierten Keypoints lassen sich nun über ihren zugehörigen Merkmalsdeskriptor in Korrespondenz setzen, indem für jeden Keypoint der ersten Bildes die Ähnlichkeitsdistanz zu allen Keypoints bzw. Deskriptoren des zweiten Bildes ausgewertet und anschließend das Paar mit der kleinsten Distanz als vorläufige Zuordnung übernommen wird. Die Distanz selbst, zunächst in Form einer euklidischen Norm, gewichtet dabei die Relevanz der gefundenen Korrespondenz. Dabei kann es auch bei unimodalen Registrieraufgaben aufgrund von ähnlichen Bildbereichen, Störungen oder zu aussageschwachen Deskriptoren zu Falsch- oder Mehrfachzuordnungen kommen. Eine Möglichkeit dies zu

vermeiden besteht in dem Vorgehen nach Lowe [40], bei dem ein *match* nur übernommen wird, sofern der zweitbeste Treffer ein definiertes Abstandsverhältnis (*matching ratio*) zum Besten einhält. Dadurch werden unsichere bzw. mehrdeutige Korrespondenzen unterdrückt.

(5) Robuste Schätzung

Für die Registrierung wird ein starres, linear-affines Transformationsmodell (Translation, Rotation und Skalierung) angenommen. Da im vergangenen Schritt Ausreißer in Form falscher Korrespondenzen nicht sicher ausgeschlossen werden können, ist zur Schätzung der Transformation ein robustes Schätzverfahren wie RANSAC [24] oder eine Hough-Transformation [20] von Nöten. Beim Hough-Voting wird für alle Kombinationen von zwei Korrespondenzpaaren ein Modell initialisiert, welches für die zugehörige Parameterkonfiguration in einem 4D-Parameterraum eine Votierung vornimmt. Diese wird dabei über die Ähnlichkeitdistanz der beteiligten Korrespondenzen gewichtet. Über eine Parzenschätzung (*kernel density estimation*) lässt sich daraufhin das beste Modell ermitteln. Diese kann durch einen RANSAC-Algorithmus ergänzt oder alternativ ersetzt werden, welcher solange mit zufällig gewählten Korrespondenzen Modellkandidaten initialisiert bis möglichst mindestens ein ausreißerfreies Modell gefunden wurde, welches am besten zu allen korrekten Korrespondenzen passt (größtes *consensus set*).

(6) Iterative Verfeinerung

Das so geschätzte Transformationsmodell kann nun als Vorwissen für ein erneutes Matching dienen, indem mögliche Korrespondenzen, die zu stark von diesem Modell abweichen, übergangen werden. So können neue Korrespondenzpaare gefunden werden, die eine darauf folgende Schätzung nochmals verfeinert. Je nach Anwendungsfall kann das Ergebnis dadurch verbessert werden, indem diese Schritte iterativ wiederholt werden, wobei die erlaubte Abweichung schrittweise verkleinert wird.

(7) Warping

Im letzten Schritt kann das ermittelte Transformationsmodell herangezogen werden, um die beiden Bilder in das selbe Koordinatensystem zu bringen. Beim *Warping* wird eines der Bilder mittels Interpolation gemäß dem Modell umgerechnet, sodass das Resultat im Idealfall genau passend über das andere Bild gelegt werden kann. Damit ist der Registrierprozess abgeschlossen.

Im Hinblick auf die in Kapitel 2.3 diskutierten lern-basierten Erweiterungen bzw. Verbesserungen für multimodale Bilddaten bieten sich vor allem zwei der obigen Prozessschritte an: Zum einen kann der Vorverarbeitung ein Verfahren vor- oder nachgeschaltet werden, welches die beiden Bilder unterschiedlicher Modalitäten in eine einheitliche Repräsentation bringt (siehe Image Analogies, Kapitel 3.4). Zum anderen kann in das Matching eingegriffen werden, indem die verwendete euklidische Distanz als Ähnlichkeitsfunktion abgelöst und durch eine applikationsspezifische Metrik ersetzt wird. Dafür geeignete Methoden werden in den Kapiteln 3.1- 3.3 vorgestellt und daraufhin genauer untersucht.

Kapitel 3

Lern-basierte Ansätze

3.1 Wörterbuch-basierter Ansatz

3.1.1 Beschreibung

Die Verwendung eines Wörterbuches (*Descriptor Dictionary*, DD) stellt ein naheliegendes Verfahren dar, um Bilddaten unterschiedlicher Modalitäten miteinander in Beziehung zu setzen. Die Idee dabei ist, für jeden Merkmalsdeskriptor $f_1 \in F_1$ eines Bildauschnittes der ersten Modalität dessen korrespondierende „Übersetzung“ $f_2 \in F_2$ in die zweite Modalität zu lernen und direkt in Form von Paaren (f_1, f_2) abzuspeichern:

$$\text{DD} = \{(f_{1,1}, f_{2,1}), ..., (f_{1,K}, f_{2,K})\}$$

$$\begin{aligned} \text{mit} \quad F_1 &= [f_{1,1}, ..., f_{1,K}] \in \mathbb{R}^{m \times K} \\ F_2 &= [f_{2,1}, ..., f_{2,K}] \in \mathbb{R}^{m \times K} \end{aligned}$$

Dabei gibt K die Größe des Wörterbuches und m die Anzahl der Merkmale und damit die Dimension der Deskriptorvektoren f an. In dieser vereinfachten Notation bezeichnet F sowohl eine Menge von Merkmalsvektoren, als auch die zugehörige Matrix, wobei jeder in F enthaltenen Vektor eine Spalte darstellt. Die Auswahl der für die Erstellung eines Wörterbuches benötigten Trainingspaare wird in Kapitel 4 erläutert.

3.1.2 Matcher

Um die Keypoints bei der Registrierung in Korrespondenz zu setzen, wird für jeden Merkmalsvektor (Deskriptor) f_1 der ersten Modalität der nächste Nachbar (euklidische Distanz) $f_{1,NN}$ im gelernten Wörterbuch gesucht und dessen Übersetzung $f_{2,est}$ als Schätzwert für den Deskriptor der anderen Modalität herangezogen.

$$(f_{1,NN}, f_{2,est}) = \arg \min_{(f_{1,DD}, f_{2,DD}) \in \text{DD}} (f_1 - f_{1,DD})^2 \quad \forall f_1 \in F_1$$

Diese Schätzungen $f_{2,est} \in F_{2,est}$ können anschließend über einen konventionellen ℓ_2-Matcher den tatsächlich gefundenen Deskriptoren $f_2 \in F_2$ der zweiten Modalität zugeordnet werden.

$$\text{matching_index}(i) = \arg\min_j (F^i_{2,est} - F^j_2)^2 \quad \forall i$$

Dabei bezeichnet F^i den Featurevektor $f \in F$ am Spaltenindex i. Wie in Kapitel 2.4 beschrieben kann zusätzlich über ein *matchingRatio* gefordert werden, dass unsichere Zuordnungen zurückgewiesen werden.

3.1.3 Einordnung

3.1.3.1 Parameter und Komplexität

Die bestimmenden Parameter dieser Methode liegen neben der Wörterbuchgröße K vor allem darin, auf welche Art und Weise die (ausschließlich positiven, also zueinander ähnlichen) Trainingspaare ausgewählt werden (näheres dazu im folgenden Kapitel). Die Speicherkomplexität liegt dabei in $\mathcal{O}(K \cdot m)$.

Darüber hinaus kann die Nearest-Neighbor-Suche im Wörterbuch alternativ zur Brute-Force-Variante mit quadratischer Laufzeit $\mathcal{O}(N^2)$ über effizientere bzw. approximative Verfahren (k-d Tree [8], ANN [35, 49]) mit Zeitkomplexität von z.T. unter $\mathcal{O}(\log N)$ deutlich beschleunigt werden.

3.1.3.2 Bewertung

Die Performance der Wörterbuchmethode hängt maßgeblich von Güte, Anzahl und Repräsentativität der Trainingspaare ab. So ermöglichen wenige Einträge in hochdimensionalen Merkmalsräumen nur eine sehr schlechte Generalisierung, da die Annahme verletzt wird, dass der nächste Nachbar im Sinne der euklidischen Distanz eine ausreichend gute Repräsentation des tatsächlichen vorliegenden Deskriptors darstellt. Mit steigender Anzahl der Einträge steigt demnach tendenziell die Qualität der resultierenden Zuordnungsentscheidungen. Da Wörterbuchpaare stets fehlerbehaftet und verrauscht sein können, besteht allerdings die Möglichkeit, dass dieser Effekt in eine Sättigung läuft oder sogar eine Trendumkehr stattfindet, wenn zu viele Einträge solcher Art eingehen.

So zeigt sich, dass die schlechte Skalierbarkeit einen großen Nachteil des Verfahrens darstellt, da hierbei keine tiefergehenden Zusammenhänge zwischen den Modalitäten gelernt werden. Es ist dementsprechend schwierig vorauszusehen, wie viel Aufwand und Speicherplatz man einplanen muss um ein bestimmtes Maß an Generalisierung zu erreichen.

Die zusätzlich benötigte Rechenzeit im Vergleich zur reinen ℓ_2-Distanz ist aufgrund der effizienten Suche mittels approximativer Suchverfahren unerheblich.

3.1.3.3 Erweiterungen

Ausgehend vom Wörterbuchansatz sind verschiedene Erweiterungen denkbar, um Verbesserungen im Bezug auf Skalierbarkeit und Generalisierbarkeit zu erreichen.

Least-Square Schätzung einer linearen Transformation Eine einfache Möglichkeit, die „Lücken“ im Wörterbuch zu schließen (die Datenbasis zu glätten), besteht darin, die Deskriptoren der beiden Modalitäten (f_1, f_2) als Argument bzw. Bild einer linearen Funktion $T(f_1)$ zu betrachten.

$$T : \mathbb{R}^m \to \mathbb{R}^m, T(f_1) = A \cdot f_1 = f_{2,est}$$

Die Least-Square-Schätzung der Matrix A kann wie folgt über die Penrose-Inverse aus den vorhandenen Wörterbucheinträgen bestimmt werden:

$$A = (F_2 F_1^T)(F_1 F_1^T)^{-1} \in \mathbb{R}^{m \times m} \tag{3.1}$$

Dabei muss das Wörterbuch mindestens m Einträge besitzen, für eine robustere Schätzung deutlich mehr.

Die lineare Approximation ist nur solange sinnvoll, wie von einem dominanten linearen Zusammenhang zwischen den Deskriptorpaaren ausgegangen werden kann. Einfache Zusammenhänge wie Konstrastunschärfe (abgeschwächte Koeffizienten auf Hauptdiagonalen) oder Kontrastinversion (stückweise starke Koeffizienten auf Nebendiagonalen) auf SIFT-Deskriptoren sollte die Approximation erfassen können, sofern diese global gültig sind.

Clustering Ein Clustering von Merkmalsvektoren innerhalb des Wörterbuchs könnte dabei helfen, den Speicherplatzverbrauch zu reduzieren. Dabei ist die alleinige Speicherung der Clusterzentren und -varianzen nicht zielführend, da sich dadurch die „Lücken“ weiter vergrößern würden. Eine Verbesserungsmöglichkeit könnte darin liegen, jedem signifikanten Clusterzentrum (ausreichende Anzahl an *inliers*) eine aus den jeweiligen *inliers* geschätzte lineare Approximation nach Gleichung 3.1 zuzuweisen. Das Clustering selbst kann beispielsweise über *k-means* oder *multiRANSAC*-Modelle [68] erfolgen. Dabei ist es problematisch, wenn die Streuung der Daten im Merkmalsraum sehr groß ist, sodass nur sehr wenige Clusterzentren zustande kommen, aus deren Inliern eine ausreichend robuste Schätzung erfolgen kann.

Sparse Dictionary Auch die Anwendung von *sparse coding*-Methoden auf das gelernte Wörterbuch könnte wirkungsvoll sowohl Speicherplatzverbrauch als auch Generalisierbar-

keit verbessern. Dabei wird nicht jedes Trainingspaar einzeln abgespeichert, sondern es werden über ein Optimierungsverfahren für jede Modalität einige wenige Basisvektoren extrahiert, aus denen die Deskriptoren beider Modalitäten schließlich über Linearkombinationen davon rekonstruiert werden können [42, 48]. Die Komponenten der Zerlegung (*sparse decomposition*) des ersten Deskriptors f_1 dienen somit der Konstruktion der Schätzung $f_{2,est}$ mit Hilfe der Basisvektoren der zweiten Modalität. Zusätzlich kann mit Hilfe von Rauschmodellen eine höhere Robustheit des Verfahrens erreicht werden [14].

3.2 Similarity-Sensitive Hashing

3.2.1 Beschreibung

Das *Similarity-Sensitive Hashing* (SSH) nach Bronstein et al. [13] ist ein Verfahren zum Erlernen einer Ähnlichkeitsmetrik auf Basis von binären Klassifikatoren (ähnlich/unähnlich). Es stellt eine Erweiterung des Locality-Sensitive Hashing (LSH) [30] und der darauf aufbauenden Approximate-Nearest-Neighbor Verfahren [35] dar und ist inbesondere auch für hochdimensionale Daten geeignet. Die Erweiterung besteht darin, mit der gelernten Metrik keine ℓ_p-Distanzen zwischen Datenpunkten abzubilden, sondern ein neues Ähnlichkeitsmaß für Daten verschiedener Modalitäten zu lernen, die ggf. auch durch unterschiedliche Deskriptorarten repräsentiert sein können. Erste Ansätze für dieses Verfahren erarbeiteten Shakhnarovich et al. [53, 54].

Das Similarity-Sensitive Hashing berechnet in Abhängigkeit der Trainingsdaten Projektionen in binäre Codes so, dass die resultierenden Hamming-Distanzen die Ähnlichkeit zwischen den Trainingspaaren widerspiegeln. Das bedeutet die Kollisionswahrscheinlichkeit der beiden Hashwerte eines Deskriptorpaares entspricht der (gelernten) Ähnlichkeit der zugehörigen Daten und verhält sich damit invers zur ihrer Hamming-Distanz [13].

Ziel des Verfahrens ist für jedes Bit je eine Projektion für beide Modalitäten zu lernen, sodass „ähnlich“-gelabelte Trainingspaare möglichst auf das gleiche, und „unähnlich“-gelabelte Trainingspaare auf unterschiedliche Vorzeichen abgebildet werden. Dies ermöglicht die Verwendung von Deskriptoren zweierlei Arten und Dimensionen. Darüber hinaus findet durch die Projektionen eine signifikante Dimensionsreduktion statt, was vor allem der Effizienz des Matchers zu Gute kommen kann.

Die einzelnen Projektionsfunktionen für jedes Bit sind das Ergebnis der Optimierung einer entsprechend gewählten Kostenfunktion, welche die Anzahl der korrekt klassifizierten Trainingspaare und damit direkt die Ähnlichkeit-Sensitivität maximiert. Das Optimierungsproblem wird dabei effizient über einen Boosting-ähnlichen Ansatz [25, 52] gelöst. Dazu werden allen K Trainingspaaren (x_k, y_k) Gewichte $w_i(k)$ zugeordnet, die initial (i=1) gleichverteilt sind. In jeder Iteration i (i-tes Bit) wird die Projektionsfunktion $h_i(x, y)$ gesucht, welche die folgende Kostenfunktion maximiert:

$$r_i = \sum_{k=1}^{K} w_i(k) s_k h_i(x_k, y_k) \in [-1, 1] \tag{3.2}$$

$s_k \in \{-1, 1\}$ gibt dabei an, ob das Trainingspaar als ähnlich ($s_k = 1$) oder unähnlich ($s_k = -1$) gelabelt ist. r_i wird nimmt den Wert 1 an, wenn durch die Abbildung $h_i(x, y)$ alle Trainingspaare richtig klassifiziert werden (perfekte Trennung) und wird kleiner, je mehr Paare (mit hohen Gewichten) falsch abgebildet werden. Gemäß dem Boosting-Ansatz werden nach jeder Iteration die Gewichte entsprechend dem Klassifikationsergebnis adaptiert, sodass falsch klassifizierte Trainingspaare im nächsten Schritt ein größeres Gewicht $w_{i+1}(k)$ bekommen.

Als Projektions- und damit auch als Trennfunktionen für die Klassifikation werden in dieser Arbeit ausschließlich *lineare* Abbildungen verwendet. Es folgt damit [13]:

$$\begin{aligned} f_i(x) &= p_i^T x + a_i & f &: \mathbb{R}^m \to \mathbb{R} \\ g_i(y) &= q_i^T y + b_i & g &: \mathbb{R}^{m'} \to \mathbb{R} \end{aligned} \tag{3.3}$$

$$\begin{aligned} h_i(x,y) &= sign(f_i(x)) \cdot sign(g_i(y)) \\ &= sign(p_i^T x + a_i) \cdot sign(q_i^T y + b_i) \qquad h_i : \mathbb{R}^m \times \mathbb{R}^{m'} \to \{-1, 1\} \end{aligned} \tag{3.4}$$

$$r_i = \sum_{k=1}^{K} w_i(k) s_k sign(p_i^T x_k + a_i) sign(q_i^T y_k + b_i) \to \text{max.} \qquad r_i \in [-1, 1] \tag{3.5}$$

Bronstein et al. [13] lösen dieses nichtlineare Optimierungsproblem approximativ über das Ausnutzen von Vorzeichenkorrelationen mit Hilfe der Summe von gewichteten Kovarianzmatrizen C der mittelwertfreien Deskriptoren aller Trainingspaare:

$$C = \sum_{k=1}^{K} w_i(k) s_k \bar{x}_k \bar{y}_k^T \in \mathbb{R}^{m \times m'} \tag{3.6}$$

Dazu wird die Singulärwertzerlegung (SVD) von C herangezogen, um die approximative Kostenfunktion $\hat{r}_i = p_i^T C q_i$ über die linken und rechten Singulärvektoren mit den größten Singulärwerten zu maximieren.

Zur approximativen Suche des Optimums von r_i werden N Paare zufälliger Linearkombinationen der linken Singulärvektoren $p_{i,cand}$ und der rechten Singulärvektoren $q_{i,cand}$ gebildet. Anschließend werden für jeden dieser Kandidaten die offsetfreien Projektionen 3.3 der Merkmalsvektoren $x_{k,proj} = p_{i,cand}^T x_k$ sowie $y_{k,proj} = q_{i,cand}^T y_k$ berechnet. Die fehlenden beiden (optimalen) Offsetparameter (a_i, b_i) können dann über eine zweidimensionale Rastersuche auf einem $B \times B$-Gitter, ggf. auf mehreren Skalen, bestimmt werden. Abschließend wird für jeden der N Kandidaten die (exakte) Kostenfunktion 3.5 ausgewertet und derjenige mit dem höchsten Wert als Projektionsfunktion $h_i(x, y)$ 3.4 für diese Iteration übernommen.

Der Wert von r_i stellt dabei ein Maß für die Güte der Projektion dar, aus dem ein Gewicht $\alpha_i = \frac{1}{2} \log(\frac{1+r_i}{1-r_i})$ für das jeweilige Bit abgeleitet werden kann. Somit eignet sich vor allem für ungleich verteilte Gewichte die *gewichtete Hammingdistanz* als Ähnlichkeitsmaß.

Der Boosting-Ansatz sorgt dafür, dass viele einfache lineare Trennfunktionen (weak learner), die allein nicht in der Lage sind, die gesamte Trainingsmenge zu separieren, zu einem starken Klassifikator (strong learner) kombiniert werden [25, 52]. Dies führt im Laufe der Iterationen zu Oszillationen in den Gewichten aufgrund von Untermengen nicht linear trennbarer

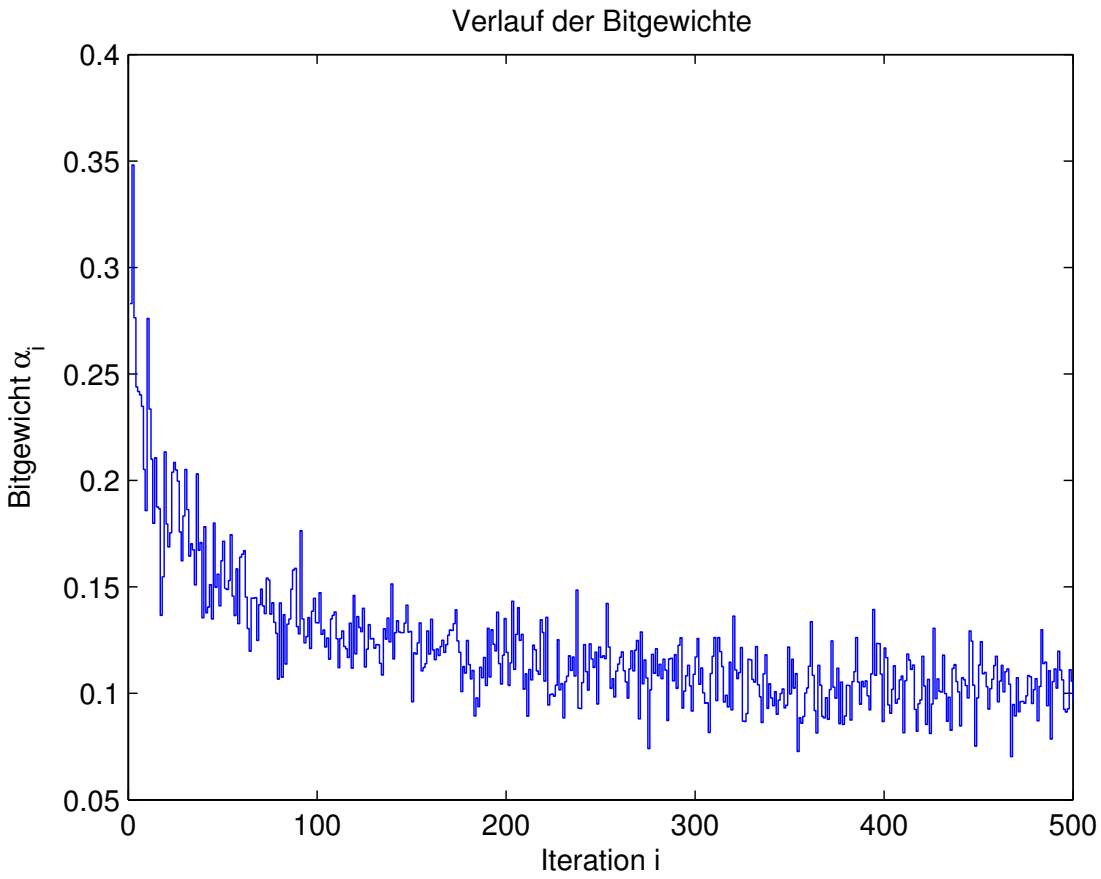


Abbildung 3.1: Beispiel für den Verlauf der zugewiesenen Bitgewichte beim Similarity-Sensitive Hashing. Die Kurve zeigt in etwa eine exponentielle Abnahme bei starken Oszillationen.

Trainingspaare. Da sich irgendwann ein Ungleichgewicht zu Gunsten der schwer trennbaren Paare einstellt, und diese in einer einzelnen Iteration nicht alle korrekt klassifiziert werden können, nehmen der Wert der Kostenfunktion r_i und damit auch die Bitgewichte α_i tendenziell mit steigender Bitanzahl ab. Bei Konvergenz des Verfahrens werden die Gewichte kleiner und gehen gegen 0. Abbildung 3.1 zeigt einen beispielhaften Verlauf der vergebenen Gewichte.

3.2.2 Matcher

Der Aufbau eines Matchers für Merkmalsvektoren auf Basis der gelernten SSH-Distanz gestaltet sich analog zur euklidischen Variante.

Es wird lediglich im Vorfeld das *Hashing*, also die Berechnung der gelernten Projektionen beider Modalitäten auf den binären Hamming-Raum $\mathbb{R}^m \rightarrow \mathbb{H}^n$ für die Deskriptoren aller gefundener N_{KP} Keypoints durchgeführt.

Werden N Deskriptoren $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{m \times N}$ und n lineare Projektionen $F = [p_1, p_2, ..., p_n]^T \in \mathbb{R}^{n \times m}$ als Matrizen aufgefasst, so kann das Hashing über eine Matrixmulti-

plikation mit anschließender Vorzeichenbestimmung via Indikatorfunktion $\mathbf{1}(\bullet)$ beschrieben und berechnet werden:

$$X_{proj} = \mathbf{1}_{(0,+\infty)}(F \cdot X) = [x_{1,proj}, ..., x_{N,proj}] \in \mathbb{H}^{n \times N}$$

Der Offset a_i kann dabei an den Projektionsvektor p_i angehängt werden, wenn an das Ende der Matrix X eine Zeile aus Einsen hinzugefügt wird. Analog erfolgt die Berechnung für die Deskriptoren y der zweiten Modalität.

Anschließend können $\mathcal{O}(N_{KP}^2)$ Hammingdistanzen über schnelle Bitoperationen ausgewertet und die jeweiligen Minima bestimmt werden. Die Distanzen pro Bit können dabei mit den gelernten Gewichten α_i gewichtet werden.

Aus Einfachheitsgründen wird für Experimente im Rahmen dieser Arbeit die gewichtete Hammingdistanz durch eine ungewichtete angenähert, indem jedes Bit i proportional zu seinem Gewicht α_i wiederholt wird. So kann auf vorhandene Hamming-Matcher (z.B. *DescriptorMatcher* aus *OpenCV* [11]) zurückgegriffen werden. Abbildung 3.2 zeigt, dass der dabei entstehende Approximationsfehler vernachlässigbar klein ist.

3.2.3 Einordnung

3.2.3.1 Parameter und Komplexität

Viele der bereits im vorherigen Abschnitt 3.2.1 erwähnten und in Tabelle 3.1 zusammengefassten Parameter nehmen vor allem Einfluss auf die Optimierung der Kostenfunktion r_i.

Da $\hat{r}_i$ nur eine Näherung hiervon darstellt, maximiert der höchstwertigste Singulärvektor nicht zwingend auch r_i. Es ist daher naheliegend, verschiedene Linearkombinationen mehrerer Singulärvektoren zu testen um sich damit dem Maximum möglichst gut anzunähern. M regelt dabei die Auswahl der Singulärvektoren, wohingegen durch N die Anzahl der zu probierenden zufälligen Linearkombinationen bestimmt wird. Mit steigendem M wird die Optimierung robuster, es tritt allerdings schon bei niedrigen Werten ein Sättigungseffekt ein [13]. Die Wahl von $M \ll m, m'$ ($M \approx 5$) führt dabei zu einer starken Reduktion der Anzahl von Suchdimensionen. Wird N erhöht, erreicht die Kostenfunktion im Erwartungswert höhere und damit bessere Werte; der Zeitaufwand für den Optimierungsprozess nimmt allerdings mit $\mathcal{O}(N)$ entsprechend zu.

Ähnlich verhält es sich erwartungsgemäß bei der Erhöhung der Anzahl von Skalen und von Testpunkten zur Optimierung der Offsetparameter (a_i, b_i) auf dem zweidimensionalen Gitter. Hierbei tragen die Testpunktanzahl und die Anzahl der Skalen mit $\mathcal{O}(B^2 \cdot \text{SCALES})$ zur Zeitkomplexität bei.

An dieser Stelle gilt es zwischen Optimierungsgüte und Rechenzeit abzuwägen, wobei das Boosting einen Ausgleich von schlechterer Optimierungsgüte über eine größere Anzahl an Bits ermöglicht. Letzteres zeigte sich in den Experimenten als guter Kompromiss.

In jeder Iteration i werden für jeden der N Kandidaten SCALES $\cdot B^2$ Kostenfunktionen r_i ausgewertet, wobei dies jeweils die Berechnung von Projektionen aller K Samples beinhal-

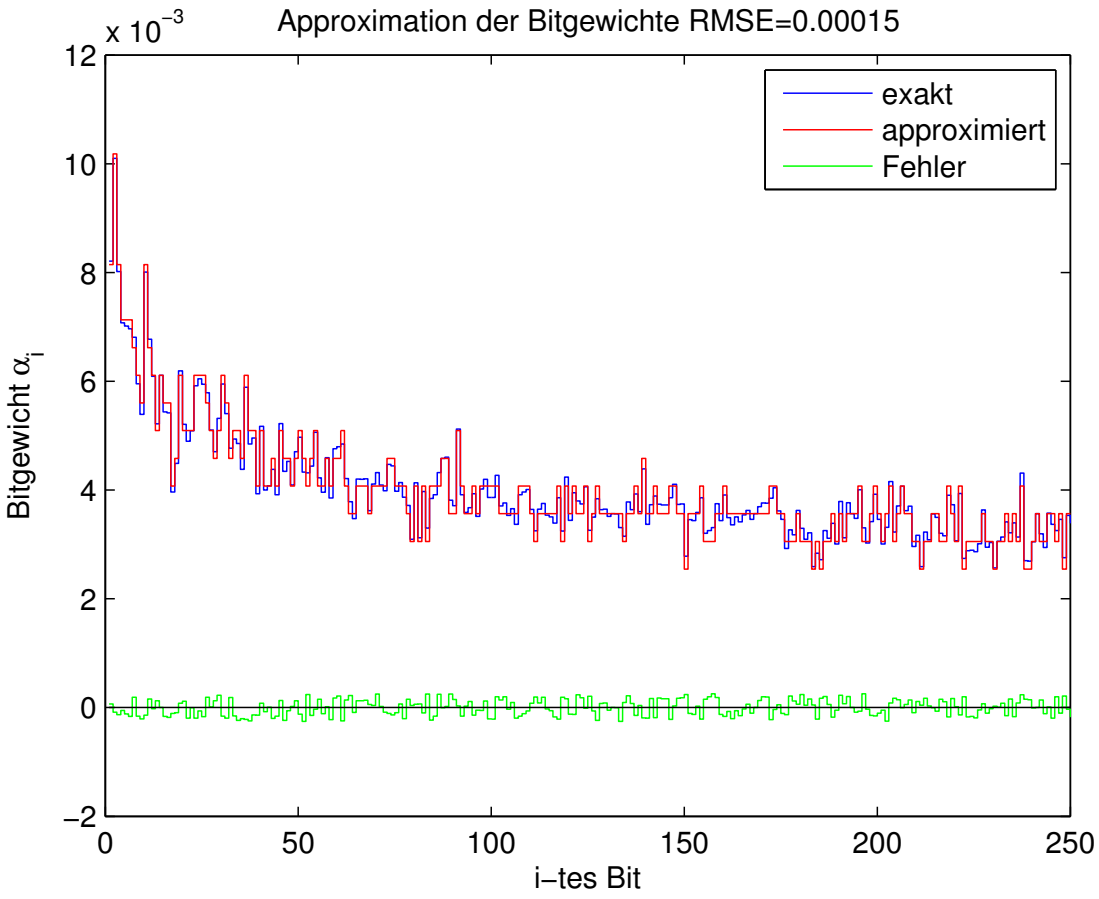


Abbildung 3.2: Visualisierung des Approximationsfehlers bei der Annäherung der Bitgewichte (blau) durch proportionale Wiederholungen von gleichgewichteten Bits(rot). Die Gewichte sind dabei auf 1 normiert. Die rote Kurve entspricht der (ebenfalls normierten) äquivalenten Anzahl an Bits. Die Differenz der beiden Kurven zeigt den Approximationsfehler (grün). Aus Gründen der Übersichtlichkeit sind die Kurven nur über den ersten 250 Bits aufgetragen. Der aufsummierte Fehler im Verhältnis zum normierten Gesamtgewicht 1 liegt hierbei für 2500 Bit bei unter 4%.

