Suche nach Fotos und Videos mithilfe von Ontologien

Rainer Marquardt

Software- und Internettechnologie
6. Semester
Universität Mannheim

Abstrakt

In dieser Ausarbeitung werde ich zwei aktuelle Arbeiten [1,2] zum Thema Erstellen von Ontologien für die Suche nach Fotos und Videos vorstellen. Im Mittelpunkt wird die Arbeit [1] zur Annotation von Fotos stehen. Es werden aber auch die Gemeinsamkeiten aufgezeigt die Ontologien für Foto- und Videosuche haben und noch allgemeine Überlegungen bezüglich einer Suche nach Multimediadaten angestellt.
Darüberhinaus werde ich einen Vergleich der in [1] 2001 gemacht wurde selbst noch einmal nachvollziehen.

Einleitung

Die Menge an Informationen die täglich produziert wird, steigt stetig. So wurden 2004 „mehr Informationen als in der bisherigen Menscheitsgeschichte produziert“ [3] und das Erdbeobachtungssystem der NASA kann täglich ein Terrabyte an Bildern produzieren [4]. Aber nicht nur für die NASA oder große Medienanstalen und Verlage wird die Flut an Informationen immer größer. Sondern, durch die stetig wachsende Akzeptanz, Ausbreitung und Erreichbarkeit des Internets, auch für Endbenutzer. So produzierte im Jahre 2003 jeder Mensch durchschnittlich 800MB an Daten [3]. Durch technische Neuerungen, insbesondere Mobiltelefone mit integrierten Kameras und Anschluss ans Internet liegt es auf der Hand, dass Multimediadaten wie Bilder und Videos daran einen immer größeren Anteil haben.

Diese Tatsachen legen es Nahe, sich Gedanken über die Strukturierung, Speicherung, den Austausch von und besonders das Suchen nach den Daten zu machen.
Die Suche nach textuellen Inhalten ist bereits umfangreich erforscht und im Internet hat sich auch bereits ein eindeutiger Marktführer herauskristallisiert: Google ist in den letzten Jahren quasi der Inbegriff der Suche im Netz geworden. Auch Forschungen über die Suche nach textuellen Ressourcen mittels Ontologien wurden bereits gemacht [11].

Die Suche nach Multimediadaten wie Fotos und Videos und der Aufbau geeigneter Strukturieren hierfür sind allerdings aktuelle Forschungsthemen, zu denen es einige Ansätze gibt.
Geht man nur vom Bild selbst aus, „sieht“ eine Maschine nichts außer eine Matrix von verschiedenfarbigen Pixeln. Um nun, ohne weitere Informationen, direkt nach einem Bild zu suchen, könnte man ein anderes (vereinfachtes) Bild als „Anfrage“ zur Suche verwenden und dann die Ähnlichkeit der Pixel in der Matrix prüfen [5].
Um mit Worten nach Bildern zu suchen, müssen Strukturen bestehen, die Worte geeignet mit den grafischen Daten verknüpfen. Zusätzliche Informationen die Aussagen über den Inhalt oder andere Eigenschaften eines Bildes treffen sind bekannt als Metadaten.
Nun stellt sich die Frage, welche Metadaten sinnvollerweise zu einem Bild annotiert werden sollen und wer dies macht.
Automatische Crawler, die für Suchmaschinen wie Google und Alta Vista das Internet nach Informationen durchsuchen, treffen i.d.R. die Annahme, dass unter anderem die textuellen Inhalte einer Website geeignete Daten sind, die mit den Bildern dieser Website verknüpft werden können [6]. Eine weitere Möglichkeit wäre es, dass Menschen, evtl. unterstützt durch automatische Verfahren, Kollektionen von Bildern und Videos annotieren.
Qualitative und quantitative Vergleiche von Suchergebnisse in automatisch und manuell erzeugten Annotation aus [1] werden später zitiert und selbst durchgeführt.

Ontologien

Zur Zeit werden Multimediainhalte häufig mit Keywords verknüpft, nach denen dann gesucht werden kann. Keywords haben aber einige Nachteile: sie haben meist eine flache Struktur, 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 Keywords zu hinterlegen.
Ein einfaches Beispiel soll dies demonstrieren: es soll mittels Keywords ein Bild annotiert werden, auf dem ein Gorilla und zwar ganz genau ein Berggorilla zu sehen ist. Die Frage ist: welche Keywords sollen nun gespeichert werden? „Berggorilla“, „Gorilla“, „Menschenaffe“, „Affe“, „Primat“,...? Dies ist der vereinfachte Stammbaum von Berggorillas, aber es sind alles Begriffe über die man das Bild des Berggorillas bei einer Suche finden können sollte, denn all das ist ein Berggorilla ja gleichzeitig. Auch nach Eigenschaften wie „pflanzenfressende Tiere“ oder „Tiere die im Wald leben“ könnte man suchen wollen. Soll das Bild unseres Berggorillas bei all diesen Anfragen gefunden werden, müsste sein gesamter Stammbaum und alle seine Eigenschaften als Keywords hinterlegt werden. Und mehr noch: gibt es nun ein weiteres Bild mit einem Flachland-Gorilla, müssten für diesen nun fast die selben Keywords erneut angelegt werden. Hieran wird die Problematik dieses Verfahrens deutlich.

