Bullin, Martin ; Henrich, Andreas

Die inhaltsbasierte Bildsuche und Bilderschließung : Ansätze und Problemfelder

Erstveröffentlichung

Bullin, Martin, Henrich, Andreas: Die inhaltsbasierte Bildsuche und Bilderschließung : Ansätze und Problemfelder. In: Bilddaten in den Digitalen Geisteswissenschaften. Hastik, Canan; Hegel, Philipp (Hg.). Wiesbaden : Harrassowitz Verlag 2020. S. 11-34.
DOI: 10.13173/9783447114608.011

Bilddaten in den Digitalen Geisteswissenschaften

Herausgegeben von
Canan Hastik und Philipp Hegel

2020
Harrassowitz Verlag · Wiesbaden

Episteme in Bewegung

Beiträge zu einer transdisziplinären Wissensgeschichte

Herausgegeben von Gyburg Uhlmann
im Auftrag des Sonderforschungsbereichs 980
„Episteme in Bewegung.
Wissenstransfer von der Alten Welt
bis in die Frühe Neuzeit“

Band 16

2020
Harrassowitz Verlag · Wiesbaden

Die Reihe „Episteme in Bewegung“ umfasst wissensgeschichtliche Forschungen mit einem systematischen oder historischen Schwerpunkt in der europäischen und nicht-europäischen Vormoderne. Sie fördert transdisziplinäre Beiträge, die sich mit Fragen der Genese und Dynamik von Wissensbeständen befassen, und trägt dadurch zur Etablierung vormoderner Wissensforschung als einer eigenständigen Forschungsperspektive bei.
Publiziert werden Beiträge, die im Umkreis des an der Freien Universität Berlin angesiedelten Sonderforschungsbereichs 980 „Episteme in Bewegung. Wissenstransfer von der Alten Welt bis in die Frühe Neuzeit“ entstanden sind.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) –
Projektnummer 191249397 – SFB 980.

Umschlaggestaltung unter Verwendung von: Gm133, Infrarotaufnahme, Montage aus 2 Einzelaufnahmen; Germanisches Nationalmuseum, Institut für Kunsttechnik und Konservierung, Aufnahme: Beate Fücker (CC BY NC ND).

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://www.dnb.de abrufbar.

Informationen zum Verlagsprogramm finden Sie unter
http://www.harrassowitz-verlag.de

Verlegt durch Otto Harrassowitz GmbH & Co. KG, Wiesbaden 2020
Gedruckt auf alterungsbeständigem Papier.
Druck und Verarbeitung: Memminger MedienCentrum AG
Printed in Germany

ISSN 2365-5666
eISSN 2701-2522
DOI: 10.13173/2365-5666

ISBN 978-3-447-11460-8
Ebook ISBN 978-3-447-39046-0
DOI: 10.13173/9783447114608

Inhalt

Die inhaltsbasierte Bildsuche und Bilderschließung: Ansätze und Problemfelder

Martin Bullin und Andreas Henrich

Bilder zu suchen und zu analysieren erweist sich als deutlich komplexer als die – ohnehin schon schwierige – Suche und Analyse von und in Textdokumenten. Der vorliegende Beitrag gibt in diesem Kontext einen Überblick über Grundlagen und Konzepte der inhaltsbasierten Bildrecherche für Anwender aus den Geistes- und Kulturwissenschaften. Der erste Teil behandelt die Geschichte und Konzepte zum inhaltsbasierten Image Retrieval: typische Anwendungsfälle, Arten von Bildern, die Rolle der Semantik, die Auswirkungen einer Segmentierung, die sensorische bzw. semantische Lücke sowie Standards der Bildrecherche. Der zweite Teil des Beitrags erläutert verschiedene Herangehensweisen zur Bildsuche und Bildanalyse. Beginnend mit klassischen Bildeigenschaften (Farbe, Textur, Form) über Segmentierungsverfahren und lokale Bildeigenschaften bis hin zu Ansätzen des Deep Learning werden verschiedene Verfahren skizziert und in ihren Stärken und Schwächen charakterisiert. Ein Blick auf einige exemplarische Anwendungen rundet den Beitrag ab.

Einleitung und Motivation

Neben Texten stellen Bilder und Objekte wichtige Artefakte in den Geistes- und Kulturwissenschaften dar. Unter die Oberkategorie ‚Bild' fallen dabei sehr verschiedene Objekttypen wie Fotografien, Druckgrafiken, Zeichnungen, Gemälde oder z.B. auch Karten.

Eine wichtige Frage ist, wie man Bilder in der digitalen Welt geeignet suchen, finden und klassifizieren kann. Hier ist zunächst zwischen einer Suche auf Basis manuell gepflegter oder automatisch erstellter Metadaten und einer Suche auf Basis des eigentlichen Bildinhalts zu unterscheiden. Bei den Metadaten wiederum sind Metadatenstandards wie DublinCore, LIDO, CIDOC CRM oder EDM, als Datenmodell der *Europeana* – um nur einige zu nennen – von eher technischen Standards wie EXIF oder NISO zu unterscheiden, die von der Kamera oder vom Scanner gelieferte technische Metadaten umfassen (wie Kameramodell, Verschlusszeit oder Aufnahmezeitpunkt).[1] Sind Metadaten in hoher Qualität vor-

1 Die Dublin Core Metadata Initiative betrachtet Fragen zu Design und Anwendung von Metadaten, URL: http://dublincore.org (05.06.2019); Das Datenformat „Lightweight Information Describing Objects" ist ein vorwiegend von Museen genutztes XML-Schema, URL: http://network.icom.museum/cidoc/working-groups/lido/what-is-lido (05.06.2019); Das CIDOC Conceptual Reference Model gibt Definitionen und eine formale Struktur zur Beschreibung

DOI: 10.13173/9783447114608.011

handen, kann eine Suche über diese Metadaten sehr effektiv sein. Man denke an eine gute inhaltliche Klassifikation mit Iconclass[2] oder an die Verfügbarkeit des per GPS erfassten Aufnahmeortes bei einer Fotografie.

Andererseits sind nicht immer hochwertige Metadaten vorhanden und oft repräsentieren die vorhandenen Metadaten auch eine bestimmte individuelle Interpretation bei ihrer Erfassung. Dies spricht dafür, neben einer auf Metadaten basierenden Suche auch eine inhaltsbasierte Suche zu unterstützen. Die Bezeichnung ‚inhaltsbasiert' ist dabei auf den ersten Blick mehrdeutig. Im Englischen spricht man von *Content Based Image Retrieval* (CBIR). Gemeint ist, dass der technische Bildinhalt und damit letztlich die Farb- oder Helligkeitswerte der einzelnen Bildpunkte Gegenstand der Betrachtung sind – und eben nicht die Metadaten.

Um die Möglichkeiten der inhaltsbasierten Bildsuche und -analyse sinnvoll einschätzen zu können, werden wir im Weiteren zunächst auf die Geschichte und Einordnung der inhaltsbasierten Bildsuche eingehen. Im dritten Abschnitt werden wir uns dann klassischen, bildglobalen Verfahren der Inhaltsbeschreibung zuwenden. Aspekte der Segmentierung von Bildern werden im vierten Abschnitt angesprochen. Im fünften Abschnitt werden Verfahren beschrieben, die lokale Bildeigenschaften adressieren, und im sechsten Abschnitt die damit verbundenen Ansätze zur Bestimmung der Bildähnlichkeit. Der siebte Abschnitt geht auf Verfahren des *Deep Learning* (DL) ein, bevor der achte Abschnitt Anwendungsbeispiele skizziert und der neunte Abschnitt die Arbeit zusammenfasst. Das Ziel ist dabei einen Überblick über Verfahren der Bildanalyse und Bildsuche sowie der Anwendungsmöglichkeiten im Bereich der Geisteswissenschaften zu geben.

Grundlagen zur Bildsuche

Die Bildsuche beziehungsweise das Themengebiet Image Retrieval vereint Elemente aus mehreren Forschungsdomänen: Hierbei wird zuerst die unterschiedliche Aussagekraft von Bildern für Mensch und Computer, für die die sensorische wie semantische Lücke eine wesentliche Rolle spielen, als Teilgebiet der *Computer Vision* betrachtet. Im Anschluss werden alle weiteren für die Bildsuche relevanten Aspekte der *Computer Vision* näher beleuchtet. Die Suchtypen, die nach Anfrageart oder Suchobjekt unterschieden werden, werden dann im Unterkapitel

von impliziten und expliziten Konzepten sowie Beziehungen, die in der Dokumentation von kulturellem Erbe genutzt werden, URL: http://www.cidoc-crm.org (05.06.2019); Das Europeana Data Model ist das von der *Europeana*-Kollektion genutzte Datenmodell, URL: https://pro.europeana.eu/resources/standardization-tools/edm-documentation (05.06.2019); Das Exchangable Image File Format bietet die Möglichkeit Metadaten in Bildern zu speichern, URL: http://exif.org (05.06.2019); Die National Information Standards Organization entwickelt und publiziert technische Standards, um digitale Informationen zu verwalten; relevant für Bilder ist insbesondere ANSI/NISO Z39.87-2006 (R2017) Data Dictionary – Technical Metadata for Digital Still Images, URL: https://www.niso.org/ (05.06.2019).

2 Iconclass ist ein Klassifizierungssystem, das für die Bereiche Kunst und Ikonographie entwickelt wurde, URL: http://www.iconclass.nl/home (31.12.2018).

Anfragearten und Anwendungen beschrieben. In den anschließenden Unterkapiteln wird darauf eingegangen, welche Informationen und Bildarten existieren und welchen Einfluss diese in verschiedenen Szenarien aufgrund von Bilddomänen sowie Domänenwissen haben.

