Text + Image Retrieval: Überblick über die Herausforderungen und Lösungsansätze des Image Retrieval über textuelle Anfragen

Milad Khojasteh

Universität Koblenz-Landau, Campus Koblenz

Proseminar: Multimediadatenbanken und Retrieval

milad@uni-koblenz.de

Abstract. Diese Arbeit stellt die Herausforderungen des Image Retrievals über textuelle Anfragen vor und gibt einen Überblick über die Lösungsansätze. Um eine möglichst gutes und schnelles Retrieval zu ermöglichen, sollten zu den Bildern Metainformationen gespeichert werden. Das automatisieren der Generierung geeigneter Metainformationen stellt sich besonders durch die Semantische Lücke als schwere Aufgabe heraus. Ontologien bieten dabei eine strukturiertere Form von Metainformationen an, um auch semantische Eigenschaften des Bildes wiedergeben zu können.

1 Einführung

Bedingt durch den Fortschritt der Informations Technologie kam es in den letzten Jahren zu großen Ansammlungen von digitalen Bildern, welche kontinuierlich wachsen. Diese Bilder können jedoch nur dann effizient genutzt werden, wenn die Benutzer durch gezielte Anfragen die gewünschten Bilder erhalten können. Es gibt verschiedene Wege Anfragen zu formulieren, wobei die am weitesten verbreitete Methode die textuellen Anfragen sind. Beim Image Retrieval besteht hier ein Bruch der Darstellungsmodalität, welche mit Informationsverlusten in Verbindung steht. Die Bilder bestehen zunächst lediglich aus Pixeldaten, während die Anfragen jedoch textueller Natur sind.

Das manuelle Annotieren von textuellen Metainformationen ist sehr zeitaufwendig, daher ist das Ziel des Image-Retrievals den Prozess der Annotierung durch Metainformationen zu automatisieren, um bei textuellen Anfragen die gewünschte Ergebnismenge zu erhalten. Durch die „semantische Lücke“ stellt es sich jedoch als schwere Aufgabe heraus, automatisch über die Pixeldaten gute Metainformationen zu extrahieren.

Der ersten Teil dieser Arbeit befasst sich zunächst mit dem Problem der „semantischen Lücke“ und seiner Signifikanz. Im zweiten Teil werden dann

verschiedene Lösungsansätze erläutert, wobei darauf geachtet wird, inwiefern die einzelnen Lösungsansätze das Problem der semantischen Lücke überbrücken können.

2 Die semantische Lücke

Das Problem der automatisierten Bildanalyse besteht hauptsächlich darin, aus den Bildinhalten eine semantische Bedeutung zu extrahieren.

Die Lösung dieses Problems enttarnt sich als eine schwere Aufgabe, da es kaum möglich ist, nur aus den visuellen Daten (Pixeldaten) eine Interpretation der dargestellten Situation zu ziehen.
Um die Semantik eines Bildes jedoch ansatzweise erfassen zu können, muss es zunächst möglich sein, die dargestellten Objekte oder Szenen aus den Pixeldaten zu erkennen und mit natürlicher Sprache in textueller Form in Verbindung zu bringen.

Die entscheidende Bedeutung von semantischen Metadaten kommt u.a. dann zum Vorschein, wenn die Anfrage Sachverhalte adressiert, die etwas mit Identifikation einer Instanz, Interpretation oder Bedeutung zutun haben, da diese keine visuellen Eigenschaften sind.

Beispiel: Das Bild in Abbildung 1 kann durch folgende Anfrage nur durch die natürlich sprachliche Beschreibung in Tabelle 1 erhalten werden: „Ein Foto, mit einem Kühlschrank des Modells Electric 76A aus dem Jahre 1950“.

