UNIVERSITÄT REGENSBURG – PHILOSOPHISCHE FAKULTÄT IV

Benutzerorientierte Evaluation von Content Based Image Retrieval-Systemen mit automatischer Beschlagwortung

Magisterarbeit im Fach Informationswissenschaft am Institut für Information und Medien, Sprache und Kultur

vorgelegt von:	Florian Greiner
Adresse:	Blaue-Lilien-Gasse 4a
	93047 Regensburg
Matrikelnummer:	114 03 97
Erstgutachter:	Prof. Dr. ChristianWolff
Zweitgutachter:	Prof. Dr. Rainer Hammwöhner
Laufendes Semester:	Wintersemester 2008/09
vorgelegt am:	23. Februar 2008

ZUSAMMENFASSUNG

Im Rahmen dieser Magisterarbeit wurde mit Hilfe einer benutzerorientierten Evaluation der Einsatz von automatischer Beschlagwortung beim Content Based Image Retrieval beurteilt. Neben einer umfassenden Erläuterung der theoretischen Ansätze beim Bildretrieval sollte anhand eines ausgewählten Beispielsystems ermittelt werden, ob die automatische Annotation von Bildmaterial Einfluss auf die Effektivität der Bildsuche und auf die Akzeptanz der Benutzer gegenüber dem System hat. Das derzeit einzige zugängliche Echtzeitsystem ALIPR wurde für die Untersuchung verwendet. Die aufgestellten Hypothesen bezüglich der Benutzerakzeptanz konnten bestätigt werden. Als Hauptgründe für die hohe Akzeptanz gegenüber automatischer Annotation, konnten die frühe Einbindung der Benutzer in den Retrievalprozess und die damit einher gehende Erhöhung der Retrievaleffektivität identifiziert werden.

ABSTRACT

This thesis analyzes the use of Automatic Annotation in Content Based Image Retrieval from a user-centered perspective. By illustrating the newest theories of Image Retrieval and especially of Automatic Annotation a brief overview is firstly presented to the reader. In a second step a real-world annotation system is evaluated with the focus on User Acceptance. The results of this study were: (1)There is a significant positive user acceptance on the use of Automatic Annotation. (2) Search with methods of Automatic Annotation is better accepted by users than search with methods of Image Retrieval. As main factors of this results the early user-involvement and accompanied by this, the higher Retrieval efficiency of Automatic Annotation are identified.

1 Einleitung

Digitale Fotografie hat in den vergangenen Jahren die analoge Fotografie beinahe vollständig abgelöst. Waren es früher analoge Kleinbildkameras, die in jedem Haushalt vorhanden waren, so wurden diese mittlerweile von ihren digitalen Pendants ersetzt. Laut einer Studie des Statistischen Bundesamtes, besaßen Anfang des Jahres 2006 42 Prozent der privaten Haushalte eine digitale Fotokamera (vgl. STATISTISCHES BUNDESAMT, 2007). Der damit einhergehende, rapide Anstieg des zur Verfügung stehenden Speicherplatzes – sowohl für die Kamera, als auch für den heimischen Computer – führt einerseits zu einer Verbesserung der Bildqualität, andererseits steigt die Zahl der gemachten Fotos.

Diese Bilderflut bringt jedoch neue Probleme und Herausforderungen mit sich. Bilddatenbanken wie Flickr oder iStockphoto sind zu riesigen digitalen Bildsammlungen angewachsen. An einem durchschnittlichen Tag werden bei Flickr mittlerweile mehr als eine Million neuer Bilder eingestellt (vgl. FORRET, 2006). Die Intention der Benutzer variiert dabei sehr stark. Einige benutzen Flickr als reine Datenablage, andere setzen die von Yahoo betriebene Bilddatenbank als persönliche Plattform zur Veröffentlichung ihrer Fotografien ein. Das Problem bleibt dasselbe: es ist schwierig, die Bilddatenbanken zu durchsuchen. Man kann sich leicht vorstellen, dass ein sinnvolles Annotations- und Retrievalsystem bei dieser Menge an Bildern unabdingbar ist.

Das Gebiet des Image Retrieval ist schon seit geraumer Zeit ein etablierter Bestandteil des modernen Information Retrieval. Seit den 80er Jahren wurde in diesem Bereich intensive Forschung betrieben (PRASAD ET AL., 1987). Mit dem Erfolg des World Wide Web und dem rapiden Fortschritt im Bereich der digitalen Fotografie, rückt dieses Thema immer mehr in den Mittelpunkt des Interesses. Im Gegensatz zum Text-Retrieval, gibt es bislang jedoch noch grundlegende Schwierigkeiten zu überwinden. Methoden zur automatischen Gewinnung von Information aus dem Bildinhalt sind noch nicht weit genug ausgereift, um zufriedenstellende Ergebnisse zu liefern. Aus diesem Grund wird oft auf unterstützende Methoden wie die Berücksichtigung von Textinformation, manuelle Indexierung oder die Verwendung von Metainformation zurückgegriffen.

Trotz diverser Ansätze ist es bislang nicht hinreichend gelungen, eine geeignete Methode zur automatischen Erschließung des Bildinhalts zu entwickeln. Im Zusammenhang mit diesem Thema wird vom Problem des ‘semantic gap’ gespro-

chen. Diese semantische Lücke kennzeichnet die fehlende Information, die zur Zusammenführung der durch automatische Inhaltserschließung gewonnenen Anhaltspunkte mit semantisch relevanter Information nötig wäre.

2 Image Retrieval

Die Forschung beschäftigt sich seit den 80er Jahren mit der inhaltlichen Erschließung von Bildmaterial und erste Ansätze zur Aufbereitung von Multimediadokumenten wurden erarbeitet. Mit der Einführung von Scannern rückten digitale Bildmedien immer mehr in den Fokus, da sie schnell weite Verbreitung gefunden hatten. In den Jahren zwischen 1990 und 2000 wurden erstmals Systeme besprochen, die den Inhalt eines Bildes oder einer Grafik automatisch erschließen und aufbereiten sollten. Ähnlich der automatischen Texterschließung, sollte dem Nutzer eine einfache Variante zur Erschließung und Suche von Bildern bereitgestellt werden. In der Literatur zum Bereich des Content-Based Image Retrieval (CBIR) wird dieser Zeitraum meist als der Beginn der modernen Inhaltserschließung von Bildern bezeichnet.

2.1 Forschungsschwerpunkte beim Image Retrieval

Die Hauptprobleme dieser Zeit lassen sich auf zwei zentrale Punkte zusammenfassen. Dies wären einerseits die Diskrepanz zwischen der digital erfassten Bildszene und der reellen Situation und andererseits die fehlende Erfassung beziehungsweise Verknüpfung der Bildszene mit inhaltlicher Information. Da bei beiden Punkten ein Mangel an (Bild-)Information vorliegt, werden diese Probleme gemeinhin als ‘sensory gap‘ und als ‘semantic gap‘ bezeichnet.

Neuerungen bei der Aufnahmetechnik von Kameras und Objektiven tragen zur Reduzierung der sensorischen Lücke entscheidend bei. Die Verbesserung von Bildsensoren die Erhöhung der Kameraauflösung, die Reduzierung von Störsignalen und Signalverlust und die Entwicklung von hochwertigen Optiken sind nur einige Beispiele zur Minderung des ‘sensory gap‘. Eine vollständige Schließung dieser sensorischen Lücke scheint mit der bisherigen Technik nur schwer vorstellbar, da noch keine Möglichkeit besteht die menschliche Wahrnehmung mit einem Algorithmus zu modellieren. Die eingesetzten Bildsensoren in Digitalkameras bilden die aufgenommene Szene auf eine Pixelmatrix ab, die der Wahrnehmung des menschlichen Auges nicht entspricht. Diese Matrix ist nur eine Annäherung an die natürlichen Szene, die mit Hilfe verschiedener mathematischer Methoden versucht das natürliche Sehen zu imitieren (vgl. SMEULDERS ET AL., 2000, pp. 1351-1352).

Das Problem des so genannten ‘semantic gap‘ stellt ein zweites Kernproblem im CBIR dar. Analog zur sensorischen Lücke, gilt auch bei der semantischen Lü-

cke, dass eine vollständige Schließung kaum realisierbar ist. Die Modellierung der menschlichen Interpretation mit mathematischen Ansätzen ist nur schwer umzusetzen. Doch gerade die Anreicherung der Bilddaten mit semantischer Information macht sich das CBIR zum Ziel. Aus diesem Grund muss versucht werden, die vorliegenden Bilddaten zur Gewinnung von semantischer Information zu nutzen und signifikante Merkmale abzuleiten (vgl. SMEULDERS ET AL., 2000, p. 1353).

2.2 Anfragemöglichkeiten

Beim Textretrieval ist das in der Praxis vorherrschende Anfrageparadigma die natürlichsprachliche Formulierung der Frage. Beim Bildretrieval hat sich noch keine der bestehenden Anfragemethoden, beziehungsweise –modalitäten, als Standard herauskristallisiert. Dem Benutzer von Bildretrievalsystemen stehen neben textbasierten Anfragemethoden auch verschiedene grafische Varianten zur Formulierung seiner Suchanfrage zur Verfügung. Abbildung 1 zeigt eine Variante zur detaillierten Einordnung verschiedener Anfragemethoden nach Smeulders (vgl. SMEULDERS ET AL., 2000, pp. 1365-1367).

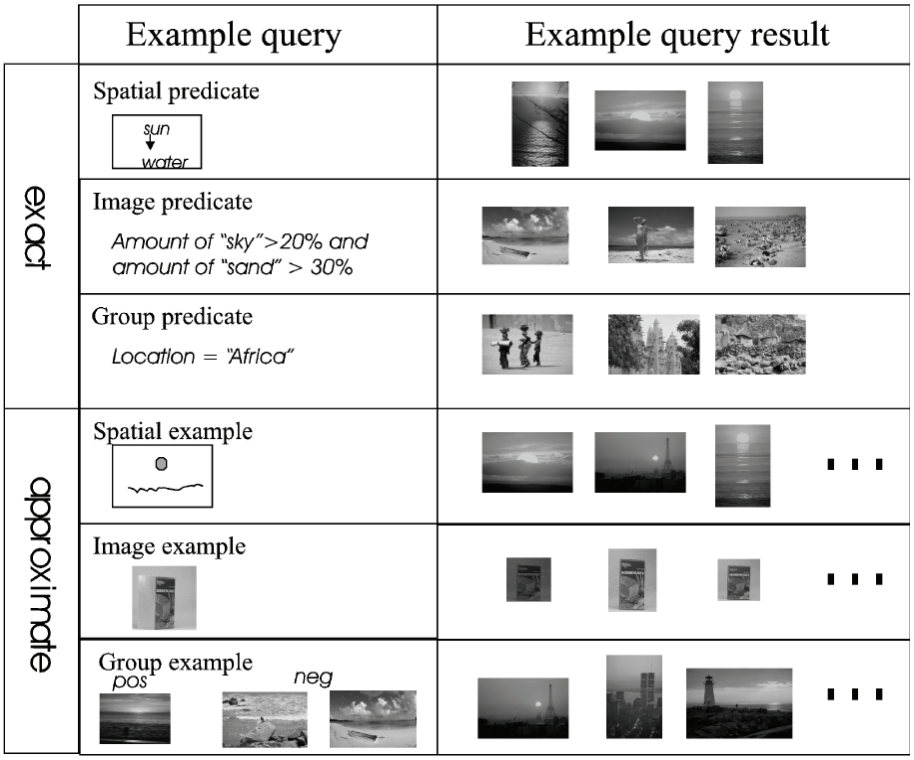

Abbildung 1: Beispiele für komplexe Anfragemethoden (SMEULDERS ET AL., 2000, p. 1366)

2.2.1 Textbasierte Methoden

Die gängigste textbasierte Methode ist die Freitextsuche, bei der die Suchanfrage in natürlichsprachlicher Form gestellt wird. Als Beispiele für Freitextsuchsysteme können Systeme wie die Flickr oder die Google Bildsuche genannt werden, wobei es sich dabei nicht um inhaltsbasierte Bildsuchsysteme handelt.

Viele der für die Freitextsuche von Bildern relevanten Forschungsbeiträge weisen Parallelen zu Beiträgen aus dem Textretrieval auf. Im Detail existieren jedoch auch viele Unterschiede. Im Gegensatz zu Texten sind Bilder nicht zwangsläufig mit textueller Information verbunden, was die Suche erschwert. Die Bilderindexierung benötigt Methoden zur Analyse der dazugehörigen textuellen Beschreibungen. Zum Bild gehörige Textbausteine wie beispielsweise der Bildtitel und beschreibende Texte zum Bild, aber auch Metadaten müssen identifiziert und verarbeitet werden. Bei der Freitextsuche nach Bildern werden häufig grafische Primitive wie die Farbe, Form und Struktur von Bildinhalten von Benutzern vernachlässigt und stattdessen komplexe Konzepte abgefragt. Neben der Verbesserung und Vereinheitlichung der Annotationsmethoden ist ein weiterer Ansatz zur Lösung dieses Problems die Anreicherung der Suchanfrage mit semantischer

Information, zum Beispiel durch Folksonomies (vgl. S. LEE & YONG, 2007). So genannte 'ESP-games' versuchen durch die Einbeziehung spielerischer Elemente Benutzer zu manueller Annotation von Bildern anzuregen (vgl. VON AHN & DABBISH, 2004, 2008). Als populäre Beispiele sind hierbei die Webanwendung Google Image Labeler (vgl. GOOGLE, 2007) beziehungsweise dessen Nachfolger von GWAP (vgl. GWAP, 2009) zu nennen. Weitere Ansätze zur Überwindung von Sprachbarrieren bei der textbasierten Suche, werden im Rahmen der Image-CLEF Initiative behandelt (vgl. GRUBINGER ET AL., 2008). Die automatische Generierung von Schlagwörtern durch textbasierte Indexierungsmethoden stellt aufgrund des hohen Aufwandes zur korrekten, manuellen Beschlagwortung einen Forschungsschwerpunkt textbasierter Suchmethoden dar (vgl. S. LEE & YONG, 2007). Als zentraler Bestandteil dieser Arbeit wird der Bereich der automatischen Beschlagwortung in einem separaten Abschnitt in Kapitel 3.2 im Detail erläutert.

2.2.2 Bildbasierte Methoden

Die Formulierung von optischen Anfragen liegt beim Bildretrieval nahe, da identische Anfrage- und Ausgabeparadigmen verwendet werden. Das heißt, die Suche nach einem Bild kann in vielen Fällen am besten durch ein optisches Beispiel „formuliert" werden. Eine Suchanfrage dieser Art wird als 'query by visual example' – kurz QBVE bezeichnet (vgl. HIRATA & KATO, 1992, pp. 56-71). Dabei kann zwischen grafischen Suchanfragen und bildlichen Suchanfragen unterschieden werden.

Bei der grafischen Suchanfrage wird vom Benutzer eine Beispielzeichnung angefertigt. Diese Zeichnung wird hinsichtlich ihrer graphischen Merkmale analysiert und ein Abgleich mit der verbundenen Bilddatenbank durchgeführt. In den Anfangsjahren der inhaltsbasierten Bildsuche war diese Art der Anfrage sehr populär. Dies wird dadurch belegt, dass als bedeutendster Repräsentant von QBVE das QBIC-System von IBM angesehen werden kann, das bereits 1995 vorgestellt wurde (vgl. ASHLEY ET AL., 1995, p. 475) und in modifizierter Form noch immer eingesetzt wird. Neben der Integration in das DB2 Projekt von IBM (vgl. IBM, 2008) bietet die Webpräsenz des Eremitage in Sankt Petersburg die QBIC Bildsuche zur Recherche in der digitalen Sammlung an (vgl. STATE HERMITAGE MUSEUM, 2003). Neuere Bildretrievalsysteme arbeiten auf Grund der vielen Nachteile von grafischen Suchanfragen nur noch selten dieser Technologie. Das auf Teilen der Flickr-Datenbank basierende Retrievr-System (vgl. LANGREITER, 2006) oder das ImgSeek–System (vgl. IMGSEEK.NET, 2009; JOHN R. SMITH & S. CHANG, 1996) bieten diese Variante der Suchanfrage jedoch weiterhin an. Die entschei-

denden Nachteile solcher QBVE-Suchanfragen sind einerseits deren hohe Komplexität und andererseits ihr fehlerhafter Abstraktionsgrad.

Je nach Benutzeroberfläche muss der Anwender das Gesuchte mit Hilfe verschiedener graphischer Werkzeuge nachzeichnen. Abgesehen vom hohen zeitlichen Aufwand einer solchen Anfrage, setzt ihre Durchführung Kenntnisse voraus, die viele Benutzer überfordern. Durch die Übertragung der Suchanfrage auf eine zweidimensionale Zeichenoberfläche treten beiderseits Interpretationsfehler auf.

> „The more complex the query i.e. the shape and structure, the more difficult it is for the end user to express and produce a meaningful visual example. “ (VENTERS ET AL., 2001, p. 515)

Das Problem der hohen Komplexität wirkt sich zusätzlich auf den Einsatz der Indexierungsmethoden aus. Die vom Benutzer erstellte Zeichnung setzt sich – wie in Abbildung 2 zu erkennen – zumeist aus wenigen geometrischen und farblichen Primitiven zusammen. Man unterscheidet zum einen zwischen ikonischen und realistischen, zum anderen zwischen kompletten und gezielten Suchanfragen. Komplexe Strukturen, Muster oder Farbverläufe sind einerseits zu aufwändig umzusetzen und andererseits zeichnerisch zu anspruchsvoll. Die Indexierungsmethoden müssen daher an das Niveau der Suchanfrage angepasst werden, um die semantische Lücke zwischen der Suchanfrage und den gesuchten Bildern zu überbrücken.

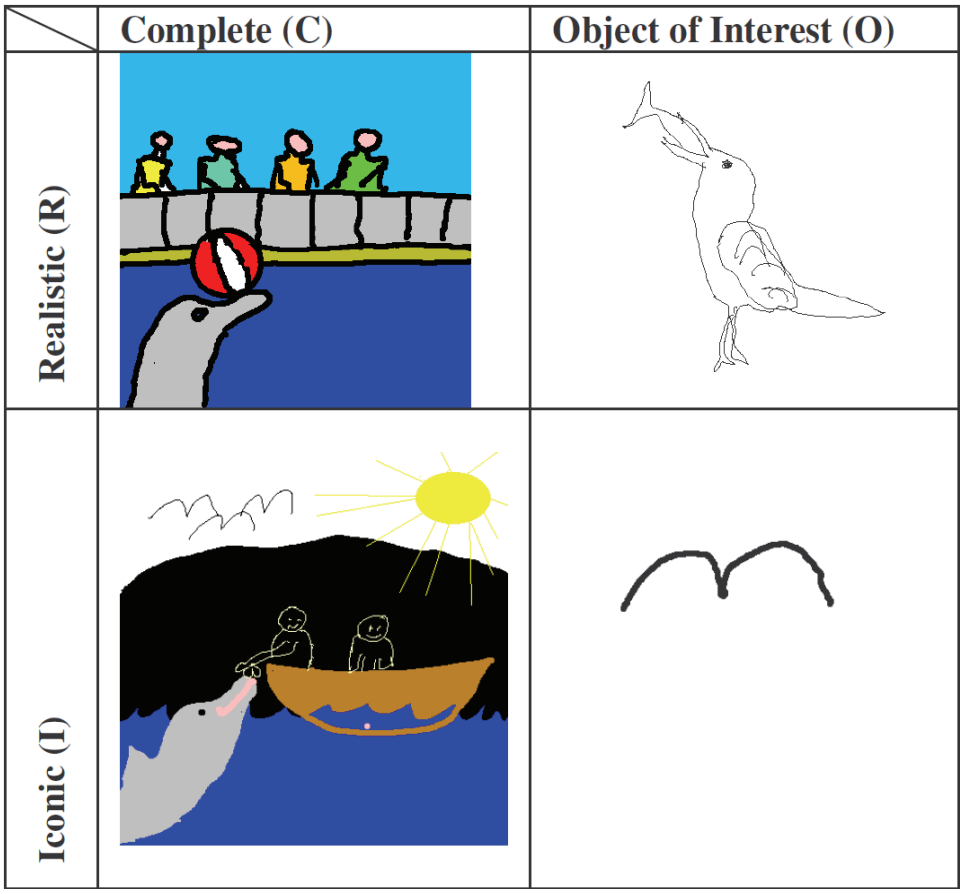

Abbildung 2: Beispiel für realistische und ikonische Suchanfragen bei grafischen Suchsystemen (HOVE, 2007, p. 8)

Suchanfragen mit Hilfe eines Beispielbildes werden von einer Vielzahl aktueller Projekte unterstützt (vgl. ALIPR, 2009; COGNISIGN LLC, 2008; LANGREITER, 2006). Dabei wird die Suchanfrage durch einen Verweis auf ein vergleichbares Bild gestellt. Das Bild wird mit den vorhandenen Indexierungsmethoden verarbeitet und eine Signatur erstellt. Diese wird mit den bereits indexierten Bildern in der Datenbank abgeglichen und die besten Ergebnisse werden zurückgeliefert. Der zeitliche und kognitive Aufwand ist im Vergleich zur behandelten grafischen Anfragemethode geringer und erfordert weniger Eigeninitiative des Benutzers. Da die Suchanfrage im selben Modus formuliert wird wie die erwarteten Ergebnisse, profitieren auch die eingesetzten Indexierungsinstrumente.

Betrachtet man die verschiedenen Anwendungsbereiche von Bildretrievalsystemen treten jedoch auch Nachteile auf. Stehen dem Benutzer keine geeigneten Beispielbilder zur Verfügung, ist es ihm nicht möglich eine Suchanfrage an das System zu stellen. Der Anwender ist also gezwungen bereits eine bildliche Vor-

stellung vom gesuchten Objekt und zusätzlich Zugriff zu einer ausreichend großen Bilddatenbank zu haben (vgl. SMEULDERS ET AL., 2000, pp. 1366-1367).

2.2.3 Zusammengesetzte Methoden

Neben einer Vielzahl von text- und bildbasierten Anfragevarianten existieren auch eine Menge zusammengesetzter Methoden. Sie bestehen aus den verschiedenen Kombinationen der bereits genannten Anfragemethoden. Zusätzlich werden häufig spezielle Interaktionsoptionen wie Relevance Feedback oder Varianten zur Anfrageverfeinerung angeboten. Durch die Verknüpfung der Vorteile der verschiedenen Anfragemodi wird ein Großteil ihrer Nachteile aufgehoben. Liegt dem Benutzer kein geeignetes Beispielbild für die Suchanfrage vor, kann er beispielsweise auf die textbasierte Methode umsteigen oder eine grafische Suchanfrage starten. Mit der effektiven Verarbeitung von gemischten Suchanfragen treten natürlich auch neue Herausforderungen auf. Das vorliegende Bildmaterial muss auf andere Weise erschlossen und den Anfragemodalitäten angepasst werden

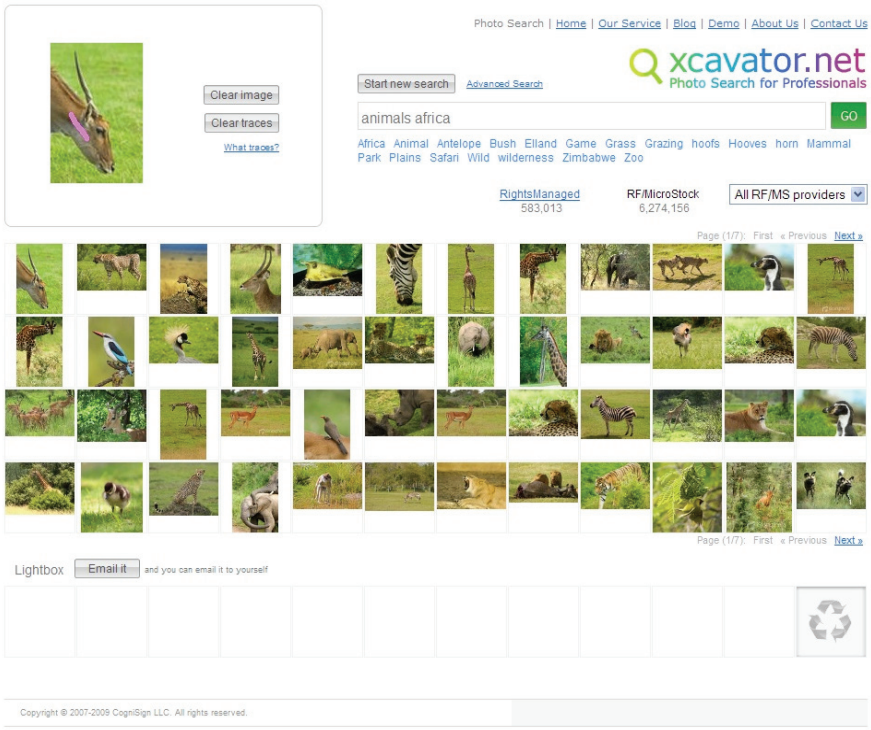

Abbildung 3: Beispiel einer Bildsuchmaschine mit kombinierten Anfragemethoden (vgl. COGNISIGN LLC, 2008)

(vgl. NATSEV ET AL., 2004, pp. 305-309). Neben der parallelen Verarbeitung der unterschiedlichen Suchanfragen existieren auch Ansätze zur einheitlichen Analyse der Bildinformation. Das heißt, die aus dem Bildinhalt extrahierte Information wird sinnvoll mit der dazugehörigen Textinformation verknüpft (vgl. DATTA ET AL., 2008, p. 14).

Die in Abbildung 3 gezeigte Benutzeroberfläche bietet dem Benutzer alle drei der erwähnten Möglichkeiten zur Formulierung der Suchanfrage. Nach Eingabe eines Schlagworts in das Textfeld werden Ergebnisse geliefert. Durch einfaches 'drag & drop' kann ein Bild aus der Ergebnismenge in das Suchfeld am linken oberen Bildschirmrand gezogen werden. Dadurch wird eine inhaltsbasierte Suchanfrage nach diesem Bild gestartet. Ist die Qualität der Ergebnisse noch nicht ausreichend, können Teilbereiche des Suchbildes grafisch markiert werden um ihre Relevanz besonders hervorzuheben. Zusätzlich können jederzeit weitere Schlagwörter zur Suche hinzugefügt werden, wobei der Benutzer durch den Einsatz einer 'tag cloud' unterstützt wird. Die Letzteren beiden genannten Suchmodalitäten dienen ferner als integrierte Relevance Feedback Funktionen, um die Suchanfrage zu verfeinern (vgl. COGNISIGN LLC, 2008).

2.3 Ausgabemöglichkeiten und Visualisierung

Die Gestaltung der Benutzeroberfläche und die Art der Ergebnispräsentation spielen bei Bildretrievalsystemen eine ebenso große Rolle, wie bei allen anderen Anwendungen. Die Einhaltung von Designprinzipien und Standards bezüglich der Systemfunktion und der äußeren Erscheinungsform tragen entscheidend dazu bei, dass der Benutzer das Bildretrievalsystem akzeptiert und wiederverwendet. Große Teile, der für die Ergebnisvisualisierung anzuwendenden Regeln und Maximen, fallen jedoch unter den Bereich der allgemeinen Softwareusability. Die gängigste Art der Ergebnispräsentation beim CBIR ist die hierarchische Ordnung der Ergebnisse nach bestimmten Kriterien. Analog zu Textretrievalsystemen ist es dabei zum Beispiel möglich, die Ergebnisse anhand ihrer Relevanz bezüglich der Suchanfrage oder anhand ihrer chronologischen Daten anzuordnen. Meist wird zur Darstellung der ermittelten Bilder auf ein Rastersystem zurückgegriffen, bei dem die Bilder der gewählten Hierarchie entsprechend angeordnet werden. Zudem existieren verschiedene abgewandelte Ordnungs- und Darstellungsmethoden, die speziell auf die vom System verwendeten Anfrage- beziehungsweise Indexierungsvarianten ausgerichtet sind (vgl. DATTA ET AL., 2008, pp. 13-15).

Chen et al. beschreiben ein Verfahren zur Darstellung von Ergebnissen anhand ihrer Bildsignatur, bei dem Bilder mit ähnlicher Signatur als zusammenge-

hörige Cluster ausgegeben werden (vgl. CHEN ET AL., 2003, pp. 194-196; X. WANG ET AL., 2004, pp. 437-439). Einen guten Überblick über unkonventionelle Darstellungsvarianten bietet Porta in seinem Paper von 2006, in dem er unter Anderem auch dreidimensionale Methoden vorstellt (vgl. PORTA, 2006, pp. 440-444).

2.4 Bildanalyseverfahren

Die Grundlage zur inhaltlichen Erschließung von Bildern stellt die quantitative Analyse des Bildmaterials dar. Es gilt die vorliegenden Bilddaten hinsichtlich bestimmter Merkmale zu untersuchen. Smeulders unterteilt diese Analyse in die Phase der Bildberechnung und die Phase der Merkmalskonstruktion (vgl. SMEULDERS ET AL., 2000, pp. 1354-1357). Bei der Bildberechnung werden die Bildinformationen abhängig von dem zu extrahierenden Merkmal analysiert und aufbereitet. Diese „raffinierten" Bilddaten werden als Signaturen bezeichnet. Die Signaturerstellung und die Merkmalskonstruktion sind häufig nicht klar zu trennen, da die verwendeten Algorithmen abhängig voneinander sind. Das heißt, ein Algorithmus zur Erstellung einer Bildsignatur erfordert meist einen bestimmten Algorithmus zur Konstruktion von Merkmalen. Diese Tatsache erklärt die Vielzahl unterschiedlicher Algorithmen und Methoden zur Merkmalsextraktion, die je nach Situation unterschiedlich effektiv eingesetzt werden können (vgl. DATTA ET AL., 2008, pp. 17-18).

2.4.1 Einfache Bildmerkmale

Zuerst wird der Begriff 'feature' im Zusammenhang mit Content Based Image Retrieval definiert:

> „A feature is defined as a distinctive characteristic of an image and a descriptor is a representation of a feature. " (MANJUNATH & W-Y. MA, 2002, p. 313)

Als 'low-level features' werden dabei in erster Linie vier Erkennungsmerkmale eines Bildes bezeichnet: Die Farbe, die Struktur, die Form und so genannte affine Bildbereiche(vgl. DATTA ET AL., 2008, p. 20). Diese vier Merkmale sind separat aus einem digitalen Bild extrahierbar und ermöglichen eine Inhaltsbeschreibung des Bildes. Diese Merkmale an sich beinhalten nur geringe semantische Information, weshalb sie – neben der gängigen Bezeichnung low-level features – auch syntaktische Merkmale genannt werden. Im Folgenden werden die üblichen Verfahren zur automatischen Gewinnung der Einzelmerkmale erläutert und deren Aussagekraft hinsichtlich des Bildinhalts diskutiert.

2.4.1.1 Farbe

Die Farberkennung stellt in der Literatur zur Inhaltserschließung von Farbbildern ein zentrales Thema dar. Je nach Anwendungssituation und verfolgter Erschließungsstrategie existieren mittlerweile sehr viele verschiedene Ansätze. Grundsätzlich stützen sich alle Varianten aber auf ein zentrales Verfahren: die Extraktion der Farbinformation aus einem Bild durch das Erstellen eines Farbhistogramms. Bereits 1991 befassten sich Swain und Ballard mit Farbhistogrammen zur Indexierung von Bildmaterial (vgl. SWAIN & BALLARD, 1991).Die Farbzusammensetzung eines digitalen Bildes wird quantitativ erfasst und in ein Histogramm übertragen werden. Durch pixelweises Auszählen wird die Häufigkeitsverteilung jeder auftretenden Farbe im Bild ermittelt und in ein Histogramm übertragen.

Image Retrieval durch den Vergleich von Farbhistogrammen bietet sehr viele Vorteile. Die Extraktion der Farbinformation ist sehr robust, da die Histogramme unabhängig von der Bildgröße und der Bildrotation erstellt werden können. Durch den überschaubaren und klaren mathematischen Ablauf der Histogrammerstellung können Algorithmen zur Farberkennung relativ einfach implementiert werden, deren Durchführung erfordert wenig Rechenaufwand und die Speicherung der Histogramme verbraucht sehr wenig Speicherplatz (vgl. KONSTANTINIDIS ET AL., 2006, p. 25).

Nichtsdestotrotz treten auch eine Reihe von Problemen auf. Da das bei digitalem Bildmaterial häufig verwendete RGB Farbmodell dem Farbraum der menschlichen Wahrnehmung nicht nahe kommt, wird die Farbinformation vor der Erstellung eines Histogramms in einen geeigneten Farbraum transformiert (vgl. GONZALEZ & WOODS, 2007, pp. 401-414). Farbmodelle wie HSV, YIQ, CIE-LUV, CIE-LAB oder HMMD werden in der Literatur neben vielen anderen angeführt und in der Praxis verwendet (vgl. J. R. SMITH, 2002a, pp. 287-290). Eine Einigung auf ein Farbmodell ist nicht absehbar, da jedes Modell spezifische Vor- und Nachteile bei der automatischen Verarbeitung aufweist.