Hintergrundwissen in Form von Ontologien kann hier eine geeignete Alternative sein. Ontologien klassifizieren Dinge, weisen ihnen Attribute zu und setzen sie in Bezug zu einander. Hätten wir also eine geeignete Ontologie von Tieren und ein Bild eines Berggorillas, müsste eine Annotation dieses Bildes lediglich mit der richtigen Ebene in der Hierarchie der Ontologie verknüpft werden. Nun stünde das ganze Hintergrundwissen bereit um das Bild unseres Gorillas über verschiedenste Methoden und Begriffe zu finden.

Glücklicherweise stehen derartige Ontologien bereits für speziellere Domänen wie Kunst (ICONCLASS), Medizin (Gene Ontology) als auch domänenübergreifend (Cyc und WordNet[7]) zur Verfügung.

Ziel und Anforderungen

Ziel ist es, Annotationen für Bilder und Videos zu erzeugen die visuelle Informationen über das Bild enthalten. Insbesondere die Inhalte sollen mit generellen Hintergrundinformation, gegeben durch existierende Ontologien, geeignet verknüpft werden, um so die Mächtigkeit der Ontologien zur Suche nutzen zu können. Dabei sollen die erzeugten Annotation und Verknüpfungen möglichst kompatibel und standardkonform sein.

Aufbau allgemein

WordNet ist eine umfassende, frei nutzbare und standardkonforme Wissensplattform und eignet sich daher sehr gut als Ressource für Hintergrundwissen auf Basis einer Ontologie. Zudem basiert WordNet auf dem RDF-Schema (RDFS) [8] und um das Ziel der Kompatibilität und Standardkonformität zu erreichen wird für den Aufbau der Annotationen in [1,2] auch RDFS verwendet.
Für die Annotation von Videos fehlen aber nach [2] in WordNet einige Eigenschaften um visuelle Inhalte zu beschreiben. Gründe hierfür sind, dass WordNet nach psycholinguistischen Erkenntnissen entworfen wurde [9] und z.B. bei der in [2] betrachteten Domäne „Beförderungsmittel“ eher die funktionalen Eigenschaften im Vordergrund stehen und weniger die visuellen. Um diesem Mangel entgegenzuwirken wird WordNet um einige Eigenschaften des MPEG-7 Standards [10] ergänzt.

Aufbau für eine Annotation von Fotos

Zunächst aber soll die Annotation von Fotos [1] im Vordergrund stehen. Betrachtet man sich ein (digitales) Bild, stellt sich die Frage, was eigentlich alles über dieses ausgesagt werden kann:

der/die inhaltlichen Betrachtungsgegenstand /- stände (*subject matter*). Z.B.: ein Gorilla in einem Wald, der eine Banane ist.
Die Umstände unter denen das Foto gemacht wurde *(photo feature)*. Z.B.: Zeit, Ort, Perspektive
Eigenschaften des (digitalen) Fotomediums *(medium feature)*. Z.B.: Format, Auflösung, Farben

Die Annotation kann nun wie ein aus UML [12] bekanntes Klassendiagramm aufgebaut werden (siehe Abb.1). Dabei setzt sich die Annotation mindestens aus *subject matter* und optional aus *medium feature* und *photo feature* zusammen. Das Hauptaugenmerk wird in [1] nun auf die inhaltlichen Aspekte (*subject matter*) gelegt.

