pixolu - Ein kollaboratives Bildsuchsystem zum Finden visuell und semantisch ähnlicher Bilder

pixolu - A Collaborative Image Search System for Finding Visually and Semantically Similar Images

Kai Uwe Barthel
Hochschule für Technik und Wirtschaft
Wilhelminenhofstraße 75A, 12459 Berlin
Tel.: +49 30 5019-2416, Fax: +49 30 5019-48-2416
E-Mail: barthel@htw-berlin.de, Internet: http://www.f4.htw-berlin.de/~barthel

Zusammenfassung:
In diesem Beitrag wird ein neues Bildsuchsystem vorgestellt, das die Schlagwortsuche, die inhaltsbasierte Bildsuche und kollaborative Filtertechniken vereinigt. Entgegen bisherigen Ansätzen wird nicht versucht, die zu einem Bild zugeordneten Schlagwörter zu optimieren, es wird vielmehr vorgeschlagen, Bildähnlichkeiten über ein Netzwerk von verknüpften Bildern zu modellieren. Im Vergleich zu bestehenden Bildsuchsystemen wird zu Beginn einer Bildsuche eine erheblich größere Menge von Bildern angezeigt. Durch eine visuelle Sortierung lassen sich bis zu 1000 Bilder gleichzeitig erfassen. Bedingt durch diese große Bildmenge kann der Nutzer schnell mehrere Bilder finden, die seiner Sucherwartung entsprechen und diese als Beispielbilder für eine Verbesserung der Suchergebnisse markieren. Durch das Erfassen der Beispielbilder vieler Suchen von unterschiedlichen Nutzern entsteht ein semantisches Netzwerk von Bildbeziehungen. Da das Netzwerk mit jeder Suche aktualisiert wird, werden die Bildsuchergebnisse im Laufe der Nutzung immer besser.

Abstract:
In this paper we propose a new image search system using keyword annotations, low-level visual features, and collaborative filtering techniques. Unlike other approaches our new system does not try to learn the degree of confidence between images and associated keywords. We rather propose to model the degree of image similarities by building a network of linked images.
Compared to normal image search systems we retrieve a much larger set of result images. Using a visually sorted display up to 1000 images can be inspected simultaneously. Due to this large number of images the user can easily find and select several candidate images that are close to his desired search result. By collecting information from many searches of different users we build a semantically network of weighted links of image relationships. By retrieving connected images the search results are improved. As the semantic network gets updated with every search, the search results do get better the more often it is used.

Hintergrund

Die Menge der verfügbaren digitalen Bilder hat in den letzten Jahren explosionsartig zugenommen. Dies führt dazu, dass die Nutzer bei einer Bildsuche nur einen kleinen Bruchteil der von einem Suchsystem gefundenen Bilder betrachten können (bzw. wollen).
Gegenwärtig erfolgt die Bildsuche noch fast ausschließlich anhand von Schlagwörtern. Für professionelle Bilddatenbanken werden Bilder manuell verschlagwortet, was aufwändig und teuer ist. Internet-Bildsuchsysteme bestimmen die Schlagwörter meist automatisch aus dem Kontext der Webseiten, was sehr häufig zu fehlerhaften Schlagwörtern führt. Schlagwörter liegen zunächst immer in einer bestimmten Sprache vor, was bei automatischen Übersetzungen zu Fehlern führen kann. Hinzu kommen Probleme durch Homonyme oder Eigennamen. So ist z. B. bei einer Bildsuche nach „Golf“ nicht klar, ob das Auto, der Sport oder die Küstenlinie gemeint ist.
Systeme, die Bilder visuell analysieren und automatisch Bildbeschreibungen liefern, existieren noch nicht und werden auch mittelfristig nicht verfügbar sein.