Die sensorische und semantische Lücke

CBIR kann dem Themenfeld der *Computer Vision* zugeordnet werden, also dem Forschungsgebiet, das sich mit dem ‚Sehvermögen' von Computern beschäftigt. *Computer Vision* adressiert die Frage, wie aus den technischen Abbildern der Realität, aufgenommen über Sensoren, die zugrunde liegenden Informationen der Realität abgeleitet bzw. rekonstruiert werden können. Dabei besteht zwischen der technischen Darstellung aus Bildpunkten in einer gewissen räumlichen Auflösung und Farbtiefe und dem Abgebildeten eine zweifache Lücke.[3]

Die *sensorische Lücke* ist der Informationsverlust, der entsteht, wenn eine Realität abgebildet wird. Diese wird neben der technischen Auflösung bei der Digitalisierung auch durch Faktoren wie den Standpunkt der Kamera, Verdeckungen oder die Beleuchtung beeinflusst. Hier geht es also darum, dass die Realität im technischen Abbild nicht perfekt und nicht vollständig abgebildet ist.

Die *semantische Lücke* beschreibt den Unterschied zwischen den Informationen, die aus den technisch repräsentierten visuellen Daten extrahiert werden können, und der Interpretation, die dieselben Daten in einem gegebenen Kontext erhalten. Dies spiegelt den Unterschied zwischen den visuellen Eigenschaften eines Bildes und der Semantik (Objekte, Beziehungen, Bedeutungen) sowie dem abstrakten Verständnis dieses Bildes in der Wahrnehmung durch einen Menschen wider. Die Objekte in einem Foto mit zwei Menschen, die sich die Hände schütteln, wären hierbei die Personen. Eine Beziehung wäre der Akt des Händeschüttelns. Die Bedeutung des Händeschüttelns könnte eine Begrüßung sein. Ist bekannt, dass die Personen Politiker sind, könnte das Bild für einen Vertragsschluss stehen. Dabei wird weiter unterschieden in die Lücke zwischen den visuellen Eigenschaften und der Objektebene (untere semantische Lücke) und die Lücke zwischen den identifizierten Objekten und der vollständigen Semantik eines Bildes (obere semantische Lücke).[4]

Bildsuche und Computer Vision

Um die Hintergründe und Möglichkeiten der Bildanalyse besser einschätzen zu können, soll nun kurz die Geschichte der *Computer Vision* beleuchtet werden. In den frühen 1970er Jahren war Computer Vision nur eine Komponente zur visuellen Aufnahme von Informationen zur weiteren Nutzung in *Artificial Intelligence*

3 Michael Grubinger, *Analysis and evaluation of visual information systems performance*, PhD Thesis, Victoria University Melbourne 2007, URL: http://vuir.vu.edu.au/1435/ (04.02.19).

4 Jonathon S. Hare u.a., „Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval", in: *Multimedia Content Analysis, Management and Retrieval 2006*, hg. von Alan Hanjalic u.a., Bellingham 2006 (SPIE Proceedings 6073), S. 75–86.

Systemen (AI).[5] Mitte der 1970er wurde die Entwicklung bzw. Ableitung von 3D-Strukturen zu einem Kernelement der *Computer Vision*, um so ein Verständnis der gesamten Szene zu schaffen. In diesem Kontext wurden Algorithmen zur Linien- und Kantenerkennung entwickelt.[6] Im Jahr 1973 führten Eschlager und Fischler sogenannte *Pictorial Structures* als Methodik ein, die Bilder in Einzelelemente und deren Verbindungen zerlegt.[7] Diese Arbeiten waren grundlegend für die Erkennung einzelner Objekte in Bildern (*Object Recognition*).[8]

In den 1980ern wurden in der *Computer Vision* viele weitere Forschungsstränge untersucht. Die für diesen Artikel wesentlichen beschäftigten sich mit der Verbesserung der Kanten- und Konturerkennung.[9] In der nächsten Dekade rückten unter anderem Ansätze zur Rekonstruktion von 3D-Modellen aus mehreren Bildern in den Vordergrund.[10] Neben den Trends in Richtung 3D wurde auch die *Bildsegmentierung* – eine Kernthematik seit den frühen Tagen der *Computer Vision* – weiter optimiert.[11]

Die Entwicklungen der *Computer Vision*, die die aktuelle Forschung im Bereich CBIR beeinflussen, fanden vor allem ab der Jahrtausendwende statt. Es kamen Feature-basierte Techniken auf, die unter anderem für DL zur Objekterkennung verwendet werden können.[12] In dieser Zeitspanne waren *Patch-based Features* zur Bilderkennung die vorwiegend erforschten Themen. Andere Forschungsstränge beschäftigten sich mit der Bilderkennung basierend auf Konturen sowie der Bildsegmentierung.[13] Ein weiterer Trend – begründet durch die Leistungssteigerung der Computer – war die zunehmende Nutzung von *Machine Learning* (ML) Verfahren für *Visual Recognition* Probleme. Das Vorliegen von vielen teilweise bereits

5 Richard Szeliski, *Computer Vision: Algorithms and Applications*, London 2010, S. 10.

6 Larry S. Davis, „A survey of edge detection techniques", in: *Computer Graphics and Image Processing* 4/3 (1975), S. 248–260.

7 Martin A. Fischler und Robert A. Elschlager, „The Representation and Matching of Pictorial Structures", in: *IEEE Transactions on Computers* 22/1 (1973), S. 67–92.

8 Pedro F. Felzenszwalb und Daniel P. Huttenlocher, „Pictorial Structures for Object Recognition", in: *International Journal of Computer Vision* 61/1 (2005), S. 55–79.

9 John Canny, „A Computational Approach to Edge Detection", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), S. 679-698; Vishvjit S. Nalwa und Thomas O. Binford, „On Detecting Edges", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8/6 (1986), S. 699–714.

10 Richard Hartley u.a., „Stereo from uncalibrated cameras", in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1992), S. 761–764.

11 David Mumford und Jayant Shah, „Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems", in: *Communications on Pure and Applied Mathematics* 42/5 (1989), S. 577–685.

12 Jean Ponce u.a., *Toward Category-Level Object Recognition*, Berlin 2006.

13 Serge Belongie u.a., „Shape Matching and Object Recognition Using Shape Contexts", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24/24 (2002), S. 509–522; Greg Mori u.a., „Recovering Human Body Configurations: Combining Segmentation and Recognition", in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Bd. 2 (2004), S. 326–333.

gelabelten – d.h. mit Annotationen versehenen – Daten vor allem im Internet unterstützt diesen Trend.[14]

Anfragearten und Anwendungen

Um effektive Suchsysteme entwickeln zu können, muss man die Suche als Prozess verstehen. Hierzu wurden in der Literatur zahlreiche Modelle vorgeschlagen, die allerdings fast ausschließlich die Suche nach Textdokumenten adressieren. Typische Modelle sehen den Ausgangspunkt der Suche in einer Aufgabe, bei deren Bearbeitung ein Informationsbedarf entsteht. Dieser Informationsbedarf muss dann zunächst verbalisiert und schließlich in die Anfrageschnittstelle der genutzten Suchlösung umgesetzt werden. Bei der Betrachtung der Ergebnisse – dem nächsten Schritt – können grob zwei Fälle unterschieden werden: Im ersten Fall sind die Ergebnisse passend und die Suchtreffer befriedigen den Informationsbedarf. Im zweiten Fall können die Ergebnisse den Informationsbedarf allenfalls teilweise erfüllen, sodass der Suchprozess iterativ fortgesetzt werden muss. Dabei können sich Änderungen im Informationsbedarf ergeben haben. Durch die Betrachtung einzelner Suchergebnisse versteht man die Aufgabe ggf. besser und setzt den iterativen Suchprozess mit einem modifizierten Informationsbedarf fort.[15]

Bedeutsam für das Verständnis dieses Prozesses ist der Charakter des Informationsbedarfs bzw. des Ziels der Suche. Jansen u.a. klassifizieren für die Suche im Web grundlegend drei Arten der Suche und ihrer Intention:[16] transaktionales, informatives und navigierendes Suchen. Eine transaktionale Suche hat nach dem Suchprozess eine Transaktion, beispielsweise einen Kauf, als Ziel. Im Hinblick auf die Bildsuche könnte man hier ggf. an Produktbilder oder Logos denken, diese Kategorie ist bei der Bildsuche aber sicher eher selten. Eine navigierende Suche hat das Finden der Website eines gewünschten Inhalts zum Ziel. Oft ist hier das gewünschte Ergebnis (eine Webseite oder ein Dokument) eigentlich schon bekannt. Man spricht dann von *Lookup* oder *Known-Item Search*. Dieser Anfragetyp ist auch im Bereich der Bildsuche oft anzutreffen. Kritisch ist hier die Frage, wie die Anfrage passend formuliert werden kann: Ob über Schlüsselworte, über ein Beispielbild oder etwa über eine Skizze. Der in geisteswissenschaftlichen Anwendungen sicher naheliegendste Typus einer Anfrage ist die informative Suche, bei der sich der Nutzer einen Überblick über ein bestimmtes Themenfeld verschaffen möchte. Marchionini bezeichnet solche Anfrageaktivitäten als *Exploratory Search*

14 Andras Ferencz u.a., „Learning to Locate Informative Features for Visual Identification", in: *International Journal of Computer Vision* 77/1-3 (2008), S. 3-24; Yann LeCun u.a., „Deep learning", in: *Nature* 521 (2015), S. 436–444.

15 Andrei Broder, „A taxonomy of web search", in: *ACM SIGIR forum* 36/2 (2002), S. 3–10.

16 Bernard J. Jansen u.a., „Determining the informational, navigational, and transactional intent of Web queries", in: Information Processing & Management 44/3 (2008), S. 1251–1266.

und unterscheidet dann weiter in die Unterbereiche *Learn* und *Investigate*.[17] Die Anfragetypen schließen sich keineswegs aus, sondern sie kommen in einem umfangreicheren Arbeitszusammenhang oft in Kombination vor.

Eine Besonderheit der Bildsuche ergibt sich dabei aus der Frage, wie eine Anfrage formuliert werden kann. Häufig wird die Suchanfrage in Form eines anderen Medientyps vorliegen als das erwartete Ergebnis: Nutzer suchen mit einem oder mehreren Schlüsselwörtern und wollen möglichst relevante Bilder erhalten. Auf der anderen Seite könnten Nutzer möglichst viele Informationen über den Inhalt eines ihnen bereits vorliegenden Bildes erhalten wollen. Beide Suchanfragen fallen in die informative Kategorie.

