TECHNISCHE UNIVERSITÄT DRESDEN

ZENTRUM FÜR INFORMATIONSDIENSTE
UND HOCHLEISTUNGSRECHNEN
PROF. DR. WOLFGANG E. NAGEL

Master-Arbeit

zur Erlangung des akademischen Grades
Master of Science

Image Retrieval für Historische Bilder

Philipp Langen
(Geboren am 26. Dezember 1994 in Münsterlingen)

Hochschullehrer: Prof. Dr. Wolfgang E. Nagel
Betreuer: Dr. Christoph Lehmann & Dr. Taras Lazariv

Dresden, 10. Juni 2020

Hier Aufgabenstellung einfügen!

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag dem Prüfungsausschuss der Fakultät Informatik eingereichte Master-Arbeit zum Thema:

Image Retrieval für Historische Bilder

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Dresden, den 10. Juni 2020

Philipp Langen

Kurzfassung

Abstract

Inhaltsverzeichnis

Motivation

Präzise und effiziente Suchwerkzeuge sind essenziell um große Datenmengen für einen Nutzer sinnvoll verwertbar zu machen. Dies gilt insbesondere auch im Bereich der Bildersuche. Die klassische Bildersuche basiert auf vom Nutzer formulierten Anfragen, mit deren Hilfe das Suchsystem eine Liste an passenden Bildkandidaten zusammenstellt und zurückgibt. Hierbei nutzt das System eine Reihe von Zusatzinformationen, sogenannten Metadaten, wie Tags, Titel, Aufnahmeort oder Datum. Eine alternativer Ansatz der Suche, der in dieser Arbeit behandelt wird, ist die inhaltsbasierte Bildersuche (engl. Content-Based Image Retrieval, kurz CBIR). Hierbei werden vom Nutzer keine Anfragen formuliert. Stattdessen dient ein Bild als Suchanfrage. Ziel ist es, Bilder mit gleichem oder ähnlichem Bildinhalt als Ergebnis zurückzugeben. Ein Vorteil dieser Herangehensweise ist, dass der Nutzer keine Informationen über den Inhalt des Suchbildes benötigt. Das System arbeitet ausschließlich mit den Pixelinformationen der Bilder. Ein weiterer Vorteil ist daher, dass weder im Suchbild noch in der Suchdatenbank Metadaten zu den Bildern vorhanden sein müssen, was die Einsatzmöglichkeiten von inhaltsbasierter Suche sehr flexibel gestaltet. Im Folgenden wird die inhaltsbasierte Suche auch als Image Retrieval bezeichnet.

Das Anwendungsgebiet dieser Arbeit ist die Suche auf historischen Bildern. Diese weit gefasste Domäne ist besonders herausfordernd, da sie sehr heterogene Daten enthält. Dabei gibt es nicht nur große Unterschiede in den abgebildeten Bildinhalten, wie Gebäude, Naturaufnahmen oder Portraits, sondern auch in den verwendeten Aufnahmeverfahren. Durch die Fortschritte der Aufnahmetechnik können historische Bilder sowohl in Form von Zeichnungen oder Malerei, aber auch als Druck oder in anfänglichen Formen der Photographie vorliegen. Da Metadaten zu historischen Bildern erst bei der Digitalisierung hinzugefügt werden können, sind diese oft gar nicht oder nur lückenhaft vorhanden. Dies macht die inhaltsbasierte Suche für diese Domäne im Vergleich zur klassischen Bildersuche zu einer besonders geeigneten Methode. Mit der Umsetzung einer unterstützenden Suche für das UrbanHistory4D Projekt [1] ergibt sich ein konkreter Anwendungsfall für diese Arbeit. Das UrbanHistory4D Projekt befasst sich mit der Erstellung interaktiver Stadtkarten. Wo vorhanden kann sich der Nutzer historische Aufnahmen anzeigen lassen, die den Ort zeigen, an dem er sich innerhalb der Karte befindet. Das Akkumulieren und Zuordnen von historischen Bildern zu diesen Plätzen ist ein wesentlicher Arbeitsanteil bei der Erstellung der Karten. Image Retrieval Systeme können helfen den Suchaufwand für die Ersteller der Karten signifikant zu reduzieren. Dabei handelt es sich um einen aktiven Forschungsbereich, in dem momentan unterschiedliche Suchsysteme analysiert werden.