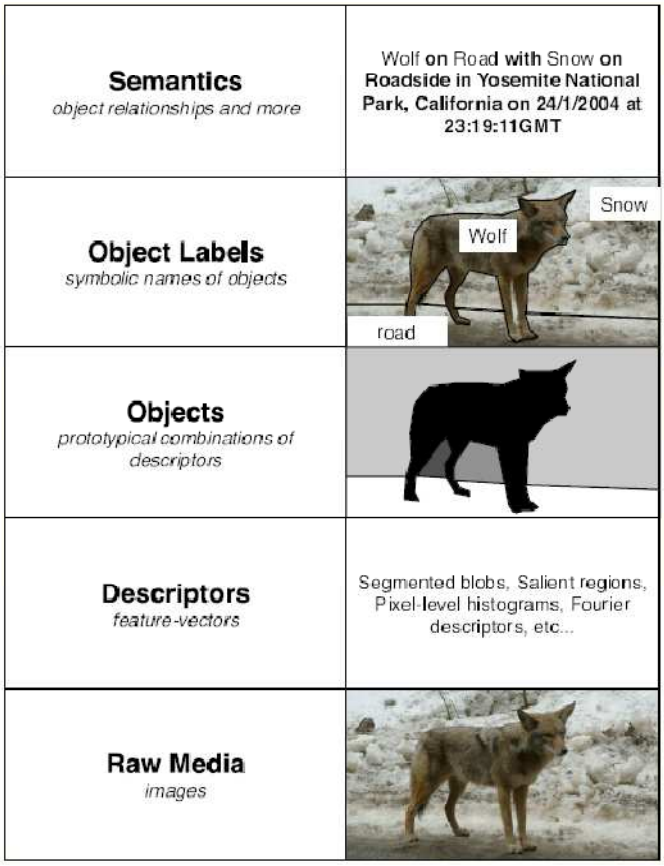
Title	Roomy Fridge
Date	circa 1952
Description	An English Electric 76A Refrigerator with an internal storage capacity of 7.6 cubic feet, a substantial increase on the standard model.
Subject	Domestic Life
Keywords	black & white, format landscape, Europe, Britain, England, appliance, kitchen appliance, food, drink, single, female, bending

[1] Tabelle 1: Metadaten des Bildes in Abbildung 1

[1] Abbildung 1: Geräumiger Kühlschrank ©Getty Images

In Bezug auf archivierte Bilder wird ein großer Anteil der Anfragen auf eindeutige Objekte definiert.
Das gleiche gilt für bewegte Bilder. 68% der Benutzeranfragen, welche auf Multimediainhalte zielen, enthalten mindestens ein spezielles Wort. Es ist jedoch schwierig durch automatische Berechnungen aus den reinen Pixeldaten eine hohe Ebene der semantischen Darstellung, zu erhalten.

Ebenen. Wie man in Abbildung 2 sehen kann, kann ein Bild auf verschiedenen Ebenen beschrieben werden.



[1] Abbildung 2: Ebenen, auf denen ein Bild beschrieben werden kann: Von den einfachen Pixeldaten zur vollen semantischen Darstellung

Auf der niedrigsten Ebene befinden sich die reinen Bilddaten (Pixeldaten).
Auf der nächsthöheren Ebene befinden sich einfache Farb-Histogramme, sogenannte descriptors im MPEG 7 Format, die spezifischen Merkmale wie z. B. die Farbe oder den Titel eines Bildes enthalten oder sogar „feature vectors“, welche sich auf das ganze Bild oder auch nur auf Bildsegmente beziehen können. Diese Beschreibungen werden oft durch „Content-Based Image Retrieval“ (CBIR) Techniken gefunden.

In der dritten Ebene sind Darstellungen von Objekten, die aus einer Kombination von „feature vectors“ oder aus anderen expliziteren Darstellungen erkannt werden.

Auf der Ebene vier befinden sich die identifizierten Objekte, welche durch abstrakte Namen oder idealerweise sogar durch Instanznamen annotiert werden. In Abbildung 2 könnte das Segment des Wolfes beispielsweise mit Wolf oder mit dem Namen des Wolfes annotiert werden.

Jedoch würde selbst das Benennen aller Objekte eines Bildes mit Instanznamen nicht die volle Semantik des Bildes repräsentieren. Dazu fehlen die Beziehungen

zwischen den Objekten, die Beziehung zur Umgebung, Aktionen bzw. Bewegungen der Objekte und weiterer Kontext.

Es gibt wenige Situationen, in denen vom reinen Bild durch automatische Berechnungen die Objekte extrahiert, benannt oder sogar andere semantischen Daten extrahiert werden können. Bei simplen Bildern ist es manchmal möglich Objekte zu extrahieren und diese durch Benennung zu spezifizieren, aber es gibt auch dort noch die Unterschiede zwischen der 4. und der 5. Ebene.

Selbst wenn die Objekte extrahiert und benannt werden, kann die Semantik des Bildes unklar sein:

Aus den Informationen, dass auf einem Bild ein Wolf, Schnee und eine Straße zu sehen sind, ist nicht zwingend ableitbar, dass es sich um einen Wolf aus einem kalifornischen Nationalpark handelt.

Daher kann man sagen, dass die Semantik eines Bildes auf einem höheren Level ist, als die simple Objektbenennung.