Allgemein gilt natürlich, dass das Suchsystem bestimmt, welche Arten der Anfrageformulierung möglich sind. Suchmaschinen wie Google erlauben bei der Bildsuche im Web die Formulierung der Anfrage mit Suchbegriffen. Die gelieferten Treffer sind zum Teil durchaus von guter Qualität, so dass der Eindruck entstehen könnte, dass die Problemstellung der Bildsuche mit Schlagworten weitestgehend gelöst ist. Die Ergebnisse werden jedoch nicht durch ein wirklich inhaltliches Verständnis der Bilder, sondern durch die Ausnutzung von Meta- bzw. Kontextinformationen erzielt. Eine Website zum Thema ‚Martin Luther' wird im Regelfall Bilder von Martin Luther bzw. kontextzugehörige Bilder beinhalten. Diese Inhalte kann sich Google zu Nutze machen und so für den Suchbegriff ‚Martin Luther' Bildergebnisse von eben solchen Webseiten anzeigen. Aufgrund der schieren Menge an Inhalten im Web, kann Google so für viele Anfragen passende Ergebnisse liefern. Um eine inhaltsbasierte Bildsuche im oben beschriebenen Sinn handelt es sich dabei aber nicht.

Um auch ohne Meta- oder Kontextinformationen eine Bildsuche durchführen zu können, wurde das Paradigma *Query-by-Example* (QbE) entwickelt.[18] Hier werden zu einem Bild ähnliche Bilder im Datenbestand gesucht. Die Ähnlichkeit wurde dabei in frühen Systemen als einfache Farbähnlichkeit oder Texturähnlichkeit (s. Abschnitt *Klassische Bildfeatures*) ermittelt.

Neben die Suche treten dabei andere Aufgabenstellungen wie die Objekterkennung in Bildern, die Klassifikation oder die automatische Annotation von Bildern. Oftmals sind hier themenbezogene Instanz- bzw. Ähnlichkeitsaspekte zentral.

Betrachtet man beispielsweise die Aufgabe ‚Embleme auf Basis der Icones (Picturae) zu suchen bzw. zu annotieren', so ergeben sich zahlreiche unterschiedliche Aspekte. Wir betrachten hier im Speziellen den *Emblematica Online*-Datensatz.[19]

17 Gary Marchionini, „Exploratory search: from finding to understanding", in: *Communications of the ACM* 49/4 (2006), S. 41–46.

18 Myron Flickner u.a., „Query by Image and Video Content: The QBIC System", in: *Computer* 28/9 (1995), S. 23–32.

19 Herzog August Bibliothek Wolfenbüttel, URL: http://www.hab.de/de/home/wissenschaft/forschungsprofil-und-projekte/emblematica-online.html (07.12.2018).

Hier sind Teile der Daten mit Iconclass-Klassen[20] zur Erfassung und inhaltlichen Erschließung der Bildinhalte versehen. *Iconclass* ist eine Ontologie für Kunst und Ikonographie. Im Datensatz der *Emblematica* sind nun verschiedene Sucharten vorstellbar.

Auf der einen Seite könnte man Kopien einer Pictura suchen wollen (s. Abb. 1.1a bis 1.1c, Farbteil), die mit dem gleichen Holzblock gedruckt wurden. Diese können farbliche Abweichungen aufgrund von Material, Zeit und weiteren Einflüssen aufweisen. Auch Erweiterungen oder Reduzierungen des Holzschnitts treten auf.

Auf der anderen Seite könnte man an Emblemen interessiert sein, deren Pictura ikonografisch ähnliche Motive zeigt, die aber in der bildlichen Darstellung durchaus deutlich differieren können (s. Abb. 1.1d bis 1.1e, Farbteil). Hier könnte eine Objekterkennung (Fuchs, Krähe, …) ebenso hilfreich sein wie eine Klassifikation auf Basis verschiedener Kriterien. Diese Beispiele verdeutlichen die Komplexität der Aufgabenstellung.

Bildarten und enthaltene Informationen

Ein wesentlicher Bereich im Themenfeld des *Image Retrieval* ist die zugrundeliegende Information, nämlich das Bild bzw. dessen Kontextinformationen. Diese bestimmen den Informationsgehalt sowie die Möglichkeiten, die zur weiteren Verarbeitung des Bildes zur Verfügung stehen. Bei den Bildarten können Bilder unterschieden werden, die am Computer erstellt, direkt fotografiert oder digitalisiert werden, beispielsweise durch einen Scanner. Die jeweiligen Geräte beeinflussen hierbei sowohl die inhaltsbezogenen Informationen wie die Metainformationen. Beispielsweise speichert eine aktuelle Digitalkamera im Regelfall ihren Typ, Fokuspunkt und weitere Werte in den Metainformationen der Bilder ab. Bei einem gescannten Bild liegen diese Informationen zum Teil nicht vor. Früher war Fotografie nur in Schwarz-Weiß möglich, heute werden Fotografien im Regelfall im RGB-Farbraum (Rot, Grün, Blau) gespeichert, was wiederum eine Steigerung der vorhandenen Informationen bedeutet. Der Informationsgehalt computererzeugter Bilder hängt wiederum von der verwendeten Software sowie dem Speicherformat ab. Hier ist jedoch keine ‚Verzerrung' durch die Retrodigitalisierung vorhanden. Die Informationen liegen also ‚klarer' vor.

Bilderdomänen und Domänenwissen

Eng verzahnt mit den Bildarten sind die Domänen. Hierbei ist die Breite der Domäne, die von einer Suchlösung abgedeckt werden muss, ein ausschlaggebendes Kriterium. Smeulders u.a. unterscheiden zwischen breiten und schmalen Domänen.[21] Eine schmale Domäne hat eine begrenzte und vorhersagbare Streuung in allen relevanten Aspekten ihrer Erscheinung. Eine breite Domäne dagegen hat

20 Vgl. Anm. 2.

21 Arnold W. M. Smeulders u.a., „Crossing the Divide between Computer Vision and Data Bases in Search of Image Databases", in: *Visual Database Systems 4*, hg. von Yannis E. Ioannidis und Wolfgang Klas, London 1998, S. 223–239.

eine sehr hohe und vor allem auch unvorhersagbare Abweichung im Erscheinungsbild – sogar für dieselbe semantische Bedeutung.

Aus der Art der Bilddomäne resultiert die Art des vorhandenen und nutzbaren Wissens. Je schmaler die Domäne, desto mehr spezifisches Wissen kann genutzt werden. Kann bei sehr breiten Domänen – wie den Bildern des Internets – kein Wissen vorausgesetzt werden, so können bei der Betrachtung von beispielsweise mittelalterlichen Emblemen Informationen wie der eingeschränkte Farbraum genutzt werden. Ein anderes Beispiel, in dem die enge Domäne gut zur Verbesserung der Sucheffektivität genutzt werden kann, ist die Suche in Briefmarkenbeständen.[22]

Klassische Bildfeatures

Als einfache – und auch recht einfach zu interpretierende – Eigenschaften von Bildern können Farb-, Textur- und Formeigenschaften in der Suche adressiert werden. Typisch ist dabei zunächst die Anwendung auf das gesamte Bild. So können Bilder z.B. aufgrund einer gut übereinstimmenden Verteilung der auftretenden Farben als ähnlich eingestuft werden.

Gerade bei der Betrachtung der Farbähnlichkeit werden aber auch Probleme dieses Ansatzes deutlich. Bilder werden in der Regel zunächst im RGB-Farbraum abgebildet, wobei zur Kodierung jedes Farbkanals zumindest ein Byte verwendet wird. Insgesamt sind damit 2^{24} = 16.777.216 verschiedene Farben darstellbar. Für das Farbähnlichkeitsempfinden ist diese Auflösung eigentlich zu fein und auch der euklidische Abstand in diesem Farbraum ist wenig geeignet, die menschliche Farbwahrnehmung angemessen umzusetzen. Die Farbverteilung im RGB-Farbraum wird in Abbildung 1.2, Farbteil, für die fünf Beispielbilder in der Mitte verdeutlicht. Die Kurven zeigen jeweils an, wie viele Bildpunkte den jeweiligen Intensitätswert für die drei Grundfarben aufweisen. Im obersten Bild sind die Verläufe sehr ähnlich, was den Schwarz-Weiß-Charakter des Bildes andeutet. Die hohen Werte bei großen Intensitäten verdeutlichen, dass das Bild insgesamt sehr hell ist. Bei den beiden folgenden Bildern wird der Rotstich im Bild deutlich. Bei den beiden Fotos im unteren Bereich der Abbildung erkennt man beim Gebirgsbild, dass es relativ viele Pixel mit geringer Rot-Intensität gibt. Ferner fällt der unterschiedliche Verlauf der Kurven auf, ohne dass eine unmittelbare Interpretation naheliegend wäre.

Die Beispiele deuten bereits an, dass der RGB-Farbraum in vielen Fällen nicht optimal für eine Farbähnlichkeitssuche ist. Ein wichtiger Schritt zur Optimierung ist daher die Farbinformation zu den einzelnen Bildpunkten in einen geeigneten Farbraum zu übertragen. Hier greift man z.B. auf den HSV-Farbraum zurück, der den Farbwert (Hue) als Winkel auf dem Farbkreis (0° für Rot, 120° für Grün, 240° für Blau), die Farbsättigung (Saturation; 100% für eine gesättigte, reine

22 Sven Siggelkow, *Feature histograms for content-based image retrieval*, Dissertation, Albert-Ludwigs-Universität Freiburg 2002.

