Inhaltsbasierte Bildsuche
(Content based image retrieval)

Ausarbeitung zum Seminar

Retrieval und Klassifikation von Multimediadaten

am Lehrstuhl Informatik IV
der RWTH Aachen

Von Benjamin Molitor
Matrikelnummer 217123
Betreuer: Daniel Keysers

25. Juni 2001

Inhaltsverzeichnis

1 Einführung

Dieser Ausarbeitung liegt primär ein Paper von Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta und Ramesh Jain unter dem Titel „Content Based Image Retrieval at the End of the Early Years“ aus dem Dezember 2000 zu Grunde, und daraus hervorgehend noch einige weitere Dokumente zu diesem Thema als Sekundärliteratur ([2], [3]). Es handelt sich dabei um einen Versuch, möglichst übersichtlich den Stand der Dinge im Bereich der inhaltsbasierten Bildsuche darzustellen. Aufgrund der riesigen Anzahl von Veröffentlichungen in den letzten Jahren ist das, wie im Paper auch dargestellt, keine leichte Aufgabe, und es wird auch keinesfalls Anspruch auf Vollständigkeit erhoben. Diese Ausarbeitung soll einige der zentralen Ideen und Tendenzen in der inhaltsbasierten Bildsuche vorstellen, ohne dabei zu ausführlich auf mathematische Hintergründe und Formeln einzugehen, und ohne sich zu sehr auf bestimmte einzelne Methoden und Ansätze zu konzentrieren. Zum Schluss wird ein kleiner Ausblick auf zukünftige (oder zum Teil sicherlich schon gegenwärtige) Entwicklungen und Forschungsgebiete gegeben.

1.1 Was ist inhaltsbasierte Bildsuche?

Inhaltsbasierte Suche von Bildern ist, im Gegensatz zu der durch das Internet weit verbreiteten textbasierten Suche, die Suche nach bestimmten Bildern in Bilddatenbanken anhand der Bilddaten selbst. Nicht Dateinamen, nicht verbale Beschreibungen und Labels von Bildern werden betrachtet, sondern die Informationen, die das Bild an sich enthält. Was macht nun diese spezielle Art der Suche so besonders interessant? Ein Sprichwort gibt - in der Essenz - die Antwort auf diese Frage:

„Ein Bild sagt mehr als 1000 Worte“ .

Es kommt oft vor, dass ein Bild in seiner Gesamtheit verbal nicht ausreichend beschrieben werden kann. Beispielsweise enthält praktisch jedes Bild eines Malers dessen ganz persönlichen Stil, der oft nur schwer in Worte gefasst werden kann, aber wenn man ein solches Bild sieht, wird man als Kenner eben doch den Künstler identifizieren können. Eine mündliche Beschreibung kann immer nur einen Teil der Information eines Bildes wiedergeben, und um zum Beispiel eine Auswahl zwischen Bildern für einen Zeitungsartikel zu treffen, wird man sich immer die betreffenden Bilder vorher anschauen wollen. Daher ist es von großem Interesse, Methoden zur Suche von Bildern in Datenbanken zu entwickeln, die sich an den Bildern selbst orientieren, zumal alternative Methoden oft wenig sinnvoll sind. Das Versehen von Bildern mit Schlagworten beispielsweise ist sehr aufwendig, bei großen Mengen von Bildern sogar beinahe unmöglich durchzuführen, da es nicht automatisiert werden kann, sondern von Hand gemacht werden muss.

1.2 Historischer Überblick

Heutzutage ist die inhaltsbasierte Bildsuche ein expandierendes Gebiet der Forschung. Nachdem zu Beginn schnelle Fortschritte in einigen speziellen Anwendungen erzielt werden konnten, konzentriert man sich heute auf die tiefer gehenden, schwierigeren Probleme.

Wahrscheinlich erstmals ausführlicher im Rahmen einer Konferenz besprochen wurden die zugrunde liegenden Ideen 1979 in Florenz, während einer Konferenz über Datenbankapplikationen von bildbezogenen Anwendungen. Viel später, 1992, wurden die zukünftigen Forschungsschwerpunkte im Rahmen eines Workshops der US National Science Foundation ziemlich genau identifiziert, wie sich in den kommenden Jahren zeigen sollte. Vor 1990 gab es kaum relevante Veröffentlichungen zur inhaltsbasierten Bildsuche, aber seit 1997 ist eine kaum überschaubare Anzahl von Arbeiten erschienen, so dass es den Rahmen dieser Arbeit schnell überschreiten würde wollte man versuchen, einen umfassenden und vollständigen Überblick zu geben. Die vorgestellten Themen sind also als eine Auswahl zu verstehen.

2 Anwendungsbereich

Für die inhaltsbasierte Bildsuche existieren viele Anwendungsmöglichkeiten, und dementsprechend auch eine große Anzahl unterschiedlicher Ansätze und Methoden. In diesem Kapitel möchte ich einen kurzen Überblick über diese geben und eine Einteilung der Anwendungsgebiete vorstellen.

2.1 Arten der Anwendung

Ein breites Spektrum von Methoden und Systemen beschäftigt sich mit dem Durchsuchen von großen Bildmengen aus nicht genauer spezifizierten Quellen. Bei der sogenannten *assoziativen Suche* hat der Benutzer zu Beginn nicht unbedingt eine klare Vorstellung dessen, was er sucht, sondern er beginnt die Suche einfach mit dem allgemeinen Ziel, etwas Interessantes zu finden. Das bedeutet, dass die Suchparameter im Verlauf der Suche verfeinert und angepasst werden müssen, dass nach Ähnlichkeiten anhand von Beispielbildern oder Skizzen gesucht werden kann, und so weiter. Eine solche Suche verläuft in hohem Maße interaktiv, da im optimalen Fall der Benutzer dem System auch ein Feedback darüber liefern sollte, wie gut die gefundenen Bilder den Suchkriterien entsprechen, so dass das System in der Lage ist, die Suchmethode und die Vorlieben des Benutzers in gewissem Maße zu erlernen (da die Suche ansonsten letztendlich nur ein reines Durchblättern der Datenbank wäre). Eines der ältesten Beispiele für ein solches System wird von T. Kato, T. Kurita und anderen in [2] beschrieben: der Benutzer fertigt eine grobe Skizze an, die dann vom System analysiert und auf ihre essentiellen Inhalte reduziert wird, bevor sie schließlich mit den Bildern in der Datenbank beziehungsweise deren Reduktionen verglichen wird, um ähnliche Bilder zu finden.

Bei einer anderen Art der Suche, der sogenannten *zielgerichteten Suche* oder *target search* möchte der Benutzer normalerweise ein ganz bestimmtes Bild finden, zum Beispiel ein spezielles Gemälde aus einem Kunstkatalog oder ähnliches. Eine Variation, die sogenannte zielgerichtete Suche anhand eines Beispiels, sucht nach Bildern die einen bestimmten Gegenstand enthalten. Mann könnte sich beispielsweise vorstellen, dass dem Suchsystem ein oder mehrere Bilder eines Stuhls gezeigt werden, und das System soll dann die Datenbank nach weiteren Bildern, die einen Stuhl enthalten, durchsuchen. Der wichtigste Unterschied zur assoziativen Suche besteht hier darin, dass der Benutzer vorher schon eine recht genaue Vorstellung dessen hat, was er eigentlich sucht.