Bilder die ein und denselben Inhalt in verschiedenen Lichtsituationen darstellen stellen ebenfalls ein Problem dar. Da bereits bei einer geringen Veränderung der Lichtsituation große Farbverschiebungen auftreten, können Bilder mit identischem Inhalt sehr unterschiedliche Histogramme aufweisen. Trotz desselben Bildinhalts können die Aufnahmen anhand der Farbmerkmale nicht als ähnlich identifiziert werden. Dieser Nachteil kann durch die Verwendung von speziellen Farbmodellen eingeschränkt werden, die den Farbkontrast verbessern und die verschiedenen Lichtsituationen nivellieren. Dieser Umstand ist ein weiteres

Argument für die Vielzahl der bereits erwähnten Farbmodelle (vgl. SMEULDERS ET AL., 2000, pp. 1354-1355; J. R. SMITH, 2002b, pp. 287-290)

Wegen der kontinuierlichen Verbesserung der Bildsensoren bzw. der digitalen Aufnahmetechnik im Allgemeinen, sind Digitalbilder nicht auf bestimmte Auflösungen oder Seitenverhältnisse normiert. Aus diesem Grund ist es unumgänglich, die erstellten Farb-Histogramme zu normalisieren, um sie vergleichbar zu machen. Ein Bild mit einer Auflösung von 1024 mal 768 Pixel (mit 786432 Pixel eine typische Bildschirmauflösung) muss also beispielsweise mit einem Bild von 3840 mal 2160 Pixel (mit 8294400 Pixel die Auflösung einer 16:9 Digitalkamera) vergleichbar gemacht werden. Der detaillierte Ablauf einer solchen Normalisierung wird von Swain und Ballard ausführlich im Zusammenhang mit computerbasiertem Bildretrieval beschrieben (vgl. SWAIN & BALLARD, 1991).

Ein weiterer Schwachpunkt der Farberkennung, der weder durch eine Normalisierung der Histogramme noch durch den Einsatz modifizierter Distanzmaße behoben wird, ist die fehlende Einbeziehung der lokalen Bildinformation. Sind die zugrunde liegenden Histogramme zweier Bilder gleich, so bedeutet dies noch lange nicht, dass die Bilder sich inhaltlich gleichen. So kann ein Bild die identische Farbinformation wie ein anderes Bild enthalten und es kann trotzdem keinerlei inhaltliche Übereinstimmung bestehen (vgl. Abbildung 4). Konstantinidis et al. bemerken hierzu:

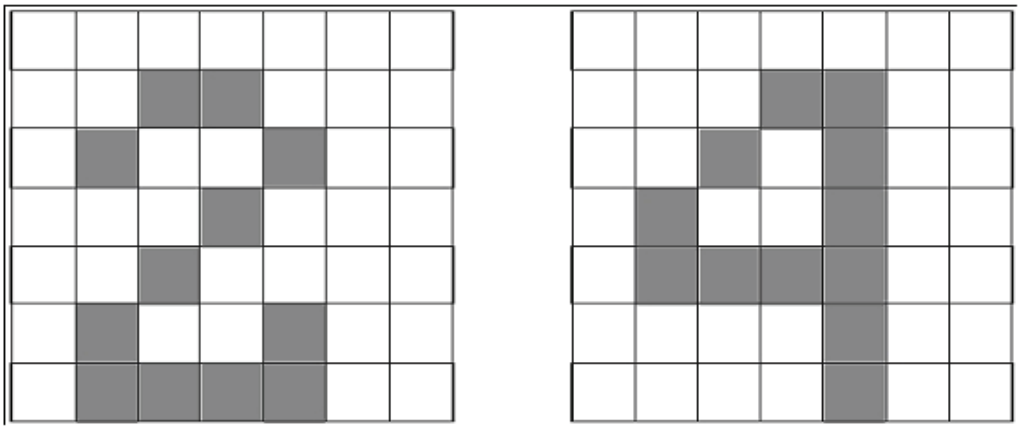

Abbildung 4: Grafik mit identischen Histogrammen, aber ohne inhaltliche Übereinstimmung (KONSTANTINIDIS ET AL., 2006, p. 28)

> „Their semantic content is noticeably different, so evidently, one cannot assume that color distribution alone is always sufficient to

> represent the pictorial content of an image, since it comprises an abstract representation of it.“ (KONSTANTINIDIS ET AL., 2006, p. 27)

Einen Lösungsansatz bietet die Aufteilung eines Bildes in mehrere Bereiche. Sie spielt vor allem bei der Gewinnung zusammengesetzter Bildmerkmale eine entscheidende Rolle, da zusätzlich zur Farbinformation auch lokale Information gewonnen wird, die Rückschlüsse auf den Bildinhalt zulässt. Eine ähnliche Methode befasst sich mit der Erstellung von mehreren Histogrammen von unterschiedlicher Auflösung. Neben einem globalen Histogramm werden Histogramme von unterschiedlich großen Teilbereichen berechnet. Diese Methode wird in der Literatur als „mehrdimensionale“ Analyse der Farbinformation bezeichnet (vgl. KONSTANTINIDIS ET AL., 2006, p. 28).

Um eine verbesserte Vergleichbarkeit von verschiedenen Farbsignaturen zu erreichen, ist eine standardisierte Zusammenfassung von Farbwerten vonnöten. Spezielle Farb- bzw. Histogramm-Deskriptoren wie beim ISO-Standard MPEG-7, sind ein Ansatz zur Lösung dieses Problems (vgl. MARTÍNEZ, 2004).

2.4.1.2 Texturerkennung

Das Erkennen und Eingrenzen einer Textur gestaltet sich schwieriger als die Erkennung von Farbmerkmalen. Im Unterschied zur Erfassung von Farbinformationen lassen sich Texturmerkmale nicht direkt aus der Pixelinformation ablesen. Deshalb sollte zuerst allgemein bestimmt werden, was als Textur oder Muster in der Bildanalyse angesehen wird:

> „Texture has no universally accepted formal definition, although it is easy to visualize what one means by texture. One can think of a texture as consisting of some basic primitives [...] whose spatial distribution in the image creates the appearance of a texture.” (MANJUNATH & W-Y. MA, 2002, p. 313)

Obwohl das menschliche Auge eine Textur relativ einfach erkennt, ist es schwierig eine universell anwendbare Definition zu finden. Manjunath und Wie-Ying sprechen in diesem Zusammenhang davon, dass Texturen nur schwer durch qualitative und quantitative Kriterien beschrieben werden können (vgl. MANJUNATH & W-Y. MA, 2002, p. 313).

Der bestimmende Wert zur Extraktion von Texturen aus Bilddaten ist die Helligkeit. Durch die Berechnung von Änderungen in den Helligkeitswerten eines Bildes können Rückschlüsse auf die vorliegenden Texturen im Bild gezogen werden. Julesz verfolgte bereits 1975 diesen Ansatz und versuchte durch den Einsatz

einer so genannten Grauwertübergangsmatrix Texturen zu extrahieren (vgl. JULESZ, 1975).

Ein häufig verwendetes Modell basiert auf einer Methode von Tamura et al., die Texturen bestimmte Attribute zuteilt. Durch psychologische Wahrnehmungstests beim Menschen unterschied Tamura ursprünglich die Attribute Körnung/Granularität, Kontrast, Gerichtetheit, Linienartigkeit, Regelmäßigkeit und Rauheit. Die mathematischen Details zur Analyse dieser Attribute können im Originaltextnachgelesen werden (vgl. TAMURA ET AL., 1978). Häufig werden diese Attribute in aktuelleren Varianten zu drei Kategorien zusammengefasst: die Granularität, die Oberflächenbeschaffenheit und die Regelmäßigkeit (vgl. MANJUNATH & W-Y. MA, 2002, pp. 318-319).

Vor allem bei der Verwendung in spezifischen Domänen liefern Texturmerkmale sehr gute Ergebnisse. Bei medizinischen Aufnahmen oder topologischen Aufnahmen sind Texturen meist so aussagekräftig, dass sie entscheidende Erkenntnisse zum Ähnlichkeitsvergleich verschiedener Bilder beitragen. Aber auch in allgemeinen Retrieval-Umgebungen lassen sich durch Texturen wertvolle semantische Erkenntnisse gewinnen (vgl. DATTA ET AL., 2008, pp. 20-21).

Neben vielen anderen Methoden zur Aufbereitung der Bildinformation hinsichtlich der Erkennung von Texturen ist die gängigste Methode die der Wavelet-Transformation. Die Schwankungen der Helligkeitswerte von Pixel zu Pixel werden dabei als Frequenz betrachtet. Durch die Anwendung der diskreten Fourrier-Transformation – einer Variante der Wavelet-Transformation – auf die Helligkeitsinformation wird eine Annäherung an die menschliche Texturwahrnehmung erzielt, da Helligkeitswerte gefiltert werden, die vom menschlichen Sehapparat nicht wahrgenommen werden können. Das Problem der unterschiedlichen Skalierung von Texturen stellte im Bildretrieval lange Zeit eine zentrale Hürde dar. Die Wavelet-Transformation und deren Varianten wie zum Beispiel so genannte Gabor Filter (vgl. A.K. JAIN & FARROKHNIA, 1990) erzielen bei Texturen von unterschiedlicher Skalierung sehr gute Ergebnisse und werden deshalb sehr häufig anderen Extraktionsmethoden vorgezogen (vgl. BACHOO & TAPAMO, 2005; KRUIZINGA ET AL., 1999).

Wie bei der Farberkennung, liegt der Fokus der Forschung derzeit auf geeigneten Methoden zur regionsbasierten Analyse von Strukturen – also einer Segmentierung der Bildinformation. Eine falsche Eingrenzung kann zur falschen Interpretation der Struktur führen oder ein Strukturmerkmal kann dadurch vollständig verloren gehen. Das Auftreten eines solches ‘scaling problem‘ kann einer-

seits durch die durchdachte Wahl der Extraktionsmethoden vermindert werden, andererseits gibt es bereits bewährte Methoden zur automatischen Auswahl von Ausschnitten hinsichtlich ihrer Struktur (vgl. CARSON ET AL., 2002).

2.4.1.3 Formerkennung

Die Formerkennung kann in der Forschung im Bereich der computergestützten Bilderkennung noch als eher „junger" Teilbereich betrachtet werden (vgl. MEHTRE ET AL., 1997). Dies lässt sich durch die Tatsache begründen, dass die Erkennung von Formen in Bildern – sowohl mathematisch, als auch im Hinblick auf die Rechenleistung – aufwändig und komplex ist. Nichtsdestotrotz sind richtig erkannte Formen ein Schlüsselmerkmal bei der semantischen Erschließung von Bildern. Analog zu den restlichen Bildmerkmalen konzentrierte man sich bei der Formerkennung zuerst auf die Analyse von globalen Formmerkmalen. Die diskrete Kurvenevolution zur Extraktion von vereinfachten zweidimensionalen Formen (vgl. LATECKI & LAKAMPER, 2000) oder Methoden zur Konturvereinfachung zur Entfernung von unnötiger Detailinformation waren erste erfolgversprechende Ansätze (vgl. CETINKAYA ET AL., 2006).

Der Einsatz von Fourrier-Transformationen und deren Varianten sind auch bei der Formerkennung ein weit verbreitetes Mittel. Fourrier-Transformationen untersuchen veränderte Bildinformation auf hervorgehobene Kanten und Kurvenbruchstücke – so genannte 'tokens' –und ordnen diese je nach Biegung und Orientierung bestimmten Klassen zu. Die auf solche Weise kategorisierten 'tokens' dienen als Merkmale für die Generierung der Formmerkmale. Die Berechnung von Fourrier-Transformationen stellt derzeit noch eine enorme Hürde für praktische Anwendungen dar, da die verwendeten Algorithmen sehr rechenintensiv sind (vgl. FOLKERS & SAMET, 2002, pp. 522-523).

Neue Ansätze wie die der so genannten 'Image Entities' gründen auf bereits bestehenden Methoden aus der Mustererkennung. Eine Datenbank aus Bild- oder Objekt-Entitäten wird mit dem Inhalt eines zu indexierenden Bildes abgeglichen. Dieses vom 'iconic indexing' abgeleitete Verfahren, bei dem nach geometrischen Primitiven gesucht wird, erlaubt eine virtuelle Nachbildung der Originalbilder die lediglich die gefundenen Entitäten und deren räumliche Anordnung enthalten. Diese Nachbildungen dienen als Basis zum Bildvergleich. Besonders erfolgversprechend scheint diese Methode bei eingeschränkten Domänen, wie zum Beispiel Architektur oder Kunst, da häufig klare Strukturen im Bildmaterial enthalten sind (vgl. DATTA ET AL., 2008, p. 22).

Mehr noch als bei den vorhergehenden Bildmerkmalen spielt bei der Formerkennung die Segmentierung eine wichtige Rolle. Durch die Eingrenzung bestimmter Objekte im Bild stellt die Formerkennung an sich bereits eine eigene Variante der Segmentierung dar. Im Vergleich zur Zuteilung eines einzelnen globalen Formmerkmals zu einem Bild, liefert diese Art der Segmentierung ein Vielfaches an semantischer Differenzierung. Dabei ist der zu Grunde liegende Rechenaufwand und die Chance einer Fehleinteilung ähnlich dem 'scaling problem' nicht zu verachten. Eine detailliertere Aufstellung von Vor- und Nachteilen der verschiedenen Segmentierungsmethoden wird in dem dafür vorgesehenen Kapitel vorgenommen (vgl. Kapitel 2.4.2).

2.4.1.4 Affine Bildbereiche

Ein Grundproblem bei der Analyse von Bildern hinsichtlich ihrer Basismerkmale stellen perspektivische Veränderungen dar. Gerade bei der Berechnung von Struktur- und Formmerkmalen sind Änderungen an Bildausrichtung, Skalierung oder Verzerrungen nur schwer zu kompensieren. Ebenso kann eine leichte Veränderung der Beleuchtungssituation zu sehr großen Unterschieden bei den Farbsignaturen führen, obwohl der Bildinhalt identisch ist. Merkmale, die invariant gegenüber solchen Einflüssen sind, werden als 'salient points' oder 'points of interest' – sinngemäß übersetzt affine Bildbereiche – bezeichnet. Jedes Bild besitzt bestimmte Punkte, die trotz einer Skalierung oder anderer Verformungen unverändert bleiben. Häufig stellen Eckpunkte oder Kanten mit markanten Kontrastabbrüchen solche unveränderlichen Punkte dar. Durch den Einsatz von Wavelet-Transformationen können solche markanten Punkte entdeckt und als Bildmerkmale verwendet werden. Neben der erwähnten Invarianz gegenüber vielen störenden Faktoren der Bildanalyse haben solche 'salient points' den weiteren Vorteil, dass eine relativ geringe Zahl von Bildpunkten ausreicht, um ein Bild zu indexieren. Bei der Erstellung einer Signatur sind also verhältnismäßig wenige Daten notwendig, was eine Verringerung des Rechenaufwandes mit sich bringt. Beim Bildretrieval besitzen 'salient points' dadurch eine hohe diskriminative Kraft, da sie bei geringem Vergleichsaufwand – also einer „kleinen Signatur" – eine hohe Unterscheidbarkeit gewährleisten (vgl. LOWE, 2004, pp. 96-114).

Nachteilig wirkt sich bei der Verwendung von 'salient points' zur Indexierung aus, dass es teilweise zu Fehleinschätzungen oder zur falschen Auswahl von Punkten kommen kann. Bei Verschmutzungen des Bildsensors oder des Objektives bei der Aufnahme kommt es zur Übertragung dieser Einschlüsse auf das Bild. Diese Störungen sind unabhängig von der Aufnahmesituation und befinden sich

immer an der gleichen Position im Bild, wodurch sie werden fälschlicherweise als markante Punkte erkannt und in die Bildsignatur aufgenommen werden. Je nach Anzahl der Einschlüsse und der Zahl an absolut extrahierten Punkten kann dieser Fehler zur Verfälschung der Bildsignatur beitragen (vgl. DATTA ET AL., 2008, pp. 22-23).

Auch bei dieser Methode ist ein Wandel von der Analyse globaler Merkmale hin zu regionalen beziehungsweise lokalen Merkmalen erkennbar, um eine Verfeinerung der Bildsignatur zu erreichen. Dabei werden markante Punkte lokal zusammengefasst und Teilsignaturen erstellt (vgl. BOUCHARD & TRIGGS, 2005, pp. 711-714).

2.4.1.5 Segmentierung

Segmentierung ist die Grundvoraussetzung bei der Erstellung regionsbasierter Signaturen. Die Bildinformation wird anhand eines vorher festgelegten Kriteriums in zusammengehörige Regionen aufgeteilt. Neben vielen anderen Kriterien zur Einteilung können Pixelwerte, Farbinformationen, Helligkeitswerte oder auch Texturmerkmale zur Aufteilung herangezogen werden (vgl. DATTA ET AL., 2008, p. 19).

Durch die Aufteilung der Bildinformation in beliebig viele Regionen ist es möglich Teilsignaturen für das Bild zu erstellen, die abhängig von ihrer räumlichen Anordnung im Bild betrachtet werden können. Erwähnte Probleme wie die Erkennung gleicher Farbmerkmale trotz unterschiedlichem Bildinhalt oder das 'scaling problem' bei Texturmerkmalen, können dadurch umgangen werden. Wird ein Bildmerkmal einem Segment zugeordnet, kann es aber zudem auch unabhängig von seiner räumlichen Anordnung mit einem Segment eines anderen Bildes verglichen werden. Dadurch können ähnliche Bildmerkmale unabhängig von ihrer Anordnung im Bild auf ihre Ähnlichkeit überprüft werden (vgl. SMEULDERS ET AL., 2000, pp. 1356-1357).

Ein anerkanntes graphentheoretisches Modell zur Segmentierung ist das von Shi und Malik vorgestellte 'normalized cut' Verfahren (vgl. Abbildung 5). Dabei wird das Bild als ein gewichteter Graph und dessen Pixel als Knoten angesehen. Die Kanten werden anhand der Ähnlichkeit zwischen den Pixeln in einem festgelegten Merkmal gewichtet. Der dem rekursiven Verfahren zugrunde liegende Algorithmus partitioniert den Graphen solange, bis ein festgelegtes Partitionslimit erreicht wurde (vgl. J. SHI & MALIK, 2000, pp. 891-894).

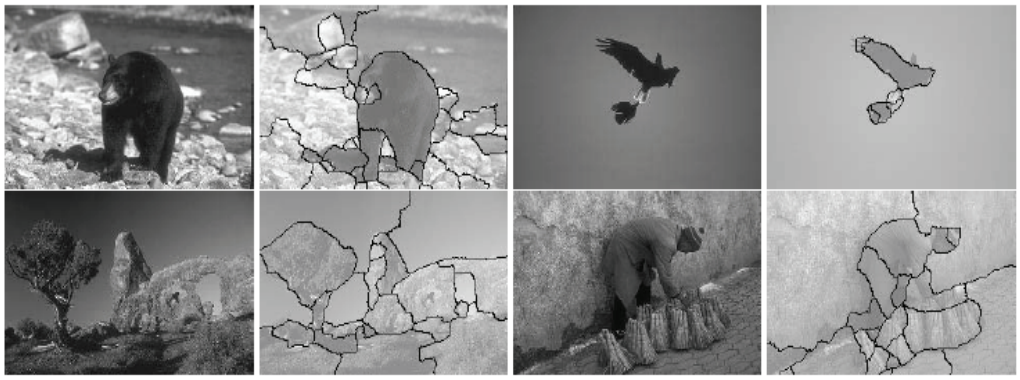

Abbildung 5: Beispiel für die Segmentierung durch ein auf dem 'normalized cut' Algorithmus beruhendes System (Ren und Malik 2003, p.17)

Weitere Verfahren zur Segmentierung eines Bildes setzen den Clustering-Algorithmen ein, die im anschließenden Kapitel besprochenen werden (vgl. Kapitel 2.6.1).

Ein Kernproblem der Segmentierung stellt die richtige Wahl der Ausschnitte dar. Es ist oft nicht möglich komplexe Strukturen oder unscharfe Übergänge klar abzugrenzen. Die Generierung von zu vielen Teilsegmenten führt zur Trennung von ähnlichen Bildbereichen. Zu wenige Regionen besitzen im Gegensatz dazu zu wenig diskriminative Kraft. Formmerkmale sind besonders abhängig von guten Segmentierungsmethoden. Werden ähnliche Formmerkmale durch falsche Segmentierung getrennt, verlieren sie ihre Aussagekraft. Es gilt abzuwägen in welchem Ausmaß Segmentierung von Nutzen für Bildretrievalsysteme ist – die dadurch gewonnenen regionalen Bildmerkmale tragen jedenfalls entscheidend zum besseren Verständnis von Bildinhalten bei (vgl. DATTA ET AL., 2008, pp. 19-20).

2.4.2 Signaturerstellung

Die Möglichkeiten zur Extraktion von Bildmerkmalen sind vielfältig und variieren je nach Anwendungsgebiet und Entwurfsschwerpunkt. Die bloße Extraktion der verschiedenen Merkmale eines Bildes erlaubt jedoch noch keinen Vergleich mit anderen Bildern. Um Bilder vergleichbar zu machen, ist es notwendig eine Merkmalssignatur zu erstellen. Grundsätzlich lassen sich Methoden zur Erstellung von Signaturen in vektorbasierte Modelle und Distributionsmodelle unterscheiden. Signatur-Histogramme und regionsbasierte Signaturen zählen zu den am häufigsten verwendeten Modellen und gehören zur Gruppe der gewichteten Vektormodelle (vgl. DATTA ET AL., 2008, pp. 23-25).

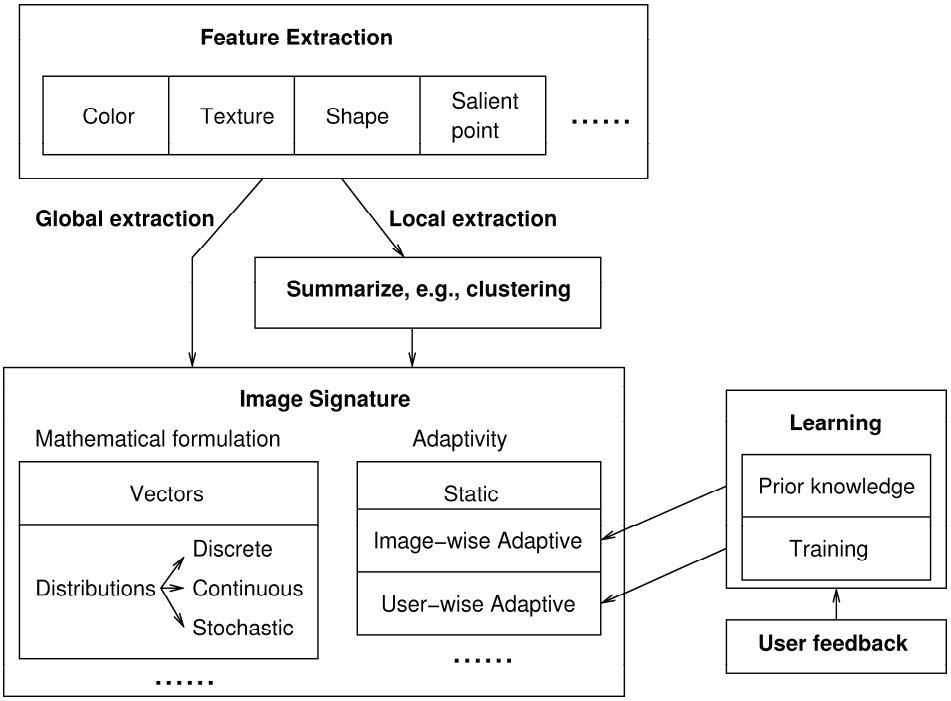

Abbildung 6: Übersicht zur Generierung von Bildsignaturen (DATTA, 2008, p.17)

Der Umstand, dass verschiedene Methoden zur Erstellung einer Signatur möglich sind, ist nicht zuletzt auch ein Grund für die Vielfalt der verschiedenen Extraktionsmethoden. Die Art der Signatur gibt teilweise auch auch die Art der Extraktionsmethoden vor. beeinflusst. Das heißt, die Erstellung einer vektorbasierten Signatur erfordert vektorbasierte Extraktionsmethoden und und der Einsatz einer Signatur nach dem Distributionsmodell verlangt nach dem Einsatz von distributionsbasierten Extraktionsmethoden. Zudem besteht die Möglichkeit gesonderte Signaturen für die verschiedenen Merkmalsausprägungen zu erstellen und ein Bild somit mit mehreren Signaturen zu verknüpfen. Abbildung 6 zeigt verschiedene Wege zur Erstellung von Bildsignaturen in Abhängigkeit von den extrahierten Bildmerkmalen.

Das Spektrum der Methoden zur Erstellung von Bildsignaturen ist sehr vielfältig, daher werden in diesem Kapitel nur einige ausgewählte Methoden vorgestellt.

2.4.2.1 Allgemeiner Stand der Forschung

Eine oft eingesetzte Methode zur Generierung einer Bildsignatur sind Histogramme. Um standardisierte Merkmalshistogramme zu erhalten, teilt man Merkmalshistogramme in eine beliebige Anzahl von Teilabschnitten, so genannte

'bins', und behandelt die Teilabschnitte als Merkmalsvektoren. Daraus lässt sich ein mehrdimensionaler Merkmalsvektor erstellen, der als Signatur dient.

Eine Weiterentwicklung dieses Vorgehens stellen 'multi-resolution histograms' dar, deren Besonderheit darin besteht, dass mehrere Histogramme mit verschieden großen Teilabschnitten als Bildsignatur verwendet werden. Durch die Variation der Durchschnittsgröße der 'bins' werden die Merkmalsvektoren mit variablem Auflösungsgrad in die Signatur aufgenommen. So ist es möglich einen variablen Detailgrad bei der Suche nach ähnlichen Bildern zu gewährleisten (vgl. HADJIDEMETRIOU ET AL., 2004, pp. 835-839).

Bei der Erstellung von Bildsignaturen von einem Wandel von globalen Bildsignaturen hin zu lokalen Signaturen sprechen. Dabei liegt der Vorteil von lokalen Bildsignaturen in einem besseren Detailgrad und einer daraus resultierenden Verfeinerung der Suchergebnisse. Durch diese Verfeinerung besteht die Gefahr die Treffermenge zu sehr einzuschränken, da es je nach Anwendungsfall nötig ist, den Detailgrad der lokalen Bildsignaturen anzupassen. Hybride Ansätze, die sowohl globale als auch lokale Signaturen verwenden, liefern hierbei gute Ergebnisse. Gerade bei Echtzeitsystemen haben sie jedoch den Nachteil, dass sie durch ihre Komplexität zu viel Rechenleistung benötigen, um ein Suchergebnis in Echtzeit zu gewährleisten. Aus diesem Grund muss für jede Anwendung eine passende Gewichtung für die hybriden Signaturen gefunden werden (vgl. J. LI & J. Z. WANG, 2006, pp. 914-915).

In Kombination mit modernen Methoden des maschinellen Lernens gibt es neue Ansätze, die sich globaler Bildsignaturen bedienen. Durch vorhergehendes Trainieren der Vergleichsalgorithmen sind die globalen Bildsignaturen hinreichend detailliert, um die Bildquellen vergleichbar zu machen.

Die Segmentierung mittels regionsbasierter Signaturen rückt mehr und mehr in den Fokus, da sie sich als sehr wirkungsvoll zur Identifikation beziehungsweise Eingrenzung von Objekten im Bild herausstellt. Eine exakte und fehlerfreie Segmentierung ist jedoch auch mit dieser Methode nicht realisierbar. Hierbei gilt es, wie bei der Segmentierung anhand von Einzelmerkmalen, ein Gleichgewicht zwischen der Genauigkeit der Segmentierung und dem dazu benötigten Aufwand zu finden (vgl. DATTA ET AL., 2008, pp. 23-25).

2.4.2.2 Earth mover's distance

Bei der Darstellung der Merkmale durch Histogramme tritt ein für diese Darstellungsmethode übliches Problem auf: Die räumliche Anordnung eines Merkmals

im Bild kann nicht abgebildet werden. Die räumliche Information der extrahierten Merkmale eines Bildes geht also verloren. Der Einsatz der 'earth mover's distance' (EMD) kann dieses Problem durch die Berücksichtigung der räumlichen Anordnung der Merkmale beheben. Zusätzlich zu den Merkmalsvektoren wird deren räumliche Anordnung zueinander in das Distanzmaß miteinbezogen. Die EMD hat ferner den Vorteil, dass durch die Koppelung mit einer Clustering-Methode Features die räumlich nahe zusammenliegen, zu Gruppen aggregiert werden können. Aufgrund dieser Eigenschaft wird die Signaturerstellung mittels der EMD als regionsbasierte Methode bezeichnet. Durch die Miteinbeziehung der räumlichen Position verschiedener Merkmalsgruppen ist es – ähnlich zur Formerkennung – möglich, Objekte im Bild zu identifizieren, die eine semantische Einordnung des Bildinhalts ermöglichen (vgl. RUBNER ET AL., 2000, pp. 104-107).

2.4.2.3 Adaptive Bildsignaturen

Beim Bildretrieval in großen, inhomogenen Bilddatenbanken können sehr viele verschiedene Kategorien von Bildern auftreten. Neben Farbfotografien können Grafiken, Schwarz-Weiß Fotografien oder eingescannte Dokumente in der Datenbank enthalten sein. Um die Bilder untereinander vergleichbar zu machen, müssen die Bildmerkmale mit den gleichen Methoden extrahiert werden. Die Aussagekraft der einzelnen Merkmale ist jedoch nicht grundsätzlich gleich. Die Farbmerkmale einer Schwarz-Weiß Aufnahme besitzen im Vergleich zu denen einer Grafik sehr wenig Aussagekraft. Das heißt, sie machen die S/W Aufnahme nicht unterscheidbar von anderen S/W Aufnahmen, die Grafik hingegen kann aufgrund ihrer Farbmerkmale sehr gut von anderen Grafiken unterschieden werden. Aus diesem Grund ist es sehr oft sinnvoll die Bildsignaturen an die Bildkategorien anzupassen und so genannte adaptive Bildsignaturen zu verwenden. Die Einteilung der Bilder in verschiedene Kategorien kann einerseits durch vorherige Benutzereingabe und andererseits durch automatische Einteilung anhand der extrahierten Bildmerkmale erfolgen (vgl. DATTA ET AL., 2008, pp. 25-26).

Eine weitere Variante stellt die Anpassung der Signatur anhand von Benutzerpräferenzen dar. Durch das Auslesen von Benutzerprotokollen oder die individuelle Angabe von Suchpräferenzen werden die Bildmerkmale unterschiedlich gewichtet und die Signatur anschließend erstellt (vgl. X. WANG ET AL., 2004).

Ein ähnliches Verfahren wird angewandt, wenn eine große Zahl von Bildmerkmalen pro Bild extrahiert wird, da dies dazu führen kann, dass die Bildsignaturen ihre Aussagekraft verlieren. Durch eine Änderung der Merkmalsgewichtung und die Zusammenstellung von 'sub-sets' kann die Retrievalqualität nachträglich

gesteigert werden. Bei der Erstellung von ‘sub-sets‘ werden nur bestimmte Teile der Bildmerkmale in die Signatur aufgenommen (vgl. WESTON ET AL., 2000, pp. 669-672).

2.5 Signaturen vergleichen

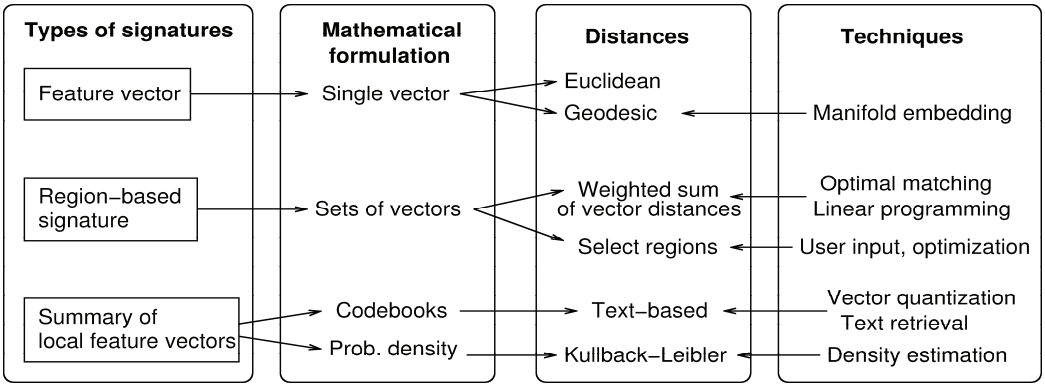

Abbildung 7: Ähnlichkeitsmaße abhängig von der Bildsignatur (DATTA, 2008, p .25)

Durch die Analyse der visuellen Merkmale von Bildern und der daraus gebildeten Signaturen sind die Bilder hinsichtlich ihres Inhalts vergleichbar. Es ist jedoch nötig geeignete Vergleichsmethoden zu finden, die den Grad der Ähnlichkeit oder der Unähnlichkeit von Bildern angemessen darstellen können. Analog zum Textretrieval muss eine geeignete Trennschärfe zwischen relevanten und irrelevanten Bilddokumenten gefunden werden um den Benutzer eines Retrievalsystems akkurate Ergebnisse zu liefern. Abbildung 7 zeigt die verschiedenen Signaturvarianten inklusive der benötigten Erstellungstechniken, Distanzmaße und mathematischen Grundlagen.

Nach Datta et al. sind dabei folgende allgemeine Kriterien ausschlaggebend für die Wahl des passenden Ähnlichkeitsmaßes (vgl. DATTA ET AL., 2008, pp. 26-27):

(1) Vereinbarkeit mit semantischen Kriterien

(2) Robustheit gegenüber Störfaktoren

(3) benötigter Rechenaufwand

(4) Invarianz gegenüber dem Hintergrund (für regionsbasierte Suchanfragen)

(5) lokale Linearität (nach der Dreiecksungleichung bei benachbarten Bildregionen)

Wie bei allen vorangegangenen Teilbereichen des CBIR gibt es auch hier eine Vielfalt verschiedener Entwürfe zur Feststellung der Ähnlichkeit von Bildern. Die

verschiedenen Designansätze werden von Datta (2008) anhand der folgenden Techniken in verschiedene Bereiche unterschieden:

(1) Betrachtung der Features als Vektoren, Non-Vektor Repräsentationen oder Zusammenstellungen

(2) Ähnlichkeitsermittlung mittels regions-basierter, globaler oder hybrider Methoden

(3) Berechnung der Ähnlichkeit im Vektorraum oder durch nichtlineare Verteilung

(4) Rolle der Bildsegmente bei der Ähnlichkeitsberechnung

(5) Stochastische, unscharfe (fuzzy) oder deterministische Ähnlichkeitsmaße

(6) Einsatz und Art von maschinellem Lernen

Die Wahl des Ähnlichkeitsmaßes ist dabei in großem Maße abhängig vom Einsatzgebiet des Retrievalsystems und den verwendeten Algorithmen zur Merkmalsextraktion, beziehungsweise zur Signaturerstellung. Aus diesem Grund findet eine Vielzahl von Distanzmaßen bei der inhaltsbasierten Bildsuche ihre Verwendung. Die populärsten Distanzmaße werden in Abbildung 8 im Detail beschrieben und ihre Vor- und Nachteile angeführt.