Das Image Retrieval Verfahren DELF (attentive DEep Local Features) [2] entwickelt von Noh, Araujo et al., welches in dieser Arbeit untersucht wird ist ein Deep Learning Ansatz. Durch den raschen Fortschritt im Bereich tiefer neuronaler Netzwerkarchitekturen der letzten Jahre erfreuen sich gelernte Ansätze immer größerer Beliebtheit. DELF erzielt auf bekannten Benchmarkdatensätzen wie Oxford5k [3] und Paris6k [4] sehr gute Ergebnisse. Besonders gut schneidet DELF im Vergleich auf dem eigens erstellten Google Landmarks Datensatz [5] ab. Dieser enthält mit über 1 Mio. Bilder und 13k unterschiedlichen Motiven eine deutlich heterogenere Mischung an Objekten als andere Benchmarks. Die gute Performanz auf diesem Datensatz lässt also hoffen, dass sich das DELF-Verfahren auch für die historische Domäne eignet.

Verwandte Arbeiten

Bei Information Retrieval handelt es sich um ein Problem aus dem Bereich der Computer Vision, welches bereits seit langem intensiv erforscht wird. In frühen Ansätzen versuchte man vor allem globale Beschreibungen von Bildern zu erstellen, um diese untereinander vergleichen zu können. Diese basierten zum Beispiel auf Farbhistogrammen oder Texturbeschreibungen [6]. Allerdings waren diese Ansätze oft sehr anfällig für Unterschiede in Beleuchtung, Skalierung und anderen Transformationen, wie sie bei unterschiedlichen Aufnahmen des selben Motivs auftreten können.
Ein wesentlicher Durchbruch gelang David G. Lowe 2004 mit der Entwicklung des SIFT-Verfahrens (Scale Invariant Feature Transform) [7]. Hierbei werden mehrere Konzepte vereint um Bildbeschreibungen zu erzeugen, die robuster gegenüber unterschiedlichen Transformationen sind. So arbeitet der SIFT Algorithmus beispielsweise nicht direkt auf den Bildern, sondern im sogenannten Scale Space. Dieser besteht auf unterschiedlich skalierten Versionen des Ursprungsbildes, auf welche wiederum unterschiedlich starke Gauß-Filter angewendete werden. Betrachtet werden schließlich Differenzbilder zwischen benachbarten Stärken der Gauß-Fitler Ergebnisse. Die Verwendung von unterschiedlich skalierten Bildversionen macht die berechneten SIFT-Merkmale deutlich robuster gegen Skalierungsunterschiede. Das SIFT-Verfahren besteht aus zwei Phasen. In der ersten Phase werden über die Suche nach lokalen Extrema bedeutsame Bildpunkte ausgewählt. Für diese werden in der zweiten Phase einzelne Deskriptoren berechnet. Das Bild wird also nicht global beschrieben, sondern über viele lokale Deskriptoren dargestellt. Die lokalen Deskriptoren ergeben sich aus Histogrammen der Gradientrichtungen umliegender Bildpunkte. Diese werden relativ zu der dominanten Gradientrichtung in der Umgebung berechnet, was die Deskriptoren invariant gegenüber Rotationen macht. Lowes Entwicklung bildet den Ursprung für viele abgeleitete Verfahren wie SURF[8], PCA-SIFT[9] und RIFT[10]. Auch in aktueller Forschung werden Image Retrieval Verfahren untersucht, die mit SIFT-Merkmalen arbeiten [11].
Der Trend bei der Entwicklung neuer Image Retrieval Systeme geht aktuell jedoch hauptsächlich in Richtung von gelernten Verfahren. Die Basis dieser Verfahren bilden tiefe CNN-Architekturen (Convolutional Neural Networks). Ein neuronales Netzwerk lässt sich als eine schichtweise Aneinanderreihung nichtlinearer Funktionen auffassen. Convolutional Neural Networks sind ein Sonderfall neuronaler Netze, welche sogenannte Convolutional Layer, zu deutsch faltende Schichten, enthalten. In diesen Schichten werden Faltungs/ bzw. Filteroperationen auf die Eingabedaten angewendet, um für das Netzwerk hilfreiche Merkmale in den Daten hervorzuheben. Dies ist durchaus vergleichbar mit den Filteroperationen, die im SIFT-Verfahren verwendet werden. Der Unterschied besteht jedoch darin, dass die Parameter der verwendeten Filtermasken, sowie aller anderen Netzparameter nicht per Hand gewählt werden, sondern in einem Trainingsverfahren für den aktuellen Anwendungsfall optimiert werden. Der Entwickler bestimmt lediglich die grobe Architektur des Netzwerks, also die Anzahl, Größe und Reihenfolge der verwendeten Schichten, sowie die Art der Operationen, die in ihnen durchgeführt werden. In Image Retrieval Systemen können CNNs eingesetzt werden, um Bilddeskriptoren zu erstellen. Hierfür werden Zwischenergebnisse des Netzwerks, also die Ausgaben einer bestimmten Schicht genutzt. An welcher Stelle im Netzwerk die Deskriptoren entnommen werden ist dabei von entscheidender Bedeutung. Zeiler und Fergus haben in ihrer Studie zur Visualisierung von CNNs gezeigt, dass die früheren Schichten von CNNs typischerweise einfache Konzepte, wie Kanten oder Ecken hervorheben. Mit wachsender Tiefe der betrachteten Netzwerkschicht steigt auch die Komplexität der Konzepte, die von den Ausgaben der