Im Hinblick auf die Lösungsansätze schlagen Hare und Lewis [1] daher vor, dass Problem der semantischen Lücke in zwei Teile aufzuteilen:

1. Lücke zwischen descriptor und Objektbenennung, also zwischen den Ebenen 2 und 3,4
2. Lücke zwischen benannten Objekten und der vollen Semantik, also zwischen der 4. Und der 5. Ebene

Viele interessante Arbeiten setzen sich ausschließlich mit Content Based Information Retrieval auseinander, was jedoch nur das erste Teilproblem lösen könnte. Laut Hare und Lewis [1] ist das zweite Teilproblem jedoch viel wichtiger, da Anfragen meistens nicht auf einzelne Objekte eines Bildes zielen.

2.1 Signifikanz der semantischen Lücke

Hare und Lewis [1] sind der Meinung, dass man sich nach den Anfragen echter Benutzer richten sollte.

Durch Analysen und Charakterisierungen der Benutzeranfragen ist herausgekommen, dass die häufigsten Anfragen semantischer Natur sind. Hiernach ist in Hinsicht auf echte Benutzeranfragen sowohl in Bezug auf stillen als auch auf bewegten Bildern das Auftreten von primitiven Eigenschaften sehr selten.

Die meisten Anfragen beinhalten demnach zeitliche, räumliche, emotionale, fachliche oder andere Aspekte, die alleine durch Content Based Information Retrieval (CBIR) Techniken keine guten Ergebnisse liefern, da diese Eigenschaften nicht visuell sind.

Wie in Kapitel 1 schon erwähnt und anhand des Kühlschrankbeispiels verdeutlicht wurde, beziehen sich Anfragen, die Sachverhalte adressieren, welche etwas mit Identifikation einer Instanz, Interpretation oder Bedeutung zutun haben, nicht zwingend auf visuelle Eigenschaften. Selbst wenn sich diese Anfragen auf visuelle Eigenschaften beziehen, sind diese in den seltensten Fällen identifizierbar. Diese nicht

visuellen Aspekte dienen normalerweise dazu, die Anfragen auf der höchsten-, nämlich der voll semantischen Ebene zu platzieren.

Beispielsweise benötigen Anfragen, die auf bestimmte Veranstaltungen zielen die volle semantische Ebene des Bildes, da jede Veranstaltung eine temporal, und spatial interpretative Beziehung zwischen Objekten ist. In diesen Situationen ist es unumgänglich die volle Semantik textuell an das Bild zu annotieren, wie das Beispiel mit dem Kühlschrank zeigt.

Eakins & Graham [2] schlagen eine alternative 3-level-Klassifikation von Anfragen vor. Diese basieren auf den primitiven Eigenschaften, abgeleiteten Eigenschaften und abstrakten Attributen, wobei Letzterer ein bedeutendes Ergebnis von high-level Argumenten über die Bedeutung und Absicht des Objektes der beschriebenen Szene einbezieht.

Ein weiteres, ebenfalls schwer zu lösendes Problem des Imageretrievals sind Anfragen, die explizit Objekte oder andere Aspekte ausschließen. Auf dieses Problem wird in dieser Arbeit jedoch nicht weiter eingegangen.

3 Lösungsansätze

Im Folgenden wird hauptsächlich auf die Indizierungsarbeit eingegangen. Diese dient dazu, um das Retrieval dieser Bilder zu unterstützen. Das Retrieval sollte möglichst schnell und effizient sein, wohingegen es bei der Indizierungsarbeit nicht so schlimm ist, wenn sie etwas länger dauert.

Bildannotation kann auch manuell geschehen, doch dies ist aufwendig und kostet viel Zeit. Außerdem ist die Qualität von manuell annotierten Metainformationen nicht zwangsläufig adäquat. Eine Lösung ist hier eine spielerische Annotationen wie z. B. das Spiel „ESP Game“ von „Games with a purpose“[1], wo zwei Personen über das Internet miteinander verbunden werden und gleich Bilder annotieren sollen, wobei nur das Wort genommen wird, welches als Erstes von beiden annotiert wurde. Dabei können Bestzeiten aufgestellt werden, was die Motivation steigern soll. Durch dieses spielerische Annotieren sollen auch die Qualität der Metainformationen gesteigert werden, da nur Metainformationen übernommen werden, die von beiden unabhängig voneinander eingegeben wurden.

Doch das manuelle Annotieren bleibt trotz dieser Versuche es spannender zu gestalten eine zeitintensive Methode. Daher wird der Versuch unternommen Bilder automatisch zu annotieren.

Ein anderer Ansatz für Image Retrieval, welches derzeit von den meisten Suchmaschinen verfolgt wird, besteht darin, die Texte in der „Umgebung“ eines Bildes mit dem Bild zu assoziieren. Mit der Umgebung können Texte gemeint sein, die sich auf der gleichen Seite in der Nähe des Bildes befinden, wie z. B.