Distance Measure	Input	Computation	Complexity	Metric	Comments
Euclidean (L^2norm)	$\vec{X}_a, \vec{X}_b \in \mathbb{R}^n$ (vectors)	$\vec{X}_a \cdot \vec{X}_b$	$\Theta(n)$	Yes	Popular, fast, L^1 also used
Weighted Euclidean	$\vec{X}_a, \vec{X}_b \in \mathbb{R}^n$ $W \in \mathbb{R}^n$ (vec. + wts.)	$\vec{X}_a^T[W]\vec{X}_b$ $[\cdot] \leftarrow$ diagonalize	$\Theta(n)$	Yes	Allows features to be weighted
Hausdorff	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$	See Eqn. 2	$\Theta(pqn)$ $(d(\cdot,\cdot) \leftarrow L^2$ norm)	Yes	Sets corr. to image segments
Mallows	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$ Signific.: S	See Eqn. 3	$\Theta(pqn)$ + variable part	Yes	The EMD is its special case
IRM	Vector sets: $\{\vec{X}_a^{(1)}, .., \vec{X}_a^{(p)}\}$ $\{\vec{X}_b^(1), .., \vec{X}_b^{(q)}\}$ Signific.: S	See Eqn. 3	$\Theta(pqn)$ + variable part	No	Much faster than Mallows computation in practise
K-L divergence	$\vec{F}, \vec{G} \in \mathbb{R}^m$ (histograms)	$\sum_x F(x) \log \frac{F(x)}{G(x)}$	$\Theta(m)$	No	Asymmetric, compares distributions

Abbildung 8: Distanzmaße beim inhaltsbasierten Bildretrieval (DATTA, 2008, p.29)

Regionsbasierte Vergleichsmaße stellen auch in diesem Bereich aufgrund detaillierterer Retrievalergebnisse den derzeitigen Trend dar. Die gängigen Probleme der regionsbasierten Ansätze treten jedoch auch hier auf. Eine angemessene Eingrenzung bestimmter Bildregionen stellt sich immer noch sehr schwierig dar (vgl. DATTA ET AL., 2008, pp. 32-33). Um die Berechnung falscher Segmente abzuschwächen, wird häufig das so genannte 'soft matching' eingesetzt Dabei werden die Abstände zwischen den regionalen Bildsignaturen um einen bestimmten Faktor verallgemeinert um eine größere Menge von Übereinstimmungen zu generieren. Weiterhin muss ein aussagekräftiges Abstandsmaß zwischen den extrahierten Regionen gefunden werden, da die Distanz von mehreren Bildmerkmalen zueinander berechnet werden muss. Die Mittelung der Einzelabstände zwischen den Merkmalspaaren der Bildregionen wird durch den Einsatz der 'earth mover's distance' oder ähnlicher Verfahren wie der IRM-Distanz (vgl. J. LI ET AL., 2000, pp. 4-5) erreicht.

2.6 Zusammengesetzte Bildmerkmale

Die Ermittlung von Einzelmerkmalen aus Bildern ist der Kernbereich des modernen Image Retrieval. Sie bilden die Basis zur Erstellung und Weiterentwicklung effektiver Bildretrievalsysteme. Trotz des rasanten Fortschritts im Bildretrieval der letzten Jahre, reicht die Extraktion von Merkmalen häufig nicht aus, um zufriedenstellende Ergebnisse bei der Bildsuche zu erzielen. Bei komplexen Bildinhalten können keine hinreichend klaren Einzelmerkmale abgeleitet werden, um ähnliche Bilder zu identifizieren (vgl. SMEULDERS ET AL., 2000, pp. 1357-1360). Häufig ist die Suchintention benutzerseitig nicht vordergründig auf spezielle Merkmale im Bild ausgerichtet. Eakins et al. zeigen in Abbildung 9 den Einfluss ver-

	Whole population			Those actually using feature in searches		
	N	Mean	S. D.	N	Mean	S. D.
Semantic (specific)	123	6.16	1.32	104	6.39	1.03
Semantic (general)	121	5.69	1.58	104	5.89	1.50
Sharpness	122	5.20	1.93	55	6.15	1.05
Cultural abstraction	123	5.14	1.76	45	5.84	1.17
Technical abstraction	122	4.60	1.94	45	5.98	1.06
Metadata	123	4.26	1.99	56	5.11	1.67
Contextual abstraction	122	4.21	1.89	38	5.47	1.58
Colour	123	3.79	2.24	30	5.59	1.62
Shape	123	3.67	1.97	22	5.05	1.43
Texture	123	3.66	2.08	14	5.46	1.66
Visual relationships	122	3.57	1.89	23	4.96	1.46
Visual extension	120	3.56	1.87	23	4.95	1.50
Emotional abstraction	122	3.04	1.87	12	5.27	1.56

Abbildung 9: Einfluss von Bildmerkmalen auf die Auswahl eines Bildes durch Benutzer (EAKINS ET AL., 2004, p. 632)

schiedener Bildmerkmale auf die Auswahl eines Bildes durch den Benutzer (vgl. Abbildung 9). Dabei sind die Farb-, Form- und Strukturmerkmale eher nebensächlich. Die hoch bewerteten semantisch übergeordneten Konzepte können nicht ohne großen Auswand aus separaten Farb-, Form- oder Strukturmerkmalen abgeleitet werden. Das 'semantic gap' kann also nicht mit den isolierten Einzelmerkmalen überwunden werden. Bereits Rui erkennt:

> „In constrained applications, such as the human face and finger print, it is possible to link the low-level features to high-level concepts (faces or finger prints). In a general setting, however, the low-level fea-

> tures do not have a direct link to the high-level concepts." (RUI & HUANG, 1999, p. 51)

Um diese Lücke zu verkleinern, müssen also Konzepte zur Weiterverarbeitung von 'low-level features' entwickelt werden. In diesem Zusammenhang sind vor allem Ansätze zur Gruppierung von Bildmerkmalen zu nennen, aber auch der Einsatz von speziellem Relevance Feedback kann das Suchergebnis verbessern.

2.6.1 Clustering von Bildmerkmalen

Die aus Bildern extrahierten Einzelmerkmale können auf verschiedene Arten gruppiert werden. Beim Clustering ist es dabei nicht erforderlich vor der Gruppierung der Einzelelemente bestimmte Klassen oder Kategorien zu definieren.

> „Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster." (A. K. JAIN ET AL., 1999, p. 265)

Aus einer undefinierten Menge von Objekten werden durch Clusteranalyse Gruppen von ähnlichen Objekten ermittelt und zusammengefasst. Bezogen auf die Bildmerkmale können so beispielsweise Klassen von ähnlichen Farb- oder Strukturmerkmalen in Gruppen eingeteilt werden. Abhängig vom Format der ermittelten Einzelmerkmale und dem angestrebten Gruppierungsziel existieren sehr viele verschiedene Methoden zum Clustering. Sie lassen sich grob in zwei Arten unterscheiden: hierarchische und heuristische Verfahren. Je nach Vorgehensweise werden sie zudem als agglomerative oder teilende Verfahren bezeichnet (vgl. MANNING ET AL., 2008, pp. 377-402).

Hierarchische Verfahren betrachten die Gesamtmenge aller Merkmale und fassen die ähnlichsten Objekte solange zu Clustern zusammen oder teilen die Gesamtmenge von Objekten solange auf, bis ein vorher bestimmtes Stoppkriterium erfüllt ist. Da sie stets die Gesamtmenge aller Merkmale betrachten sind sie im Vergleich zu heuristischen Verfahren speicherintensiver. Beim CBIR ist die hierarchisch agglomerative Methode die am meisten verwendete hierarchische Methode (vgl. DATTA ET AL., 2008, p. 33).

Heuristische Verfahren bestimmen eine beliebige Menge von Einzelmerkmalen zufällig als Clusterzentren und teilen die restlichen Objekte abhängig von ihrer Nähe zum Zentrum den Clustern zu. Daraufhin werden die Clusterzentren

neu berechnet und die Objekte abermals zugeteilt. Iterativ wird solange damit fortgefahren, bis keine Änderung mehr eintritt. Der populärste Ansatz beim partitionierenden Clustering ist das so genannte 'k-means' Verfahren. Die Menge der Cluster k wird durch einen Schwellenwert bestimmt, der den maximalen Abstand eines Einzelmerkmals zum Clusterzentrum festlegt. 'k-means' ist aus Sicht der Bildverarbeitung ein sehr robustes und primitives Verfahren. Komplexere Gruppierungen sind mit dieser Variante des Clustering nur schwer zu erreichen. Eine Verbesserung bietet das so genannte 'expectation-maximization' Verfahren, ein auf dem 'k-means' Ansatz basierendes statistisches Verfahren. Es gruppiert die Einzelmerkmale durch Schätzung ('expectation') der Wahrscheinlichkeiten, ob ein Merkmal zu einer vorgegebenen Anzahl von Clustern zugeordnet werden kann. Daraufhin werden die Clustermittelpunkte neu berechnet und das Verfahren schrittweise solange wiederholt bis jedes Objekt mit einer hohen Wahrscheinlichkeit dem richtigen Cluster zugeteilt wurde ('maximization'). Den Abbruch der Iterationen legt dabei entweder ein Schwellenwert bei den Wahrscheinlichkeitsberechnungen oder ein maximales Limit bei den Wiederholungen fest (vgl. A. K. JAIN ET AL., 1999, pp. 274-296).

Neben diesen populären Beispielen existieren noch unzählige Alternativen und Varianten zur Gruppierung mittels Clustering, die Hastie et al. in einen umfassenden Überblick zu Clusteringverfahren und Data Mining in seinem Buch 'The Elements of Statistical Learning' beschreiben(vgl. HASTIE ET AL., 2003, pp. 453-479).

Clustering kann gerade bei großen unstrukturierten Bilddatenbanken seine Stärken ausspielen, da keine Vorkenntnisse über die vorliegenden Daten bestehen müssen. Bei beiden Varianten existieren aber auch einige Nachteile. Die Anzahl der Cluster muss beim heuristischen Ansatz beispielsweise vor Beginn der Berechnungen festgelegt werden. Bei der Wahl einer unangemessenen Clustermenge kann das zu falschen Einteilungen führen. Die Cluster können dadurch zu allgemein oder zu spezifisch sein. Beim hierarchischen Clustering tritt bei der Wahl eines unangebrachten Schwellenwertes dasselbe Problem auf. Es gilt daher immer abzuwägen, ob die Größe beziehungsweise die Anzahl der Cluster repräsentativ für die Menge der Einzelmerkmale ist (vgl. DATTA ET AL., 2008, pp. 33-37).

2.6.2 Klassifikation

Unter einer Klassifikation wird die systematische Ordnung von Merkmalen zu Klassen verstanden. Dabei unterscheidet man zwischen diskriminativen und generativen Modellen zur Klassifikation. Entgegen dem Clustering werden bei der

Klassifikation die Merkmale einer oder mehreren bereits bestehenden Klassen zugeordnet (vgl. DUDA ET AL., 2000, p. 12f).

Als klassifizierende Merkmalsausprägungen beim Bildretrieval können beispielsweise bestimmte Farb-, Form- oder Strukturvarianten dienen. Häufig geht Klassifizieren mit Methoden aus dem maschinellen Lernen einher. Die eingesetzten Klassifikationsmethoden werden vor ihrer Anwendung auf die extrahierten Bildmerkmale mit vorverarbeiteten Daten trainiert. So werden beispielsweise für jede gewünschte Klasse bestimmte Trainingssets angelegt, anhand derer der Algorithmus bereits ein Repertoire an klassifizierenden Merkmalen lernt. Aus diesem Grund ist Klassifizierung besonders effektiv, wenn die Bildmerkmale bereits erlernten Kategorien einzuordnen sind (vgl. DATTA ET AL., 2008, pp. 36-37).

Gängige Modelle zur diskriminativen Klassifikation von Bildermerkmalen im CBIR sind die 'Support Vector Machines' – kurz SVM – (vgl. PANDA & E. Y. CHANG, 2006, p. 318ff; S. TONG & E. CHANG, 2001, p. 108ff) und Entscheidungsbäume (vgl. MACARTHUR ET AL., 2000, p. 2f). Die populärsten generativen Modelle basieren auf dem Einsatz des Satzes von Bayes (vgl. VAILAYA ET AL., 2001, p. 120f).

Mit der wachsenden Größe der Bilddatenbanken werden Gruppierungsverfahren beim Bildretrieval immer wichtiger, um eine übersichtliche und strukturierte Ergebnismenge zu erhalten. Klassifikation kann dazu einen entscheidenden Beitrag leisten. Ein begrenzendes Kriterium für die Qualität eines Klassifikationsverfahrens beim CBIR sind dabei meist die Trainingsdaten. Um effektives maschinelles Lernen in diesem Bereich anzuwenden sind viele, mit zusätzlicher Information aufbereitete Daten nötig. Diese Zusatzinformation wird manuell von einem Experten hinzugefügt und wird als 'ground truth' bezeichnet. Diese subjektive Einteilung ist zugleich der größte Kritikpunkt beim Klassifizieren. Die manuell zugeteilten Klassen basieren oft auf zu komplexen semantischen Konzepten und können deshalb nicht sinnvoll mit den Bildmerkmalen verknüpft werden (vgl. MÜLLER ET AL., 2002, pp. 41-46).

2.7 Relevance Feedback

Die Effektivität eines Bildretrievalsystems ist neben den Verfahren zur Extraktion und Verarbeitung von Bildinformation zu einem großen Teil abhängig vom Grad der Benutzerinteraktion. Wie bereits im Kapitel zu den Anfragemöglichkeiten behandelt wird diese Form der Interaktion als Relevance Feedback (RF) bezeichnet und dient zur Verbesserung der Benutzeranfragen (vgl. Kapitel 2.2). Relevance Feedback kann nicht nur zur Verbesserung der Suchanfrage eingesetzt werden, sondern dient ebenso als Werkzeug zur Verbesserung der Bildanalyse oder Er-

gebnisausgabe. Aufgrund seiner zentralen Stellung bei modernen Bildanfragesystemen und seiner vielseitigen Einsatzmöglichkeiten, sollen in diesem Kapitel nochmals einige spezielle Varianten des Relevance Feedback erörtert werden.

Zur Verbesserung von Bildanalysemethoden und der Ähnlichkeitsmaße wird Relevance Feedback mit Methoden zum maschinellen Lernen kombiniert. Die Auswahl richtiger Ergebnisse aus der Ergebnismenge durch den Benutzer wird für das Training der Analysemethoden herangezogen und kann als aktiver Lernprozess betrachtet werden. Bei jeder Suchanfrage lernt das System seine Analysemethoden und Vergleichsmaße an die Benutzerbedürfnisse anzupassen (vgl. J. HE ET AL., 2004, pp. 15-18).

Klassisches Relevance Feedback bietet dem Benutzer eine Ergebnismenge an, aus der er richtige und falsche Ergebnisse auswählen kann, um die Suchanfrage zu verfeinern. Durch mehrmaliges Wiederholen dieser Prozedur erhält er nach einigen Schritten eine Ergebnismenge die seinen Vorstellungen entspricht. Häufiges Wiederholen birgt jedoch die Gefahr, dass die Geduld des Benutzers strapaziert wird. Alternative Feedbackmethoden versuchen dieses Problem zu umgehen. Ein Bespiel hierfür sind semantische Kategorisierungsmethoden, die versuchen die Anfrage des Benutzers einer semantischen Klasse zuzuordnen bevor der Retrievalprozess gestartet wird (vgl. YANG ET AL., 2005, pp. 416-417). Einen anderen Ansatz verfolgt ein Verfahren zur Verbesserung des Feedback, bei dem die Anzahl der Durchgänge verringert wird, indem die vorhergehenden Benutzeranfragen aufgezeichnet und in die Verfeinerung der ersten Anfrage miteinbezogen werden (vgl. HOI & LYU, 2004, pp. 25-27).

Die wachsende Relevanz von regionalen Bildsignaturen hat auch Auswirkung auf Relevance Feedback Methoden. Jing liefert beispielsweise Ansätze zur Einbeziehung der regionsbasierten Bildmerkmale von positiv gekennzeichneten Ergebnissen (vgl. JING, 2002).

Relevance Feedback nimmt beim Text- und Multimediaretrieval eine Schlüsselposition ein. Richtig eingesetzt kann es Retrievalprozesse effektiv und zuverlässig machen. Andererseits dürfen die eingesetzten Methoden den Benutzer keinesfalls irritieren. Das Gleichgewicht zwischen kognitiver Beanspruchung des Nutzers und einer Effizienzsteigerung lässt sich nur schwer einhalten. Aus diesem Grund ist die direkte Einbeziehung des Anwenders bei den derzeitig eingesetzten Bildretrievalsystemen meist nur sehr gering. Einen guten Überblick über das Feld des RF und Strategien zum Umgang mit diesem speziellen Problem bieten Zhou

und Huang in ihrem Artikel ‘Relevance feedback in image retrieval: A comprehensive review’ (vgl. ZHOU & HUANG, 2003, pp. 538-539).

3 Automatische Annotation

Inhaltsbasiertes Bildretrieval hat seine Ursprünge in Disziplinen wie der Bildverarbeitung, der Mustererkennung, dem maschinellen Sehen und des Information Retrieval. Nachdem sich CBIR über 20 Jahre in der Forschungslandschaft etabliert hat, ist es heute ein etablierter selbstständiger Forschungszweig. Die ursprüngliche Hauptintention, Bilder lediglich anhand ihrer inhaltlichen Bestandteile zu identifizieren und zu vergleichen ist nicht mehr der einzige Bereich mit dem man sich beschäftigt. Analog zu anderen Disziplinen entwickeln sich Fragestellungen, die ihren Schwerpunkt auf anderen Bereichen ansiedeln. Hierzu zählen Themen wie der Einsatz von CBIR zur persönlichen Authentifizierung, zur Überprüfung von Urheberrechtsverletzungen oder zur Einordnung von Bildern anhand ästhetischer Kriterien (vgl. DATTA ET AL., 2008, pp. 45-50).

Ein weiterer vielversprechender Zweig ist die Anreicherung von Bildmaterial mit Metainformation. Mit Metainformation versehene Bilder werden bereits für das Training von Algorithmen auf dem Bereich des maschinellen Lernens im CBIR eingesetzt. Hinreichend annotierte Bilddatenbanken sind zwar vorhanden, die Menge und die Annotationsqualität der beinhalteten Bilder sind jedoch begrenzt. Zudem erfordert die Erstellung solcher Datenbanken einen erheblichen manuellen Aufwand. Betrachtet man proprietäre Bilddatenbanken im Netz bietet sich ein anderes Bild. Teils sind Bilder ohne jegliche Metainformation hinterlegt, teils ist die vorhandene Information fehlerhaft und unzureichend. Die manuelle gemeinschaftliche Annotation durch die Benutzer – so genannte 'Folksonomies' – wird in vielen 'Social Networks' angewandt, aber nicht konsequent genug durchgeführt. Allein die Menge der erstellten Bilder macht eine sorgfältige manuelle Annotation unmöglich. Der Ansatz durch Methoden aus dem Bildretrieval relevante Information automatisch mit Bildern zu verknüpfen liegt daher nahe.

3.1 Grundlagen der automatischen Annotation

Grundsätzlich lassen sich Verfahren zur automatischen Annotation in zwei Bereiche aufteilen. Die gängigsten Systeme ziehen zusätzliches Wissen über Bilder heran, indem sie den damit verbundenen Text analysieren. Sie zählen zum Bereich der textbasierten Annotation. Bildbasierte Annotation dagege,n benutzt den Bildinhalt um Informationen über das Bild zu gewinnen (vgl. DATTA ET AL., 2008, pp. 43-45).

3.1.1 Textbasierte Annotation

Textbasierte Annotation lässt sich in die Unterbereiche der ontologiebasierten und der kontextbasierten Methoden aufteilen. Wie der Name bereits vorgibt, werden bei den ontologiebasierten Methoden verschiedene Ontologien zur Ermittlung von semantisch relevanten Schlagwörtern aus den umgebenden Textbausteinen verwendet.

Kontextbasierte Methoden verwenden die das Bild umgebende Textinformation und extrahieren daraus passende Schlagwörter. Dabei werden neben den Textinformationen auch Metainformationen wie Bildtitel, Autor oder Verlinkungen miteinbezogen und entsprechend gewichtet. Das populärste Beispiel für kontextbasierte Annotation ist die Google Bildsuche (vgl. Google 2009). Textbasierte Annotation besitzt den Nachteil, dass es nur eingesetzt werden kann, wenn Zusatzinformation vorliegt. Aufgrund der geringen Relevanz im Bezug auf den praktischen Teil wird nicht näher auf diese Methode eingegangen (vgl. R. SHI ET AL., 2007).

3.1.2 Bildbasierte Annotation

Erste Versuche zur Verknüpfung von textueller Information mit Bildern durch inhaltsbasierte Analyse wurden bereits Ende der Neunziger gestartet (vgl. Y. MORI ET AL., 1999).

> „The quality of the low-level descriptors used in any CBIR or automatic image annotation system is dominant over any other component. “ (HERVÉ & BOUJEMAA, 2007, p. 173)

Der grundlegende Ablauf von automatischer Annotation hat bis heute Bestand und lässt sich in zwei Kategorien einteilen. Auf der einen Seite Modelle zur Verknüpfung von Bildteilen mit Wörtern, auf der anderen Seite Modelle zur beaufsichtigen Klassifizierung von Bildern. Abbildung 10 zeigt ein Schema zur Verknüpfung von Wort und Bild mittels einer wahrscheinlichkeitsbasierten Methode. Zentraler Bestandteil beider Varianten bleibt die Extraktion der ‘low-level features’ durch Methoden aus dem Content Based Image Retrieval, was ihre ausführliche Erläuterung in dieser Arbeit rechtfertigt.

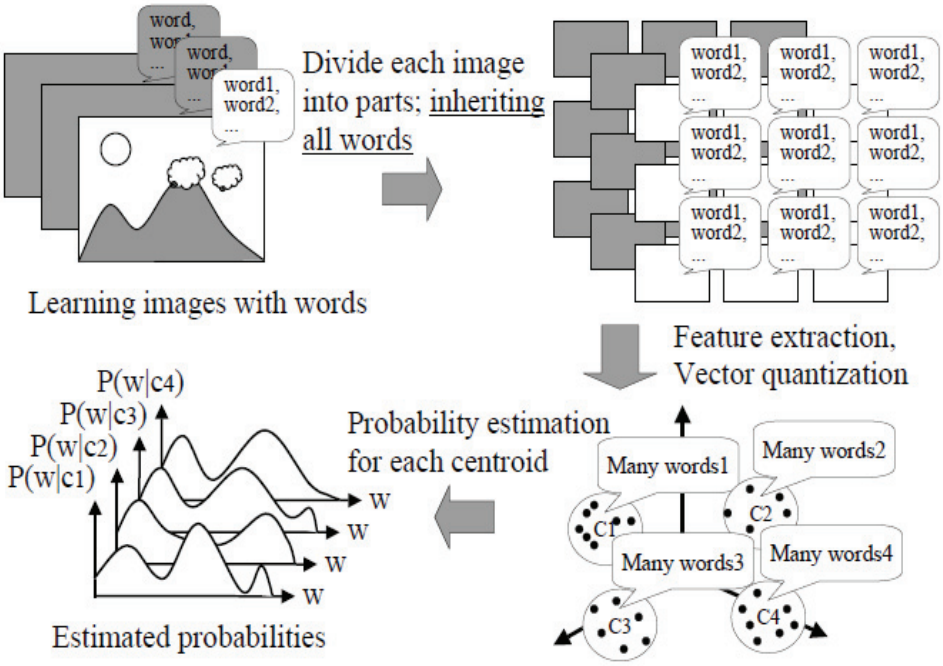

Abbildung 10: Schema zur automatischen Annotation von Bildern (MORI ET AL., 1999, p. 3)

> „The quality of the low-level descriptors used in any CBIR or automatic image annotation system is dominant over any other component."(HERVÉ & BOUJEMAA, 2007, p. 173)

3.1.3 Verknüpfung von Wort und Bild

Die Zuordnung von Textinformation zu Bildern setzt Vorarbeit voraus. Ein Trainingsset mit manuell beschlagworteten Bildern dient zur Vorbereitung eines Retrievalsystems. Bildmerkmale werden für die Bilder aus dem Set extrahiert und mit der textuellen Information verknüpft. Eine Bildsignatur wird erstellt und für das Retrieval abgelegt. Wird eine Suchanfrage mit einem Bild ohne Textinformation gestartet, werden inhaltlich ähnliche Bilder ermittelt und deren textuelle Information mit dem neuen Bild verknüpft. Mit jedem neu annotierten Bild wird gleichzeitig das zugrunde liegende Trainingsset erweitert. Üblicherweise wird das System nicht vollkommen automatisiert betrieben, um mögliche Fehler des Algorithmus zu beschränken. Bei diesem Ansatz wird zwischen unbeaufsichtigtem und beaufsichtigtem Lernen unterschieden (vgl. BARNARD ET AL., 2003, pp. 1109-1111). Ein sensibler Punkt beim Einsatz solcher Verfahren ist die manuelle Beschlagwortung der Trainingsbilder. Die Schlagwörter müssen von Experten zuge-

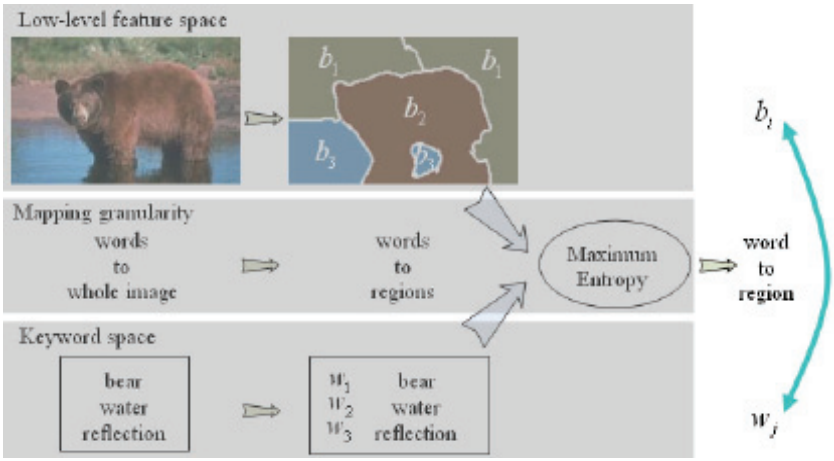

Abbildung 11: Schematischer Ablauf von Wort-Bildmerkmal Verknüpfungen (LI & SUN, 2005, p.40)

teilt werden und sind dadurch zu einem gewissen Grad subjektiv ausgewählt. Ein weiterer Nachteil ergibt sich aus der entstehenden semantischen Lücke zwischen Bildmerkmalen und den von den Experten zugeteilten Wörtern. Um diesen Effekt zu minimieren werden die Trainingsbilder mit Schlagwörtern aus einem kontrollierten Vokabular annotiert. Eine weitere Möglichkeit ist der Einsatz von lexikalischen Datenbanken zur Verbesserung der Annotation. So wird beispielsweise WordNet zur Ermittlung geeigneter Schlagwörter vorgeschlagen (vgl. JIN ET AL., 2005, pp. 706-712).

Typischerweise werden bei der Verknüpfung mit Wörtern die Bilder in Regionen aufgeteilt. In jeder Region werden die Merkmalsvektoren berechnet und die entsprechenden Wörter zugeordnet – diese Teilbereiche werden als 'bag of features' bezeichnet. Erreicht das System eine ausreichend präzise Segmentierung, ist diese Kombination von Schlagwörtern und Bildmerkmalen sehr effektiv. Ein Beispiel für die Segmentierung und die Zuteilung von Schlagwörtern kann Abbildung 11 entnommen werden. Durch den Einsatz heuristischer Methoden wie der Maximum-Likelihood-Schätzung kann die Wahrscheinlichkeit von Wort-Bildmerkmal Kombinationen berechnet und für spätere Vergleiche verwendet werden (vgl. W. LI & SUN, 2005, pp. 39-43).

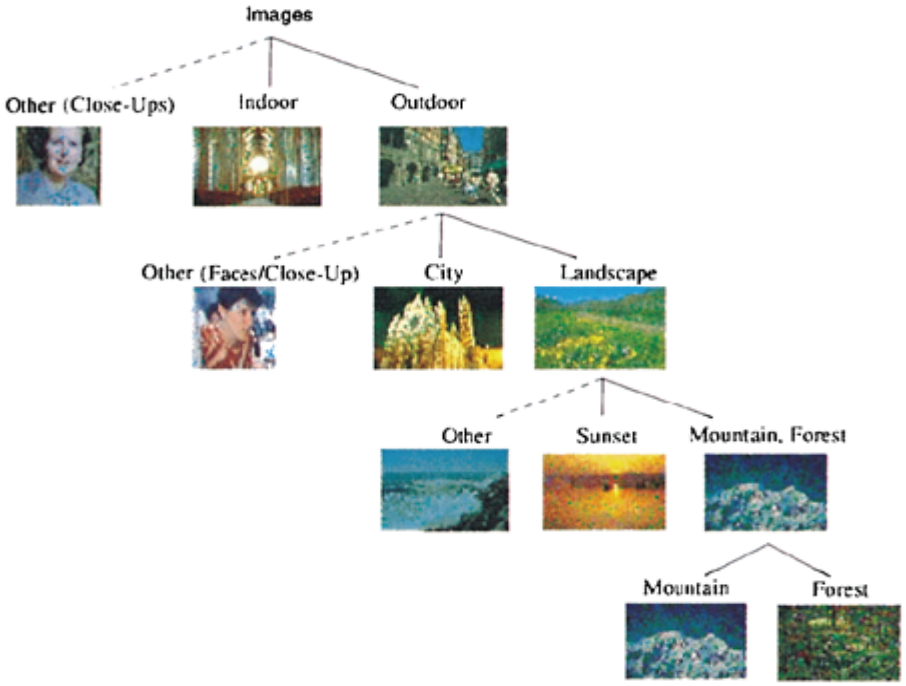

Abbildung 12: Kategorisierung anhand semantischer Konzepte (VAILAYA ET AL., 2001, p. 118)

3.1.4 Kategorisierung

Schon früh wurde versucht Bilder in Kategorien einzuteilen und daraus Rückschlüsse für ihre Annotation zu ziehen (vgl. Abbildung 12). Die Einteilung von Bildern in Farbfotos und Schwarz-Weiß-Bilder kann als einfaches Beispiel einer solchen Einteilung dienen. Die durch eine solche Einteilung gewonnene Information kann zur Beschlagwortung herangezogen werden. Die Zuteilung von Konzepten wie Farbe/Schwarz-weiß, Tag/Nacht oder Außenaufnahme/Innenaufnahme ist dabei ein Mittel zur Erzeugung von semantischer Information (vgl. VAILAYA ET AL., 2001, p. 118).

Anstatt dieses Top-Down-Ansatzes versuchen Li und Wang semantische Information durch die statistische Analyse der Bildmerkmale für die Annotation zu gewinnen. Dabei kommt ein multidimensionales Hidden-Markov-Modell in Kombination mit maschinellem Lernen zum Einsatz. Manuell zusammengestellte Trainingssets für insgesamt 600 semantische Konzepte dienen zum Training ei-

nes statistischen Modells zur Zuweisung von Schlagwörtern. Für jedes Bild eines Sets werden Merkmale für Bildausschnitte in verschiedenen Auflösungen extrahiert und ein statistisches Modell für jedes Konzept erstellt. Die erstellten Konzeptmodelle werden manuell mit passenden Schlagwörtern annotiert und in der Datenbank hinterlegt. Es kann als semantisches Lexikon für das Retrievalsystem betrachtet werden. Will man ein Bild mit dem System annotieren, werden dessen Merkmale ebenfalls extrahiert und mit den abgelegten statistischen Modellen

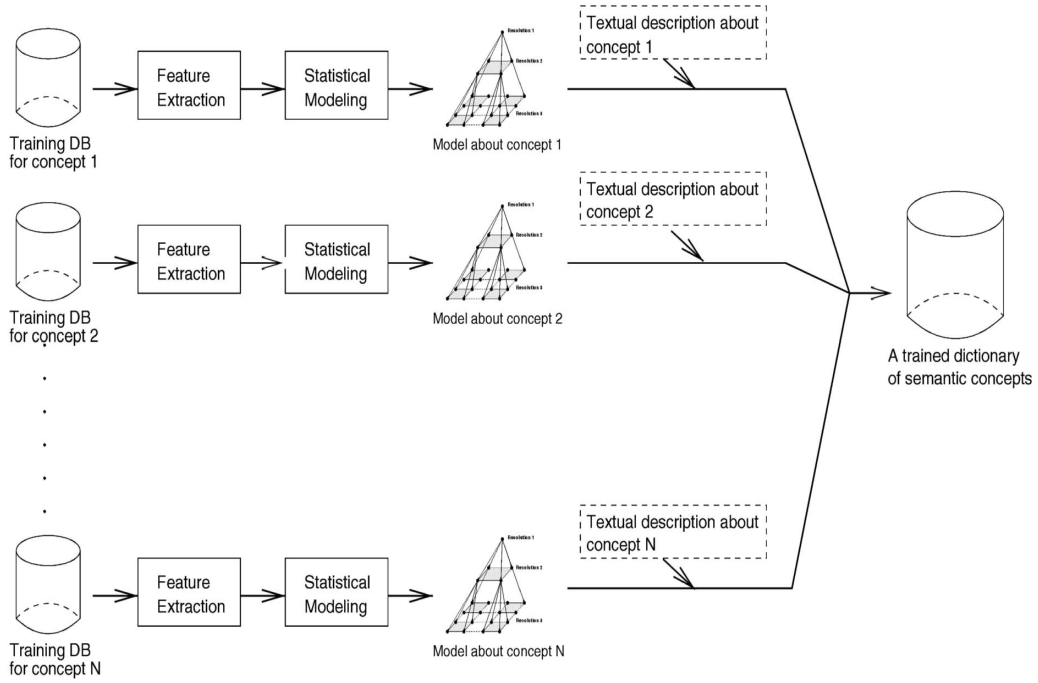

Abbildung 13: Statistische Modellierung semantischer Konzepte anhand ihrer Bildmerkmale (LI & WANG, 2003, p. 1077)

abgeglichen. Die Tags der Modelle mit der größten Ähnlichkeit werden dem Bild dann zugeordnet. Ein Schema dieses findet sich in Abbildung 13 und im dazugehörigen Text (vgl. J. LI & J. Z. WANG, 2003, pp. 1076-1781).