Farbe) sowie die Helligkeit (Value) unterscheidet. Wenn man in diesem dreidimensionalen Farbraum den Farbwert in 16 mögliche Klassen und die Sättigung sowie die Helligkeit in jeweils 4 mögliche Klassen aufteilt, dann entstehen 16·4·4 = 256 Farbklassen. Zählt man die Pixel eines Bildes, die in die jeweilige Farbklasse fallen, so kann man ein Histogramm bilden, dass die Farbverteilung eines Bildes interpretierbar macht. In Abbildung 1.2, Farbteil, findet sich ein solches Histogramm jeweils auf der linken Seite, wobei wir zur besseren Übersichtlichkeit eine 8:2:2-Aufteilung mit nur 32 Farbklassen statt der üblichen 16:4:4-Aufteilung gewählt haben. Hier wird die Unterscheidung der beiden unteren Bilder durch die dominanten Farbklassen sehr plausibel. Bei den leicht rotstichigen Bildern ergeben sich sehr ähnliche Histogramme. Lediglich beim ersten Bild versagt das Verfahren, weil sehr helle, vom Menschen als weiß empfundene Bildpunkte fast zufällig einer Farbe zugeteilt werden. Hierbei wird die Schwäche des HSV-Models offengelegt, dass der Farbwert bei weißen Bildpunkten keine Rolle spielt, da eine maximale Helligkeit vom Farbwert unabhängig immer Weiß ergibt. Damit zeigt sich klar, dass der Farbraum passend zur Bildkollektion gewählt werden muss.[23]

Neben den Farbeigenschaften eines Bildes spielt für die menschliche Wahrnehmung auch die Textur eine große Rolle. Unter der Textur versteht man eine kleinräumige Oberflächenstruktur, gleich ob natürlich oder künstlich, regelmäßig oder unregelmäßig. Beispiele für Texturen unterschiedlicher Charakteristik könnten eine Baumrinde, ein Strickmuster, eine Holzmaserung oder die Oberfläche eines Schwamms sein. Um die Textur eines Bildes zu erfassen wird im Kontext der Bildsuche in der Regel eine statistische Texturanalyse genutzt, die die Textur anhand bestimmter Attribute – wie z.B. der lokalen Grauwertvarianz, der Regelmäßigkeit, Grobkörnigkeit, Orientierung und des Kontrastes – beschreibt. Einige Maße dazu sind z.B. im MPEG-7 Standard definiert.[24]

Da die Textureigenschaften eines Bildes weitgehend unabhängig von den Farbeigenschaften sind, werden Farbbilder für die Texturanalyse üblicherweise zunächst in Graustufenbilder konvertiert. Bei der Betrachtung der Graustufenbilder stellen sich dann die Fragen: Welche Strukturen möchte man als Textur bezeichnen? Wo im Bild befinden sich diese Strukturen?

Wie Texturen in einem Bild vorkommen, hängt stark von der Skalierung ab. Stellt das Bild nur eine Aufnahme eines Teppichs dar, so wird man das Bild im Allgemeinen durch eine einzige Textur gut beschreiben können. Haben wir es aber mit einem Bild zu tun, das beispielsweise einen Raum mit Personen, Möbeln, einem Parkettboden, Gardinen, etc. zeigt, so treten in dem Bild zahlreiche unterschiedliche Texturen auf. In diesem Fall erscheint es sinnvoll, das Bild zunächst in Segmente einzuteilen und für diese einzeln die Textureigenschaften zu bestimmen. Wir werden uns im folgenden Abschnitt mit Fragen der Bildseg-

23 Horst Eidenberger, *Handbook of Multimedia Information Retrieval*, Wien 2012, Kapitel 5.2.
24 Bangalore S. Manjunath u.a., *Introduction to MPEG-7: multimedia content description interface*, Bd. 1, Chichester 2002.

mentierung beschäftigen. Das Kriterium für das Auftreten einer Textur in einer Bildregion ist dabei die signifikante, regelmäßige Variation der Grauwerte.

Eine dritte Art von Bildeigenschaften, die zur Suche verwendet werden, sind Formeigenschaften. Zur Ermittlung dieser Eigenschaften wird oft zunächst eine Kantenerkennung durchgeführt. Dazu können entsprechende Masken eingesetzt werden, die in Bildern solche Bildpunkte hervorheben, deren Umgebung größere Sprünge hinsichtlich der Farbe oder der Helligkeit aufweist. Ein wesentliches Verfahren in diesem Kontext ist der *Canny-Algorithmus* der sich in verschiedene Faltungsoperationen gliedert und ein Bild ableitet, das die Kanten des Ausgangsbildes darstellt.[25] Für die Kanten können nun Statistiken zur Ausrichtung und Länge abgeleitet werden, um die Formen in einem Bild zu charakterisieren. Auf diese Weise können z.B. Bilder aus Städten sehr gut von Naturbildern unterschieden werden.

Eine große Schwäche der bisher betrachteten Verfahren ist, dass sie globale Bildeigenschaften beschreiben. Ein Farbhistogramm sieht für ein Bild, in dem die Bildpunkte beliebig vertauscht werden, genauso aus, wie für das Ursprungsbild. Ob sich z.B. 10% rote Pixel durch einen roten Kreis oder wild über das Bild verstreute rote Punkte ergeben, ist dem Histogramm nicht zu entnehmen.

Trotz dieser Schwäche können Verfahren auf Basis der Farb-, Form- oder Texturähnlichkeit in speziellen Fällen durchaus zielführend sein. Man denke z.B. an eine Ähnlichkeitssuche in einem großen Bestand von Markenzeichen und Firmenlogos.

Segmentierung

Ein Weg, den fehlenden Ortsbezug der im Abschnitt *Klassische Bildfeatures* beschriebenen Verfahren zu überwinden, ist die Segmentierung von Bildern. Soll z.B. mit dem Bild eines Markenzeichens über QbE nach Bildern gesucht werden, die dieses Markenzeichen enthalten, so versagen Farb- oder Textureigenschaften, die sich auf ein gesamtes Bild beziehen, zwangsweise. Der Grund ist wie oben beschrieben, dass das Anfragebild – also das Markenzeichen – auch im Falle eines Treffers nur mit einem kleinen Teil des Zielbildes übereinstimmt, was in einem Farbhistogramm kaum zum Ausdruck kommt. Ein Lösungsansatz ist, das Bild in Teile zu segmentieren, die möglichst gut mit signifikanten Bildregionen bzw. Objekten korrespondieren.

Der Begriff der Segmentierung ist verwandt mit dem Begriff der Objekterkennung. Dabei wird im Wesentlichen nach dem Ziel der Segmentierung unterschieden. Die Suche nach genau einer Objektklasse wird als *Object Detection* bezeichnet. Es wird nach einem bestimmten Objekt gesucht und nur geprüft, ob dieses dargestellt wird oder nicht. Die *Object Recognition* dagegen beschäftigt sich damit, überhaupt Objekte zu finden und sucht meist nach mehr als einer Klas-

25 John Canny, „A Computational Approach to Edge Detection", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), S. 679–698.

se. Ein Beispiel hierzu findet sich in Abbildung 1.3, Farbteil, in der Mitte und rechts, wobei z.B. die Klassen ‚Pferd', ‚Person' und ‚Hund' mit einer bestimmten Wahrscheinlichkeit erkannt werden. Dabei sind die Segmentierung und die Objekterkennung allerdings oft weder vollständig noch korrekt. So werden bei einem niedrigen Schwellenwert (*Threshold*) (mittlere Spalte) beispielsweise eine nicht vorhandene Handtasche, ein Rucksack sowie ein Frisbee, jedoch weiterhin nicht der Stein im Hintergrund als Objekte erkannt.

Abbildung 1.3, Farbteil, zeigt dabei auf der rechten Seite ein recht weit entwickeltes Verfahren auf Basis von umfangreichen Trainingsdaten, in denen Bildbereiche ausdrücklich mit Kategorien wie ‚Person', ‚Pferd' oder ‚Hund' annotiert sind. Es sind auch wesentlich einfachere Verfahren denkbar, die ohne Trainingsdaten auskommen. Ein sehr einfaches Verfahren wäre das Unterteilen eines Bildes in fünf gleich große, überlappende Bereiche wie in der linken Spalte von Abbildung 1.3, Farbteil, dargestellt. In Spezialfällen wie beispielsweise dem Scan einer Dokumentensammlung, bei dem immer vier gleich große Bilder gemeinsam gescannt wurden, könnte diese Segmentierung sinnvoll sein. Geht man jedoch weiter und betrachtet den Fall der *Emblematica*, reicht das Verfahren nicht aus. Hier zeigt sich auch die Mehrstufigkeit einer Segmentierung. Zunächst geht es um die Erkennung und Segmentierung der Pictura im Emblem und anschließend um die Erkennung von Objekten innerhalb der Pictura.

Allgemein kann man im Bereich der Segmentierung bei den einfachen Verfahren zwischen Ansätzen, die auf einer Kantenerkennung basieren, und sogenannten Wachstumsverfahren unterscheiden.[26] Bei der Kantenerkennung ist das Problem, dass die erkannten Kanten in der Regel zunächst keine geschlossenen Formen bilden. Hier müssen also kleinere Lücken in einem nachgelagerten Schritt über entsprechende Heuristiken geschlossen werden, um Segmente zu bilden. Die Wachstumsverfahren starten mit Saatpixeln, z.B. sehr hellen oder sehr dunklen Stellen im Bild. Von dort wird das Segment erweitert, bis eine bestimmte Größe erreicht ist oder die Helligkeit einen bestimmten Schwellenwert über- bzw. unterschreitet. Interpretiert man die Helligkeit im Bild als Höhe, so wird deutlich, warum diese Verfahren auch als *Water Flow Algorithm* charakterisiert werden.

Von solchen allgemeinen Verfahren sind spezielle Verfahren zu unterscheiden, die hinsichtlich einer Kollektion und eines darin auftretenden Problems optimiert sind. Im Fall der *Emblematica* können zur Entwicklung eines Verfahrens Informationen genutzt werden, wie beispielsweise, dass die Pictura von vielen weißen Bereichen umgeben ist, dass sie unregelmäßiger ist als die umgebenden Elemente oder auch dass sie immer eine ähnliche Größe aufweist.

Andere Verfahren kombinieren die Segmentierung direkt mit der Objekterkennung. Einige Erkennungssysteme verwenden dazu Klassifizierer und Lokali-

26 Walter Pätzold, „Digitale Bildbearbeitung", in: *Taschenbuch der Medieninformatik* hg. von Kai Bruns und Klaus Meyer-Wegener, Leipzig 2005, S. 164–165.

sierer, um die Erkennung durchzuführen. Sie wenden das Modell für eine Klasse (z.B. Pferd oder Person) auf ein Bild an mehreren Stellen und in unterschiedlichen Maßstäben an. Bereiche mit hoher Übereinstimmung werden dann als Segmente identifiziert und mit der Klasse annotiert. Andere aktuelle Verfahren wie YOLO wenden ein einzelnes neuronales Netzwerk auf das gesamte Bild an.[27] Dieses Netz teilt das Bild in Regionen auf und prognostiziert Begrenzungsrahmen und Wahrscheinlichkeiten für jede Region. Diese Begrenzungsrahmen (Segmente) werden mit den vorhergesagten Wahrscheinlichkeiten gewichtet. Die Segmente der zwei rechten Spalten in Abbildung 1.3, Farbteil, wurden mit diesem Verfahren bestimmt.