Tabelle 3.1: Parameter beim Similarity-Sensitive Hashing

Parameter	Beschreibung	Wert
n	Anzahl Iterationen (Bits)	$500 - 2500$
M	Anzahl ausgewählter Singulärvektoren	5
N	Anzahl zufälliger Linearkombinationen	8
B^2	Anzahl Gitterpunkte (a, b)-Optimierung	32^2
SCALES	Anzahl Skalen (a, b)-Optimierung	2
weightThres	Gewichtschwellwert für approx. Bestimmung von C	0.8
K	Anzahl Trainingspaare	$10^4 - 10^6$
m	Anzahl Elemente der Deskriptorvektoren x	512
m'	Anzahl Elemente der Deskriptorvektoren y	512

tet (siehe Formel 3.5). Dies kann mit Hilfe von Parallelisierung der N Auswertungen und Integralbild-Techniken [13] bei der Optimierung von (a_i, b_i) beschleunigt werden.

Als sehr zeitintensiv stellte sich die Berechnung der gewichtet aufsummierten Kovarianzmatrizen C in Verbindung mit großen Merkmalsvektoren ($m \approx m'$ Elemente) heraus, da diese sich mit $\mathcal{O}(m^2)$ auf die Rechenzeit pro Iteration auswirken. Deshalb wurde C in allen Experimenten näherungsweise berechnet, indem nur ein gewisser Prozentsatz der K Trainingspaare mit den größten Gewichten zu C beitragen (*weight trimming* [26]). Der Anteil wird durch den Parameter *weightThres* festgelegt. Der Unterschied zur exakten Berechnung von C wird in Kapitel 5.3.2 experimentell untersucht.

Wie bei Boosting-Algorithmen üblich, trägt die Anzahl der Iterationen (Anzahl der Bits) n und die Anzahl der Trainingspaare K linear mit $\mathcal{O}(Kn)$ zum Zeitaufwand bei, wenn eine der beiden Größen konstant bleibt. Die Anzahl der zu einer ausreichend diskriminativen Trennung der Trainingspaare benötigten Iterationen hängt stark vom Anwendungsfall und den Trainingsdaten selbst ab.

Beim Training geht das Verfahren sehr speichereffizient vor. Es kann direkt auf den integerwertigen Eingabe-Trainingsdaten gearbeitet werden; zusätzlicher Speicher wird abgesehen von der Ausgabe effektiv nur für $1 \times K$-Gewichte, die $m \times m'$ Kovarianzmatrix C, sowie für das B^2-Feld der Rastersuche benötigt.

Für die im Rahmen dieser Arbeit verwendeten Trainingsdaten und Bit-Anzahlen (siehe Werte in Tabelle 3.1) sind Trainingszeiten in der Größenordnung von 10 Stunden auf dem verwendeten Testrechner (siehe Kapitel 5) nötig.

3.2.3.2 Bewertung

Das Similarity-Sensitive Hashing beruht auf einem diskriminativen Lernverfahren, welches Klassifikatoren auf der Statistik von Merkmalsvektoren bimodaler Datenpaare trainiert. Dies führt im Gegensatz zum nicht parametrisch gelernten Wörterbuch dazu, dass der statistische Zusammenhang zwischen den Trainingsdaten gelernt wird, und das Verfahren damit -

ausreichend viele hochwertige und repräsentative Daten vorausgesetzt - besser generalisiert. Die gelernten Projektionen bewerten dabei implizit über die jeweiligen Koeffizienten, welche Elemente der Merkmalsvektoren für Ähnlichkeit relevant sind. Die Experimente zeigten, dass hierbei Güte und Anzahl der verfügbaren Trainingsdaten einen großen Einfluss haben. Dies wird in Kapitel 4 und 5 genauer untersucht und beschrieben.

Ein Vorteil des SSH-Verfahrens beseht darin, dass Daten nicht nur aus verschiedenen Modalitäten, sondern auch in unterschiedlichen Repräsentationen in Beziehung gesetzt werden können.

Die hohe Speichereffizienz macht das Verfahren auch für große Datenmengen mit hochdimensionalen Merkmalsvektoren anwendbar. Eher schlägt die für das Training benötigte Zeit zu Buche, welche jedoch durch eine angemessene Wahl der Optimierungsparameter im Rahmen gehalten werden kann.

Dem paarweisen Matching von Merkmalsvektoren, wie es im Falle der Registrierung von Nöten ist, geht die Auswertung der gelernten Projektionsfunktionen voraus. Da diese in einer einzigen Matrixmultiplikation je Modalität umgesetzt werden kann und die Hammingdistanzen anschließend sehr effizient berechnet werden können, gibt es gegenüber dem ℓ_2-Matcher im Hinblick auf den Berechnungsaufwand in der Praxis keine Nachteile.

3.3 Random Forests for Metric Learning

3.3.1 Beschreibung

Einen etwas anderen Ansatz zum Erlernen eines applikationsspezifischen Ähnlichkeitsmaßes beschreiben Xiong et al. in *Random Forests for Metric Learning with Implicit Pairwise Position Dependence* [65]. Auch hier wird das *metric learning*-Problem auf ein paarweises Klassifikationsproblem zurückgeführt, jedoch ist der Zusammenhang mit klassischen Metriken direkter, da dem Training des Klassifikators unter anderem eine ℓ_1-Distanz zwischen Merkmalsvektoren zugrunde liegt (siehe Gleichung 3.7). Die genaue Herangehensweise wird im Folgendem beschrieben.

Ist eine heterogene Datenmenge mit nichtlinearen Ähnlichkeitsrelationen (in Form von positiven oder negativen Labels als *constraints* bzw. Randbedingungen für die Trainingspaare) gegeben, so steht man prinzipiell vor der Entscheidung, ob über dem Merkmalsraum eine einzelne, globale Metrik oder mehrere lokale Metriken trainiert werden sollen [7, 3].

Eine globale Metrik sollte dabei über dem gesamten Merkmalsraum alle vorgegebenen Randbedingungen der Trainingsmenge erfüllen. Dies kann unter anderem durch Transformation der Daten in einen einfacher zu diskriminierenden Unterraum (bspw. RCA [5, 6], PCCA [46]) bzw. das Erlernen einer optimal parametrierten Mahalanobis-Distanz (ITML [19], LMNN [55] und weitere) erfolgen. Da sich die genannten Verfahren größtenteils nur in der Optimierungsmethode unterscheiden, bleibt das Grundproblem bestehen, dass eine globale Metrik je nach Form der zugrundeliegenden Daten nicht alle *constraints* gleichzeitig erfüllen kann. Dafür skaliert eine Solche sehr gut mit der Anzahl der Daten und ist effizient zu berechnen und zu speichern.

Mehrere lokale Metriken, sog. *multi-metric*-Ansätze erzielen dagegen auch bei starken Nichtlinearitäten eine angemessene Abbildung der realen Ähnlichkeitsverhältnisse. Alle von der Trainingsmenge geforderten *constraints* sollen eingehalten werden, indem eine größere Anzahl an Metriken, bspw. für Untermengen des paarweisen Merkmalraums oder im Extremfall für jedes einzelne Trainingspaar optimiert wird [28, 27]. Obwohl die resultierende Metrik sich dadurch verbessert, sind die meisten Verfahren auf dieser Basis für große und hochdimensionale Daten nicht geeignet, da sie eine hohe Zeit- und Speicherplatzkomplexität aufweisen [65].

Im Idealfall möchte man demnach einen Kompromiss aus beiden Varianten haben, um die jeweiligen Nachteile möglichst zu kompensieren [3]. Mit der *Random Forest Distance* (RFD) schlagen Xiong et al. [65] eine solche Methode vor, die Eigenschaften beider Ansätze vereint und optimal ausnutzt. Die Idee besteht darin, nicht nur den relativen Abstand eines Merkmalpaares, sondern auch deren absolute Position (Mittelwert) im Merkmalsraum in die Ähnlichkeitsbetrachtung mit einzubeziehen. Dies ermöglicht eine entsprechende Abbildungsfunktion $\phi(x_i, x_j)$.

$$\phi(x_i, x_j) = \begin{bmatrix} |x_i - x_j| \\ \frac{1}{2}(x_i + x_j) \end{bmatrix}, \phi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{2m} \tag{3.7}$$

Durch Einbettung in einen neuen *paarweisen Merkmalsraum* mit der doppelten Anzahl an Dimensionen passt sich eine darauf aufbauende Metrik implizit den zugrundeliegenden Daten an.

Xiong et al. zeigen, dass die Anwendung von Mahalanobismetriken auf diesem neuen Merkmalsraum einige theoretische Nachteile mit sich bringt und daher nicht sinnvoll ist [65]. Daher interpretierten sie das Anpassen eines Ähnlichkeitsmaßes in ein Klassifikationsproblem um und wählten für die neue Metrik *Random Forests* [12] als zugrundeliegenden Klassifikator. Diese weisen im Vergleich zu anderen Klassifikationsverfahren eine hohe Leistungsfähigkeit auf und kommen gut mit Daten hoher Dimensionalität und nichtlinearen Entscheidungsgrenzen klar [16, 12].

Die Idee hinter Random Forests besteht darin, eine Menge (*forest*) von Entscheidungsbäumen auf den gelabelten Trainingsdaten zu lernen, die dazu in der Lage sind, Deskriptorpaare als „ähnlich“ oder „unähnlich“ zu klassifizieren. Dabei dient die Anzahl aller Bäume dieser Menge, die das Label „unähnlich“ prädizieren, als Distanzmaß. Das Training und die Klassifikation finden auf dem mit Gleichung 3.7 gebildeten Merkmalsraum statt.

Insgesamt werden T voneinander unabhängige Entscheidungsbäume trainiert. An jeder Verzweigung (*split*) eines solchen Baumes werden $\sqrt{m}$ zufällige Merkmale gezogen, anhand derer die optimale Trennfunktion in Bezug auf die verbliebene Untermenge aller Trainingsdaten bestimmt wird. Als Kostenfunktion eignen sich dabei Entropie- oder andere statistische Maße. Dieses Vorgehen wird rekursiv wiederholt, bis alle verbliebenen Trainingsbeispiele eines Knotens zur selben Klasse gehören, oder keine weitere Aufteilung möglich ist. An jedem Blatt steht demnach die Entscheidung für eine Klasse fest [18].

Im Hinblick auf die *metric learning*-Problemstellung bedeutet dies, dass bei jeder Entscheidung nicht nur paarweise Unterschiede in den ursprünglichen Merkmalen eingehen, sondern auch über die absolute Position im Merkmalsraum getrennt werden kann. Der durch den jeweiligen Teilbaum entstehende Einfluss auf die Gesamtmetrik ist daher nur in dem entsprechenden Halbraum gültig. Die Metrik passt sich damit der lokalen Struktur der Daten an und ist somit in der Lage, auch heterogene Ähnlichkeitsrelationen zu reflektieren [65].

3.3.2 Matcher

Ein Keypoint-Matcher auf Basis der *Random Forest Distance* unterscheidet sich dahingehend von den bisherigen Ansätzen, als dass keinerlei Vorverarbeitung auf den *einzelnen* Featuredeskriptoren möglich ist, sondern über die Einbettungsfunktion 3.7 immer auf *Deskriptorpaaren* gearbeitet und klassifiziert werden muss.

Dies hat zur Folge, dass bei einem Brute-Force Matching (ohne weiteres Vorwissen) für jedes mögliche Kandidatenpaar die Abbildung 3.7 berechnet und der Random Forest ausgewertet

Tabelle 3.2: Parameter beim Random Forest Training

Parameter	Beschreibung	Wert
T	Anzahl Entscheidungsbäume im Random Forest	$>= 100$
$\phi(x_i, x_j)$	Einbettungsfunktion (mit/ohne Absolutposition)	Gleichung 3.7
K	Anzahl Trainingspaare	> 10000
m	Anzahl Elemente der Deskriptorvektoren x und y	512

werden muss, um die gewünschte Ähnlichkeitsdistanz über die Stimmen der jeweiligen Bäume zu erhalten. Erst dann kann für jeden Keypoint der beste Match im Sinne der kleinsten RF-Distanz (d.h. des Klassifikationsergebnisses) herausgesucht werden.

3.3.3 Einordnung

3.3.3.1 Parameter und Komplexität

Die zur Klassifikation benötigten Entscheidungsbäume des Random Forest wurden wie oben beschrieben fest konfiguriert. Interessante Parameter darüber hinaus sind zum einen die Anzahl T der Bäume und zum anderen die Art der Einbettungsfunktion $\phi(x_i, x_j)$.

Sowohl beim Training als auch bei der Auswertung hängt die Zeitkomplexität linear von T ab. Dabei sollte sich mit steigendem T die Trennbarkeit der Daten und damit die Genauigkeit der Metrik verbessern. Wird T zu groß, könnten sich *over-train*-Effekte zeigen.

Nachteilig im Bezug auf die Speicherkomplexität wirkt sich die Einbettung der Daten aus, da hierfür Speicher doppelt so großer Dimensionalität im Vergleich zu den Trainingsdaten angelegt werden muss. Da sich bei Daten großer Homogenität und Linearität die Einbringung der Absolutposition als kontraproduktiv erweisen kann [65], könnte auch die Untersuchung einer alternativen Einbettungsfunktion von Interesse sein, bei der lediglich der Differenzbetrag der Deskriptorpaare eingeht.

Das Training eines einzelnen Entscheidungsbaums trägt dabei mit $\mathcal{O}(K \log(K))$ zum Berechnungsaufwand bei [65]. Hierbei bietet sich eine Parallelisierung sowohl auf *forest*- als auch auf Baumebene an, da Bäume und Teilbäume unabhängig voneinander aufgestellt werden können. Insgesamt verhält sich das Verfahren im Training zeiteffizient.

Aufgrund der oben diskutierten Eigenheiten des Matchers müssen zur Laufzeit pro Registrierungsvorgang $\mathcal{O}(N_{KP}^2 \cdot T)$ Entscheidungsbäume ausgewertet werden, wobei N_{KP} die Anzahl der detektierten Keypoints bezeichnet.

3.3.3.2 Bewertung

Obwohl sich die *Random Forest Distance* in ihrer Herangehensweise vom Similarity-Sensitive Hashing unterscheidet, lässt sich dennoch feststellen, dass die Diskriminanz auf den Daten ebenfalls durch Klassifikation anhand einer Serie linearer Entscheidungsfunktio-

nen erreicht wird. Daher gelten alle Einschätzungen bezüglich der Generalisierungsfähigkeiten und Anforderungen an die Trainingsdaten aus Kapitel 3.2.3.2 auch für die RFD.

Ein entscheidender Nachteil bei der Verwendung von RFD zur Korrespondenzfindung ergibt sich aus der oben angesprochenen Problematik, dass jedes mögliche Deskriptorpaar die Auswertung durchlaufen muss. Lässt man bei angenommen $N_{KP} = 3000$ für jeden Keypoint die RFD zu allen Kandidaten (sequentiell) auswerten, beträgt die Berechnungszeit bei $T = 100$ Bäumen auf dem verwendeten Testrechner (siehe Kapitel 5) etwa zwei Stunden pro Bild.

Diese Eigenschaft macht nicht nur das Verfahren für diese Anwendung unpraktikabel, sondern auch Experimente sehr zeitaufwändig. Im Zuge dessen wurden testweise verschiedene Random Forest Implementierungen (u.a. von Breiman und Cutler, Treebagger in Matlab) herangezogen, diese zeigten allerdings dasselbe ungünstige Verhalten. Daher erscheint es nicht sinnvoll, RFD über einige wenige Tauglichkeitstests hinaus zu untersuchen.

3.4 Image Analogies

3.4.1 Beschreibung

Die grundlegende Schwierigkeit bei der Registrierung multimodaler Bilddaten besteht darin, dass dieselben abgebildeten Strukturen ein unterschiedliches Aussehen aufweisen können. Wäre es möglich, eine applikationsspezifische Filteroperation zu erlernen, welche auf robuste Art zwischen den charakteristischen Erscheinungsbildern der beteiligten Modalitäten abbildet, so könnte damit ein konventionelles unimodales Registrierverfahren im Bezug auf multimodale Daten erweitert bzw. unterstützt werden. Dazu wird das Bild der ersten Modalität gefiltert und damit in die Darstellungsart der Zweiten überführt, wodurch gemeinsam vorhandene Strukturen und Hintergründe eine hohe Ähnlichkeit aufweisen. Über eine unimodale Registrierung zwischen gefiltertem Bild und dem der zweiten Modalität kann auf diese Weise anschließend die Beziehung zwischen den Originalbildern ermittelt werden.

Vielversprechenden Möglichkeiten für das Erlernen von Filtern bzw. Bildtransformationen dieser Art bieten zum Beispiel Image Analogies (IA) [33]. Diese Methode vereint performante Verfahren aus dem Bereichen der Textursynthese [63, 2] und des Texturtransfers [21] und verallgemeinert diese auf Bildpaare. Dies bedeutet, dass Analogien zwischen einem Trainings-Bildpaar (A, A') auf ein neues Bild B übertragen werden können. Die Erfassung solcher Analogien zwischen dem Trainings-Quellbild A und Trainings-Filterbild A' ermöglicht ein einfaches Markov-Netzwerk (MRF) auf Basis von Pixelintensitäten in einer Nachbarschaft auf mehreren Auflösungsstufen. Es werden damit direkte Zusammenhänge zwischen Gruppen von Intensitätswerten vom Quellbildpaar auf das Zielbildpaar gelernt. Bei der Synthese des gefilterten Bilds B' kommt ein autoregressiver Algorithmus zum Einsatz, welcher pixel- und zeilenweise Nachbarschaften aus B in A sucht und die jeweils am besten passende korrespondierende Pixelintensität aus A' einbaut. Um Kohärenz im Zielbild B' zu gewährleisten werden dabei auch bereits synthetisierte Pixel über eine kausale (L-förmige) Nachbarschaftsmaske in B' berücksichtigt. Wird das Verfahren auf mehreren Ebenen einer

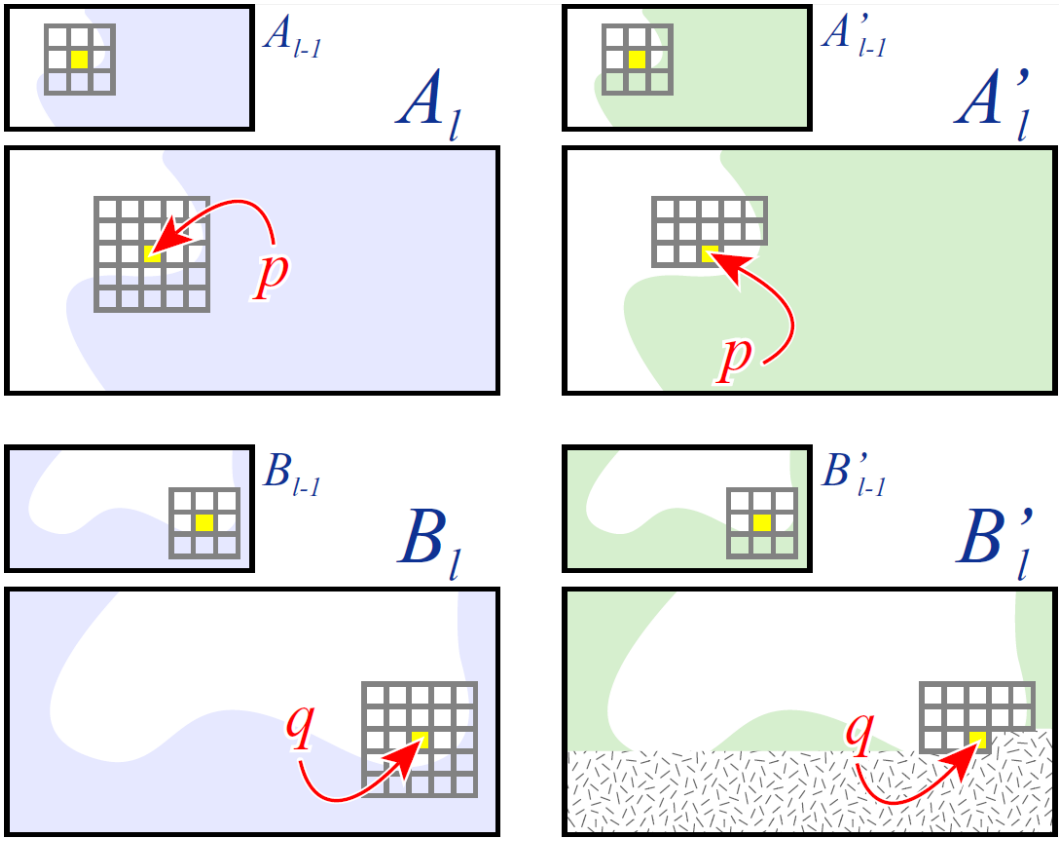


Abbildung 3.3: Zur Synthese des Pixels q in B' werden die zugehörigen Nachbarschaften in B_l, B'_l und deren reduzierten Versionen B_{l-1} und B'_{l-1} herangezogen, um den besten Treffer zu den Nachbarschaften des korrespondierenden Pixels p in den Trainingsbildern A zu finden [33].

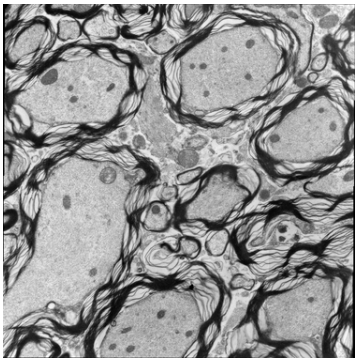
Auflösungspyramide durchgeführt, können anfänglich grobe, großflächige Bildelemente und anschließend iterativ immer detailliertere und höher aufgelöste Strukturen erfasst und synthetisiert werden ohne auf zu große und rechenaufwändige Nachbarschaften zurückgreifen zu müssen. Das pixelweise Vorgehen ist in Abbildung 3.3 illustriert. Zusätzlich ist eine spezielle Behandlung an Rändern und der niedrigsten Auflösungsstufe von Nöten, damit stets auf gültigen Pixelintensitäten in B' gerechnet werden kann.

Während Cao et al. [15] dieses Verfahren in Verbindung mit bildbasierter (direkter) Registrierung zur Anwendung in der korrelativen Mikroskopie (Abbildung 3.4) vorschlagen, soll in dieser Arbeit der Nutzen für keypoint-basierte Registriermethoden mit konventionellem ℓ_2-Matcher genauer untersucht werden.

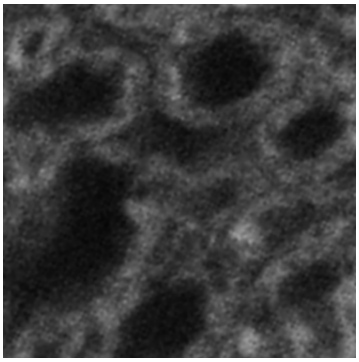
3.4.2 Einordnung

3.4.2.1 Parameter und Komplexität

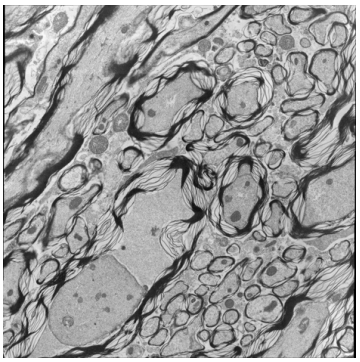
Zur Evaluation wurde die Image Analogies Implementierung von Cao et al. [15] herangezogen, die im Wesentlichen auf der Originalsoftware von Hertzmann et al. [33] basiert. Beide



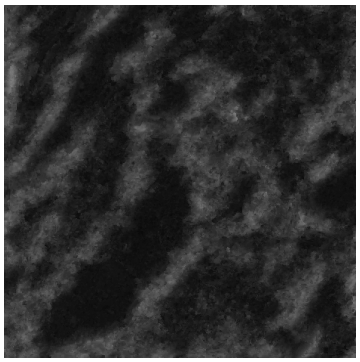
(a) Training: TEM



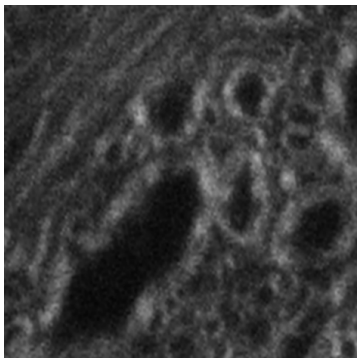
(b) Training: Confocal



(c) TEM-Testbild



(d) gefiltertes Testbild



(e) Ground Truth

Abbildung 3.4: Image Analogies für die Anwendung in der korrelativer Mikroskopie: IA erlernt aus dem Trainingspaar (Bilder a und b) einen Filterprozess, der TEM-Bilddaten (Transmissionselektronenmikroskopie) auf Lichtmikroskopiedaten abbildet. Im Vergleich zur Ground Truth (Bild e) können die dominanteren Strukturen dabei recht gut erfasst werden, wohingegen ein Großteil der kleineren Details unterdrückt bzw. falsch synthetisiert werden und das Bild insgesamt zur Undeutlichkeit neigt. Bildmaterial aus [15].

Tabelle 3.3: Parameter bei Image Analogies

Parameter	Beschreibung	Wert
N^2	Bildgröße	$< 500^2$px
σ_{NB}	Größe der berücksichtigten Nachbarschaft	$5 - 11$
L	Höhe der Auflösungspyramide	$3 - 5$
ϵ_{coh}	Kohärenzparameter	$0 - 10$

Varianten sind sehr experimentell gehalten und bieten eine Vielzahl von Parametern zur Konfiguration des Lernprozesses an (Tabelle 3.3). Den größten Einfluss haben hierbei die Größe der Nachbarschaft, die Höhe und Art der Auflösungspyramide sowie die Gewichtung der Kohärenz. Zur Beschleunigung der Suche wird ein ANN-Verfahren eingesetzt. Darüber hinaus können bis zu 239 Trainingspaare übergeben werden, wobei die Software ab einer bestimmten Anzahl zu berücksichtigender Nachbarschaften nur instabil läuft. Die Zeitkomplexität hängt mit $\mathcal{O}(N^2 \log N^2)$ von der Größe der Trainings- und des Filterbildes (alle $N \times N$) ab. Bei konstanten Bilddimensionen geht die Größe der Nachbarschaft σ_{NB} quadratisch ein.

Die Auflösungspyramide unterstützt dabei die Suche für jeweils nächste Ebene, bei zu starken Reduktion L gehen jedoch auch gröbere Strukturen verloren. Dies kann unter Umständen verhindern, dass die zugehörigen Details in den darauf folgenden Schritten richtig eingefügt werden und das entstehende Bild somit eher einer zufällig generierten Textur gleicht.

3.4.2.2 Bewertung

Ähnlich wie bei Wörterbuchalgorithmen basieren mit Image Analogies gelernte Filter direkt auf den zur Verfügung gestellten Trainingsbildpaaren. Hierbei findet keinerlei Abstraktion der Trainingsdaten statt, was die Generalisierungsfähigkeit stark einschränkt. Strukturen müssen stets in derselben Intensität, Ausrichtung und Skalierung trainiert worden sein, um bei der Filteroperation synthetisiert werden zu können. Zwar ermöglichen kleinere Nachbarschaften $\sigma_{\text{NB}} \approx 3$ bessere Treffer im Sinne der euklidischen Distanz, dennoch fehlt bei der Synthese möglicherweise der Kontext, in welche Art von Umgebung ein Pixel eingefügt werden soll. Denn je nach Kontext können ähnliche Gruppen von Intensitäten zu sehr verschiedenen Strukturen gehören, die sich über die Modalitäten hinweg anders abbilden. Durch eine größere Nachbarschaft $\sigma_{\text{NB}} \geq 5$ kann dieses Problem reduziert werden; dies setzt jedoch voraus, das Strukturen dieser Art auch genau so in den Trainingsbildern vorkommen. Zwar können durch das Vorschalten eines Preprocessing-Schrittes zumindest Varianzen in der Helligkeitsverteilung zwischen Trainings- und Testbildern klein gehalten werden, jedoch müssen darüber hinausgehende Variationen wie geometrische Verzerrungen oder Rotationen durch die Trainingsdaten (Kapitel 4.2) abgedeckt werden.

Von großer Wichtigkeit ist es außerdem, dass Quell- und Filterbild des Trainingsbildpaares mit einer ausreichenden Genauigkeit registriert sind, da sonst die örtliche Zuordnung nicht korrekt ist und somit Bildelemente an falschen Orten synthetisiert werden.

Ein weiterer Freiheitsgrad besteht darin, in welcher Richtung die Transformation durchgeführt werden soll. Je nach Auflösungsunterschieden, Rauschen und Strukturdetails kann dabei eine Richtung sinnvoller sein, um Artefakte zu vermeiden.