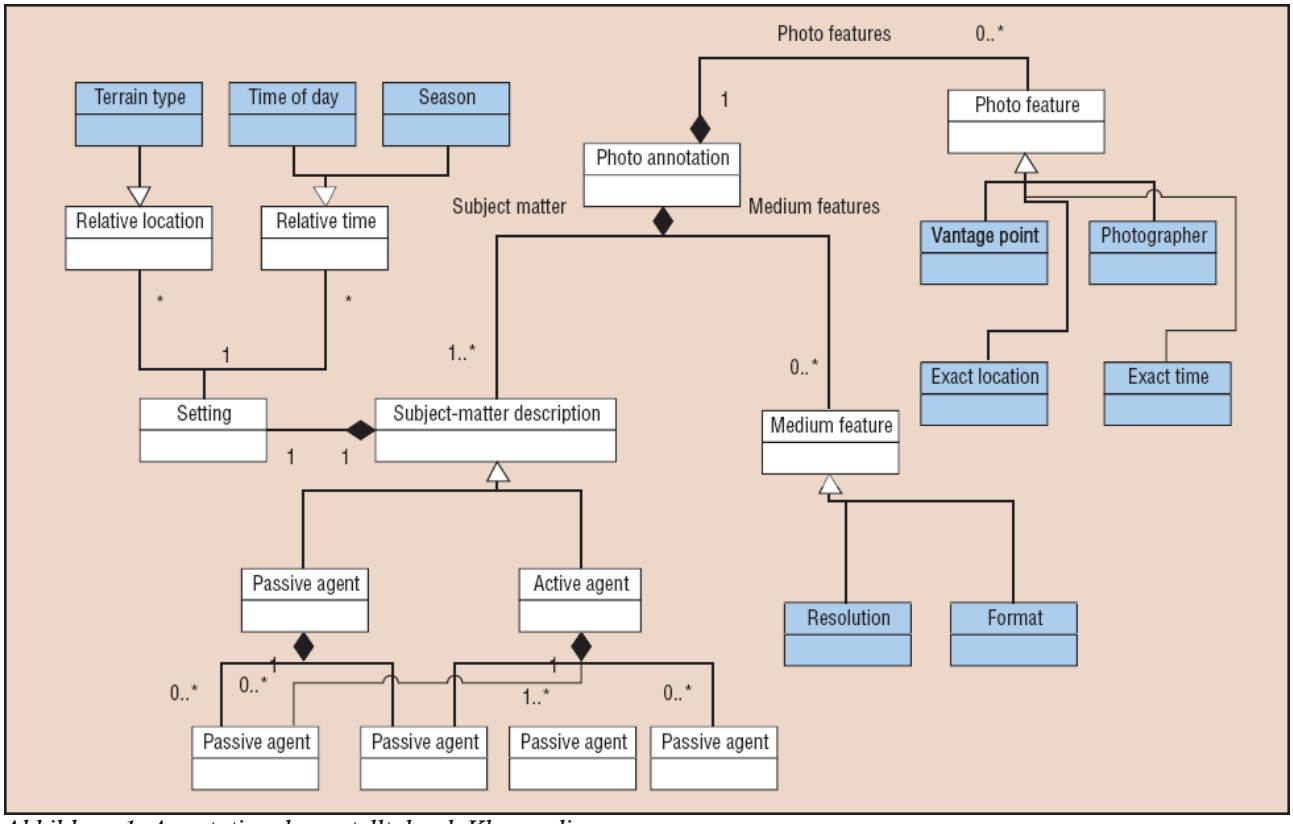

Abbildung 1: Annotation dargestellt durch Klassendiagramm

Subject matter lässt sich nach A. Tam und C. Leung weiter strukturieren [13] in

agent samt *agent-modifiers*. Z.B.: *agent* „Affe“ mit *modifier* „Farbe: orange“
action. Z.B: „essen“
object. Z.B.: „Banane“
setting. Z.B.: „Wald in der Dämmerung“

In [1] werden anhand dieses Schemas von *subject matter* noch zwei Subklassen gebildet:

ein *passive agent* ist ein Schema mit einem *single agent*, optionalen *modifiers* in einem *setting*
ein *active agent* ist ein Schema mit einem *single agent*, *modifiers*, der optional mit einem *object* eine *action* tätigt in einem *setting*

Die 4 Eigenschaften von *subject matter* der Annotation können nun mit Werten belegt werden. RDFS gibt hierfür einen *range,* also einen Wertebereich, vor aus dem Werte gewählt werden können. Darüberhinaus wird RDFS an dieser Stelle auch um Standard- bzw. Vorgabewerte erweitert. Auch über diese kann die Ontologie durchsucht werden.

In dem betrachteten Beispiel „Affen“ stellt WordNet den biologischen, hierarchisch gegliederten Stammbaum von Tieren zur Verfügung. Die Eigenschaft *agent* kann mit Werten aus dem Wertebereich *species* belegt werden. Die *modifiers* können mit Werten aus *animal_charecteristics* belegt werden.
RDFS sieht vor, dass für den *range* nur ein Wert gewählt werden kann – mehrere Werte sollen über eine Superklasse angelegt werden. Aufgrund von disjunkten Wertebereichen wird dies in [1] als unpraktikabel angesehen und es werden mehrere Werte zur Auswahl erlaubt. Aber auch wenn es in einige Fällen umständlich sein mag, so stellt sich meiner Meinung nach die Frage, ob es sinnvoll ist z.B. die Auswahl mehrerer Werte für eine Eigenschaft wie *agent* zu erlauben. In unserem Fall könnte also ein *single agent* gleichzeitig mit verschiedenen *species* belegt werden, was eher nicht sinnvoll wäre.