Bei diesen Verfahren verschmelzen somit Segmentierung und Ähnlichkeitsbetrachtung. Sie greifen dabei auf die in den folgenden Abschnitten beschriebenen Techniken zurück.

Lokale Features (Detektoren und Deskriptoren)

Den lokalen Features liegt die Idee zugrunde, dass Objekte eher an markanten Details als an globalen Bildeigenschaften erkannt werden können. Verfahren, die lokale Features nutzen, arbeiten in der Regel mehrstufig. In der ersten Stufe wird für ein Bild eine Menge solcher markanter Details in Form sogenannter *Merkmalspunkte* (Keypoints) identifiziert. Das können für ein Bild Hunderte bis zu einigen Tausend Keypoints sein. In der zweiten Phase werden Beschreibungen (in der Regel Beschreibungsvektoren) für die lokalen Bildcharakteristika in der engeren Umgebung für jeden Keypoint berechnet. Ein Bild wird nun also durch sehr viele Vektoren beschrieben, die jeweils charakteristische lokale Bildeigenschaften darstellen. Schließlich müssen Ähnlichkeitswerte zwischen den Beschreibungsvektoren berechnet werden, um festzustellen, ob zwei Bilder ähnliche markante Details enthalten.

Ein großer Vorteil dieser Verfahren ist, dass sie praktisch keine Trainingsdaten benötigen. Die Verfahren sind besonders erfolgreich in der Objekterkennung. Wenn das Bild eines Objektes gegeben ist, dann können recht effektiv Bilder ermittelt werden, die dieses Objekt enthalten. Dabei nutzt das Verfahren aus, dass es asymmetrisch arbeiten kann. Es werden Bilder gesucht, die eine signifikante Anzahl von Beschreibungsvektoren enthalten, die sehr ähnlich zu den Beschreibungsvektoren aus dem Anfragebild sind. Dass im Bild noch zahlreiche weitere Keypoints vorkommen, zu denen es im Objektbild keine Entsprechung gibt, spielt hier keine Rolle. Es werden ja nicht ähnliche Bilder gesucht, sondern Bilder, die das gesuchte Objekt enthalten. Um die Genauigkeit des Verfahrens weiter zu erhöhen, können dabei noch Verfahren nachgelagert werden, die prüfen, ob die

27 Joseph Redmon und Ali Farhadi, *YOLOv3: An Incremental Improvement*, Ithaca 2018, URL: http://arxiv.org/abs/1804.02767 (04.01.19).

jeweils zueinander passenden Keypoints auch in einer ähnlichen geometrischen Anordnung in beiden Bildern vorkommen.[28]

Einen Meilenstein in der Entwicklung der lokalen Features bilden die durch die *Scale-Invariant Feature Transform* (SIFT)-Methodik gewonnen Features. Diese bereits 1999 eingeführten Features werden noch heute in ihrer ursprünglichen Methodik aber auch in angepasster Form in der *Computer Vision* eingesetzt.[29] Wie der Name bereits vermuten lässt, sind diese Features robust gegen Skalierung, d.h. die Größe, in der ein Objekt in einem Bild vorkommt, sollte die Erkennung kaum beeinträchtigen. Ebenso sind sie rotationsinvariant und robust gegen die Positionierung im Bild.[30] Wie diese Invarianz erzeugt wird, wird im Folgenden kurz erläutert.

Im ersten Schritt werden auf das Bild unterschiedlich starke *Gaußfilter* (Weichzeichner) angewendet. Die dadurch entstehende Abfolge, bestehend aus dem ursprünglichen Bild und immer stärker weichgezeichneten Versionen davon, wird als Oktave bezeichnet. In einem zweiten Schritt wird das Bild auf 1/4 der Pixel reduziert (in Höhe und Breite halbiert) und wieder durch *Gaußfilter* weichgezeichnet, um so eine neue Oktave zu generieren. Diese Operationen werden mehrmals wiederholt. Die Kombination der Oktaven wird als *Scale Space* bezeichnet. Im Anschluss wird zwischen jedem nebeneinanderliegenden *Gaußfilter*-Paar in den Oktaven die *Difference of Gaussians* (DoG) berechnet; die Bilder werden voneinander subtrahiert, um nur noch die Feinheiten zu erhalten, die durch das Weichzeichnen verloren gegangen sind. Im nächsten Schritt werden Maxima beziehungsweise Minima gesucht. Hierbei werden die direkt umliegenden Nachbarpixel eines Pixels (8 Pixel) sowie die Nachbarpixel auf der Achse der Ergebnisbilder der DoG betrachtet (also im vorherigen und nachfolgenden DoG-Bild). Da so jeweils 9 Pixel von diesen Bildern mitbetrachtet werden, hat man insgesamt 26 Nachbarpixel für jedes betrachtete Pixel. Wenn der Wert für ein Pixel immer kleiner beziehungsweise größer ist als der aller Nachbarpixel, ist dieses ein Kandidat für einen Keypoint.

Nach einer weiteren Nachbearbeitung dieser Punkte wird jedem gefundenem Keypoint eine Referenzorientierung zugewiesen. Diese Orientierung basiert auf der maximalen Veränderung zu den umliegenden Pixeln (Stärke des Grauwertabfalls bzw. -anstiegs in der Richtung). Dazu wird der Umkreis in 36 Sektoren von je 10° eingeteilt und die Orientierung so rotiert, dass die maximale Steigung nach oben zeigt. Auf diese Weise wird die *Rotationsinvarianz* sichergestellt. Nun werden zu dem direkten Umfeld des Keypoints (typischerweise 16x16 Pixel) für

28 Mahmoud Hassaballah u.a., „Image Features Detection, Description and Matching", in: *Image Feature Detectors and Descriptors. Foundations and Applications*, hg. von Ali Ismail Awad und Mahmoud Hassaballah, Cham 2016 (Computational Intelligence 630), S. 11–45.

29 Arnold W. M. Smeulders u.a., „Content-Based Image Retrieval at the End of the Early Years", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (2010), S. 1349–1380.

30 David G. Lowe, „Object Recognition from Local Scale-Invariant Features", in: *Proceedings of the International Conference on Computer Vision* (1999), S. 1150–1157.

16 (je 4x4) Pixel große Teilbereiche die Steigungen für 8 Sektoren von je 45° berechnet. Daraus ergibt sich ein Beschreibungsvektor mit $16 \cdot 8 = 128$ Komponenten.[31] (Vgl. Abb. 1.4, Farbteil.)

Neben der *Positionsinvarianz*, die sich aus der Betrachtung der Keypoints unabhängig von ihrer Lage ergibt, und der *Rotationsinvarianz*, die sich durch die Referenzorientierung ergibt, stellt das Verfahren die *Skalierungsinvarianz* durch die Betrachtung verschiedener Skalierungs- sowie Unschärfeebenen sicher. Das Verwerfen von „schwachen" Keypoints sowie das Normalisieren der Beschreibungsvektoren (Histogramme) führen zu Keypoints, die weitestgehend unabhängig von ‚Rauschen' im Bild sowie globalen Beleuchtungseinflüssen sind und sich damit sehr gut für die Objekterkennung eignen.

Ausgehend von SIFT wurden zahlreiche Varianten und Optimierungen entwickelt. Bei den *Speeded Up Robust Features* (SURF) handelt es sich um eine verbesserte und insbesondere beschleunigte Variante. Dabei werden die wesentlichen mathematischen Verfahren durch Näherungen ersetzt, die sich schneller berechnen lassen ohne das Ergebnis nennenswert zu verschlechtern.[32]

Binary Robust Independent Elementary Features (BRIEF) wurden für Echtzeitszenarien konzipiert. Weitere Einsatzgebiete sind die Nutzung auf mobilen Endgeräten oder bei sehr großen Datensätzen.[33] Aufwendige Vergleiche werden durch eine Binärisierung der Feature-Vektoren performanter gestaltet.

Features from Accelerated Segment Test (FAST) ist ein Feature- bzw. Keypoint-Detektor, der Ecken um ein Pixel sucht. Dazu werden die Pixel auf einem Kreis mit Radius 3 um das Pixel herum geprüft. Liegen von diesen 16 Pixeln 12 entweder über oder unter einem Schwellenwert, wird eine Ecke unterstellt. Durch weitere Optimierungen können die Beschreibungen in Echtzeit ermittelt werden.[34] Als Weiterentwicklung von FAST, das keine Information bezüglich der Orientierung enthält und damit nicht rotationsinvariant ist, wurde *Oriented FAST* als Teil von *Oriented FAST* and Rotated BRIEF (ORB) eingeführt. Hierbei wird der FAST Detektor um Informationen zur Orientierung erweitert.[35]

---

31 David G. Lowe, *Method and Apparatus for Identifying Scale Invariant Features in an Image and Use of Same for Locating an Object in an Image*, U.S. Patent US6711293B1, application granted 2004.

32 Herbert Bay u.a., „Speeded-Up Robust Features (SURF)", in: *Computer Vision and Image Understanding* 110/3 (2008), S. 346–359.

33 Michael Calonder u.a., „BRIEF: Binary Robust Independent Elementary Features", in: *Computer Vision – ECCV 2010*, hg. von Kostas Daniilidis u.a., Berlin 2010 (Lecture Notes in Computer Science 6314), S. 778–792.

34 Edward Rosten und Tom Drummond, „Fusing Points and Lines for High Performance Tracking", in: *Proceedings of the 10th IEEE International Conference on Computer Vision* (2005), Bd. 2, S. 1508–1515; Edward Rosten u.a., „Faster and better: A Machine Learning Approach to Corner Detection", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32/1 (2010), S. 105–119.