Die in dieser Einteilung letzte Variante ist die *kategorieorientierte Suche*. In diesem Fall wird üblicherweise nach Bildern gesucht, die in eine durch irgendwelche Parameter bestimmte Kategorie fallen. Ein vorstellbares Szenario wäre hier, dass der Benutzer bereits einige Bilder aus einer Kategorie hat und nach weiteren dazu gehörigen Bildern suchen möchte, oder es wird ganz allgemein die zu suchende Kategorie - zum Beispiel „Bilder, die Wasser enthalten" - angegeben und das System liefert einen oder mehrere zufällig ausgewählte Repräsentanten aus der Datenbank.

Die gerade beschriebene Einteilung in drei verschiedene Bereiche ist, wie gesagt, nur eine von vielen möglichen. Eine andere Sichtweise identifiziert fünf typische Arten der Benutzung eines Bildsuchsystems:

- Suche nach einem bestimmten Bild
- Generelles Durchsuchen, um eine interaktive Auswahl zu treffen
- Suche nach einem Bild, das zu einem allgemeinen Thema passt
- Suche nach Bildern, um einen Text zu illustrieren
- Suche nach ästhetischen Gesichtspunkten

Aber auch das ist nur eine weitere von vielen anderen möglichen Einteilungen.

2.2 Der Bilderbereich (*image domain*)

Wenn man die Menge aller Bilder betrachtet, die für eine Applikation oder ein System von Bedeutung sind, bietet der Begriff des sogenannten Bilderbereiches eine Möglichkeit, eine gewisse Einteilung vorzunehmen. Je nachdem, wie der Bilderbereich für eine gegebene Bilddatenbank oder eine Suchanfrage einzuordnen ist, kann man Entscheidungen über die verwendbaren Suchmethoden treffen. Dabei kommt es selten vor, dass diese Einordnung ganz klar und eindeutig ist, sondern sie liegt zumeist irgendwo zwischen den beiden Grundkategorien, dem *engen* und dem *weiten* Bilderbereich.

Enger Bilderbereich (*narrow domain*): In einem engen Bilderbereich haben alle relevanten Aspekte der betrachteten Bilder nur eine geringe und gut vorhersagbare Variabilität. Damit einher geht normalerweise, dass auch die Bedeutung der Bilder klar definiert ist. Erforderlich für einen engen Bilderbereich sind möglichst identische Aufnahmebedingungen für alle Bilder, ohne Verdeckungen oder andere Störungen. Die Bildsuche hier ist eher als eine Klassifikation von Daten zu betrachten. Ein Beispiel wäre eine Menge von Frontalaufnahmen von Gesichtern, die alle vor einem hellen Hintergrund bei gleichmäßiger Beleuchtung gemacht wurden.

Weiter Bilderbereich (*broad domain*): Der weite Bilderbereich befindet sich am entgegengesetzten Ende des Spektrums: die Variabilität der relevanten Eigenschaften ist unbeschränkt und unvorhersagbar, selbst für Bilder, die von ihrer Bedeutung her identisch sind. Zusätzlich können Bilder auch mehrere Bedeutungen haben, oder ihre Semantik ist möglicherweise nur teilweise beschrieben. Bei der Bildsuche in einem weiten Bilderbereich kann man eher von Retrieval sprechen, und die meisten interessanten Anwendungen zum Thema Bildsuche sind auch in diesem Bereich zu finden.

Beispiele für einen weiten Bilderbereich sind Sammlungen von Bildern aller Art ohne bestimmtes Schema, und die weiteste und allgemeinste Klasse von Bildern ist die Menge aller Bilder, die im Internet verfügbar sind.

2.3 Die sensorische Kluft (*sensory gap*)

Der Begriff der sensorischen Kluft beschreibt das Problem, dass zwischen dem tatsächlichen Objekt in der realen Welt und der Information, die aus einer Aufnahme dieses Objektes oder der Szene gewonnen wird, oft ein großer Unterschied besteht. Je weniger über die genauen Aufnahmebedingungen bekannt ist, desto schwerwiegender wird dieses Problem, denn es ist durchaus vorstellbar, dass ein Bild je nach Beleuchtung sehr unterschiedlich aussehen kann, auch wenn es immer dasselbe Objekt enthält. Ein anderer Grund für die Existenz dieser Kluft ist, dass die Aufnahmen verschiedener 3-dimensionaler Objekte als 2-dimensionale Bilder identisch sein können, da die Information über die dritte Dimension verloren geht. Ein Ziel von inhaltsbasierten Bildsuchsystemen muss daher sein, diesen Graben so gut wie möglich zu überbrücken, das heißt zum Beispiel durch möglichst gute Kenntnis des Bilderbereiches und der Aufnahmebedingungen zu erreichen, dass so wenig Fehlinterpretationen wie möglich zustande kommen. Eine Einteilung in folgende Gebiete kann nützlich sein, um dieses Wissen über den Bilderbereich strukturiert betrachten und auswerten zu können:

syntaktisch: Gleichheit oder Ähnlichkeit im Buchstäblichen (bzw. pixelbezogenen) Sinn, beispielsweise anhand von Farbwerten.

bezüglich menschlicher Wahrnehmung: Gleichheit oder Ähnlichkeit nach den Gesetzen menschlicher Wahrnehmung. Der RGB-Farbraum kann daher hier nicht verwendet werden, es müssen die zu diesem Zweck entworfenen CIE-Lab oder Munsell-Farbräume benutzt werden.

physikalisch: Physikalische Gesetze betreffend Beleuchtung, Oberflächenreflexion etc. werden für das Feststellen von Ähnlichkeiten zu Rate gezogen.

geometrisch: Geometrische und topologische Regeln, wie beispielsweise die, dass Objekte kleiner werden, je näher sie dem Horizont (fast immer eine horizontale Linie über das gesamte Bild) sind, oder dass der Horizont eine virtuelle Linie ist, die alle perspektivischen Fluchtpunkte enthält.

kategorisch: Fast nur in schmalen Bilderbereichen verwendbar: Ähnlichkeit durch Einordnen in Kategorien, z.B. die allgemeine Kategorie „Teekanne“ mit den Merkmalen Henkel, Nase, ...

kulturell: Kulturelle und auch sprachliche Aspekte werden zur Erkennung von Ähnlichkeiten benutzt. Zum Beispiel haben Werkzeuge eine bestimmte Größe, um auch benutzbar sein zu können, und Innenräume haben viele rechte Winkel.

Generell kann man die einfache Regel aufstellen, dass ein enger Bilderbereich eine gesonderte Betrachtung dieser Bereiche vereinfacht, während ein weiter Bilderbereich sie entsprechend erschwert.

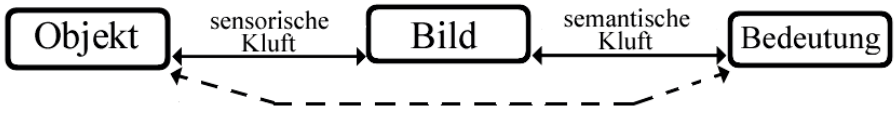

Abbildung 1: Sensorische und semantische Kluft

2.4 Die semantische Kluft (*semantic gap*)

Die semantische Kluft kommt dann zum Tragen, wenn die Information, die aus einer Szene gewonnen wird und die Interpretation, die ein Benutzer aus denselben Daten zieht (beziehungsweise die er dem abgebildeten Objekt beimisst), nicht übereinstimmen. Der Grund für ihre Existenz ist zugleich die Erklärung, warum eine Überbrückung nur schwer zu erreichen ist: ein Bild mündlich vollständig zu beschreiben ist eine wahrscheinlich unmögliche Aufgabe, und auch die datenbezogene Interpretation, die ein Bildsuchsystem macht ist im Grunde nichts anderes als eine abstrakte Beschreibung des Inhalts, und daher nie eindeutig (eine Ausnahme bilden vielleicht sehr enge Bilderbereiche). Auch können gleiche Bilder oder Objekte für verschiedene Menschen ganz unterschiedliche Bedeutungen haben, je nach ihren ganz persönlichen Erfahrungen, Meinungen und so weiter. Die semantische Kluft ist, laut [1], der Hauptgrund dafür, dass Bildsuchsysteme oft noch recht unbefriedigende Ergebnisse liefern, da zwar schon sehr viel Arbeit und Zeit für die Interpretation der Bilddaten auf höherer Ebene verwendet wurde, aber das Grundlegende Problem der eindeutigen Objekterkennung in einem einzelnen Bild noch immer nicht zufriedenstellend gelöst ist. Abbildung 1 auf Seite 6 veranschaulicht noch einmal den Bezug zwischen sensorischer und semantischer Kluft.