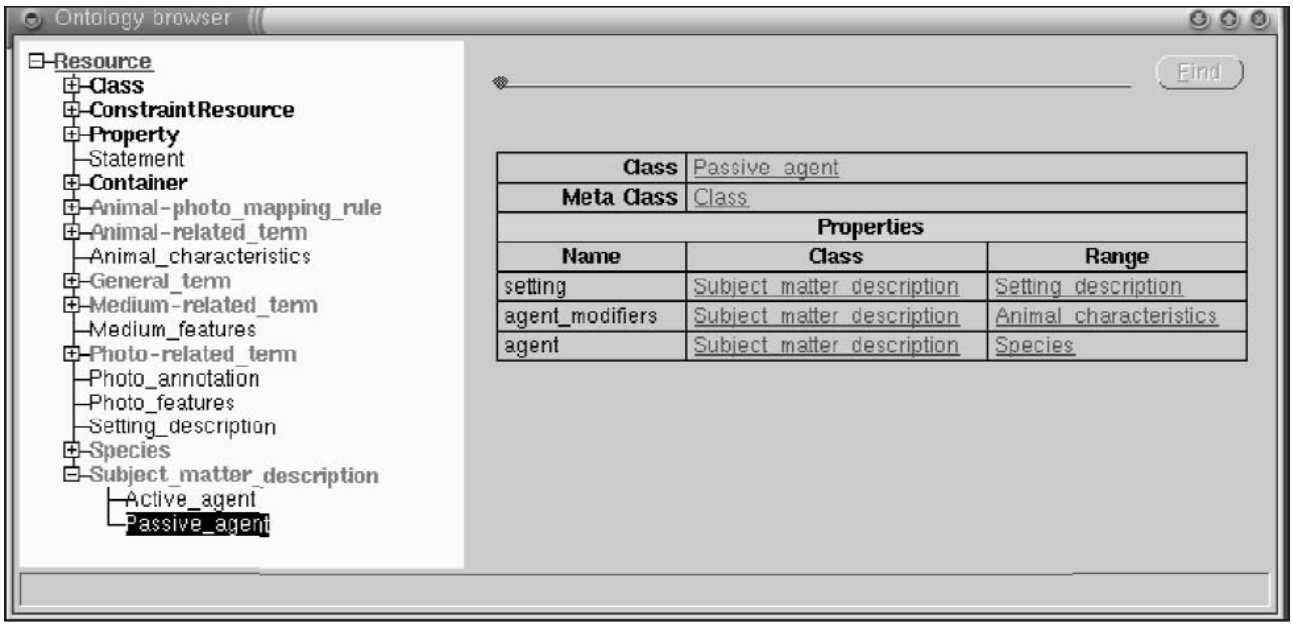

Abbildung 2: Klasse "passive_agent" mit Eigenschaften aus einem WordNet Wertebereich

Einschränkend muss erwähnt werden, dass nicht unbedingt alles, was jemand über ein Foto aussagen möchte in der Ontologie Platz finden kann. Für diese zusätzlichen Daten braucht es geeignete syntaktische Vorlagen wie einfache Textelemente, die aber auch durchsuchbar sein müssen. Weiterhin können nicht alle *subject matter* Eigenschaften mit Ontologien von Domänen verknüpft werden. Beispielsweise das *object,* das unser *agent* in der Hand hält, könnte eine Banane oder irgendetwas vollkommen anderes sein. Aufgrund der zahllosen Möglichkeiten ist es nach [1] unmöglich hierfür eine einschränkende Auswahl aus einer Ontologie bereitzustellen.

Experiment

Annotation

In [1] wurde ein Tool kreiert, mit dem Daten zu den drei Klassen *subject matter*, *photo feature* und *medium feature*, aus denen sich die Annotation zusammensetzt, angelegt werden können. Bei Wertebereichen die mit Ontologien verknüpften sind, können die Werte per Tastatur eingetippt werden und sobald es sich um einen gültigen Wert handelt wird dieser fett hervorgehoben. Alternativ können die Werte aus einer hierarchischen Auswahl gewählt werden.
Einfache Textdaten können direkt hinterlegt werden.
Ein Foto kann mit mehreren Annotationen verknüpft werden.

Suche

Das Interface zur Suche ähnelt dem zur Annotation. Gesucht werden können Annotationen die den Werten der Anfragemaske entsprechen oder spezieller sind. Wird also eine *agent* „Affe“ gesucht, der mit „seiner Hand etwas am Kopf macht“ *(modifier)* so findet das Tool alle Affenspezies, also Schimpansen, Gorillas, Orang-Utans, usw. die ihre Hand irgendwie an ihrem Kopf haben.
Nun könnte man den *agent* wie auch die *modifier* spezialisieren: ein „Schimpanse“ (statt „Affe“) der sich am Kopf *„kratzt“* (statt „irgendwas“ mit der Hand am Kopf zu machen). In umgekehrter Richtung entspräche dies einer Generalisierung. Auch können Eigenschaften als Wildcard leergelassen werden – z.B. das *object*: finde einen Affen, der „irgendetwas“ isst.