Schicht beschrieben werden können.
In dem in [12] beschriebenen Image Retrieval System von Babenko, Slesarev et al. wird als Modell ein CNN bestehend aus fünf faltenden gefolgt von drei vollvernetzten Schichten (im Englischen fully-connected layer) genutzt. Als Deskriptoren werden die Ausgaben, der ersten bzw. zweiten vollvernetzten Schicht verwendet. In einer vollvernetzten Schicht hat jeder Wert der Eingabe Einfluss auf jeden Wert in der Ausgabe. Die Ausgaben solcher Schichten werden also von der gesamten Bildeingabe beeinflusst und können daher als globale Deskriptoren verstanden werden. Bereits diese sehr direkte Herangehensweise erzielt leichte Verbesserung gegenüber den gängigen algorithmischen Verfahren zur Zeit der Veröffentlichung.
Razavian, Sullivan et al. stellen in [13] ein System auf Basis der in [14] beschriebenen Netzwerkarchitektur vor. Das Modell besteht ebenfalls aus fünf faltenden und drei vollvernetzten Schichten. Die Deskriptoren stammen aus den Ausgaben der letzten faltenden Schicht. Anders als bei Babenko, Slesarev et al. werden in diesem System mehrere Deskriptoren pro Bild erstellt. Hierfür werden systematisch Teilbilder aus Bildbereichen unterschiedlicher Größe generiert. Anschließend werden die Teilbilder auf eine feste Größe skaliert und als Eingabe in das Netzwerk gegeben. So kann für jeden betrachteten Bildbereich ein eigener lokaler Deskriptor erstellt werden. In ihren Experimenten stellen die Autoren fest, dass die Verwendung von lokalen Deskriptoren gegenüber einer globalen Betrachtung zu einer signifikanten Verbesserung der Retrievalperformanz führt. Der überwiegende Teil aktueller Retrieval Systeme setzt auf die Erstellung von lokalen Deskriptoren.
Eine interessante Frage bei der Konzeption von Image Retrieval Systemen, die mit lokalen Deskriptoren arbeiten ist, wie man entscheidet, welche Bildregionen am sinnvollsten zu betrachten sind. Das ONE-Verfahren [15] von Xie, Hong et al. nutzt ein VGG-19 [16] Modell und extrahiert Deskriptoren aus der vorletzten vollvernetzten Schicht. Als Eingaben in das Netzwerk dienen sogenannte Object Proposals. Dabei handelt es sich um Bildausschnitte, welche Regionen umschließen, in denen Objekte vermutet werden. Die Autoren testen sowohl manuell annotierte, sowie automatisch extrahierte Object Proposals und erzielen mit beiden Ansätzen ähnlich gute Ergebnisse. Für die automatische Bestimmung von Object Proposals nutzen sie das Selective Search Verfahren [17].
Das Delf-Verfahren [2] welches in dieser Arbeit untersucht wird basiert ebenfalls auf lokalen Deskriptoren. Die Deskriptoren werden aus einer faltenden Schicht aus dem hinteren Teil eines ResNet-50 [18] Modells extrahiert. Als Eingabe in das Netzwerk werden Bilder in ihrer Gesamtheit betrachtet. Da bis zur Extraktionsschicht keine vollvernetzten Schichten genutzt werden kann für jeden extrahierten Wert zurückgerechnet werden, von welchen Bereichen des Ursprungsbildes er beeinflusst wurde. Dies erlaubt es die Ausgaben der Extraktionsschicht in einzelne lokale Deskriptoren zu unterteilen. Um auszuwählen, welche der lokalen Deskriptoren zur Darstellung eines Bildes genutzt werden sollen, werden die lokalen Deskriptoren in ein weiteres neuronales Netz gegeben. Diesen Netz hat die Aufgabe zu bewerten, wie geeignet die einzelnen Deskriptoren zur Beschreibung des Gesamtbildes sind. Auf Grund dieser Bewertung werden die wichtigsten Deskriptoren zu jedem Bild erhalten, wogegen schlecht bewertete Deskriptoren verworfen werden. Der konzeptionelle Unterschied bei der Auswahl der Deskriptoren im Vergleich zum ONE-Verfahren ist also, das die Auswahl auf Grund der bereits berechneten Deskriptoren geschieht, anstatt auf Grund des Ursprungsbildes. Die Funktionsweise des Delf-Verfahrens wird in Kapitel **??** ab Seite **??** im Detail erklärt.
Bevor Neuronale Netze für die Erstellung von Deskriptoren genutzt werden können müssen ihre Para-