Die eigentliche Herausforderung für Bildsuchsysteme, die auf einem weiten Bilderbereich arbeiten, wird hier als die Aufgabe beschrieben, das Suchsystem durch die Auswertung des Feedbacks des Benutzers an den engen Bildbereich anzupassen, den der Benutzer vor Augen hat, und damit überhaupt erst realistische Bedingungen für eine wirkungsvolle Überbrückung sowohl der sensorischen als auch der semantischen Kluft zu schaffen.

3 Bildverarbeitung

Inhaltsbasierte Bildsuche benötigt keine Beschreibung des Inhalts eines Bildes in seiner Gesamtheit (was ja auch, wie bereits erwähnt, nur schwer möglich ist), sondern Ziel ist lediglich, ähnliche Bilder in einem von Benutzer definierten Sinn zu finden. Die Beschreibung des Inhaltes sollte demnach primär diesem Ziel dienen. Fast immer findet darum als erster Schritt bei der Bildsuche eine Vorverarbeitung der Bilddaten statt, die die im jeweiligen System als relevant eingestuften Daten auf eine interne Datenstruktur abbildet, anhand derer dann die eigentlichen Vergleiche und die Suche stattfindet. Als hilfreich in diesem Zusammenhang hat sich das Benutzen von Invarianzen herausgestellt, also von Merkmalen und Eigenschaften in Bildern, die sich für bestimmte Objekte oder Bildinhalte nicht oder nur sehr wenig verändern, selbst wenn die Aufnahmebe-

dingungen variieren, und so zur Identifikation dieser Objekte beitragen können. Invarianzen kann man in allen der im folgenden vorgestellten Bereiche finden, die zumeist für eine Vorverarbeitung der Bilddaten betrachtet werden.

3.1 Farbe

Farbe spielt bei der Verarbeitung von Bilddaten eine große Rolle, schon allein weil der 3-dimensionale Farbraum ein sehr großes Unterscheidungspotential im Vergleich zum 1-dimensionalen Bereich der Grauwerte bietet. Die in Bildern auftretenden Farben sind von den unterschiedlichsten Faktoren abhängig, und können daher auch für gleiche Objekte stark variieren. So sind zum Beispiel die Ausrichtung eines Objektes in Relation zu der oder den Lichtquellen von Bedeutung, die Farbe und Intensität der Beleuchtung und auch bestimmte Materialeigenschaften wie beispielsweise Lichtreflexion. Das allgemein bekannte RGB-Farbmodell kann gerade aus diesen Gründen auch nur bedingt erfolgreich eingesetzt werden, da es nicht an der subjektiven, menschlichen Wahrnehmung im psychologischen Sinn orientiert ist, sondern an den drei verschiedenen Arten von Lichtrezeptoren im Auge (rot, grün, blau), also der rein physikalischen Wahrnehmung von Farben. Oft bringt die Benutzung von Gegenfarbmodellen, die einen Farbwert nicht mehr durch (R, G, B), sondern durch (R-G, 2B-R-G, R+G+B) darstellen, schon eine deutliche Verbesserung in Bildsuchsystemen. Diese Repräsentation hat den großen Vorteil, dass durch ein verstärktes Augenmerk auf die dritte Komponente R+G+B, die die Information über die Helligkeit enthält, gegenüber den ersten beiden Komponenten, die die Farbnuance wiedergeben, die Farbinformation besser aus menschlicher Sicht ausgewertet werden kann. Das kommt daher, dass für Menschen die Helligkeit bei der Wahrnehmung von Farbunterschieden eine größere Rolle spielt als der Farbton. In anderen Fällen sind andere Farbmodelle nützlicher, zum Beispiel der bereits erwähnte Lab-Farbraum. Im Lab-Farbmodell beispielsweise werden Farben so repräsentiert, dass ihre euklidische Distanz der menschlichen Wahrnehmung von Farbunterschieden entspricht.

3.2 Lokale Formen

Bildverarbeitung unter dem Aspekt von lokalen Formen dient dem Zweck, möglichst viele auffällige (und damit potenziell wichtige) geometrische Details im Bild zu finden und so zu verstärken, dass sie im eigentlichen Suchprozess zu besseren Ergebnissen führen. Hier sollte auf jeden Fall angemerkt werden, dass sich lokale Formen nicht betrachten lassen, ohne farbliche Aspekte zu berücksichtigen, da es oft gerade Farbkontraste sind, die Kanten und damit Formen und eventuell sogar Objekte erst erkennbar machen. Aus diesem Grund ist es in den seltensten Fällen sinnvoll, bei der Vorverarbeitung der Bilddaten eine strenge Trennung zwischen Farbe und Form zu machen; in [1] wird sogar verstärkt eine integrierte Sichtweise (auch unter Einbeziehung von Textur) gefordert, um die Voraussetzungen für brauchbare Unterscheidungskriterien in Bilddatenbanken mit einer großen Anzahl von Bildern zu schaffen.

3.3 Textur

Der Begriff „Textur“ wird in der Literatur nicht einheitlich verwendet, laut [1] wird aber als Textur im Allgemeinen die Struktur der einzelnen Teile eines Bildes bezeichnet, die auch darin bestehen kann, dass gewisse Bereiche ganz ohne erkennbares Muster (also zufällig texturiert) sind. Was Texturen von den gerade beschriebenen lokalen Formen unterscheidet ist zumeist ihre Größe und Anzahl: sie sind meist zu klein und zahlreich, um noch effektiv einzeln betrachtet werden zu können. Um die Textur einzelner Bereiche eines Bildes zu beschreiben (und damit auch vergleichbar zu machen) wird die meiste Forschungsarbeit derzeit in Richtung statistischer oder generativer Methoden geleistet. Anwendung finden texturbasierte Methoden unter anderem bei der Auswertung von Satellitenbildern, wo die Textur zum Beispiel Auskunft über Vegetation geben kann. Auch hier sollte zur Betonung noch einmal angemerkt werden, dass Farbe, Form und Textur eigentlich nicht getrennt voneinander betrachtet werden können, sondern möglichst immer im Zusammenhang.

4 Features

Bei der inhaltsbasierten Bildsuche werden die Bilder meistens zuerst nach bestimmten Kriterien in Bereiche unterteilt, bevor dann Eigenschaften oder Features dieser einzelnen Bereiche berechnet und als Beschreibung ihres Inhaltes gespeichert werden. Der Begriff „Feature“ (dt.: Merkmal) beschränkt sich allerdings nicht nur auf diesen Bereich. So bezeichnet man auch zum Beispiel die Farbwerte eines Bildes oder lokale Textureigenschaften als Merkmale. Die Unterteilung von Bildern in Bereiche kann auf sehr unterschiedliche Art erfolgen, je nach Bedarf und Möglichkeiten. Oft wird die Komplexität und Schwierigkeit, aber auch die Wichtigkeit gerade dieses Arbeitsschrittes unterschätzt. Viele der in den letzten Jahren erschienenen Veröffentlichungen beschäftigen sich vor allem damit, wie die Bildsuche *nach* der Berechnung einer geeigneten Segmentierung effektiv zu gestalten ist, und lassen dabei außer Acht, dass eben diese Unterteilung der Bilddaten noch immer ein in weiten Teilen offenes Problem darstellt, für das es keinen einfachen oder allgemein gültigen Ansatz gibt.