Abbildung 3: suche nach "einem affen, der irgendetwas mit der hand am kopf macht"

Aufbau für eine Annotation von Videos

wir haben nun gesehen, wie Annotation von Fotos basierend auf Ontologien erstellt werden können. Nun stellt sich die Frage, wie wir dieses Prinzip auf Videos erweitern können. Zunächst ist also zu überlegen, was ein Foto von einem Video unterscheidet: eigentlich ist ein Video ja nur eine Folge von einzelnen Bildern. Da es sich hierbei i.d.R um mehrere Bilder pro Sekunde handelt, wäre es etwas zu extrem aufgefasst wirklich jedes

einzelne Bild eines Videos zu betrachten. Denn eine (kurze) Sequenz aufeinander folgender Bilder zeigt ja meist die selben Inhalte (siehe Abb. 3). Unsere Grundlage zur Annotation bilden also *shots*: „shot ist defined as a continous camera recording,...“ [14]. Zusätzlich kann es bei Videos auch Ton geben. Die Eigenschaften von Audio bei der Annotation von Videos mit einfließen zu lassen, wäre sicher auch ein interessanter Gesichtspunkt, aber uns geht es hier hauptsächlich um den Aufbau einer visuellen Ontologie (VO) zur Annotation.
Um ein Video zu annotieren, müssen wir also eine Folge von *shots* annotieren. Alle *shots* können mitunter aber recht viele sein – Abhilfe schaffen hier Verfahren der Zusammenfassung von Videos z.B. nach [14], die automatisch die wichtigsten *shots* eines Videos selektieren.

Abbildung 4: Veranschaulichung von Shots

Wie bereits zuvor erwähnt, reichen in [2] die von WordNet bereitgestellten visuellen high-level Eigenschaften *visibility, naturalness, envoirement* und *material* nicht aus und werden daher um die low-level Eigenschaften *color, shape* und *motion* von MPEG-7 erweitert. Von der betrachteten Domäne *conveyance* (Beförderungsmittel) gibt es 564 Klassen in WordNet. Jede dieser Klassen wird mit den 7 Eigenschaften bestückt und mit Werten gefüllt deren Wertebereich von WordNet bestimmt wird (siehe Abb. 4). Eine kleine Erweiterung von Hoogs en Stein [18] wird hierbei auch übernommen: der ursprüngliche Wertebereich für *visibility* umfasst die Werte *invisible* und *visible*. *Visible* lässt sich aber noch weiter verfeinern durch *viewable* – für Dinge die mit blossem Auge sichtbar sind – und *visualisable* für Dinge die mit technischer Hilfe sichtbar gemacht werden können. Bei der Auswahl von Werten steht wieder der Vorteil von Ontologien zur Verfügung: generelle Werte können weiter spezialisiert werden, spezielle Werte auf allgemeinere zurückgeführt werden.

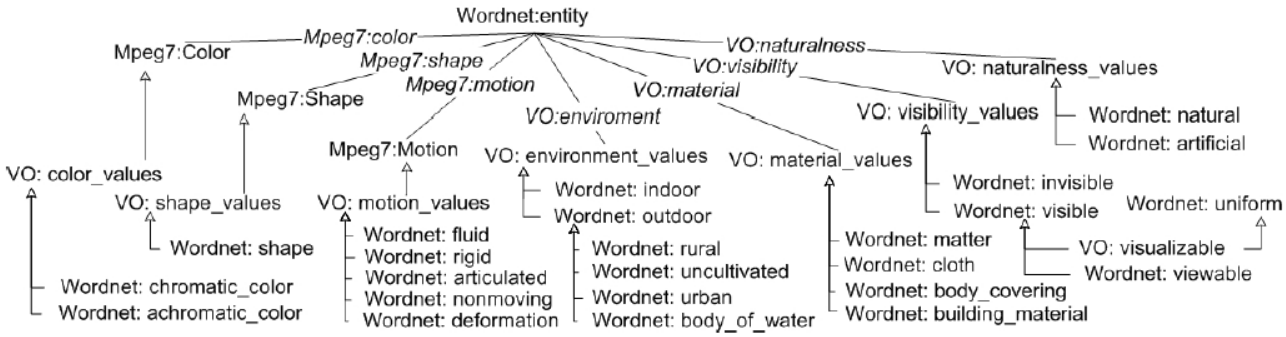

Abbildung 5: RDFS Graph der Video Eigenschaften aus WordNet und MPEG-7 und Werten aus WordNet

Experiment

In [2] wurde mit einer Kollektion von 40 Shots eines TRECVID 2003 [15] Nachrichtenvideos gearbeitet. Die Zerlegung eines ganzen Videos in Shots ist also bereits vorausgesetzt. Auf allen Shots waren Boote, Flugzeuge, Schiffe oder Züge zu sehen. Von jeder Sorte gab es 10 Shots.