Eine Erhöhung des Kohärenzparameters ϵ_{coh} kann das Erscheinungsbild des synthetisierten Bilds visuell konsistenter gestalten, jedoch werden dabei Intensitäten eingefügt, die nicht lokal mit dem Quellbild übereinstimmen und das Ergebnis daher verfälschen können. Dies kann insbesondere bei schlecht gewählten Trainingsdaten und hohen Werten von ϵ_{coh} zu einer starken Fehlerfortpflanzung führen und ist daher für eine Registrierungsanwendung unbrauchbar.

Der Berechnungsaufwand und damit die Rechenzeit ist dabei je nach Trainings- und Testbildergröße nicht unerheblich. So benötigt die Filterung eines 500×500-Bildes mit 239 Trainingsbildpaaren auf dem Testrechner (siehe Kapitel 5) etwa 45 min.

Kapitel 4

Auswahl der Trainingsdaten

Die Tauglichkeit und Robustheit eines gelernten Modells für die Anwendung hängt maßgeblich von der Güte und Art der für das Lernverfahren verwendeten Trainingsdaten ab. Aufgrund der statistischen Natur der Verfahren besteht dabei eine grundlegende Annahme darin, dass Trainings- und Testdaten von derselben Statistik sind. In diesem Kapitel werden Vorgehensweisen zur Auswahl von Trainingspaaren beschrieben, die dem Lernprozessen und Experimenten in Kapitel 5 zu Grunde liegen.

4.1 Wörterbuch und diskriminative Verfahren

Während für Wörterbuchalgorithmen lediglich korrespondierende Merkmalsdeskriptoren, also Daten positiver Trainingspaare benötigt werden, hängt die Performance diskriminativer Lernverfahren ebenso stark von den verfügbaren Negativbeispielen ab. Die Auswahl solcher Daten kann auf unterschiedlichen Wegen erfolgen.

4.1.1 Positive Trainingspaare

Eine Grundvoraussetzung zur Gewinnung von korrespondierenden Merkmalsdeskriptoren besteht darin, dass eine Menge von bereits aufeinander registrierten Trainingsbildpaaren existiert. Im Falle des Funduskopie-Datensatzes kann diese leicht erfüllt werden, da diese Bilder mit dem konventionellen Lowe-ℓ_2-Matching registriert werden können. Zwar verbleibt dabei ein kleiner Restfehler; interne Untersuchungen haben allerdings gezeigt, dass eine manuelle Ausrichtung ein schlechteres Resultat liefert. Anschließend können die Bilder beider Modalitäten jeweils paarweise verarbeitet werden.

Für jedes Trainingsbildpaar werden dazu analog zum Registrierprozess (Kapitel 2.4) Keypoints detektiert. Die Keypointpositionen in homogenen Koordinaten werden im Folgenden als $r_1 = \begin{pmatrix} x_1 & y_1 & 1 \end{pmatrix}^T$ (Modalität 1) und $r_2 = \begin{pmatrix} x_2 & y_2 & 1 \end{pmatrix}^T$ (Modalität 2) bezeichnet. Um örtlich korrespondierende Keypoints zu selektieren können die Keypointpositionen der ersten Modalität über die affine Transformationsmatrix $H_{3\times3}$ aus der Registrierung in das Koordi-

natensystem des zweiten Bildes über $r_{1,t} = Hr_1$ transformiert werden. Im Anschluss können die Keypoints mit diesen Positionen über ihre örtliche Distanz zu denen an den Positionen r_2 zugeordnet werden. Dieses Matching wird jeweils für Keypoints derselben oder ähnlicher Skalengröße durchgeführt.

Um die Qualität der Trainingsdaten zu erhöhen, werden Mehrfachzuordnungen und Zuordnungen, die örtlich zu weit auseinanderliegen, eliminiert. Darüber hinaus können die Keypoints vorab nach ihrem *score*, also dem Wert der Harrisfunktion und damit ihrer Güte, gefiltert werden um Punkte in Umgebungen mit sehr viel Rauschen und wenig Struktur zurückzuweisen. Des Weiteren kann die Einhaltung eines Mindestabstands zwischen den vorläufigen Korrespondenzen gefordert werden, um gleiche bzw. sehr ähnliche Trainingspaare zu vermeiden. Die einzelnen Parameter werden hierbei auf eine möglichst große Relevanz und Repräsentativität der resultierenden Trainingsdaten abgestimmt. Abbildung 4.1 (oben) zeigt beispielhaft ein korrespondierendes Trainingspaar, Abbildung 4.2 das Ergebnis des Auswahlverfahrens.

Im Anschluss werden für alle beteiligten Keypointpaare $(r_1, r_2)_i$ die zugehörigen Deskriptoren $(f_1, f_2)_i$ extrahiert und in ein Wörterbuch (siehe Kapitel 3.1) abgespeichert.

4.1.2 Negative Trainingspaare

Um eine ausreichend diskriminative Ähnlichkeitsmetrik zu erlernen ist es darüber hinaus nötig, den Ähnlichkeitsbereich mit unähnlichen Negativbeispielen abzustecken. Die Auswahl der negativen Trainingspaare hat einen großen Einfluss auf das resultierende Distanzmaß und wurde daher auf verschiedene Arten durchgeführt und experimentell untersucht. Die Verfahren werden im Folgenden kurz beschrieben. Das Vorgehen läuft dabei bis zum Matching analog zum vorherigen Abschnitt ab.

Zufällige Zuordnungen Eine einfache Möglichkeit zur Gewinnung von Negativbeispielen besteht darin, die Keypoints eines bimodalen Bildpaares anhand von zufällig permutierten Zuordnungsindizes einander zuzuweisen. Um zu vermeiden, dass dabei eigentlich ähnliche Trainingspaare als Negative gelabelt werden, können zusätzliche Nebenbedingungen gefordert werden. So ist es zum einen sinnvoll, Zuordnungen zu eliminieren, die sich örtlich (abhängig von den beteiligten Keypoint-Skalen) zu nahe sind; zum anderen kann durch Bevorzugung von Paaren mit sehr unterschiedlichen Keypoint-Skalen die Wahrscheinlichkeit minimiert werden, dass sich die Keypointumgebungen tatsächlich ähneln. Auf Abbildung 4.1 (unten) sind die Bildausschnitte einer solchen negativen Korrespondenz zu sehen.

Zufällige Permutation der Positivbeispiele Wurde bereits ein Wörterbuch aus Positivpaaren erstellt, so können daraus mit Hilfe von zufälligen Zuordnungen Negative generiert werden. Dies hat den Vorteil, dass von jedem solchen Negativpaar jeweils zwei der Positiven betroffen sind und voneinander getrennt werden. Dies kommt der Relevanz des Negativbeispiels zu Gute. Auf der anderen Seite kann mit diesem Verfahren schwer verhindert werden, dass zwei sehr ähnliche Positivpaare fälschlicherweise als negativ gelabelt werden.

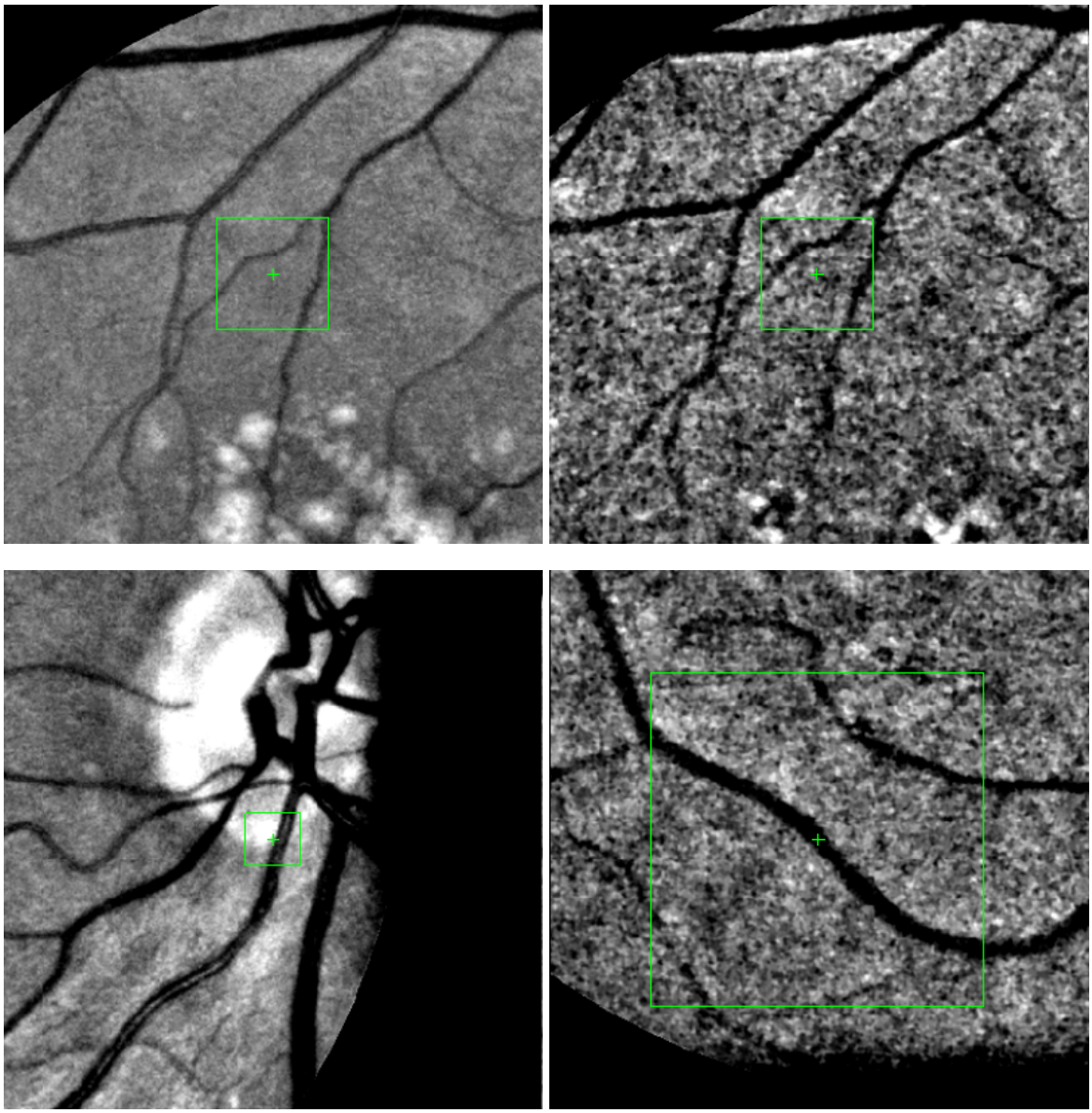


Abbildung 4.1: Visualisierung von Beispielkorrespondenzen bei der Fundusbildgebung. (a) und (b) zeigen die Bildausschnitte eines ähnlichen Deskriptorpaares, (c) und (d) eines unähnlichen Paares, welches zufällig über einen möglichst großen Skalenunterschied zugeordnet wurde. Links ist jeweils die Photographie, rechts das Autofluoreszenzbild (FAF) zu sehen.

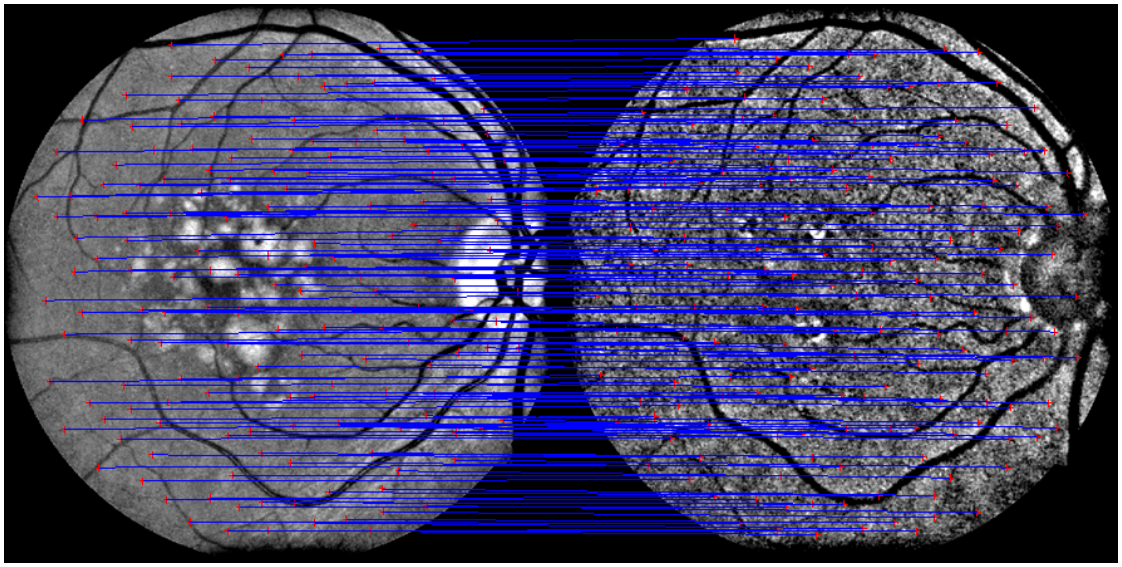


Abbildung 4.2: Visualisierung gefundener positiver Korrespondenzen auf einem Fundus-Bildpaar. Links ist die Photographie, rechts das Fluoreszenzbild zu sehen. Örtlich Korrespondierende Keypoints sind mit blauer Linie verbunden.

Bootstrapping Auch ein iteratives Vorgehen kann verwendet werden, um Trainingsbeispiele in Abhängigkeit einer bereits gelernten Ähnlichkeitsmetrik hinzuzufügen. Dabei kann im Idealfall der zugrundeliegende Registrierungsprozess nachgebildet werden, indem die Ähnlichkeitswerte eines jeden Keypointdeskriptores der einen Modalität zu allen Deskriptoren der anderen Modalität ausgewertet werden. Anschließend können die Keypointpaare mit den kleinsten Distanzen, die weit genug von der korrekten Position entfernt sind, als Negative hinzugefügt werden. Eine andere Bootstrapping-Variante benutzt die bisher gelernte Metrik, um aus einer Menge von Negativ-Kandidaten (z.B. aus zufälliger Permutation) solche mit großer Relevanz für das Lernverfahren herauszufiltern. Anschließend wird das Distanzmaß mit der vergrößerten Trainingsmenge neu trainiert. Das Bootstrapping kann beliebig oft wiederholt werden.

Jedes der genannten Verfahren bringen mindestens eines der folgenden Probleme mit sich. Zum einen garantiert eine örtliche Mindestdistanz zwischen den Keypoints im Allgemeinen keine Unähnlichkeit. Insbesondere bei Bildmaterial mit sich wiederholenden Bildstrukturen kann diese Annahme nicht gemacht werden. Dies kommt vor allem in Verbindung mit einem Bootstrapping-Verfahren zum Tragen, da solche Paare tendenziell eine niedrige Distanz aufweisen und daher als sehr relevant eingestuft und als Negative übernommen werden. Zum anderen sind sich Negativpaare auf Basis stark verschiedener Keypoint-Skalen zwar mit hoher Wahrscheinlichkeit auch tatsächlich unähnlich, allerdings ist dafür die Relevanz solcher Trainingspaare für das Lernverfahren nicht unbedingt hoch. Während das erste Problem durch die Zurückweisung von negativen Trainingspaaren mit geringer euklidischer Distanz in gewissen Grenzen reduziert werden kann, bleibt die Frage der Qualität und Repräsentativität Trainingsdaten hingegen weiterhin offen.

Auf Auswirkung der verschiedenen Methoden zur Trainingsdatenauswahl beim Similarity-Sensitive Hashing wird in Kapitel 5.3.2 eingegangen.

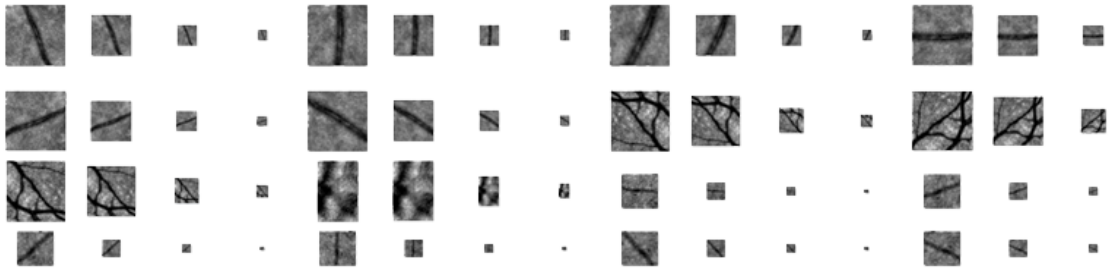


Abbildung 4.3: Trainingsdatenbeispiel für Image Analogies. Die Patches beinhalten relevante Strukturen aus der Fundusbildgebung (nach dem Preprocessing) in verschiedenen Rotationen und Skalen (nur eine Modalität gezeigt).

4.2 Image Analogies

Im Gegensatz zu den anderen vorgestellten Methoden werden für das Lernen eines Image Analogy Filters Bildpaare bzw. Paare von korrespondierenden Bildpatches verwendet.

Aufgrund der in Kapitel 3.4 angesprochenen fehlenden Invarianz des Verfahrens gegenüber Intensität, Skalierung und Rotation stellen diese Aspekte die Anforderungen an taugliche Trainingsdaten dar. Demzufolge wird in den Experimenten je nach Datensatz neben ganzen Bildern auch mit manuell selektierten Patches gearbeitet, welche in mehreren synthetisch generierten Rotationen und Auflösungsstufen vorliegen. Dazu werden möglichst Bildausschnitte mit signifikanten Strukturen gewählt um mit der maximal möglichen Anzahl an Trainingsbildern eine möglichst hohe Repräsentativität zu erreichen (siehe Abbildung 4.3). Intensitätsschwankungen sollen durch einen Preprocessing-Schritt vermieden werden.

Da nicht alle verwendeten Bilddatensätze in registrierter Form vorlagen, wurden diejenigen Bildpaare, an denen das konventionelle Verfahren scheiterte, über manuell selektierte Korrespondenzen und eine Least-Square Schätzung registriert.

Kapitel 5

Auswertung

In diesem Kapitel sollen alle durchgeführten Experimente mit den in Kapitel 3 vorgestellten Verfahren beschrieben und deren Resultate präsentiert werden. Um vergleichbare Ergebnisse zu erhalten wurden die Verfahren allesamt in die in Kapitel 2.4 dargelegte Prozesskette integriert und die Experimente auf demselben Testrechner (Intel Core i5 M560 mit 4x2.67GHz, 8GB RAM mit MATLAB R2013a auf Windows 7) durchgeführt. Nach einer kurzen Darstellung der Verfahrensbewertung erfolgt die Beschreibung der experimentellen Untersuchung aller Ansätze zusammen mit der Diskussion der Ergebnisse. Abschließend sollen die Verfahren im Hinblick auf die gezeigten Eigenschaften bei den verwendeten Bilddaten analysiert und Schlüsse bezüglich der Tauglichkeit für des multimodale Registrierproblem gezogen werden.

5.1 Datensätze

Zu Beginn sollen die verwendeten Bilddatensätze kurz vorgestellt werden. Der Datensatz „Fundus I“ beinhaltet 112 Bildpaare aus der Fundusbildgebung, jeweils bestehend aus einer Farbphotographie und einer Autofluoreszenzaufnahme (FAF) des Augenhintergrundes (siehe Abbildung 1.2) Beide Bilder haben die Auflösung $1440 \times 1440\,\mathrm{px}^2$. Die Trainingsdaten wurden aus 80 der Bildpaare entnommen, die restlichen 32 dienten als Testbilder für die Registrierung.

Da die hohe Auflösung der originalen Fundusbilder eine lange Rechenzeit bei der Anwendung von IA-Filtern bedeutet, umfasst der Datensatz „Fundus II“ $500 \times 500\,\mathrm{px}^2$ große Ausschnitte aus dem Zentrum der Fundusbilder in bereits registrierter und vorverarbeiter Form. So können daraus direkt Trainingsdaten für Image Analogies gewonnen und die Korrektheit der Zuordnungen einfach überprüft werden.

Darüber hinaus dienten weitere Bilder aus der korrelativen Mikroskopie für Experimente mit Image Analogies, darunter Hautgewebe und Aufnahmen von Zellen des menschlichen Gehirns [15]. Aufgrund der begrenzten Anzahl verfügbarer Trainings- und Testbilder konnte auf diesen Daten neben der Euklid-Baseline lediglich das Image Analogies Verfahren evaluiert werden.

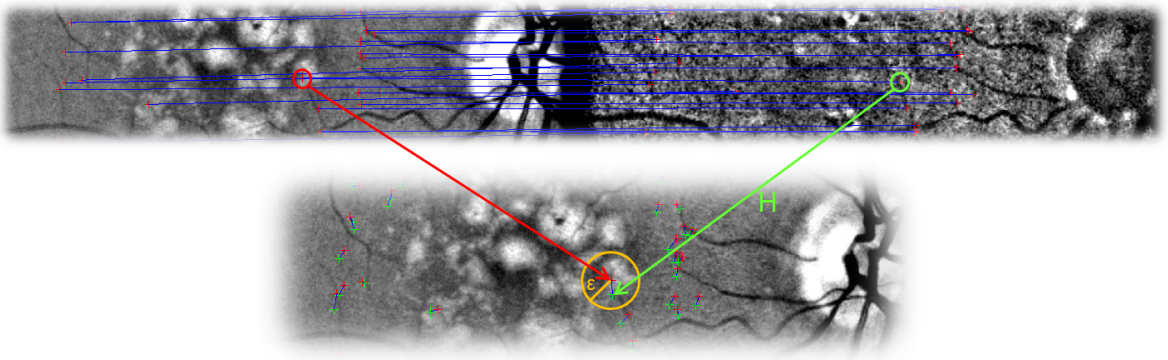


Abbildung 5.1: Visualisierung einer korrekten Korrespondenz auf einem Fundus-Bildpaar. Keypoints der einen Modalität werden in das Koordinatensystem der anderen transformiert (grün). Befindet sich der resultierende Punkt innerhalb einer ϵ-Umgebung (orange) des anderen Keypoints (rot), wird die Korrespondenz als *true positive* gewertet.

5.2 Evaluationsmaße

Zur einheitlichen quantitativen Bewertung der untersuchten Verfahren wird jede gefundene Korrespondenz auf Korrektheit überprüft, indem zunächst einer der zugehörigen Keypoints über die vorab bekannte Homographiematrix H („Ground Truth“) in das jeweils andere Koordinatensystem transformiert wird. Anschließend kann die euklidische Distanz zwischen dem Korrespondenzpaar berechnet werden. Liegt diese innerhalb eines Schwellwertes ϵ, welcher abhängig von der typisch erreichten Registriergenauigkeit und Verzeichnungseffekten angemessen gewählt werden muss, so gilt die Korrespondenz als gültig (*true positive*). Andernfalls wird diese als *false positive* gezählt. Abbildung 5.1 veranschaulicht dieses Vorgehen an einem Fundus-Bildpaar. Als ϵ-Radius wurden hierbei 30 px angesetzt.

Trägt man die Verteilung der *true positives* und *false positives* über dem zugehörigen Ähnlichkeitswert (*similarity score*), so lässt sich anhand deren Überscheidung die Trennfähigkeit der untersuchten Metrik einschätzen (Abbildung 5.2). Zur genaueren Auswertung kann die absolute Anzahl der gefundenen Treffer (pro Bild) über dem Anteil der Falschzuordnungen bis zu einem bestimmten Ähnlichkeitswert aufgetragen werden. Führt man dies in absteigender Sortierung für alle *similarity scores* durch, so erhält man die zugehörige ROC-Kurve (Receiver Operating Characteristic [23], siehe Abbildung 5.4). Da robustere bzw. diskriminativere Metriken für eine bestimmte Trefferanzahl eine geringere Fehlerquote aufweisen und damit stets links bzw. über einer schlechteren Vergleichskurve liegen, lassen sich anhand dieser Kurve verschiedene Verfahren einfach vergleichen.

Darüber hinaus lässt sich ein Verfahren anhand der Gesamtanzahl gefundener und korrekter Korrespondenzen (pro Bild) bewerten, indem die erreichte Höhe der ROC-Kurve bei 100% *false positive rate* abgelesen wird. Anhand dieser Anzahl lässt sich beispielsweise einschätzen, ob einem Verfahren aufgrund der veränderten Metrik oder fehlender Strukturen Zuordnungen entgehen, möglicherweise zugunsten einer besseren Klassifikationsrate.

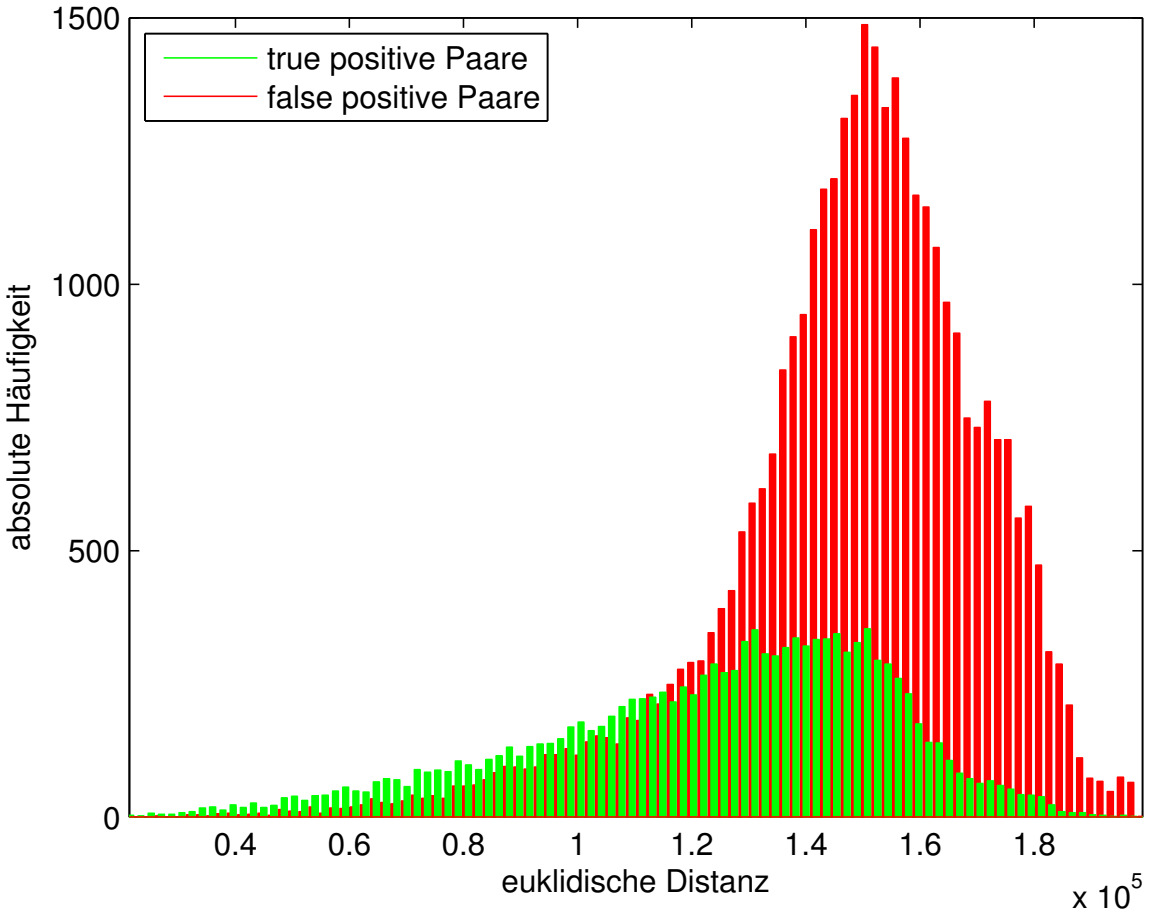


Abbildung 5.2: Verteilung der Matching-Distanzen bei der Euklid-Baseline. Die Distanzen tatsächlich korrekten Korrespondenzen sind in grün, die falscher Korrespondenzen in rot aufgetragen. Die euklidische Distanz nimmt von links (hoher Ähnlichkeitswert) nach rechts (niedriger Ähnlichkeitswert) zu.

5.3 Ergebnisse

Zunächst soll anhand des Vergleichsverfahrens (Euklid-Baseline) das Zustandekommen der ROC Kurve genauer erläutert werden. Wie oben beschrieben, liegen die Korrespondenzen nach dem Experiment abhängig vom Schwellwert ϵ in klassifizierter Form vor.

Abbildung 5.2 zeigt das Histogramm der euklidischen Distanzen richtig und falsch klassifizierter Korrespondenzen auf dem Datensatz „Fundus I“. Zwar gibt es eine leichte Tendenz der korrekten Zuordnungen in Richtung niedriger Distanzen, dennoch weisen die Verteilungen aufgrund ihrer breiten Varianz eine hohe Überschneidung auf. Das Zuordnungsergebnis der verwendeten Metrik ist auf diesen Daten demnach mit vielen Ausreißern durchsetzt, welche zum Teil sehr niedrige Distanzen besitzen.