35 Ethan Rublee u.a., „ORB: An Efficient Alternative to SIFT or SURF", in: *Proceedings of the 2011 IEEE International Conference on Computer Vision*, S. 2564–2571.

Ähnlichkeit und Matching

Im Abschnitt *Anfragearten und Anwendungen* haben wir bereits dargelegt, dass sehr unterschiedliche Anfragen bzw. Informationsbedürfnisse in der Bildanalyse und Bildsuche adressiert werden können. Eine wichtige Grundoperation ist dabei immer ein Anfragebild und ein Bild aus einem Datenbestand im Hinblick auf ihre Ähnlichkeit oder Passung für die aktuelle Fragestellung zu vergleichen. Im Fall der Farbhistogramme aus dem Abschnitt *Klassische Bildfeatures* wird z.B. gerne der sogenannte Histogramm-Schnitt gewählt, um die Ähnlichkeit von zwei Bildern zu messen. Dazu wird zu jeder Farbklasse der minimale Histogrammwert der beiden zu vergleichenden Bilder gewählt und über diese Minima die Summe über alle Farbklassen gebildet. Ist diese Summe hoch, so finden sich offensichtlich viele Pixel in gleichen Farbklassen, was einer hohen Ähnlichkeit entspricht.

Bereits für globale Farbeigenschaften lassen sich auch deutlich leistungsfähigere Ähnlichkeitsmaße anwenden, wie z.B. die *Earth Mover's Distance* (EMD), bei der berücksichtigt wird, dass der Unterschied zwischen einem hellen und einem dunklen Rot kleiner ist, als zwischen Rot und Blau.[36] Für die im vorhergenden Abschnitt vorgestellten lokalen Features stellt sich das Problem aber nochmals anders dar, weil hier die beiden zu vergleichenden Bilder nicht durch jeweils einen Beschreibungsvektor, sondern durch eine Menge von Beschreibungsvektoren für die Keypoints repräsentiert werden. Daher soll im Folgenden auf Ansätze zum Matching in diesem Kontext eingegangen werden.

Der naheliegende erste Ansatz zum Vergleich von Bildern auf Basis von SIFT- oder BRIEF-Features beziehungsweise Keypoints wird auch *Brute-Force Matcher* genannt. Dieses Verfahren vergleicht jeden Beschreibungsvektor des einen Bildes mit jedem Beschreibungsvektor des anderen Bildes. Der hieraus resultierende Rechenaufwand ist allerdings sehr hoch.

In Abbildung 1.4, Farbteil, unten werden zur Veranschaulichung ein Emblem sowie die Seite, die dieses Emblem enthält, dargestellt. Jedes Feature der einen Quelle (Suchbild, hier das linke Bild) wird mit jedem Feature der anderen Quelle (Referenzbild, hier das rechte Bild) verglichen. Unterhalb der Bilder in Abbildung 1.4 ist dieser Vorgang exemplarisch für ein Feature des Referenzbildes dargestellt. Dieser Vergleich wird in der Regel mit Hilfe eines Ähnlichkeitsmaßes wie dem oben skizzierten Histogramm-Schnitt durchgeführt. Nachdem die Ähnlichkeiten von dem Feature des Suchbildes zu den Features des Referenzbildes berechnet wurden, wird das Feature im Referenzbild als Treffer betrachtet, das die geringste Distanz aufweist, sofern diese einen vorgegebenen Schwellenwert nicht unterschreitet. Anschließend wird für jedes weitere Feature des Suchbildes analog verfahren. Die Passung des Referenzbildes zum Suchbild kann dann durch die Anzahl der Features gemessen werden, zu denen ein passendes Gegenstück gefunden wurde.

36 Christian Beecks, Distance-based Similarity Models for Content-based Multimedia Retrieval, PhD Thesis, RWTH Aachen University 2013.

In den Beispielen in Abbildung 1.4, Farbteil, wurden jeweils die 20 besten Feature-Matches durch Linien verbunden. Dies verdeutlicht, dass das Verfahren in einem klassischen Anwendungsfall (oben) recht gut funktioniert. Hier werden 18 der 20 Features korrekt angezeichnet. Bei dem unteren Beispiel aus dem *Emblematica*-Datensatz werden dagegen nur vier Features korrekt zugeordnet. Die Übertragung eines im klassischen Bereich der Objekterkennung recht gut funktionierenden Ansatzes auf geisteswissenschaftliche Bildsammlungen ist damit nicht ohne Weiteres möglich.

Wie bereits angedeutet, sind für größere Datenbestände andere Verfahren als *Brute-Force-Matching* notwendig. Eine Herangehensweise ist der *Bag of Visual Words* (BoVW). ‚Wörter' bezeichnen hier nicht Wörter im eigentlichen Sinn, sondern einzelne charakteristische Keypoints, die durch SIFT, SURF oder andere lokale Algorithmen zur Feature-Generierung gefunden werden. Die Bezeichnung greift die Analogie zum Text-Retrieval auf, in dem ein Dokument oft durch einen Bag of Words repräsentiert wird – d.h. nur durch seine Wörter, unabhängig von Grammatik, Position etc. Dabei kann man sich in der Regel auf eine kleine Zahl von für das Dokument charakteristischen Wörtern beschränken, deren Häufigkeiten in einem dünn besetzten Vektor repräsentiert werden. Für derartige Vektoren existieren zudem im Text Retrieval mit invertierten Listen effiziente Indexstrukturen, die eine schnelle Suche ermöglichen. Analog dazu wird nun für den *Bag of Visual Words* die Vorkommenshäufigkeit von Visual Words in einem Vektor dargestellt.[37]

Der erste Schritt ist hier die *Merkmalsextraktion* (Feature Extraction). Dabei werden vorhandene Verfahren wie die Generierung von SIFT-Features eingesetzt. Zur Erzeugung eines Vokabulars von visuellen Wörtern werden für einen gesamten Datenbestand von Bildern die Features extrahiert. Die einzelnen Features werden dann geclustert. Dies wird durch Verfahren wie das *k-Means-Clustering* umgesetzt. Eine Herausforderung hierbei ist die Festlegung des Parameters k – der Anzahl der Cluster und damit der Größe des visuellen Vokabulars. Sind passende Cluster-Center bestimmt, kann ein Bild über einen Vektor der Dimension k beschrieben werden, der angibt, wie viele Keypoints des Bildes zu welchem Cluster gehören. Neben der Verbesserung der Performance wird mit diesem Verfahren auch eine durchaus gewünschte Abstraktion von zu feinen visuellen Details hin zu visuellen Konzepten – eben den visuellen Wörtern – erreicht.

(Deep) Learning – Convolutional Neural Networks

Seit einigen Jahren finden *Deep Learning-Verfahren* bzw. *neuronale Netze* in zahlreichen Szenarien und auch in der Bildanalyse mehr und mehr Anwendung. Ein Neuron wird dabei durch eine mathematische Funktion repräsentiert, die

37 Jun Yang u.a., „Evaluating Bag-of-Visual-Words Representations in Scene Classification", in: *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR '07),* New York 2007, S. 197–206.

Eingangssignale bewertet und ein entsprechendes Ausgangssignal erzeugt. Eine wesentliche Voraussetzung für die Anwendung ist dabei die Verfügbarkeit umfangreicher Trainingsdaten, um die Funktionen bzw. deren Parameter lernen zu können. Im Fall der Bildanalyse bedeutet dies die Verfügbarkeit eines großen Datenbestands mit annotierten Trainingsdaten. Eine weitere Aufgabe besteht darin, die Architektur des Netzes zu optimieren. Dabei werden aktuell recht tiefe Architekturen mit vielen Ebenen von Neuronen eingesetzt. [38]

Deep Learning ist aber kein neues Konzept, sondern schon Jahrzehnte bekannt, es gewann jedoch in den letzten Jahren durch das Thema *Big Data* und die Berechnung auf speziellen Clustern (auf Basis von High-End-Grafikkarten) neue Bedeutung in der praktischen Umsetzbarkeit. Hinzu kam die Verfügbarkeit entsprechender Software (z.B. Google's Tensorflow).

Deep Learning generell versucht mit Hilfe eines Netzes von künstlichen Neuronen und deren Verbindungen einen Datenoutput für einen bestimmten Dateninput zu generieren. In unserem Szenario könnten beispielsweise extrahierte Features eines Bildes oder auch die Bildinformationen selbst als Input des neuronalen Netzes dienen. Der Output könnte eine bestimmte Klassifikation sein, die ggf. eine Segmentierung einschließen kann.

Alle Inputwerte (Zahlen) eines Neurons werden mit Gewichten multipliziert und aggregiert und berechnen so den Wert eines Neurons des ersten sogenannten *Hidden Layers*. Dies wird für alle Neuronen dieser Ebene mit unterschiedlichen Gewichten durchgeführt. Anschließend wird der Ausgabewert des Neurons durch eine Aktivierungsfunktion, beispielsweise *Rectified Linear Units* (ReLU) berechnet. Diese ist unter anderem für die Nicht-Linearität des Modells verantwortlich. Für alle weiteren *Hidden Layers* ist es dasselbe Vorgehen. Zum Schluss wird aus den Werten des letzten *Hidden Layers* das Ergebnis, der Output, wie zum Beispiel die Wahrscheinlichkeit für mehrere Klassen wie ‚Fuchs', ‚Engel', ‚Himmel' oder ‚Kopf' berechnet. An dieser Stelle setzt das *Lernen* (Learning) des *Deep-Learning* an, indem mit einer Verlustfunktion für einen Trainingsdatenbestand die Abweichung von den korrekten Klassen (bei Abb. 1.5: Fuchs 1, Engel 0, Himmel 0, Kopf 0) und den ausgegeben Klassen des Algorithmus (beispielsweise Fuchs 0,3, Engel 0,3, Himmel 0,3, Kopf 0,1) berechnet wird. Es wird in mehreren Iterationen (Epochen) versucht, diesen Verlust zu minimieren. Dazu werden die Gewichte im Netz rekursiv aktualisiert, wobei die Aktualisierung jeweils so vorgenommen wird, dass der Fehler im nächsten Durchlauf reduziert wird (Gradient Descent). Wichtig ist hier, dass zu viel Training auch zu schlechteren Ergebnissen führen kann, da dann ein Overfitting auftritt: Das Netz lernt die Testbilder ‚auswendig', kann aber neue Bilder nicht mehr erkennen.