4.1 Gruppierung der Daten

Eine Methode, deren Anwendung sehr zur Erleichterung der Interpretation des Bildinhalts beiträgt, die aber meist sehr schwer durchzuführen ist, ist die sogenannte *starke Segmentierung*: Wenn man auf der Suche nach einem bestimmten Objekt ist, wäre es natürlich besonders wünschenswert, eben dieses Objekt im Bild komplett als eigenen Bereich zu erkennen und zu markieren, also alle Pixel im Bild auszuwählen, die das Objekt darstellen (wenn T der markierte Bereich im Bild und O das tatsächliche Objekt in der realen Welt ist, soll hier $T = O$ gelten). Typischerweise kann man dieses Verfahren aber nur in sehr schmalen Bilderbereichen anwenden, wo die Anzahl der möglichen Interpretationen sehr beschränkt ist und die Sicherheit, bei der Segmentierung keinen Fehler zu begehen, ausreichend groß ist (eine falsche Auswahl führt bei starker Segmentierung sehr schnell zu einer falschen Interpretation des Bildes, da die semantische Bedeutung sehr eng mit dem ausgewählten Segment verbunden beziehungsweise eindeutig für bestimmte Segmentierungen ist).

Wenn starke Segmentierung nicht erreichbar ist oder ihre Anwendung zu oft zu Fehlern führt, kann man sich eventuell mit der sogenannten *schwachen Segmentierung* helfen: Bei diesem Verfahren werden in sich auf eine bestimmte Weise homogene Regionen (zum Beispiel in Hinsicht auf Farbe, Textur, Form, usw.) des Bildes ausgewählt, in der Hoffnung, dabei Teile der tatsächlichen Objekte zu markieren (in der oben eingeführten Notation möchte man also erreichen, dass $T \subset O$ ist). Ein wichtiger Unterschied zur starken Segmentierung ist hier, dass es durchaus möglich ist, dass nicht das gesuchte Objekt insgesamt erkannt und markiert wird, sondern nur Teile davon. Damit ist die schwache Segmentierung auch dann anwendbar, wenn Teile eines Objektes im Bild verdeckt oder aus anderen Gründen nicht zu sehen sind (starke Segmentierung versagt in diesem Fall). Sie wird in vielen Systemen benutzt, entweder um direkt zum Ziel bzw. zur Interpretation des Bildinhaltes zu gelangen oder als Vorverarbeitungsschritt für andere, fortgeschrittenere Methoden der Segmentierung.

Wenn ein Objekt eine beinahe konstante Form und Bedeutung hat, wie zum Beispiel ein Auge, eine Ampel oder auch Buchstaben, nennt man es ein *Zeichen* oder *Sign*. Zeichen sind hilfreich bei der Zuweisung einer Bedeutung zu einem gegebenen Bild, da sie semantisch eindeutig sind. Allerdings gilt auch hier wie bei der starken Segmentierung die strenge Bedingung, dass das Erkennen unter allen möglichen Umständen fehlerfrei funktionieren muss, denn sonst wird der Inhalt des Bildes schnell falsch interpretiert.

Die schwächste Form der Gruppierung von Bilddaten ist eine einfache *Partitionierung* des Datenfeldes, ohne jede Rücksicht auf Inhalt und Bedeutung, in der obigen Notation symbolisiert durch $T \neq O$. Dabei ist die Wahl der Einteilung letztendlich frei wählbar, möglich (wenn auch nicht unbedingt sinnvoll) wäre auch, das ganze Bild zu wählen. Dass aber auch dieses Verfahren seine Anwendung findet, liegt daran, dass bestimmte Gruppen von Bildern sich oft an gewissen Regeln orientieren. So dürften zum Beispiel viele Landschaftsaufnahmen der Konvention folgen, dass der Himmel etwa das obere Drittel des Bildes einnimmt, und damit würde eine Einteilung in oberes Drittel und untere zwei Drittel durchaus Sinn machen. Eine andere Art, wie diese Methode der Segmentierung angewendet wird, ist die Aufteilung des Bildes in Quadrate oder Rechtecke gleicher Größe, die dann jeweils gesondert auf dominante Features untersucht werden, die dann am Ende über das gesamte Bild aufsummiert werden.

Welche Art der Segmentierung (oder welche Kombination der beschriebenen Methoden) sinnvoll ist, hängt sehr stark vom betrachteten Bilderbereich ab und muss auf jeden Fall sehr sorgfältig bedacht werden, da bei schlechter Segmentierung die weiteren Arbeitsschritte deutlich erschwert oder gar unmöglich gemacht werden.

4.2 Globale und akkumulative Features

Als nächster Schritt werden für die gefundenen Segmente die sogenannten Features oder Merkmale berechnet. Eine erste, noch recht allgemeine Form solche Eigenschaften zu finden stellen akkumulative Features dar: sie sammeln im einfachsten Fall die rein räumliche Information, die durch die Segmentierung des Bildes entsteht, ohne ansonsten weiter auf die eigentlichen Bilddaten der einzelnen Segmente einzugehen. Ein Spezialfall davon sind die globalen Features, die aus dem gesamten Bild berechnet werden. Eine besonders einfache, aber den-

noch oft effektive Möglichkeit, globale Features zu sammeln und auszuwerten, ist die Verwendung von *Histogrammen*. Hier sind allerdings nicht nur Farbhistogramme gemeint, sondern auch Feature-Histogramme, die die Häufigkeit des Auftretens bestimmter Features in einem Bild festhalten und dabei jede räumliche Information verlieren. Ein großer Vorteil bei der Verwendung solcher Histogramme ist sicherlich, dass sie sehr robust gegenüber Translation - und in gewissem Maße auch Rotation - von Objekten sind: wenn ein bestimmtes Objekt durch eine einzigartige Kombination von Features gekennzeichnet ist, dann werden diese auch im Histogramm zu finden sein, gleichgültig wo sich das Objekt im Bild befindet. Wenn das Histogramm zusätzlich noch um eine oder mehrere Dimensionen erweitert wird, um beispielsweise auch räumliche Informationen zu speichern, werden die Möglichkeiten zum Vergleich von Bildern noch gesteigert. Allerdings sollte man auch bedenken, dass eine zu große Anzahl von Dimensionen die Arbeit mit dem Histogramm uneffektiv werden lassen kann, da sie nicht mehr schnell genug durchführbar ist.

Das Ziel der Berechnung von Merkmalen für ein Bild ist es meist, alle (relevanten) Informationen eines Bildes in Featurewerten zusammenzufassen, wobei allerdings einige Dinge beachtet werden müssen, damit dieses Verfahren auch für die Bildsuche von Nutzen ist. Zunächst ist es ein nicht zu unterschätzendes Problem zu erkennen, welche Informationen in einem Bild tatsächlich für die Interpretation relevant sind, denn sowohl das Auslassen relevanter Daten als auch das hinzunehmen irrelevanter Daten erschweren die Suche anhand der generierten Features. Um die Featurewerte auch effektiv nutzen zu können, müssen noch weitere Anforderungen erfüllt werden, die keinesfalls trivial sind, auf die ich aber nicht näher eingehen werde. Es ist möglich, durch diese Art der Sammlung von Bilddaten eine Kompression der Bildinformation zu erreichen, wenn nämlich die Bit-Größe der errechneten Features kleiner ist als die des ursprünglichen Bildes, was für den weiteren Verlauf der Suche auf Grund der geringeren zu verarbeitenden Datenmenge von Vorteil sein kann.