Zur genaueren Untersuchung dieser falschen Korrespondenzen können die zugehörigen Bildausschnitte herangezogen werden. Abbildung 5.3 zeigt auf, dass Falschzuordnungen kleinster Distanzen sich im Bild wiederholende Strukturen (bspw. Verlauf von Adern im Fundusbild) zur Ursache haben. Dies ist insbesondere bei Keypoints kleiner Skalen der Fall,

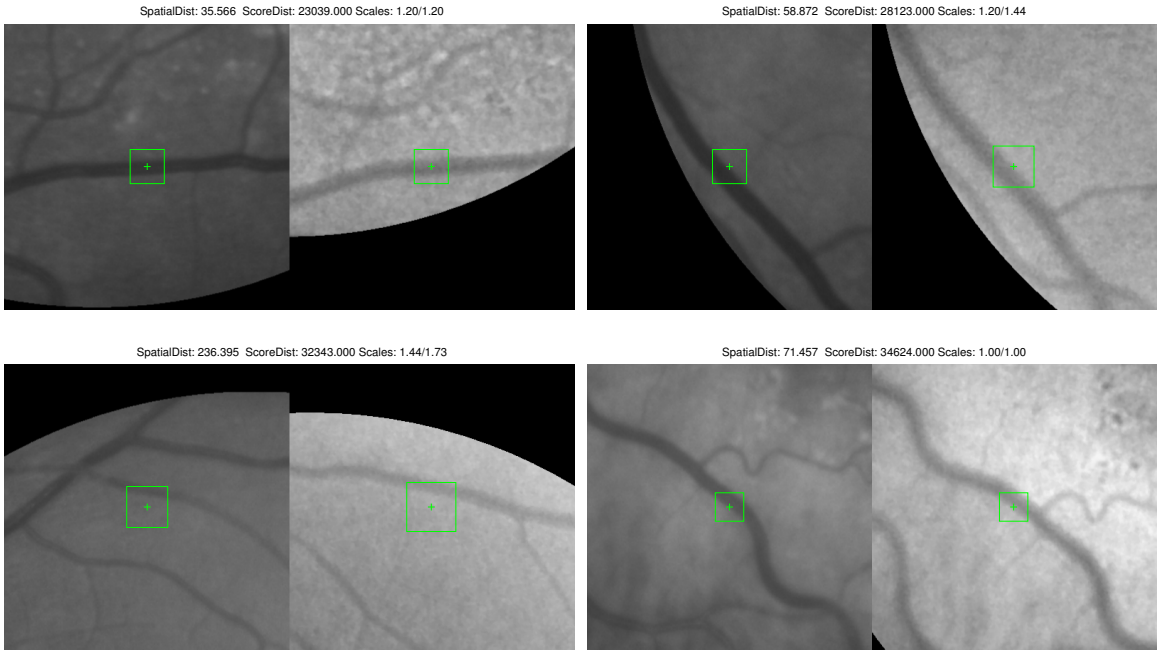


Abbildung 5.3: Die Abbildung zeigt vier Bildausschnitte (Patches), auf denen die Deskriptorumgebungen der Keypoints der vier Falschzuordnungen mit den kleinsten Distanzen dargestellt sind. Dabei gibt *SpatialDist* die örtliche- und *ScoreDist* die Merkmalsdistanz an. Die grün eingezeichnete Umrandung beinhaltet den jeweils Bereich, auf dem der SIFT-Deskriptor berechnet wird. Als Schwellwert wird $\epsilon = 30$px angenommen.

bei denen der Kontext im Bild nicht so sehr zum Tragen kommt. In einem anderen Anwendungsfall könnte als weiterer Grund auch eine zu kleine Wahl des ϵ-Wertes in Frage kommen.

Trägt man nun wie oben beschrieben für jeden Distanzwert die Anzahl der gefundenen, korrekten Zuordnungen über der Fehlerrate auf, so erhält man die in Abbildung 5.4 gezeigte ROC-Kurve. Dadurch lässt sich die Güte des untersuchten Verfahrens, sowohl im Bezug auf die Klassifikation (Verlauf und Anstiegsrate der Kurve), als auch auf die Gesamtanzahl der gefundenen korrekten Zuordnungen (erreichte Höhe der Kurve) bewerten und mit anderen Methoden vergleichen. Im Folgendem sollen auf diese Weise die in Kapitel 3 eingeführten Verfahren näher evaluiert werden.

5.3.1 Wörterbuch-basiert

Wie in Kapitel 4.1.1 beschrieben, spielt bei der Erstellung des Wörterbuch neben der Anzahl an Einträgen vor allem die zulässige Nachbarschaftsgröße, also der maximale örtliche Abstand eines Korrespondenzpaares, um als „ähnlich“ gelabelt zu werden, eine ausschlaggebende Rolle. Da mit einer größeren Nachbarschaft pro Bildpaar mehr zulässige Korrespondenzen entstehen, sind diese beiden Parameter nicht unabhängig voneinander. Dies hat, auch aufgrund der begrenzten Anzahl an Trainingsbildern, zur Folge, dass für die Experimente nicht jeder Parameter beliebig in Relation zum anderen variiert werden kann. Daher muss

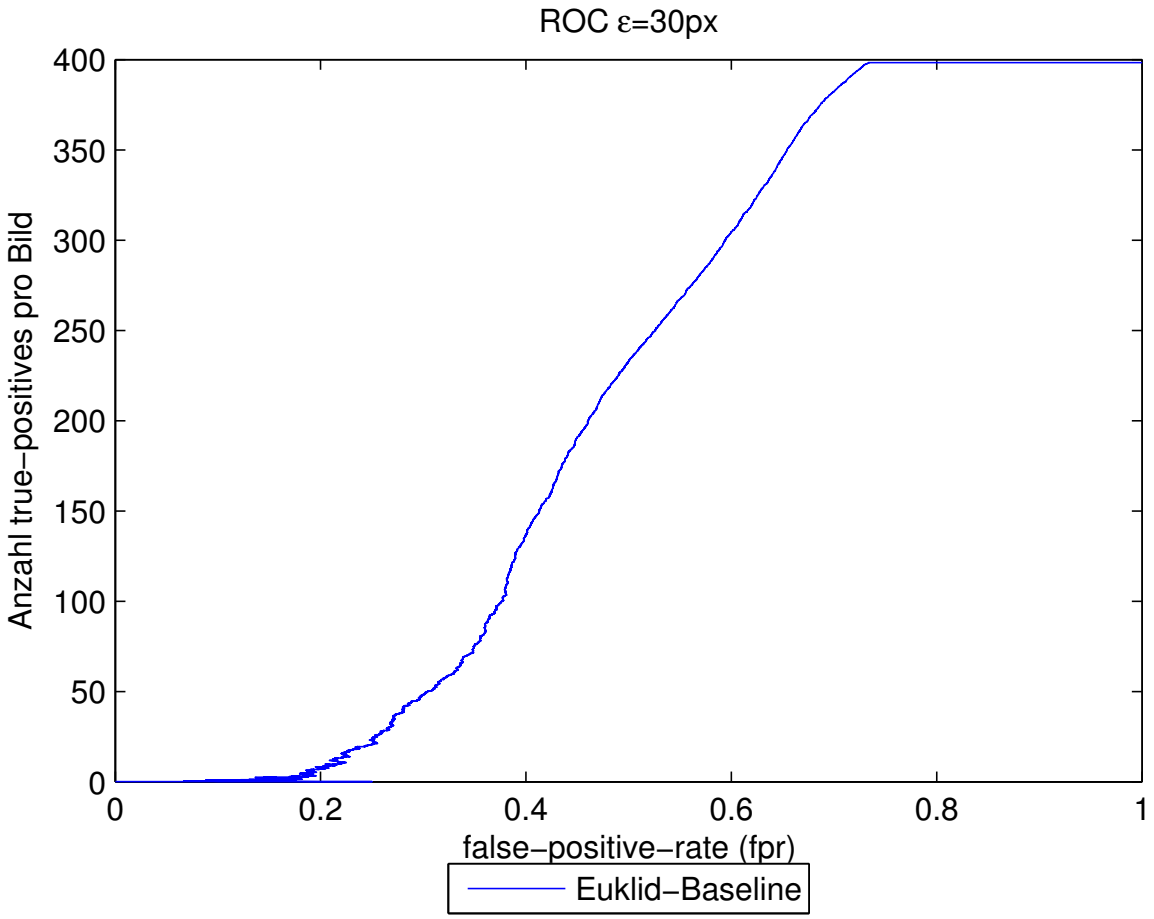


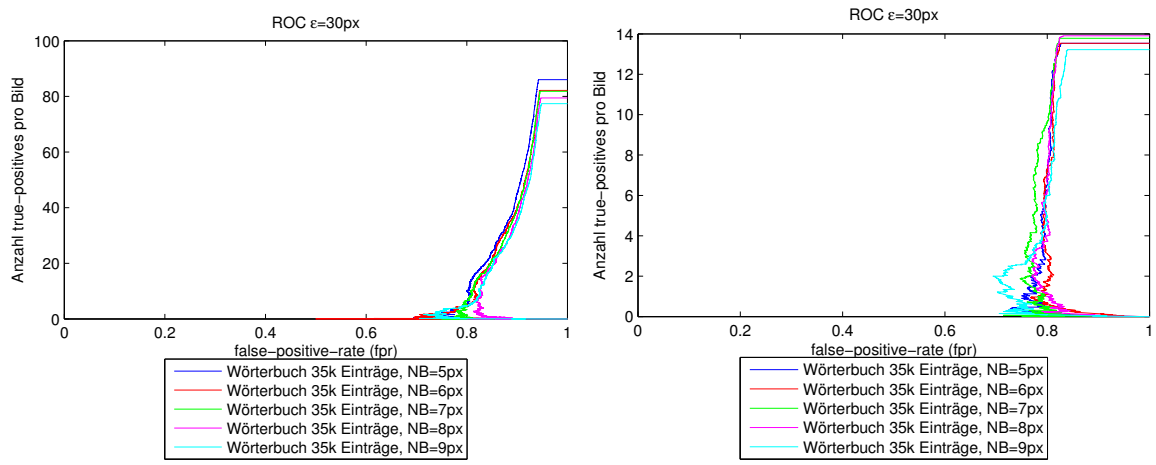
Abbildung 5.4: Die ROC-Kurve zum Experiment mit dem Vergleichsverfahren Euklid-Baseline, aufgenommen auf dem Datensatz „Fundus I“, ohne *matching ratio*.

auf der einen Seite beim Vergleich von Wörterbüchern unterschiedlicher Nachbarschaftsgröße die Anzahl der Einträge auf die des kleinsten beteiligten Wörterbuches angeglichen werden, indem ein entsprechender Anteil zufällig gezogen wird. Auf der anderen Seite kann bei der Untersuchung des Einflusses der Wörterbuchgröße bei konstant bleibender Nachbarschaftsgröße die Anzahl der Einträge nicht beliebig vergrößert werden.

Ziel der folgenden Experimente ist, Wörterbücher auf Basis ausgewählter Trainingsdatenkonfigurationen in Abhängigkeit der genannten Parameter sowie unter zusätzlicher Anwendung einer *matching ratio* [40] (siehe Kapitel 2.4, Schritt 4) zu untersuchen. Dabei soll insbesondere auf die Auswirkung der einzelnen Parameter auf die Güte der Korrespondenzfindung eingangen werden. Wie in Kapitel 3.1 beschrieben, wird neben einem reinen Wörterbuchverfahren auch die lineare Approximation betrachtet. Abschließend soll ein Vergleich der besten Konfigurationen mit der Euklid-Baseline vorgestellt werden.

Einfluss der zulässigen Nachbarschaftsgröße

Zu Beginn soll die Bedeutung der zulässigen Nachbarschaftsgröße bei der Trainingsdatenauswahl (siehe vorheriges Kapitel) für den Wörterbuchalgorithmus ermittelt werden. Abbildung 5.5 zeigt dazu die ROC-Plots der entsprechenden Verfahren. Der linke Plot zeigt

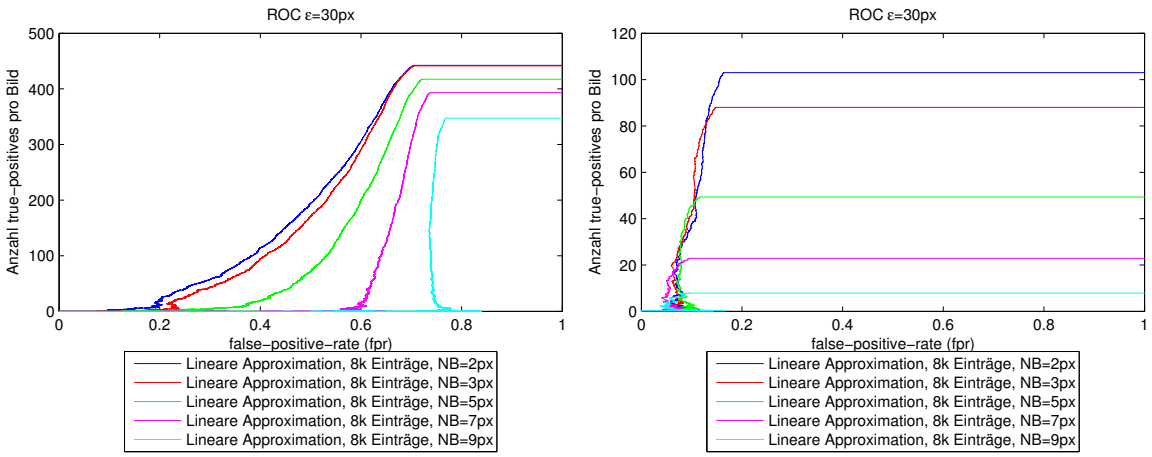


(a) ohne *matching ratio* (b) *matching ratio* 1.2

Abbildung 5.5: Einfluss der zulässigen Nachbarschaftsgröße beim Wörterbuch

jeweils das Ergebnis ohne Verwendung von Lowes *matching ratio* [40], es findet also keinerlei Zurückweisung von zu ähnlichen und damit evtl. mehrdeutigen Zuordnungen statt. Dagegen ist das Verhältnis 1.2 (rechter Plot) ein heuristisch ermittelter Wert, welcher auf die vorliegenden Daten abgestimmt wurde. Für dieses Experiment wird die Anzahl der Einträge im Wörterbuch auf 35000 konstant belassen, während die zulässige Nachbarschaftsgröße im Bereich von 5 px bis 9 px variiert wird. Es ist zu beobachten, dass in allen Fällen die Kurven ab einer *false-positive-rate* von etwa 80% ein sehr ähnliches Anstiegsverhalten zeigen. Die Metriken sorgen demnach auch bei niedrigen Distanzen für viele Falschzuordnungen. Das höhere *matching ratio* bewirkt hierbei einen steileren Anstieg der Kurve und damit eine bessere Klassifikation bei gleicher Fehlerrate, allerdings auf Kosten der Gesamtanzahl gefundener Treffer (nur etwa $\frac{1}{6}$). Die Auswirkungen der Nachbarschaftsgröße geht größtenteils im Rauschen unter, es lässt sich lediglich feststellen, dass die größte zulässige Nachbarschaft von 9 px in beiden Fällen das schlechteste Ergebnis im Sinne der wenigsten Treffer liefert, wobei die Unterschiede nicht übermäßig stark ausgeprägt sind.

Im Gegensatz dazu wirkt sich die Nachbarschaftsgröße bei der linearen Approximation auf den Trainingsdaten erheblich stärker aus, wie Abbildung 5.6 zeigt. Ausgehend von einer gut funktionierenden Trainingsdatenkonfiguration (blaue Kurve: 8000 Paare, Nachbarschaftsgröße von 2 px) ist die Performance der resultierenden Metrik bei schrittweiser Erhöhung der Nachbarschaftsgröße bis zu einem Radius von 9 px gezeigt. Hierbei ist deutlich zu beobachten, dass die Qualität der Schätzung abnimmt, je toleranter die Trainingspaare in Bezug auf die örtliche Abweichung selektiert werden. Zum einen verschiebt sich dabei ohne *matching ratio* der Anstieg der Trefferanzahl in Richtung höherer Fehlerquoten (von 20% bis hin zu 75%), zum anderen sinkt auch die absolute Anzahl korrekter Korrespondenzen pro Bild von 450 auf 350. Beim höheren *matching ratio* unterscheiden sich die Kurven sehr wenig im Verlauf, lediglich die Gesamtanzahl sinkt mit größer werdender Nachbarschaft von über 100 auf unter 10 Treffer pro Bild ab. Der Großteil davon wird bei einer Fehlerrate von etwa 10% gefunden. Als Ursache kommt dabei infrage, dass eine größere Anzahl örtlich zu schlecht übereinstimmender Deskriptoren sich negativ auf die global geschätzte lineare

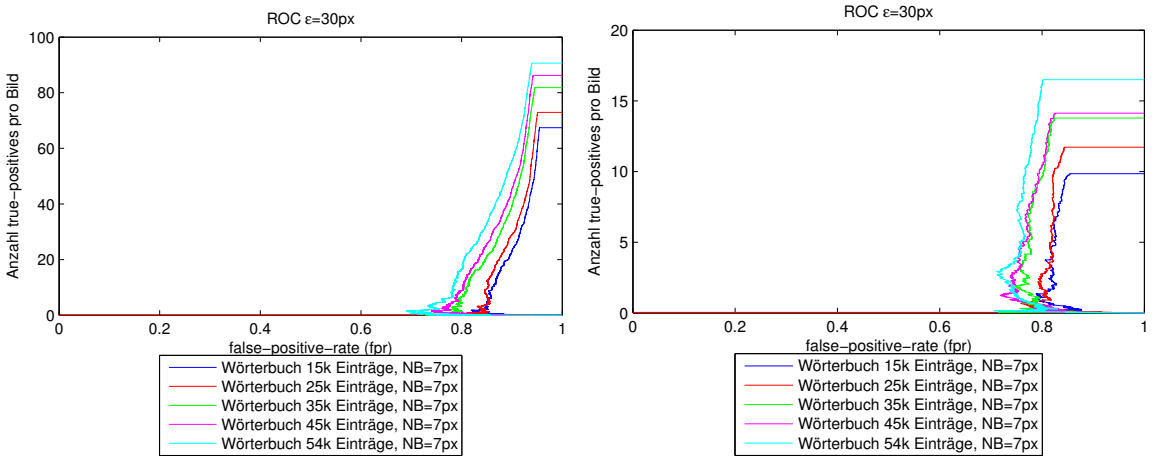


(a) ohne *matching ratio* (b) *matching ratio* 1.2

Abbildung 5.6: Einfluss der zulässigen Nachbarschaftsgröße beim Wörterbuch mit linearer Approximation

Transformation auswirkt und die darauf basierende Prädiktion insgesamt entsprechend verschlechtert. Beim reinen Wörterbuch hingegen wirken sich ein solcher Fehler nur lokal auf den jeweiligen Eintrag aus ist daher eher zu kompensieren.

Einfluss der Wörterbuchgröße

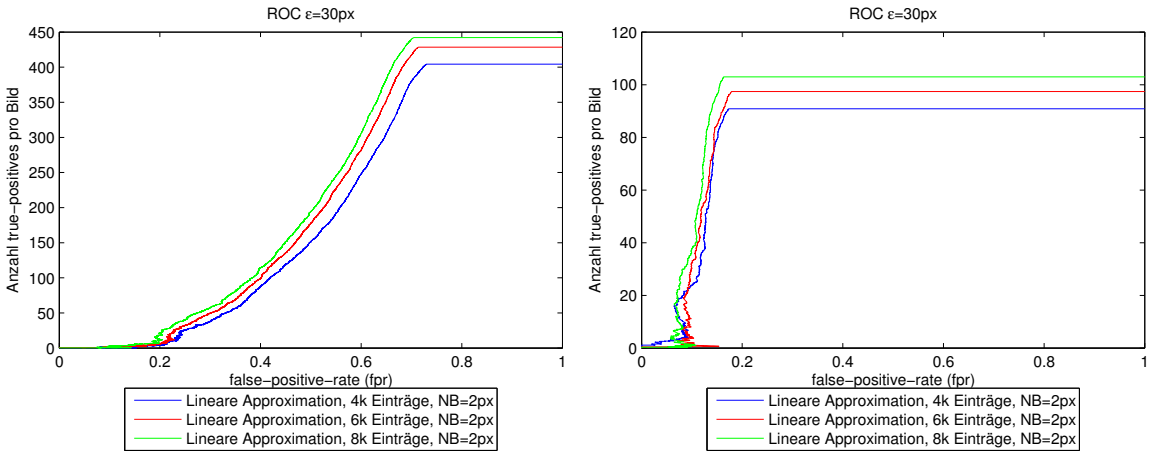


(a) ohne *matching ratio* (b) *matching ratio* 1.2

Abbildung 5.7: Einfluss der Anzahl an Einträgen beim Wörterbuch

Im Anschluss soll nun die Anzahl der Wörterbucheinträge als der zweite einflussreiche Parameter bei der Trainingsdatenauswahl untersucht werden. Dazu wird ausgehend von einer konstanten zulässigen Nachbarschaftsgröße von 7 px die Wörterbuchgröße von 15000 auf 54000 Einträge erhöht. Abbildung 5.7 zeigt, dass das Matchingverfahren erwartungsgemäß von einer größeren Anzahl an Einträgen sowohl hinsichtlich der Fehlerrate als auch der Gesamtanzahl an Treffern profitiert. Dies ist bei beiden gezeigten *matching ratios* der Fall, die Erhöhung des Verhältnisses auf 1.2 bewirkt wiederum einen steileren Anstieg der Kurve und

in diesem Bereich damit mehr Treffer bei gleicher Fehlerrate, zu Lasten der Gesamttrefferanzahl. Im untersuchten Parameterbereich kann die Fehlerquote zwar in gewissem Maße vermindert werden, größere Performancesprünge sind allerdings nicht zu beobachten. Auch ein wie in Kapitel 3.1 diskutierter Sättigungseffekt ist an dieser Stelle nicht zu verzeichnen.



(a) ohne *matching ratio* (b) *matching ratio* 1.2

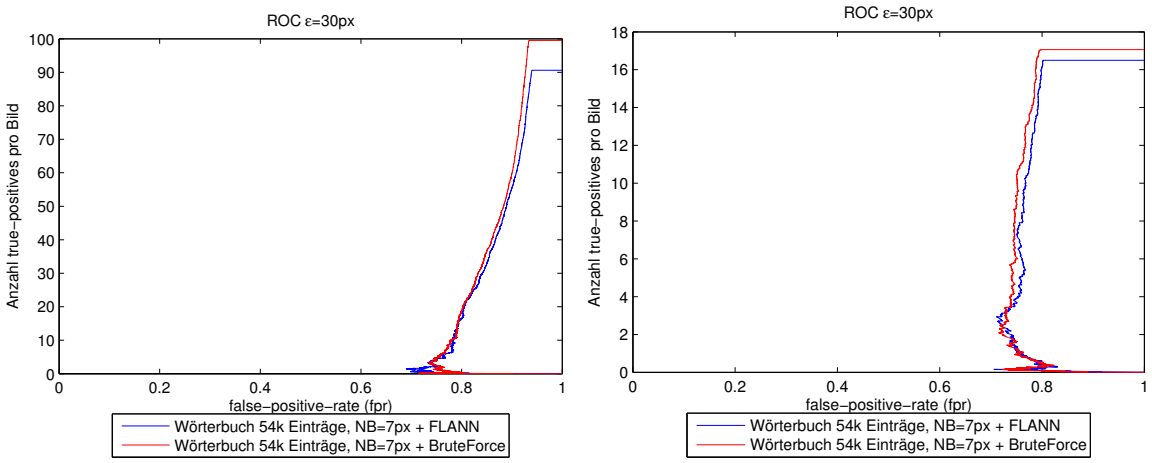
Abbildung 5.8: Einfluss der Anzahl an Einträgen beim Wörterbuch mit linearer Approximation

Auch die Güte der linearen Approximation hängt von der Anzahl der verfügbaren Trainingspaaren ab, wie auf Abbildung 5.8 gezeigt. Wird die Wörterbuchgröße von 4000 auf 8000 Einträge bei gleicher Nachbarschaft ($2\,\text{px}$) erhöht, so wirkt sich dies bei beiden *matching ratios* positiv aus, was an der Verschiebung der Kurve nach oben und nach links erkennbar ist. Im Idealfall sollte die Schätzung der Transformation nach diesen Erkenntnissen auf einer möglichst großen Auswahl an Trainingsdaten mit geringer zulässigen Nachbarschaft berechnet werden.

Einfluss des Suchverfahrens

Wie in Kapitel 3.1 erwähnt, kann die Nearest-Neighbor-Suche in hochdimensionalen Räumen sehr aufwändig sein. Da alle bisherigen Wörterbuchexperimente auf einem approximativen Suchverfahren (FLANN [49]) durchgeführt wurden, soll an dieser Stelle der Unterschied zur exakten Suche experimentell untersucht werden. Abbildung 5.9 zeigt die zugehörigen ROC-Kurven, die anhand eines Wörterbuchs mit 54000 Einträgen und einer $7\,\text{px}$ Nachbarschaft aufgezeichnet wurden. Hier ist unabhängig vom *matching ratio* erkennbar, dass das Suchverfahren den Verlauf und die Steigung der Kurve kaum beeinflusst. In beiden Fällen findet das exakte Suchverfahren (rote Kurve) allerdings einige wenige korrekte Korrespondenzen mehr ($5 - 10\%$). Für den damit verbundenen um Faktor 12 größeren Zeitaufwand ($60\,\text{s}$ vs. $5\,\text{s}$ pro Bild) ist dieser Nutzen allerdings sehr teuer erkauft.

Da der einfache Wörterbuchansatz in allen durchgeführten Experimenten eine nur sehr begrenzte Tauglichkeit aufweist, soll dieser im Folgenden nicht weiter analysiert und der Fokus stattdessen auf die lineare Approximation gelegt werden, welche ein deutlich größeres Po-

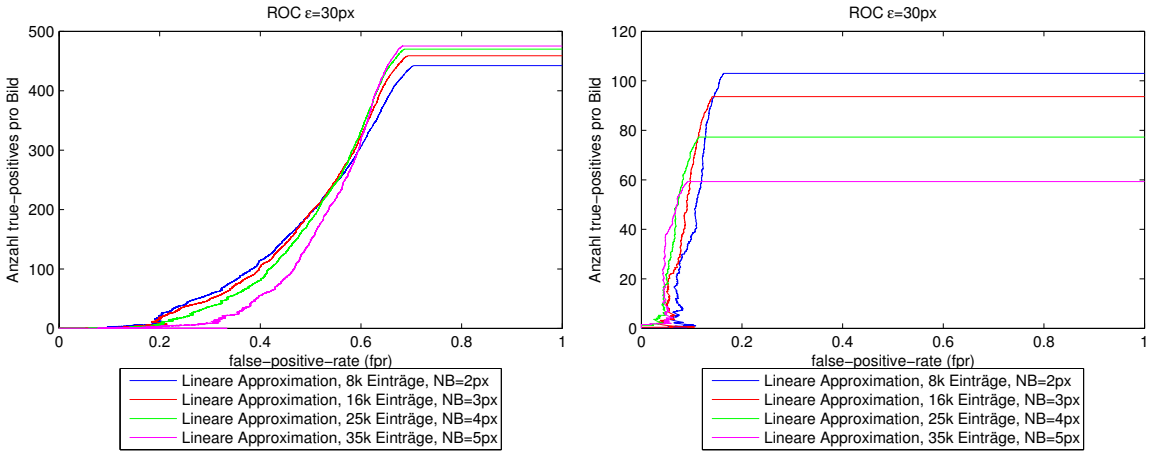


(a) ohne *matching ratio* (b) *matching ratio* 1.2

Abbildung 5.9: Einfluss des verwendeten Suchverfahrens bei der Wörterbuchsuche

tential zeigt. Ein Überblick über die Leistung beider Verfahren zur Baseline ist im Abschnitt „Vergleich der besten Konfigurationen“ zu finden.

Alternative Konfigurationen für lineare Approximation



(a) ohne *matching ratio* (b) *matching ratio* 1.2

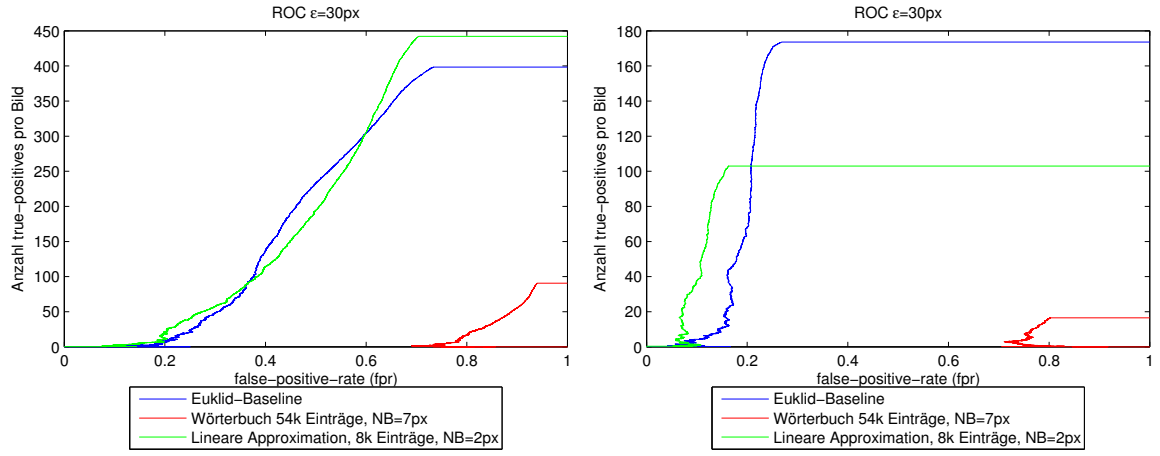
Abbildung 5.10: Einfluss verschiedener Trainingsdaten beim Wörterbuch mit linearer Approximation

Da sich die beiden Parameter Nachbarschafts- und Wörterbuchgröße allerdings, wie eingangs erwähnt, nicht beliebig zueinander variieren lassen, kann eventuell ein Trade-off gefunden werden, indem man den Performanceverlust aufgrund von Trainingsdaten aus größeren Nachbarschaften durch eine höhere Anzahl zu kompensieren versucht. Zu diesem Zweck sind auf Abbildung 5.10 die ROC-Kurven verschiedener Konfigurationen von Trainingsdaten aufgezeichnet. Die Parameter sind dabei so gewählt, dass zu jeder Nachbarschaftsgröße die dabei maximal mögliche Anzahl an Trainingkorrespondenzen ins Wörterbuch übernommen wird. Die Experimente mit und ohne *matching ratio* zeigen dabei ein umgekehrtes Verhalten:

Während ohne die lineare Approximation bei toleranteren Nachbarschaften und dementsprechend größeren Wörterbüchern ein schlechteres Anstiegsverhalten bei einer leichten Erhöhung der Gesamttrefferanzahl zeigt, sorgt dies beim Verhältnis 1.2 eher für steilere Kurven, die früher abflachen und damit weniger Treffer repräsentieren.