Annotation

Das Annotieren der einzelnen Shots geschieht nach [16]: aus den Shots werden Keyframes bestimmt und für eine Region dieses Frames werden alle Eigenschaften außer *envoirement* bestimmt. Aufgrund stetig steigender Qualität von Detektoren ist annehmbar, dass es in naher Zukunft sehr gute Detektoren für die 6 sichtbaren Eigenschaften geben wird. Es wurde daher davon ausgegangen, dass es bereits perfekte Detektoren gibt und die Eigenschaften wurden manuell bestimmt – getestet wurde also nur die Qualität der Annotation.
Anhand dieser Werte wird eine Liste mit möglichen Konzepten aus *conveyance* bereitgestellt aus der ein passendes Konzept für die Annotation gewählt werden kann. Die benachbarten Regionen legen nach dem selben Verfahren die Werte für *envoirement* fest.

Suche

Die Suche nach Videos bzw. Shots verläuft genauso: die Ontologie wird anhand einer Anfrage nach den 7 Eigenschaften durchsucht und liefert eine Liste möglicher Shots, aus denen dann ein passender Shot gewählt werden kann.

Auswertungen

In [2] wird davon ausgegangen, dass die Annotation halbautomatisch stattfindet und sowohl eine Zerlegung eines Videos in Shots, als auch die Bestimmung der Eigenschaften automatisch stattfindet. Ab hier verläuft die Suche und Annotation fast gleich: anhand der ermittelten oder abzufragenden Eigenschaften werden mögliche Konzepte aus der Ontologie für die Annotation bzw. Shots mit passenden Annotationen für die Suche bereitgestellt. Aus der jeweiligen Liste kann dann manuell gewählt werden.
In [1] wird von einer manuellen Annotierung ausgegangen und es wird sowohl getestet wie anwenderfreundlich diese ist, als auch, wie gut die Suchergebnisse der Suche nach Bildern ist.

Die Kriterien nach denen in [2] die Annotierung bewertet werden, sind Verlässlichkeit und Präzision. Verlässlichkeit ist der Prozentsatz, dass mindestens ein zur Auswahl gestelltes Konzept relevant ist. Präzision ist die Menge der relevanten Konzepte im Verhältnis zu allen Konzepten die zur Auswahl bereitgestellt werden. Im Versuch lag die Verlässlichkeit im Schnitt bei 93%, wohingegen die Präzision im Schnitt nur bei 8% lag. Immerhin wurden die 564 Konzepte von *conveyance* auf durchschnittlich 57 reduziert, aus denen in vertretbarem Aufwand manuell ein Konzept bestimmt werden kann.

In [1] sollten zum Test der Anwenderfreundlichkeit 6 Studenten einige Annotationen anlegen und auch nach Bildern suchen. Beim Annotieren gab es mit *subject matter* einige Probleme, da z.B. nicht klar war, wie ein Bild mit mehreren Affen behandelt werden soll – zu Erinnerung, es gibt diese 4 Eigenschaften: *(single) agent + modifier, action, object* und *setting*. Evtl. müsste das Template um so etwas wie *interacting agents* erweitert werden.

Verwirrend war auch der Gebrauch von Vorgabewerten. Einige waren überflüssig oder die Nutzer widersprachen ihnen und änderten sie. Andere Dinge wie „lips pressed together“ konnten nicht ausgedrückt werden, da den Leuten nicht klar war, dass sie pro Bild mehrere Annotation, z.B. eine mit *passive agent* und eine mit *active agent* anlegen konnten.
Allgemein gab es beim Annotieren und Suchen einige Schwierigkeiten, die darauf zurückzuführen sind, dass einige Kenntnisse der Domäne notwendig gewesen wären.

Das Forscherteam selbst hat die Suche selbst auch getestet und diese mit anderen Suchmaschinen verglichen. An dieser Stelle prüfe ich die Ergebnisse der Suchmaschinen heute (05/2006) noch einmal nach.

Die Ergebnisse werden einerseits wieder anhand der Präzision bewertet, die auch hier als der Prozentsatz der relevanten Bilder aus der Menge aller gefundenen Bilder definiert ist. Zum Anderen wird als *recall* betrachtet, wie viele relevante Bilder, aus der Menge aller vorhandenen relevanten Bilder, gefunden werden.
Zunächst wurde nach großen Affen gesucht. Das Tool das die entwickelte Ontologie durchsuchte lieferte 100% Präzision und 100% Recall. Eine Anfrage an Alta Vista nach „great + apes“ fand 13 Bilder, davon waren 6 von richtigen Affen. Es ist nun schwer den Recall zu bestimmten, denn es ist weder bekannt, wie viele relevante Bilder sich im gesamten Web noch wie viele relevante Bilder sich davon im Index der Suchmaschine befinden. Man kann aber wohl guten Gewissens davon ausgehen, dass eine sehr große Menge an relevanten Bildern vorhanden ist. Daher ist der Recall für diese Anfrage eher schwach und die Präzision ist mit 50% auch nicht sehr gut. Die Suche nach „great apes“ lieferte 45.000 Bilder. Auch hier, wie gehabt, ist der Recall schwer zu schätzen. Bei Betrachtung der ersten 200 Bilder handelte es sich bei 64 davon um große Affen – also eine Präzision von 32%.
Ich habe diese Anfragen 2006 nun bei Google und Alta Vista wiederholt. Wie bereits erwähnt, lässt sich der Recall schwer bestimmen und ich lasse ihn außen vor. Für die Präzision habe ich jeweils die ersten 200 Bilder betrachtet, da in allen Fällen mehr als 200 gefunden wurden. Auf beiden Suchmaschinen habe ich die Suche nach der genauen Wortgruppe „great apes“ durchgeführt, was bedeutet: „great AND apes“, und auch in dieser Reihenfolge. Berücksichtigt habe ich alle Bilder auf denen fotografierte, lebende Affen zu sehen waren.
Google fand ca. 1900 Bilder mit einer Präzision von ca. 32%. Alta Vista fand 648 Bilder mit einer Präzision von ca. 60%. Googles Präzision war deswegen niedriger, weil zahlreiche Bilder mit Diagrammen, Zeichnungen, Landkarten oder auch Mitarbeitern des „Great Ape Projects“ gefunden wurden. Bei Alta Vista kamen viele Fotos von ein und der selben Website - es waren aber alles brauchbare Fotos. Bei beiden Suchmaschinen haben sich einige Bilder wiederholt, z.B. aufgrund von verkleinerten Vorschaubildern – diese habe ich dennoch mitgezählt.