4.3 Hervorstechende Features (*salient features*)

Hervorgehobene Features werden vor allem zur Beschreibung von Bildern benutzt, die mittels schwacher Segmentierung unterteilt wurden. Dabei werden nur die hervorstechendsten Regionen beziehungsweise Segmente betrachtet und ausgewertet, um das Bild zu interpretieren. Im Extremfall wird die Information des Bildes sogar auf einige Punkte reduziert. Auch hier stellt die Auswahl von bestimmten Teilen des Bildes (welche Bereiche sind besonders hervorstechend?) und die gleichzeitige Vernachlässigung der restlichen Daten sowohl ein Problem als auch einen Vorteil dieser Methode dar: Zum einen wird der Inhalt des Bildes bei diesem Vorgehen oft sogar stark komprimiert, aber gerade aus diesem Grund ist es besonders wichtig, die Features mit großer Sorgfalt auszuwählen, damit sie auch die tatsächlich relevanten Informationen des Bildes enthalten. Anschaulich vorstellen kann man sich hervorstechende Features als die besonderen Punkte oder Bereiche eines Bildes, die am längsten noch erkennbar und vom Rest unterscheidbar sind, wenn man das Bild schrittweise verwischt. Dies ist aber nur eine Art, hervorstechende Features zu beschreiben. Abbildung 2 auf Seite 11 zeigt ein Beispielbild, in dem unterschiedliche Features markiert sind.

Wenn eine der möglichen Interpretationen eines Bildes oder eines Segmentes in einem Bild so sehr überwiegt, dass sie als *die* Bedeutung angesehen werden

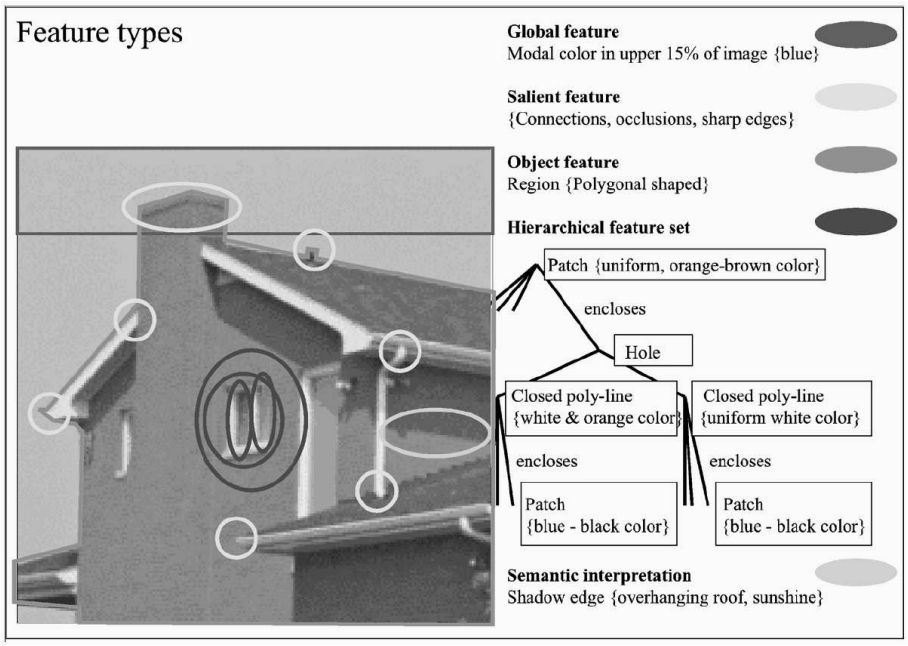

Abbildung 2: Beispielbild mit verschiedenen Feature-Arten

kann, dann spricht man von einem ein *Zeichen* oder *sign*. Zeichen sind besonders deswegen von großem Interesse, weil ihre Bedeutung eindeutig ist und damit auch die Interpretation des gesamten Bildes oft stark erleichtert wird. Beispiele für die Verwendung von Signs sind die automatische Erkennung von Symbolen auf Landkarten oder auch die Texterkennung in einem Bild („optical character recognition“ oder OCR).

4.4 Objekt-Features und Bild-Layout

Im ersten Teil dieses Kapitels wurde die starke Segmentierung vorgestellt. Wenn man nun ein Bild speziell im Hinblick auf die enthaltenen Objekte beschreiben oder die entsprechende Information im Bild verstärken möchte, bietet sich die starke Segmentierung auf den ersten Blick an, da sie tatsächlich ganze Objekte im Bild als einzelne Segmente markiert. Allerdings wurde auch auf die großen Schwierigkeiten hingewiesen, die eine automatische starke Segmentierung in weiten Bilderbereichen mit sich bringt, und aus diesem Grund sind verstärkt andere Wege gesucht worden, um die objektspezifischen Daten eines Bildes zu verarbeiten. Eine wichtige Feststellung in diesem Zusammenhang ist beispielsweise, dass es oft nicht wichtig ist, wo sich ein Objekt im Bild befindet, wenn man erkennt *dass* es im Bild vorhanden ist. Außerdem erlaubt die in vielen Fällen interaktive Suche auch, dass gelegentliche Fehler vorkommen, so dass große Präzision (die zumeist nur mit sehr zeitaufwendigen Methoden erreicht werden kann) gegen rechnerische Effizienz und damit Geschwindigkeit und Benutzbarkeit eingetauscht werden kann: Wenn ein Fehler gemacht wird, teilt der Benutzer dies dem System mit und formuliert damit gleichzeitig seine Anfrage genauer, was dann zu besseren Suchergebnissen führen sollte.

Ein Problem, das besonders bei der Suche nach (3-dimensionalen) Objekten in (2-dimensionalen) Bildern beziehungsweise bei der Beschreibung dieser Objekte aufgrund ihrer Konturen im Bild auftritt, ist der Blickwinkel: wenn irgendwie möglich sollten invariante Beschreibungen der Konturen gefunden werden, die natürlich - je nachdem vom welcher Seite man ein Objekt fotografiert - zunächst einmal sehr unterschiedlich sein können.

Wenn man nun eine geeignete Möglichkeit gefunden hat, einzelne Bereiche oder gar Objekte in einem Bild durch Features zu beschreiben, kann es auch interessant werden, zusätzlich zu ihrer reinen Existenz noch eine irgendwie geartete Beziehung zwischen diesen Features festzustellen und beim Speichern der Daten zu beachten. Das können zum Beispiel räumliche Verhältnisse sein, oder man könnte die Features nach einer bestimmten Hierarchie geordnet speichern. Diese Art der Beachtung der Anordnung von Objekten in einem Bild (des *Layouts*) kann vor allem dann von großer Bedeutung sein, wenn eben gerade die räumliche Position ansonsten gleicher Objekte zueinander die Bedeutung des Bildes ausmacht, und auch in allgemeineren Fällen dürfte dieser Ansatz zur Erhöhung der Unterscheidbarkeit verschiedener Bilder beitragen.

