Ein Verfahren zur automatischen Erstellung eines visuellen Wörterbuchs für die Bildsuche

Magdalena Rischka
Institut für Informatik
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Deutschland
rischka@cs.uni-duesseldorf.de

ZUSAMMENFASSUNG

Das Internet bietet eine enorme Anzahl an Bildern. Bildsuchmaschinen stehen vor der Herausforderung Bilder effektiv und effizient zu erschließen. Die klassischen Arten der Bildsuche, die stichwort- und die inhaltsbasierte Bildsuche, haben Nachteile. Ein Retrieval-Modell, welches die Vorteile beider Sucharten integriert und die Nachteile ausschließt, ist die auf einem visuellen Wörterbuch basierende Bildsuche. Ein visuelles Wörterbuch ist dabei eine Menge von Stichwort-zu-visueller-Beschreibung Beziehungen. Wir präsentieren ein Verfahren zur automatischen Erstellung eines visuellen Wörterbuchs aus einer Trainingsmenge von annotierten Bildern. Dabei werden verschiedene Modelle von visuellen Beschreibungen untersucht und anschließend evaluiert. Wir zeigen, dass eine kompakte visuelle Beschreibung existiert, die verglichen mit multiple-Instanzen visuellen Beschreibungen bessere Retrieval-Ergebnisse liefert und gleichzeitig die Anfragezeit drastisch senkt.

Schlüsselwörter

image search, visual dictionary, visual words, visual phrases

1. EINLEITUNG

Das heutige World Wide Web stellt einen großen und ständig wachsenden Datenbestand von Bildern dar und bildet somit eine gute Basis für die Suche nach gewünschten Bildern. Es gibt zwei klassische Arten der Bildsuche: die stichwortbasierte und die inhaltsbasierte Bildsuche. Die stichwortbasierte Bildsuche basiert auf Annotationen und Metadaten der Bilder. Die Anfrageformulierung erfolgt textuell, somit schnell und unkompliziert. Bei der Verarbeitung der Anfrage sucht das System nach Bildern, die, grob gesagt, die Stichwörter aus der Anfrage beinhalten. Einen Nachteil hat diese Suchart jedoch: der Erfolg der Suche hängt von der Qualität der Annotationen und Metadaten der Bilder ab. Je nachdem, ob Bilder manuell vom Benutzer oder automatisch mit Hilfe eines Algorithmus annotiert wurden, weisen diese unterschiedliche Schwächen auf, z.B. die Subjektivität des Beschreibenden, abstrakte Formulierungen oder falsche Stichwortzuordnungen, sowie Unvollständigkeit der Beschreibung. Aufgrund dieses Nachteils versucht man heutzutage, fern von den Annotationen, auf das Bild selbst einzugehen und somit den Inhalt des Bildes zu erschließen. Die inhaltsbasierte Bildsuche basiert demnach auf visuellen Eigenschaften des Bildes, z.B. bzgl. der Farbe, der Textur, Form usw. Eine Anfrage wird mittels einem Beispielbild gestellt, das Retrieval-System sucht dann nach Bildern, die dem Anfragebild ähnlich sind, bezogen auf den, dem System zugrundeliegenden Deskriptor und das Ähnlichkeitsmaß. Der Nachteil dieser Suchart betrifft die Anfrageformulierung mittels dem Anfragebild - ein Anfragebild liegt dem Benutzer in der Regel nicht vor, dieses wird schließlich gesucht. Gewünscht ist daher ein Retrieval-System, welches die Vorteile beider Sucharten integriert, d.h. eine textuelle Anfrageformulierung mit einer inhaltsbasierten Bildsuche kombiniert. Eine Lösung ist das Modell des visuellen Wörterbuchs als eine Menge von Stichwort-zu-visueller-Beschreibung Beziehungen. Bei der Bildsuche auf der Basis des visuellen Wörterbuchs wird nun eine Anfrage textuell gestellt, dann die Stichwörter aus der Anfrage in dem visuellen Wörterbuch nachgeschlagen und deren Übersetzung, d.h. eine visuelle Beschreibung des Stichwortes, für die anschließende inhaltsbasierte Bildsuche verwendet. Die Entwicklung eines Verfahrens zur automatischen Erstellung eines visuellen Wörterbuchs ist Gegenstand dieses Papers. Wir geben zunächst einen Überblick über verwandte Arbeiten, beschreiben dann das entwickelte Verfahren, evaluieren visuelle Beschreibungen und schließen mit einer Schlussfolgerung und einem Ausblick.