meter in einem Trainingsverfahren optimiert werden. Während dem Training muss das Netzwerk eine Aufgabe lösen. Wie erfolgreich das Netzwerk dabei ist wird mit Hilfe einer Fehlerfunktion dargestellt. Das Netz versucht seine Parameter so anzupassen, dass die Fehlerfunktion minimiert wird. Im Fall der bereits vorgestellten Verfahren wird dabei eine Dummy-Aufgabe, typischerweise die Klassifikation von Bildern, gestellt. In der letzten Zeit wurden jedoch einige Ansätze veröffentlicht, die versuchen neuronale Netze direkt an Image Retrieval Aufgaben zu trainieren. Radenović, Tolias und Chum stellen in [**?**] eine solchen Ansatz vor. Während dem Training arbeiten sie dabei mit Bildpaaren (i, j). Diese Paare bilden entweder ein korrektes Match, also enthalten den gleichen Bildinhalt oder ein inkorrektes Match. Beide Bilder durchlaufen ein identisches Netzwerk und werden durch die jeweiligen Netzwerkausgaben repräsentiert $(\mathbf{d}(i), \mathbf{d}(j))$. Für die Optimierung wird eine spezielle Fehlerfunktion L definiert. Falls es sich bei den Bildern um ein korrektes Match handelt sollten sich die Netzwerkrepräsentationen der Bilder möglichst ähneln.

$$L_{korr}(i, j) = \frac{1}{2} \parallel \mathbf{d}(i) - \mathbf{d}(j) \parallel^2 \tag{0.1}$$

Falls es sich jedoch um ein inkorrektes Match handelt, sollten die Repräsentationen einen Mindestabstand τ voneinander einhalten.

$$L_{inkorr}(i, j) = \frac{1}{2} \max \left[0, \tau - \parallel \mathbf{d}(i) - \mathbf{d}(j) \parallel^2\right] \tag{0.2}$$

Die Autoren testen ihr Verfahren auf unterschiedlichen CNN Architekturen, wie VGG [16] und AlexNet [19] und erzielen damit sehr gute Ergebnisse auf gängigen Retrievalbenchmarks. Beim direkten Training für Retrievalaufgaben handelt es sich um eine vielversprechende neue Forschungsrichtung im Retrievalbereich, die gerade intensiv untersucht wird.