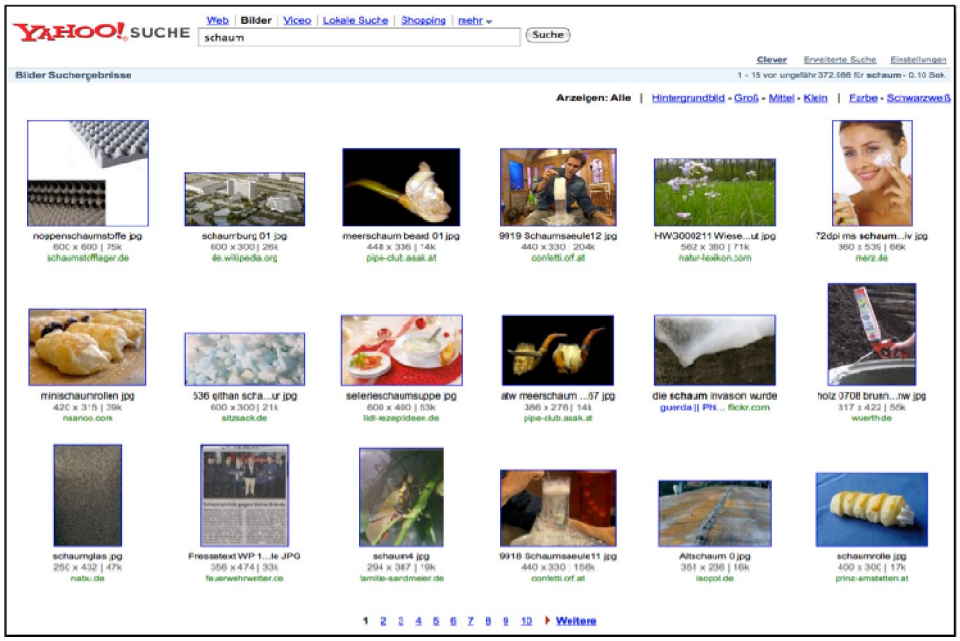

Abbildung 1: Ergebnis der Yahoo-Bildersuche mit dem Suchbegriff „Schaum“

Abbildung 1 zeigt als Beispiel das Ergebnis einer Yahoo-Bildersuche mit dem Suchbegriff „Schaum“. In diesem Fall verfügt das Suchsystem über fast 400.000 Ergebnisbilder, die in Sets von meist 20 Bildern pro Webseite angezeigt werden.

Werden die visuellen Inhalte der Bilder für eine Bildsuche mitverwendet, spricht man von *inhaltsbasierter Bildsuche* oder engl. „*Content Based Image Retrieval*“ (CBIR). Aus den Bilddaten werden elementare statistische Eigenschaften (Featuredaten bzw. Featurevektoren) automatisch extrahiert und anhand dieser werden Bilder mit ähnlichen Eigenschaften gesucht (Abb. 2b). Typischerweise werden mehrere unterschiedliche Featuretypen, die z. B. die Form, die Farbe oder die Textur des Bildes beschreiben, gewichtet kombiniert.
Der Status Quo der Bildsuchverfahren lässt sich folgendermaßen zusammenfassen. Eine gezielte Suche nach spezifischen Bildern ist aktuell weder mit der schlagwortbasierten noch mit der inhaltsbasierten Bildsuche zufriedenstellend möglich.

- Werden Bilder mit Schlagwörtern gesucht, so sind die Ergebnisse oft unbefriedigend, da die Schlagwörter häufig unvollständig, fehlerhaft, mehrdeutig (Abb. 2a) oder schlecht übersetzt sind.

- Inhaltsbasierte Bildsuchsysteme finden vielfach Bilder, die zwar ähnliche Farben und Formen, aber eine gänzlich andere inhaltliche Bedeutung besitzen (Abb. 2c). Bilder, die ähnliche Inhalte darstellen, die aber unterschiedlich aussehen, können nicht gefunden werden (Abb. 2d).