2. VERWANDTE ARBEITEN

In der Literatur existieren zwei weitverbreitete Definitionen des Begriffs *visuelles Wörterbuch*. Die erste Definition beschreibt das Konzept der Zuordnungen von Stichwort zu visueller Beschreibung, die zweite betrifft die Quantisierung des Deskriptor-Raums in Partitionen, sogenannte *visuelle Wörter*. Jeder Deskriptor wird dann mit seinem zugehörigen visuellen Wort repräsentiert. Alle Partitionen bilden das visuelle Wörterbuch. Oft werden beide Konzepte kombiniert [1, 4]. [1] verwendet eine gut vorbereitete Trainingsmenge, SCD und HTD (MPEG-7 Standard) Deskriptoren und beschreibt ein Stichwort mit einer konstanten Anzahl von visuellen Wörtern. [4] entwickelt ein visuelles Wörterbuch auf der Grundlage von SIFT-Deskriptoren und daraus abgelei-

23^{rd} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 31.05.2011 - 03.06.2011, Obergurgl, Austria.

teten visuellen Wörtern und stellt jedes Stichwort mit einer Gaußschen Mischverteilung dar. Das Konzept der visuellen Wörter wird mit der Idee der *visuellen Phrase* als ein Paar adjazenter visueller Wörter erweitert. Basierend auf SIFT wird in [6] das Modell der visuellen Phrase untersucht und dabei die Verbesserung des Retrievals nachgewiesen. Wir verwenden den Begriff des visuellen Wörterbuchs um die erste Definition auszudrücken. Falls das Konzept der zweiten Definition und ihre Erweiterung gemeint ist, sprechen wir von visuellen Wörtern und visuellen Phrasen.

3. DAS VERFAHREN ZUR ERSTELLUNG EINES VISUELLEN WÖRTERBUCHS

In diesem Kapitel präsentieren wir das entwickelte Verfahren zur automatischen Erstellung eines visuellen Wörterbuchs aus einer Trainingsmenge von annotierten Bildern. Das Verfahren basiert auf der Idee, die Trainingsbilder einmal bzgl. der Ähnlichkeit ihrer Annotationen und einmal bzgl. ihrer visuellen Ähnlichkeit zu gruppieren, dann die Trainingsbilder, die bzgl. der beiden Aspekte zueinander ähnlich sind, d.h. bzgl. beider Aspekte zusammen gruppiert wurden, aufzusuchen und aus diesen schließlich Korrelationen zwischen Stichwörtern und visuellen Bildmerkmalen abzuleiten.

3.1 Anforderungen an das visuelle Wörterbuch

Das visuelle Wörterbuch kann man sich wie ein herkömmliches Wörterbuch vorstellen, welches aus einer Menge von Einträgen besteht. In dem visuellen Wörterbuch sollen Objekte und visuelle Zusammenhänge, wie z.B. Tiere, Gegenstände, Gebäude, Logos, Symbole, etc. verwaltet werden. Jeder Eintrag ist ein Paar aus einem Stichwort, der das Objekt benennt und einer dazugehörigen visuellen Beschreibung des Objektes. Stichwörter sollen in der Grundform vorliegen - wir sprechen dann von Termen -, und es soll die Polysemie der Terme unterstützt werden. Eine visuelle Beschreibung stellt eine Einheit dar, die für die inhaltsbasierte Bildsuche verwendet wird. Diese soll nur die für dieses Objekt relevanten visuellen Charakteristika erfassen, die allen Perspektiven und Erscheinungsformen des Objektes gemeinsam sind. Zudem soll diese aus Effizienzgründen kompakt, sowie zu der Repräsentation der Bilder kompatibel sein.

3.2 Das konzeptuelle Modell des Verfahrens

Das konzeptuelle Modell des Verfahrens ist in Abbildung 1 dargestellt. Grundlage zum Erlernen des visuellen Wörterbuchs bildet die Trainingsmenge von annotierten Bildern, die beliebig und ohne zusätzliche Vorbearbeitung gewählt werden kann. Ausgehend von dieser werden zunächst einmal zwei Ziele verfolgt: die Gruppierung von ähnlichen Bildern auf der Basis der semantischen Ähnlichkeit ihrer Annotationen und die Gruppierung von ähnlichen Bildern bezüglich ihrer visuellen Ähnlichkeit. Dazu werden die Annotationen sowie die Bilder unabhängig voneinander in eine interne Repräsentation überführt und auf der Basis eines definierten Ähnlichkeitsmaßes gruppiert. Aus den beiden Gruppierungen wird dann das visuelle Wörterbuch erstellt. Dazu wird zunächst einmal das Vokabular für das visuelle Wörterbuch bestimmt. Für jeden Term des Vokabulars werden Trainingsbilder ermittelt, die diesen Term in der Annotation enthalten und bzgl. der Ähnlichkeit von Annotationen und der visuellen Bildmerkmale ähnlich sind. Es findet also ein Abgleich der Gruppierungen statt. Aus den ermittelten Trainingsbildern eines Terms wird schließlich die visuelle Beschreibung des Terms gelernt und zusammen mit dem Term als ein Eintrag in dem visuellen Wörterbuch abgespeichert. Wir erhalten das visuelle Wörterbuch aus Stichwort-zu-visueller-Beschreibung Einträgen.

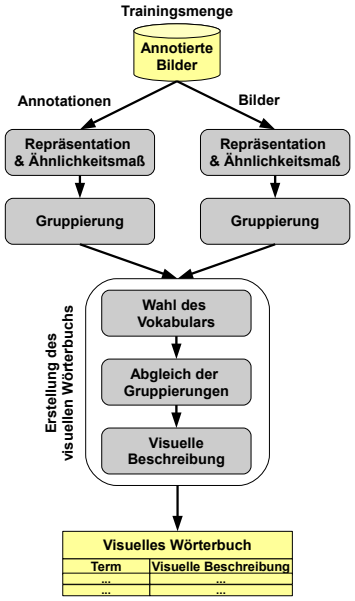

Abbildung 1: Das konzeptuelle Modell des Verfahrens

3.3 Repräsentation und Ähnlichkeitsmaß von/ für Annotationen und Bilder

Für einen semantischen Vergleich müssen Annotationen und Bilder in eine interne Darstellung überführt werden.