Statistische Modelle in Verbindung mit maschinellem Lernen werden in ähnlichen Arbeiten zur Kategorisierung von Bildern ebenfalls angeführt. Datta kombiniert diese Methoden mit dem semantischen Lexikon von WordNet, um das Schlagwortrepertoire gezielt zu erweitern (vgl. DATTA ET AL., 2007, pp. 982-983).

3.1.5 Diskussion

Automatische Annotation ist eine extrem vielversprechende Teildisziplin des Bildretrieval. Die Möglichkeit der automatischen Beschlagwortung würde eine enorme Erleichterung bei der effizienten Arbeit mit Bildern bedeuten. Aufgrund der vertrauten Anfragemethode wäre die Umstellung für Benutzer geringer als bei rein inhaltsbasierten Systemen. Nichtsdestotrotz ist das Verfahren in der Praxis nur äußerst schwer umzusetzen. Den momentanen Flaschenhals bildet – wie so häufig auf dem Feld des CBIR – die Überbrückung der semantischen Lücke. Dieser Umstand hat mehrere Faktoren.

Die für viele Annotationssysteme erforderliche Segmentierung ist, unabhängig von den angewendeten Segmentierungsverfahren, immer noch zu ungenau. Die Annäherung an die menschliche Wahrnehmung ist auf diesem Bereich noch nicht ausreichend gelungen.

> „We humans segment objects better than machines, having learned to associate over a long period of time, through multiple viewpoints, and literally through a 'streaming video' at all times, which partly accounts for our natural segmentation capability." (DATTA ET AL., 2008, p. 44)

Zudem ist es noch nicht möglich die Annotationssysteme vollkommen unabhängig von Benutzereingaben zu machen. Obwohl diese Tatsache nicht bei jedem System als Nachteil gewertet werden kann, sind Eingriffe besonders während der Trainingsphase aufgrund ihrer Subjektivität und aus Effizienzgründen ungünstig.

3.2 Systeme zur automatischen Annotation

Aus den vorangehenden Punkten geht hervor, dass der Einsatz von automatischer Beschlagwortung rege diskutiert wird. Um klare Aussagen zur Perspektive von automatischer Annotation treffen zu können, müssen neben der Analyse der theoretischen Grundlagen auch bestehende Systeme betrachtet werden. Dabei kann man die Systeme in Bezug auf verschiedene Faktoren unterscheiden. Wie bereits behandelt, unterscheidet man grob zwischen textbasierten und inhaltsbasierten Systemen. Als weitere allgemeine Attribute können die Stufe der Automatisierung, der Anwendungsbereich und die verwendete Datenbasis der Systeme genannt werden. Eine Auflistung aller Unterscheidungskriterien kann anhand der spezifischen Ausrichtung einiger Anwendungen ohnehin nicht erfolgen. Im Fokus

dieser Arbeit sollen insbesondere inhaltsbasierte Systeme stehen, die auf den Einsatz in großen Bildarchiven mit vielen potentiellen Nutzern ausgerichtet sind. Kurz – es soll sich um möglichst große, universelle Systeme handeln.

3.2.1 Prototypen

Ein Großteil der in der Fachliteratur behandelten Systeme baut aus diversen Gründen auf kleinen Bilddatenbanken auf. Als Motive werden dabei meist rechtliche Probleme und Kapazitätsprobleme angeführt. Hervé und Boujemaa bemerken dazu:

> „Moreover, and paradoxically, the availability of large-scale image databases for research purposes is compromised by the uncertainty of copyright ownership. This leads researchers to work on a small number of commonly available databases." (HERVÉ & BOUJEMAA, 2007, p. 170)

Aus diesem angeführten Grund ist es nicht verwunderlich, dass die erstellten Systeme den Sprung vom Prototypenstatus zu einem Produktivsystem nur sehr selten bewältigen. Die Fachliteratur beschränkt sich bei der Vorstellung dieser Systeme meist auf die Erläuterung der zugrunde liegenden Theorien, beispielhafte Beschreibungen der Systemfunktionen und die ausgiebige Darlegung der durchgeführten Tests und Benchmarks. Einen aussagekräftigen und objektiven Eindruck über die Funktionsweise und die Güte eines solchen Systems zu gewinnen ist für Außenstehende faktisch nicht möglich. Der vielversprechende Ansatz der Gruppe um Nuno Vasconcelos und Gustavo Carneiro lässt sich neben unzähligen anderen zu diesen Prototypenystemen zählen (vgl. CARNEIRO ET AL., 2007; A B CHAN ET AL., 2006).

3.2.2 ALIPR

Bisher ist es lediglich einer Forschungsgruppe gelungen, eine Echtzeitanwendung zu realisieren. Nach der Implementierung des ALIP-Systems [1] (Automatic Linguistic Indexing of Pictures) zur Beschlagwortung von Bildern wurde es 2005 unter dem Namen ALIPR (Automatic Linguistic Indexing of Pictures in Realtime) entwickelt und vorgestellt (vgl. J. LI & J. Z. WANG, 2005).

3.2.2.1 Bildanalyseverfahren

Um die Echtzeitverarbeitung der Suchanfragen zu gewährleisten, musste die Methode zur Merkmalsextraktion und zur Annotation entsprechend angepasst werden. Dabei galt es einen ausgewogenen Kompromiss zwischen Rechenaufwand und Retrievalqualität zu treffen. Zur Reduzierung der Rechenzeit werden bei ALIPR lediglich die Farb- und die Strukturmerkmale als 'low-level features' extrahiert. Die Autoren heben jedoch explizit hervor, dass sich auch andere Bildmerkmale wie Formmerkmale oder affine Bildpunkte an das vorgeschlagene Modell anpassen lassen. Die Bildmerkmale werden nach der Segmentierung des Bildes aus den einzelnen Regionen extrahiert. Zu dieser Segmentierung werden auf Wavelet-Transformation und 'k-means-clustering' basierende Methoden eingesetzt (vgl. J. LI & J. Z. WANG, 2006, p. 94). Diese Verfahrensweise ist eng an das von Wang, Li und Wienerhold vorgeschlagene CBIR-System SIMPLIcity angelehnt (vgl. J. Z. WANG ET AL., 2001). Dadurch werden von den Entwicklern aber auch Nachteile in Kauf genommen:

> „Unfortunately, this method is more suitable for recognizing scenes, and thus, we expect the method will be insufficient for recognizing individual objects, given the great variations a type of objects (e.g., dogs) can appear in pictures. Although object names are often assigned by the system, the selection is mostly based on statistical correlation with scenes." (J. LI & J. Z. WANG, 2006, p. 94)

Je nach Bildinhalt wird für jedes Bild eine unterschiedlich große Menge von Segmenten identifiziert die anschließend anhand ihrer Bildmerkmale miteinander verglichen werden. Die regionale Merkmalsmodellierung bei ALIPR basiert auf einem statistischen Modell zur diskreten Wahrscheinlichkeitsverteilung und deshalb muss die Signaturmodellierung ebenfalls an dieses Modell angepasst werden. Ansonsten können die Bilder nicht untereinander verglichen werden. Für dieses Problem wurde ein spezielles statistisches Modell entwickelt – das so genannte 'discrete distribution-clustering'. Kombiniert mit einem Verfahren zum 'mixture modelling' ist es damit möglich die Bilder zu inhaltlich beziehungsweise semantisch ähnlichen Clustern zu gruppieren. Diese aus dem Trainingsset gewonnenen Cluster dienen als semantische Konzepte für den Vergleich mit unbeschlagworteten Bildern (J. LI & J. Z. WANG, 2006, pp. 94-95).

3.2.2.2 Annotationsmethode

Bei der Verarbeitung eines neuen Bildes werden wie bereits erwähnt dessen Bildmerkmale extrahiert und auf gleiche Weise aufbereitet wie die Trainingsbil-

der. Dann wird für jedes semantische Konzept die Wahrscheinlichkeit *p1* berechnet, dass das Suchbild dem jeweiligen Cluster zugeteilt werden kann. Li und Wang gehen davon aus, dass die Wahrscheinlichkeit einem neuen Bild eindeutig ein Konzept zuordnen zu können sehr gering ist.

> „The posterior probability decreases slowly across the concepts, suggesting that the most likely concept for each image is not strongly favored over the others. It is therefore important to quantify the posterior probabilities rather than simply classifying an image into one concept" (J. LI & J. Z. WANG, 2006, p. 99)

Die Schlagwörter der einzelnen Konzepte werden mit dem identischen Verfahren dem Suchbild zugeordnet. Gewichtet nach *p1* wird für jedes Schlagwort die Wahrscheinlichkeit *p2* errechnet, ob es zutreffend für das Suchbild ist. Vereinfacht ausgedrückt: Tritt ein beliebiges Schlagwort mehrmals in Konzepten mit hoher Wahrscheinlichkeit *p1* auf, steigt auch die Wahrscheinlichkeit *p2* – also die Wahrscheinlichkeit, dass das Schlagwort dem Bild zugeteilt wird. Ein festgelegter Schwellenwert bestimmt im letzten Schritt über die Aufnahme der wahrscheinlichsten Schlagwörter in die Annotation. Jedes Bild wird automatisch mit den

Abbildung 14: Beispiel für die Beschlagwortung des ALIPR-Systems (J. LI & J.Z. WANG, 2006 p. 104)

neun wahrscheinlichsten Wörtern versehen (vgl. J. LI & J. Z. WANG, 2006, pp. 99-100). Beispiele für die Annotationsvorschläge des Systems können in Abbildung 14 betrachtet werden.

3.2.2.3 ALIPR in der Praxis

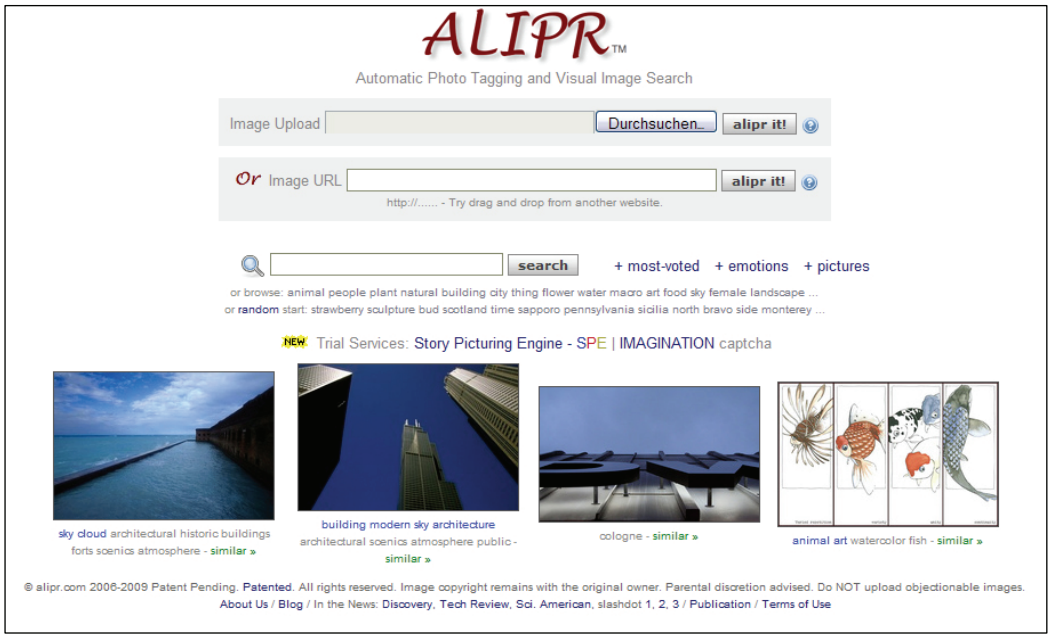

Abbildung 15:ALIPR Web-Schnittstelle (ALIPR, 2009)

Da das ALIPR-System in Echtzeit arbeitet, ergeben sich neue Einsatzszenarien. ALIPR ist das erste automatische Bildannotationssystem das öffentlich im Web zugänglich ist. Nachdem das System mit ausreichend vielen Bilddaten trainiert wurde und durch manuelle Überprüfung der Beschlagwortung gewährleistet werden konnte, dass das System ausreichend gute Ergebnisse liefert, wurde es im Oktober 2006 öffentlich zugänglich gemacht. Unter der Webadresse www.alipr.com können Suchanfragen über das ALIPR-System getätigt werden (vgl. Abbildung 15). Dabei kann das System in zwei Unterbereiche gegliedert werden – die automatische Annotation und das Bildretrieval.

Die Suchanfragemaske wurde analog zu bekannten Retrievalsystemen wie der Google-Suche und ähnlichen Anwendungen, schlicht gehalten. Aufgrund der verschiedenen Anfragemöglichkeiten werden bei ALIPR dem Anwender jedoch drei Anfragefelder angeboten. Zwei Anfragefelder dienen zur Adresseingabe für die bildliche Suche, über das dritte Feld kann die Suche mit Schlagwörtern gestartet werden. Zudem kann der Benutzer zur Suche aus einer Reihe beispielhafter Schlagwörter wählen oder durch die Auswahl eines von vier Beispielbildern mit der Suche beginnen.

Zentraler Bestandteil des Systems ist jedoch die bildliche Anfragekomponente. Dabei können Bilder die lokal oder im Web zugänglich sind, vom System beschlagwortet werden. Um die Annotationsqualität zu erhöhen, werden dem Benutzer die fünfzehn wahrscheinlichsten Schlagwörter vorgeschlagen und er darf die für ihn zutreffenden Wörter auswählen. Per Hand können zusätzliche Schlagwörter eingegeben werden, falls der Benutzer die vorgeschlagenen Wörter als nicht ausreichend für die Beschreibung erachtet. Weiterhin können der Bildtitel, ein Verweis zu ähnlichen Bildern und die Copyright-Informationen angegeben werden. Das Bild wird daraufhin in der Datenbank abgelegt und dem Benutzer werden die Ergebnisse der Suchanfrage geliefert. Dabei werden die Bilder anhand der vergebenen Schlagwörter verglichen (vgl. ALIPR, 2009).

3.2.2.4 Bisherige Tests mit dem ALIPR System

Im Fall des ALIPR Systems wurde bisher lediglich die Systemperspektive durch umfangreiche Tests evaluiert. Anhand drei verschiedener Bildkollektionen werden bei ALIPR Recall und Precision der automatischen Annotation erhoben. Dabei werden nicht die Ergebnisse einer bildlichen Suchanfrage betrachtet, sondern die Zuteilung von Schlagwörtern zum Suchbild. Ein Vergleich zu anderen Systemen wird aufgrund des hohen Arbeitsaufwandes und der speziellen Stellung des Systems im Bezug auf die Echtzeitannotation nicht vorgenommen (vgl. J. LI & J. Z. WANG, 2006, p. 100).

Corel Stock Photo Library

Die Tests werden mit der am häufigsten verwendeten Bildsammlung – der Corel Stock Photo Library – durchgeführt, die bereits zum Training des Annotationssystems verwendet wird. Diese Sammlung besteht aus 599 Bildkategorien, die jeweils 100 Bilder enthalten. Zum Training des Systems werden jeweils nur 80 Bilder aus jeder Kategorie eingesetzt. Die restlichen 20 Bilder werden zur Evaluation herangezogen. Li und Wang bemerken zum Einsatz dieses Bildsets:

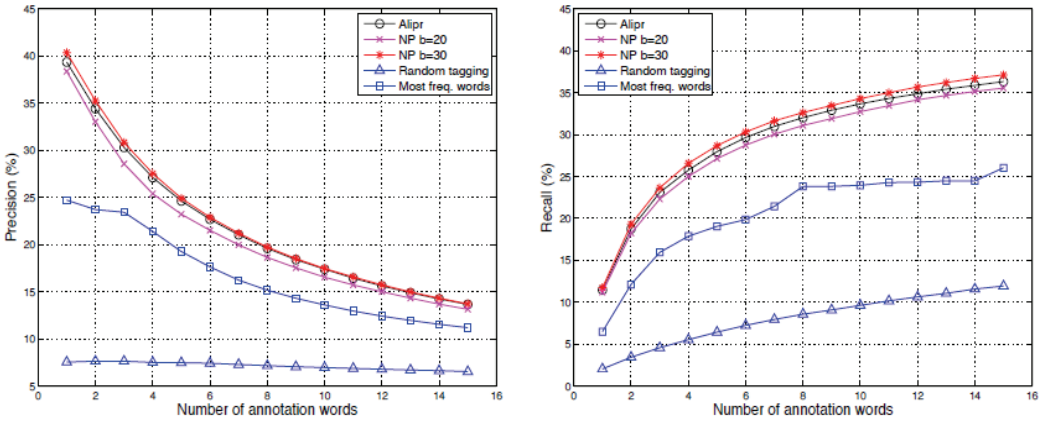

Abbildung 16: Recall und Precision der Annotation von ALIPR anhand des Corel Test-sets (LI & WANG, 2006, p. 111)

> „Performance achieved in this case, however, is optimistic because the Corel images are known to be highly clustered, that is, images in the same category are sometimes extraordinarily alike." (J. LI & J. Z. WANG, 2006, p. 100)

Zur Einordnung der Systemleistung werden alternative Annotationsmethoden verwendet. Die Bilder werden mittels parameterfreier Statistiken, durch zufällige Annotation und nach der Wortfrequenz der Schlagwörter in den ermittelten Klassen zugeteilt. Eine detaillierte Erläuterung dieser Verfahren findet sich im Originaltext (vgl. J. LI & J. Z. WANG, 2006, pp. 100-102). Wie Abbildung 16 zeigt, liefert das eingesetzte Verfahren gute Ergebnisse im Vergleich zur zufälligen Annotation beziehungsweise zur Annotation anhand der Wortfrequenz. Annotation durch parameterfreie Statistik liefert zwar vergleichbar gute Ergebnisse, wird jedoch von den Autoren als rechenintensiver und weniger robust eingestuft:

> „The NP approach is computationally more intensive during annotation than ALIPR because in ALIPR, we only need distances between a test image and each prototype, while the NP approach requires distances to every training image." (J. LI & J. Z. WANG, 2006, p. 102)

Flickr-Datenset

Die Annotation wird durch den Einsatz einer manuell zusammengestellten Datenbank von 54700 Bildern aus der Flickr-Sammlung überprüft. Dabei werden sowohl qualitative als auch quantitative Aussagen getroffen. Bei der qualitativen Analyse wurden von den Autoren speziell fünf Schwachpunkte des Systems hervorgehoben (vgl. J. LI & J. Z. WANG, 2006, p. 104):

a) when the way an object is taken in the picture is very different from those in the training
b) when the picture is fuzzy or of extremely low resolution or low contrast
c) if the object is shown partially
d) if the white balance is significantly off
e) if the object or the concept has not been learned

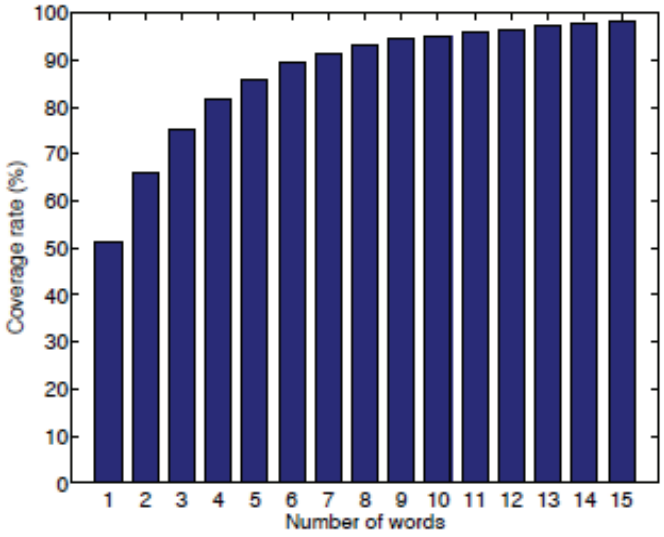

Abbildung 17: Abdeckungsrate der fünfzehn wahrscheinlichsten Schlagwörter (LI & WANG, 2006, p. 104)

Zur quantitativen Analyse wird die webbasierte Variante des ALIPR-Systems eingesetzt. Durch die manuelle Auswahl relevanter Schlagwörter wurden von Experten über 5000 der Flickr-Bilder annotiert. Dabei wird vor allem das Verhältnis der richtig erkannten im Vergleich zu den falsch erkannten Bildern betrachtet. Als großer Schwachpunkt dieser Methode wird die Subjektivität des Experten bei der Zuteilung der Schlagwörter gesehen. Um diesen Effekt auszugleichen, werden die Experten zum einen nicht in die Systementwicklung eingebunden und zum anderen müssen festgelegte Heuristiken und Auswahlregeln befolgt werden. Die Untersuchungen zeigen, dass für 98,13% aller Bilder mindestens eines der fünfzehn Schlagwörter zutreffend ist. Für eine Abdeckung von 80% reichen bereits die vier wahrscheinlichsten Schlagwörter aus. Weiterhin werden im Schnitt 4,1 zutreffende Wörter gefunden (vgl. J. LI & J. Z. WANG, 2006, pp. 102-105).

ALIPR-Datenset

Als dritte Testkollektion wurde die stetig wachsende Menge der über die ALIPR-Website eingereichten Bilder verwendet. Die Auswahl der Schlagwörter erfolgt dabei nicht wie im vorhergehenden Experiment durch einen Experten, sondern

# of checked tags	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# of images	3277	2824	2072	1254	735	368	149	76	20	22	3	1	2	3	3
(%)	30.3	26.1	19.2	11.6	6.8	3.4	1.4	0.7	0.2	0.2	0.	0.	0.	0.	0.

Abbildung 18: Durchschnitt der von Benutzern korrekt eingestuften Schlagwörter (LI & WANG, 2006, p. 106)

wird dem Benutzer überlassen. Dabei kann er – wie bereits beschrieben – auch zusätzliche Schlagwörter hinzufügen. Aufgrund dieser unkontrollierten Dateneingabe werden von den Autoren lediglich subjektive Beobachtungen erwähnt. Dabei unterstellen sie den Benutzern sehr strenge Auswahlkriterien bei der Schlagwortvergabe und bewusst schwierig gewählte Bildinhalte. So werden beispielsweise nur Schlagwörter gewählt, die den Hauptgegenstand des Bildes beschreiben oder Suchanfragen mit manipuliertem Bildinhalt gestellt. Im Durchschnitt werden pro Bild 2,24 Wörter als korrekt zugeordnet und 1,67 Wörter manuell angefügt. (vgl. J. LI & J. Z. WANG, 2006, pp. 105-106)

4 Akzeptanzevaluation des ALIPR-Systems

In den vorhergehenden Kapiteln werden viele verschiedene Theorien zur Gestaltung eines Bildretrievalsystems betrachtet. Die Auseinandersetzung mit den theoretischen Grundlagen des Bildretrieval und der automatischen Annotation wirft aber die Frage auf, wie ein nach diesen Theorien entworfenes System bewertet werden kann. Es müssen Möglichkeiten gefunden werden, die Qualität und die Effektivität einer solchen Applikation zu bewerten. In dieser Arbeit soll die Qualität des vorgestellten ALIPR-Systems und der dabei eingesetzten Bildanalyseverfahren überprüft werden. Der Fokus liegt hierbei auf der Bewertung Systemkomponente zur automatischen Annotation durch die Benutzer.

4.1 Evaluation im Information Retrieval

Bevor eine geeignete Methode zur Überprüfung dieser Fragen entwickelt werden kann, muss zuerst ein Blick auf die übliche Vorgehensweise auf diesem Gebiet geworfen werden. Die Evaluation von interaktiven Systemen ist eine zentrale Disziplin auf dem Gebiet des Information Retrieval. Einen ausführlichen Überblick über die allgemeinen Verfahren zur Evaluation im Information Retrieval findet sich in diversen Publikationen (vgl. BAEZA-YATES, 2005, pp. 73-98; MANNING ET AL., 2008, pp. 139-161). Durch die speziellen Voraussetzungen bei der Evaluation von bildbasierten Retrievalsystemen gilt es die Evaluationsmethoden an die Umstände anzupassen, die generellen Verfahren bleiben jedoch die Gleichen.

4.1.1 Evaluationsmethoden beim Bildretrieval

Die Bewertung der Effektivität, beziehungsweise der Qualität, erfolgt bei jedem Bildretrievalsystem anhand einer Evaluation. Dadurch kann zum einen die Leistung der verschiedenen Systeme untereinander verglichen werden, zum anderen können Systemdefizite aufgedeckt und verbessert werden. Die Evaluationsmethoden können in systemzentrierte und benutzerzentrierte Ansätze aufgeteilt werden (vgl. DÍAZ ET AL., 2008, pp. 1294-1295; TURPIN & SCHOLER, 2006, pp. 11-13).

4.1.1.1 Systemzentrierte Evaluation

Die Evaluation beim bildbasierten Retrieval konzentriert sich in erster Linie auf die Erhebung der Systemeffektivität anhand der erzielten Ergebnisse. Durch die

Analyse der aus dem Information Retrieval bekannten Gütemaße Precision und Recall können Systeme objektiv verglichen werden (vgl. BAEZA-YATES, 2005, pp. 74-82). Wie bei dem beschriebenen Test des ALIPR-Systems mit der Corel Bildsammlung gezeigt wurde, ist es notwendig, bestimmte Grundlagen für die Erhebung dieser Gütemaße zu schaffen. Hierzu zählen ein angemessener Testdatenbestand und eine eindeutig definierte Menge relevanter und irrelevanter Schlagwörter – beim CBIR als 'ground truth' bekannt (vgl. DATTA ET AL., 2008, p. 51). Die Berechnung von Recall und Precision hat den Vorteil, dass das System meist vollkommen automatisiert geprüft werden kann. Es gilt als objektives Mittel zur Ermittlung der Systemgüte von Retrieval Systemen. Aufgrund des häufigen Einsatzes von Recall und Precision in allen Bereichen des Information Retrieval, existieren viele verschiedene Varianten zur Berechnung dieser Maße und vergleichbarer Alternativen (vgl. BAEZA-YATES, 2005, pp. 75-82). Die gängigste Methode zur Messung von Recall und Precision ist beim Bildretrieval die 'Mean Average Precision', bei dem das Ergebnisranking in die Berechnung mit einbezogen wird (vgl. DATTA ET AL., 2008, p. 52). Um zuverlässige Vergleiche zwischen verschiedenen Bildretrievalsystemen zu ziehen, ist es notwendig, dass die Systeme dieselbe Testkollektion verwenden. Eine Initiative zur Standardisierung der Kollektionen wie sie im Textretrieval mit TREC realisiert wurde (vgl. VOORHEES, 2005), hat sich beim Bildretrieval noch nicht durchgesetzt. Daher werden mehrere Bildsammlungen zum Vergleich der Systeme herangezogen. Die gängigsten Kollektionen sind die bereits erwähnte Corel Stock Photo Library, die WANG Kollektion, die IRMA-10000 Kollektion und einige andere (vgl. DESELAERS ET AL., 2008, pp. 8-10). Eine weiterführende, detaillierte Abhandlung zum effektiven Einsatz von Recall und Precision bei bildbasierten Retrievalsystemen findet sich zudem bei Huijsmans und Sebe (vgl. HUIJSMANS & SEBE, 2005).

4.1.1.2 Benutzerzentrierte Evaluation

Der zentrale Kritikpunkt bei der systemzentrierten Evaluation ist die mangelnde Einbindung der Benutzerbelange. Deshalb ist es naheliegend, sich bei der Evaluation von Systemen mit hoher Benutzereinbindung sich auf die Benutzerperspektive zu konzentrieren. Nielsen formulierte in seiner Publikation über Usability Engineering von 1993:

> „User Testing with real users is the most fundamental usability method and is in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the concrete interface being tested." (NIELSEN, 1993, p. 165)

Al-Maskari et al. erwähnen mehrere Studien, bei denen aus Benutzersicht, trotz einer Steigerung von Recall und Precision keine messbaren Effektivitätsverbesserungen erreicht werden konnten. Unter anderem führen sie die Benutzerzufriedenheit als ausschlaggebenden Faktor zur Bewertung eines Retrievalsystems an und überprüfen diese Theorie am Beispiel eines bildbasierten Retrievalsystems. Es wird nachgewiesen, dass Benutzer bereits bei relativ niedriger Systemeffektivität ausreichende Qualität feststellen und mit dem Retrievalergebnis zufrieden sind (vgl. AL-MASKARI ET AL., 2006). Es ist daher sinnvoll, spezielle Theorien für benutzerzentrierte Evaluationen zu entwickeln und neue Ansätze zur Qualitätsbewertung von Retrievalsystemen zu finden.

> „Developing user-centric benchmarks is a next generation challenge for researchers in CBIR and associated areas." (DATTA ET AL., 2008, p. 53)

4.1.2 Akzeptanz als Bewertungskriterium

Software-Ergonomie befasst sich seit Jahren ausführlich mit der Frage welche Erkenntnisse durch die Einbeziehung der Benutzer in den Evaluierungsprozess gewonnen werden können. Begriffe wie 'Usability Testing' beschreiben nichts anderes, als die benutzerorientierte Evaluation der Gebrauchstauglichkeit eines Softwareprodukts. Die Ausrichtung benutzerorientierter Evaluation ist nach dieser Definition jedoch allgemeiner. Hier wird nicht nur die von Al-Maskari et al. erwähnte, subjektiv erlebte Effektivität eines Produkts, sondern es werden alle Teilbereiche seiner Benutzungsfreundlichkeit evaluiert. Die vom Benutzer subjektiv empfundene Effektivität eines Systems kann durch die Anfertigung einer Akzeptanzstudie ermittelt werden (vgl. SCHWEIBENZ & THISSEN, 2002, pp. 118-119). Die Anfertigung einer solchen Studie setzt die Zusammenstellung eines geeigneten theoretischen Modells voraus – einige der populärsten Vertreter werden im Folgenden vorgestellt.

4.1.2.1 Technology Acceptance Model

Das von Davies et al. aufgestellte 'Technology Acceptance Model' (TAM) setzt die subjektive Akzeptanz mit der Benutzerfreundlichkeit eines Systems gleich. Nach diesem Modell werden zwei Hauptkomponenten für die Akzeptanz eines Benutzers gegenüber einem System identifiziert: 'perceived usefulness' und 'perceived ease-of-use'. Darunter kann einerseits die wahrgenommene Nützlichkeit eines Systems für eine bestimmte Tätigkeit und andererseits die wahrgenommene Bedienfreundlichkeit eines Systems verstanden werden. Der dabei vom Benutzer gewonnene subjektive Gesamteindruck kann als Gradmesser für die Usability

eines Systems gesehen werden (vgl. DAVIS ET AL., 1989). Das Modell basiert grundsätzlich auf der so genannten 'Theory of reasoned action', die bereits 1975aufgestellt wurde (vgl. FISHBEIN & AJZEN, 1975). Um verlässliche und aussagekräftige Untersuchungen mit dem TAM zu gewährleisten, wurde es unter anderem von Wixom und Todd nochmals angepasst (vgl. WIXOM & TODD, 2005).

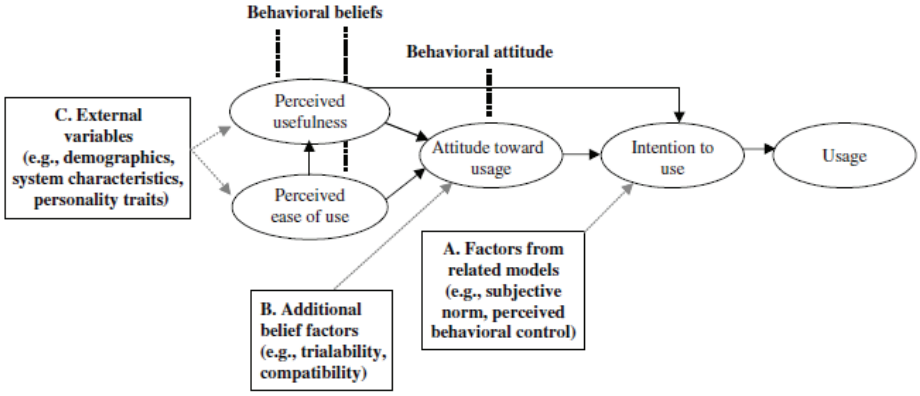

Abbildung 19: TAM und mögliche Erweiterungen (vgl. WIXOM & TODD, 2005, p. 87)

Als akzeptanzbeeinflussende Hauptfaktoren, werden die Informationsqualität und die Systemqualität identifiziert, die sich wiederum in mehrere Faktoren aufteilen lassen. Die Systemqualität lässt sich laut Wixom und Todd in fünf Bereiche unterteilen:

> „For system quality, *reliability* refers to the dependability of system operation, *flexibility* refers to the way the system adapts to changing demands of the user, *integration* refers to the way the system allows data to be integrated from various sources, *accessibility* refers to the ease with which information can be accessed or extracted from the system, and *timeliness* refers to the degree to which the system offers timely responses to requests for information or action." (WIXOM & TODD, 2005, p. 90)

Informationsqualität wird folgendermaßen definiert:

> „Information quality is shaped by four dimensions: *completeness* represents the degree to which the system provides all necessary information; *accuracy* represents the user's perception that the information is correct; *format* represents the user's perception of how well the information is presented; and *currency* represents the user's perception of the degree to which the information is up to date." (WIXOM & TODD, 2005, p. 91)

Abbildung 19 zeigt die Verbesserungsvorschläge der Autoren am ursprünglichen TAM, die zur Neuformulierung des Modells anregen.