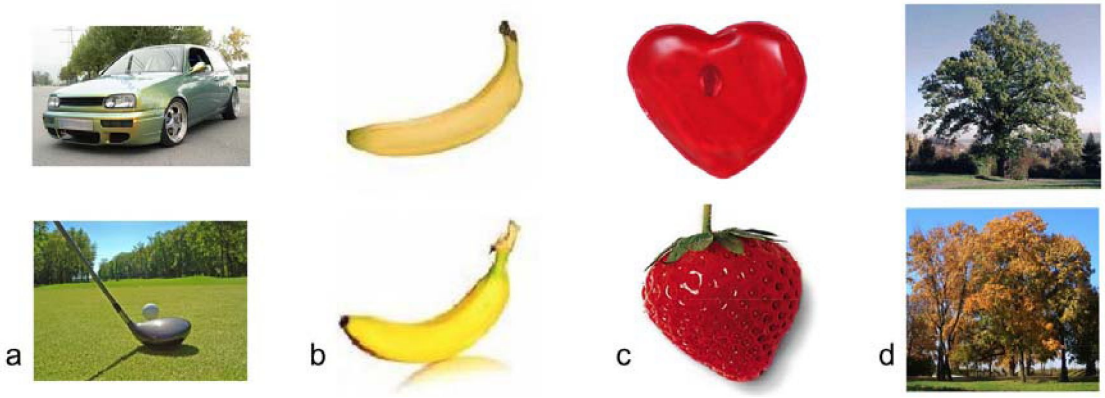

Abbildung 2: Beispiele zum Status Quo der schlagwortbasierten und inhaltsbasierten Bildsuche

Stand der Technik
Aufgrund der genannten Probleme der schlagwort- und inhaltsbasierten Bildsuchverfahren wird kontinuierlich weiter an Verbesserungen geforscht. Zunächst seien andere aktuelle Forschungsansätze vorgestellt. Das Prinzip der inhaltsbasierten Bildsuche besteht unverändert darin, den visuellen Bildinhalt anhand automatisch aus den Bildern extrahierter Low-Level-Featuredaten zu beschreiben. Diese Featuredaten können zwar die visuelle Erscheinung eines Bildes beschreiben, aber keine Aussagen über den semantischen Inhalt eines Bildes machen. Dieses Hauptproblem der inhaltsbasierten Bildsuche wird auch als *semantische Lücke* (engl. *semantic gap*) bezeichnet. Hiermit wird ausgedrückt, dass zwischen der Art, wie Menschen Bilder beschreiben und den Verfahren, mit denen Computern visuellen Bildinhalt repräsentieren, ein sehr großer – momentan noch unüberbrückter – Unterschied besteht.

Erste Forschungsarbeiten zur inhaltsbasierten Bildsuche befassten sich mit der Optimierung der Featureextraktionsverfahren [1]. Ein guter Überblick über den aktuellen Stand der inhaltsbasierten Bildsuche ist in [2] zu finden. Um die Qualität der inhaltsbasierten Bildsuchsysteme zu verbessern, hat sich in der letzten Zeit der Forschungsfokus hin zu Ansätzen orientiert, mit denen versucht wird, die semantische Lücke zu verkleinern [3]. Hierzu gehören die folgenden Kategorien:

- Einbeziehung von ontologischen Ansätzen
- Maschinelle Lernverfahren
- Relevanz-Feedback-Verfahren
- Verknüpfung von Featuredaten und semantischen Informationen

Bei der Einbeziehung von ontologischen Ansätzen wird versucht, formale Relationen zwischen Eigenschaften – in diesem Fall den lokalen Bildeigenschaften – zu modellieren [4]. Dies kann dann bespielweise bedeuten, dass bei Landschaftsaufnahmen vorausgesetzt wird, dass sich ein blauer Himmel oberhalb von einer grünen Landschaft befinden muss. Für bestimmte Bildtypen – wie z. B. medizinische Bilder – kann die Verwendung von Ontologien sinnvoll sein. Für beliebiges Bildmaterial ist es jedoch sehr schwierig, entsprechend allgemeingültige Ontologien zu entwickeln.

Inhaltsbasierte Bildsuche mit maschinellen Lernverfahren wie beispielsweise Support-Vector Maschinen (SVM) [5, 6] sind in idealer Weise geeignet, um die hochdimensionalen Featureräume zu clustern. Gerade bei der Klassifizierung von Bildsets mit einer begrenzten Anzahl von Bildklassen lassen sich sehr gute Ergebnisse erzielen. Der Einsatz von SVMs kann die Ergebnisse von inhaltsbasierten Bildsuchsystemen verbessern, die dargestellten generellen Probleme bestehen jedoch nach wie vor.