Wir bereiten auf und bereinigen zuerst die Annotationen, erstellen dann einen Index mit dem Indexvokabular und leiten daraus für jede Annotation einen Annotationsvektor gemäß der *tf-idf* Gewichtung. Als Ähnlichkeitsmaß wählen wir das Kosinusmaß.

Als Grundlage für die Repräsentation von Bildern wählen wir *Scale Invariant Feature Transform* (SIFT)[2], da es in der Literatur als eins der robustesten Features gilt. Eine auf rohen SIFT-Features basierende Bilddarstellung ist schwer zu handhaben und aus Gründen der Effizienz ungeeignet. Um alle Bilder einheitlich zu repräsentieren wenden wir daher die Technik der visuellen Wörter an. Mit dem Clusteringalgorithmus K-Means basierend auf der Euklidischen Distanz wird der 128-dimensionale Deskriptor-Raum der SIFT-Keypoints in 1000 Partitionen, die visuellen Wörter, zerlegt. Jedem Deskriptor wird gemäß dem Nächsten-Nachbar-Prinzip das entsprechende visuelle Wort zugeordnet. Ein Bild wird schließlich mit einem Histogramm der visuellen Wörter dargestellt, indem das i-te Bin die Vorkommenshäufigkeit des i-ten visuellen Wortes in dem Bild misst. Weiterhin verwenden wir auch das Konzept der visuellen Phrase für die Bilddarstellung. Eine visuelle Phrase vp_{ij} ist ein nichtgeordnetes Paar (Menge) von zwei visuellen Wörtern vw_i, vw_j. Für die Eigenschaft der räumlichen

Nähe übernehmen wir die in [5] definierte Bedingung. In einem Bild liegt eine visuelle Phrase vp_{ij} vor, falls in dem Bild zwei Keypoints kp_a, kp_b existieren und für diese folgendes gilt: das visuelle Wort von kp_a ist vw_i und von kp_b ist vw_j und die Euklidische Distanz *distanz* zwischen den (x, y) Positionen der Keypoints erfüllt die Bedingung:

$$distanz(kp_a, kp_b) < s_a \cdot \lambda \qquad \text{oder} \qquad distanz(kp_a, kp_b) < s_b \cdot \lambda \tag{1}$$

wobei s_a und s_b die Skalierung der Keypoints und λ ein Parameter ist, welcher das Auftreten der visuellen Wörter Paare kontrolliert. Den experimentellen Ergebnissen aus [5] folgend setzen wir $\lambda = 4$. Analog zu visuellen Wörtern erstellen wir auch für visuelle Phrasen ein Histogramm, welches das Vorkommen der 500.500 visuellen Phrasen in einem Bild zählt. In [6] wurde gezeigt, dass Retrieval-Systeme, die auf beiden Bilddarstellungen, der visuellen Wörter und der visuellen Phrasen, basieren, die besten Ergebnisse liefern. Wir folgen dieser Erkenntnis und repräsentieren jedes Trainingsbild mit zwei Histogrammen, der visuellen Wörter und der visuellen Phrasen:

$$b = \left(aHist^{VW}, aHist^{VP}\right) \tag{2}$$

Für die Bestimmung der Ähnlichkeit zweier Bilder verwenden wir ein Ähnlichkeitsmaß, das auf dem *Histogrammschnitt hs* zweier Histogramme basiert:

$$hs(nHist_i^X, nHist_j^X) = \sum_{l=1}^{k} min(nHist_i^X[l], nHist_j^X[l]) \tag{3}$$

wobei $nHist^X$ die normalisierte Version des absoluten Histogramms $aHist^X$ darstellt. Die Ähnlichkeit zweier Bilder b_i und b_j ergibt sich dann mit:

$$\begin{aligned} \ddot{a}hnlichkeit(b_i, b_j) = (1-\alpha) \cdot hs(nHist_i^{VW}, nHist_j^{VW}) \\ + \alpha \cdot hs(nHist_i^{VP}, nHist_j^{VP}) \end{aligned} \tag{4}$$

Für den Wert des Gewichts α orientieren wir uns an dem Paper [6], in welchem der Einfluß unterschiedlicher Gewichtswerte auf die Retrieval-Resultate untersucht wird. Es zeigt sich, dass das Optimum bei dem Wert $\alpha = 0.75$ liegt.

3.4 Gruppierung von Annotationen und von Bildern

Für die Gruppierung der Annotationen wenden wir den in [3] vorgeschlagenen Clusteringalgorithmus *Clustering by Committee* (CBC) an.