Da Image Retrieval Systeme meist auf großen Bilddatenbanken eingesetzt werden und somit für eine Suchanfrage viele Bilder miteinander verglichen werden müssen, ist es sinnvoll Bildrepräsentationen so kompakt wie möglich zu gestalten, um die Laufzeit der Suche zu verbessern. Insbesondere bei Verfahren, die lokale Deskriptoren erstellen und häufig hunderte oder tausende Merkmale pro Bild extrahieren, kann mit einer guten Kodierung viel Rechenzeit gespart werden. Ein beliebter Ansatz zur Erstellung kompakter Darstellungen aus lokalen Deskriptoren ist das BOVW-Modell (Bag-of-Visual-Words) [20], erstmals vorgestellt im Kontext von Textklassifikation von McCallum und Nigam. Hierbei werden zunächst alle aus einem Datensatz extrahierte Deskriptoren mittels Clusteranalyse (bspw. K-Means-Clustering [21]) in Gruppen eingeteilt. Deskriptoren, die dem gleichen Cluster zugeordnet werden, werden dabei auf das selbe "visuelle Wort" abgebildet. Als Beschreibung des Gesamtbilds dient ein Histogramm über die im Bild enthaltenen visuellen Wörter. Bei diesem Verfahren geht durch Quantisierung ein Teil der Information verloren. Das ebenfalls auf Clustering basierte VLAD-Verfahren [22] von Jégou, Douze et al. versucht diese Information nutzbar zu machen, indem es statt der Vorkommen die Quantisierungsfehler akkumuliert, die beim abbilden auf die nächsten visuellen Worte entstehen.

Um eine Suchanfrage mit einer Rangliste der ähnlichsten Bilder zum Suchbild beantworten zu können, werden die Deskriptoren der Bilder in der Datenbank mit denen des Suchbildes verglichen. Als Metrik dient hierbei meist die euklidische Distanz. Auf kleinen Datensätzen ist es laufzeittechnisch sinnvoll alle Kombinationen von Such- und Datenbankbildern zu vergleichen. Häufig werden bei größeren Datensätzen jedoch Methoden der approximierten nächsten Nachbarsuche (ANN) verwendet. Diese garantieren zwar kein optimales Ergebnis, erlauben jedoch deutlich schnellere Verarbeitung von Suchanfragen. So

gibt es zum Beispiel Ansätze Deskriptoren mit Hilfe spezieller Hashfunktionen zu vergleichen. Diese werden so konstruiert, dass ähnliche Deskriptoren auf die gleichen bzw. möglichst ähnliche Hashcodes abgebildet werden, während gleichzeitig die Kollisionswahrscheinlichkeit für sehr unterschiedliche Deskriptoren minimal gehalten wird. Wang et al. beschreiben in ihrer Studie [23] unterschiedliche Konzepte für die Erstellung solcher Hashfunktionen.

Literaturverzeichnis

[1] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. A 4D Information System for the Exploration of Multitemporal Images and Maps using Photogrammetry, Web Technologies and VR/AR. *Virtual Archaeology Review*, 10:1, 07 2019.

[2] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. pages 3476–3485, 10 2017.

[3] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[5] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, Hartmut Neven, and Jay Yagnik. Tour the World: A Technical Demonstration of a Web-Scale Landmark Recognition Engine. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, page 961–962, New York, NY, USA, 2009. Association for Computing Machinery.

[6] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.

[7] David Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004.

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up Robust Features. volume 3951, pages 404–417, 07 2006.

[9] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more Distinctive Representation for Local Image Descriptors. volume 2, pages II–506, 05 2004.

[10] Cordelia Schmid and J. Ponce. Semi-Local Affine Parts for Object Recognition. *BMVC04*, 08 2004.

[11] Miaojing Shi, Yannis Avrithis, and Hervé Jégou. Early Burst Detection for Memory-Efficient Image Retrieval. 06 2015.

[12] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural Codes for Image Retrieval. volume 8689, 04 2014.

[13] Ali S. Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.

[15] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, page 3–10, New York, NY, USA, 2015. Association for Computing Machinery.

[16] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*, 09 2014.

[17] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective Search for Object Recognition. *Int. J. Comput. Vision*, 104(2):154–171, September 2013.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 7, 12 2015.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.

[20] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *AAAI 1998*, 1998.

[21] James B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. 1967.

[22] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating Local Descriptors into a Compact Image Representation. pages 3304 – 3311, 07 2010.

[23] Jingdong Wang, Heng Shen, Jingkuan Song, and Jianqiu Ji. Hashing for Similarity Search: A Survey. 08 2014.