[1] www.gwap.com

Überschriften oder aber auch der Dateiname. Dieser Ansatz umgeht das mühselige Annotieren, jedoch kann es auch zu Fehltreffern kommen, wenn Webseiten nicht dementsprechend aufgebaut sind. Dennoch ist es derzeit die Technik, die von großen Suchmaschinen wie Google für Image Retrieval eingesetzt werden, da die restlichen Ansätze noch nicht ausgereift genug sind.

In Abschnitt 2.2 wird ein Überblick über die Lösungsansätze gegeben, die Bilder automatisch zu annotieren.

Ein weiterer Ansatz sind Ontologien, die sich am ehesten an die Semantik des Bildes orientieren. Dieser Ansatz wird im Abschnitt 2.3 dieser Arbeit erläutert.

3.2 Automatische Bildannotationen

Im Gegensatz zur Annotation durch Menschen ist es bei automatischer Bildannotation schwer semantische Eigenschaften aus dem Bild zu extrahieren, um diese an das Bild zu annotieren.

Automatisch annotierte Bilder werden daher oft durch Content Based Image Retrieval (CBIR) Techniken gefunden, da CBIR nach Formen, Konturen oder Histogrammen sucht. Oft werden beim Content Based Image Retrieval die Anfrage in Form von Beispielbildern, Farbhistogrammen oder Sketches gestellt. Bei Anfragen in textueller Form müssen die durch CBIR Algorithmen erkannten Objekte extrahiert und mit textuellen Metainformationen in Verbindung gebracht werden. Hierbei entsteht ein Modalitätenbruch, den es zu überbrücken gilt. Die Lösung zu diesem Problem würde das erste Teilproblem der semantischen Lücke lösen, doch nicht das zweite Teilproblem, welches wichtiger in Bezug auf echte Benutzeranfragen ist, wie in Kapitel 1.2 erläutert wurde. Um auch das zweite Teilproblem zu lösen, müssten die extrahierten Metainformationen, die lediglich Objektnamen erhalten durch Relationen, Emotionen, Aktionen, etc. ergänzt werden.

Die heutigen Techniken kann man in zwei Kategorien aufteilen, nämlich in die, die ein Bild in Regionen aufteilen und die, die eine Szene orientierte Herangehensweise probieren, indem globale Informationen mit einbezogen werden, wobei der segmentierende Ansatz in den meisten Arbeiten verfolgt wurde.

Hierbei geht es hauptsächlich darum, einzelne Objekte durch Kombinationen von descriptoren, Farbhistogrammen oder Ähnlichen zu erkennen und die Namen der Objekte in textueller Form an das Bild zu annotieren.

Es ist jedoch anzumerken, dass das erfolgreiche Lösen dieses Problems lediglich die unwichtigere Lücke überbrücken würde, nämlich das in Kapitel 1 erwähnte erste Teilproblem.

Im Folgenden gebe ich einen Überblick über die verschiedenen Lösungsansätze zur automatischen Bildannotation.

Es gibt viele verschiedene Ansätze das Problem der automatischen Bildannotation zu lösen. Der häufigste Ansatz besteht darin, aus einer großen Menge von manuell annotierten „Trainingsbildern“ zu „lernen“ und neue Bilder auf der Basis des Gelernten automatisch zu annotieren.

Co-occurence Model. Zu diesen Ansätzen gehört z. B. das co-occurrence Model von Mouri [3], welches er bereits im Jahre 1999 vorgestellt hat. Er schlägt vor Bilder durch einen Segmentierungsalgorithmus, wie den „Normalized cuts algorithmus“ oder andere Algorithmen, welche z. B. auf Farbhistogrammen und anderen Eigenschaften von Bildausschnitten basieren, in sogenannte blobs zu segmentieren. Ein blob ist ein Vektor mit den extrahierten Eigenschaften, wie z. B. Farbe, Position, Größe, Textur etc. eines Bildausschnittes. Die Menge aller ähnlichen blobs wird als ein Visterm bezeichnet. Die Ähnlichkeit wird durch das sogenannte k-means Verfahren ermittelt. Dazu werden zunächst k zufällige Punkte im Vektorraum über die Merkmale gewählt, die den Mittelwert ihrer Cluster definieren. Daraufhin werden alle blobs per nearest neighbour Suche einem Cluster zugeordnet und der Mittelpunkt der Cluster wird neu bestimmt. Wenn beispielsweise aus vier verschiedenen Bildern Bildsegmente mit Tigern extrahiert wurden, ist die Menge der blobs dieser Bildausschnitte ein Visterm.