4.1.2.2 Unified theory of acceptance and use of technology

Venkatesh und Morris erweitern, basierend auf acht Ansätzen zur Akzeptanzmodellierung, die Akzeptanztheorie zu ihrem Modell der 'Unified Theory of Acceptance and Use of Technology'. Dieses in der Kurzform UTAUT genannte Akzeptanzmodell gliedert die akzeptanzbeeinflussenden Faktoren in vier Bereiche auf: die erwartete Systemleistung, den zu erwarteten Arbeitsaufwand, soziale Einflüsse und erleichternde Gegebenheiten. Neben diesen Hauptfaktoren werden noch weitere Faktoren wie die allgemeine Einstellung gegenüber neuen Technologien, Selbstvertrauen oder Verängstigung in die Untersuchung miteinbezogen. Zur Gewichtung der Faktoren postulieren die Autoren Einflussgrößen wie das Alter, das Geschlecht, die Erfahrung und den Grad der freiwilligen Nutzung seitens der Testpersonen. Für jeden Faktor werden Fragebogenitems vorgeschlagen, welche teilweise von den untersuchten Modellen abgeleitet und teilweise eigens für die Theorie entworfen werden. Abbildung 20 zeigt die verwendeten Items gegliedert nach den einzelnen Faktoren (vgl. VENKATESH & MORRIS, 2003, pp. 427-460).

Table 16. Items Used in Estimating UTAUT

Performance expectancy

U6:	I would find the system useful in my job.
RA1:	Using the system enables me to accomplish tasks more quickly.
RA5:	Using the system increases my productivity.
OE7:	If I use the system, I will increase my chances of getting a raise.

Effort expectancy

EOU3:	My interaction with the system would be clear and understandable.
EOU5:	It would be easy for me to become skillful at using the system.
EOU6:	I would find the system easy to use.
EU4:	Learning to operate the system is easy for me.

Attitude toward using technology

A1:	Using the system is a bad/good idea.
AF1:	The system makes work more interesting.
AF2:	Working with the system is fun.
Affect1:	I like working with the system.

Social influence

SN1:	People who influence my behavior think that I should use the system.
SN2:	People who are important to me think that I should use the system.
SF2:	The senior management of this business has been helpful in the use of the system.
SF4:	In general, the organization has supported the use of the system.

Facilitating conditions

PBC2:	I have the resources necessary to use the system.
PBC3:	I have the knowledge necessary to use the system.
PBC5:	The system is not compatible with other systems I use.
FC3:	A specific person (or group) is available for assistance with system difficulties.

Self-efficacy

I could complete a job or task using the system…

SE1:	If there was no one around to tell me what to do as I go.
SE4:	If I could call someone for help if I got stuck.
SE6:	If I had a lot of time to complete the job for which the software was provided.
SE7:	If I had just the built-in help facility for assistance.

Anxiety

ANX1:	I feel apprehensive about using the system.
ANX2:	It scares me to think that I could lose a lot of information using the system by hitting the wrong key.
ANX3:	I hesitate to use the system for fear of making mistakes I cannot correct.
ANX4:	The system is somewhat intimidating to me.

Behavioral intention to use the system

BI1:	I intend to use the system in the next <n> months.
BI2:	I predict I would use the system in the next <n> months.
BI3:	I plan to use the system in the next <n> months.

Abbildung 20: Fragebogenitems zur Bestimmung von UTAUT (VENKATESH & MORRIS, 2003, p. 460)

4.2 Theoretisches Modell der Studie

Betrachtet man die verschiedenen Retrievalmethoden, die das ALIPR-System anbietet, stellt sich die Frage, welche der beiden Methoden von potentiellen Benutzern bevorzugt wird. Es wurde bereits behandelt, dass die Güte eines Systems einerseits aus der Systemperspektive und andererseits aus der Benutzerperspektive bewertet werden kann. Da ALIPR das erste System ist, das automatische Annotation im Web ermöglicht, ist es zugleich das erste System, das von einem großen Benutzerkreis verwendet wird. Bereits wenige Wochen nach seiner Veröffentlichung verzeichnete die Website mehrere tausend Zugriffe pro Tag (vgl. J. LI & J. Z. WANG, 2006, p. 105). Aus diesem Grund ist die benutzerseitige Sicht unter informationswissenschaftlicher Perspektive ebenso relevant.

Im Zusammenspiel mit den Theorien zur Akzeptanz ergeben sich, hinsichtlich des Einsatzes von ALIPR, interessante Fragestellungen. Der Fokus der Untersuchung liegt dabei auf den unterschiedlichen Anfrage- und Retrievalparadigmen, die das System anbietet. Der Einfluss der Systemqualität – dass heißt von Faktoren aus dem Bereich des Software-Engineering – soll möglichst minimiert werden. Da die Anfragen mit Hilfe derselben Schnittstelle erfolgen, kann davon ausgegangen werden, dass diese Faktoren vernachlässigt werden können. Aufgrund des begrenzten zeitlichen Spielraumes, der begrenzten Kapazitäten und der spezifischen Anforderungen des Themas muss aus den bestehenden Akzeptanzmodellen ein vereinfachtes Modell erstellt werden.

Die Untersuchung verschiedener Retrievalkonzepte – einerseits das tagbasierte, andererseits das bildbasierte Retrieval – und innovativer Anfragemodelle, wie der automatischen Annotation, kann nicht problemlos auf ein Modell zur Akzeptanz gegenüber Informationstechnologie übertragen werden. Nach eingehender Studie der vorangegangenen Theorien wurde ein Modell gewählt, das sich an den Faktoren des ursprünglichen 'Technology Acceptance Model' orientiert, aber auch Ansätze der Folgemodelle miteinbezieht.

Als Hauptfaktoren für die Akzeptanz der Benutzer gegenüber den von ALIPR angebotenen Suchparadigmen werden die subjektiv wahrgenommene Nützlichkeit und die subjektiv wahrgenommene Schwierigkeit identifiziert. Weiterhin sollen systemseitige Effektivitätsmaße als Vergleichsgrößen eingesetzt werden. Die weiteren akzeptanzbeeinflussenden Faktoren werden durch die gezielte Anpassung der Evaluation so weit als möglich miteinbezogen oder neutralisiert.

4.3 Fragestellungen und Hypothesen

Vor der praktischen Modellierung eines geeigneten Instruments zur Akzeptanzevaluation gilt es, geeignete Fragestellungen zu formulieren und eine entsprechende Methode zur Überprüfung dieser Thesen zu finden.

4.3.1 Kernhypothesen

Wie beschrieben, werden von ALIPR standardmäßig Vorschläge zur Beschlagwortung eines nicht annotierten Bildes gemacht. Dem Benutzer steht es danach frei, aus fünfzehn vorgeschlagenen Wörtern zu wählen und manuell weitere zutreffende Wörter und andere Zusatzinformationen anzugeben. Die Auswahl der vorgeschlagenen Tags ist der erste Interaktionsschritt, der für die Untersuchung von Interesse ist, da der Benutzer im Gegensatz zu gängigen Retrievalsystemen, schon vor der Ergebnispräsentation in den Retrievalprozess eingreifen muss. Es gilt die Frage zu beantworten, ob der Benutzer die frühe Einbindung akzeptiert oder dadurch gestört wird. Somit lässt sich die erste Hypothese formulieren:

(1) Die Akzeptanz der Benutzer gegenüber dem Interaktionsschritt zur automatischen Schlagwortvergabe ist hoch.

Nach der Auswahl der vom System vorgeschlagenen Schlagwörter werden ähnlich beschlagwortete Bilder ausgegeben. Sind die gelieferten Ergebnisse nicht ausreichend, besteht neben verschiedenen Formen zur Verfeinerung der Suchanfrage die Möglichkeit, Ergebnisse anhand inhaltsbasierter Merkmale anzufordern. Es stehen sich also zwei Anfragemethoden gegenüber, die von den gleichen Bildmerkmalen ausgehend, zwei grundlegend verschiedene Ansätze des Retrieval anbieten – das durch inhaltsbasierte Analyse unterstützte, tagbasierte und das rein inhaltsbasierte Retrieval. Ein Vergleich dieser beiden Paradigmen ist sowohl aus der System-, als auch aus der Benutzersicht aufschlussreich. Davon ausgehend, dass durch die frühe Einbeziehung der Benutzer die tagbasierte Anfragemethode qualitativ bessere Ergebnisse liefert, als die inhaltsbasierte Methode, lauten die weiteren Hypothesen:

(2) Die durch die tagbasierte Anfragemethode gelieferten Ergebnisse sind signifikant besser als die der inhaltsbasierten Anfragemethode

und darauf aufbauend:

(3) Die Akzeptanz der Benutzer gegenüber der tagbasierten Methode ist signifikant höher, als die Akzeptanz gegenüber der inhaltsbasierten Anfragemethode

4.3.2 Nebenbeobachtungen

Neben der Überprüfung der drei angeführten Haupthypothesen gibt es eine Reihe zusätzlicher Fragestellungen. Vor allem müssen Wege gefunden werden, um den Einfluss weiterer akzeptanzbeeinflussender Faktoren zu identifizieren und gegebenenfalls zu kompensieren. Es wird versucht, diese Faktoren durch die Analyse der persönlichen Meinungen der Benutzer herauszuarbeiten und somit den Umfang der Untersuchung im kontrollierbaren Rahmen zu halten. Dabei gilt der Fokus folgenden Fragestellungen:

a) Welche Vor-, oder Nachteile identifizieren die Benutzer bei den verschiedenen Retrievalmethoden des ALIPR Systems
b) Welchen subjektiven Eindruck haben die Benutzer vom ALIPR System
c) Haben die Benutzer Verbesserungsvorschläge

Zusätzlich sollen zur besseren Einordnung der Ergebnisse, in Bezug auf die 'Unified Theory of Acceptance and Use of Technology', demografische Angaben abgefragt werden. Hierzu zählen das Alter, das Geschlecht und der Beruf der Teilnehmer.

Parallel zur Untersuchung der Benutzerakzeptanz, sollen zusätzlich Rückschlüsse zur Auswahl der vorgeschlagenen Tags gezogen werden. Neben der Auswahl der automatisch generierten Tags, bietet ALIPR die Eingabe zusätzlicher Schlagwörter an. Um die Güte der automatischen Annotation systemseitig zu evaluieren, ist es hilfreich, für jeden Benutzer die gleiche Grundmenge von Schlagwörtern anzunehmen. Aus diesem Grund werden bei der Evaluation keine manuell hinzugefügten Schlagwörter zugelassen. Im Hinblick auf die Ausführungen zum 'semantic gap', ist der Vergleich zwischen manueller Beschlagwortung und der semi-automatischen Beschlagwortung bei ALIPR aber lohnend. Erhebt man, unabhängig von den bei ALIPR ausgewählten Tags, die manuellen Benutzervorschläge, kann dieser Vergleich trotzdem gezogen werden. Dabei stehen drei Fragestellungen im Mittelpunkt:

d) In welchem Verhältnis variiert die Anzahl der manuell vergebenen Schlagwörter im Vergleich zur Menge der ausgewählten Schlagwörter bei ALIPR
e) Wie viele Übereinstimmungen bestehen zwischen den beiden Schlagwortmengen
f) Inwiefern unterscheiden sich die manuell vergebenen Schlagwörter von den automatisch vergebenen Schlagwörtern

4.3.3 Auswahl geeigneter Evaluationsinstrumente

Bei der Auswahl eines geeigneten Messinstruments muss überlegt werden, wie die theoretisch postulierten Faktoren und Einflussgrößen bestmöglich im vorgegebenen Rahmen erhoben werden können. Dabei stehen alle Instrumente zur quantitativen und qualitativen Erhebung von Benutzerverhalten zur Auswahl. Im Gegensatz zu anderen Bereichen der Akzeptanzforschung, steht auf dem Gebiet des bildbasierten Retrieval noch kein standardisiertes Untersuchungsinstrument zur Verfügung. Daher ist es, analog zur Formulierung des theoretischen Modells, nötig, ein geeignetes Instrument zu konstruieren. Unter Berücksichtigung gängiger Untersuchungsinstrumente aus dem Bereich des Bildretrieval, muss ein Weg gefunden werden, den speziellen Gegebenheiten dieser Untersuchung Rechnung zu tragen. Hierbei rücken vor allem benutzerzentrierte Ansätze in den Vordergrund Eine detaillierte Einteilung verschiedener benutzerzentrierter Evaluationsverfahren beim Usability Engineering findet sich in Abbildung 21.

Method Name	*Lifecycle Stage*	*Users Needed*	*Main Advantage*	*Main Disadvantage*
Heuristic evaluation	Early design, "inner cycle" of iterative design	None	Finds individual usability problems. Can address expert user issues.	Does not involve real users, so does not find "surprises" relating to their needs.
Performance measures	Competitive analysis, final testing	At least 10	Hard numbers. Results easy to compare.	Does not find individual usability problems.
Thinking aloud	Iterative design, formative evaluation	3–5	Pinpoints user misconceptions. Cheap test.	Unnatural for users. Hard for expert users to verbalize.
Observation	Task analysis, follow-up studies	3 or more	Ecological validity; reveals users' real tasks. Suggests functions and features.	Appointments hard to set up. No experimenter control.
Questionnaires	Task analysis, follow-up studies	At least 30	Finds subjective user preferences. Easy to repeat.	Pilot work needed (to prevent misunderstandings).
Interviews	Task analysis	5	Flexible, in-depth attitude and experience probing.	Time consuming. Hard to analyze and compare.
Focus groups	Task analysis, user involvement	6–9 per group	Spontaneous reactions and group dynamics.	Hard to analyze. Low validity
Logging actual use	Final testing, follow-up studies	At least 20	Finds highly used (or unused) features. Can run continuously.	Analysis programs needed for huge mass of data. Violation of users' privacy.
User feedback	Follow-up studies	Hundreds	Tracks changes in user requirements and views.	Special organization needed to handle replies.

Abbildung 21: Methoden zur Usability-Evaluation nach Nielsen (vgl. NIELSEN, 1993, p. 224)

In Bezug auf die benutzerzentrierte Evaluation der Akzeptanz gegenüber den Interaktionsvarianten von ALIPR, bieten sich in erster Linie quantitative Verfahren zur Erhebung an. Durch deren Einsatz können präzise und leicht reproduzierbare Ergebnisse gewonnen werden. Sie garantieren die objektive und zuverlässige Erhebung von Benutzermeinungen. Bei der quantitativen Evaluation

eines Systems unterscheidet man zwischen summativer und formativer Evaluation. Formative Evaluationen werden hauptsächlich während des Designprozesses eingesetzt, um Teilkomponenten eines Systems zu prüfen. Summative Untersuchungen beziehen sich auf die Gesamtleistung eines bereits fertigen Systems. Zudem sollten quantitative Evaluationen bestimmte Rahmenbedingungen erfüllen, um gültige Ergebnisse zu ermitteln (vgl. NIELSEN, 1993, pp. 165-170).

Bortz und Döring führen dabei drei Gütekriterien einer quantitativen Untersuchung an:

> „Die Qualität eines Tests bzw. eines Fragebogens lässt sich an drei zentralen Kriterien der Testgüte festmachen: Objektivität, Reliabilität und Validität." (BORTZ & DÖRING, 2003, p. 195)

Die Objektivität eines Tests gibt an, zu welchem Grad die Testperson und die Testergebnisse voneinander unabhängig sind. Reliabilität bezieht sich auf die Messgenauigkeit eines Tests und Validität beschreibt, wie gut der Test in der Lage ist, die postulierten Faktoren zu messen (vgl. BORTZ & DÖRING, 2003, pp. 195-202; SARODNICK & BRAU, 2006, p. 170).

Um die aufgestellten Hypothesen und Fragestellungen unter Berücksichtigung dieser Kriterien zu überprüfen, wurde die schriftliche Befragung der Benutzer mittels eines Fragebogens gewählt.

4.3.4 Fragebogenentwurf

Beim Entwurf des Fragebogens ist es nötig, mehrere Faktoren in dessen Gestaltung mit einzubeziehen. Zunächst muss entschieden werden, in welcher Form der Fragebogen den Testpersonen angeboten wird.

4.3.4.1 Modalität

Die gängigen Alternativen bei der Fragebogengestaltung sind einerseits die schriftliche Ausfertigung und andererseits die digitale Variante. Beide Varianten besitzen Vor- und Nachteile die es abzuwägen gilt: Online-Fragebögen besitzen den Vorteil, dass sie problemlos einer sehr großen Menge von Testpersonen zugänglich gemacht werden können und die erhobenen Daten bereits in digitaler Form zur Verarbeitung bereitliegen. Dafür muss in Kauf genommen werden, dass keine Möglichkeit zur Überwachung während der Bearbeitungszeit besteht. Klassische Papierfragebögen bieten diese Möglichkeit und darin liegt auch ihr größter Vorteil. Nachteilig wirkt sich der hohe Aufwand zur Übertragung der Daten ins

digitale Format, sowie die eingeschränkte lokale und temporäre Reichweite der schriftlichen Variante aus (vgl. SARODNICK & BRAU, 2006, pp. 172-173). Ein Vergleich der Vor- und Nachteile verschiedener Befragungsmethoden kann aus Abbildung 22 entnommen werden. Einen ausführlichen Überblick zur Auswahl eines geeigneten Verfahrens bietet Kaya (vgl. KAYA, 2007, pp. 52-54).

> „Test tasks should normally be given to the users in writing. Not only does this ensure that all users get the tasks described the same way, but having written tasks also allows the user to refer to the task description during the experiment instead of having to remember all the details of the task." (NIELSEN, 1993, p. 186)

Aufgrund der speziellen Ausrichtung der Evaluation, bei der neben der Bearbeitung des Fragebogens auch die Interaktion mit dem ALIPR System von den Testpersonen gefordert war, fällt die Entscheidung auf die schriftliche Variante der Befragung. Die Benutzer sollen sich ausschließlich auf ihre Aufgabe bei ALIPR konzentrieren und nicht zwischen verschiedenen Fenstern wechseln müssen. Der Entwurf eines elektronischen Befragungssystems mit integrierter ALIPR-Abfrage wird aufgrund des zeitlichen Aufwands verworfen. Es kommt ein summativer, teilstandardisierter Fragebogen zum Einsatz, bei dem sowohl Freitextantworten, als auch Ratingskalen verwendet wurden.

Beurteilungs-kriterium	Schriftliche Befragung	Telefonische Befragung	Persönliche Befragung	Internet- Befragung
Datengenauigkeit	sehr hoch	mittel bis sehr hoch	mittel bis sehr hoch	sehr hoch
Erhebbare Datenmenge pro Erhebungsfall	gering	mittel bis sehr groß	sehr groß	sehr groß
Flexibilität	gering	mittel bis sehr hoch	sehr hoch	hoch
Repräsentativität	gering bis hoch	hoch	sehr hoch	sehr hoch
Kosten pro Erhebungsfall	sehr gering	gering	mittel bis hoch	hoch
Zeitbedarf pro Erhebungsfall	mittel	niedrig bis sehr niedrig	mittel bis sehr hoch	niedrig bis sehr niedrig
Interviewer-Bias	sehr gering	hoch	sehr hoch	sehr gering

Abbildung 22: Vor- und Nachteile verschiedener Befragungsmethoden (KAYA, 2007, p. 54)

4.3.4.2 Umfang der Befragung

Der Umfang des Fragebogens muss zum einen ausreichend umfangreich für die Untersuchung der aufgestellten Hypothesen sein, zum anderen darf der Bearbeitungsaufwand dadurch nicht zu sehr ansteigen.

> „[...] it is normally best to keep the questionnaire short to maximize the response rate." (NIELSEN, 1993, p. 36)

Muss sich die Testperson über eine sehr lange Zeitspanne mit der Bearbeitung des Fragebogens auseinandersetzen, kann dies zur Verfälschung der Ergebnisse führen. Fehlende Motivation oder schlichte Langeweile können die Benutzer dazu verleiten, Aufgaben fehlerhaft oder unreflektiert zu bearbeiten. Schlimmstenfalls bricht die Testperson die Bearbeitung ab und der Fragebogen kann nicht ausgewertet werden. Dies gilt im selben Maße für die einzelnen Teilbereiche eines Fragebogens. Für die Bearbeitung des ALIPR-Fragebogens werden 30 Minuten veranschlagt. Keine der gestellten Teilaufgaben sollte mehr als 10 Minuten in Anspruch nehmen.

4.3.4.3 Benutzerinstruktion

Durch die Anleitung der Testpersonen vor Beginn der Evaluation, soll gewährleistet werden, dass die Befragung ohne Eingreifen des Testleiters durchgeführt werden kann. Jede Intervention seitens des Testleiters während der Befragung, kann die Testergebnisse verfälschen und sollte möglichst vermieden werden. In Bezug auf das Gütekriterium der Objektivität, ist der Ablauf der Befragung ohne Intervention des Testleiters eine Grundvoraussetzung. Die Instruktion kann in zwei Unterbereiche gegliedert werden – eine formelle Begrüßung und eine Einleitung.

Im Begrüßungsteil werden das Thema der Evaluation, die Bearbeitungsdauer und Datenschutzfragen erläutert. Zudem werden die Versuchsteilnehmer darauf hingewiesen, subjektiv und nicht leistungsorientiert zu antworten.

> „The main exception from the rule that users should not be helped is when the user is clearly stuck and is getting unhappy with the situation. The experimenter may also decide to help a user who is encountering a problem that has been observed several times before with previous test users." (NIELSEN, 1993, p. 190)

Der Einführungsteil enthält eine detaillierte Beschreibung des Testumfelds und Instruktionen zur Versuchsdurchführung. Der Testaufbau und der Testablauf werden ebenso beschrieben, wie das getestete System an sich. Obwohl der

Kontakt zu den Testpersonen während des Tests vermieden werden soll, wird im Falle von groben Unklarheiten auf den Testleiter verwiesen, um die vollständige Bearbeitung des Fragebogens zu gewährleisten. Der Testleiter soll in diesem Fall entscheiden, ob die Testperson trotz der Intervention in die Datenerhebung aufgenommen werden kann (vgl. NIELSEN, 1993, pp. 188-190).

4.3.4.4 Formulierung der Aufgaben

Für die aufgestellten Hypothesen müssen geeignete Aufgabenstellungen und Fragen gefunden werden. Durch die Formulierung der Aufgaben muss ein Weg gefunden werden, die Benutzer mit realistischen Szenarien zu konfrontieren, die klare Anweisungen enthalten.

> „The basic rule for test tasks is that they should be chosen to be as representative as possible of the uses to which the system will eventually be put in the field." (NIELSEN, 1993, p. 185)

Die Testperson wird dazu angeregt, sich in ein bestimmtes Szenario zu versetzen, bei dem der Einsatz eines Systems zur automatischen Beschlagwortung realistisch erscheint. Dadurch soll gewährleistet werden, dass Einflussfaktoren wie die Verängstigung oder die Verwirrung der Testpersonen durch die Aufgabenstellung, abgeschwächt werden.

Für den ALIPR-Fragebogen wird ein Szenario mit journalistischem Hintergrund gewählt. Der Benutzer muss sich in die Rolle eines Bildjournalisten versetzen, der verschiedene Aufgaben im Zusammenhang mit seiner Bildsammlung bewältigen muss. Für jede Teilaufgabe werden bestimmte Tätigkeiten aus dem Bereich der Bildannotation, beziehungsweise des Bildretrieval, mit diesem Szenario verbunden. Der Benutzer muss bei der ersten Teilaufgabe vier Bilder aus seiner Bildsammlung manuell mit Schlagwörtern versehen und im zweiten Abschnitt mit denselben Bildern eine automatische Beschlagwortung mit Hilfe des ALIPR Systems durchführen.

4.3.4.5 Verwendete Bilder

Die Auswahl der Bilder für den ALIPR-Fragebogen ist ein zentraler Punkt bei der Erstellung der Befragung. Trotz der zufriedenstellenden Annotationsleistung des ALIPR-Systems (vgl. Kapitel 3.2.2), variiert die Qualität der Annotation von Bild zu Bild deutlich. Um Akzeptanzergebnisse zu erzielen, die unabhängig von der Retrievalleistung sind, müssten Bilder mit identischer Annotationsqualität gewählt werden. Eine objektive Einschätzung der Annotationsqualität kann jedoch nicht getroffen werden, da die Einteilung der vorgeschlagenen Tags in relevante

Abbildung 23: Getty Themenkategorien (vgl. GETTY IMAGES, 2009)

und nicht-relevante erst durch die Benutzer erfolgt. Daher stehen zwei Alternativen für die Auswahl von Bildern zur Verfügung:

(1) Bilder mit identischer Annotationsqualität werden gewählt; die Einteilung der relevanten und nicht-relevanten Schlagwörter erfolgt vor Beginn der Befragung durch den Evaluationsleiter, der sich durch die Auseinandersetzung mit Theorien zur automatischen Annotation als Annotations-Experte qualifiziert

(2) Die Bilder werden, unabhängig von ihrer Annotationsqualität, anhand anderer Kriterien ausgewählt und die Akzeptanzergebnisse in Abhängigkeit zur Annotationsqualität betrachtet

Li und Wang beschreiben in ihren Untersuchungen zur Annotationsqualität des ALIPR-Systems die kritische Auseinandersetzung der Benutzer mit den vom System vorgeschlagenen Tags (vgl. Kapitel 3.2.2.4). Da das Problem der subjektiven Einteilung von relevanten und nicht-relevanten Schlagwörtern durch die erste Variante der Bildauswahl nicht vollständig gelöst werden kann und zudem die Gefahr besteht, dass der Annotations-Experte wegen der Fokussierung auf das Thema keine repräsentative Auswahl treffen kann, fällt die Entscheidung auf die letztere Variante.

Als entscheidendes Kriterium zur Auswahl von Bildbeispielen wurde ihre thematische Zuordnung gewählt. Durch die Analyse von Methoden zur Anreicherung von Bildern mit Metadaten wie dem IPTC-NAA Standard (vgl. IPTC, 2008), aber auch den Ordnungssystemen und Klassifizierungsverfahren von verschiedenen Bilddatenbanken im Netz, wie beispielsweise Getty Images oder Flickr (vgl. GETTY IMAGES, 2009; YAHOO!, 2009), werden geeignete Themen

identifiziert und ausgewählt. Abbildung 23 zeigt das sehr fein granulierte Kategoriensystem der Getty Bilddatenbank.

Aus Kategorien mit ähnlichem Abstraktionsgrad werden vier repräsentative Kategorien ausgewählt:

a) Tiere

b) Architektur

c) Menschen

d) Pflanzen

Zudem wird versucht, die Auswahl der Bilder in Abhängigkeit von der jeweiligen Aufnahmesituation (Portrait, Panorama, Gruppenfoto, Nahaufnahme, usw.) zu treffen, um möglichst unterschiedliche Beispiele zu erhalten.

Vor der endgültigen Auswahl von Beispielbildern muss eine geeignete Bilddatenbank gesucht werden, die Bilder kostenfrei für wissenschaftliche Zwecke zur

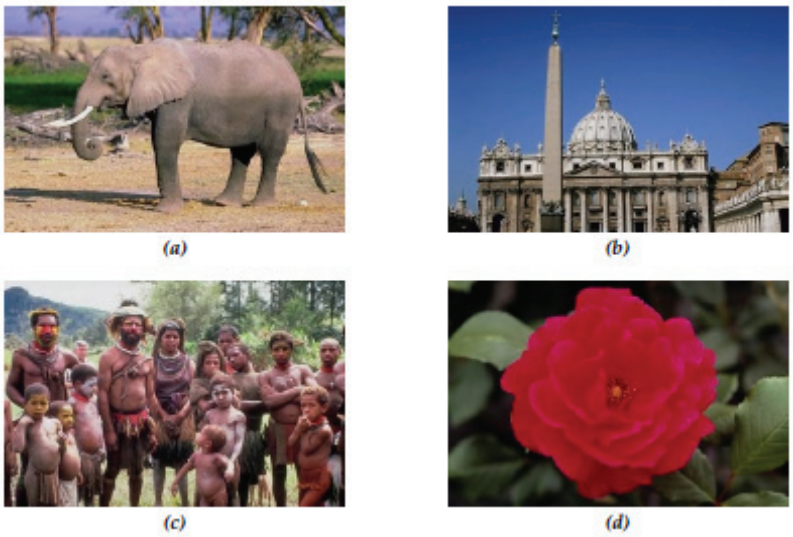

Abbildung 24: Im ALIPR-Fragebogen verwendete Bilder (vgl. Anhang A)

Verfügung stellt. Neben den genannten Bilddatenbanken im Web, wird auch die Verwendung von im Bildretrieval verwendeten Testkollektionen geprüft. Bilddatenbanken im Netz haben den Nachteil, dass die Aufnahmequalität, die Bildauflösung und andere Eigenschaften von Bild zu Bild stark variieren. Viele Testkollektionen, wie das Corel-Testset, stehen nicht zur kostenfreien Verfügung und können aus diesem Grund nicht herangezogen werden. James Z. Wang stellt auf der

Website seiner Forschungsgruppe zwei Bildsets zur Verfügung, um frühere Testergebnisse zu überprüfen (vgl. J. Z. WANG, 2004). Das kleinere der beiden Sets bietet sich aufgrund der Vorsortierung nach bestimmten Kategorien und der identischen Bildeigenschaften an. Zum einen können aus diesem Set für jede gewählte Bildkategorie geeignete Bilder gewonnen werden, zum anderen ist durch die vorhergehende Verwendung der Bilder bei der Evaluation von SIMPLIcity gewährleistet, dass die Bilder bei der inhaltsbasierten Bildsuche verwertbare Ergebnisse liefern.

Um den Umfang der Befragung nicht zu sprengen, ist es erforderlich, pro Kategorie nur ein Bild zu wählen. Die Aufgabenstellung erfordert eine intensive Auseinandersetzung mit dem Inhalt des Bildes und die Bearbeitung nimmt daher für jedes Bild viel Zeit in Anspruch. Im Rahmen von 30 Minuten ist die Bearbeitung von weiteren Bildern nicht realisierbar. Die ausgewählten Beispielbilder sind in Abbildung 24 aufgeführt.

4.3.4.6 Eingesetzte Skalen

Für die Beantwortung der Items können die Testpersonen auf einer sechsstufigen Likert-Skala ihre Meinung ausdrücken. Dabei wird der Grad der Zustimmung in Abstufungen von „stimme voll und ganz zu“ bis „stimme überhaupt nicht zu“ eingeteilt. Die Entscheidung für eine geradzahlige Skala soll die Teilnehmer der Studie dazu zwingen, eine eindeutige Tendenz anzugeben. Beim Einsatz von ungeraden Likert-Skalen kommt es häufig zum Phänomen der zentralen Tendenz, bei dem die Benutzer aus diversen Gründen – zum Beispiel Desinteresse oder hohe kognitive Belastung – dazu neigen, das neutrale Mittel der Skala anzukreuzen (vgl. BORTZ & DÖRING, 2003, p. 184).

> „Geradzahlige Ratingskalen verzichten auf eine neutrale Kategorie und erzwingen damit vom Urteiler ein zumindest tendenziell in eine Richtung weisendes Urteil [...]. Diese Vorgehensweise empfiehlt sich, wenn man mit Verfälschungen der Urteile durch eine übermäßige zentrale Tendenz [...] der Urteiler rechnet.“ (BORTZ & DÖRING, 2003, p. 180)

Die Skalenabstufungen werden durch symbolische Abbildungen, die dem jeweili-

Abbildung 25: Symbolische Beschriftung des Zustimmungsgrades (vgl. Anhang A)

gen Zustimmungsgrad entsprechen, beschriftet. Dadurch können einerseits undeutliche Formulierungen der verschiedenen Zustimmungsgrade vermieden werden. Andererseits werden die Passagen des Fragbogens, die Items enthalten, aufgelockert – das heißt, sie erscheinen den Testpersonen nicht so lang (vgl. BORTZ & DÖRING, 2003, p. 177).

4.3.4.7 Zusammenstellung der Fragen

Die Auswahl geeigneter, so genannter Items zur Feststellung der Benutzerakzeptanz muss wohl überlegt sein. Um die Validität eines Fragebogens zu gewährleisten, müssen die Fragen dem abgefragten Faktor entsprechend formuliert werden. Wird dies nicht erreicht, ist die Messung eines Effekts nicht möglich oder fehlerhaft. Standardisierte Fragebögen zur Messung von Usability, wie TAM, UTAUT, QUIS (vgl. NORMAN & SHNEIDERMAN, 2002) oder SUMI (vgl. HUMAN FACTORS RESEARCH GROUP, 2007), verweisen ausgiebig auf die Validität der eingesetzten Items zur Messung bestimmter Faktoren (vgl. SARODNICK & BRAU, 2006, pp. 175-181). Da die Prüfung auf Validität bei selbst formulierten Fragen einen nicht zu bewältigenden Mehraufwand bedeuten würde – die Eichung von Items muss mit sehr großen Stichproben (N > 300) durchgeführt werden (vgl. SARODNICK & BRAU, 2006, p. 170) – wird bei der Zusammenstellung des ALIPR-Fragebogens auf Items aus standardisierten Fragebögen zurückgegriffen (vgl. CHOI & RASMUSSEN, 2002; DAVIS ET AL., 1989; vgl. MCGILL & HOBBS, 2008; SPINK, 2002; VENKATESH & MORRIS, 2003).