5 Interpretation und Ähnlichkeit

Erst wenn diejenigen der bis hierhin beschriebenen Schritte der Vorverarbeitung erfolgreich durchgeführt worden sind, die man in seinem Suchsystem benutzen möchte, kann man zum eigentlichen Suchvorgang übergehen: man muss den Features, die die Informationen der Bilder enthalten, auf irgendeine Art Bedeutung geben. Dafür gibt es grundsätzlich zwei Wege: entweder wird das Bild aufgrund seiner Features eindeutig interpretiert (im einfachsten vorstellbaren Fall bedeutet das, ob ein Bild einer Suchanfrage entspricht oder nicht), oder man definiert ein Ähnlichkeitsmaß und stellt zu jedem Bild in der Datenbank durch einen Vergleich der Features fest, wie ähnlich es dem gesuchten ist.

5.1 Semantische Interpretation

Um bei der inhaltsbasierten Bildsuche Erfolg zu haben sollte man versuchen, die inhaltliche Bedeutung des Bildes anhand der vorhandenen Features so weit wie möglich zu interpretieren. In den meisten Fällen - vor allem dann, wenn ein weiter Bilderbereich betrachtet wird - sollte man dabei jedoch immer beachten, dass man fast immer nur einen Teil aller möglichen Interpretationen finden kann. Ein wichtiges und zugleich schwer zu erreichendes Ziel ist es bei der Zuweisung einer Semantik daher, möglichst die Interpretationen zu finden, die für den aktuellen Kontext relevant sind. Eine schwächere Form der semantischen Interpretation versucht, um diesem Problem der Auswahl aus dem Weg zu gehen, aus den gesammelten Features nur annäherungsweise eine Teilmenge von möglichen Bedeutungen zu finden, die für das gegebene Suchszenario interessant sein könnten.

Ein gutes Beispiel, bei dem tatsächlich eine starke semantische Interpretation möglich und auch zumeist sinnvoll ist, sind Bilder, in denen Zeichen (*signs*) eindeutig identifiziert werden können, da diese wie bereits beschrieben zumeist eine eindeutige Interpretation des Bildinhaltes zulassen. Sogenannte schwache

semantische Interpretationen sind besonders in interaktiven Szenarien sinnvoll, da sich das Suchsystem hier durch das Feedback des Benutzers im Verlauf der Suche an dessen Anforderungen anpassen und so immer bessere Ergebnisse erzielen kann.

5.2 Ähnlichkeit von Features

Eine andere Möglichkeit besteht darin, die beobachteten Features nicht auf ihre inhaltliche Bedeutung hin zu untersuchen, sondern sie anhand einer Ähnlichkeitsfunktion, die zunächst einmal durch das Vorwissen über den Bilderbereich bestimmt wird, paarweise zu vergleichen. Im besten Fall haben dann Bilder, die aufgrund dieser Funktion als ähnlich eingestuft werden, auch eine ähnliche Bedeutung, aber das ist nur dann der Fall, wenn das Ähnlichkeitsmaß passend für die gegebene Menge von Bildern und möglichen Interpretationen gewählt wurde.

Ähnlichkeiten lassen sich auf unterschiedlichen Ebenen finden, von denen ich hier drei besonders hervorheben und etwas näher beschreiben möchte:

- Ähnlichkeit von *Objektsilhuetten*
- Ähnlichkeit von *Strukturen und Layout*
- Ähnlichkeit *hervorstechender Merkmale*

Die Silhuetten ganzer Objekte lassen sich dann besonders gut betrachten und vergleichen, wenn eine starke Segmentierung gelingt, aber das ist sehr selten der Fall. Auch die bereits bei der Generierung von Featurewerten für Formen und Objekte auftretenden Probleme, die besonders durch verschiedene Blickwinkel und auch durch Verdeckungen im Bild zustande kommen, erschweren diese Vorgehensweise. Der Ansatz, für bestimmte Formen invariante, robuste Features zu identifizieren und zu betrachten, könnte sich hier als hilfreich erweisen.

Für den Vergleich zwischen unterschiedlichen Layout-Features wird eine große Anzahl zum Teil sehr unterschiedlicher Methoden eingesetzt, die auch in [1] nur aufgelistet und kaum näher beschrieben werden. Grundsätzlich haben aber alle diese Methoden das Ziel, auf eine sinnvolle Art und Weise die Ähnlichkeit von Bildern im Bezug auf die topologische Anordnung von Objekten beziehungsweise Features im Bild zu definieren.

Wenn die hervorstechenden Features zweier Bilder miteinander verglichen werden sollen, stehen Bildsuchsystemen grundsätzlich mehrere Wege zur Verfügung. Zunächst einmal können die besonders hervorgehobenen Punkte oder Regionen zweier Bilder in Hinblick auf ihre Farbe, Textur und Form miteinander verglichen werden, wobei auch hier immer die Suche nach einer nützlichen Definition für Abstände eine zentrale Rolle spielt. Eine andere Möglichkeit besteht darin, alle hervorstechenden Punkte eines Bildes in einem Histogramm zu speichern, beispielsweise im Hinblick auf einige ausgewählte Eigenschaften wie die Farbe im Innern im Gegensatz zur Farbe außerhalb der betreffenden Bildregion. Verglichen wird dann nach dem Vorhandensein derselben Menge von hervorstechenden Features. Auch hier gibt es noch eine große Menge weiterer Ansätze.

6 Interaktion

Das Prinzip der interaktiven Suche wurde zu Beginn bereits vorgestellt. Um bei dieser Art der inhaltsbasierten Suche gute Ergebnisse erzielen zu können ist in hohem Maße die aktive Mitarbeit des Benutzers am Suchvorgang nötig, um die Suchkriterien ständig zu verbessern und den Vorstellungen des Benutzers anzupassen. Die Interaktion in diesem Fall ist ein komplexes Zusammenspiel zwischen dem Benutzer, den Bildern und ihrer semantischen Bedeutung.

6.1 Der Anfrageraum (*query space*)

Der Anfrageraum ist, ganz allgemein, definiert als das zielabhängige 4-Tupel $\{I_Q, F_Q, S_Q, Z_Q\}$. I_Q ist die für die Suchsitzung relevante Auswahl von Bildern aus dem großen Bildarchiv I. Typischerweise wird diese Auswahl anhand von Standardverfahren getroffen, wie zum Beispiel Bilder eines bestimmten Malers, Bilder von einer bestimmten Webseite etc. Die zweite Komponente F_Q beschreibt die Auswahl der für die Suchanfrage relevanten Features, als Teilmenge der gesamten Features F der betrachteten Bilder. Meist ist ein Benutzer nicht immer in der Lage, diese Auswahl selbst zu treffen, da er normalerweise nicht entscheiden kann, ob für die Beschreibung von Formen am besten Momente oder Fourier-Koeffizienten verwendet werden sollen, aber er sollte zumindest eine generelle Aussage über die Klasse der relevanten Features machen können (Form, Textur, ...). Desweiteren sollte der Benutzer eine Ähnlichkeitsfunktion S_Q auswählen, die sinnvollerweise Parameter wie zum Beispiel Gewichte für unterschiedliche Features enthalten sollte, um an verschiedene Bildmengen und Suchziele angepasst werden zu können. Z_Q schließlich ist eine Menge von Bezeichnern oder *Labels*, die die zielabhängige Semantik der Features beschreibt.

Zu Beginn einer Anfrage, wenn kein Vorwissen existiert, sollte der Anfrageraum Q_0 so initialisiert sein, dass er keine besonderes Vorlieben oder Ähnlichkeiten enthält; diese entwickeln sich dann erst im Verlauf der Interaktion mit dem Benutzer.