Es ist zu beachten, dass derzeit keine perfekte Segmentation existiert. Also ein Tiger kann in mehrere Teile aufgeteilt sein oder es wurde nur ein Teil des Tigers erkannt und der Rest wird zu einem anderen Objekt gezählt. Außerdem sind Segmentierungen bei ähnlichen Objekten fehleranfällig [4]. Ein Flugzeug und ein Vogel, die beide von unten aufgenommen wurden können in ihrer Farbe und ihrer Form gleich aussehen. Sie würden trotzdem zu einem Visterm zusammengefasst.

Anstelle von Segmentierungs-Algorithmen, die das Bild in ähnliche Bildausschnitte aufteilen, können Bilder auch in viereckige Bildausschnitte aufgeteilt werden. Ähnliche Vierecke werden dann auch zu einem sogenannten diskreten Visterm zusammengefasst.

Zunächst besteht kein Zusammenhang zwischen den Visterms und den Wörtern, die manuell an das Trainingsbild annotiert wurden. Die Aufgabe besteht darin, die Zusammengehörigkeit dieser Visterms zu den Wörtern zu finden.

Dabei beruht das co-occurence Model auf Wahrscheinlichkeiten. Dazu wird die co-occurence Tabelle aufgestellt, in der aufgeführt ist, wie oft die Kombination aus Visterm und einem bestimmten Wort, welches an dieses Bild annotiert wurde, vorkommt. Es ist also wichtig, dass die Testbilder hohe Vorkommen von einzelnen Objekten und den entsprechenden Schlüsselwörtern haben. Mithilfe der co-occurence Tabelle werden jeweils für die Wörter die zugehörigen Visterms ermittelt, indem die wahrscheinlichste Kombination, also die Kombination mit den häufigsten Vorkommen in Relation zu den anderen Wörtern zu diesem Visterm, errechnet wird. Dies kann auch in die andere Richtung für Visterms gemacht werden, um die entsprechenden Wörter zu finden.

maschinelles Übersetzungsmodell. Duygulu [5] hat im Jahre 2002 ein maschinelles Übersetzungsmodel vorgestellt, dass die Visterms in die entsprechenden Wörter übersetzt. Die Übersetzung geschieht mit einer maschinellen Übersetzung, welches die Zusammengehörigkeit von Visterm und Wörtern mit dem Expectation- and Maximization Algorithmus (EM) ableitet. Das Übersetzungsmodell wurde ursprünglich von Brown zur „Wort-zu-Wort-Übersetzung“ erfunden.

Der EM-Algorithmus berechnet dazu ebenfalls die Wahrscheinlichkeiten möglicher Übersetzungen, bis die wahrscheinlichste Korrespondenz erreicht ist. Dazu wird zunächst im E-Schritt die Näherung von der Wahrscheinlichkeit, dass ein Wort zu einem blob gehört bestimmt und im M-Schritt daraus die Wahrscheinlichkeit berechnet, dass das gesuchte Wort genau das Wort ist, welches unter der Bedingung, dass ein blob im selben Cluster liegt, abgeleitet wurde. Aus diesem Wert kann wiederum die Wahrscheinlichkeit, dass ein Wort zu einem blob gehört besser angenähert werden.

Um bessere Übersetzungen zu erhalten, wurde vorgeschlagen, die uneindeutigen Bildsegmente, mit dem Wert „null“ zu versehen, damit es nicht zu inadäquaten Retrieval Ergebnissen kommt.

Eine gute Übersetzung ist jedoch nur dann möglich, wenn genügend Daten vorhanden sind. Dazu hat Duygulu die Corel Database erschaffen, welche sehr große Mengen an annotierten Bildern besitzen. Die Corel Database wurden in der Literatur zu einer häufig verwandten Benchmark.

Doch das maschinelle Übersetzungsmodell beinhaltet genauso wenig Informationen über den Kontext wie das co-occurence Modell.

Cross Media Relevance Model. Joen [6] vervollständigte die Resultate von Duygulu, indem er die Übersetzung von Visterms zu Wörtern auf „cross lingual Information Retrieval“ zurückführte und das Cross Media Relevance Model (CMRM) zur Übersetzung benutzt. CMRM wurde ursprünglich für das Cross-Lingual Information Retrieval durch Lavrenco und Croft im Jahre 2001/2002 entwickelt.