Sehr beeindruckende Ergebnisse bei der Suche nach Bildduplikaten können mit sogenannten Keypoint-Extraktionsverfahren erzielt werden [7]. Hierbei wird ein „digitaler Fingerabdruck“ eines Bildes erzeugt, der auch bei sehr starken Bildmanipulationen, wie Beschneidungen, Farb- und Kontraständerungen oder geometrischen Manipulationen erhalten bleibt. Mit diesen Verfahren lassen sich zu einem Anfragebild sehr gut Duplikate bestimmen, allerdings ist es nicht möglich, Bilder zu finden, die zwar sehr ähnlich aussehen, aber nicht den identischen Inhalt darstellen.

Relevanz-Feedback (RF)-Systeme wurden vorgeschlagen, um Bildsuchergebnisse durch die Mithilfe des Nutzers zu verbessern [8-10]. Hierbei bewertet der Nutzer die Suchergebnisse und gibt an, ob diese seinen Erwartungen entsprechen oder nicht. Anhand dieses Feedbacks werden dann entweder die Gewichte der unterschiedlichen Features neu angepasst oder es wird eine neue Suchfrage mit einem veränderten Featurevektor initiiert. Ein Problem des RF besteht darin, dass das Suchsystem den Grund des Feedbacks nicht kennt: Es ist unbekannt, welche Bildeigenschaft (z.B. Farbe oder Form) ausschlaggebend für das positive oder negative Feedback waren. Häufig sind RF-Systeme iterativ gestaltet, d. h. der Nutzer muss mehrfach nacheinander viele Bilder bewerten, um ein besseres Suchergebnis zu erhalten. Das größte Problem der RF-Systeme besteht darin, dass es den Nutzern typischerweise zu mühsam ist, bei vielen Bildern jeweils anzugeben, ob diese dem angestrebten Suchergebnis entsprechen oder nicht.

Eine weitere Gruppe von Verfahren kombiniert die Verwendung der Low-Level-Featuredaten mit den Schlagwörtern. Ein großer Teil dieser Ansätze verfolgt das Ziel einer automatischen Verschlagwortung der Bilder. Wenyin et al. stellten ein solches Verfahren für eine semi-automatische Bildverschlagwortung vor [11]; sie benutzten ein RF-System, das unverschlagworteten Bildern im Fall eines positiven Feedbacks automatisch die Schlagwörter der Suchanfrage zuweist. Pan et al. schlagen ein vollautomatisches Verschlagwortungsverfahren vor, das auf der Ähnlichkeit der Featuredaten basiert [12]. Bilder ohne Schlagworte erben gewissermaßen die semantischen Eigenschaften von visuell ähnlichen Bildern. Dieses Verfahren funktioniert gut bei Bilddatenbanken, die viele ähnliche Bilder enthält, die unter ähnlichen Bedingungen aufgenommen wurden. Stammen die Bilder (wie im Internet) aus unterschiedlichen Quellen, ist dieser Ansatz sehr problematisch, da ähnliche Featuredaten eben nicht automatisch auch eine ähnliche semantische Bedeutung implizieren (Abb. 2c). Wang et al. beschreiben ein System, das es dem Benutzer erlaubt, wahlweise eine Schlagwort- oder eine inhaltsbasierte Suche durchzuführen [13]. Lu et al. [15] und Zhou et al. [14] schlagen vor, Schlagwörter und Featuredaten zu kombinieren. Indem sie RF-Techniken verwenden, erzeugen sie gewichtete Verbindungen zwischen den Bildern und den Schlüsselwörtern. Dieser Ansatz hat jedoch die gleichen Probleme wie die klassischen inhaltsbasierten Bildsuchsysteme: Wenn mehrere Bilder eine starke Verbindung zum gleichen Schlagwort besitzen, bedeutet dies nicht, dass diese Bilder auch semantisch ähnlich sind. Dieser Fall tritt insbesondere bei Homonymen (Worte, die für verschiedene Begriffe stehen) auf. Unser Ansatz vermeidet dieses Problem, indem wir nicht die Bilder mit Schlagwörtern verknüpfen, sondern die Bilder untereinander vernetzen.