Die Gruppierung von ähnlichen Bildern bedeutet, Bilder, die dasselbe Objekt beinhalten, in eine Gruppe zu fassen. Da wir von nicht vorbearbeiteten Trainingsbildern ausgehen, liegen diese Bilder also in der Regel etwas „verschmutzt“ vor, d.h. sie beinhalten neben dem Hauptobjekt ggf. noch andere irrelevante Objekte oder einen Hintergrund. Dadurch kann es leicht zu dem Problem kommen, dass zwei Bilder, die wir intuitiv nicht gruppiert hätten, weil diese unterschiedliche Hauptobjekte haben, trotzdem einen höheren Ähnlichkeitswert haben, als zwei Bilder, die dem menschlichen Empfinden nach ähnlich sind. Bei der Wahl eines Gruppierungsverfahrens müssen wir diese Problematik einbeziehen. Clusteringverfahren, die die Trainingsbilder in Partitionen zerlegen, sind nicht geeignet, es könnte nämlich passieren, dass Bilder aufgrund für uns falsch erscheinenden Gemeinsamkeiten, wie dem Hintergrund, zusammengefasst und dann bzgl. des relevanten Objektes nicht mehr gruppiert werden. Am besten wäre, man hätte visuelle Beschreibungen von den, in der Trainingsmenge enthaltenen Objekten und würde diese als Clusterzentren nehmen, um die Trainingsbilder anhand dieser Clusterzentren überlappend zu gruppieren. Die visuellen Beschreibungen sind aber genau das was wir suchen. Hätten wir solche Beschreibungen, dann wäre die Gruppierung hier überflüssig. Man kann trotzdem versuchen solche visuellen Beschreibungen zu simulieren, indem man das was den Trainingsbildern gemeinsam ist, extrahiert. Sind zwei Bilder bzgl. einem Objekt ähnlich und teilen damit die Charakteristika des Objektes, dann müssen die gemeinsamen Charakteristika auch in der kompakten Bilddarstellung der visuellen Wörter und der visuellen Phrasen verankert sein, nämlich als Durchschnitt der visuellen Wörter und visuellen Phrasen Histogramme der beteiligten Bilder. Der Durchschnitt *ds* zweier Bilder b_i, b_j ist wie folgt definiert:

$$\begin{aligned} ds(b_i, b_j) &= \left(ds^{VW}(b_i, b_j), ds^{VP}(b_i, b_j)\right) \qquad \text{mit} \\ ds^X(b_i, b_j) &= (m(1), ..., m(k)) \qquad \text{und} \\ m(l) &= min(aHist_i^X[l], aHist_j^X[l]) \end{aligned} \tag{5}$$

wobei bei $X = VW$ ist $k = 1000$ und $X = VP$ ist $k = 500.500$. Für die Gruppierung der Trainingsbilder berechnen wir die Durchschnitte der ähnlichsten Bilder, betrachten diese als Pseudo-Objekte und damit als Centroide, und clustern die Trainingsbilder gemäß einem Schwellwert überlappend an diese Durchschnitte. Wir erhalten eine Menge von Gruppen visuell ähnlicher Bilder $\left\{G_l^{visuell} \mid 1 \leq l \leq n\right\}$.

3.5 Wahl des visuellen Wörterbuch Vokabulars

Als nächstes müssen wir klären, welche Stichwörter in das visuelle Wörterbuch aufgenommen werden. Als Stichwörter kommen natürlich nur Terme aus dem Indexvokabular in Frage. Die Übernahme aller Terme als Stichwörter ist jedoch nicht sinnvoll, denn nicht alle Terme bzw. die den Termen zugrundeliegenden Wörter beschreiben Objekte oder beinhalten einen visuellen Aspekt. Wir betrachten daher die Gruppen, die wir durch das Clustering von Annotationen erhalten haben. Wir nehmen an, dass innerhalb einer Annotationsgruppe die Terme, die in den meisten Annotationen vorkommen, etwas mit dem visuellen Inhalt der zugehörigen Bilder zu tun haben müssen. Für jede Annotationsgruppe werden daher diese hochfrequenten Terme bestimmt. Dazu wird zunächst der Term mit der höchsten Annotationshäufigkeit ermittelt und dann noch weitere, deren Annotationshäufigkeit größer ist als 0.7 mal die maximale Häufigkeit. Die Vereinigung der so erhaltenen Terme bildet dann das Vokabular des visuellen Wörterbuchs, also im Grunde die Einträge. Um die Forderung nach der Unterstützung der Polysemie von Termen zu realisieren, werden die betroffenen Terme mehrmals, nur mit unterschiedlichem Kontext, in dem visuellen Wörterbuch aufgeführt. Der jeweilige Kontext eines Terms ergibt sich aus der Gruppe, genauer aus den anderen Termen der Gruppe, zu der der Term gehört. Als Kontext wird der Centroid der Gruppe verwendet. Wir erhalten somit eine Seite des visuellen Wörterbuchs, nämlich eine Menge von Einträgen, die jeweils ein Objekt repräsentieren und aus einem Term und seinem Kontextvektor bestehen.