Die endgültige Formulierung wird, sofern sie nicht bereits eindeutig zutrifft, dem Thema Bildretrieval entsprechend angepasst. Besondere Beachtung finden dabei Items, die aufgrund ihrer Einordnung in den standardisierten Fragebogen, eindeutig den Hauptfaktoren der eigenen Studie zugeordnet werden können (vgl. Kapitel 4.2). Zusätzlich zu den geschlossenen Fragen wird am Ende der Befragung noch eine Reihe von offenen Fragen gestellt, um den Benutzern die Möglichkeit eines persönlichen Feedbacks zu geben.

> „After the test, the user should be debriefed and allowed to make comments about the system.“ (NIELSEN, 1993, p. 184)

Die Abfrage von demografischen Daten beschränkt sich auf die Ermittlung des Alters, des Geschlechts und des Berufs. Alle Fragebogenitems können der finalen Fassung des Tests im Anhang entnommen werden (vgl. Anhang A)

4.3.5 Stichprobenkonstruktion

Die korrekte Auswahl von Testpersonen und die Stichprobengröße sind im Bereich der Usability-Evaluation ein viel diskutiertes Feld. Die Formulierung der Hypothesen und die Festlegung des Untersuchungsgegenstandes bestimmen diese Variablen bereits wesentlich. Zudem fließen organisatorische, zeitliche und finanzielle Beschränkungen in die Auswahl der Testpersonen und die Festlegung der Stichprobengröße mit ein (vgl. BORTZ & DÖRING, 2003, pp. 602-604).

Verschiedene Varianten der Stichprobenerhebung stellen Bortz und Döring folgendermaßen dar:

> „Um mit Hilfe einer Stichprobenerhebung (anstelle einer Vollerhebung) gültige Aussagen über eine Population treffen zu können, muss die Stichprobe repräsentativ sein, d. h., sie muss in ihrer Zusammensetzung der Population möglichst stark ähneln. Eine Stichprobe ist (merkmals)spezifisch repräsentativ, wenn ihre Zusammensetzung hinsichtlich einiger relevanter Merkmale der Populationszusammensetzung entspricht. Sie ist global repräsentativ, wenn ihre Zusammensetzung in nahezu allen Merkmalen der Populationszusammensetzung entspricht." (BORTZ & DÖRING, 2003, pp. 397-398)

Lässt man zeitliche und organisatorische Beschränkungen unbeachtet, soll die Untersuchung des ALIPR-Systems den Vergleich zwischen zwei sehr spezifischen Anfragemethoden behandeln. In diesem Fall besteht kein Anspruch auf die repräsentative Darstellung der Ergebnisse. Allein durch die spezifische Anwendungssituation kann davon ausgegangen werden, dass Großteile der Gesamtpopulation ein System wie ALIPR nicht im Alltag verwenden würden. Aus diesem Grund werden die Testpersonen für die Stichprobe aus einem Personenkreis gewonnen, der das System potenziell auch anwenden würde. Um aufwändige Verfahren zur Identifikation der Benutzertauglichkeit während des Tests zu vermeiden, wird vor der Befragung festgelegt, welche Kriterien die Testpersonen erfüllen müssen, um für die Befragung zugelassen zu werden. Nielsen stützt diese Vorgehensweise:

> „The main rule regarding test users is that they should be as representative as possible of the intended users of the system." (NIELSEN, 1993, p. 175)

Dabei stehen vor allem fachliche Vorkenntnisse im Vordergrund. Da das untersuchte System eine Webschnittstelle hat, sollen die Versuchsteilnehmer auf dem Feld der Informationstechnologie bereits Erfahrungen gesammelt haben.

Hinsichtlich der Stichprobengröße muss ein ausgewogener Kompromiss zwischen dem zur Befragung nötigen Aufwand und dem daraus gezogenen Nutzen gefunden werden. Nielsen stellt in seinem Buch 'Usability engineering' verschiedene Thesen über die Stichprobengröße bei Untersuchungen in Abhängigkeit zum erzielten Mehrwert auf. Abbildung 26 zeigt Nielsens Kosten-Nutzen-Rechnung, die in ihren Grundzügen auch auf den organisatorischen und zeitlichen Aufwand übertragen werden kann (vgl. NIELSEN, 1993, pp. 175-179).

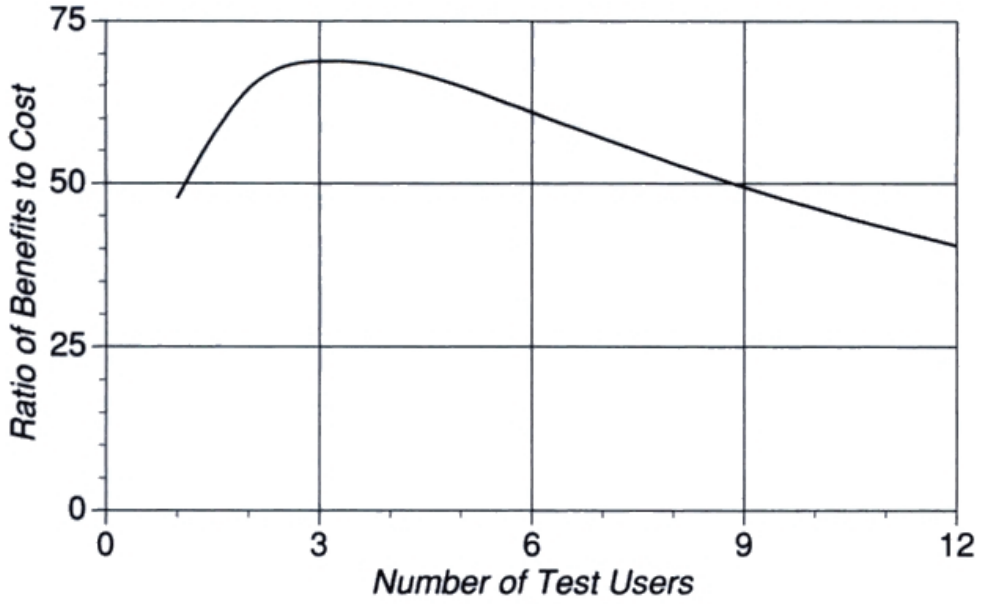

Abbildung 26: Nielsens Kosten-Nutzen-Diagramm bezüglich der Stichprobengröße (NIELSEN, 1993, p. 174)

Nach der Abwägung der genannten Begebenheiten wird für die Evaluation des ALIPR-Systems eine Stichprobengröße von mindestens 20 Personen angestrebt. Dabei sollen die Teilnehmer informationstechnisches Hintergrundwissen vorweisen und somit potentielle Nutzer des Systems sein. Objektive Kriterien, welche die Eignung der Testpersonen für die Untersuchung bestätigen, sind beispielsweise die berufliche Tätigkeit oder der regelmäßige Einsatz von bildbasierten Informationstechnologien. Die endgültige Auswahl der Probanden anhand dieser Kriterien erfolgt durch den Versuchsleiter.

4.3.6 Untersuchungsdurchführung

Die Durchführung der Untersuchung muss vor dem Beginn der ersten Tests geplant und strukturiert werden. Dadurch kann sich der Versuchsleiter bereits auf etwaige Probleme vorbereiten und Irritationen seitens der Testpersonen von Beginn an vermeiden. Zu dieser Phase zählt einerseits die Vorbereitung der Unterlagen, des Arbeitsplatzes und der Software, aber auch das Einstudieren von eventuellen Antworten oder Reaktionen gegenüber den Testpersonen (vgl. BORTZ & DÖRING, 2003, pp. 130-132). Neben der reibungslosen Durchführung des Tests, sollte vor allem auf die Bereitstellung identischer Bedingungen für alle Versuchsteilnehmer geachtet werden. Nur dadurch kann gewährleistet werden, dass die erzielten Ergebnisse untereinander vergleichbar sind. Ein zentrales Problem in Bezug auf dieses Thema, stellen so genannte Versuchsleiterartefakte dar, die so weit als möglich vermieden werden müssen (vgl. BORTZ & DÖRING, 2003, pp. 82-85)

4.3.6.1 Pretest

Häufig treten während der Durchführung von Benutzertests auch Probleme auf, die nicht vom Versuchsleiter vorhergesehen werden können. Dies kann dazu führen, dass die Versuchsergebnisse massiv verfälscht werden und die Studie unbrauchbar wird (vgl. COLLINS, 2003). Abbildung 27 zeigt gängige Probleme, die während der Durchführung einer Untersuchung bei den Testpersonen auftreten können. Ein übliches Verfahren dies zu vermeiden, ist die Durchführung eines Pretests, bei dem etwaige Probleme identifiziert werden können (vgl. KROMREY, 2006, pp. 359-360). Da sich der erstellte Fragebogen, wie bereits erwähnt, aus

Traditional model	Task-focused model
(1) Problems with survey questions: – that are misunderstood – that cannot be answered, either at all or accurately – that respondents will not answer	(1) Comprehension problems resulting from: – use of vocabulary – complex sentence structure – not understanding the nature of the task and the rules about how to respond
(2) Problems with survey interviewers: – do not read the questions as worded – probe directively – bias answers as a result of the way interviewers relate to respondents (for example, differences in ethnicity, age, social class, gender) – record answers inaccurately	(2) Validity problems resulting from: – respondents interpreting the same question in different ways, or – in the same way but not in the way the researcher intended
	(3) Processing difficulties: – respondents may be unwilling or unable to retrieve the information necessary to answer the question
	(4) Pronunciation or communication difficulties: – these may affect both interviewers and respondents

Abbildung 27: Häufig auftretende Probleme während der Durchführung einer Evaluation (vgl. Collins, 2003, p. 230)

angepassten Items anderer Akzeptanzuntersuchungen zusammensetzt und in der Praxis bisher nicht verwendet wurde, liegt die Durchführung eines Pretests nahe.

4.3.6.2 Testumgebung

Traditionell werden Usability-Tests in speziell dafür ausgestatteten Labors durchgeführt. Die Testpersonen befinden sich dadurch in einem kontrollierten Umfeld und können sich auf die Bearbeitung der gestellten Aufgaben konzentrieren (vgl. KAIKKONEN ET AL., 2005, p. 5). Darüber hinaus wird durch die Durchführung der Benutzertests in einem Usability-Labor gewährleistet, dass alle Testpersonen unter identischen Bedingungen arbeiten.

Bei der geplanten Untersuchung finden sich keine triftigen Argumente für die Durchführung eines Feldtests – das heißt, für die Durchführung eines Experiments im wirklichen Arbeitsumfeld der Benutzer. Aufgrund der unkontrollierten Bedingungen und des erheblichen Mehraufwands von Feldstudien, ist die Durchführung von Labortests die geeignete Alternative. Der Lehrstuhl für Informationswissenschaft stellt seit 2007 ein Usability-Labor zur Verfügung, das die Voraussetzungen zur Durchführung der ALIPR-Befragung hinreichend erfüllt.

4.3.7 Datenanalyse

Um eine detaillierte Datenaufbereitung zu garantieren, wird das Statistikprogramm SPSS zur Analyse der quantitativ erhobenen Daten herangezogen (vgl. SPSS INC., 2009). Dabei soll vor allem auf die Methoden zur deskriptiven Statistik und zur Prüfung von Zusammenhangshypothesen zurückgegriffen werden.

Bei den übrigen Daten, wie beispielsweise den offenen Fragen am Ende des Fragebogens, sowie den Aufzeichnungen und Protokollen während der Testphase, wird die intellektuelle Analyse im Vordergrund stehen. Die Auswertung der manuell vergebenen Tags erfolgt mittels des Tabellenkalkulationsprogramms Microsoft Excel. Das Rechenzentrum der Universität Regensburg stellt die aufgeführte Software für wissenschaftliche Zwecke kostenlos zur Verfügung (vgl. RZ UNI REGENSBURG, 2008).

4.4 Ergebnisse

Die Auswertung der Ergebnisse kann in mehrere Bereiche aufgeteilt werden. Dazu zählen die Analyse des Pretests, die Stichprobenbeschreibung, die Auswertung der Kernhypothesen, sowie die Analyse der Nebenfragestellungen.

4.4.1 Ergebnisse des Pretests

Um den reibungslosen Ablauf der eigentlichen Tests zu gewährleisten, werden vor Beginn der Testphase mehrere Pretests durchgeführt. Dabei kommen Testpersonen zum Einsatz, die Erfahrung sowohl auf dem Bereich der quantitativen Evaluation, als auch im Zusammenhang mit Usability-Engineering vorweisen können. Die Durchführung dieser Tests trägt nachhaltig zur Verbesserung der Testdurchführung, aber auch des Fragebogens an sich bei. Vor allem können irreführende oder undeutliche Formulierungen der Fragebogenitems identifiziert werden.

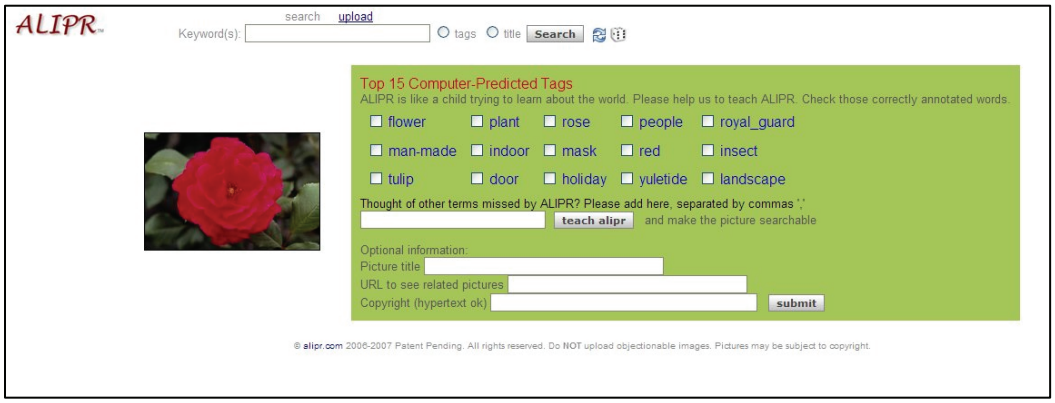

Abbildung 28: Ausgewählter Einstiegspunkt für die Testpersonen (vgl. ALIPR, 2009)

Für die spätere Durchführung der Befragung können wertvolle Erkenntnisse über den optimalen Umgang mit Benutzerfragen während des Tests gewonnen werden. Durch die Ermittlung von irritierenden Formulierungen oder unklaren Aufgabenstellungen kann der Fragebogen dahingehend verbessert werden, dass eine Intervention des Versuchsleiters weitestgehend vermieden wird. Zu den Änderungen zählen unter anderem formelle Abwandlungen bei der schriftlichen Benutzerinstruktion und eine verbesserte Strukturierung der Instruktion seitens des Versuchsleiters. Auffällig ist die hohe Ablehnung der Pretest-Teilnehmer gegenüber der Formulierung der manuell vergebenen Schlagwörter in englischer Sprache. Ursprünglich war angedacht, die manuelle Beschlagwortung ausschließlich in englischer Sprache durchzuführen, um einen präzisen Vergleich zu den automatisch zugeordneten Tags von ALIPR ziehen zu können. Trotz der Gefahr der falschen Disambiguierung bei der Übersetzung, wird den Testpersonen für die Benutzertests freigestellt die Schlagwörter in englischer Sprache oder in ihrer Muttersprache zu formulieren.

Zusätzlich wird die Interaktion der Testpersonen mit dem ALIPR-System erst beim Schritt zur Auswahl der Annotationsvorschläge gestartet. Dadurch

kann einerseits Zeit für die Bearbeitung der Teilaufgaben gewonnen und andererseits verhindert werden, dass mögliche Probleme mit der Bedienung der ALIPR-Schnittstelle Einfluss auf die Benutzerakzeptanz haben. Die Bedienung des ALIPR-Systems startet mit dem in Abbildung 28 dargestellten Interaktionsschritt und beschränkt sich somit auf einige wenige Eingaben.

4.4.2 Stichprobenbeschreibung

Die Durchführung der Benutzerbefragung erfolgte vom 09. Dezember bis zum 04. Januar 2008 an der Universität Regensburg und fand im Usability-Labor des Lehrstuhls für Informationswissenschaft statt. Jeder Testteilnehmer wurde separat getestet. Die angestrebte Bearbeitungszeit von 30 Minuten wurde bei keinem der Probanden überschritten.

Die erhobene Stichprobe umfasst insgesamt 24 Testpersonen – davon sind 13 Personen männlich und 11 Personen weiblich. Das Durchschnittsalter liegt bei 25 Jahren. Der älteste Teilnehmer der Befragung ist 32 Jahre, der jüngste 18 Jahre alt. Die Altersverteilung der Testpersonen weist eindeutig auf das universitäre Umfeld der Befragung hin. Aufgrund der Ausführungen zur Stichprobenkonstruktion ist diese Verteilung jedoch unbedenklich (vgl. Kapitel 4.3.5)

Alle Testpersonen arbeiten oder studieren im Bereich der Informationstechnologie und erfüllen damit die Zulassungsvoraussetzung für den Benutzertest. Zudem geben alle Teilnehmer an, vorher noch nicht mit dem ALIPR-System gearbeitet zu haben. Dadurch ist gewährleistet, dass die Akzeptanzbeurteilung der Benutzer gegenüber den Retrievalformen unabhängig von vorhergehenden Erfahrungen ist und vollkommen aus den Eindrücken der Testphase resultiert.

Die Auswertung der quantitativen Daten unterteilt sich in mehrere Phasen. Zu Beginn stehen die Kernhypothesen im Fokus der Auswertung. Im Anschluss werden die Nebenbeobachtungen diskutiert.

4.4.3 Ergebnisse zu den Kernhypothesen

Die Ergebnisse der Auswertung der Items zur Akzeptanz der verschiedenen Interaktionsmethoden werden für jeden Aufgabenteil einzeln analysiert. Dabei werden vor allem die positiv, beziehungsweise negativ auffälligen Items hervorgehoben. Im Anschluss folgen die Bewertung der Gesamtfaktoren und die Einordnung der aufgestellten Hypothesen.

Die Größe der Stichprobe und die explorative Analyse der erhobenen Daten lassen den Rückschluss zu, dass bei der Evaluation keine Normalverteilung vor-

liegt. Daher erfolgt die statistische Auswertung mittels deskriptiver Statistiken und nicht-parametrischer Tests, wie dem Wilcoxon Vorzeichenrangtest für abhängige Stichproben (vgl. DANCEY & REIDY, 2004, pp. 523-530)

4.4.3.1 Bewertung der automatischen Schlagwortvergabe

Die Testpersonen bewerten den Interaktionsschritt zur Auswahl der automatisch generierten Schlagwörter fast ausschließlich positiv. Dabei werden die vorgeschlagenen Wörter als besonders verständlich, leicht unterscheidbar, in der Menge ausreichend, gut einzuordnen, inspirierend und zutreffend für den Bildinhalt eingestuft. Zudem werden die Zeitersparnis und die Erleichterung bei der Beschlagwortung durch den Einsatz des Annotationssystems positiv bewertet. Negativ wird lediglich die Unterscheidungsfähigkeit der Schlagwörter bewertet – das heißt, die Benutzer glauben nicht, dass die Schlagwörter zur Unterscheidung des Bildes im Vergleich zu anderen Bildern beitragen.

Neutral beurteilen die Testpersonen den Nutzen des Systems für die eigene Arbeit, sowie die Übereinstimmung der generierten Schlagwörter mit der eigenen Wahl. Die Erfassung des gesamten Bildinhalts durch die Schlagwörter wird ebenfalls neutral gesehen. Diese Fragebogenitems weisen keine signifikante Abweichung in eine Richtung auf.

	Anzahl	Signifikanz	Mittelwert
SW sind verständlich (SS)	24	,000	5,3750
SW leicht in richtig/falsch unterscheidbar (SS)	24	,000	5,2083
System erspart mir Zeit (SS)	24	,000	4,8333
System ist Erleichterung (SS)	24	,000	4,7083
SW-Menge ist ausreichend (SN)	24	,001	4,5000
SW kann man dem Thema zuordnen (SN)	24	,001	4,4167
System würde ich auch benutzen (SN)	24	,021	4,2500
SW inspirieren zu neuen manuellen SW (SS)	24	,002	4,1250
SW zutreffend für das Bild (SN)	24	,022	4,0417
System ist nützlich für eigene Arbeit (SN)	24	,681	3,6250
SW entsprechen der Wahl der Benutzer (SS)	24	,582	3,6250
SW beschreiben alle Bestandteile des Bildes (SS)	24	1,000	3,5000
SW erleichtern Unterscheidung des Bildes zu anderen (SN)	24	,017	2,8750
subjektive Nützlichkeit (SN – 6 Items)	24	,131	4,2600
subjektive Schwierigkeit (SS – 7 Items)	24	,013	4,2142
Gesamtfaktor Akzeptanz (SN und SS – 13 Items)	24	,003	4,2400

Tabelle 1: Mittelwerte der Items zur Akzeptanz der automatischen Beschlagwortung

Hinsichtlich der Faktoren zur subjektiven Nützlichkeit und zur subjektiven Schwierigkeit werden vor allem letztere sehr gut bewertet. Fünf der sieben positiven Items können dem Faktor der subjektiv wahrgenommenen Schwierigkeit zugeordnet werden. Die Testpersonen empfinden die automatische Beschlagwortung also in erster Linie als Erleichterung. Vergleicht man die Mittelwerte der beiden errechneten Faktoren, wird zwar auch die Nützlichkeit des Interaktionsschritts sehr gut bewertet – ein signifikanter Effekt konnte jedoch nicht festgestellt werden.

> **Hypothese 1:** Die Akzeptanz der Benutzer gegenüber dem Interaktionsschritt zur automatischen Schlagwortvergabe ist hoch.

Darauf aufbauend kann postuliert werden, dass die Akzeptanz der Benutzer gegenüber dem Interaktionsschritt der automatischen Beschlagwortung sich auf

einem signifikant hohen Niveau befindet. Tabelle 1 schlüsselt das Ergebnis in seine Teilfaktoren und in die einzelnen Items auf.

4.4.3.2 Auswertung der Retrievaleffizienz

Die Retrievaleffizienz wird durch die Berechnung der Precision der ersten zehn Ergebnisse bewertet. Dabei fließt das Ranking der Ergebnisse nicht in die Analyse mit ein, da durch die unterschiedliche Auswahl der automatisch generierten Schlagwörter keine homogenen Ergebnismengen entstehen.

	N	Spannweite	Precision in Prozent
Anzahl relevanter Bilder TAG 1a	24	9	**87,9 %**
Anzahl relevanter Bilder TAG 1b	24	10	**52,9 %**
Anzahl relevanter Bilder TAG 1c	24	8	**13,3 %**
Anzahl relevanter Bilder TAG 1d	24	9	**82,1 %**
Anzahl relevanter Bilder IB 1a	24	3	**14,2 %**
Anzahl relevanter Bilder IB 1b	24	8	**38,8 %**
Anzahl relevanter Bilder IB 1c	24	4	**20,8 %**
Anzahl relevanter Bilder IB 1d	24	5	**80,0 %**
Precision TAG Gesamt	24	10	**59,0 %**
Precision IB Gesamt	24	8	**38,5 %**

Tabelle 2: Precision der ersten 10 Ergebnisse

Generell variiert die Bewertung der Retrievalergebnisse je nach Bildbeispiel sehr stark. Die Spannweite der Bewertungen ist bei fünf von acht Beispielen größer als 7. Dabei fällt auf, dass alle vier Beispiele der tagbasierten Retrievalmethode eine solch große Spannweite aufweisen.

> **Hypothese 2:** Die durch die tagbasierte Anfragemethode gelieferten Ergebnisse sind signifikant besser als die der inhaltsbasierten Anfragemethode

Vergleicht man die durchschnittlichen Precision-Werte, schneiden beim tagbasierten Retrieval drei Beispielbilder im Vergleich zum bildbasierten Retrieval besser ab. Besonders deutlich wird dieser Unterschied bei Bild 1a, bei dem die tagbasierte Methode eine Precision von 87,9 Prozent und die inhaltsbasierte Methode nur einen Wert von 14,2 Prozent erreicht. Die inhaltsbasierte Retrievalmethode erreicht lediglich bei Bild 1c eine höhere Precision als das tagbasierte Ret-

rieval. Dabei fällt das Ergebnis bei beiden Methoden mit 13,3 Prozent und 20,8 Prozent unterdurchschnittlich aus. Bei Bild 1d sind die Ergebnisse beider Methoden annähernd gleich und mit 82,1 Prozent und 80 Prozent auch auf einem hohen Niveau. Einen genauen Überblick über die Streuung und die Precision der Bildbeispiele bietet Tabelle 2.

Betrachtet man die Gesamt-Precision der beiden Methoden, schneidet die tagbasierte Methode deutlich besser ab, als die inhaltsbasierte Methode. Ein Wert von 59,0 Prozent steht einem Wert von 38,5 Prozent gegenüber. Die statistische Überprüfung, ergibt in Bezug auf die Gesamt-Precision keinen signifikanten Unterschied der Ergebnismengen. Die aufgestellte Hypothese kann nicht bestätigt werden.

4.4.3.3 Auswertung von tag- und inhaltsbasiertem Retrieval

Methode ist effizient (SN)	24	**4,63**	,71094
Methode ist eine Erleichterung bei der Bildsuche (SS)	24	**5,00**	,78019
Ergebnisse der tagbasierten Suche sind relevant (SN)	24	**4,79**	,83297
Methode liefert Ergebnisse in angemessener Zeit (SN)	24	**5,79**	,50898
Ergebnisse sind leicht in richtig/falsch einzuteilen (SS)	24	**5,13**	,85019
Ergebnisse unterscheiden sich inhaltlich (SS)	24	**4,54**	,88363
Ergebnisse liefern neue Erkenntnisse zum Thema (SN)	24	**3,79**	1,25036
Methode liefert das was VPN erwartet haben (SS)	24	**4,29**	,75060
Methode ist für Arbeit von Nutzen (SN)	24	**5,17**	,63702
subjektive Nützlichkeit (SN – 5 Items)	24	**4,66**	-
subjektive Schwierigkeit (SS – 4 Items)	24	**4,96**	-
Gesamtfaktor Akzeptanz (SN und SS – 9 Items)	24	**4,79**	-

Tabelle 3: Mittelwerte zur Akzeptanzbewertung der tagbasierten Methode

Für die Bewertung der beiden Retrievalmethoden wurden neun identische Fragebogenitems herangezogen. Betrachtet man die Mittelwerte der beiden Methoden im Einzelnen, fällt bei der Akzeptanzbewertung der tagbasierten Methode auf, dass die Items durchweg positiv bewertet werden. Dabei können signifikante Effekte bei sieben von neun Items festgestellt werden. Der Mittelwert des Erkenntnisgewinns durch die gelieferten Ergebnisse übersteigt mit 3,72 nur knapp den Median der Stichprobe. Besonders positiv werden die Schnelligkeit des Systems, der Nutzen für die Arbeit und die leichte Unterscheidbarkeit der Ergebnisse in richtige und falsche bewertet. Obwohl die Erleichterung bei der Suche ebenfalls sehr positiv bewertet wird, ist dieses Item – neben dem Item zur Bewertung des Erkenntnisgewinns – das einzige, das keinen signifikanten Effekt aufweist.

Für die Bewertung der beiden Retrievalmethoden wurden neun identische Fragebogenitems benutzt. Betrachtet man die Mittelwerte der beiden Methoden im Einzelnen, fällt bei der Akzeptanzbewertung der tagbasierten Methode auf, dass die Items überwiegend als positiv bewertet werden. Die subjektive Schwierigkeit des Systems wird mit einem Mittelwert von 4,96 etwas besser bewertet, als die subjektive Nützlichkeit mit 4,66. Der Gesamtfaktor wird mit einem Mittelwert von 4,79 bewertet. Die Akzeptanz gegenüber dem tagbasierten Retrieval kann daher positiv gesehen werden.

Die inhaltsbasierte Methode schneidet in zwei von neun Fällen unterdurchschnittlich ab. Die restlichen sieben Items werden positiv bewertet. Die Schnelligkeit des Systems, sowie die Einteilung der Ergebnisse in richtige und falsche, werden signifikant positiv bewertet. Die restlichen Items weisen weder eine gerichtet signifikante, noch eine ungerichtet signifikante Tendenz auf.

Analog zur tagbasierten Methode wird die subjektive Schwierigkeit besser eingeschätzt als die subjektive Nützlichkeit. Auch hier ist der Unterschied der beiden Mittelwerte mit 0,23 nur sehr gering. Der Akzeptanzwert gegenüber dem inhaltsbasierten Retrieval ist mit dem Mittelwert 4,09 ebenfalls positiv. Die berechneten Mittelwerte und die Standardabweichung der Items können in Tabelle 3 und 4 jeweils im Detail betrachtet werden.

	Anzahl	Mittelwert	Std.abweichung
Methode ist effizient (SN)	24	**3,33**	1,37261
Methode ist eine Erleichterung bei der Bildsuche (SS)	24	**3,63**	1,37722
Ergebnisse der tagbasierten Suche sind relevant (SN)	24	**3,71**	1,23285
Methode liefert Ergebnisse in angemessener Zeit (SN)	24	**5,46**	,65801
Ergebnisse sind leicht in richtig/falsch einzuteilen (SS)	24	**5,13**	,89988
Ergebnisse unterscheiden sich inhaltlich (SS)	24	**4,17**	1,20386
Ergebnisse liefern neue Erkenntnisse zum Thema (SN)	24	**4,08**	1,17646
Methode liefert das was VPN erwartet haben (SS)	24	**3,38**	1,17260
Methode ist für Arbeit von Nutzen (SN)	24	**3,96**	1,39811
subjektive Nützlichkeit (SN – 5 Items)	24	**3,99**	-
subjektive Schwierigkeit (SS – 4 Items)	24	**4,22**	-
Gesamtfaktor Akzeptanz (SN und SS – 9 Items)	24	**4,09**	-

Tabelle 4: Mittelwerte zur Akzeptanzbewertung der inhaltsbasierten Methode

4.4.3.4 Akzeptanzvergleich von tag- und bildbasierter Suche

Im Vergleich der beiden Retrievalmethoden schneidet die inhaltsbasierte Retrievalmethode deutlich schlechter ab. Sieben Items haben bei der tagbasierten Methode einen höheren Mittelwert erreicht. Die Einteilung der Ergebnisse in richtige und falsche wird bei beiden Verfahren identisch, der Erkenntnisgewinn durch die Ergebnisse beim inhaltsbasierten Retrieval sogar höher bewertet.

	Signifikanz TAG vs. IB	Mittelwert TAG	Mittelwert IB
Methode ist effizient (SN)	**,001**	4,63	3,33
Methode ist eine Erleichterung bei der Bildsuche (SS)	**,000**	5,00	3,63
Ergebnisse der tagbasierten suche sind relevant (SN)	**,003**	4,79	3,71
Methode liefert Ergebnisse in angemessener Zeit (SN)	**,021**	5,79	5,46
Ergebnisse sind leicht in richtig/falsch einzuteilen (SS)	**1,000**	5,13	5,13
Ergebnisse unterscheiden sich inhaltlich (SS)	**,116**	4,54	4,17
Ergebnisse liefern neue Erkenntnisse zum Thema (SN)	**,175**	3,79	4,08
Methode liefert das was VPN erwartet haben (SS)	**,002**	4,29	3,38
Methode ist für Arbeit von Nutzen (SN)	**,000**	5,17	3,96
subjektive Nützlichkeit (SN – 5 Items)	**,000**	4,66	3,99
subjektive Schwierigkeit (SS – 4 Items)	**,001**	4,96	4,22
Gesamtfaktor Akzeptanz (SN und SS – 9 Items)	**,000**	4,79	4,09

Tabelle 5: Vergleich von tagbasierter (TAG) und inhaltsbasierter (IB) Methode

Hypothese 3: Die Akzeptanz der Benutzer gegenüber der tagbasierten Methode ist signifikant höher als die Akzeptanz gegenüber der inhaltsbasierten Anfragemethode.

Die statistische Auswertung zeigt, dass signifikant höhere Ergebnisse einzig von der tagbasierten Variante erzielt werden. Sechs Items werden von den Testpersonen signifikant besser bewertet. Die Teilfaktoren Nützlichkeit und Schwierigkeit und der Gesamtfaktor Akzeptanz werden ebenso signifikant besser bewertet. Damit lässt sich eine eindeutige Tendenz der Testpersonen zum tagbasierten Retrievalmodus feststellen. Die aufgestellte Hypothese trifft zu.

4.4.4 Weitere Befunde

Neben den durch Signifikanztests überprüften Kernhypothesen, enthält der Fragebogen eine Reihe von zusätzlichen Fragestellungen. Durch die Auswertung dieser – größtenteils offenen – Fragen und die Überprüfung der Testprotokolle können weitere Thesen formuliert werden.

4.4.4.1 Analyse der manuell vergebenen Schlagwörter

Vor der Auswahl der von ALIPR vorgeschlagenen Tags wurden alle Benutzer aufgefordert, für jedes Beispielbild manuell Schlagwörter zu vergeben und aufzuschreiben.

Die Analyse der manuellen Auswahl liefert einen präziseren Einblick in die Auswahlkriterien der Benutzer. Gerade der Vergleich der manuellen mit der automatischen Auswahl verspricht interessante Ergebnisse. Die die Auswahl der ALIPR-Schlagwörter wurde vom Versuchsleiter protokolliert und kann nun in Tabelle 6 zum Vergleich herangezogen werden.