In diesem Ansatz wird nicht versucht explizit Visterms zu Wörter zu übersetzen. Es wird versucht durch die Verteilung der Vorkommen von „blobs“ und Wörtern in gemeinsamen Bildern die zu annotierenden Wörter, zu finden. Also wenn durch die Trainingsdaten „gelernt“ wurde, dass bestimmte Wörter häufig zusammen in einem Bild vorkommen, sind die Wahrscheinlichkeiten hoch, dass diese auf dem zu annotierenden Bild ebenfalls gemeinsam abgebildet sind. Daher kann es hier auch passieren, dass ein Wort an ein Bild annotiert wird, obwohl das entsprechende „blob“ nicht gefunden wurde.

Wenn beispielsweise die Wörter Wiese und Wald bei den „Trainingsdaten“ oft mit Tiger auftauchen, ist die Wahrscheinlichkeit, dass das „blob“ einer Wiese mit dem Schlüsselwort Tiger in einem Bild sind, gegeben (auch wenn die Wahrscheinlichkeit nicht so hoch ist, wie dass das „blob“ Tiger mit dem Schlüsselwort Tiger in einem Bild sind). Deshalb wird bei einem Bild mit Tigern, wo keine Wiese erkannt wurde

dennoch nach der Wahrscheinlichkeit geprüft, dass das „blob“ Wiese im Bild abgebildet ist.

Das bedeutet, dass die Annotation durch CMRM auf das Vorkommen der Kombinationen von „blobs“ in den Trainingsdaten basiert. Dabei gibt es drei verschiedene Formen der Annotation, nämlich „fixed Annotation CMRM“ (FACMRM), „probabilistic Annotation CMRM“ (PACMRM) und „direct Retrival CMRM“ (DRCMRM).

Beim FACMRM wird die „harte Annotation benutzt“, also es wird eine bestimmte Anzahl von Wörtern an das Bild angehangen. Beispielsweise die fünf Wörter, für die die berechnete Wahrscheinlichkeit am höchsten ist.

Beim PACMRM wird dagegen eine bestimmte Wahrscheinlichkeitsschwelle gesetzt und nur Wörter werden an das Bild annotiert, deren Wahrscheinlichkeit höher ist als diese Schwelle.

Die meisten Herangehensweisen zur automatischen Bildannotation arbeiten mit „harten“ Annotationen.

Allerdings bemerkte Joen beim Vergleichen von PACMRM mit FACMRM, dass diese „harte“ Annotation zu fehlerhaften Annotationen führen können. So kann es beispielsweise passieren, dass ein Bild mit einem falschen Wort annotiert wird, weil dieses falsche Wort die nächst größte Wahrscheinlichkeit besitzt, dass es an das Bild passt. Duygulu versuchte das Problem durch die Zusammenfassung von ähnlichen Schlüsselwörtern zu umgehen, wodurch zusätzlich die Chancen beim Retrieval erhöht werden, so viele relevante Bilder wie möglich zu erhalten.

Eine andere Herangehensweise ist DRCMRM, welches aus den abgeleiteten Zusammenhängen zwischen „blobs“ und Wörtern die textuelle Query in eine Menge von „blobs“ übersetzt und damit suchen kann. Dies ist nur möglich, weil beim CMRM nicht die „blobs“ explizit zu Wörtern übersetzt werden.

Continuous-space Relevance Model und multiple Bernoulli Modell. Das „Continuous-space Relevance Model“ (CRM) von Lavrenko [7] und das multiple Bernoulli Modell (MBRM) von Feng, Manmatha und Lavrenko [8] sind ähnlich wie das maschinelle Übersetzungsmodell, doch sie berechnen die Wahrscheinlichkeiten mit stetigen dichten Wahrscheinlichkeitsverteilungen.

Da die Abschätzung der Annotationswahrscheinlichkeit beim CRM auf die relative Häufigkeit eines Wortes in der Annotation eines Bildes basiert, kann es bei Bildern mit verschiedenen Anzahlen von annotierten Wörtern zu schlechteren Retrieval Ergebnissen kommen. Daher wurde das Verfahren zum „Normalized CRM“ erweitert, welches eine Anzahl von „null“ Wörtern an ein Bild hängt, bis es gleichviele annotierte Wörter besitzt wie das zu vergleichende Trainingsbild.

Beim MBRM müssen im Gegensatz zum CRM nicht für jedes Wort die relativen Wahrscheinlichkeiten verglichen werden. Durch diesen Ansatz bekommt man viele gute Ergebnisse, jedoch ist das Retrieval relativ langsam.

Die Aufteilung des Bildes in viereckige Bildsegmente hat sich bei den Retrievalergebnissen als besser herausgestellt als die Segmentierung durch Algorithmen, die auf Eigenschaften wie z. B. Farbe beruhen. Dies liegt vermutlich daran, dass die Segmentierungsalgorithmen noch zu fehleranfällig sind.