6.2 Spezifikation der Anfrage

Um eine Anfrage q in Q zu spezifizieren wurden viele unterschiedliche Interaktionsmethoden vorgeschlagen. Grundsätzlich lässt sich aber jede Anfrage in eine der beiden im Folgenden beschriebenen großen Kategorien einordnen. Zum einen gibt es die *exakte Anfrage*, die als Ergebnis eine Menge von Bildern liefert, die einer Anzahl von vorgegebenen Suchkriterien entsprechen. Exakte Anfragen können in drei unterschiedlichen Formen gestellt werden:

Exakte Anfrage anhand eines räumlichen Prädikats: Nur anwendbar in schmalen Bilderbereichen. Eine Anfrage könnte beispielsweise darin bestehen, dass nach Bildern gesucht wird, die eine Sonne oberhalb einer Wasserfläche zeigen.

Exakte Anfrage anhand von Bildprädikaten: Meistens eine Beschreibung globaler, also bildweiter Prädikate. Ein Beispiel wäre eine Anfrage nach Bildern mit „mehr als 50 % Himmel und mehr als 30 % Sand“.

Exakte Anfrage anhand von Gruppenprädikaten: Eine Anfrageform, die ein Element z aus Z_Q benutzt, wobei Z_Q ein Satz von Kategorien ist, der

I_Q partitioniert. Beispielsweise könnte man nach Bildern suchen, die in einer bestimmten Umgebung, z.B. Afrika, aufgenommen wurden.

Im Gegensatz dazu liefert eine *Näherungsanfrage* oder *approximate query* als Ergebnis nicht eine begrenzte Anzahl von Bildern, sondern eine Rangfolge der Bilder aus I_Q, geordnet nach ihrer Ähnlichkeit mit den spezifizierten Suchkriterien. Auch hier gibt es drei unterschiedliche Formen der Anfrage:

Näherungsanfrage anhand eines räumlichen Beispiels: Eine solche Anfrage führt zu Bildern, die einer vorgegebenen räumlichen Struktur entsprechen. Eine sinnvolle Möglichkeit ist beispielsweise, eine grobe Skizze (ein Kreis über einer horizontalen Linie zum Beispiel) als Beispiel anzugeben.

Näherungsanfrage anhand eines Bildbeispiels: Hier wird dem System einfach ein komplettes Bild präsentiert, und es wird dann auf die gewünschte Art nach ähnlichen Bildern im Feature-Raum gesucht. Unterschieden wird noch danach, ob das Beispielbild in I_Q enthalten ist (dann können die Beziehungen der Bilder untereinander schon vorberechnet werden) oder nicht.

Näherungsanfrage anhand eines Gruppenbeispiels: Die Anfrage wird durch eine Menge von Bildern spezifiziert, wobei überlicherweise positive *und* negative Beispiele gegeben werden.

Abbildung 3 auf Seite 16 gibt für jede dieser sechs genannten Anfragekategorien ein anschauliches Beispiel. Diese Einteilung stammt aus [1], es stellt sich jedoch die Frage, ob sie in dieser Form sinnvoll ist: auch eine exakte Anfrage muss ein Maß für die Ähnlichkeit benutzen, und dann zum Beispiel alle Bilder, die zu mehr als 80% dem Originalbild ähneln, als Suchergebnis zurückgeben.

6.3 Interaktion und Feedback

Wenn man exakte Anfragen betrachtet, ist es durchaus angebracht, den Ablauf der Anfragesession als einen iterativen Prozess zu beschreiben: in jedem Schritt erneuert der Benutzer seine Anfrage, oder er passt sie an. Bei Näherungsanfragen allerdings sollte man die interaktive Anfragesitzung in ihrer Gesamtheit betrachten, während derer das System den Anfrageraum Q ständig nach dem Feedback des Benutzers anpasst und damit verändert (diese Unterscheidung wird in [1] ausdrücklich gemacht, ist aber nicht unbedingt ganz einsichtig). Bei der Interaktion besteht die Aufgabe des Benutzers vor allem darin, dem System ein Relevanz-Feedback zu geben: Wie gut entsprechen die gefundenen Bilder den spezifizierten Kriterien? Auf die Frage, wie dieses Feedback geschieht, gibt es viele Antworten, die jedoch alle gemeinsam haben, dass sie einen sinnvollen Mittelweg zwischen möglichst großem Informationsgehalt und möglichst kleinem Aufwand für den Benutzer suchen. Abbildung 4 auf Seite 17 zeigt ein Beispiel für das Ergebnis einer Anfrage, bei der im ersten Schritt die rot markierten Segmente als positive Beispiele für das Merkmal *Blätter* gegeben sind. Das Suchsystem erkennt alle grün markierten Segmente ebenfalls als Blätter. Im nächsten Schritt (rechtes Bild) sind die rot markierten Segmente negative Beispiele (keine Blätter), und das System liefert ein mit weniger Fehlern behaftetes Ergebnis.

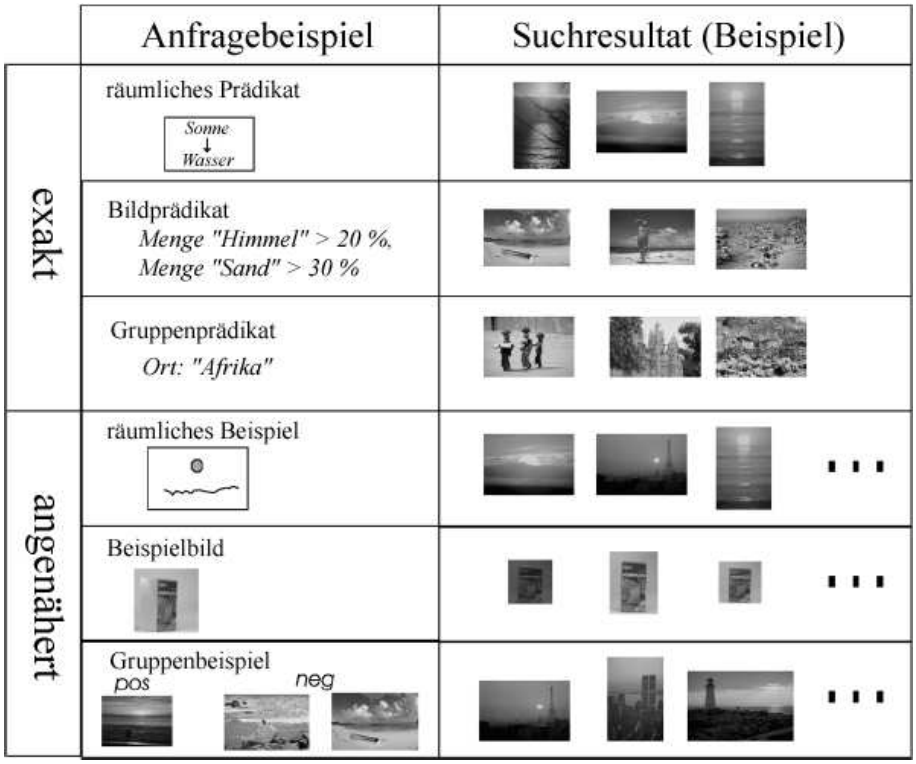

Abbildung 3: Die verschiedenen Arten der Anfrage

7 Systemaspekte

Die interessantesten Anwendungen der inhaltsbasierten Bildsuche sind zumeist gerade diejenigen, die auf einer sehr großen Menge an Bildern operieren, da in diesem Fall auch erst das Erlernen genereller Gesetze aus dem Datensatz Sinn macht. Für solch große Bilddatenbanken kann aber die rein rechnerische Performance eines Bildsuchsystems nicht mehr außer acht gelassen werden, denn eine rein lineare Speicherung der Features würde schnell zu inakzeptabel langen Suchzeiten führen: Wenn man sich beispielsweise eine Datenbank vorstellt, die etwa 100000 verschiedene Bilder enthält, deren Features jeweils in einem 50-dimensionalen Vektor gespeichert sind (und jedes Bild enthält natürlich mehrere Features), bekommt man eine ungefähre Idee davon was es heißen würde, bei jeder Suchanfrage die Features aller Bilder linear zu vergleichen und auf Ähnlichkeit zu untersuchen. Eine weit verbreitete Methode zum schnelleren Auffinden der relevanten Daten ist die Indizierung der Daten, die auf unterschiedliche Arten vorgenommen werden kann. Drei Beispiele, die in [1] näher vorgestellt werden, sind die Indizierung über räumliche Partitionierung, über Datenpartitionierung oder anhand abstandsbasierter Techniken. Alle diese Methoden haben aber das Ziel, die Daten in einer Baumstruktur abzuspeichern und damit möglichst eine Suche in $O(\log N)$ durchzuführen, wobei N die Anzahl der Bilder in der Datenbank ist.