	Anzahl Tags Manuell	Anzahl Tags automatisch	Übereinstimmungen manuell-automatisch
Bild 1a	2,35	3,35	1,43
Bild 1b	3,13	3,00	0,61
Bild 1c	2,83	1,65	0,22
Bild 1d	2,91	3,43	2,30
Gesamt	2,81	2,86	1,14

Tabelle 6: Durchschnitt der zugeteilten Schlagwörter und Übereinstimmungen

Die Menge der zugeteilten Schlagwörter variiert von Bild zu Bild sehr stark. Bei Bild 1d werden lediglich 13 verschiedene Wörter von den Testpersonen zugeteilt. Für Bild 1b werden dagegen mit 27 verschiedenen Tags mehr als doppelt so viele vergeben. Man kann annehmen, dass der Bildinhalt bei Bild 1d eindeutiger durch Wörter zu beschreiben ist, als der von Bild 1b. Durchschnittlich werden 20,5 verschiedene Wörter pro Bild vergeben. Es besteht also eine höhere Vielfalt bei der Vergabe der manuellen Schlagwörter.

Der Vergleich der durchschnittlich zugeteilten Tags bei der manuellen und der automatischen Beschlagwortung ergibt kein klares Bild. Zwar können bei den einzelnen Beispielbildern geringe Unterschiede festgestellt werden, der Gesamtdurchschnitt ist aber beinahe identisch. Die Zuteilung von ungefähr drei Schlagwörtern pro Bild scheint für die Benutzer bei beiden Methoden ausreichend zu sein, um ein Bild zu beschreiben. Vergleicht man diese Werte mit den Untersuchungen von Li und Wang – 2,24 ALIPR-Schlagwörter werden pro Bild ausgewählt – liegt die Anzahl der ausgewählten ALIPR-Tags leicht über dem Durchschnitt. Im Schnitt stimmen 1,14 Wörter der manuellen Beschlagwortung mit der automatischen Beschlagwortung überein. Der Höchstwert der Übereinstimmungen liegt bei vier identischen Schlagwörtern (vgl. Tabelle 7).

4.4.4.2 Auswertung der offenen Fragen

Die Betrachtung der Benutzeranmerkungen deckt sich im Wesentlichen mit den quantitativen Beobachtungen. Beide Retrievalmethoden werden generell positiv bewertet, wobei eine klare Tendenz zum tagbasierten Retrieval festzustellen ist. Besonders hervorgehoben werden die Einfachheit und Schnelligkeit beider Retrievalmethoden sowie die unerwartet hohe Ergebnisqualität.

Bei der bildbasierten Methode wurde mehrmals auf die Besonderheit hingewiesen, dass Bildinhalte, die durch Wörter nur schwer zu beschreiben sind, anhand dieses Verfahrens effektiver gesucht werden können. Bei der tagbasierten Methode wurde neben der hohen Retrievalqualität vor allem die Möglichkeit zur reinen Beschlagwortung durch die Benutzer begrüßt. Der Wunsch nach einer vollautomatischen Beschlagwortung und Kategorisierung von nichtbeschlagworteten Bildsammlungen wurde in diesem Zusammenhang häufig geäußert. Die detaillierten Annotationsvorschläge, wie zum Beispiel ‘elephant‘ bei Bild 1a oder ‘rose‘ bei Bild 1d, wurden von den Testpersonen besonders positiv bewertet. Zudem wurde die Kombination beider Suchmethoden bei der Suche nach relevantem Bildmaterial angeregt.

Manuell vergebene Schlagwörter				Automatische Beschlagwortung			
Bild 1a	**Bild 1b**	**Bild 1c**	**Bild 1d**	**Bild 1a**	**Bild 1b**	**Bild 1c**	**Bild 1d**
(N=19)	**(N=27)**	**(N=23)**	**(N=13)**	**(N=15)**	**(N=15)**	**(N=15)**	**(N=15)**
Africa	church	Australia	**red**	**animal**	**building**	**people**	**plant**
savannah	statue	aboriginies	nature	**wild-life**	historical	building	**flower**
grey	column	tribe	blossom	grass	water	face	**rose**
animal	**landmark**	nature	green	**tree**	**sky**	landscape	people
desert	architecture	inhabitants	**rose**	rock	ocean	indoor	royal-guard
oasis	church	natives	leaves	man-made	man-made	office	man-made
ivory	monument	aboriginies	close-up	antelope	boat	historical	indoor
wild-life	sightseeing	black person	big	lion	people	animal	mask
steppe	cathedral	traditional	pink	**elephant**	ship	wild-life	**red**
holidays	Rome	group	**flower**	indoor	**landmark**	rock	insect
landscape	landmark	family	green	old	beach	male	tulip
jungle	St. Peter	New Guinea	**plant**	dish	flower	city	door
wild	obelisk	wild		reptile	sea	modern	holiday
lonesome	white	jungle		**landscape**	balloon	work	yuletide
elephant	dome	human		lake	landscape	ruin	landscape
big	square	indians					
green	religion	children					
baby	islam	Africa					
tree	mosque	**people**					
	minaret						
	building						
	city						
	Vatican						
	sky						
	blue						
	palace						

Tabelle 7: Zugeteilte bzw. akzeptierte Schlagwörter im Vergleich (Übereinstimmungen wurden hervorgehoben)

Nachteilig wird hauptsächlich der hohe Anteil von irrelevanten Ergebnissen beider Retrievalkonzepte bewertet. Bildergebnisse, die thematisch nicht dem Anfragebild zuzuordnen sind, werden hier als besonders störend empfunden. Zudem wird von einigen Benutzern die Intransparenz der Retrievalmechanismen bemän-

gelt. Interessant ist auch die Forderung nach Mechanismen zum Relevance Feedback im Laufe der Ergebnispräsentation.

Im Vergleich der beiden Methoden wird beim tagbasierten Ansatz vor allem der mangelhafte Detailgrad der Beschlagwortung angeführt. Die zum Teil sehr allgemeinen Annotationsvorschläge werden hierbei als wenig hilfreich für das Retrieval angesehen. Mehrdeutigkeiten bei der Schlagwortvergabe werden ebenfalls bemängelt. Es wird bemängelt, dass das System beim tagbasierten Retrieval teilweise zu abhängig von einzelnen Schlagwörtern ist. Bei der Auswahl mehrerer Annotationsvorschläge würden die Schlagwörter unterschiedlich stark in das Retrieval einbezogen. Bei der bildbasierten Methode wird neben der hohen Fehlerrate vor allem die hohe Abhängigkeit des Systems von dominanten Bildmerkmalen als Problem identifiziert. Mehrere Testpersonen verweisen beispielsweise darauf, dass dominante Farbmerkmale – wie der hohe Blauanteil bei Bild 1b – zu großen Einfluss auf die Ergebnismenge haben.

5 Diskussion

Neben der sachlichen Analyse des Themas gibt es einige Ergebnisse und Anmerkungen, die nach einer ausführlichen Diskussion verlangen. Dabei wird versucht, Erklärungen für die Testergebnisse zu finden und mögliche Tendenzen zu identifizieren. Eine kritische Analyse der Untersuchungsmethode und des untersuchten Systems sollen zudem einen Überblick und Verbesserungsvorschläge für zukünftige Untersuchungen verschaffen.

5.1 Ergebnisinterpretation

Die Analyse wird dabei, analog zur Ergebnisauswertung, erst auf die einzelnen Kernhypothesen und die Nebenbefunde angewandt. Abschließend erfolgt eine übergreifende Auswertung der Befragung.

5.1.1 Diskussion der Kernhypothesen

Bei der statistischen Auswertung der Ergebnisse kann im ersten Schritt eine hohe Akzeptanz gegenüber der automatischen Annotation festgestellt werden. Im Vergleich der beiden Retrievalmethoden schneidet die tagbasierte Methode signifikant besser ab als die bildbasierte Methode, obwohl sich die Qualität der Ergebnisse nicht signifikant unterscheidet. Es bleibt die Frage, wieso die Testpersonen die automatische Beschlagwortung, trotz des zeitlichen Mehraufwands, positiv aufnehmen und weshalb die tagbasierte Methode höher bewertet wird als die inhaltsbasierte.

5.1.1.1 Automatische Bildbeschlagwortung

Die hohe Akzeptanz gegenüber der automatischen Schlagwortvergabe lässt auf den Wunsch der Benutzer schließen, ein solches System zur Beschlagwortung von Bildern auch wirklich zur Verfügung zu haben. Durch die Formulierung eines realistischen Benutzungsszenarios wird vielen Benutzern das Problem der Bildbeschlagwortung erst bewusst. Die Erleichterung bei der Beschlagwortung und die Zeitersparnis durch den Einsatz des Systems werden signifikant positiv bewertet. Die Bewertung der subjektiv wahrgenommenen Schwierigkeit des Systems fällt eindeutig positiv aus. Daraus lässt sich schließen, dass die Auswahl der Schlagwörter den Benutzern leicht fällt und sie sich sehr schnell mit der ihnen unbekannten Methode vertraut machen können.

Trotz der Tatsache, dass die Gesamtakzeptanz signifikant hoch ist, kann nicht allumfassend davon gesprochen werden, dass der Benutzer das System akzeptiert. Die genaue Betrachtung der subjektiven Nützlichkeit und der offenen Fragen ergibt ein differenzierteres Bild. Die Qualität der angebotenen Schlagwörter und die hohe Fehlerrate des Systems werden kritisch gesehen. So wird beispielsweise der mangelhafte Detailgrad der vorgeschlagenen Tags bemängelt. Man könnte sagen, der Nutzen des Systems wird erkannt, es sind jedoch noch Verbesserungen vorzunehmen.

5.1.1.2 Vergleich der Retrievaleffizienz

Die Bewertung der Retrievaleffizienz seitens der Benutzer fällt – wie bereits behandelt – sehr unterschiedlich aus. Die Anzahl der als relevant erkannten Bilder schwankt teilweise extrem. Bei Beispielbild 1b – einem Bild eines historischen Gebäudes – schwanken die Relevanzbewertungen der Ergebnisse von 0 bis 10. Die subjektive Meinung der Testpersonen über die Relevanz und Irrelevanz der Bildbeispiele ist in dieser Hinsicht sehr unterschiedlich. Dieser Effekt lässt sich durch die unterschiedliche Sichtweise der Benutzer erklären. Die Testpersonen gehen von unterschiedlichen semantischen Konzepten bei der Bewertung der Bilder aus. Ein Benutzer erachtet zum Beispiel alle Bilder als relevant, die dem semantischen Konzept „Gebäude“ entsprechen. Ein anderer Teilnehmer wählt lediglich die Bilder aus, die dem viel spezifischeren Konzept „Kathedrale“ zuzuordnen sind. Der Abstraktionsgrad der Testpersonen ist in diesem Fall nicht identisch. Um diesen Effekt zu verringern, wurden bei der Durchführung der Tests möglichst spezifische Retrievalaufgaben gestellt. Trotz dieser Maßnahme konnte kein vollständiger Ausgleich geschaffen werden.

Die hohen Schwankungen der durchschnittlichen Precision-Werte erklären sich durch die wechselhafte Retrievalleistung der beiden Methoden. Die Ergebnisse beider Methoden sind in hohem Maß abhängig vom Bildinhalt. Da die Beschlagwortungsvorschläge der tagbasierten Methode ebenso auf der inhaltlichen Analyse des Bildinhalts beruhen, tritt dieser Effekt bei beiden Methoden auf. Die Auswahl der relevanten Schlagwörter vor dem Retrieval verbessert die Precision der tagbasierten Methode jedoch erheblich. Im Grunde kann dieses Eingreifen der Testpersonen als vorzeitiges Relevance Feedback angesehen werden. Bietet das Relevance Feedback – wie bei Bild 1a, durch die Auswahl des Tags ‘elephant‘ – die Möglichkeit einer sehr detaillierten Suchverfeinerung, ist dieser Effekt natürlich hoch. Ein Gegenbeispiel stellt Bild 1c dar, da hier nur einer der fünfzehn Vorschläge für den Bildinhalt zutreffend ist. Dieser zutreffende Begriff ist zudem sehr

allgemein. Die Konsequenz ist eine geringere Precision bei der tagbasierten Methode.

Da kein signifikanter Unterschied zwischen den beiden Methoden festgestellt werden kann, gilt es weitere Untersuchungen in diese Richtung anzustellen und die Entwicklung beider Retrievalmethoden voranzutreiben.

5.1.1.3 Vergleich der Benutzerakzeptanz

Die hohe Akzeptanz gegenüber der tagbasierten Retrievalvariante wird durch mehrere Faktoren beeinflusst. Der deutliche Unterschied der Akzeptanzbewertung verleitet zu der Annahme, dass ein Zusammenhang zwischen der hohen Akzeptanz der Benutzer und der erzielten Retrievalqualität besteht, da die tagbasierte Methode bessere Ergebnisse als die inhaltsbasierten Suche erzielt. Dem widerspricht, dass keine signifikant bessere Precision festgestellt werden konnte. Ein gewisser Einfluss auf die Bewertung der Testpersonen kann dadurch nicht vollständig abgesprochen werden. Betrachtet man das Retrieval als zweistufigen Prozess von Anfrageformulierung und Ergebnispräsentation, kann zumindest bei der zweiten Stufe unterstellt werden, dass die Ergebnisqualität in die Akzeptanzbewertung mit einfließt.

Ein weiterer Vorteil für die tagbasierte Methode ergibt sich durch die Analyse der Anfragemethoden der beiden Retrievalkonzepte. Bei der inhaltsbasierten Suche besteht die Anfrage aus der Einreichung eines Beispielbildes, wohingegen bei der tagbasierten Methode nach der Einreichung des Bildes ein weiterer Zwischenschritt erfolgt. Der Benutzer erhält bereits vor der ersten Ergebnispräsentation die Möglichkeit, die Suchanfrage zu verfeinern. Der daraus resultierende Qualitätsvorteil bei der Ergebnispräsentation hat damit auch Gewicht bei der Bewertung der Akzeptanz.

Nichtsdestotrotz ist der Vergleich der beiden Retrievalmethoden legitim. Neben der Qualitätsverbesserung bedeutet die frühe Einbindung des Benutzers auch einen höheren kognitiven und zeitlichen Aufwand. Dieser Mehraufwand wird von den Testpersonen jedoch nicht als Belastung wahrgenommen. Es scheint, dass der Benutzer die Möglichkeiten zur Einflussnahme auf das Retrieval generell positiv bewertet. Zudem wird die subjektiv wahrgenommene Effektivität beider Methoden bei der Akzeptanzbewertung bewusst miteinbezogen. Bei der statistischen Auswertung schneiden aber nicht nur effektivitätsbezogene Items besser ab, sondern auch eine Vielzahl anderer Items. Daher ist nicht allein die Effektivität des Retrievalsystems ausschlaggebend für die Akzeptanz der Benutzer gegenüber einem System.

Das Gesamtergebnis beider in dieser Studie untersuchten Methoden kann als allgemeiner Wunsch der Benutzer nach alltagstauglichen Alternativen zu textbasierten Bildretrievalmethoden gewertet werden. Dabei würde beim aktuellen Entwicklungsstand die tagbasierte Methode der inhaltsbasierten Methode vorgezogen werden.

5.1.2 Diskussion der Nebenbefunde

Die zusätzlich bei der Untersuchung behandelten Fragestellungen bestätigen die aufgestellten Kernhypothesen und geben einen detaillierten Einblick in die Entscheidung der Testpersonen.

5.1.2.1 Manuelle vs. Automatische Schlagwortvergabe

Die Beschreibung des Bildinhalts durch die Benutzer lieferte – wie bereits beschrieben – eine sehr große Menge an Schlagwörtern. Die Tatsache, dass dabei ausschließlich Personen mit informationstechnischem Hintergrund an der Befragung teilnahmen, wirkte sich positiv auf die Qualität der Beschlagwortung aus. Die Beobachtungen von Li und Wang im Bezug auf die Auswahl der ALIPR-Tags können nur teilweise bestätigt werden. Die Auswahlkriterien der Testpersonen sind von Fall zu Fall sehr unterschiedlich. Eine leichte Tendenz in die Richtung der von Li und Wang aufgestellten Hypothesen konnte beobachtet werden. So wurde bei Bild 1b häufig ‘building‘ als einzig zutreffendes Wort gewählt und die Schlagwörter ‘historical‘, ‘sky‘, ‘man-made‘ und ‘landmark‘ nicht berücksichtigt. Um klare Aussagen zu diesem Punkt treffen zu können, wären jedoch weitere Untersuchungen vonnöten.

Im Vergleich zu den automatisch generierten Schlagwörtern finden sich viele Tags, die den Bildinhalt sehr detailliert beschreiben. Darunter fallen auch Prognosen, die nicht allein durch die Analyse des Bildinhalts erschlossen werden können. Die Testpersonen bedienen sich ihres Weltwissens, um Annahmen über den potentiellen Aufnahmeort, die kulturelle Bedeutung oder soziale Einstufungen des Bildinhalts zu treffen. Dazu zählen Schlagwörter wie ‘Africa‘, ‘religion‘ oder ‘family‘. Bei diesem Vergleich zeigen sich deutlich die bestehenden Probleme der Überwindung des ‘semantic gap‘. Die eingesetzten Schlagwörter und Kategorien bei der automatischen Annotation sind noch nicht differenziert genug, um solche Konzepte abzubilden. Die Betrachtung der von ALIPR vorgeschlagenen Wörter zeigt aber auch, dass die Überwindung dieser Schwelle nicht unmöglich ist. Teilweise konnten sehr detaillierte Tags, wie ‘elephant‘ oder ‘landmark‘ zugewiesen werden.

Im Umkehrschluss kann aber festgestellt werden, dass Benutzer dazu neigen, sich auf die Hauptbestandteile eines Bildes zu konzentrieren und den restlichen Inhalt zu vernachlässigen. In diesem Punkt hat die objektive Betrachtungsweise des ALIPR-Systems Vorteile gegenüber der subjektiven Sicht der Benutzer. Tags wie ‘tree‘ bei Bild 1a oder ‘sky‘ bei Bild 1b wurden von fast allen Benutzern als relevant markiert, obwohl bei der manuellen Schlagwortvergabe nicht auf diese Bildbereiche eingegangen wurde. Hier zeigt sich, neben der Verringerung des manuellen Arbeitsaufwands, ein weiterer Vorteil von automatischer Beschlagwortung. Dem Benutzer wird es dadurch erleichtert, ein grundlegendes Problem der manuellen Annotation von Bildern zu umgehen.

5.1.2.2 Ergebnisse der offenen Fragen

Die Angaben der Benutzer bei den offenen Fragen decken sich häufig mit den Antworten auf bestimmte Fragebogenitems. Es scheint, als wollen die Benutzer Aspekte hervorheben, die ihnen besonders wichtig bei der Beurteilung eines solchen Systems sind. Die allgemein positive Bewertung der Einfachheit und der Schnelligkeit des Systems besitzt nur bedingte Aussagekraft, da der Retrievalprozess nicht vollständig von den Benutzern durchgeführt wird. Die Auswahl der Beispielbilder erfolgt bereits vor der Einbindung der Benutzer durch den Testleiter.

Klar feststellbar ist die bereits behandelte Fokussierung der Benutzer auf detaillierte semantische Konzepte im Bild. Der Wunsch nach einer Steigerung des Detailgrads wird sowohl für die Ergebnismenge, als auch für die vorgeschlagenen Tags geäußert. Das Tätigkeitsfeld der Testpersonen ist auch hier klar erkennbar, da teilweise sehr detaillierte Anmerkungen zu den Retrievalmethoden gemacht werden. Der Hinweis auf die hohe Abhängigkeit der inhaltsbasierten Methode gegenüber farblichen Bildmerkmalen oder die Forderung nach detaillierteren Verfahren zum Relevance Feedback sind Beispiele hierfür.

Bei der tagbasierten Methode wird von vielen Benutzern das Potenzial zur vollautomatischen Annotation hervorgehoben. Die Möglichkeiten zur automatischen Kategorisierung und zur Anreicherung des Bildmaterials mit Metainformation werden dabei als besonders erstrebenswert angesehen.

5.2 Verbesserungsvorschläge

Die umfassende Untersuchung und Auswertung des ALIPR-Systems hinsichtlich seiner Benutzerakzeptanz führt zu einer Reihe von Anmerkungen und Verbesse-

rungsvorschlägen. Diese beziehen sich einerseits auf die angewandte Methode und andererseits auf das untersuchte System.

5.2.1 Anmerkungen zur Untersuchungsmethode

Das bei der Untersuchung angewandte, theoretische Modell und das eingetzte Instrument hat sich für die Erhebung der Benutzerakzeptanz bewährt. Nach dem Abschluss der Benutzerbefragung können für einige Bereiche Anmerkungen zur Verbesserung ähnlicher Untersuchungen gemacht werden.

5.2.1.1 Zusammenstellung der Stichprobe

Die gewählte Stichprobe setzt sich – wie bereits erwähnt – aus Personen aus dem Bereich der Informationstechnologie zusammen. Bei der Untersuchung eines prototypischen Bildretrievalsystems bietet sich diese homogene Teilnehmerstruktur an, da Laien bei der Benutzung dieses Systems schnell überfordert werden können. Nachdem ein Großteil der Benutzer die einfache Bedienbarkeit des Systems besonders hervorgehoben hat, wäre die Überprüfung der Hypothesen mit einer größeren und inhomogeneren Gruppe von Testpersonen definitiv interessant. Unter Einbeziehung der Vorkenntnisse der Benutzer zur Einteilung in mehrere unabhängige Stichproben, wäre eine weitere Untersuchung der Benutzerakzeptanz sicherlich eine Herausforderung.

5.2.1.2 Auswahl und Menge der Bilder

Die ermittelte Precision der beiden Retrievalmethoden kann nur bedingt als repräsentativ gesehen werden. Die geringe Anzahl der Beispielbilder erlaubt keine verlässliche Signifikanzbewertung. Um eine präzise Bewertung der Retrievalqualität zu ermöglichen, ist eine weitaus umfassendere Bildmenge nötig, die alle von ALIPR verwendeten Bildkategorien abdeckt. Für die Berechnung dieser „subjektiven“ Precision wäre dabei die Ermittlung der Benutzerkompetenz hinsichtlich der Einteilung von relevanten und irrelevanten Bildern ebenfalls von Interesse. Es gilt als fraglisch, ob die Ermittlung des Precision-Wertes durch den Einsatz einer schriflichen Erhebung in diesem Fall sinnvoll wäre. Kosten und Aufwand einer solch umfangreichen Befragung wären nur schwer durch den daraus gewonnenen Nutzen aufzuwiegen.

Die Auswahl der Bilder für die Untersuchung wurde intellektuell vorgenommen. Dabei wurden möglichst verschiedene semantische Konzepte für die Bilder verwendet, um die Leistung der Retrievalmethoden bei unterschiedlichen Bildszenen zu testen. Zudem wurde bei der Auswahl auf die jeweilige Aufnahme-

situation beziehungsweise auf den Bildstil geachtet. Ein quantitativer Vergleich hinsichtlich dieser Kategorien war aber anhand der vier Bildbeispiele nicht möglich. Um valide Aussagen zur Retrievalleistung des ALIPR-Systems in Abhängigkeit zum Bildstil zu treffen wäre der Vergleich mehrerer Bildbeispiele aus jeder Kategorie nötig. Der Einsatz eines Benutzertests ist auch hier eine Frage der Kosten-Nutzen-Rechnung, da eine sehr große Anzahl von Bildern verwendet und die Retrievalleistung manuell bewertet werden müsste. In Bezug auf die Retrievalleistung des ALIPR-Systems muss man sich daher auf die Aussage der Entwickler verlassen, die die Stärke der tagbasierten Methode mehr in der Szenenerkennung als in der Identifikation spezieller Objekte im Bild sehen (vgl. J. LI & J. Z. WANG, 2006, p. 4).

5.2.1.1 Item-Gewichtung

Die Untersuchung der Einzelkomponenten des ALIPR-Systems fokussiert sich in erster Linie auf signifikante Unterschiede bei der Benutzerakzeptanz. Dabei lassen sich klare Ergebnisse berechnen und es können Tendenzen identifiziert und diskutiert werden. Eine Aussage über den Einfluß der einzelnen Items auf den Gesamtfaktor Akzeptanz lässt sich jedoch nicht treffen. Bei der durchgeführten Untersuchung wurde die gleichmäßige Gewichtung aller Items beschlossen, da im vorhandenen Rahmen keine klare Vorhersage über den Einfluß der Teilfaktoren getroffen werden konnte. Eine Gewichtung nach der Auswertung der Ergebnisse wurde ausgeschlossen, um nicht den Anschein einer nachträglichen Ergebnisanpassung zu erwecken. Durch den Entwurf eines geeigneten theoretischen Modells zur Gewichtung der akzeptanzbeeinflussenden Faktoren unter Einbeziehung der vorhandenen Akzeptanztheorien, könnte der Gesamtfaktor für folgende Untersuchungen jedoch genauer bestimmt werden. Die Einbeziehung weiterer Teilfaktoren neben der subjektiven erlebten Schwierigkeit und Nützlichkeit wie beim UTAUT-Modell wäre hier ein zu bedenkender Lösungsansatz.

Aufbauend auf den Überlegungen zur Item-Gewichtung kann neben der generellen Signifikanz auch über die Berechnung von Effektstärken nachgedacht werden. Dadurch kann die Korrelation der Einzelitems zwischen den verschiedenen Retrievalmethoden berechnet werden. Neben einer Aussage über signifikante Tendenzen können besonders einflussreiche Teilfaktoren identifiziert werden und spezifischere Aussagen über die Benutzerpräferenzen getroffen werden.

5.2.1.2 Validität und Reliabilität

Validität und Reliabilität von selbst erstellten, nicht-standardisierten und nicht-geeichten Fragebögen müssen bei der Bewertung der Untersuchungsergebnisse kritisch gesehen werden. In der Fachliteratur werden – sofern vorhanden – standardisierte und geeichte Fragebögen für die Erhebung von Usabilityfaktoren empfohlen (vgl. SCHWEIBENZ & THISSEN, 2002, p. 119). Für das bildbasierte Retrieval existiert bis dato kein geeichtes Verfahren zur Erhebung der Benutzerakzeptanz. Aus diesem Grund war die selbstständige Konstruktion eines passenden Untersuchungsinstruments unumgänglich. Auch wenn die eingesetzten Items größtenteils aus validierten Fragebögen entnommen wurden, mussten sie für das spezifische Szenario angepasst werden. Für zukünftige Untersuchungen sollten die verwendeten Items ausgiebig hinsichtlich ihrer Validität geprüft werden, um den erhobenen Ergebnissen mehr Aussagekraft zu verleihen. Die bereits erwähnte, hohe Anzahl von Testpersonen für die Eichung eines Fragebogens macht diesen Punkt zu einer sehr anspruchsvollen Aufgabe. Dabei ist zu beachten, dass der validierte Fragebogen möglichst universell auf Systeme aus dem inhaltsbasierten Bildretrieval anwendbar ist.

Hinsichtlich der Reliabilität gibt es trotz der sehr strukturierten Befragungsdurchführung ebenfalls Verbesserungspotential. Beispielsweise könnte durch den Einsatz eines schriftlichen Ablaufprotokolls für den Testleiter der Testablauf weiter standardisiert werden. Eine noch genauere Formulierung der Benutzerinstruktion kann für einige Punkte der Befragung ebenfalls hilfreich sein. Aufgrund des sehr spezifischen Themas der Studie ist die Verbesserung der Reliabilität für zukünftige Studien nur sinnvoll, wenn auf systemunabhängige Punkte eingegangen wird.

5.2.2 Anmerkungen zum ALIPR-System

Durch die Studie der zentralen Bestandteile des ALIPR-Systems aus Sicht der Benutzer lassen sich nicht nur Aussagen über die Zusammenstellung der Untersuchungsmethode oder über die Benutzerpräferenz treffen. Neue Erkenntnisse hinsichtlich der Verbesserung und Weiterentwicklung des ALIPR-Systems ergeben sich ebenso. Dabei wird vor allem auf Komponenten eingegangen, die der tagbasierten Methode des Systems zuzuordnen sind, da darauf auch der Fokus der Arbeit liegt.

5.2.2.1 Verbesserung der Bildanalyse

Beide Suchvarianten des ALIPR-Systems können durch den Einsatz von verbesserten Bildanalysemethoden in ihrer Retrievalqualität verbessert werden. Die derzeitige Version des Systems bedient sich der Farb- und der Texturmerkmale der analysierten Bilder und setzt auf eine mehrdimensionale Segmentierungsmethode. Methoden zur Formerkennung werden vom System bisher nicht eingesetzt. Der Austausch der bisherigen Segemtierungsmethode durch eine geeignetes Verfahren zur Formerkennung könnte die Erkennungsrate steigern. Dabei gilt es eine Methode zu finden, auf die das zur Kategorisierung verwendete statistische Modell angewandt werden kann. Durch die Erhöhung der Rechnerleistung kann zudem über eine Verbesserung der bestehenden Extraktionsmethoden nachgedacht werden. Eine spezifische Empfehlung lässt sich aufgrund der fehlenden Einsicht in den Aufbau des Systems und die praktische Einbindung der Retrievalalgorithmen nicht aussprechen.

5.2.2.2 Ausbau der Schlagwortkategorien

Ein Hauptkritikpunkt bei der Benutzerbefragung war der mangelhafte Detailgrad der automatischen Annotation. Das System kann aus mehreren Gründen keine detailliertere Annotation bewerkstelligen. Zum einen lässt die Echtzeitverarbeitung keine leistungsfähigere Bildanalyse und –kategorisierung zu. Dieser Umstand kann mit wachsender Rechenleitung und der kontinuierlichen Verbesserung der Analyseverfahren negiert werden. Zum anderen basiert das eingesetzte Annotationsverfahren auf einer zu kleinen Wortliste. Die Entwickler geben an, dass die durch zusätzliche manuelle Annotation gewonnenen Schlagwörter in die Wortliste aufgenommen werden, dadurch steigt jedoch die Gefahr das fehlerhafte Tags aufgenommen werden. Die neuen Schlagwörter müssen also vor der Aufnahme in den Annotationskatalog überprüft werden. Kann die Wortliste durch alternative Methoden, wie zum Beispiel den Einsatz von Ontologien und Wortnetzen in Kombination mit hochqualitativen Trainingssets erweitert werden, würde dieser Schritt entfallen. Ohne eine Verbesserung der Annotationsgenauigkeit scheint das System jedenfalls nicht tauglich für den Einsatz in großen Umfang.

5.2.2.3 Fokussierung auf die Annotation

Betrachtet man die angebotenen Retrievalmethoden, stellt sich die Frage welchen Mehrwert die Suche von Bildern mittels der automatischen Annotation gegenüber der rein textbasierten Suche bietet. Der Zwischenschritt zur Auswahl geeigneter Wörter ist zeitlich und kognitiv fordernd. Erhält der Benutzer bei der

automatischen Annotation keine geeigneten Tags wurde der Suchaufwand unnötig erhöht. Wie im vorhergehenden Punkt erwähnt ist die automatische Annotation noch zu ungenau um durchgehend relevante Schlagwörter zu generieren. Daher wird der Benutzer im Alltag nach einigen Fehlversuchen mit ALIPR mit einiger Wahrscheinlichkeit wieder auf die herkömmliche textbasierte Methode vertrauen.

Wird ALIPR dagegen verwendet, um den Bildinhalt unabhängig von seiner Retrievalkomponente mit Schlagwörtern zu versehen, stellt es eine Alternative zur manuellen Anreicherung von Bildern mit Metainformation dar. Auch hier gilt der Einwand der mangelhaften Genauigkeit der Tags, aber entgegen dem Einsatz beim Retrieval verspricht der Einsatz des Systems zur Unterstützung der manuellen Annotation eine kognitive und zeitliche Entlastung. Wird das System in dieser Richtung konsequent verbessert und die Einbindung des Benutzers weiter reduziert, stellt es eine vielversprechende Entwicklung auf dem Bereich der semantischen Informationsanreicherung dar. Der Schritt zur vollständigen Automatisierung des Annotationsprozesses ist in jedem Fall noch nicht absehbar.

6 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden die theoretischen Grundlagen des Bildretrieval im Bezug auf die automatische Annotation von Bildern behandelt. Durch den Entwurf einer Studie zur Benutzerakzeptanz gegenüber automatischer Annotation am Beispiel des ALIPR-Systems wurde ein praktischer Bezug hergestellt. Die Auswertung der Benutzerevaluation ergab eine hohe Akzeptanz der Benutzer gegenüber der automatischen Beschlagwortung und der damit verbundenen Methoden. Durch die kritische Auseinandersetzung mit den Ergebnissen, konnten Rückschlüsse zur Verbesserung der Untersuchungsmethode und des Systems gezogen werden. Anschließend wurde die Perspektive des ALIPR-Systems hinsichtlich der automatischen Anreicherung von Bildmaterial mit Metainformation diskutiert.

Eakins formuliert die Ansprüche von Benutzern eines Bildretrievalsystems folgendermaßen:

> „The ability to retrieve images by their semantic content is a clear priority for users of image databases. Lower level issues are generally considered less important.[...] System design must take account of this, while trying to introduce newer, more efficient interaction to those users." (EAKINS ET AL., 2004, p. 637)

Das ALIPR-System kann die hohen Erwartungen der Benutzer in den meisten Fällen noch nicht erfüllen. Der Einsatz von neuen Retrievalmethoden wird von den Benutzer trotzdessen positiv angenommen. Dies sollte als klares Indiz für das Potential des verwendeten Systems gesehen werden. Die Anreicherung von Bilddaten mit semantisch wertvollen Metadaten ist dabei die wichtigste Komponente des Systems.

Eine qualitativ hochwertig annotierte Bilddatenbank kann einen entscheidenden Beitrag für die Weiterentwicklung von neuen Technologien des Semantic Web leisten und birgt großes Potential für den Einsatz im Alltag. Automatische Beschlagwortung dient dabei als Schlüsseltechnologie auf dem Bereich des Bildretrieval. Obwohl das 'semantic gap' durch automatische Annotation nicht geschlossen werden kann, trägt diese Technologie enorm zu Verkleinerung der Kluft zwischen Daten und Inhalt bei.