Inference network Modell. Metzler und Manmatha [9] schlugen im „inference network Modell“ vor, aus dem Netzwerk Rückschlüsse zu ziehen, wie Regionen und deren Annotationen zu verbinden sind. Dies würde die „Trainingsdaten“ ab einer bestimmten Menge an annotierten Bildern im Netzwerk überflüssig machen. Es ist jedoch wichtig, dass die Annotationen im Netzwerk gut sind.

Semantic Space. Eine ganz andere Herangehensweise an das Problem der automatischen Bildannotation verfolgt der „Semantic Space“ von Hare und Lewis [1], welches Methoden aus der lineare Algebra benutzt, um Bilder mit Wörtern bzw. Ausdrücken zu assoziieren.

Die Herangehensweise basiert auf Latent Semantic Indexing (LSI). LSI ist ein Retrievalverfahren, das hauptsächlich bei der Suche nach textuellen Dokumenten eingesetzt wird und semantisch ähnliche Dokumente zu den der Anfrage zu finden. Das Schlüsselwort bzw. die Kombination aus Schlüsselwörtern müssen dafür nicht zwingend in den gefundenen Dokumenten vorkommen. Es wird also nach der Semantik, und nicht nur nach dem Vorkommen von Zeichenketten in den Dokumenten gesucht. Um nach der Semantik suchen zu können, wird der Text zunächst so weit reduziert, bis das Dokument nur noch Content Words besteht. Dazu werden Wörter wie z. B. der, die, das, ein, und, etc., die in jedem Text vorkommen können rausgestrichen. Die Beziehungen zwischen Dokumenten und Worte werden in einer Matrix dargestellt. In dieser Matrix kann man ablesen, wie häufig ein Wort in einem Dokument aus der Dokumentsammlung vorkommt. Mittels LSI wird diese Darstellung in einen semantischen Raum überführt, in dem Worte und Dokumente über latente semantische Konzepte dargestellt werden. Die Semantik ist im Gebiet der Sprachforschung als Sinn und Bedeutung von u.a. Wörtern und deren Beziehungen untereinander definiert. In der Ausgangsmatrix werden die semantischen Verbindungen zwischen Worten und Dokumenten explizit angegeben. LSI bildet daraus Konzepte, die die zugehörigen Worte bzw. Dokumente zusammenfassen. Somit führt LSI zu einer Dimensionsreduktion.

Der durch lineare algebraische Techniken kreierte semantische Raum von den Bildern und den Schlüsselwörtern wird dazu benutzt, um die Bilder je nach ihrem Vektor an bestimmte Orte im Raum zu positionieren. Dabei haben ähnliche Bilder auch ähnliche Positionen. Durch Anfragen werden Positionen des semantischen Raums angefragt, wodurch die Bilder zurückgeliefert werden, die an der Position oder „in der nähe“ sind.

3.3 Ontologien

Ontologien versuchen das wichtigere, zweite Teilproblem der semantischen Lücke zu überbrücken, indem sie auch die Relationen zwischen den Objekten modellieren. Außerdem sind sie interoperabel zwischen verschiedenen Systemen und dadurch vielseitig einsetzbar.

Derzeit werden Ontologien hauptsächlich für textuelle Dokumente eingesetzt. Da die Menge an Multimediadaten weiter zunimmt und dadurch die Bereitstellung von effektiven Retrievalmethoden wichtiger wird, besteht ein steigendes Interesse für den Einsatz von Ontologien für Multimediasammlungen.

Bei den annotierten Metadaten von Bildern handelt es sich oft um eine flache Struktur. Es existiert also keine zugrunde liegende Hierarchie oder Verknüpfungen und es ist nicht möglich alles Wissen, das man z. B. über ein bestimmtes Objekt auf einem Foto hat, in Form von Schlüsselwörtern zu hinterlegen. Eine Ontologie besteht jedoch aus Klassen, Attributen, Relationen zwischen den Klassen und Instanzen, womit es die zugrunde liegende Hierarchie und Verknüpfungen darstellen kann.

Will man beispielsweise ein Bild annotieren, auf dem ein Berggorilla zu sehen ist, würde man im Idealfall bei der automatischen Annotation das Wort Berggorilla an das Bild hängen. Doch ein Berggorilla ist u.a. auch noch ein Gorilla, Menschenaffe, Affe, Primat, Pflanzenfresser, ein Tier, das im Wald lebt, usw. Außerdem müsste jedes Foto mit einem Flachlandgorilla mit fast den gleichen Schlüsselwörtern annotiert werden.