3.6 Abgleich der Gruppierungen

Für jeden Eintrag des visuellen Wörterbuchs muss nun eine Menge von Trainingsbildern bestimmt werden, aus der die visuelle Beschreibung des Terms gelernt werden soll. Das bedeutet, es müssen die Bilder bestimmt werden, die sowohl bzgl. des Terms als auch visuell bzgl. des beinhaltenden Objekts ähnlich sind. Dazu werden Bilder, die diesen Term in der Annotation enthalten, aus der Annotationsgruppe des Terms genommen und es wird daraus eine Gruppe $G^{eintrag}$ gebildet. Diese Gruppe wird dann mit jeder Gruppe $G_l^{visuell}$ visuell ähnlicher Bilder abgeglichen. Beim Abgleich wird der Mengendurchschnitt jeweils zweier Gruppen gebildet, indem die Bilder übernommen werden, die in der Gruppe $G^{eintrag}$ und in der Gruppe $G_l^{visuell}$ vorkommen. Der resultierende Mengendurchschnitt zweier Gruppen muss mindestens zwei Bilder beinhalten, sonst können keine gemeinsamen Charakteristika gelernt werden. Als Resultat des Abgleichs erhalten wir wiederum, ggf. überlappende, Gruppen von Bildern. Die Bilder innerhalb einer solchen Gruppe sind nun visuell als auch bzgl. des Terms und seinem Kontext ähnlich. Jeder Eintrag des visuellen Wörterbuchs besteht nun aus einem Term, seinem Kontextvektor und der Menge der Bildgruppen aus welcher eine visuelle Beschreibung im nächsten Schritt hergeleitet wird.

3.7 Visuelle Beschreibungen

Als nächstes muss die rechte Seite des visuellen Wörterbuchs, die Seite der visuellen Beschreibungen, bestimmt werden. Wir betrachten einen Eintrag, also einen Term, des visuellen Wörterbuchs und die ihm zugehörige, im letzten Abschnitt bestimmte Menge von Bildgruppen. Es gibt mehrere Möglichkeiten aus der Menge der Bildgruppen eine visuelle Beschreibung abzuleiten. Im Folgenden stellen wir einige Arten von visuellen Beschreibungen in der Reihenfolge der eigenen Entwicklung und Untersuchung vor.

3.7.1 Alle Bilder

Die erste und einfachste Methode eine visuelle Beschreibung anzugeben ist, die Bildgruppen zu vereinigen und die so erhaltene Menge an Trainingsbildern als Repräsentation des Terms zu verwenden. Bei der Bildsuche zu diesem Term finden dann mehrere inhaltsbasierte Bildsuchen statt, indem jedes dieser Trainingsbilder als Anfragebild verwendet wird. Bei dieser multiple-Instanzen visuellen Beschreibung erhalten wir jedoch zunächst für jedes Anfragebild ein Ranking von Bildern als Ergebnis. Es stellt sich also die Frage, wie das Endergebnis aus den Ergebnissen der einzelnen Anfragen berechnet werden soll. Für die Angabe des Endergebnisses werden drei Strategien untersucht.

Bei der ersten Strategie wird das beste Resultat als Endergebnis ausgegeben. Dazu wird die Güte der einzelnen Ergebnisse mittels einem Qualitätsmaß berechnet. Eine solche Berechnung erfordert allerdings zu wissen, welche Bilder des Ergebnisrankings für den Anfrageterm relevant und welche irrelevant sind. Dafür müssten die Bilder in der Bilddatenbank kategorisiert oder mit Termen versehen sein. Von diesem Fall kann man in der Realität jedoch nicht ausgehen. Diese Strategie ist auf einer Bilddatenbank also praktisch nicht anwendbar, lediglich auf einer vorbereiteten Testmenge. Aus Gründen des Performance Vergleichs wird diese trotzdem aufgeführt und untersucht.
(*AlleBilder-BesterScore*)

Jedes Bild aus der Bilddatenbank hat für jedes Anfragebild der visuellen Beschreibung, also in jedem der einzelnen Ergebnisse, eine Rankingposition und einen Ähnlichkeitswert zum Anfragebild. Bei der zweiten Strategie wird für jedes Bild aus der Bilddatenbank der maximale Ähnlichkeitswert aus seinen Ähnlichkeitswerten zu allen Anfragebildern ausgewählt, die Bilder dann entsprechend ihrem maximalen Ähnlichkeitswert sortiert und als Endergebnis ausgegeben (*AlleBilder-MaxÄhnlichkeit*).

Eine dritte Lösung zur Bestimmung des Endergebnisses ist, für jedes Bild aus der Bilddatenbank das arithmetische Mittel ihrer Rankingpositionen aus den einzelnen Ranking-Ergebnissen zu berechnen, dann die Bilder bezüglich diesem arithmetischen Mittel aufsteigend zu sortieren und dieses Ranking als Endergebnis auszugeben.
(*AlleBilder-DurchschnittsRank*)

3.7.2 Durchschnitte

Bei der letzten visuellen Beschreibung werden nicht wirklich Charakteristika des Objektes gelernt, diese stellt also keine visuelle Beschreibung in unserem gewünschten Sinne dar. Wir gehen davon aus, dass die Ähnlichkeit zweier ähnlicher Bilder auf einer gemeinsamen Teilmenge der visuellen Wörter und visuellen Phrasen basiert. Wir extrahieren daher die Gemeinsamkeiten zweier ähnlicher Bilder, indem wir den Durchschnitt ihrer Histogramme gemäß der Formel 5 bilden. Für jede Bildgruppe aus der Menge der Bildgruppen werden paarweise Durchschnitte der Trainingsbilder aus der Bildgruppe berechnet. Die visuelle Beschreibung besteht dann aus allen gebildeten Durchschnitten, d.h. jeder Durchschnitt dient bei der inhaltsbasierten Bildsuche als ein Anfragebild und es finden mehrere Anfragen statt.