38 Alex Krizhevsky u.a., „ImageNet Classification with Deep Convolutional Neural Networks", in: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Bd. 1, hg. von Fernando Pereira u.a., Red Hook 2012, S. 1097–1105; Ian Goodfellow u.a., *Deep Learning*, Cambridge 2016.

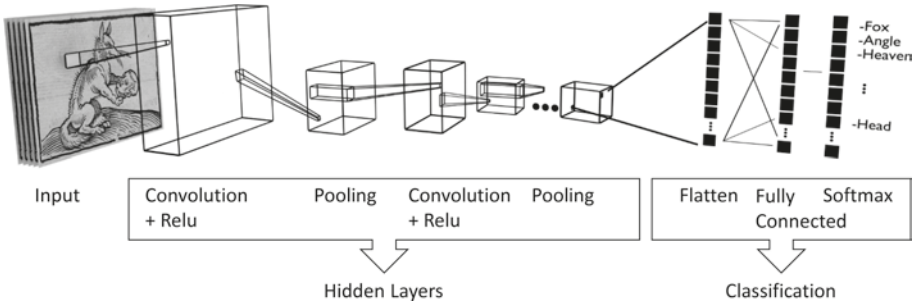

Abb. 1.5: Beispiel für eine Architektur eines *Deep Learning*-Netzwerks; Abbildung, Martin Bullin (CC BY-NC-ND), Pictura by permission of University of Glasgow Library, Archives & Special Collections.

Ein weiterer Punkt speziell für die Bildanalyse ist, dass die Bilder für die neuronalen Netze aufbereitet werden müssen. Hierbei ist es nicht sinnvoll, die Bilder pixelweise mit ihren RGB-Werten als Input für die sogenannten *Fully-Connected-Layers* zu nutzen. Daher werden dem eigentlichen neuronalen Netz weitere Ebenen zur Extraktion und zum Lernen von relevanten Features vorgelagert. Dieser Vorgang wird in der Regel durch Convolution-Layers umgesetzt. Hierbei werden die Pixelwerte des Bildes mit einer Faltungsmatrix multipliziert. Dies erinnert an den *Gaußfilter*, der bei SIFT angewendet wird, und in der Tat unterscheidet sich der *Deep Learning*-Ansatz bei der Merkmalsextraktion nicht wesentlich von dem der klassischen Features. Im Gegensatz zu den klassischen Features werden die Filter aber bei DL durch das Netz selbst trainiert, gelernt und verbessert.

Als Input werden die pixelweisen Graustufen oder auch RGB-Werte genutzt (vgl. Abb. 1.5, Input). Diese werden mit einem Filter multipliziert und ergeben so das erste Zwischenergebnis. Anschließend wird die ReLU-Funktion angewendet und alle negativen Werte auf null gesetzt. Hierdurch wird die Nichtlinearität des Netzes geschaffen. Anschließend wird durch *Pooling* (auch *Subsampling*) die Datenmenge reduziert. Hierbei können einfache Funktionen genutzt werden wie beispielsweise die Max-Funktion: Für einen Pixelblock von 2 x 2 Pixeln wird der höchste Wert ausgewählt und dies für alle Pixelblöcke wiederholt. Mit iterativer Ausführung der Schritte kann die Genauigkeit des Ergebnisses verbessert werden. In Abbildung 1.5 wurden zwei Iterationen genutzt. Anschließend wird das Ergebnis eindimensional weiterverarbeitet. Hier setzt dann ein neuronales Netz ein, wie es bereits zuvor beschrieben wurde.

Für die Nutzung von Verfahren des *Deep Learning* ist neben einem entsprechenden umfangreichen Satz an Trainingsdaten auch eine angemessene Hardware vonnöten, da die Verfahren sehr rechenintensiv sind. Auch die Verwendung eines passenden Netzes ist wichtig. Dieses kann man selbst konzipieren, oft werden aber bewährte Architekturen genutzt oder angepasst. Ein Beispiel ist das bereits

im Abschnitt *Segmentierung* erwähnte YOLO (s. auch Abb. 1.3, Farbteil, rechts).[39] Dabei verfügen diese Netze über sehr viele Ebenen. Ein Vorteil der Ebenen ist, dass dazwischen neue Informationen, sogenannte Repräsentationen, gebildet werden, die eine Abstraktion der eigentlichen Eingangssignale sind. Diese Repräsentationen kann man grob mit den visuellen Wörtern beim BoVW vergleichen. Man bezeichnet dies auch als *Representation Learning*. So können DL-Modelle zum Teil recht gut vom ursprünglichen Trainingsdatenbestand abstrahieren. Die Ebenen ermöglichen auch, den Aufwand und den Bedarf an Trainingsdaten zu reduzieren, da für einen neuen Datenbestand nicht unbedingt das ganze Netz sondern nur einzelne Ebenen neu trainiert werden müssen.

Anwendungsbeispiele

In den vorangegangenen Abschnitten wurden bereits verschiedene Anwendungsbeispiele erwähnt. Im Folgenden werden wir dies noch ausbauen, um die Möglichkeiten und Anwendungsbereiche der Methoden exemplarisch aufzuzeigen.

Chanjong Im u.a. haben mit Hilfe von *Deep Learning Verfahren* untersucht, inwieweit es möglich ist, die Produktionsart der Bilder in Büchern des 19. Jahrhunderts festzustellen.[40] Hier wurde zwischen den Klassen ‚Holzstich' und ‚Lithographie' unterschieden. Gelernt wurde mit einem vergleichsweise kleinen Datensatz von jeweils 2.235 Bildausschnitten pro Klasse. Die relativ niedrige Accuracy von 63% wird dabei auf eben diese geringe Anzahl an Trainingsbildern sowie den Ansatz, Bildausschnitte anstelle von ganzen Bildern zu nutzen, zurückgeführt.

Die *Bayerische Staatsbibliothek* (BSB) hatte bereits 2017 1,2 von 11 Mio. Bänden digitalisiert. Die inhaltliche Erschließung hinkt aufgrund des Umfangs allerdings hinterher. Um die Bilder aus den Werken zugänglich zu machen, wurde daher auf Methoden des *Image Retrieval* zurückgegriffen. Es wurden ML-Methoden genutzt, um aussagekräftige Bilder zu selektieren. Hierdurch konnten 43 Mio. Bilder identifiziert werden. Ebenfalls wurde eine QbE-Funktionalität umgesetzt, die es zulässt, ähnliche Bilder zu einem selbst gewählten bzw. hochgeladenen Bild zu suchen. Diese visuelle Suche basiert auf Deskriptoren, die Farb-, Kanten- und Texturmerkmale zusammenfassen, wobei die Gewichtung der Kriterien vom Nutzer angepasst werden kann.[41]

Ein anderes Beispiel hatte die Suche nach bestimmten ikonographischen Elementen im bereits erwähnten *Emblematica Online*-Bestand zum Ziel. Dabei handelt es sich um eine weltweite Sammlung von Emblemen, die durch die Universität von Illinois bereitgestellt werden. Hier liegen 33.268 Embleme vor, die teilweise bereits durch Kategorien nach Iconclass ausgezeichnet sind. Im Ideal-

39 Redmon und Farhadi, *YOLOv3*.

40 Chanjong Im u.a., „Deep Learning Approaches to Classification of Production Technology for 19th Century Books", in: *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen"*, hg. von Rainer Gemulla u.a., Aachen 2018 (*CEUR Workshop Proceedings* 2191), S. 150–158.

41 Markus Brantl u.a., „Visuelle Suche in historischen Werken", in: *Datenbank Spektrum* 17 (2017), S. 53–60.

fall sind diese Auszeichnungen auch einem Bildsegment zugeordnet, was jedoch nicht immer der Fall ist. Hollender u.a. haben ein Werkzeug vorgestellt, das es Forschern ermöglichen soll, Ähnlichkeit, Identitäten und weitere Beziehungen von Emblemen zu teilen und zu nutzen. Hierbei ist es oft von Relevanz, ob die Abbildungen von demselben Holzschnitt stammen, der sich aber über die Zeit verändert haben kann.[42] Die große Datenmenge sowie die bereits ausgezeichneten Daten legen hier nahe, DL-Verfahren einzusetzen. Erste Ansätze in einer Masterarbeit sowie einem Projekt an der Universität Bamberg haben bisher aber auf Grund der starken Abweichung der Picturae von normalen Bilddatensätzen keine zufriedenstellenden Klassifikationsergebnisse liefern können. Vielmehr zeigten sich hier lokale Merkmale gegenüber den mit DL-Verfahren erzielten Ergebnissen sogar leicht überlegen. Dies deutet an, dass DL-Verfahren nicht in allen Szenarien das Mittel der Wahl sind und auch einer fundierten Anpassung und Adaption bedürfen. Die Arbeiten von Bermeitinger u.a. zur Objektklassifizierung in Bildern neoklassischer Kunstgegenstände mittels DL zeigen zwar bessere Ergebnisse, auch hier liegen die Erkennungsraten aber in einem Bereich, der für eine praktische Anwendung noch zu niedrig liegt.[43]

Zusammenfassung und Ausblick

Bereits aus den Überlegungen zur sensorischen und semantischen Lücke lassen sich für die inhaltsbasierte Bildsuche und Bilderschließung verschiedene Konsequenzen ableiten: Sofern die (Retro-)Digitalisierung selbst beeinflusst werden kann, sollte darauf geachtet werden, die sensorische Lücke durch eine professionelle und zielgerichtete Digitalisierung so gering wie möglich zu halten. Bei der Analyse der Bilder kann dann mit gewissen Erfolgen gerechnet werden, sofern man die untere semantische Lücke adressiert. Objekterkennung und inhaltsbasierte Bildsegmentierung können heute zum Teil beachtliche Erfolge vorweisen. Um die obere semantische Lücke zu adressieren ist zumindest eine umfangreiche Modellierung des Domänenwissens und des Bildkontextes erforderlich. Letztlich sind aber auch dann nur begrenzte Erfolge zu erwarten, da die Interpretation eines Bildes häufig großen Spielraum lässt. Wichtig ist in jedem Fall die zielgerichtete Auswahl der eingesetzten Technologien für den konkreten Anwendungszweck. Dabei sollten die Anforderungen im Anwendungskontext ebenso berücksichtigt werden wie die Verfügbarkeit von Trainingsdaten.