	Ontologie aus [1]	Alta Vista *<great + ape>* 2001	Alta Vista *<great ape>* 2001	Google *<„great ape“>* 2006	Alta Vista *<„great ape“>* 2006
Recall / Treffer	100%	13	45000	1900	648
Präzision	100%	50%	32%	32%	60%

Tabelle 1: Suche über die Ontologie im Vergleich zu bekannten Suchmaschinen mit automatisch generiertem Suchindex

Hierbei muss allerdings im Auge behalten werden, dass es sich in [1] um manuell erzeugte Annotationen handelt während die betrachteten Suchmaschinen mit automatisch generierten Suchindizes arbeiten.
Daher wurde nun mit der Suchmaschine gettyone.com verglichen. Diese durchsucht Kollektionen, die von Hand, auf Basis von Keywords annotiert werden. „great ape“ war bei [1] als Keyword unbekannt und lieferte daher gar keine Treffer. „ape“ lieferte 521 Treffer, davon waren 10 Bilder nicht von lebenden (fotografierten) Affen. 64 der verbleibenden Bilder waren von anderen Primaten.
Meine Suche 2006 fand inzwischen 2 Bilder für „great ape“ auf denen auch lebende Affen zu sehen waren. Die Suche nach „ape“ fand nun 1754 Bilder. Von den ersten 240, entsprachen 14 nicht den o.g. Kriterien.

	Ontologie aus [1]	gettyone.com *<great ape>* 2001	gettyone.com *<ape>* 2001	gettyone.com *<great ape>* 2006	gettyone.com *<ape>* 2006
Recall / Treffer	100%	0	521	2	1754
Präzision	100%	-	85%	100%	94%

Tabelle 2: Suche über die Ontologie im Vergleich zu manuell annotierten Kollektionen auf Basis von Keywords

Dies sind bisher eigentlich gute Ergebnisse. Probleme gab es in [1] nun aber bei spezielleren Anfragen, z.B. Einschränkungen der *agent-modifiers.* Nun wurde nach einem Schimpansen gesucht, der sich mit der Hand am Kopf kratzt (siehe auch Abb. 3). Über die erzeugte Ontologie war dies einfach und Recall und Präzision lagen wieder bei 100%. Mit Keywords sind solche komplexeren Suchen schwer zu formulieren. So war es auch in gettyone.com. Unter anderem wurden diese Anfragen probiert: „chimpanzee scratching“, „chimpanzee AND hand AND hand“. Es wurden in [1] oft nur wenige Bilder oder Ergebnisse mit stark schwankender Präzision gefunden. Die Affen kratzten sich z.B. woanders als am Kopf, oder taten was anderes am Kopf als kratzen. Diese Beobachtungen konnte ich beim eigenen Versuch nachvollziehen (siehe Tab. 3).

	Ontologie aus [1]	gettyone.com <chimpanzee scratching> 2006	gettyone.com <chimpanzee AND hand AND head> 2006
Recall / Treffer	100%	11	4
Präzision im Vgl. zur Anfrage („irgendwo kratzen“, „hand irgendwie am kopf“)	100%	8 (73%)	2 (50%)
Präzision im Vgl. zu „mit der Hand am Kopf kratzen“	100%	6 (55%)	0 (0%)

Tabelle 3: Suche über die Ontologie im Vergleich zu manuell annotierten Kollektionen auf Basis von Keywords mit komplexeren Suchanfragen