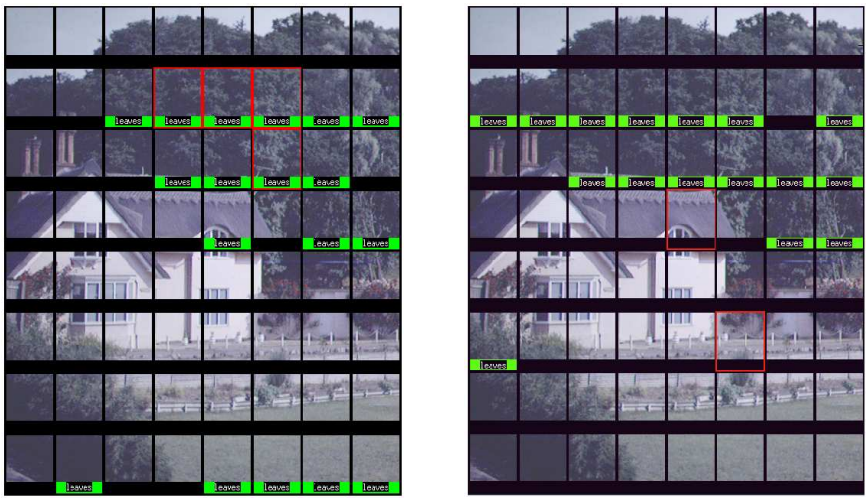

Abbildung 4: Anfrage mit positivem und negativem Feedback

7.1 Bewertung von Systemen

Um eine Aussage über die Nützlichkeit praktischer Anwendungen von inhaltsbasierten Bildsuchsystemen machen zu können ist die Bewertung von Systemen eine schwierige, aber nötige Aufgabe, die oft vernachlässigt wird. Häufig werden beispielsweise Werte für die Effizienz eines Systems angegeben, die sich auf einen ganz bestimmten Datensatz in einem einzigen Szenario beziehen, ohne dabei wirkliche Vergleichsmöglichkeiten zu anderen Systemen zu bieten. Aus dem Gebiet der allgemeinen Informationssuche (*information retrieval*) sind einige Ansätze übernommen worden, wobei besonders die Maße *precision* und *recall* von Bedeutung sind: sei q die Anfrage, die ein Benutzer an das Suchsystem stellt, und seien $A(q)$ die Menge der Bilder, die das System als Antwort liefert und $R(q)$ die Gesamtmenge der Bilder, die der Benutzer als relevant in Bezug auf seine Anfrage einstuft. Dann ist p (*precision*) der Anteil der gefundenen Bilder, der tatsächlich relevant ist, repräsentiert durch

$$p = \frac{|A(q) \cap R(q)|}{|A(q)|}, \tag{1}$$

und r (*recall*) ist der Anteil der gefundenen, tatsächlich relevanten Bilder an der Menge aller relevanten Bilder, repräsentiert durch

$$r = \frac{|A(q) \cap R(q)|}{|R(q)|}. \tag{2}$$

Es hat sich jedoch herausgestellt, dass diese Maße nicht immer ausreichend sind, da die Bildsuche ein weitaus komplexeres System darstellt, deren Ergebnisse meist deutlich differenzierter und weniger eindeutig sind als beim Information Retrieval beispielsweise in Textdokumenten. Der wichtigste Grund dafür ist, dass die Relevanz bestimmter Bilder nicht eindeutig festgelegt ist: je nach Benutzer können unterschiedliche Ansichten darüber existieren, welche Bilder zu

einer bestimmten Anfrage passen und welche nicht. In einigen speziellen Fällen erweisen *precision* und *recall* sich jedoch durchaus als nützlich. Bewertungskriterien für Bildsuchsysteme sind daher nötigerweise meist sehr komplex und haben mit vielen Problemen zu kämpfen, von denen als Beipiel nur der Begriff der Relevanz eines Suchergebnisses genannt sei, da diese ja nur durch die rein subjektive Definition eines Benutzers entsteht.

8 Zusammenfassung

Die inhaltsbasierte Bildsuche ist ein relativ junges Forschungsgebiet, das ziemlich plötzlich erschienen ist und sich noch immer in sehr starker Bewegung befindet. Für beide Phänomene - das plötzliche Erscheinen sowie die anhaltende Dynamik - fallen als Gründe und treibende Kräfte besonders die schnelle und breitgefächerte Entwicklung und Verbreitung von digitalen Sensoren und Aufnahmegeräten, die ständig fallenden Preise für Speichermedien und natürlich das rasante Wachstum des Internets auf. Daraus resultierend ist auch zu erwarten, dass die inhaltsbasierte Bildsuche weiter in die verschiedensten Richtungen wachsen wird, zum Beispiel im Bezug auf neue Zielgruppen, neue Benutzungszwecke, neue Arten der Benutzung, neue Wege der Interaktion, größere Datenmengen und neue Lösungsansätze.

Gerade auch weil das Internet eine so zentrale Rolle in der Entwicklung der inhaltsbasierten Bildsuche gespielt hat und sogar in immer noch wachsendem Maße spielt, hat sich die Interaktion als ein sehr interessantes Gebiet herausgestellt, und sie ist aus diesem Grund auch im Rahmen dieser Ausarbeitung ausführlich behandelt worden. Ferner wird wahrscheinlich ein wichtiges Ziel sein, möglichst gute Lösungen für die Probleme der semantischen und sensorischen Kluft (*semantic / sensory gap*) zu finden, da hier auch nicht auf bereits aus der allgemeinen Informationssuche bekannte Methoden und Mittel zurückgegriffen werden kann, sondern es sich um einzigartige Probleme bei der inhaltsbasierten Bildsuche handelt. Ein Ansatz, der aber bis jetzt erst in sehr einfacher Form genutzt wird, ist das Integrieren verschiedenster anderer Informationsquellen in die Suchanfrage. Beispiele dafür sind der Kontext, den Bildern angehängte Bezeichner *(labels)*, zugehörige Texte (in die Bilder eventuell eingebunden sind) und vieles mehr.

Literatur

[1] A. W. M. Smeulders, M. Worring, A. Gupta und R. Jain, „Content Based Image Retrieval at the End of the Early Years“ , *IEEE Transactions on pattern analysis and machine intelligence,* Vol. 22, No. 12, Dezember 2000

[2] T. Kato, T. Kurita, N. Otsu und K. Hirata, „A Sketch Retrieval Method for Full Color Image Database-Query by Visual Example“ , *In Proceedings. 11th IAPR International Conference on Pattern Recognition,* pp. 530-533, 1992

[3] R. W. Picard and T. P. Minka, „Vision Texture for Annotation“, *M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 302*, Multimedia Systems: Special Issue on Content-based Retrieval