Wie bei der ersten visuellen Beschreibung, erhalten wir auch hier eine Menge von einzelnen Ergebnissen und müssen diese zu einem Endergebnis berechnen. Wir wenden dazu die drei beschriebenen Strategien an (*Durchschnitte-BesterScore*, *Durchschnitte-MaxÄhnlichkeit*, *Durchschnitte-DurchschnittsRank*).

3.7.3 Bestes Bild

Die bisher vorgestellten visuellen Beschreibungen sind problematisch: sie bestehen aus mehreren Anfrageinstanzen und weisen daher eine zeitaufwändige Anfrageverarbeitung auf. Eine kompakte Darstellung der visuellen Beschreibung, d.h. eine Darstellung, die aus nur einer Anfrageinstanz besteht, wäre von Vorteil. Eine einfache Lösung wiederum ist, das beste Trainingsbild aus den Trainingsbildern eines Eintrags als visuelle Beschreibung zu wählen. Um das beste Trainingsbild zu bestimmen, vereinigen wir die Bildgruppen und stellen mit jedem Bild aus der Vereinigung eine Anfrage an die ganze Trainingsmenge. Mit einem Qualitätsmaß wird jedes Anfrageergebnis bewertet und das Anfragebild mit der besten Güte, d.h. mit dem höchsten Score des Ergebnisses für die visuelle Beschreibung übernommen (*BestesBild*). Das gewählte Trainingsbild kann jedoch ein lokales Optimum darstellen und in der Suche auf der Bilddatenbank versagen. Weiterhin zeigt sich auch hier das Problem, dass keine Charakteristika von Objekten aus den ähnlichen Trainingsbildern gelernt werden.

3.7.4 Durchschnitte kompakt - Anzahl

Um eine kompakte Darstellung der visuellen Beschreibung zu erhalten, die die gemeinsamen Charakteristika des Objektes ausdrückt, kommen wir auf das Konzept der Durch-

schnitte zurück. Wie in der zweiten visuellen Beschreibung beschrieben, bilden wir zunächst Durchschnitte der paarweisen Trainingsbilder pro jede Bildgruppe. Wir nehmen an, dass visuelle Wörter und visuelle Phrasen, die in den meisten Durchschnitten auftreten, für das Objekt relevanter sind, als die, die seltener vorkommen. Wir erstellen daher eine visuelle Beschreibung aus zwei Histogrammen, der visuellen Wörter und der visuellen Phrasen, und zählen für jedes visuelle Wort und jede visuelle Phrase, in wievielen Durchschnitten es vorkommt. Diese absolute Durchschnitts-Frequenz bildet dann den Wert des jeweiligen visuellen Wortes oder der visuellen Phrase in den Histogrammen (*DurchschnitteKompakt-Anzahl*).

3.7.5 Durchschnitte kompakt - Summe

Um die Wichtigkeit jedes visuellen Wortes und jeder visuellen Phrase innerhalb eines Durchschnitts zu betonen, wird anstatt der Anzahl der Durchschnitte eine Summe der Durchschnitte gebildet. Genaugenommen werden wieder zwei Histogramme der visuellen Wörter und visuellen Phrasen erstellt und jedes Bin des Histogramms ist die Summe der entsprechenden Bins der Histogramme aller Durchschnitte. (*DurchschnitteKompakt-Summe*)

3.7.6 Durchschnitte kompakt - Gewichtete Summe

Der nächsten visuellen Beschreibung liegt die folgende Frage zugrunde: gibt es visuelle Phrasen, die für ein Objekt spezifisch sind, d.h. ist der Anteil der Bilder zu einem Term und einer visuellen Phrase an allen Bildern, die diese visuelle Phrase beinhalten, besonders hoch? Wir berechnen für jede visuelle Phrase vp und dem zugrundeliegenden Term t des Eintrags das Gewicht:

$$g^{VP}(t, vp) := \frac{\#B(t, vp)}{\#B(vp)} \tag{6}$$

mit $B(t, vp)$ stellt die Menge aller Trainingsbilder zu dem Term t, d.h. die Vereinigung der Bilder aus den Bildgruppen zu t, die die visuelle Phrase vp beinhalten, dar. Wir übernehmen die zuvor definierte visuelle Beschreibung *Durchschnitte-Kompakt-Summe* und gewichten den Häufigkeitswert jeder visuellen Phrase vp mit $g^{VP}(t, vp)$. (*DurchschnitteKompakt-GewichteteSumme*)