Schlussfolgerung

Besonders die Vergleiche der Suche nach Fotos zeigen deutlich: manuelle Annotationen liefern deutlich präzisere Ergebnisse als die automatisch erzeugten Annotationen der großen Suchmaschinen. Manuelle Annotationen auf Basis von Keywords liefern gute Ergebnisse für einfache Standardanfragen mit möglichst wenigen Suchbegriffen. Bei komplexeren Suchwünschen scheitern sie aber.
Beachtet werden muss aber: zur Annotation auf Basis von Ontologien, muss das Hintergrundwissen in geeigneter Struktur für die zu betrachtende Domäne vorliegen. Und es ist klar: manuelle Annotationen allgemein werden von Menschen durchgeführt und menschliche Arbeitskraft ist am kostenintensivsten. Zur (halb-) automatischen Annotation müssen geeignete Detektoren vorliegen, oder es müssen Möglichkeiten geschaffen werden, vorhandene, auf keywordbasierte Annotationen in Ontologien zu überführen.

Ein weiterer Aspekt: Durch die Vorherrschaft von Google bei der Suche im Internet haben Nutzer inzwischen auch auch die Erwartung an eine Suche, dass sie genauso funktionieren sollte wie bei Google. Nämlich über ein Textfeld, in das Begriffe eingetippt werden können, woraufhin sofort passende Ergebnisse angezeigt werden [3]. Diese Erwartungshaltung müsste wohl auch eine auf Ontologien basierte Multimediasuche befrieden. So wäre es noch nötig, die Nutzeranfrage aus einem Textfeld geeignet auf die Ontologiesuche in [1], mit verteilten Feldern zur Eingabe von Eigenschaften aus vorgegebenen Wertebereichen, zu transferieren.

Zusammenfassend sind die Vorteile von Ontologien bei der Suche von Multimediadaten also:

> Sie können helfen Fotos und Videos durch die Verknüpfung mit vorhandenem Hintergrundwissen zu annotieren. Sie eignen sich gut zur manuellen oder zur halbautomatischen Annotation.
> Sie unterstützen die Generalisierung und Spezialisierung und verbessern und erweitern somit sowohl die Möglichkeiten der Annotation als auch der Suche
> Im Gegensatz zu Keywords legen Ontologien genau die Beziehungen zwischen den Bestandteilen eines Fotos fest. Ein großer Affe der neben einem kleinen Baum sitzt ist eben etwas anderes als ein kleiner Affe der auf einem großen Baum sitzt. Umstände die sich mit Keywords so gut wie nicht ausdrücken lassen.

Referenzen

1. A. Th. Schreiber, B. Dubbeldam, J. Wielemaker: *Ontology-Based Photo Annotation,* University of Amsterdam, 2001
2. L. Hollink, M. Worring, A. Th. Schreiber: *Building a Visual Ontology for Video Retrieval,* University of Amsterdam, 2005
3. Violeta Trkulja: *Suche ist überall, Semantic Web setzt sich durch, Renaissance der Taxonomien,* Password, 2005
4. V. Gudivada, V. Raghavan: *Content-Based Image Retrieval Systems*, IEEE Computer, Bd. 28, Nr. 9, 1995
5. Thomas Käster: *Inteligente Bildersuche durch den Einsatz inhaltsbasierter Techniken,* Universität Bielefeld, 2005
6. Google FAQ: *wie funktioniert die Bildsuche,* www.google.de/intl/de/help/faq_images.html, 2006
7. WordNet: http://wordnet.princeton.edu, Princeton University
8. Dan Brickley, R.V. Guha: *Resource Description Framework Schema (RDFS),* W3C, 2004

9. http://de.wikipedia.org/wiki/Wordnet, Wikipedia, 2006
10. Moving Picture Expert Group (MPEG): *MPEG-7 Standard,* 2002
11. S. Handschuh and S. Staab: *Annotation for the Semantic Web,* IOS Press, 2003.
12. Object Management Group (OMG): *UML 1.x,* 1997
13. A.M. Tam and C.H.C. Leung: *Structured Natural-Language Description for Semantic Content Retrieval*, J. American Soc. Information Science, 2001
14. Stephan Kopf, Thomas Haenselmann, Dirk Farin, Wolfgang Effelsberg: *Automatic Generation of Summaries for the Web,* University of Mannheim, 2004
15. TRECVID 2003: http://www-nlpir.nist.gov/projects/tv2003/tv2003.html
16. M.A. Hoang, J. Geusebroek, and A.W.M. Smeulders: *Color texture measurement and segmentation,* In Proceedings of the 2nd international workshop on Texture Analysis and Synthesis, 2002

Bildquellen

Abb. 1,2,3: A. Th. Schreiber, B. Dubbeldam, J. Wielemaker: *Ontology-Based Photo Annotation* University of Amsterdam, 2001

Abb. 4: Stephan Kopf, Thomas Haenselmann, Dirk Farin, Wolfgang Effelsberg: *Automatic Generation of Summaries for the Web,* University of Mannheim, 2004

Abb. 5: L. Hollink, M. Worring, A. Th. Schreiber: *Building a Visual Ontology for Video Retrieval,* University of Amsterdam, 2005