Welche Konfiguration hierbei zu bevorzugen ist, hängt stark vom Anwendungsfall ab. Betrachtet man die Metriken beim *matching ratio* von 1.2, so kann entweder eine gewählt werden, die über 50 korrekte Korrespondenzen bei einer Fehlerrate von weniger als 10% findet (magentafarbene Kurve) oder alternativ eine gemäß der blauen Kurve, die doppelt so viele Treffer bei etwa der doppelten Fehlerrate erreicht. Während bei der ersten Metrik ein RANSAC-Algorithmus [24] mit wenigen Iterationen als Schätzverfahren für das geometrische Modell in Frage kommt, ist bei letzterer aufgrund der höheren Ausreißerquote eher ein Hough-Voting [20] geeignet. Dieses kann dabei von der größeren Anzahl an *Votes* profitieren, vorausgesetzt die falschen Korrespondenzen sind einigermaßen uniform über den Parameterraum verteilt.

Vergleich der besten Konfigurationen



(a) ohne *matching ratio* (b) *matching ratio* 1.2

Abbildung 5.11: Vergleich der besten Wörterbuchverfahren zur Euklid-Baseline

Abschließend sollen die besten gefundenen Konfigurationen mit der Euklid-Baseline verglichen werden. Abbildung 5.11 zeigt dazu die zugehörigen ROC-Kurven für beide *matching ratios* in einem gemeinsamen Plot.

Hier wird deutlich, dass die Gesamtperformance des reinen Wörterbuches (rot) mit nur einem Bruchteil der Gesamttreffer und deutlich höheren Fehlerraten unzureichend ist und sich daher für diese Anwendung nicht eignet. Da die globale lineare Approximation der Transformation (grün) zwischen den Deskriptoren der beiden Modalitäten ein deutlich besseres Verhalten zeigt, liegt die Vermutung nahe, dass die Lücken zwischen den Einträgen zu groß sind und daher die Generalisierung auf Basis der Nearest-Neighbor Interpolation versagt. Abhilfe würde entsprechend nur eine massive Erhöhung der Trainingsdaten schaffen, um ausreichend viele mögliche Korrespondenzfälle abzudecken. Dies ist allerdings sowohl aufgrund der Ver-

fügbarkeit von Trainingsdaten, als auch wegen des benötigten Speicherplatzverbrauchs nur begrenzt möglich.

Bei deaktiviertem *matching ratio* (linker Plot) ist das lineare Modell je nach Score in etwa auf der Höhe der Euklid-Baseline (blau). Abgesehen von einigen Schwankungen kann die Approximation bei einem ähnlichen Anstiegsverhalten insgesamt etwa 12.5% mehr korrekte Korrespondenzen liefern. Erhöht man das *matching ratio* auf 1.2, so erreicht das lineare Modell im niedrigen Distanzbereich bis etwa 100 korrekten Korrespondenzen dieselbe Trefferanzahl bei einem um $5 - 10\%$ kleinerem Fehler. Die euklidische Metrik hingegen ist darüber hinaus in der Lage noch weitere Treffer (ca. 70% mehr) zu finden. Auch hier gilt es wiederum je nach Anwendungsfall abzuwägen, welche Kurvenform bzw. Metrik jeweils die bessere Eignung aufweist.

Geschätzte Transformationsmatrix

Die nebenstehende Abbildung 5.12 zeigt die anhand der Trainingsdaten aus dem letzten Experiment geschätzte Transformationsmatrix. Diese weist eine ausgeprägte Diagonalstruktur auf, wobei die Koeffizienten auf der Diagonale alle unterhalb von 1 liegen. Wie in Kapitel 3.1 bereits erwähnt, deutet dies darauf hin, dass der dominante Anteil an Strukturen in der anderen Modalität ähnlich aussieht und sich lediglich in der Stärke der Kontraste unterscheiden, diese also mit mehr Rauschen und unschärfer vorliegen. Da mit einem linearen Modell im Allgemeinen nicht alle Fälle eines komplexeren multimodalen Zusammenhangs abgedeckt werden können, unterliegt das Verfahren gerade in Verbindung mit möglichen Kontrastinversionen und fehlenden Details stets einem gewissen Prädiktionsfehler. So stammen die darüber hinaus erkennbaren Muster in der Transformationsmatrix sowie die leicht hervortretenden Nebendiagonalen und teilweise auch negativen Einträge möglicherweise aus dem Versuch, aus vielschichtigen nichtlinearen Relationen ein einzelnes lineares Modell zu schätzen. In diesem Fall stellen alle Datenpunkte abseits einer vorherrschenden linearen Beziehung Ausreißer dar, welche die Schätzung entsprechend verschlechtern. Als Beispiele hierfür kommen unter anderem ortsvariantes Rauschen oder Verzerrungen in Betracht. In dieser Hinsicht ebenfalls problematisch sind im Falle der Fundusbildgebung Strukturen, welche auf der Photographie sehr ähnlich sind, sich beim Autofluoreszenzbild hingegen unterschiedlich verhalten können. Wie in Kapitel 3.1.3.3 angedeutet, haben *sparse dictionary*-Methoden ein größeres Potential zur Generalisierung und die Fähigkeit, solche nichtlinearen Ähnlichkeitsverhältnisse besser abzudecken.

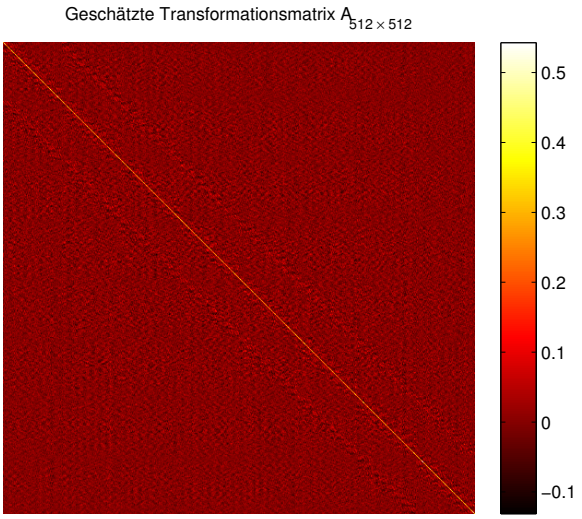


Abbildung 5.12: Visualisierung der geschätzten 512×512 Transformationsmatrix aus dem Least-Square-Verfahren

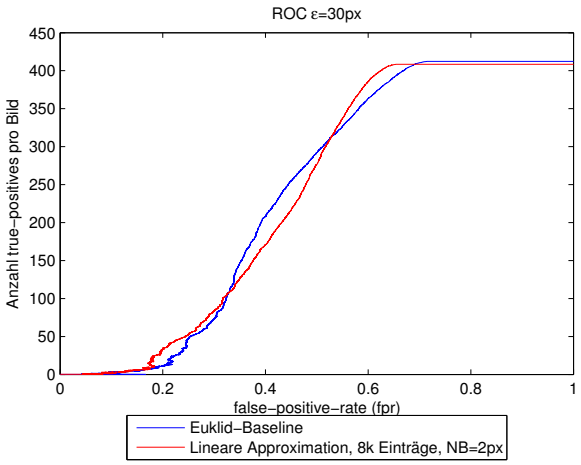


Abbildung 5.13: Vergleich der linearen Approximation zur Euklid-Baseline mit *matching ratio* 1.035

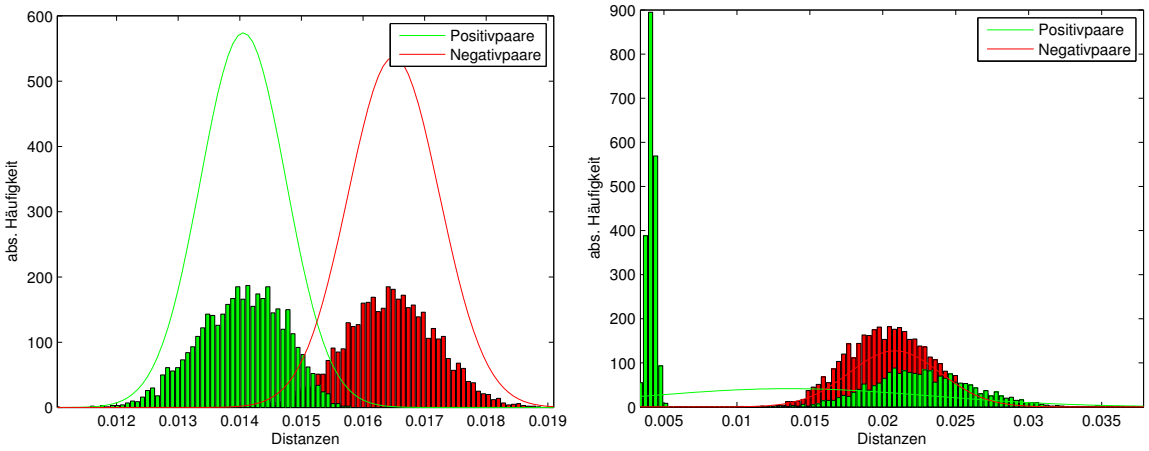
Dies liefert eine Erklärung für das Verhalten der ROC-Kurven der einfachen euklidischen Distanz im Vergleich zu der mit vorgeschalteter Transformation. Zwar ist das lineare Modell in der Lage, das Matching ausgewählter Korrespondenzpaare (der Inlier) zu verbessern; für andere Datenpunkte ist das Modell allerdings unpassend und über- bzw. unterschätzt Beziehungen zwischen den Elemente des Merkmalsdeskriptors. Dies hat zur Folge, dass das Ergebnis mit Transformation auf den Inliern diskriminativer ist, was sich in den früher ansteigenden Kurven bei hohen *matching ratios* zeigt. Dafür rücken aufgrund der abschwächenden Diagonalkoeffizienten die Distanzen auch aller anderen Korrespondenzmöglichkeiten aneinander, was ohne *matching ratio* zwar zu einer größeren Trefferanzahl führt; bei einem Verhältnis von 1.2 werden davon allerdings sehr viele als mehrdeutig klassifiziert und zurückgewiesen. Bei einer *matching ratio* von etwa 1.035 (Abbildung 5.13) kompensieren sich die genannten Effekte, was sich im sehr ähnlichen Verlauf der ROC-Kurven widerspiegelt.

5.3.2 Similarity-Sensitive Hashing

Einleitend soll zur Verifizierung der im Rahmen dieser Arbeit entstandenen Implementierung des Similarity-Sensitive Hashing ein Experiment mit synthetisch generierten Daten vorangestellt werden.

Synthetisch generierte Daten

Um die Fähigkeiten der applikationsspezifischen Metrik zu untersuchen werden Daten mit einer bimodalen Ähnlichkeitsrelation betrachtet. Dazu wird eine vorgegebene Anzahl positiver Trainingspaare generiert, indem 64-dimensionale Zufallsvektoren ihrer verrauschten Kopie zugeordnet werden, welche in allen Elementen um einen von zwei möglichen konstanten Offsets ± 50 verschoben ist. Beispielweise ist also der Wert 100 in der ersten Modalität ähnlich zu 50 und 150 in der anderen Modalität. Die Varianz des Rauschens wird zu $\sigma = 10$ gewählt, wobei die Vektorelemente ganzzahlige Werte im Bereich $[0, 255]$ annehmen können. Als Negativbeispiele werden Paare von unabhängigen Zufallsvektoren herangezogen. Nach dem Lernprozess können auf dieselbe Art Testdaten erzeugt werden, um die entstandene Metrik darauf zu bewerten.



(a) Similarity-Sensitive Hashing (500bit)

(b) Euklidische Distanz

Abbildung 5.14: Verteilung der Distanzen positiver und negativer Paare auf synthetisch generierten bimodalen Daten.

Abbildung 5.14 zeigt die Verteilung der resultierenden Distanzen zwischen den Testpaaren sowohl für die SSH- als auch die euklidische Distanz. Hier wird deutlich, dass die gelernte Distanz im Gegensatz zur ℓ_2-Distanz dazu in der Lage ist, die Daten zu trennen. Die euklidische Distanz hingegen kann erwartungsgemäß nur positiven Paaren, die (durch das Rauschen) nahe beieinanderliegen, eine geringe Distanz zuweisen. Insgesamt sind die Verteilungen aber zum einen nicht konsistent (die zweite Mode der Positivpaare hat höhere Distanzen als die der Negativpaare) und weisen zum anderen eine hohe Überlappung auf. Daher ist die Verwendung der euklidischen Distanz als Ähnlichkeitsmetrik auf diesen Daten nicht sinnvoll. Die SSH-Distanz kann hingegen mit der Bimodalität umgehen und ordnet auch den Negativpaaren Distanzen in einem annehmbaren höheren Bereich zu. Dies legt nahe,

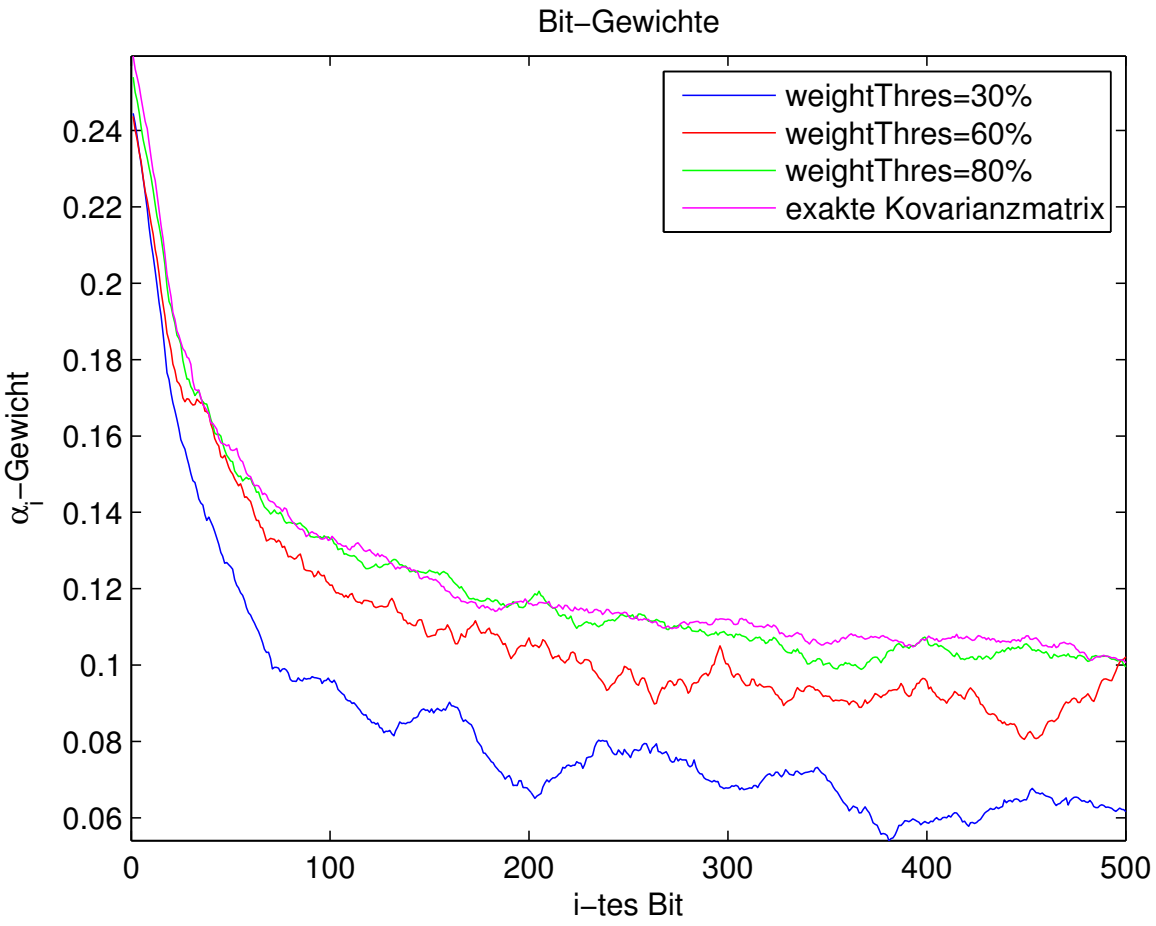


Abbildung 5.15: Die Abbildung zeigt den Verlauf der vergebenen Bitgewichte und damit die erreichte Klassifikatorgüte bei jedem Bit. Die Kurven sind dabei mit unterschiedlich genauen Approximationen der Kovarianzmatrix C aufgenommen. Für die Schwelle beim Abschneiden der Gewichtsanteile sind die Werte 30%, 60%, 80% und 100% (exakte Berechnung) angesetzt. Für die Klarheit der Darstellung wurde auf den Daten eine gleitende Glättung vorgenommen.

dass das Trainingsset (10^4 Positivpaare, 10^5 Negativpaare) für diesen Merkmalsraum ausreichend repräsentativ ist. Mit optimierten Parametern und einem längerem Bitstring dürfte sich die Diskriminanz weiter verbessern lassen. Dieses Ergebnis lässt die Annahme zu, dass die SSH-Implementierung korrekt arbeitet und für Experimente auf realen Bilddaten verwendet werden kann.

Parametereinflüsse

Um den Einfluss der Optimierungsparameter beim Similarity-Sensitive Hashing einzuschätzen, können die während des Lernprozesses vergebenen Bit-Gewichte α_i als Maß für die Güte der gefundenen Projektionen herangezogen werden. Je höher die Gewichte, desto sicherer ist die spätere Klassifikation bei gleicher Bitanzahl.

Zu Beginn soll die in Kapitel 3.2 erwähnte Verwendung einer approximierten Kovarianzmatrix C und deren Einfluss auf die Optimierung untersucht werden. Dazu werden bei an-

sonsten konstanten Parametern und gleichen Trainingsdaten sowohl mit dem exakten als auch mit unterschiedlich genau approximierten Ansätzen für C die Metriken gelernt und der Verlauf der α_i-Gewichte verglichen. Abbildung 5.15 zeigt auf, dass die Qualität der Projektionen nach einigen Iterationen stärker nachlässt, je mehr Gewichtsanteil im Zuge des *weight trimming* nicht miteinbezogen wird. Da bei allen Ansätzen für die erste Iteration C exakt bestimmt wird und der Gewichtsverlauf zur Unterdrückung hochfrequenter Oszillationen für eine klarere Darstellung einer Glättung unterzogen wurde, werden die ersten 32 Bits hier nicht betrachtet. Im weiteren Verlauf reduzieren sich die Bitgewichte bei einer eher groben Approximation mit einer Schwelle von $weightThres = 30\%$ (blaue Kurve) etwa auf die Hälfte der Gewichte aus der exakten Bestimmung der Kovarianzmatrix C (magentafarbene Kurve). Dies bedeutet, dass für dieselbe Klassifikationsgüte die doppelte Anzahl an Bits nötig ist. Dafür benötigt die exakte Berechnung mit $200\,\mathrm{s}$ etwa die zehnfache Rechenzeit. Einen guten Kompromiss bietet eine Schwelle von etwa 80% (grün eingezeichnet), welche nur unwesentlich schlechtere Ergebnisse liefert als die exakte Variante. Die Rechenzeit pro Bit (siehe Tabelle 5.1) hängt hierbei stark von der Ungleichverteilung der Gewichte ab und liegt daher zu Beginn bei ca. $100\,\mathrm{s}$; im Laufe der ersten 200 Iterationen reduziert sich diese mit etwa $21\,\mathrm{s}$ auf das Niveau der 30%-Schwelle. Der anfängliche Mehraufwand trägt damit gerade bei größeren Bitanzahlen ($n > 2500$) kaum zur Gesamtdauer bei.

Als nächster Parameter soll die Anzahl N der Zufallsvektoren analysiert werden. Abbildung 5.16 zeigt die Auswirkung einer Variation von N nach oben und unten auf die resultierenden Bitgewichte. Alle anderen Parameter sowie die Trainingsdaten sind hierbei konstant belassen. Auch hier lässt sich der generelle Trend beobachten, dass gelernte Projektionen mit größerem $N = 32$ im Gegensatz zu denen mit kleinerem $N = 2$ zu höheren Güte (etwa 12% Verbesserung) neigen. Der nur sehr geringe Unterschied zwischen $N = 16$ und $N = 32$ von durchschnittlich 2% deutet auf eine schon eintretende Sättigung hin. Dies ist ab einem bestimmten Wert von N zwangsläufig der Fall, da das Optimum im Sinne der Trennbarkeit mit einer ansonsten konstanten Parameterkonfiguration stets irgendwann erreicht ist und das Durchtesten weiterer Linearkombinationen darüber hinaus nicht weiter nützlich ist. Im Hinblick auf die benötigte Rechenzeit pro Bit (Tabelle 5.1) bietet sich ein Wert von $N = 8$ als guter Kompromiss an.

Tabelle 5.1: Rechenzeiten pro Bit in Sekunden beim Similarity-Sensitive Hashing für Variationen in den Optimierungsparametern $weightThres$ und N von der Standardkonfiguration (Tabelle 3.1). Die Zeiten wurden nach einer „Einschwingphase“ von 200 Iterationen gemessen.

$weightThres$	
30%	$19\,\mathrm{s}$
60%	$20\,\mathrm{s}$
80%	$21\,\mathrm{s}$
100%	$200\,\mathrm{s}$

N	
2	$10\,\mathrm{s}$
8	$20\,\mathrm{s}$
16	$38\,\mathrm{s}$
32	$70\,\mathrm{s}$

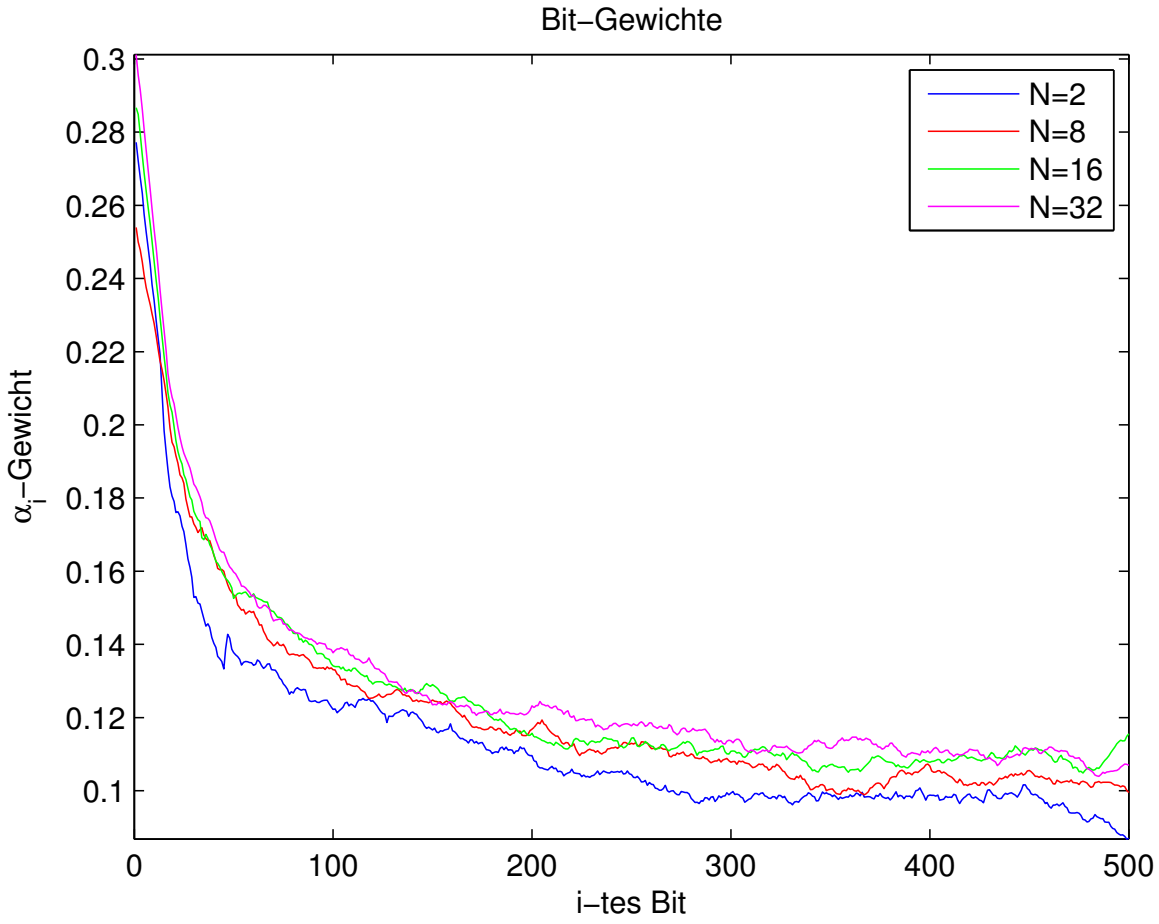


Abbildung 5.16: Die Abbildung zeigt den Verlauf der vergebenen Bitgewichte und damit die erreichte Klassifikatorgüte bei jedem Bit in Abhängigkeit der Anzahl N an zu testenden Linearkombinationen. Den dargestellten Kurven liegt jeweils N mit den Werten 2, 8, 16 und 32 zugrunde. Für die Klarheit der Darstellung wurde auf den Daten eine gleitende Glättung vorgenommen.

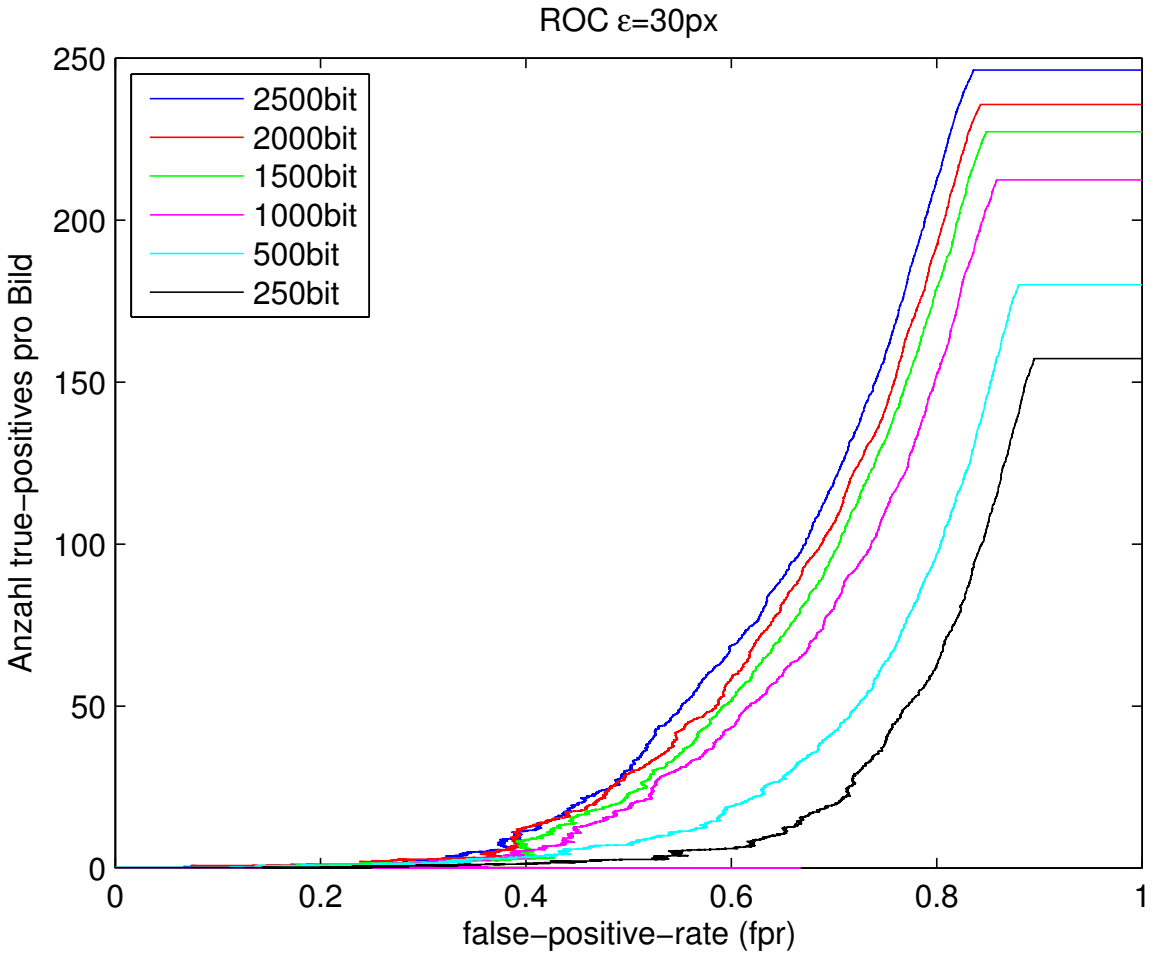


Abbildung 5.17: Die Abbildung zeigt die resultierenden ROC-Plots der gelernten Similarity-Sensitive Hashing Metriken. Die Kurve repräsentiert hierbei Metriken auf Basis von Bitstrings der Länge 250, 500, 1000, 1500, 2000 sowie 2500.