3.7.7 Durchschnitte kompakt - TFIDF

Für die folgende visuelle Beschreibung übernehmen wir die Idee der tf-idf Gewichtung für Dokumentvektoren. Mit Hilfe der inversen Dokumenthäufigkeit eines Terms, hier inverse Bildhäufigkeit eines visuellen Wortes oder einer visuellen Phrase, wollen wir die Häufigkeiten der visuellen Wörter und visuellen Phrasen, die in sehr vielen Trainingsbildern vorkommen, schwächer, und die die seltener vorkommen, stärker gewichten. Analog zum Text-Retrieval bilden wir also eine Summe aller Durchschnitte, wie in der visuellen Beschreibung *DurchschnitteKompakt-Summe* beschrieben, berechnen dann für jedes visuelle Wort vw und jede visuelle Phrase vp das idf Gewicht:

$$g^{IDF}(vw) := log\frac{\#B}{\#B(vw)} \tag{7}$$

wobei B ist die Menge aller Trainingsbilder und $B(vw)$ die Menge der Trainingsbilder, die das visuelle Wort vw beinhalten. Analog für vp. Die aus der Summe der Durchschnitte entstandenen Histogramme werden dann mit diesen Gewichten multipliziert: jedes Bin zu einem visuellem Wort vw mit $g^{IDF}(vw)$ und jedes Bin zu einer visuelle Phrase vp mit $g^{IDF}(vp)$. (*DurchschnitteKompakt-TFIDF*)

3.7.8 Durchschnitte kompakt - Gewichtetes TFIDF

Die Gewichte aus den beiden letzten visuellen Beschreibungen werden im Folgenden kombiniert. Wir erstellen wieder die Summe aller Durchschnitte und gewichten dann jeden Häufigkeitswert des jeweiligen visuellen Wortes vw mit $g^{IDF}(vw)$, und jeden Häufigkeitswert einer visuellen Phrase vp mit dem kombinierten Gewicht:

$$g^{VP-IDF}(t, vp) := \frac{\#B(t, vp)}{\#B(vp)} \cdot log\frac{\#B}{\#B(vp)} \tag{8}$$

(*DurchschnitteKompakt-GewichtetesTFIDF*)

4. EVALUATION

4.1 Trainings- und Testmenge

Für die Test- und Trainingsmenge werden Bilder und Annotationen zu 50 Objekten aus dem World Wide Web gesammelt. Für jedes der 50 Terme werden jeweils 10 Trainingsbilder und ca. 30 Testbilder heruntergeladen. Als Objekte werden Tiere, Früchte, Gegenstände, Gebäude und Symbole gewählt. Fast alle Bilder liegen in einer Auflösung von ca. 400×400 Pixel vor.

4.2 Testdurchführung

Die vorgestellten visuellen Beschreibungen werden hinsichtlich der Qualität des Retrievals und der Anfrageeffizienz analysiert, um so aus den daraus gewonnenen Ergebnissen und Erkenntnissen die beste für das visuelle Wörterbuch auswählen zu können. Dazu wird für jede visuelle Beschreibung zuerst ein visuelles Wörterbuch aus der Trainingsmenge gelernt und dieses dann in der Anwendung der Bildsuche eingelesen. Für jeden Eintrag des visuellen Wörterbuchs, also jeden Term (im jeweiligen Kontext), wird eine Anfrage auf der Testmenge durchgeführt, dabei die Anfragezeit gemessen und schließlich aus dem erhaltenen Ranking-Ergebnis die Güte des Ergebnisses mit dem Maß *Score*, der im Folgenden erläutert wird, berechnet. Um die visuellen Beschreibungen letztlich miteinander vergleichen zu können, wird für jede visuelle Beschreibung, also jedes Wörterbuch, das arithmetische Mittel der Anfragezeiten und der Scores über allen Einträgen gebildet.

4.3 Bewertungsmaß

In [6] wird für die Evaluation des Retrieval-Systems ein Maß *Score* benutzt. Score bewertet die Top-20 zurückgegebenen Bilder, indem jedes relevante Bild entsprechend des Intervalls, in dem seine Rankingposition liegt, gewichtet wird, die Gewichte aller relevanten Bilder summiert und schließlich auf den Bereich $[0, 1]$ normalisiert werden. Die Autoren des Papers begründen, dass die meisten Benutzer nur die ersten beiden Ergebnisseiten, mit jeweils 10 Bildern pro Seite, betrachten und daher nur die Top-20 der zurückgegebenen Bilder zu einer Anfrage die relevantesten für den Benutzer sind. Wir stimmen mit der Argumentation überein und übernehmen dieses Maß für die Qualitätsbewertung der visuellen Beschreibungen.