Dieses Hintergrundwissen über bestimmte Objekte und die Beziehungen zu anderen Objekten können einfach in Form einer Klassenstruktur, ähnlich zu den UML-Klassendiagrammen gegeben sein. Das Bild mit dem Berggorilla müsste dazu lediglich mit der entsprechenden Klasse assoziiert werden, wobei idealerweise die Attribute für diese Instanz noch angepasst werden. Abgesehen davon kann man die Kontextinformationen, wie z. B. wann, wo und von wem das Bild aufgenommen wurde auch in Ontologieklassen speichern.

Hare und Lewis [1] haben in ihrem „Sculpteur project“ Metainformationen von Multimedia Objekte von Museen an eine Ontologie Basis, nämlich an das konzeptuelle Referenz Modell CIDOC übertragen um semantisches Retrieval zu unterstützen, welches zusätzlich noch mit „Content based“ Techniken kombiniert wurde.

4 Zusammenfassung und Ausblick

Es gibt gute Lösungsansätze durch automatische Bildannotationen die Lücke zwischen descriptor und Objektbenennung durch mithilfe von low-level Bildinformationen zu überbrücken. Diese Technik stößt jedoch bei der Extraktion von semantischen Eigenschaften eines Bildes an ihre Grenzen. Es sieht so aus, dass mithilfe von Ontologien die Lücke zwischen Objektbenennung und Semantik des Bildes schließbar sei. Dieser Ansatz steht jedoch in Hinsicht auf Multimediadaten noch in Kinderschuhen und muss noch weiter reifen, um die semantische Lücke wirklich schließen zu können.

Derzeit funktionieren Ontologien für Bilder jedoch nur semi-automatisch. Eine Kombination der „Content based“ Techniken zur Erkennung von Objekten und deren Benennung und Ontologien könnte jedoch eine Lösung sein, das Problem der semantischen Lücke vollständig zu lösen. Allerdings sind aktuelle Objektbenennungen zwar relativ gut, aber nicht immer korrekt bzw. vollständig und aktuelle Annotations-zu-Ontologie Übersetzer noch langen nicht ausgereift.

Der aktuelle Trend geht jedoch dahin, dass auch bei Multimediadaten versucht, wird semantische Eigenschaften zu extrahieren. Die Auswertung der Benutzeranfragen zeigt, dass dies ein wichtiger Aspekt für weitere Arbeiten des Image Retrievals ist.

Eine weitere Möglichkeit gute Image Retrieval Ergebnisse zu erzielen besteht in dem Ansatz der Relevance Feedback Methode. Dieser Ansatz umgeht die Probleme des CBIR hinsichtlich fehlender Beispielbilder, indem es textuelle Anfragen mit Anfragen in Form von Beispielbildern kombiniert. Dazu gibt der Benutzer die textuelle Anfrage ein, wählt aus den zurückgelieferten Bildern Beispielbilder aus, die für ihn relevant sind und kann noch eine Anfrage, diesmal in Form von Beispielbildern abschicken.

5 Literaturverzeichnis

1. Hare, J. S., Lewis, P. H., Enser, P. G. B. and Sandom, C. J. (2006) Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In: *Multimedia Content Analysis, Management and Retrieval 2006*, 17-19 January, San Jose, California, USA. pp. 607309-1.
2. J. Eakins and M. Graham, “Content-based image retrieval,” Tech. Rep. JTAP-039, JISC, 2000.
3. Y. Mori, H. Takahashi, and R. Oka, “Image-to-word transformation based on dividing and vector quantizing images with words,” in Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM’99), 1999.
4. Barnard, Kobus and Duygulu, Pinar and Guru, Raghavendra and Gabbur, Prasad and Forsyth, David (2003) *The effects of segmentation and feature choice in a translation model of object recognition.* In IEEE Conf. on Computer Vision and Pattern Recognition.
5. P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” in ECCV ’02: Proceedings of the 7th European Conference on Computer Vision-Part IV, pp. 97–112, Springer-Verlag, (London, UK), 2002.

6. J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in SIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 119–126, ACM Press, (New York, NY, USA), 2003.
7. V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in Advances in Neural Information Processing Systems 16, S. Thrun, L. Saul, and B. Sch¨olkopf, eds., MIT Press, Cambridge, MA, 2004.
8. S. L. Feng, R. Manmatha, and V. Lavrenko, “Multiple bernoulli relevance models for image and video annotation.,” in CVPR (2), pp. 1002–1009, 2004.
9. D. Metzler and R. Manmatha, “An inference network approach to image retrieval.,” in Enser et al.,45 pp. 42–50.