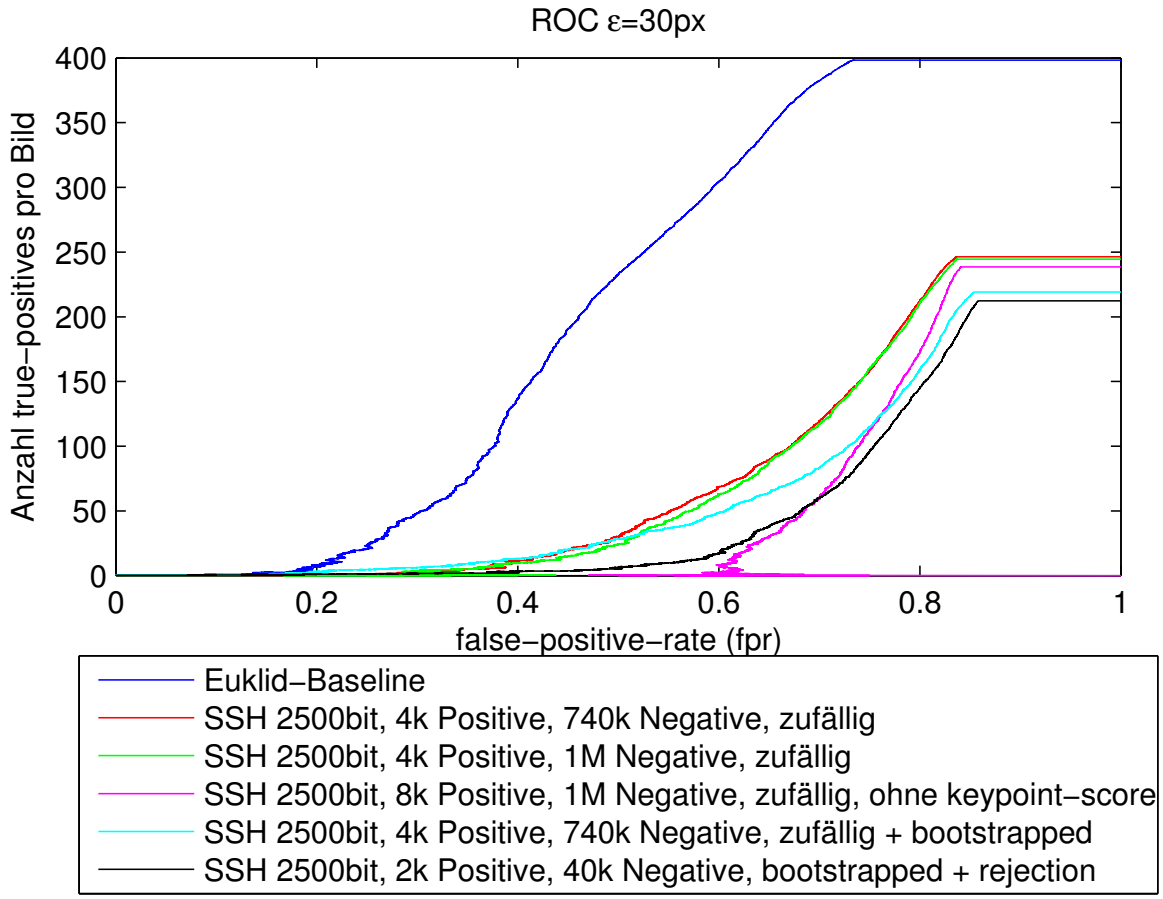


Abbildung 5.18: ROC-Kurven von SSH-Metriken auf Basis verschiedener Trainingsdaten.

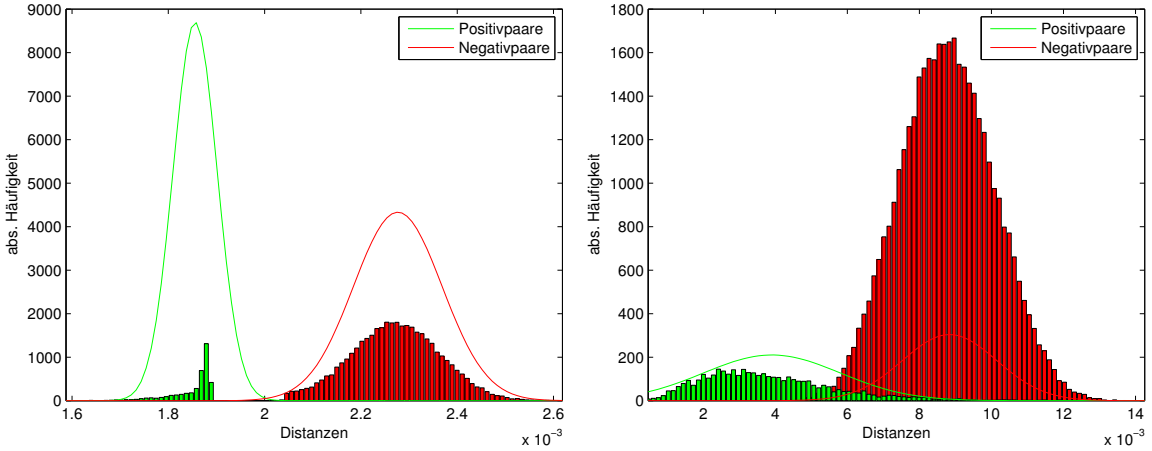
Zum Abschluss der Parameteruntersuchungen soll der Einfluss der Länge des Bitstrings n und damit der Anzahl der Klassifikatoren eingeschätzt werden. Dazu wird die mittels Standardkonfiguration (Tabelle 3.1) gelernte Metrik hergenommen und Testläufe mit schrittweiser reduzierter Anzahl an Bits durchgeführt. Die resultierenden ROC-Kurven sind auf Abbildung 5.17 dargestellt. Erwartungsgemäß verhält sich der Matcher mit steigender Bitstringlänge besser, was an der stetigen Verschiebung der Kurve nach oben links erkennbar ist. Die Kurvenform bleibt dabei sehr ähnlich. Vergleicht man die beiden Randfälle (250 vs. 2500 Bits), so werden bei einer durchschnittlich 20% weniger Fehlerquote etwa 100 Treffer mehr gefunden. Auch hier ist ein Sättigungseffekt an der immer näher beieinanderliegenden Kurven der hohen Bit-Anzahlen zu beobachten. Da zudem der Trainingsprozess mit 2500 Bit bei ca. $8 \cdot 10^6$ Trainingspaaren auf dem Testrechner etwa $13\,\mathrm{h}$ in Anspruch nimmt, wurde auf eine weitere Erhöhung von n verzichtet.

Ergebnisse mit verschiedenen Trainingsdaten

Wie in Kapitel 4 angesprochen hat die Auswahl der für das Erlernen der SSH-Metrik verwendeten Trainingdaten einen starken Einfluss auf das Ergebnis. Abbildung 5.18 zeigt die resultierenden ROC-Kurven für Matcher auf Basis von SSH-Metriken, die mit verschiedenen Konfigurationen von Trainingsdaten gelernt wurden. Die blaue Kurve repräsentiert das Ergebnis des Euklidschen Matchers. Die Kurven aller untersuchten SSH-Metriken zeigen

einen deutlich späteren und flacheren Anstieg und liegen auf in der Gesamtanzahl korrekter gefundener Korrespondenzen deutlich unter der Euklid-Baseline. Die Metrik auf Basis von 4000 positiven Trainingspaaren und ca. $8 \cdot 10^5$ zufällig zugeordneten Negativpaaren (rote Kurve) zeigt beste Verhalten bei der SSH-Methode. Eine Erhöhung der Anzahl an negativen Trainingspaaren auf etwa 10^6 wirkt sich kaum aus (grüne Kurve). Fügt man alle verfügbaren Positivpaare unabhängig vom *keypoint score* hinzu, verschlechtert sich durch die nachlassende Qualität der Trainingsdaten das Ergebnis deutlich (magentafarbene Kurve). Alle weiteren untersuchten Methoden zur Auswahl von Trainingsdaten (siehe Kapitel 4) wie Bootstrapping (cyanfarbene Kurve), Bootstrapping mit Zurückweisung zu allgemeingültiger Positivpaare und Negativpaaren, die eine geringe euklidische Distanz aufweisen (schwarze Kurve) und weitere untersuchte Kombinationen (nicht abgebildet) davon führen ebenfalls zu schlechteren Ergebnissen.

Visualisierung und Analyse der Trennfähigkeit

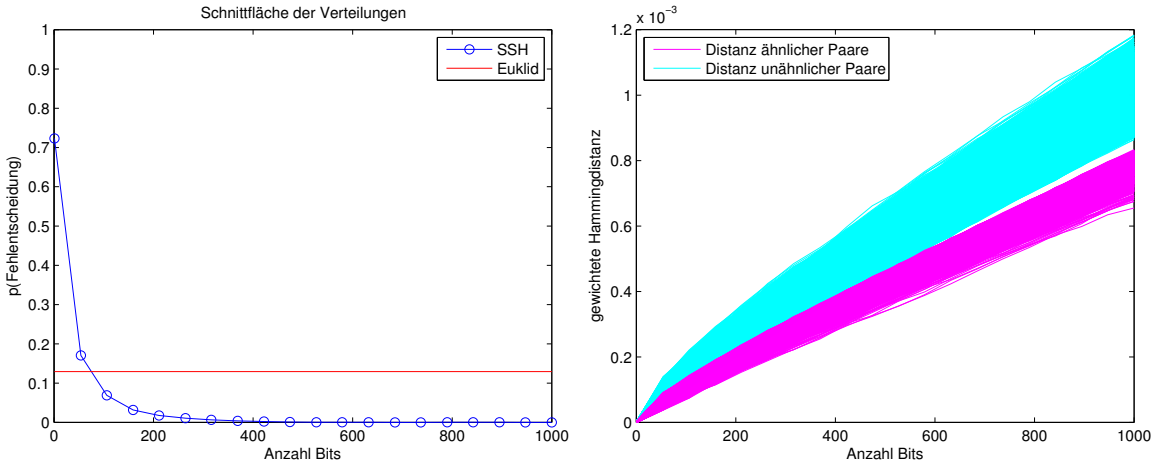


(a) Similarity-Sensitive Hashing (2500bit) (b) Euklidische Distanz

Abbildung 5.19: Verteilung der Distanzen positiver und negativer Paare auf „Fundus I"-Trainingsdaten.

Um die mit der am besten funktionierenden SSH-Metrik erreichte Diskriminanz zu beurteilen, sind auf Abbildung 5.19 die Verteilungen der Distanzen positiver und negativer Trainingspaare sowohl für die SSH-Metrik (links), als auch für die euklidische Distanz (rechts) aufgetragen. Dabei ist erkennbar, dass das Similarity-Sensitive Hashing in dieser Konfiguration sehr gut in der Lage ist, die vorliegenden Trainingsdaten (4000 Positive, $8 \cdot 10^6$ Negative auf Basis zufälliger Zuordnungen) gemäß ihrem Label (niedrige Distanz für ähnliche Paare, hohe Distanz für unähnliche) zu trennen. Für eine klarere Darstellung ist auf dem Plot nur eine zufällig gezogene Untermenge von 40000 Negativpaaren abgebildet. Die auf denselben Wertebereich normierte euklidische Distanz weist hierbei eine wesentlich höhere Varianz sowohl bei den Positiv- als auch bei den Negativpaaren auf und stellt damit aufgrund der starken Überlappung die weitaus weniger robuste Metrik dar.

Betrachtet man das Verhalten der SSH-Metrik in Abhängigkeit der Anzahl verwendeter Bits (Abbildung 5.20) so fällt auf, dass diese bereits nach weniger als 100 Bits eine kleinere



(a) Verlauf der Überschneidungsfläche

(b) Verlauf der Distanzen

Abbildung 5.20: Die Abbildung zeigt den Verlauf der Schnittfläche der Distanzverteilungen positiver und negativer Trainingspaare (links) sowie eine Visualisierung der Entwicklung der Distanzen (rechts) mit steigender Anzahl an Trennfunktionen (Bits).

Schnittfläche der Verteilungen und damit eine geringer Wahrscheinlichkeit der Fehlklassifikation [9, S. 41] im Vergleich zur euklidschen Distanz aufweist. Ab etwa 500 Bits geht diese gegen 0, was einer perfekten Trennung entspricht. Abbildung 5.20 (b) zeigt auf, dass die Distanzen der unähnlich gelabelten Trainingspaare schneller ansteigt als die der ähnlichen, sodass die Trennung deutlich sichtbar wird. Obwohl diese Diskriminanz bereits mit unter 1000 Bits erreicht wird, profitiert der Matcher bei der Registrierung in gewissen Grenzen dennoch von einer größeren Anzahl an Trennfunktionen, wie auf Abbildung 5.17 zu sehen ist.

Da sich die überlegene Trennfähigkeit nicht in den Resultaten des Matchings widerspiegelt, liegt die Vermutung nahe, dass die verwendeten Trainingsdaten nicht repräsentativ genug für diese Anwendung sind. Daher soll an dieser Stelle das Verhalten der SSH-Metrik bei der Korrespondenzfindung während des Registriervorgangs genauer untersucht werden.

Die Visualisierung der über die SSH-Distanz zugeordneten Keypoints mit sehr niedrigen Distanzen (Abbildung 5.21) zeigt, dass die gelernte Metrik Bildausschnitte durchaus auf sinnvolle Art in Korrespondenz setzt. Die schwache Performance der SSH-Metriken deutet darauf hin, dass die gelernte Distanz zu tolerant in Bezug auf Bildbereiche mit ähnlichen dominanten Strukturen agiert. Unterschiede in feineren Details wie dünne Adern (Abbildung 5.21, unten rechts) wirken sich kaum auf die Distanz aus. Auffällig ist auch hier, dass vor allem Keypoints kleinerer Skalen mit wenig Umgebungskontext von den Verwechslungen betroffen sind.

Betrachtet man die Korrespondenzfindung eines zufällig ausgewählten Keypoints im Detail (Abbildung 5.22) so fällt auf, dass die Verhältnisse der Distanzen zwischen bestem und nachfolgender Treffer bei der euklidischen Distanz sich viel weiter von 1 entfernen und die Zuordnung daher diskriminativer und damit deutlich sicherer ist. Die SSH-Distanz hingegen ordnet vielen Keypoints der anderen Modalität ein sehr ähnliches Abstandsmaß auf

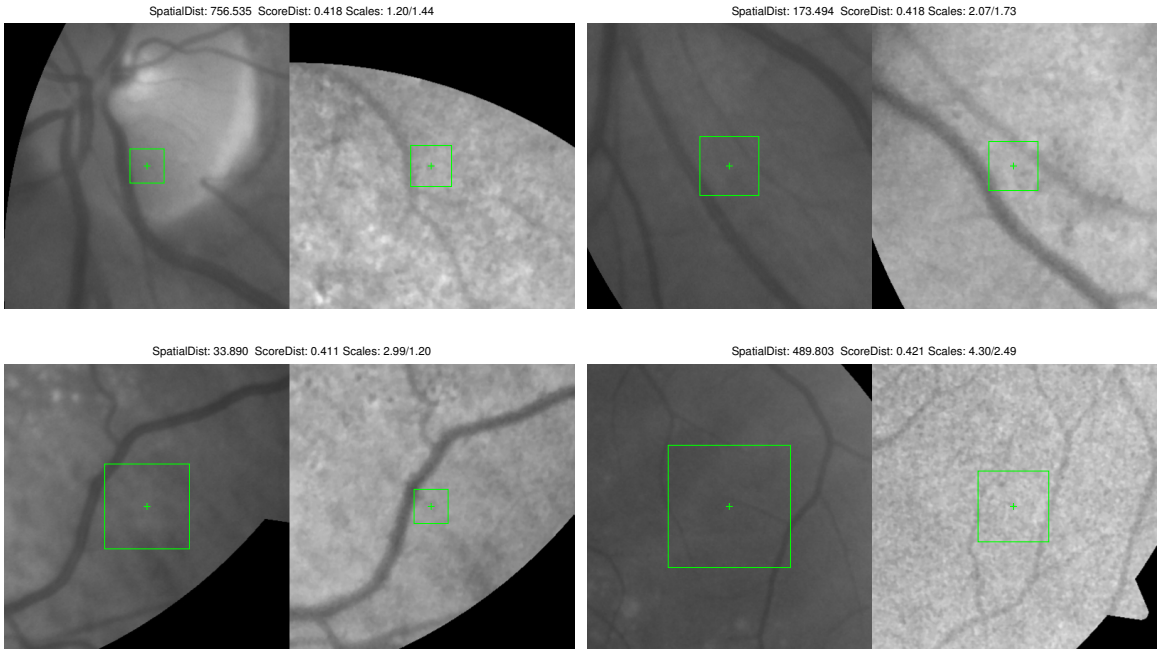


Abbildung 5.21: Die Abbildung zeigt vier Bildausschnitte (Patches), auf denen die Deskriptorumgebungen der Keypoints der vier Falschzuordnungen mit den kleinsten Distanzen dargestellt sind. Dabei gibt *SpatialDist* die örtliche- und *ScoreDist* die Merkmalsdistanz an. Die grün eingezeichnete Umrandung beinhaltet den jeweils Bereich, auf dem der SIFT-Deskriptor berechnet wird. Als Schwellwert wird $\epsilon = 30$px angenommen.

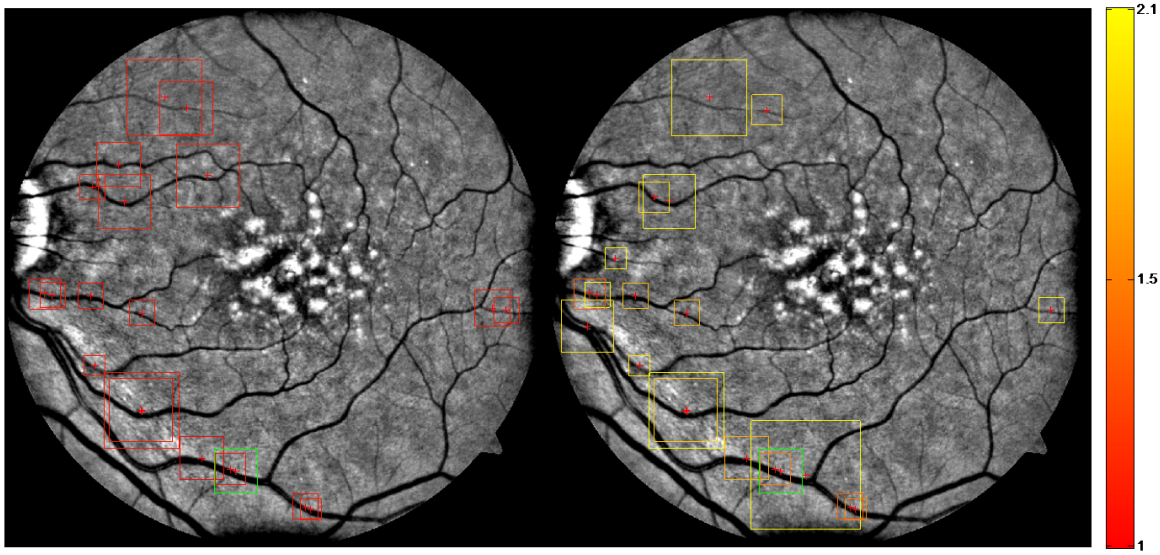


Abbildung 5.22: Die Abbildung visualisiert 20 der Kandidaten mit niedrigsten Distanzen für die Korrespondenzfindung eines zufällig gewählten Keypoints, an dem sowohl die gelernte SSH-Metrik (links), als auch die euklidsche Distanz (rechts) die korrekte Zuordnung (grün eingezeichnet) liefern. Dazu sind die Bildauschnitte der zugehörigen Deskriptoren farblich entsprechend dem Verhältnis der Distanzen zum besten Treffer markiert.

niedrigem Niveau zu und weist keinen ausgeprägten Peak bei der korrekten Korrespondenz auf. Aufgrund dessen führt selbst die Anwendung sehr kleiner *matching ratios* zur Zurückweisung eines Großteils der Korrespondenzen, was die darauf aufbauende Schätzung der geometrischen Transformation entsprechend verschlechtert und daher nicht wünschenswert ist.

Abbildung 5.23 zeigt auf, dass dieses Verhalten keine Ausnahmeerscheinung ist, sondern einem generellen Trend entspricht. Zeichnet man die Distanzverhältnisse der beiden Metriken über eine größere Stichprobe von korrekt zugeordneten Keypoints gemittelt auf, so wird deutlich, dass die ℓ_2-Distanz zum besten Treffer hin wesentlich spitzer zuläuft. An dieser Stelle wird klar, dass sich hierbei der korrekte Treffer besser von den anderen Zuordnungsmöglichkeiten absetzen kann. Die gelernte Metrik des Similarity-Sensitive Hashing hingegen neigt dazu, Ähnlichkeit zu tolerant zu sehen und ist in dieser Hinsicht für diese Anwendung bei keypoint-basierten Registrierverfahren weniger geeignet. Denn dieses Verhalten ist gerade im Zusammenspiel mit ähnlichen oder wiederkehrenden Strukturen im Bild bei der Korrespondenzfindung problematisch.

In Anbetracht dieser Erkenntnisse lässt sich zusammenfassend feststellen, dass es mit keiner der untersuchten Methoden möglich ist, eine Trainingsdatenkonfiguration zu erstellen, mit der eine darauf gelernte SSH-Metrik die für die Korrespondenzfindung nötige Diskriminanz erreicht. Die Ursachen hierfür liegen zum einen in der begrenzten Anzahl qualitativ hochwertiger Positivbeispiele und zum anderen in der fehlenden Methodik, negative Trainingsbeispiele auszuwählen, die sowohl relevant, als auch tatsächlich unähnlich sind. Wie in Abbildung5.24 skizziert, kommt es daher entweder zu dem Fall, dass der Toleranzbereich aufgrund zu spärlicher und wenig relevanter Negativpaare zu groß wird, oder andererseits durch eine Bootstrapping-Variante falsch gelabelte Negativbeispiele im Trainingsprozess zu Widersprüchen und Oszillationen führen, die den Ähnlichkeitsbereich zu stark einschränken. Eine Methode, die beide Fälle vermeidet und eine bessere Verteilung der Trainingspaare im Merkmalsraum erreicht, konnte im Rahmen dieser Arbeit nicht gefunden werden.

5.3.3 Random Forest Distance

Aufgrund in Kapitel 3.3.3.2 erläuterten Einschränkungen bei der experimentellen Untersuchung des Matchers auf Basis der Random Forest Distance werden im Folgendem lediglich die Ergebnisse zweier RFD-Matcher vorgestellt, welche mit unterschiedlichen Trainingsdaten trainiert wurden.

Abbildung 5.25 zeigt die ROC-Kurven der jeweils mit $T = 100$ Bäumen trainierten Metriken im Vergleich zur Euklid-Baseline. Die Metrik mit zufällig gezogenen Negativenpaaren (rot) weist dabei im Hinblick auf den Kurvenverlauf keine signifikanten Unterschiede zu der mit auf Basis von Bootstrapping generierten Trainingsdaten (grün) auf. Beide verlaufen deutlich unterhalb der Euklid-Baseline bei etwa einer über 60% höheren *false-positive-rate*. Ein Grund für die schlechte Performance der RFD-Metriken liegt sicherlich in der geringen *forest size* T. Im Hinblick auf das Similarity-Sensitive Hashing (Abbildung 5.17) entspricht das Resultat etwa dem einer 250 Bit SSH-Metrik. Die Diskriminanz der Verfahrens und da-

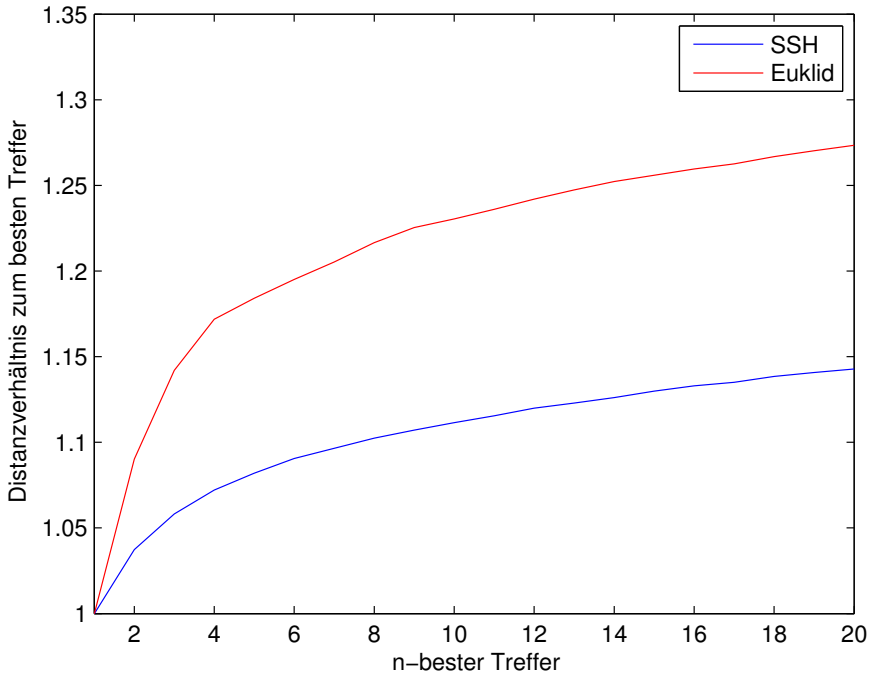


Abbildung 5.23: Anhand des Verhältnisses der Distanzen zwischen dem besten und nachfolgenden Treffern lässt sich einschätzen, wie sehr sich die gefundene von allen anderen möglichen Zuordnungen absetzt. Die Werte sind dabei über einer zufällig gezogene Stichprobe von 1000 Keypoints des Datensatzes „Fundus I“ gemittelt, welche von beiden Metriken korrekt zugeordnet werden konnten. Anhand der Steilheit der Kurve zum besten Treffer hin lässt sich die Diskriminanz der Metriken beim Registrierprozess vergleichen.

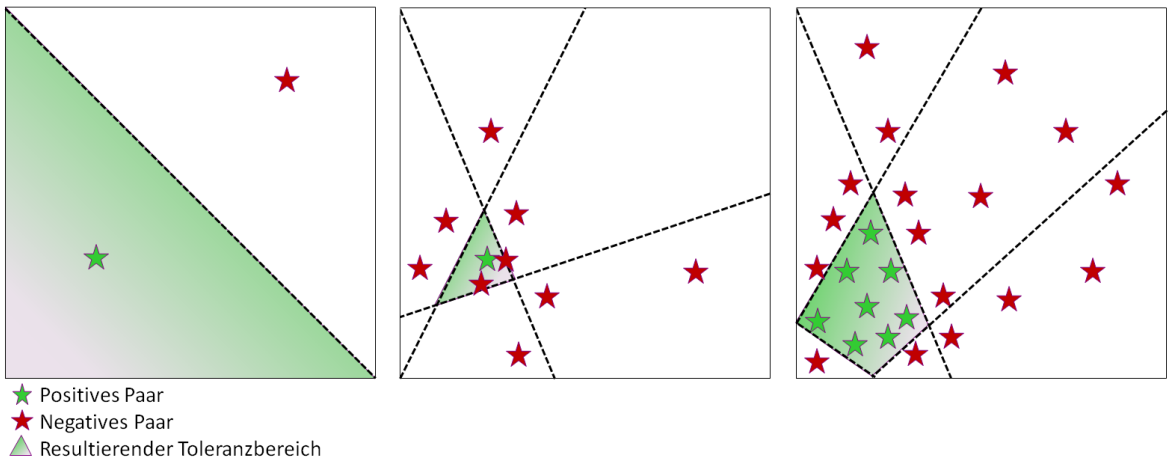


Abbildung 5.24: Eine vereinfachte Skizze der in Abhängigkeit der verfügbaren Trainingspaare gelernten Trennfunktionen zeigt, wie sowohl zufällig gezogene und damit wenig relevante (links) als auch per Bootstrapping hinzugefügte und möglicherweise inkorrekt gelabelte Negativpaare (Mitte) Einfluss nehmen auf den resultierenden Ähnlichkeits- bzw. Toleranzbereich im Merkmalsraum. Optimale Trainingsdaten sollten den Grenzbereich der Ähnlichkeit gut abdecken und damit sowohl relevant sein als auch in größerer Anzahl zur Verfügung stehen (rechts).

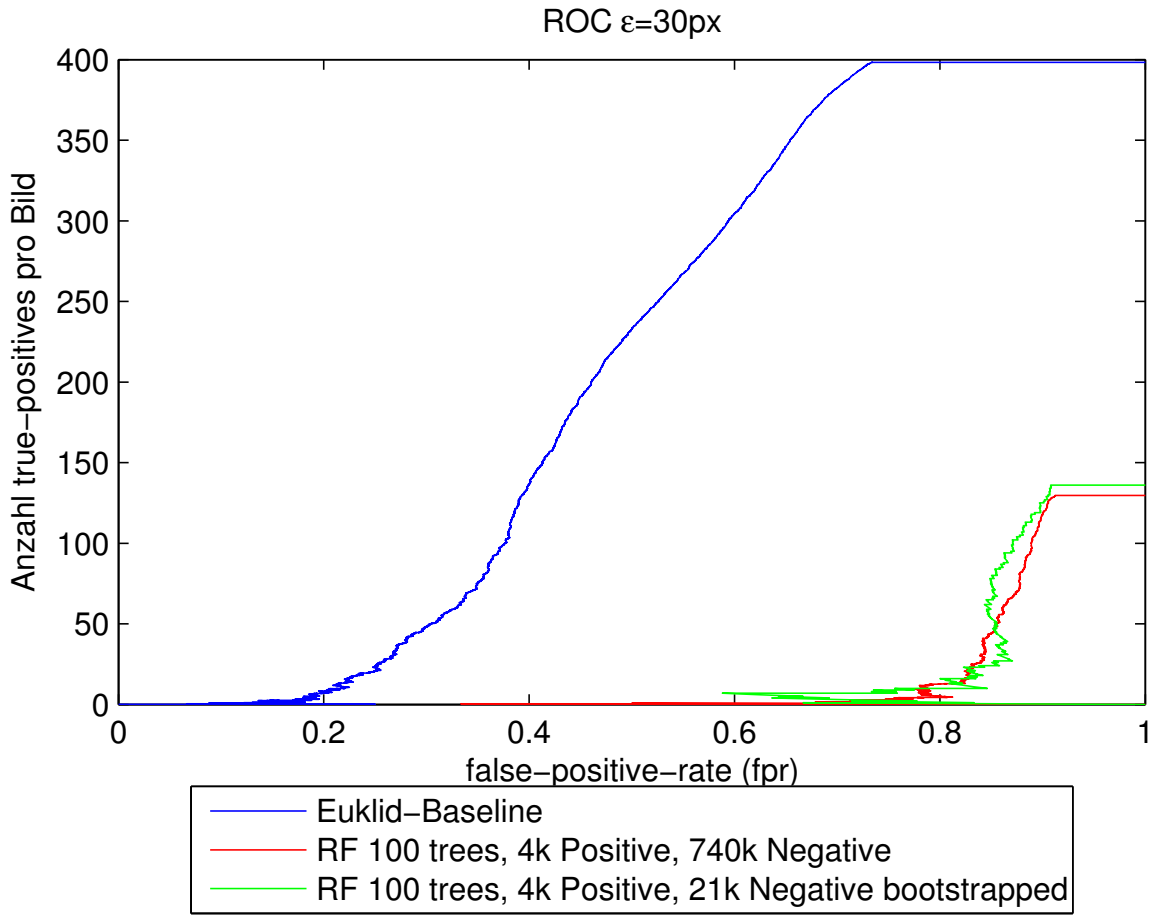


Abbildung 5.25: ROC-Kurven von RFD-Metriken auf Basis verschiedener Trainingsdaten.

mit die Klassifikationsleistung der RFD-Metrik dürfte sich demnach mit steigender *forest size* an die des SSH-Verfahrens heranbringen lassen. Dabei schränkt der für die Korrespondenzfindung benötigte Zeitaufwand die sowohl die Experimente im Rahmen dieser Arbeit als auch die Tauglichkeit des Verfahrens stark ein. Daher wurde auf weitere Experimente mit größerer Anzahl an Bäumen verzichtet.