4.4 Testergebnisse

Als Testergebnis erhalten wir die zwei Diagramme in Abbildung 2. Das obere Diagramm stellt den durchschnittlichen Score und das untere die durchschnittliche Anfragezeit für jede visuelle Beschreibung dar. Die besten durchschnittlichen Scores erreichen die visuellen Beschreibungen *AlleBilder-BesterScore*, *Durchschnitte-BesterScore*, die aus multiplen Instanzen und der Endergebnis-Strategie *BesterScore* bestehen. Dabei sieht man, dass die auf Durchschnitten basierende visuelle Beschreibung ein besseres Retrieval-Ergebnis liefert, die durchschnittliche Anfragezeit sich gleichzeitig aber verdoppelt. Wie bereits erwähnt ist diese Endergebnis-Strategie nur ein theoretisches Modell. Die zwei praktisch realisierbaren Endergebnis-Strategien verhalten sich je nach visueller Beschreibung unterschiedlich: *MaxÄhnlichkeit* schneidet bei *AlleBilder* besser und bei *Durchschnitte* schlechter ab als *DurchschnittsRank*. Diese multiple-Instanzen visuellen Beschreibungen mit den Strategien *MaxÄhnlichkeit* und *DurchschnittsRank* werden jedoch von den eine-Instanz, auf Durchschnitten basierenden visuellen Beschreibungen bzgl. dem durchschnittlichen Score deutlich übertroffen. Von den besten multiple-Instanzen visuellen Beschreibung *AlleBilder-MaxÄhnlichkeit*, *Durchschnitte-DurchschnittsRank* zu den besten eine-Instanz, *DurchschnitteKompakt-GewichteteSumme* und *DurchschnitteKompakt-TFIDF* haben wir einen Zuwachs des durchschnittlichen Scores von 0.07 und die Anfragezeit sinkt dabei drastisch um das 9- bzw. 17-fache. Die eine-Instanz, auf Durchschnitten basierenden visuellen Beschreibungen weisen einen deutlich besseren, um ca. 0.12 höheren, durchschnittlichen Score gegenüber *BestesBild* auf, sind untereinander mit Unterschieden von bis 0.02 aber relativ ähnlich. Die besten unter ihnen, *DurchschnitteKompakt-GewichteteSumme* und *DurchschnitteKompakt-TFIDF* liefern zudem den besten durchschnittlichen Score unter allen praktisch realisierbaren visuellen Beschreibungen. *DurchschnitteKompakt-TFIDF* maximiert den durchschnittlichen Score und minimiert gleichzeitig die Anfragezeit, ist daher am besten für das visuelle Wörterbuch geeignet.

5. SCHLUSSFOLGERUNG UND AUSBLICK

Von den gestellten Anforderungen an das visuelle Wörterbuch werden die Grundform der Terme mit dem Stemming-Schritt in der Aufbereitungsphase und die Polysemie mit dem CBC Clustering, dem Kontextvektor und den damit verbundenen Mehreinträgen eines Terms, realisiert. Mit der erwähnten besten visuellen Beschreibung ist das Erfassen der Charakteristika des Objektes mit dem Konzept der Durchschnitte, die Kompaktheit und Effizienz mit der eine-Instanz Darstellung und die Kompatibilität zu der Bilddatenbank mit den Histogrammen der visuellen Wörter und Phrasen erfüllt. Zukünftig, um die Anfragezeiten der eine-Instanz visuellen Beschreibungen von ca. 12 Sekunden weiter zu reduzieren, kann man geeignete effiziente Indexstrukturen und Algorithmen für die Bildsuche untersuchen und einsetzen. Um die Qualität des Retrievals weiter zu verbessern, könnte man versuchen auch Farbeigenschaften und ihre Relevanz für Objekte miteinzubeziehen, d.h. diese für die visuelle Beschreibung zu lernen und in der Bildsuche einzusetzen.

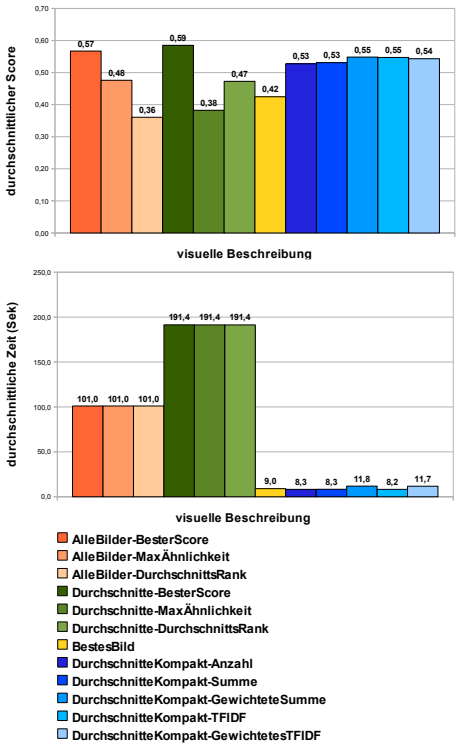

Abbildung 2: Durchschnittlicher Score und durchschnittliche Anfragezeit der visuellen Beschreibungen

6. LITERATUR

[1] C. Hentschel, S. Stober, A. Nürnberger, and M. Detyniecki. Adaptive multimedial retrieval: Retrieval, user, and semantics. chapter Automatic Image Annotation Using a Visual Dictionary Based on Reliable Image Segmentation, pages 45–56. Springer-Verlag, Berlin, Heidelberg, 2008.

[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[3] P. A. Pantel. *Clustering by Committee*. PhD thesis, University of Alberta, 2003.

[4] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Visual tag dictionary: interpreting tags with visual words. In *Proceedings of the 1st workshop on Web-scale multimedia corpus*, WSMC '09, pages 1–8, New York, NY, USA, 2009. ACM.

[5] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, pages 75–84, New York, NY, USA, 2009. ACM.

[6] Q.-F. Zheng and W. Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:7:1–7:19, October 2008.