7 Abbildungsverzeichnis

8 Tabellenverzeichnis

9 Literaturverzeichnis

VON AHN, L., & DABBISH, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 319-326). Vienna, Austria: ACM. doi: 10.1145/985692.985733.

VON AHN, L., & DABBISH, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58-67. doi: 10.1145/1378704.1378719.

ALIPR. (2009). ALIPR - Automatic Real-time Image Tagging and Searching. Letzter Zugriff: 13. Februar 2009, unter http://www.alipr.com/.

AL-MASKARI, A., CLOUGH, P., & Sanderson, M. (2006). Users' effectiveness and satisfaction for image retrieval. In *Proceedings of the LWA 2006 Workshop* (pp. 84-88). Hildesheim, Germany: White Rose Research Online. Letzter Zugriff: 3. Februar 2009, unter http://eprints.whiterose.ac.uk/4504/.

ASHLEY, J., FLICKNER, M., HAFNER, J., LEE, D., NIBLACK, W., & PETKOVIC, D. (1995). The query by image content (QBIC) system. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data* (p. 475). San Jose, USA: ACM. doi: 10.1145/223784.223888.

BACHOO, A. K., & TAPAMO, J. (2005). Texture detection for segmentation of iris images. In *Proceedings of the 2005 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries* (pp. 236-243). White River, South Africa: South African Institute for Computer Scientists and Information Technologists. Letzter Zugriff: 13. Februar 2009, unter http://portal.acm.org/citation.cfm?id=1145701&dl=GUIDE&coll=GUIDE&CFID=22251539&CFTOKEN=50875279.

BAEZA-YATES, R. (1999). *Modern information retrieval*. New York, USA: ACM Press.

BARNARD, K., DUYGULU, P., FORSYTH, D., FREITAS, N. D., BLEI, D. M., & JORDAN, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107-1135.

BORTZ, J., & DÖRING, N. (2003). *Forschungsmethoden und Evaluation. für Human- und Sozialwissenschaftler* (3. ed.). Berlin, Germany: Springer Berlin.

BOUCHARD, G., & TRIGGS, B. (2005). Hierarchical part-based visual object categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005* (Vol. 1, pp. 710-715). Cambridge, USA: MIT Press. doi: 10.1109/CVPR.2005.174.

CARNEIRO, G., CHAN, A. B., MORENO, P. J., & VASCONCELOS, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 394-410. doi: 10.1.1.88.3490.

CARSON, C., BELONGIE, S., GREENSPAN, H., & MALIK, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to im-

age querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 1026-1038. doi: 10.1.1.27.1819.

CETINKAYA, B., ASLAN, S., SENGUN, Y. S., COBANKAYA, N. O., & ILGIN, D. E. (2006). Contour Simplification with Defined Spatial Accuracies. In *Workshop of the ICA Commission on Map Generalisation and Multiple Representation* (pp. 1-7). Vancouver, United States.

CHAN, A. B., MORENO, P. J., & VASCONCELOS, N. (2006). Using Statistics to Search and Annotate Pictures: an Evaluation of Semantic Image Annotation and Retrieval on Large Databases. In *Proceedings of Joint Statistical Meetings* (pp. 1-9). Seattle.

CHEN, Y., WANG, J. Z., & KROVETZ, R. (2003). Content-based image retrieval by clustering. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval* (pp. 193-200). Berkeley, California: ACM. doi: 10.1145/973264.973295.

CHOI, Y., & RASMUSSEN, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing and Management*, *38*(5), 695-726. doi: 10.1016/S0306-4573(01)00059-0.

COGNISIGN LLC. (2008). Search Stock Photography - Powerful Image Recognition Technology for Searching Photos and Images by Xcavator.net. Letzter Zugriff: 13. Februar 2009, unter http://xcavator.net/PhotoSearch#tags/image:Animals%20africa/15392038/%23C87F00/3,0/0.

COLLINS, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229-238. doi: 10.1023/A:1023254226592.

DANCEY, C. P., & REIDY, J. (2004). *Statistics Without Maths for Psychology: Using Spss for Windows*. Upper Saddle River, USA: Prentice-Hall. Letzter Zugriff: 19. Februar 2009, unter http://portal.acm.org/citation.cfm?id=993817.

DATTA, R., GE, W., LI, J., & WANG, J. Z. (2007). Toward Bridging the Annotation-Retrieval Gap in Image Search. *IEEE MultiMedia*, *14*(3), 24-35.

DATTA, R., JOSHI, D., LI, J., & WANG, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, *40*(2), 1-60. doi: 10.1145/1348246.1348248.

DAVIS, F. D., BAGOZZI, R. P., & WARSHAW, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, *35*(8), 982-1003.

DESELAERS, T., KEYSERS, D., & NEY, H. (2008). Features for image retrieval: an experimental comparison. *Information Retrieval*, *11*(2), 77-107. doi: 10.1007/s10791-007-9039-3.

DÍAZ, A., GARCÍA, A., & GERVÁS, P. (2008). User-centred versus system-centred evaluation of a personalization system. *Information Processing & Management*, *44*(3), 1293-1307. doi: 10.1016/j.ipm.2007.08.001.

DUDA, R. O., HART, P. E., & STORK, D. G. (2000). *Pattern Classification: Pattern Classification Pt.1* (2. ed.). New York, USA: Wiley & Sons.

EAKINS, J. P., BRIGGS, P., & BURFORD, B. (2004). Image Retrieval Interfaces: A User Perspective. In *Image and Video Retrieval* (p. 49). Heidelberg, Germany: Springer Berlin. Letzter Zugriff: 20. Januar 2009, unter http://www.springerlink.com/content/jb4a8tf5du37w50h.

FENG JING, MINGJING LI, HONG-JIANG ZHANG, & BO ZHANG. (2002). Region-based relevance feedback in image retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 4, pp. 145-148). IEEE Computer Society. doi: 10.1109/ISCAS.2002.1010410.

FISHBEIN, M., & AJZEN, I. (1975). *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading, USA: Addison-Wesley.

FOLKERS, A., & SAMET, H. (2002). Content-based image retrieval using Fourier descriptors on a logo database. In *Proceedings of the 16th International Conference on Pattern Recognition, 2002.* (Vol. 3, pp. 521-524). Washington, USA: IEEE Computer Society. doi: 10.1109/ICPR.2002.1047991.

FORRET, P. (2006). A picture a day: Flickr's storage growth at blog.forret.com. *blog.forret.com*. Blog, . Letzter Zugriff: 11. Februar 2009, unter http://blog.forret.com/2006/10/a-picture-a-day-flickrs-storage-growth/.

GETTY IMAGES. (2009). Search. *Search*. Suchportal, . Letzter Zugriff: 20. . Februar 2009, unter http://www.gettyimages.com/Search/Search.aspx?contractUrl=2&language=en-US&family=editorial&assetType=image&p=&src=standard.

GONZALEZ, R. C., & WOODS, R. E. (2007). *Digital Image Processing* (3. ed., p. 954). Upper Saddle River, USA: Prentice-Hall.

GOOGLE. (2007). Google Image Labeler. Letzter Zugriff: 11. Februar 2009, unter http://images.google.com/imagelabeler/.

GRUBINGER, M., CLOUGH, P., HANBURY, A., & MÜLLER, H. (2008). Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In *Advances in Multilingual and Multimodal Information Retrieval* (pp. 433-444). Heidelberg, Germany: Springer-Verlag. Letzter Zugriff: 13.Februar 2009, unter http://portal.acm.org/citation.cfm?id=1428850.1428919&coll=Portal&dl=GUIDE&CFID=17682836&CFTOKEN=57339162.

GWAP. (2009). gwap.com - Home. Letzter Zugriff: 11.Februar 2009, unter http://www.gwap.com/gwap/.

HADJIDEMETRIOU, E., GROSSBERG, M., & NAYAR, S. (2004). Multiresolution histograms and their use for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(7), 831-847. doi: 10.1109/TPAMI.2004.32.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. H. (2003). *The Elements of Statistical Learning* (Corrected.). New York, USA: Springer.

HE, J., TONG, H., LI, M., ZHANG, H., & ZHANG, C. (2004). Mean version space: a new active learning method for content-based image retrieval. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval* (pp. 15-22). New York, USA: ACM. doi: 10.1145/1026711.1026715.

HERVÉ, N., & BOUJEMAA, N. (2007). Image annotation: which approach for realistic databases? In *Proceedings of the 6th ACM international conference on Image and video retrieval* (pp. 170-177). Amsterdam, Netherlands: ACM. doi: 10.1145/1282280.1282310.

HIRATA, K., & KATO, T. (1992). Query by Visual Example - Content based Image Retrieval. In *Proceedings of the 3rd International Conference on Extending Database Technology: Advances in Database Technology* (pp. 56-71). London, England: Springer-Verlag. Letzter Zugriff: 13. Januar 2009, unter http://portal.acm.org/citation.cfm?id=649863.

HOI, C., & LYU, M. R. (2004). A novel log-based relevance feedback technique in content-based image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 24-31). New York, USA: ACM. doi: 10.1145/1027527.1027533.

HOVE, L. (2007). Evaluating Use of Interfaces for Visual Query Specification. In *Proceedings of the NOKOBIT 2007*. Bergen: Tapir Akademisk Forlag. doi: 10.1.1.101.3456.

HUIJSMANS, D., & SEBE, N. (2005). How to complete performance graphs in content-based image retrieval: add generality and normalize scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(2), 245-251. doi: 10.1109/TPAMI.2005.30.

HUMAN FACTORS RESEARCH GROUP. (2007). What is SUMI? *What is SUMI?* Letzter Zugriff: 11.Februar 2009, unter http://sumi.ucc.ie/whatis.html.

IBM. (2008). IBM - DB2 - Data server - database software - database management - open source. Letzter Zugriff: 13.Februar 2009, unter http://www-01.ibm.com/software/data/db2/.

IMGSEEK.NET. (2009). imgSeek. Letzter Zugriff: 12. Februar 2009, unter http://www.imgseek.net/.

IPTC. (2008). IPTC Standard Photo Metadata 2008. International Press Telecommunications Council. Letzter Zugriff: 19. Februar 2009 unter http://www.iptc.org/std/photometadata/2008/specification/.

JAIN, A. K., MURTY, M. N., & FLYNN, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264-323. doi: 10.1145/331499.331504.

JAIN, A., & FARROKHNIA, F. (1990). Unsupervised texture segmentation using Gabor filters. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 14-19). Washington, USA: IEEE Computer Society. doi: 10.1109/ICSMC.1990.142050.

JEON, J., & MANMATHA, R. (2004). Using Maximum Entropy for Automatic Image Annotation. In *Image and Video Retrieval*, Lecture Notes in Computer Science (pp. 2040-2041). Heidelberg, Germany: Springer Berlin. Letzter Zugriff: 22. Januar 2009, unter http://www.springerlink.com/content/lnk059udw8wqccth.

JIN, Y., KHAN, L., WANG, L., & AWAD, M. (2005). Image annotations by combining multiple evidence \& wordNet. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 706-715). Hilton, Singapore: ACM. doi: 10.1145/1101149.1101305.

JING, F., LI, M., ZHANG, H., & ZHANG, B. (2002). Region-based relevance feedback in image retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 4, pp. 145-148). doi: 10.1109/ISCAS.2002.1010410.

JULESZ, B. (1975). Experiments in the visual perception of texture. *Scientific American*, *232*(4), 34-43.

KAIKKONEN, A., KEKÄLÄINEN, A., CANKAR, M., KALLIO, T., & KANKAINEN, A. (2005). Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. *Journal of Usability Studies*, *1*(1), 4-17.

KAYA, M. (2007). Verfahren der Datenerhebung. In *Methodik der empirischen Forschung* (pp. 49-64). Gabler. Letzter Zugriff: 9. Februar 2009, unter http://dx.doi.org/10.1007/978-3-8349-9121-8_4.

KONSTANTINIDIS, K., GASTERATOS, A., & ANDREADIS, I. (2006). The Impact of Low-Level Features in Semantic-Based Image Retrieval. In *Semantic-Based Visual Information Retrieval* (1. ed.). Hershey, USA: IRM Press.

KROMREY, H. (2006). *Empirische Sozialforschung*. UTB ; 1040 : Soziologie. Stuttgart: Lucius & Lucius.

KRUIZINGA, P., PETKOV, N., & GRIGORESCU, S. E. (1999). Comparison of Texture Features Based on Gabor Filters. In *Proceedings of the 10th International Conference on Image Analysis and Processing* (p. 142). Washington, USA: IEEE Computer Society. Letzter Zugriff: 13. Februar 2009, unter http://portal.acm.org/citation.cfm?id=840748.

LANGREITER, C. (2006). retrievr - search by sketch / search by image. Letzter Zugriff: 12. Februar 2009, unter http://labs.systemone.at/retrievr/about.

LATECKI, L., & LAKAMPER, R. (2000). Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(10), 1185-1190. doi: 10.1109/34.879802.

LEE, S., & YONG, H. (2007). TagPlus: A Retrieval System using Synonym Tag in Folksonomy. In *International Conference on Multimedia and Ubiquitous Engineering* (pp. 294-298). Washington, USA: IEEE Computer Society. doi: 10.1109/MUE.2007.201.

LI, J., & WANG, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1075-1088. doi: 10.1.1.6.3902.

LI, J., & WANG, J. Z. (2005). ALIP: the Automatic Linguistic Indexing of Pictures system. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 1208-1209). Washington, USA: IEEE Computer Society. doi: 10.1109/CVPR.2005.67.

LI, J., & WANG, J. Z. (2006). Real-time computerized annotation of pictures. In *Proceedings of the 14th annual ACM international conference on Multimedia* (pp. 911-920). Santa Barbara, USA: ACM. doi: 10.1145/1180639.1180841.

LI, J., WANG, J. Z., & WIEDERHOLD, G. (2000). IRM: integrated region matching for image retrieval. In *Proceedings of the eighth ACM international conference on Multimedia* (pp. 147-156). Marina del Rey, USA: ACM. doi: 10.1145/354384.354452.

LI, W., & SUN, M. (2005). Automatic Image Annotation Using Maximum Entropy Model. In *Natural Language Processing – IJCNLP 2005*, Lecture Notes in Computer Science (pp. 34-45). Heidelberg, Germany: Springer Berlin. Letzter Zugriff: 22. Februar 2009, unter http://dx.doi.org/10.1007/11562214_4.

LOWE, D. G. (2004). Distinctive image features unter scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91-110. doi: 10.1.1.2.8899.

MACARTHUR, S. D., BRODLEY, C. E., & SHYU, C. (2000). Relevance feedback decision trees in content-based image retrieval. *Proceedings of the IEEE Workshop on Content-based Access of lmage and Video Libraries, 2000*, 68-72. doi: 10.1.1.32.1775.

MANJUNATH, B., & MA, W. (2002). 12 Texture Features for Image Retrieval. In *Image Databases : Search and Retrieval of Digital Imagery* (1. ed., pp. 313-344). New York, USA: Wiley-Interscience.

MANNING, C. D., RAGHAVAN, P., & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. New York, USA: Cambridge University Press.

MARTÍNEZ, J. M. (2004). MPEG-7 Overview. *www.chiariglione.org*. Letzter Zugriff: 13. Februar 2009, unter http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm#E11E2.

MARUYAMA, T. (2006). Real-time K-Means Clustering for Color Images on Reconfigurable Hardware. In *Proceedings of the 18th International Conference on Pattern Recognition* (pp. 816-819). Washington, USA: IEEE Computer Society. Letzter Zugriff: 13. Januar 2009, unter http://portal.acm.org/citation.cfm?id=1170748.1172355&coll=Portal&dl=GUIDE&CFID=17682836&CFTOKEN=57339162.

MCGILL, T., & HOBBS, V. (2008). How students and instructors using a virtual learning environment perceive the fit between technology and task. *Journal of Computer Assisted Learning*, *24*(3), 191-202. doi: 10.1111/j.1365-2729.2007.00253.x.

MEHTRE, B. M., KANKANHALLI, M. S., & LEE, W. F. (1997). Shape measures for content based image retrieval: a comparison. *Information Processing and Management: an International Journal*, *33*(3), 319-337.

MORI, Y., TAKAHASHI, H., & OKA, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*. Letzter Zugriff: 22. Februar 2009, unter http://citeseer.ist.psu.edu/368129.html.

MÜLLER, H., MARCHAND-MAILLET, S., & PUN, T. (2002). The Truth about Corel - Evaluation in Image Retrieval. In *Image and Video Retrieval*, Lecture Notes in Computer Science (pp. 38-49). Heidelberg, Germany: Springer Berlin. Letzter Zugriff: 20. Februar 2009, unter http://dx.doi.org/10.1007/3-540-45479-9_5.

NATSEV, A., RASTOGI, R., & SHIM, K. (2004). WALRUS: a similarity retrieval algorithm for image databases. *IEEE Transactions on Knowledge and Data Engineering*, *16*(3), 301-316. doi: 10.1109/TKDE.2003.1262183.

NG, H. P., ONG, S. H., FOONG, K. W. C., GOH, P. S., & NOWINSKI, W. L. (2006). Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm. In *Proceedings of the 2006 IEEE Southwest Symposium on Image Analysis and Interpretation* (pp. 61-65). Washington, USA: IEEE Computer Society. Letzter Zugriff: 13. Februar 2009, unter http://portal.acm.org/citation.cfm?id=1338450.1339137&coll=Portal&dl=GUIDE&CFID=17682836&CFTOKEN=57339162.

NIELSEN, J. (1993). *Usability engineering* (1. ed.). San Francisco, USA: Morgan Kaufmann.

NORMAN, K., & SHNEIDERMAN, B. (2002). Questionnaire For User Interaction Satisfaction. Letzter Zugriff: 11. Februar 2009, unter http://www.lap.umd.edu/QUIS/index.html.

PANDA, N., & CHANG, E. Y. (2006). Efficient top-k hyperplane query processing for multimedia information retrieval. In *Proceedings of the 14th annual ACM international conference on Multimedia* (pp. 317-326). Santa Barbara, USA: ACM. doi: 10.1145/1180639.1180712.

PORTA, M. (2006). Browsing large collections of images through unconventional visualization techniques. In *Proceedings of the working conference on Advanced visual interfaces* (pp. 440-444). Venezia, Italy: ACM. doi: 10.1145/1133265.1133354.

PRASAD, B. E., GUPTA, A., TOONG, H. D., & MADNICK, S. E. (1987). A Microcomputer-Based Image Database Management System. *IEEE Transactions on Industrial Electronics*, *IE-34*(1), 83-88. doi: 10.1109/TIE.1987.350929.

REN, X., & MALIK, J. (2003). Learning a classification model for segmentation. *Proceedings of the 9th International Conference on Computer Vision*, *1*, 10-17. doi: 10.1.1.60.2137.

RUBNER, Y., TOMASI, C., & GUIBAS, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, *40*(2), 99-121. doi: 10.1023/A:1026543900054.

RUI, Y., & HUANG, T. S. (1999). Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, *10*, 39-62. doi: 10.1.1.106.2928.

RZ UNI REGENSBURG. (2008). PC Softwareindex - Alphabetischer Index. Letzter Zugriff: 12. Februar 2009, unter http://www.uni-regensburg.de/Einrichtungen/RZ/Benutzer/Allgemein/PCSI/alpha/index.phtml.

SARODNICK, F., & BRAU, H. (2006). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung* (1. ed.). Bern, Schweiz: Huber.

SCHWEIBENZ, W., & THISSEN, F. (2002). *Qualität im Web: Benutzerfreundliche Webseiten durch Usability-Evaluation* (1. ed.). Heidelberg, Germany: Springer Berlin.

SHI, J., & MALIK, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888-905.

SHI, R., LEE, C., & CHUA, T. (2007). Enhancing image annotation by integrating concept ontology and text-based bayesian learning model. In *Proceedings of the 15th international conference on Multimedia* (pp. 341-344). Augsburg, Germany: ACM. doi: 10.1145/1291233.1291307.

SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., & JAIN, R. (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349-1380.

SMITH, J. R. (2002). Color for image retrieval. In *Image Databases: Search and Retrieval of Digital Imagery* (1. ed., pp. 285-311). New York, USA: Wiley-Interscience.

SMITH, J. R., & CHANG, S. (1996). VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia* (pp. 87-98). New York, USA: ACM. doi: 10.1145/244130.244151.

SPINK, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management*, *38*(3), 401-426. doi: 10.1016/S0306-4573(01)00036-X.

SPSS INC. (2009). SPSS, Data Mining, Statistical Analysis Software, Predictive Analysis, Predictive Analytics, Decision Support Systems. Letzter Zugriff: 12. Februar 2009, unter http://www.spss.com/.

STATE HERMITAGE MUSEUM. (2003). The State Hermitage Museum: Digital Collection -- Powered by IBM. Letzter Zugriff: 13. Februar 2009, unter http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English.

STATISTISCHES BUNDESAMT. (2007). 23% der Haushalte besitzen einen MP3-Player, 59% ein DVD-Gerät. Letzter Zugriff: 11. Februar 2009, unter http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Presse/pm/2007/02/PD07__051__IKT.psml.

SWAIN, M. J., & BALLARD, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11-32.

TAMURA, H., MORI, S., & YAMAWAKI, T. (1978). Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, *8*(6), 460-473. doi: 10.1109/TSMC.1978.4309999.

TONG, S., & CHANG, E. (2001). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 107-118). Ottawa, Canada: ACM. doi: 10.1145/500141.500159.

TSAI, C., MCGARRY, K., & TAIT, J. (2006). Qualitative evaluation of automatic assignment of keywords to images. *Information Processing and Management: an International Journal*, *42*(1), 136-154.

TURPIN, A., & SCHOLER, F. (2006). User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 11-18). Seattle, USA: ACM. doi: 10.1145/1148170.1148176.

VAILAYA, A., MEMBER, A., FIGUEIREDO, M. A. T., JAIN, A. K., ZHANG, H., & MEMBER, S. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, *10*(1), 117-130. doi: 10.1.1.6.56.

VENKATESH, V., & MORRIS, M. (2003). User Acceptance of Information Technology: Toward a Unified View. *Management Information Systems Quarterly*, *27*(1), 18.

VENTERS, C., HARTLEY, R., COOPER, M., & HEWITT, W. (2001). Query by Visual Example: Assessing the Usability of Content-Based Image Retrieval System User Interfaces. In *Advances in Multimedia Information Processing — PCM 2001*, Lecture Notes in Computer Science (pp. 514-521). Heidelberg, Germany: Springer Berlin. Letzter Zugriff: 13. Februar 2009, unter http://dx.doi.org/10.1007/3-540-45453-5_66.

VOORHEES, E. M. (2005). *TREC*. Digital libraries and electronic publishing. Cambridge, USA: MIT Press.

WANG, J. Z. (2004). Wang Group: Modeling Objects, Concepts, and Aesthetics in Images. Letzter Zugriff: 11. Februar 2009, unter http://wang.ist.psu.edu/docs/related.shtml.

WANG, J. Z., LI, J., & WIEDERHOLD, G. (2001). SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(9), 947-963.

WANG, X., MA, W., HE, Q., & LI, X. (2004). Grouping web image search result. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 436-439). New York, USA: ACM. doi: 10.1145/1027527.1027632.

WESTON, J., MUKHERJEE, S., CHAPELLE, O., PONTIL, M., POGGIO, T., & VAPNIK, V. (2000). Feature selection for SVMs. *Advances in Neural Information Processing Systems*, *13*, 668-674. doi: 10.1.1.36.6307.

WIXOM, B. H., & TODD, P. A. (2005). A Theoretical Integration of User Satisfaction and Technology Acceptance. *Information Systems Research*, *16*(1), 85-102. doi: 10.1287/isre.1050.0042.

YAHOO! (2009). Popular Tags on Flickr. Letzter Zugriff: 13. Februar 2009 unter http://www.flickr.com/photos/tags/.

YANG, C., DONG, M., & FOTOUHI, F. (2005). Semantic feedback for interactive image retrieval. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 415-418). Hilton, Singapore: ACM. doi: 10.1145/1101149.1101240.

ZHOU, X. S., & HUANG, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, *8*(6), 536-544. doi: 10.1007/s00530-002-0070-3.

Anhang A: Fragebogen

Benutzerstudie im Fach Informationswissenschaft

Akzeptanz gegenüber automatischer Beschlagwortung und inhaltsbasierter Bildsuche

Florian Greiner

1 Einleitung

1.1 Begrüßung

Liebe Probandin, lieber Proband,
im Rahmen meiner Magisterarbeit im Fach Informationswissenschaft an der Universität Regensburg, soll die

> „Akzeptanz von Benutzern gegenüber inhaltsbasierter Bildsuche und automatischer Beschlagwortung“

untersucht werden.

Ich bitte Sie, mich bei dieser Evaluierung zu unterstützen und die Fragen spontan und vollständig zu beantworten. Selbstverständlich ist Ihre Teilnahme freiwillig und anonym. Das Bearbeitung aller Aufgaben dauert circa 30 Minuten. Die aus dem Fragebogen gesammelten Informationen werden von mir ausgewertet und dienen als Grundlage eines Forschungsprojekts.

Beachten Sie bitte Folgendes:

- Bei dem Test geht es um Ihre persönliche Einschätzung. Es gibt keine richtigen oder falschen Antworten.
- Überlegen Sie bitte nicht erst, welche Antwort den „besten Eindruck“ machen könnte, sondern antworten Sie spontan so, wie es für Sie persönlich gilt.
- Lassen Sie bitte keine Aussage unbeantwortet.

Ich danke Ihnen bereits im Voraus für Ihre Mitarbeit!
Florian Greiner

1.2 Einführung

Die Evaluation beschäftigt sich mit einem System zur automatischen Erfassung von Bildinhalten. Hierbei wird einerseits die Möglichkeit zur automatischen Annotation und andererseits die inhaltsbasierte Bildsuche im Fokus der Studie stehen.

Um Ihnen den Umgang mit dem System zu erleichtern, werden Sie während der Evaluation kurz mit dessen Bedienung vertraut gemacht. Sollten nach der Einführung noch Unklarheiten bezüglich der Bedienung des Systems bestehen, so wenden Sie sich bitte an den Versuchsleiter.

Ihnen werden im Laufe der Evaluation drei verschiedene Arbeitsaufgaben gestellt. Das System arbeitet mit einer englischsprachigen Datenbank. Daher bitte ich Sie die Aufgaben ebenfalls in englischer Sprache zu bearbeiten. Lesen Sie die Aufgaben gründlich durch und beginnen Sie im Anschluss daran mit deren zügiger Abarbeitung. Sollten vor oder während der Aufgabe Fragen auftreten, werde ich Ihnen diese natürlich umgehend beantworten.

Nach dem erfolgreichen Abschluss einer Aufgabe bitte ich Sie, mir dies kurz mitzuteilen und daraufhin mit der Beantwortung des zugehörigen Fragenteils zu beginnen. Haben Sie die Fragen beantwortet, können Sie mit der Bearbeitung der nächsten Aufgabe fortfahren.

Während der Arbeit an den Aufgaben wird ihre Bildschirmeingabe aufgezeichnet. Die Aufzeichnungen werden selbstverständlich vertraulich und anonym behandelt.

2 Frageteil

2.1 Automatische Annotation

Sie arbeiten als freier Bildjournalist für eine Lokalzeitung. Im Rahmen Ihrer Arbeit haben sie eine große Bildsammlung aufgebaut. Um die Bilder leichter auffindbar zu machen, beschließen Sie die Aufnahmen mit Schlagwörtern zu versehen.

(a)

(b)

(c)

(d)

Abbildung 1: **Bildsammlung**

1. Schreiben Sie die für sie passenden Schlagwörter zu den Abbildungen auf.

(a) _____

(b) _____

(c) _____

(d) _____

Anmerkung: *Sie erhalten nun eine kurze Einführung in die Bedienung des ALIPr-Systems.*

Nachdem Sie die vier Abbildungen aus Ihrem Fundus bereits für Artikel verwendet haben, wollen Sie nach ähnlichen Bildern suchen. Dafür benutzen Sie das kürzlich von Ihnen entdeckte ALIPr-System. Anhand der automatisch von *http://www.alipr.com* generierten Schlagwörter kann eine Suchanfrage nach ähnlichen Bildern gestartet werden.

1. Markieren Sie erst alle für sie zutreffenden Schlagwörter.

2. Bewerten Sie folgende Aussagen. $+++$ steht für „stimme voll und ganz zu“ und $---$ für „stimme überhaupt nicht zu“. Zutreffendes bitte in dem dafür vorgesehenen Feld ◯ ankreuzen.

Die vom Programm vorgeschlagenen Wörter sind zutreffend für den Inhalt des Bildes.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Schlagwörter helfen das Bild einem Thema zuzuordnen.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Menge aller vom Programm angegebenen Schlagwörter ist ausreichend.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die gefundenen Schlagwörter hätte ich auch gewählt.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Man versteht was die Schlagwörter bedeuten.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die gefundenen Schlagwörter helfen mir, noch bessere Schlagwörter zu vergeben.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Ein Verfahren, das meinen Bildern automatisch Schlagwörter zuordnet, würde ich für die Beschlagwortung von Bildern auch benutzen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die vom Programm angegebenen Schlagwörter ermöglichen eine eindeutige Unterscheidung zu ähnlichen Bildern.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Automatische Beschlagwortung stellt eine Erleichterung bei der Beschlagwortung von Bildern dar.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die gefundenen Schlagwörter beschreiben alle Bestandteile des Bildes.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Automatische Beschlagwortung erspart mir Zeit bei der Arbeit mit Bildern

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die gefundenen Schlagwörter sind einfach in richtige und falsche zu unterteilen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Automatische Beschlagwortung wäre mir in meinem Arbeitsumfeld von Nutzen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

2.2 Tagbasierte Bildsuche

Fahren sie fort indem Sie auf „submit“ klicken.

1. Wie viele der ersten zehn angezeigten Ergebnisse erachten sie als relevant?
(a):
(b):
(c):
(d):

2. Bewerten Sie folgende Aussagen. $+++$ steht für „stimme voll und ganz zu“ und $---$ für „stimme überhaupt nicht zu“. Zutreffendes bitte in dem dafür vorgesehenen Feld ◯ ankreuzen.

Die Bildsuche ist effizient.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Bildsuche stellt eine Erleichterung bei der Suche nach Bildern dar.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Bildsuche liefert Ergebnisse, die für meine Nachforschungen relevant sind.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Bildsuche liefert Ergebnisse in angemessener Zeit.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Ergebnisse der Bildsuche sind einfach in richtige und falsche zu unterscheiden.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die Bildsuche liefert ausreichend unterscheidbare Ergebnisse zum Thema.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die gefundenen Bilder liefern neue Erkenntnisse zu meinem Thema.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die gefundenen Bilder entsprechen dem was ich mir vorgestellt habe.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Eine solche Bildsuche wäre mir bei der Arbeit mit Bildern von Nutzen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

2.3 Inhaltsbasierte Bildsuche

Das ALIPr-System ist weiterhin in der Lage, eine sogenannte inhaltsbasierte Bildsuche durchzuführen. Dabei wird der Bildinhalt hinsichtlich Kriterien wie Farbe, Form und Struktur durchsucht und daraufhin Bilder mit ähnlichen Inhalten gesucht. Die automatisch ermittelten Schlagwörter werden dabei außer Acht gelassen.

1. Starten Sie die inhaltsbasierte Suche durch klicken auf „search for visually similar pictures“.

2. Wie viele der ersten zehn angezeigten Ergebnisse erachten sie als relevant?

(a):

(b):

(c):

(d):

2. Bewerten Sie folgende Aussagen. $+++$ steht für „stimme voll und ganz zu“ und $---$ für „stimme überhaupt nicht zu“. Zutreffendes bitte in dem dafür vorgesehenen Feld ◯ ankreuzen.

Inhaltsbasierte Bildsuche ist effizient.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Inhaltsbasierte Bildsuche stellt eine Erleichterung bei der Suche nach Bildern dar.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Inhaltsbasierte Bildsuche liefert Ergebnisse, die für meine Nachforschungen relevant sind.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Inhaltsbasierte Bildsuche liefert Ergebnisse in angemessener Zeit.

$+++$	$++$	$+$	$-$	$--$	$---$
◯	◯	◯	◯	◯	◯

Die inhaltsbasierte Bildsuche ist einfach zu benutzen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die Ergebnisse der inhaltsbasierten Bildsuche sind einfach in richtige und falsche zu unterscheiden.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die inhaltsbasierte Bildsuche liefert ausreichend unterscheidbare Ergebnisse zum Thema.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Der Ablauf einer inhaltsbasierten Bildsuche ist einfach zu erlernen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die gefundenen Bilder liefern neue Erkenntnisse zu meinem Thema.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Die gefundenen Bilder entsprechen dem was ich mir vorgestellt habe.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Der Ablauf von inhaltsbasierter Bildsuche ist nachvollziehbar.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

Inhaltsbasierte Bildsuche wäre mir bei der Arbeit mit Bildern von Nutzen.

+++	++	+	−	−−	−−−
○	○	○	○	○	○

4. Welche der beiden dargestellten Suchmethoden würden Sie bevorzugen? Verwenden Sie die Abkürzung *TB* für die tagbasierte Methode und *IB* für die inhaltsbasierte Methode.

(a):

(b):

(c):

(d):

5. Was hat Ihnen am besten bezüglich der Suchmethoden gefallen?

6. Was hat Sie am meisten bezüglich der Suchmethoden gestört?

7. Haben Sie noch weitere Anmerkungen oder spezielle Anregungen zu den Suchmethoden?

2.4 Demografischer Teil

Zum Schluss noch einige Fragen zu Ihnen.

1. Wie alt sind Sie?

2. Geschlecht? männlich ○ weiblich ○

3. Welchen Beruf üben sie aus?

Vielen Dank für Ihre Mitarbeit!

Erklärung

Ich versichere, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Hilfsmittel angefertigt zu haben.

Regensburg, den 23. Februar 2009