Eine erste Analyse der falschen Zuordnungen mit niedrigsten Distanzen bei der RFD-Metrik (Abbildung 5.26) deutet auf dieselben Probleme hin, welches im letzten Abschnitt im Zusammenhang mit dem Similarity-Sensitve Hashing diskutiert wurden. Eine Ähnlichkeit der Korrespondenzen ist zwar vorhanden, was für die prinzipielle Funktionsweise des Verfahrens spricht, allerdings ist auch die RFD-Metrik in dieser Hinsicht zu ähnlichkeitstolerant, um ein für keypoint-basierte Registrierverfahren ein ausreichend diskriminatives Ergebnis zu liefern. Zusammen mit den in Kapitel 3.3.3.2 dargelegten Analogien zum Similarity-Sensitive Hashing lässt sich sagen, dass die Ursachenanalyse im vorigen Abschnitt auch auf die Korrespondenzfindung mittels Random Forest Distance zutrifft und von einer weiteren Untersuchung an dieser Stelle keine darüber hinausgehenden Erkenntnisse zu erwarten sind.

Nach diesen Ergebnissen der RFD-Metriken auf Basis der im Rahmen dieser Arbeit verwendeten Trainingsdaten ist demnach eine Tauglichkeit des Verfahrens für den Anwendungsfall

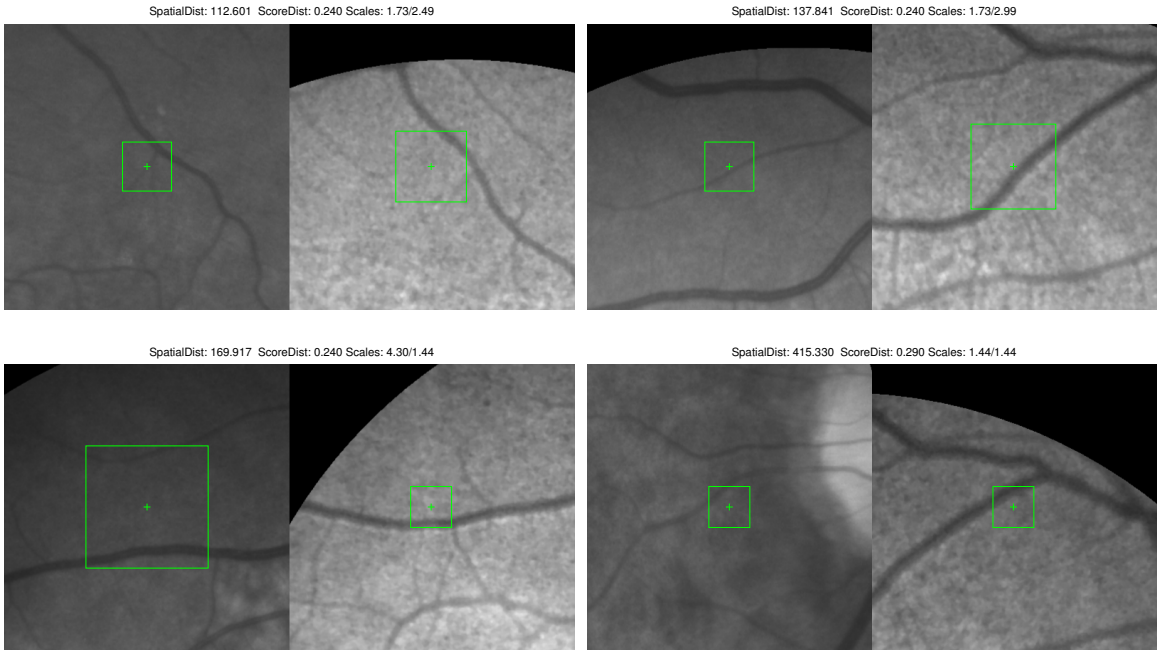


Abbildung 5.26: Die Abbildung zeigt vier Bildausschnitte (Patches), auf denen die Deskriptorumgebungen der Keypoints der vier Falschzuordnungen mit den kleinsten Distanzen dargestellt sind. Dabei gibt *SpatialDist* die örtliche- und *ScoreDist* die Merkmalsdistanz an. Die grün eingezeichnete Umrandung beinhaltet den jeweils Bereich, auf dem der SIFT-Deskriptor berechnet wird. Als Schwellwert wird $\epsilon = 30$px angenommen.

keypoint-basierter Registrierung nicht gegeben. Zudem stellt der benötigte Rechenaufwand einen gravierenden Nachteil gegenüber der anderen untersuchten Verfahren dar.

5.3.4 Image Analogies

Abschließend sollen die Ergebnisse des Image Analogies Verfahrens anhand von Funduskopie- und korrelativen Mikroskopiebildern vorgestellt und die Methode auf ihre Tauglichkeit für keypoint-basierte Registrierverfahren in Abhängigkeit des Bildmaterials untersucht werden.

Fundusbildgebung

Die Trainingsbildpaare des Datensatzes „Fundus II“ für den Image Analogies Filter entstanden durch die manuelle Selektion von relevanten Bildauschnitten und deren Rotation und Skalierung (siehe Kapitel 4.2). Als Richtung für den Filterprozess wurde die Synthese der Photographie aus dem FAF-Bild gewählt. Abbildung 5.27 veranschaulicht die Gründe für diese Entscheidung: bei der Generierung des Autofluoreszenzbildes (rechts unten) entsteht tendenziell deutlich mehr und sehr grobkörniges Rauschen, welches zum Teil stark auf vorhandene Strukturen übergreift. Dem liegt zu Grunde, dass der Filter nur schwer lernen kann, welche Bildelemente in welchem Kontext verrauscht bzw. in gestörter oder verzerrter Form sichtbar sind. Dagegen führt die Filterung in der anderen Richtung (links unten) zu einer

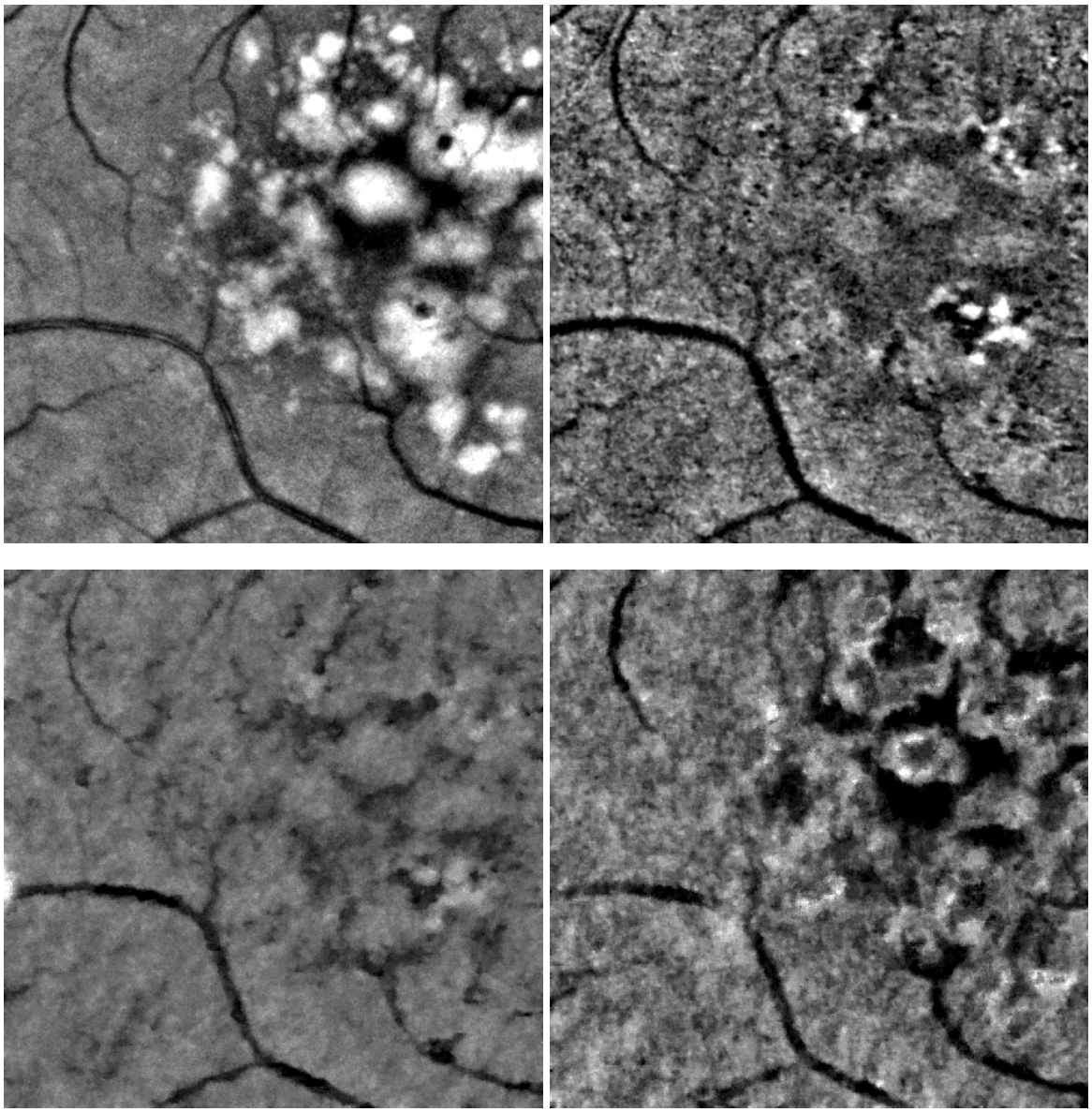


Abbildung 5.27: Die obere Reihe zeigt Ausschnitte der Originalbilder aus dem Datensatz „Fundus II“ beider Modalitäten. Je nach Filterungsrichtung kann aus diesen jeweils die Photographie (links unten) oder das Autofluoreszenzbild (rechts unten) über den Image Analogies Filter synthetisiert werden. Quell- und Zielbild sind hier jeweils diagonal angeordnet.

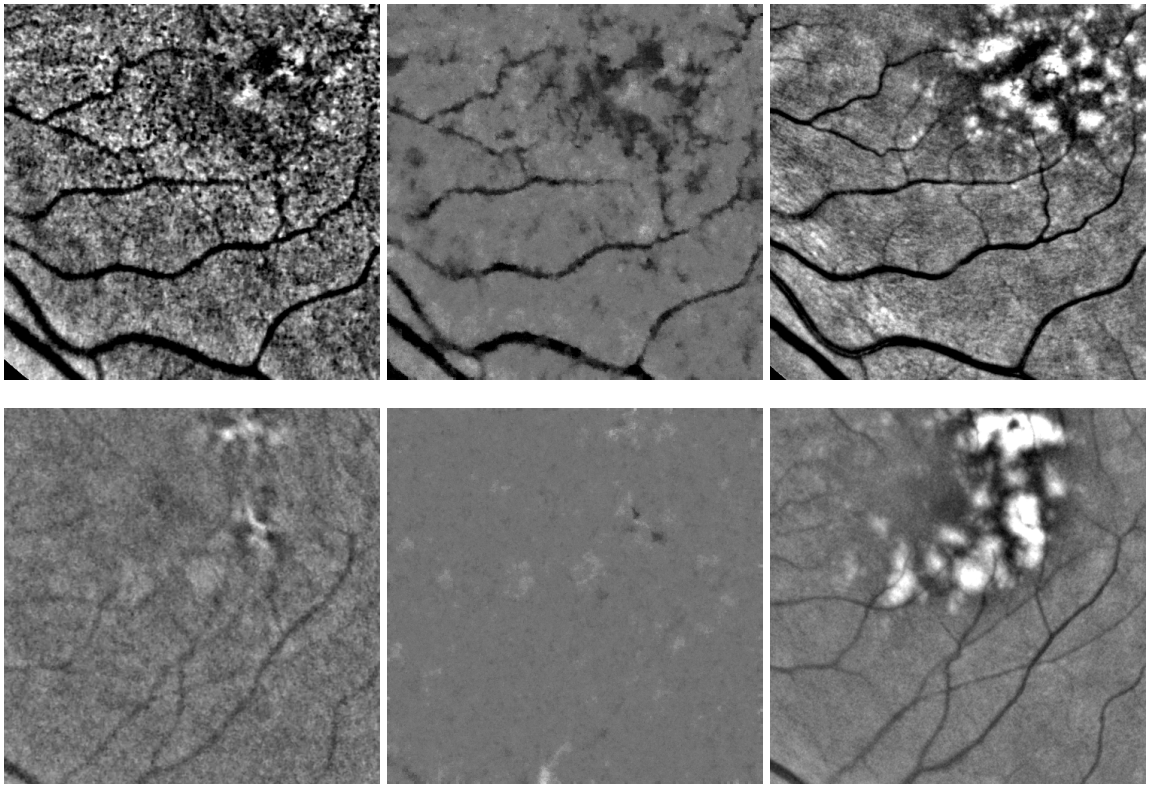


Abbildung 5.28: Zwei Beispiele für die Resultate der IA-Filterung aus dem „Fundus II“-Datensatz: Die erste Spalte zeigt das Autofluoreszenz-Quellbild, die zweite das Ergebnis der Filterung und die letzte Spalte die darauf zu registrierende Photographie.

Glättung, da die Photographiebilder eine höhere Bildqualität und wenig Rauschen aufweisen. Bei beiden Filterprozessen verliert das Bild deutlich an Schärfe und Strukturen gehen zum Teil verloren oder bleiben in Form von falsch synthetisierten Artefakten bestehen, obwohl diese in der anderen Modalität nicht vorkommen. Insgesamt macht die Synthese der Photographie (links unten) einen geeigneteren Eindruck, da viele irrelevante Details durch Glättung entfernt werden und dominante Strukturen übrig bleiben, über die eine Registrierung möglich sein sollte.

Um einen visuellen Eindruck der gefilterten Bilder zu Gewinnen, sind auf Abbildung5.28 zwei weitere Beispiele zusammen mit den zugehörigen Originalbildern beider Modalität abgebildet. Dabei wird deutlich, dass das Ergebnis der Synthese je nach Bild eine sehr unterschiedliche Güte aufweisen kann. Der in der ersten Zeile gezeigte Bildausschnitt enthält dominante Adern, die auch im Autofluoreszenzbild (links) sehr scharf abgebildet sind. Diese kann der IA-Filter einigermaßen gut synthetisieren, wobei Schärfe und Kontrastintensität stark von der lokalen Dicke und Ausrichtung der Adern abhängen. Dünnere Strukturen wie auch das grobkörnige Rauschen fallen dabei eher der Glättung zum Opfer. Die Bildqualität nimmt insgesamt stark ab. Dass manche Strukturen nicht vorhanden sind, kann zum einen daran liegen, dass betroffene Stellen im Fluoreszenzbild zu stark verrauscht sind oder zum anderen, dass diese von den Trainingsdaten nicht abgedeckt sind. Das zweite Beispiel (untere Zeile) demonstriert, dass der so gelernte Filter bei Quellbildern mit stärkerem, aber feinkör-

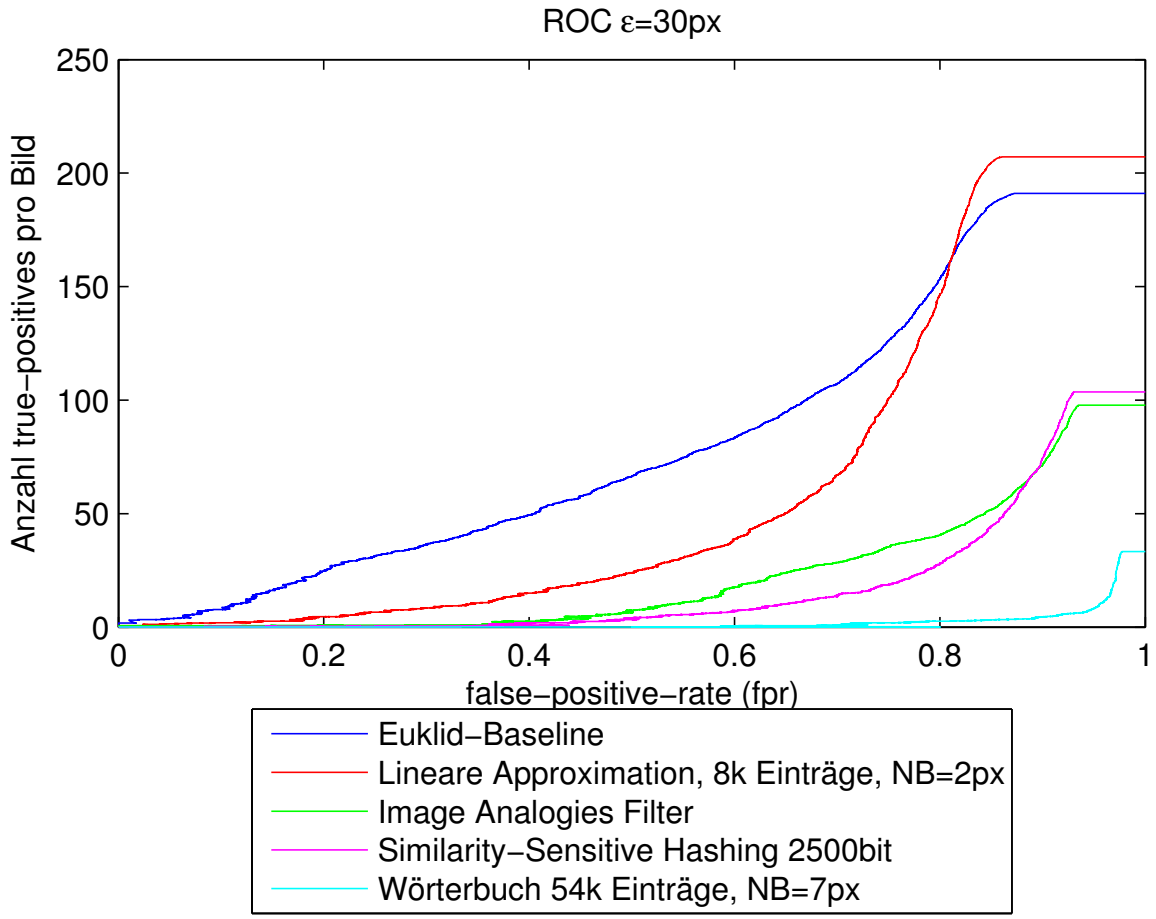
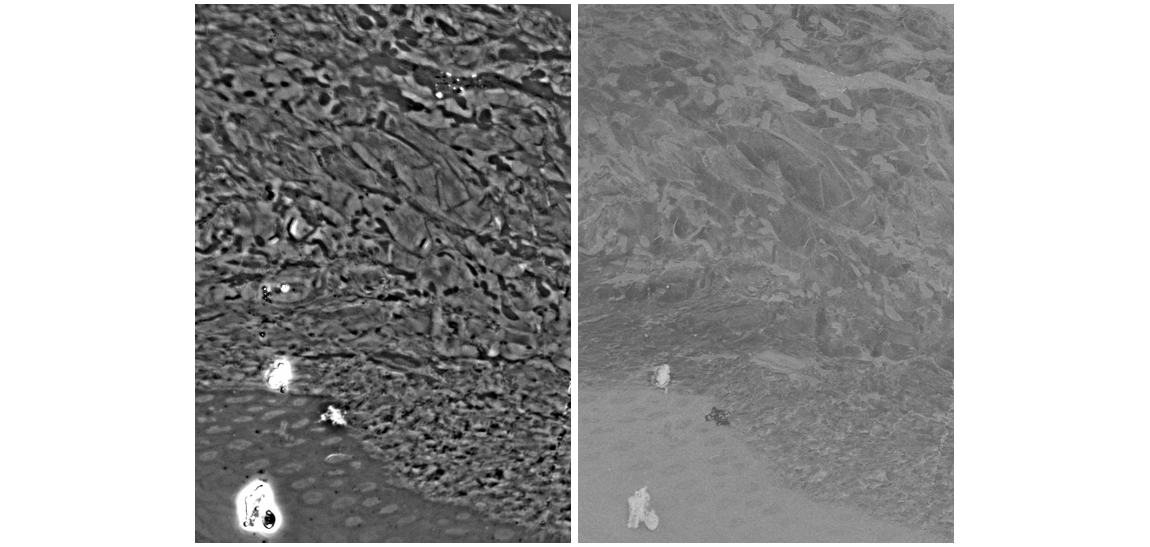


Abbildung 5.29: ROC-Plot der untersuchten Verfahren auf dem „Fundus II“-Datensatz.

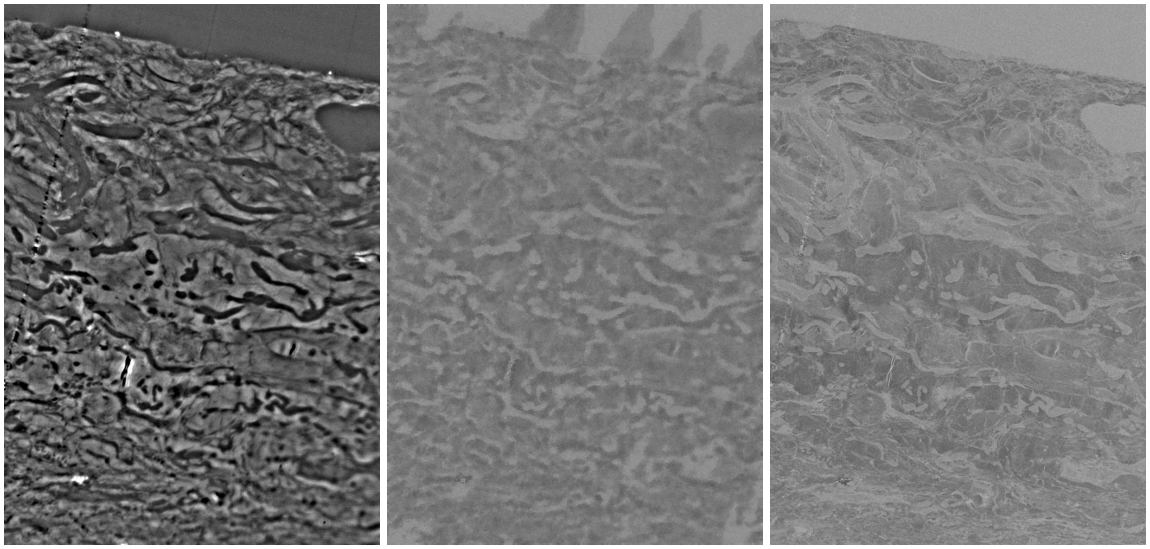
nigem Rauschanteil, vorwiegend feinerer Strukturen mit wenig Kontrast und einer etwas anderen Beleuchtungssituation nicht mehr in der Lage ist, charakteristische Bildelemente richtig zu synthetisieren. Es ist hier lediglich ein fleckiger Hintergrund erkennbar. Für eine Registrieranwendung ist ein solches Resultat untauglich.

Die starke Abhängigkeit der Ergebnisse des Image Analogies Filters von der Größe, Ausrichtung und Intensität der im Quellbild vorhandenen Strukturen legt den Schluss nahe, dass die eingeschränkte Anzahl und die damit einhergehende begrenzte Repräsentativität der Trainingsbildpaare in Verbindung mit der schlechten Fähigkeit des Verfahrens zur Generalisierung den Flaschenhals für die Tauglichkeit darstellt. Die Qualität der IA-Filters dürfte sich daher - eine entsprechende Güte vorausgesetzt - mit steigender Anzahl an Trainingsdaten deutlich verbessern lassen. Aufgrund der in Kapitel 3.4 genannten Limitierungen ist weitergehende Untersuchung mit der verwendeten Implementierung an dieser Stelle nicht möglich.

Dieses Verhalten bei der Bildsynthese spiegelt sich auch in den quantitativen Evaluationsergebnissen wider. Abbildung 5.29 zeigt zusammenfassend die Resultate aller im Rahmen dieser Arbeit untersuchten lern-basierten Verfahren (außer RFD) auf dem „Fundus II“-Datensatz bei deaktiviertem *matching ratio*. Die ungenügende Qualität der Bildsynthese beim IA-Filter (grüne) verschlechtert die Korrespondenzfindung im Vergleich zur Euklid-Baseline (blau) deutlich sowohl im Hinblick auf die Fehlerraten als auch auf die Gesamtan-



(a) (b)



(c) (d) (e)

Abbildung 5.30: Image Analogies Filter bei Aufnahmen von Hautgewebe mittels korrelativer Mikroskopie. Das Bildpaar (a) und (b) in der ersten Zeile diente als Trainingsanalogie für den gelernten Filter, der anschließend auf ein anderes Bild derselben Modalität (c) angewendet wurde. Das Filterergebnis (d) dient als Schätzung des Erscheinungsbildes in der anderen Modalität, welches mit dem Originalbild (e) unimodal registriert werden kann.

zahl korrekter gefundener Treffer. Das IA-Verfahren liegt hier etwa gleich auf mit dem Similarity-Sensitive Hashing (magentafarben). Ähnlich wie auf dem Datensatz „Fundus I“ erreicht die lineare Approximation einige Treffer mehr als der ℓ_2-Matcher, das Wörterbuch selbst (cyanfarben) schneidet am schlechtesten ab.

Korrelative Mikroskopie

Um zu die Fähigkeiten des Image Analogies Verfahrens bei ausreichend repräsentativen Trainingsdaten zu demonstrieren, soll an dieser Stelle ein Experiment mit Bilddaten aus der

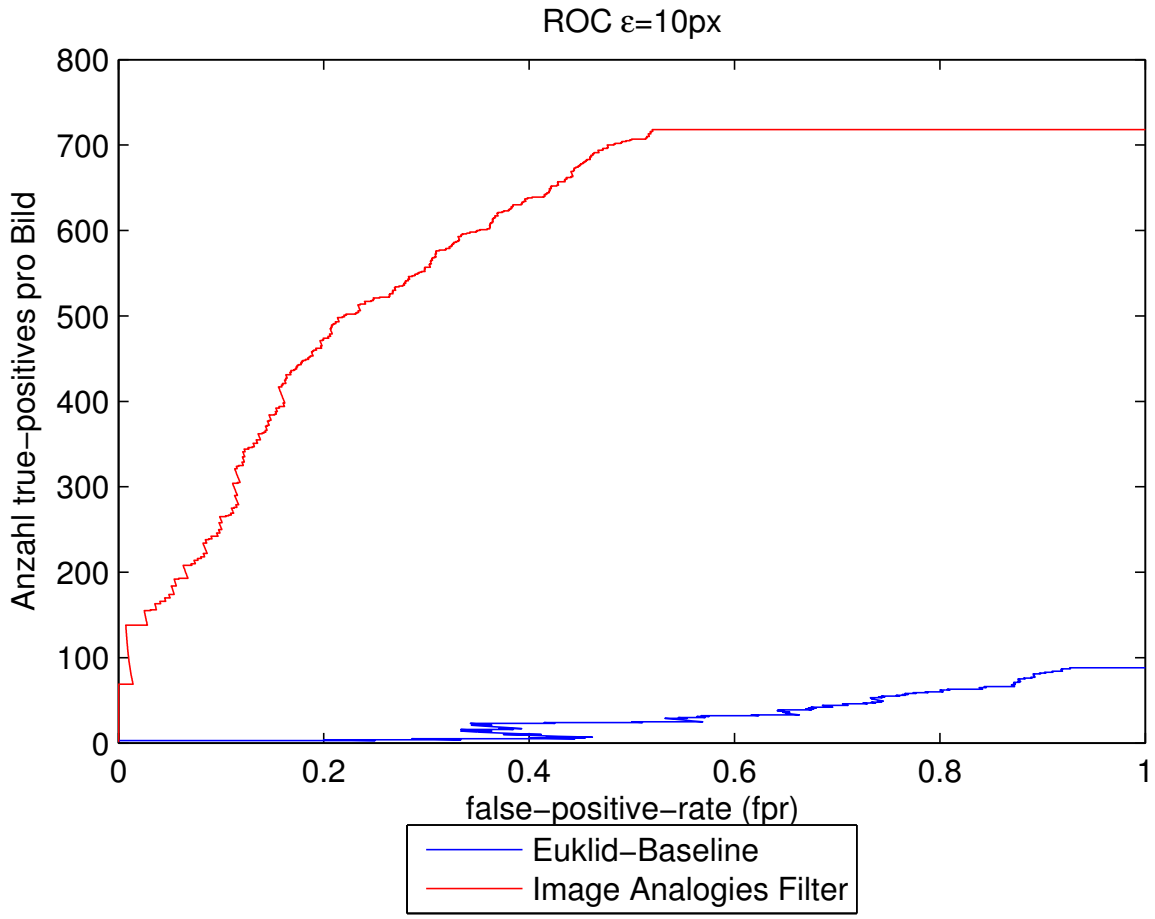


Abbildung 5.31: ROC-Plot zum Vergleich der Resultate der Korrespondenzfindung auf der Hautaufnahme (Abbildung 5.30) mit (rot) und ohne (blau) Untersützung durch den gelernten Image Analogies Filter.

korrelativen Mikroskopie vorgestellt werden. Trainings- und Testbild stammen aus derselben Aufnahme, die vertikal auseinander geschnitten wurde. So ist eine hohe Ähnlichkeit der Strukturen auf beiden Bildern sichergestellt. Die Bildpaare zeigen zwei verschiedene mikroskopische Aufnahmen von menschlicher Haut. Abbildung 5.30 zeigt auf, dass der IA-Filter charakteristische Strukturen sehr zuverlässig wiedergeben kann und das synthetisierte Bild (d) gut mit der Aufnahme in der anderen Modalität (e) übereinstimmt. Lediglich flächige Elemente mit konstanten Intensitäten wie im oberen Bildbereich sind von schwerwiegenderen Artefakten betroffen, da diese im Trainingsbildpaar nicht vorkommen.