42 Kurt Hollender u.a., *Annotation of Digitized Emblematica (Illinois Annotation Experiment Final Report)*, URL: http://www.openannotation.org/Partners.html (05.06.2019).

43 Bernhard Bermeitinger u.a., „Object Classification in Images of Neoclassical Artifacts Using Deep Learning", in: *Digital Humanities 2017: Conference Abstracts*, hg. von Rhian Lewis u.a., Montréal 2017, S. 395–397.

Literaturverzeichnis

Bay, Herbert, Tinne Tuytelaars und Luc Van Gool, „Speeded-Up Robust Features (SURF)", in: *Computer Vision and Image Understanding* 110/3 (2008), S. 346–359.

Beecks, Christian, Distance-based Similarity Models for Content-based Multimedia Retrieval, PhD Thesis, RWTH Aachen University 2013.

Belongie, Serge, Jitendra Malik und Jan Puzicha, „Shape Matching and Object Recognition Using Shape Contexts", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24/24 (2002), S. 509–522.

Bermeitinger, Bernhard, Simon Doing, Maria Christoforaki, André Freitas und Siegfried Handschuh, „Object Classification in Images of Neoclassical Artifacts Using Deep Learning", in: *Digital Humanities 2017: Conference Abstracts*, hg. von Rhian Lewis, Cecily Raynor, Dominic Forest, Michael Sinatra und Stéfan Sinclair, Montréal 2017, S. 395–397.

Brantl, Markus, Klaus Ceynowa, Thomas Meiers und Thomas Wolf, „Visuelle Suche in historischen Werken", in: *Datenbank Spektrum* 17 (2017), S. 53–60.

Broder, Andrei, „A taxonomy of web search", in: *ACM SIGIR forum* 36/2 (2002), S. 3–10.

Calonder, Michael, Vincent Lepetit, Christoph Strecha und Pascal Fua, „BRIEF: Binary Robust Independent Elementary Features", in: *Computer Vision – ECCV 2010*, hg. von Kostas Daniilidis, Petros Maragos und Nikos Paragios, Berlin 2010 (Lecture Notes in Computer Science 6314), S. 778–792.

Canny, John, „A Computational Approach to Edge Detection", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), S. 679–698.

CIDOC Conceptual Reference Model, URL: http://www.cidoc-crm.org, (05.06.2019).

Davis, Larry S., „A survey of edge detection techniques", in: *Computer Graphics and Image Processing* 4/3 (1975), S. 248–260.

Dublin Core Metadata Initiative, URL: http://dublincore.org (05.06.2019)

Eidenberger, Horst: *Handbook of Multimedia Information Retrieval*, Wien 2012.

Europeana Data Model, URL: https://pro.europeana.eu/resources/standardization-tools/edm-documentation, (05.06.2019).

Exchangable Image File Format, URL: http://exif.org (05.06.2019).

Felzenszwalb, Pedro F. und Daniel P. Huttenlocher, „Pictorial Structures for Object Recognition", in: *International Journal of Computer Vision* 61/1 (2005), S. 55–79.

Ferencz, Andras, Erik G. Learned-Miller und Jitendra Malik, „Learning to Locate Informative Features for Visual Identification", in: *International Journal of Computer Vision* 77/1–3 (2008), S. 3–24.

Fischler, Martin A. und Robert A. Elschlager, „The Representation and Matching of Pictorial Structures", in: *IEEE Transactions on Computers* 22/1 (1973), S. 67–92.

Flickner, Myron, Harpreet Sawhney, Wayne Nibleck, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele und Peter Yanker, „Query by Image and Video Content: The QBIC System", in: *Computer* 28/9 (1995), S. 23–32.

Goodfellow, Ian, Yoshua Bengio und Aaron Courville, *Deep Learning*, Cambridge 2016.

Grubinger, Michael, *Analysis and evaluation of visual information systems performance*, PhD Thesis, Victoria University Melbourne 2007, URL: http://vuir.vu.edu.au/1435/ (04.02.19).

Hare, Jonathon S., Paul H. Lewis, Peter G. B. Enser und Christine J. Sandom, „Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval",

in: *Multimedia Content Analysis, Management and Retrieval 2006*, hg. von Alan Hanjalic, Nicu Sebe und Edward Y. Chang, Bellingham 2006 (SPIE Proceedings 6073), S. 75–86.

Hartley, Richard, Rajiv Gupta und Tom Chang, „Stereo from uncalibrated cameras", in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1992), S. 761–764.

Hassaballah, Mahmoud, Aly Amin Abdelmgeid und Hammam A. Alshazly, „Image Features Detection, Description and Matching", in: *Image Feature Detectors and Descriptors. Foundations and Applications*, hg. von Ali Ismail Awad und Mahmoud Hassaballah, Cham 2016 (Computational Intelligence 630), S. 11–45.

Herzog August Bibliothek Wolfenbüttel, URL: http://www.hab.de/de/home/wissenschaft/ forschungsprofil-und-projekte/emblematica-online.html, (07.12.2018).

Hollender, Kurt, Jacob Jett, Jessica Nicholas, Jordan Vannoy, Timothy Cole, Myung-Ja Han, Thomas Kilton und Mara Wade, *Annotation of Digitized Emblematica (Illinois Annotation Experiment Final Report)*, URL: http://www.openannotation.org/Partners.html (05.06.2019).

Iconclass, URL: http://www.iconclass.nl/home, (31.12.2018).

Im, Chanjong, Junaid Ghauri, John Rothman und Thomas Mandl, „Deep Learning Approaches to Classification of Production Technology for 19th Century Books", in: *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen"*, hg. von Rainer Gemulla, Simone Paolo Ponzetto, Christian Bizer, Margret Keuper und Heiner Stuckenschmidt, Aachen 2018 (*CEUR Workshop Proceedings* 2191), S. 150–158.

Jansen, Bernard J., Danielle L. Booth und Amanda Spink, „Determining the informational, navigational, and transactional intent of Web queries", in: Information Processing & Management 44/3 (2008), S. 1251–1266.

Krizhevsky, Alex, Ilya Sutskever und Geoffrey E. Hinton, „ImageNet Classification with Deep Convolutional Neural Networks", in: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Bd. 1, hg. von Fernando Pereira, Christopher J. C. Burges, Léon Bottou und Kilian Q. Weinberger, Red Hook 2012, S. 1097–1105.

LeCun, Yann, Yoshua Bengio und Geoffrey Hinton., „Deep learning", in: *Nature* 521 (2015), S. 436–444.

Lightweight Information Describing Objects, URL: http://network.icom.museum/cidoc/ working-groups/lido/what-is-lido (05.06.2019).

Lowe, David G., *Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image*, U.S. Patent US6711293B1, application granted 2004.

—, „Object Recognition from Local Scale-Invariant Features", in: *Proceedings of the International Conference on Computer Vision* (1999), S. 1150-1157.

Manjunath, Bangalore S., Philippe Salembier und Thomas Sikora, *Introduction to MPEG-7*: *multimedia content description interface*, Bd. 1, Chichester 2002.

Marchionini, Gary, „Exploratory search: from finding to understanding", in: *Communications of the ACM* 49/4 (2006), S. 41–46.

Mori, Greg, Xiaofeng Ren, Alexei A. Efros und Jitendra Malik, „Recovering Human Body Configurations: Combining Segmentation and Recognition", in**:** *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Bd. 2, S. 326–333.

Mumford, David und Jayant Shah, „Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems“, in: *Communications on Pure and Applied Mathematics* 42/5 (1989), S. 577–685.

Nalwa, Vishvjit S. und Thomas O. Binford, „On Detecting Edges“, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8/6 (1986), S. 699–714.

National Information Standards Organization, URL: http://dublincore.org, (05.06.2019).

Pätzold, Walter, „Digitale Bildbearbeitung“, in: *Taschenbuch der Medieninformatik* hg. von Kai Bruns und Klaus Meyer-Wegener, Leipzig 2005, S. 164–165.

Ponce, Jean, Martial Hebert, Cordelia Schmid und Andrew Zisserman, *Toward Category-Level Object Recognition*, Berlin 2006.

Redmon, Joseph und Ali Farhadi, *YOLOv3: An Incremental Improvement*, Ithaca 2018, URL: http://arxiv.org/abs/1804.02767 (04.01.19).

Rosten, Edward, Reid Porter und Tom Drummond, „Faster and better: A Machine Learning Approach to Corner Detection“, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32/1 (2010), S. 105–119.

— und Tom Drummond, „Fusing Points and Lines for High Performance Tracking“, in: *Proceeding of the 10th IEEE International Conference on Computer Vision* (2005), Bd. 2, S. 1508–1515.

Rublee, Ethan, Vincent Rabaud, Kurt Konolige und Gary Bradski, „ORB: An Efficient Alternative to SIFT or SURF“, in: *Proceedings of the 2011 IEEE International Conference on Computer Vision*, S. 2564–2571.

Siggelkow, Sven, *Feature histograms for content-based image retrieval*, Dissertation, Albert-Ludwigs-Universität Freiburg 2002.

Smeulders, Arnold W. M., Marcel Worring, Simone Santini, Amarnath Gupta und Ramesh Jain, „Content-Based Image Retrieval at the End of the Early Years“, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (2010), S. 1349–1380.

—, Martin L. Kersten und Theo Gevers, „Crossing the Divide between Computer Vision and Data Bases in Search of Image Databases“, in: *Visual Database Systems 4*, hg. von Yannis E. Ioannidis und Wolfgang Klas, London 1998, S. 223–239.

Szeliski, Richard, *Computer Vision: Algorithms and Applications*, London 2010

Yang, Jun., Yu-Gang Jiang, Alexander G. Hauptmann und Chong-Wah Ngo, „Evaluating Bag-of-Visual-Words Representations in Scene Classification“, in: *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR '07)*, New York 2007, S. 197–206.