Auch anhand der zugehörigen ROC-Kurven in Abbildung 5.31 wird deutlich, dass ein Matching über die konventionelle euklidische Distanz in diesem multimodalen Fall scheitert, da eine Ähnlichkeit über ℓ_2-Distanzen von SIFT-Deskriptoren hier nicht gegeben ist. Der unimodale Registrierprozess profitiert daher enorm von der Transformation des Bildes in die zweite Modalität durch den IA-Filter (rote Kurve). Statt knapp 100 werden hierbei über 700 korrekte Zuordnungen registriert, bei deutlich früherem und steilerem Anstieg der Kurve. Insbesondere werden die besten 70 Treffer beim IA-gefiltertem Bild werden beinahe ohne Fehlzuordnungen gefunden.

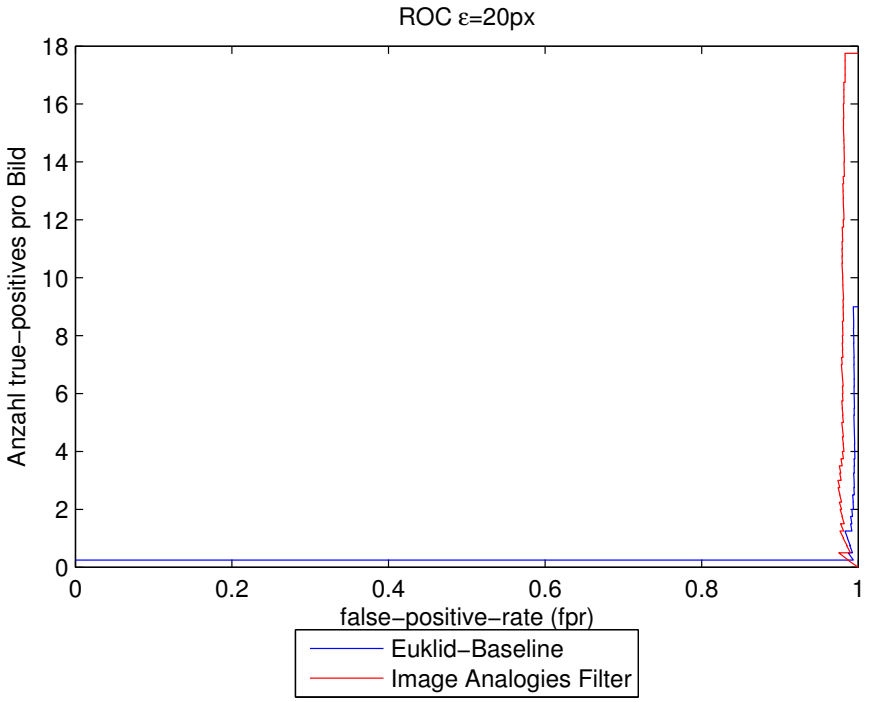


Abbildung 5.32: ROC-Plot zum Vergleich der Resultate der Korrespondenzfindung auf der Zellgewebeaufnahme (Abbildung 5.30) mit (rot) und ohne (blau) Untersützung durch den gelernten Image Analogies Filter.

Abschließend soll an dieser Stelle die Tauglichkeit des Image Analogies Verfahrens zusammen mit keypoint-gestützter Registrierung für multimodale TEM- und Konfokalmikroskopischen Aufnahmen (Abbildung 3.4) untersucht werden. Die Bilddaten stammen hierbei von Cao et. al [15], welche die IA-Methode in Verbindung mit direkten Registrierverfahren vorschlagen. Abbildung 5.32 zeigt auf, dass die Korrespondenzfindung mit euklidischem Matcher (blau) sowohl allein, als auch mit Anwendung des IA-Filters mit knapp unter 100% *false positive rate* keine brauchbaren Ergebnisse liefert. Im ersten Fall liegt dies erwartungsgemäß an den sehr unterschiedlichen Erscheinungsbildern der abgebildeten Zellstrukturen in den beiden Modalitäten. Die Filterung mittels Image Analogies verdoppelt zwar in etwa die Anzahl der korrekten gefundenen Korrespondenzen bei ca. 3% geringerer Fehlerrate, es kann hier allerdings nicht von einem robustem Verfahren gesprochen werden. Abbildung 5.33 zeigt die Ursachen dieses Verhaltens anhand der synthetisierten Bilder auf. Zwar werden lokal, insbesondere beim ersten (obere Zeile) und dritten (untere Zeile) Beispiel, Strukturen wie Zellwände einigermaßen korrekt synthetisiert; insgesamt tauchen allerdings vor allem an flächigen Bildelementen an vielen Stellen Artefakte auf. Dies verletzt die grundlegende Annahme keypoint-gestützter Registrierverfahren, dass an korrespondierenden Stellen der Bilder 2D-Struktur und damit Keypoints vorliegen, die einander zugeordnet werden können. Ein bildbasiertes Registrierverfahren verhält sich dagegen in einem solchen Fall robuster [15], da sich der Fehler global auf das gesamte Bild verteilt und beim Minimum einer entsprechend gewählten Kostenfunktion durchaus die korrekte Registrierung vorliegen kann. Zudem sind direkte Verfahren weniger auf kontrastreiche Strukturen bzw. scharfe Kanten angewiesen, was sich in diesem Fall als vorteilhaft erweist. Beim zweiten Beispielbild (Abbildung 5.33,

mittlere Zeile) wird erneut die fehlende Repräsentativität der Trainingsbilder deutlich, da der Großteil der vorkommenden Strukturen bei der Filterung verloren geht.

Zusammenfassend lassen diese Ergebnisse den Schluss zu, dass bei einem geeigneten Anwendungsbereich mit wenig Variation der Strukturen und Aufnahmebedingungen in den multimodalen Daten der Image Analogies Ansatz eine Registrierung über ein koventionelles unimodales Verfahren ermöglichen bzw. deutlich verbessern kann. Sind die Trainingsdaten nicht repräsentativ genug, beispielsweise aufgrund schwankender Beleuchtungs- und Rauschverhältnisse oder einer Vielzahl unbekannter (nicht trainierter) vorkommender Strukturen, so stellten sich Image Analogies im Rahmen dieser Untersuchungen als untauglich bzw. sogar kontraproduktiv zur Unterstützung keypoint-basierter Registrierverfahren heraus.

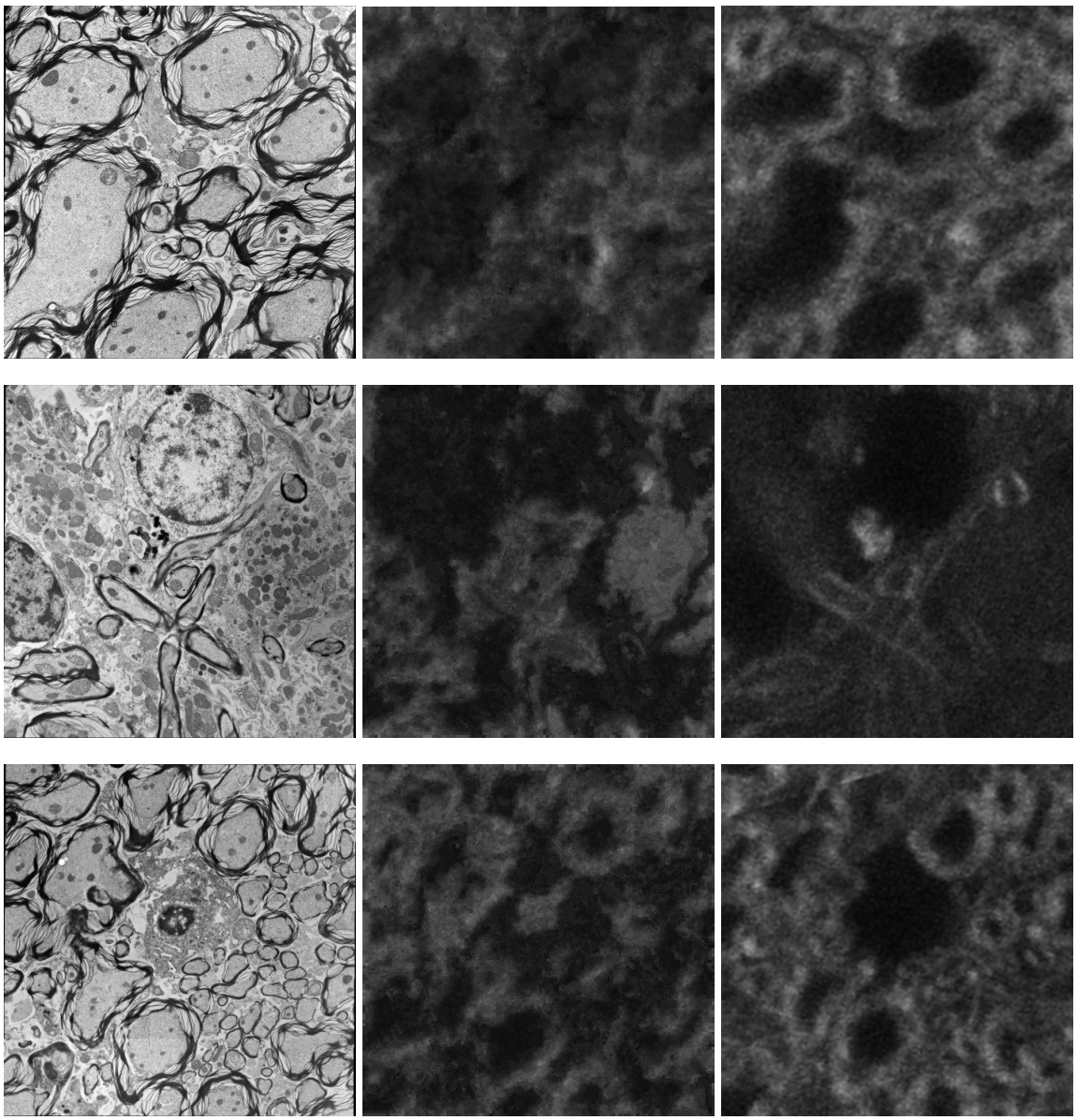


Abbildung 5.33: Drei Resultate des Image Analogies Filters bei Bilddaten aus der korrelativen Mikroskopie. Die erste Spalte zeigt jeweils das Bild des Transmissionselektronenmikroskops (TEM), die zweite das über den IA-Filter berechnete Konfokalmikroskopbild und die dritte das tatsächlich aufgenommene, darauf zu registrierende Bild. Bildmaterial aus [15].

Kapitel 6

Ausblick

Abschließend soll ein kurzer Ausblick über Erweiterungen der untersuchten Verfahren sowie weitere mögliche Vorgehensweisen zum Thema „lern-basierte multimodale Registrierung“ gegeben werden.

Eine essentielle Grundlage des im Rahmen dieser Arbeit verwendeten keypoint-gestützten Registrierframeworks stellt der verwendete Merkmalsdeskriptor dar. Der SIFT-Deskriptor besitzt einige geometrische und intensitätsbezogene Invarianzen, die je nach Art multimodaler Bilddaten vorteilhaft oder aber von Nachteil sein können. So können problemspezifische Deskriptoren [29, 37, 36] zuverlässigere (diskriminativere) Beschreibungen extrahieren, die für Ähnlichkeit charakteristisch sind. Neben Gradienteninformation können dafür auch zusätzliche Merkmale wie lokale Entropie oder Textur [44, 43] in Frage kommen. Ein darauf aufbauendes Lernverfahren könnte demnach ein robusteres Ähnlichkeitsmaß erlernen. Darüber hinaus ist es denkbar, zusätzliche modalitätsspezifische Informationen in die Ähnlichkeitsbetrachtung miteinzubeziehen, um eine diskriminativere Metrik zu erhalten. So könnten beispielsweise bei Funduskopiedaten statt der Verwendung einer Grauwertprojektion und gleicher Deskriptoren für beide Modalitäten auf dem Photographiebild farbabhängige Deskriptoren [51] berechnet werden und diese über eine SSH-Metrik mit Standard-SIFT Deskriptoren vom Autofluoreszenzbild in Korrespondenz gesetzt werden.

Eine alternative Möglichkeit des Trainings bzw. der Trainingsdatenauswahl für hochdimensionale Deskriptoren besteht darin, die Lernverfahren auf einem durch eine Dimensionreduktion verkleinertem Merkmalsraum durchzuführen. Dies kann beispielweise durch eine Hauptkomponentenanalyse (PCA) [36] oder die Zerlegung eines Deskriptors nach den örtlichen *bins* erfolgen. Auf dem resultierenden niedrigdimensionalen Merkmalsraum könnten die Trainingsdaten repräsentativer sein, was sich positiv auf die gelernte Metrik auswirkt.

Ist eine aussagekräftige applikationsspezifische Ähnlichkeitsmetrik auf Basis geeigneter Deskriptoren gefunden, so kann durch die Einbettung in ein iteratives keypoint-basierte Registrierframework wie „Dual Bootstrap“ [66] (Kapitel 2.3) ein hoher Grad an Robustheit und Registriergenauigkeit erreicht werden.

Die Experimente mit Wörterbuchverfahren haben gezeigt, dass vor allem die begrenzte Anzahl an Einträgen die Leistungsfähigkeit stark einschränkt. Wie in Kapitel 3.1 angedeutet,

kann der einfache Wörterbuchansatz gegebenenfalls durch die Verwendung von *sparse coding* Verfahren in Bezug auf Generalisierbarkeit und Skalierbarkeit signifikant verbessert werden. Entsprechende Methoden wurden von Mairal et al. [42] sowie Cao et al. [14] vorgeschlagen und können im Hinblick auf multimodale, keypoint-basierte Registrierung untersucht werden.

Beim Image Analogies Verfahren kann gegebenenfalls eine größere Anzahl an Korrespondenzen und damit eine bessere Schätzung des geometrischen Transformationsmodells erreicht werden, indem IA-Filter in beide Richtungen gelernt werden und die Registrierung anschließend auf beiden Bildpaaren durchgeführt wird. So können die gefundenen Korrespondenzen beider Suchrichtungen in eine einzelne, robustere Gesamtschätzung eingehen.

Dass der Image Analogies Ansatz vor allem wegen der schlechten Generalisierbarkeit nur in Grenzen für Registrieranwendungen tauglich ist, liegt zum größten Teil an dem sehr einfachen Inferenzmodell, welches die zu synthetisierenden Pixelintensitäten in Abhängigkeit der Nachbarschaften prädiziert. Durch den Einsatz eines entsprechend trainierten neuronales Netzwerks (*Gated Boltzmann machine*) mit zusätzlichen versteckten Variablen kann die statistische Beziehung zwischen Quell- und Zielbild erfasst werden (Memisevic et al. [45]). So kann das Modell unter Ausnutzung von Korrelationen und besser generalisieren und sich daher robuster gegenüber Variationen in den Nachbarschaften verhalten.

Literatur

[1] ASHBURNER, J. ; FRISTON, K.J.: Rigid body registration. In: FRACKOWIAK, R.S.J. (Hrsg.) ; FRISTON, K.J. (Hrsg.) ; FRITH, C. (Hrsg.) ; DOLAN, R. (Hrsg.) ; FRISTON, K.J. (Hrsg.) ; PRICE, C.J. (Hrsg.) ; ZEKI, S. (Hrsg.) ; ASHBURNER, J. (Hrsg.) ; PENNY, W.D. (Hrsg.): *Human Brain Function*. 2nd. Academic Press, 2003

[2] ASHIKHMIN, Michael: Synthesizing Natural Textures. In: *Proceedings of the 2001 Symposium on Interactive 3D Graphics*. New York, NY, USA : ACM, 2001 (I3D '01), 217–226

[3] BABENKO, Boris ; BRANSON, Steve ; BELONGIE, Serge: Similarity Metrics for Categorization: from Monolithic to Category Specific. In: *International Conference on Computer Vision (ICCV)*. Kyoto, Japan, 2009

[4] BANTERLE, Francesco ; ARTUSI, Alessandro ; DEBATTISTA, Kurt ; CHALMERS, Alan: *Advanced High Dynamic Range Imaging: Theory and Practice*. 1st. Natick, MA, USA : A. K. Peters, Ltd., 2011

[5] BAR-HILLEL, A. ; HERTZ, T. ; SHENTAL, N. ; WEINSHALL, D.: Learning Distance Functions using Equivalence Relations. In: *Proceedings of the International Conference on Machine Learning*, 2003

[6] BAR-HILLEL, Aharon ; HERTZ, Tomer ; SHENTAL, Noam ; WEINSHALL, Daphna: Learning a Mahalanobis Metric from Equivalence Constraints. In: *J. Mach. Learn. Res.* 6 (2005), Dezember, 937–965

[7] BELLET, Aurélien ; HABRARD, Amaury ; SEBBAN, Marc: A Survey on Metric Learning for Feature Vectors and Structured Data. In: *CoRR* abs/1306.6709 (2013)

[8] BENTLEY, Jon L.: Multidimensional Binary Search Trees Used for Associative Searching. In: *Commun. ACM* 18 (1975), September, Nr. 9, 509–517

[9] BISHOP, Christopher M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2006

[10] BODENSTEINER, Christoph ; HÜBNER, Wolfgang ; JÜNGLING, Kai ; MÜLLER, Jürgen ; ARENS, Michael: Local multi-modal image matching based on self-similarity. In: *ICIP*, IEEE, 2010, 937-940

[11] BRADSKI, G.: The opencv library. In: *Doctor Dobbs Journal* 25 (2000), Nr. 11, S. 120–126

[12] BREIMAN, Leo: Random Forests. In: *Mach. Learn.* 45 (2001), Oktober, Nr. 1, 5–32

[13] BRONSTEIN, Michael M. ; BRONSTEIN, Alexander M. ; MICHEL, Fabrice ; PARAGIOS, Nikos: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: *CVPR*, IEEE, 2010, 3594-3601

[14] CAO, T. ; JOJIC, V. ; MODLA, S. ; POWELL, D. ; CZYMMEK, K. ; NIETHAMMER, M: Robust Multimodal Dictionary Learning. In: *MICCAI*, 2013

[15] CAO, Tian ; ZACH, Christopher ; MODLA, Shannon ; POWELL, Debbie ; CZYMMEK, Kirk ; NIETHAMMER, Marc: Registration for Correlative Microscopy Using Image Analogies. In: *Proceedings of the 5th International Conference on Biomedical Image Registration*. Berlin, Heidelberg : Springer-Verlag, 2012 (WBIR'12), 296–306

[16] CARUANA, Rich ; NICULESCU-MIZIL, Alexandru: An Empirical Comparison of Supervised Learning Algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA : ACM, 2006 (ICML '06), 161–168

[17] CHUNG, Albert C. S. ; WELLS, William M. III ; NORBASH, Alexander ; GRIMSON, W. Eric L.: Multi-modal Image Registration by Minimising Kullback-Leibler Distance. In: *Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention-Part II*. London, UK, UK : Springer-Verlag, 2002 (MICCAI '02), 525–532

[18] CUTLER, Adele ; CUTLER, D. R. ; STEVENS, John R. ; CHA ZHANG, Yunqian M. (Hrsg.): *Ensemble Machine Learning: Methods and Applications*. Springer, 2012. – 157ff S.

[19] DAVIS, Jason V. ; KULIS, Brian ; JAIN, Prateek ; SRA, Suvrit ; DHILLON, Inderjit S.: Information-theoretic Metric Learning. In: *Proceedings of the 24th International Conference on Machine Learning*. New York, NY, USA : ACM, 2007 (ICML '07), 209–216

[20] DUDA, Richard O. ; HART, Peter E.: Use of the Hough transformation to detect lines and curves in pictures. In: *Commun. ACM* 15 (1972), Januar, Nr. 1, 11–15

[21] EFROS, Alexei A. ; FREEMAN, William T.: Image Quilting for Texture Synthesis and Transfer. In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA : ACM, 2001 (SIGGRAPH '01), 341–346

[22] ELLI, Alexandra F. ; THOMAS, Christian ; BÖKER, Christian ; WIEDERSPAHN, Markus: Korrelative Licht- und Elektronenmikroskopie (CLEM). In: *Optik & Photonik* 7 (2012), Nr. 1, 32–36

[23] FAWCETT, Tom: An Introduction to ROC Analysis. In: *Pattern Recogn. Lett.* 27 (2006), Juni, Nr. 8, 861–874

[24] FISCHLER, Martin A. ; BOLLES, Robert C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In: *Commun. ACM* 24 (1981), Juni, Nr. 6, 381–395

[25] FREUND, Yoav ; SCHAPIRE, Robert E.: *A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting*. 1997

[26] FRIEDMAN, Jerome H.: Greedy Function Approximation: A Gradient Boosting Machine. In: *Annals of Statistics* 29 (2000), S. 1189–1232

[27] FROME, Andrea ; SHA, Fei ; SINGER, Yoram ; MALIK, Jitendra: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: *In ICCV*, 2007

[28] FROME, Andrea ; SINGER, Yoram ; MALIK, Jitendra: Image Retrieval and Classification Using Local Distance Functions. In: SCHÖLKOPF, Bernhard (Hrsg.) ; PLATT, John (Hrsg.) ; HOFFMAN, Thomas (Hrsg.): *NIPS*, MIT Press, 2006, 417-424

[29] GHASSABI, Zeinab ; SEDAGHAT, Amin ; SHANBEHZADEH, Jamshid ; FATEMIZADEH, Emad: An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors. In: *EURASIP J. Image and Video Processing* 2013 (2013), 25

[30] GIONIS, Aristides ; INDYK, Piotr ; MOTWANI, Rajeev: Similarity Search in High Dimensions via Hashing. In: *Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1999 (VLDB '99), 518–529

[31] GLOCKER, Ben ; KOMODAKIS, Nikos ; TZIRITAS, Georgios ; NAVAB, Nassir ; PARAGIOS, Nikos: Dense image registration through MRFs and efficient linear programming. In: *Medical Image Analysis* 12 (2008), Nr. 6, 731-741

[32] HABER, Eldad ; MODERSITZKI, Jan: Intensity gradient based registration and fusion of multi-modal images. In: *Methods of Information in Medicine, Schattauer Verlag*, 2006, S. 726–733

[33] HERTZMANN, Aaron ; JACOBS, Charles E. ; OLIVER, Nuria ; CURLESS, Brian ; SALESIN, David H.: Image Analogies. In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA : ACM, 2001 (SIGGRAPH '01), 327–340

[34] HUANG, Jing ; YOU, Suya ; ZHAO, Jiaping: Multimodal image matching using self similarity. In: *AIPR*, IEEE, 2011, 1-6

[35] INDYK, Piotr ; MOTWANI, Rajeev: Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. New York, NY, USA : ACM, 1998 (STOC '98), 604–613

[36] KE, Yan ; SUKTHANKAR, Rahul: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: *CVPR (2)*, 2004, 506-513

[37] KELMAN, Avi ; SOFKA, Michal ; STEWART, Charles V.: Keypoint Descriptors for Matching Across Multiple Image Modalities and Non-linear Intensity Variations. In: *CVPR*, IEEE Computer Society, 2007

[38] LEE, Daewon ; HOFMANN, Matthias ; STEINKE, Florian ; ALTUN, Yasemin ; CAHILL, Nathan D. ; SCHÖLKOPF, Bernhard: Learning similarity measure for multi-modal 3D image registration. In: *CVPR*, IEEE, 2009, 186-193

[39] LEVENTON, Michael E. ; GRIMSON, W. Eric L.: Multi-modal Volume Registration Using Joint Intensity Distributions. In: *Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*. London, UK, UK : Springer-Verlag, 1998 (MICCAI '98), 1057–1066

[40] LOWE, David G.: Distinctive Image Features from Scale-Invariant Keypoints. In: *Int. J. Comput. Vision* 60 (2004), Nr. 2, 91–110

[41] MAES, Frederik ; COLLIGNON, André ; V, Dirk ; MARCHAL, Guy ; SUETENS, Paul: Multimodality Image Registration by Maximization of Mutual Information. In: *IEEE transactions on Medical Imaging* 16 (1997), S. 187–198

[42] MAIRAL, Julien ; BACH, Francis ; PONCE, Jean ; SAPIRO, Guillermo: Online Dictionary Learning for Sparse Coding. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA : ACM, 2009 (ICML '09), 689–696

[43] MANJUNATH, B. S. ; MA, W. Y.: Texture Features for Browsing and Retrieval of Image Data. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 18 (1996), August, Nr. 8, 837–842

[44] MANJUNATH, B. S. ; OHM, J. R. ; VASUDEVAN, V. V. ; YAMADA, A.: Color and Texture Descriptors. In: *IEEE Trans. Cir. and Sys. for Video Technol.* 11 (2001), Juni, Nr. 6, 703–715

[45] MEMISEVIC, Roland: Unsupervised learning of image transformations. In: *In Computer Vision and Pattern Recognition. IEEE Computer Society*, 2007

[46] MIGNON, Alexis ; JURIE, Frédéric: PCCA: A new approach for distance learning from sparse pairwise constraints. In: *CVPR*, IEEE, 2012, 2666-2672

[47] MIKOLAJCZYK, Krystian ; SCHMID, Cordelia: Scale & Affine Invariant Interest Point Detectors. In: *International Journal of Computer Vision* 60 (2004), Nr. 1, 63-86

[48] MONACI, Gianluca ; V, Pierre ; MAILHÉ, Boris ; LESAGE, Sylvain ; GRIBONVAL, Rémi: Learning Multi-Modal Dictionaries. In: *IEEE Transactions on Image Processing* (2006)

[49] MUJA, Marius ; LOWE, David G.: Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In: *International Conference on Computer Vision Theory and Application VISSAPP'09)*, INSTICC Press, 2009, S. 331–340

[50] PLUIM, Josien P. W. ; MAINTZ, J. B. A. ; VIERGEVER, Max A.: Image Registration by Maximization of Combined Mutual Information and Gradient Information. In: *IEEE Trans. Med. Imaging* 19 (2000), Nr. 8, 809-814

[51] SANDE, Koen van d. ; GEVERS, Theo ; SNOEK, Cees: Evaluating Color Descriptors for Object and Scene Recognition. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (2010), September, Nr. 9, 1582–1596

[52] SCHAPIRE, Robert E.: A Brief Introduction to Boosting. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1999 (IJCAI'99), 1401–1406

[53] SHAKHNAROVICH, Gregory: *Learning task-specific similarity*, Massachusetts Institute of Technology, Diss., 2006

[54] SHAKHNAROVICH, Gregory ; VIOLA, Paul ; DARRELL, Trevor: Fast pose estimation with parameter sensitive hashing. In: *In ICCV*, 2003, S. 750–757

[55] SHEN, Chunhua ; KIM, Junae ; WANG, Lei: Scalable Large-Margin Mahalanobis Distance Metric Learning. In: *CoRR* abs/1003.0487 (2010)

[56] SO, Ronald W K. ; CHUNG, Albert C S.: Learning-based multi-modal rigid image registration by using Bhattacharyya distances. In: *Conf Proc IEEE Eng Med Biol Soc* 2011 (2011), 2642-5

[57] STATHAKI, Tania: *Image Fusion: Algorithms and Applications*. Academic Press, 2008

[58] SZELISKI, Richard: Image Alignment and Stitching: A Tutorial. In: *Found. Trends. Comput. Graph. Vis.* 2 (2006), Januar, Nr. 1, 1–104

[59] SZELISKI, Richard: *Computer Vision: Algorithms and Applications*. 1st. New York, NY, USA : Springer-Verlag New York, Inc., 2010. – 188f S

[60] VEDALDI, Andrea: *Scale Invariant Feature Transform (SIFT) in VLFeat documentation*. Version: August 2014

[61] WACHINGER, Christian ; NAVAB, Nassir: Manifold Learning for Multi-Modal Image Registration. In: *Proceedings of the British Machine Vision Conference*, BMVA Press, 2010, S. 82.1–82.12. – doi:10.5244/C.24.82

[62] WACHINGER, Christian ; NAVAB, Nassir: Entropy and Laplacian images: Structural representations for multi-modal registration. In: *Medical Image Analysis* 16 (2012), Nr. 1, 1-17

[63] WEI, Li-Yi ; LEVOY, Marc: Fast texture synthesis using tree-structured vector quantization. In: *SIGGRAPH*, 2000, 479-488

[64] WELLS, William M. ; VIOLA, Paul A. ; ATSUMI, Hideki ; 0002, Shin N. ; KIKINIS, Ron: Multi-modal volume registration by maximization of mutual information. In: *Medical Image Analysis* 1 (1996), Nr. 1, 35-51

[65] XIONG, Caiming ; JOHNSON, David ; XU, Ran ; CORSO, Jason J.: Random Forests for Metric Learning with Implicit Pairwise Position Dependence. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA : ACM, 2012 (KDD '12), 958–966

[66] YANG, Gehua ; STEWART, Charles V. ; MICHAL SOFKA, Chia-Ling T.: Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2007), November, Nr. 11, 1973–1989

[67] ZHANG, Jie ; RANGARAJAN, Anand: A Unified Feature-Based Registration Method for Multimodality Images. In: *ISBI*, IEEE, 2004, 724-727

[68] ZULIANI, Marco ; KENNEY, Charles S. ; MANJUNATH, B. S.: The multiRANSAC algorithm and its application to detect planar homographies. In: *ICIP (3)*, IEEE, 2005, 153-156