Han et al. [16] schlagen ein System vor, bei dem semantische Korrelationen zwischen Bildern anhand Nutzerfeedback bestimmt wird. Hierzu wird das Verhältnis aus der Anzahl, wie oft zwei Bilder bei einer Suchanfrage gleichzeitig positives Feedback erhalten und der Anzahl, wie oft beide Bilder zwar gleichzeitig angezeigt werden, bei denen aber nur ein Bild positives Feedback erhält, bestimmt. Dieses Verfahren hat diverse Probleme, die eine effiziente und gut nutzbare Realisierung erheblich erschweren: Einerseits entsteht ein extrem großer Speicherbedarf, da die semantischen Korrelationen für alle möglichen Bildpaare gespeichert werden müssen. Ein weiteres Problem besteht im RF-Ansatz. Da Nutzer typischerweise nur wenige positive Beispiele markieren, bedeutet dies, dass semantische Korrelationen häufig fehlerhaft bestimmt werden. Da bei RF-Systemen nur wenige Ergebnisbilder angezeigt werden können, um den Nutzer nicht zu viel Bewertungen zuzumuten, bedeutet dies gleichzeitig, dass auch nur wenige semantische Relationen bestimmt werden können.

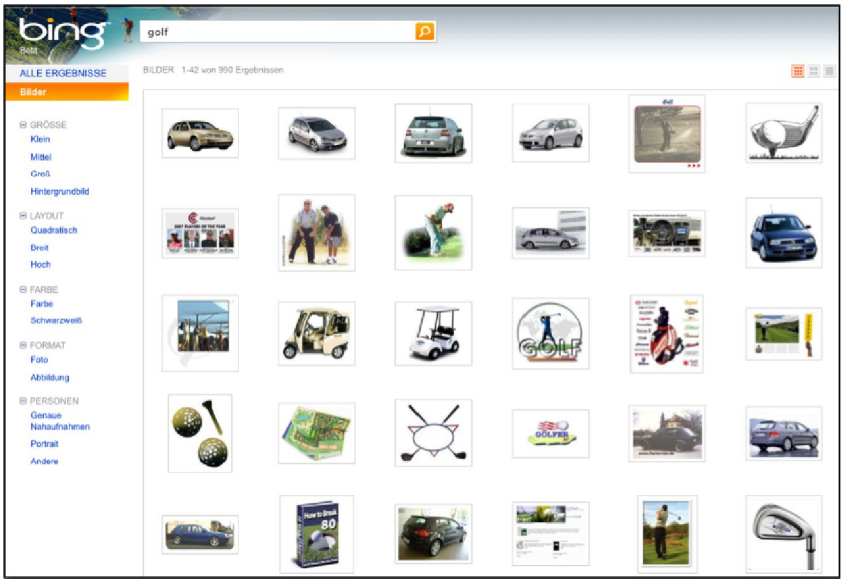

Abbildung 3: Visuelle Ähnlichkeitssuche mit Microsofts „Bing"-Bildersuche, (Suche nach den dem ersten Bild ähnlichen „Golf"-Bildern). Es ist zu erkennen, dass auch semantisch unähnliche Bilder wie z. B. Golfspieler gefunden werden.

Aktuell beginnen einige der Internet-Bildsuchsysteme wie z. B. Google oder Bing von Microsoft (Abb. 3) Bildsuchsysteme einzusetzen, die Schlagwortsuche und inhaltsbasierte Bildsuche kombinieren. Zu einem über eine Schlagwortsuche gefundenen Bild lassen sich andere visuell ähnliche Bilder anzeigen. Die Bilder, die hierbei gefunden werden, sind häufig Duplikate oder visuell sehr ähnliche Bilder, aber eben auch Bilder, die inhaltlich völlig andere Dinge darstellen. Semantisch ähnliche Bilder, die sehr anders aussehen, lassen sich jedoch nicht finden.

Prinzip des neuen Verfahrens

Obwohl die text- und inhaltsbasierte Bildsuche seit Jahren weiterentwickelt wurde, bestehen die dargestellten Probleme nach wie vor. Trotz dieser eigentlich eher negativen Voraussetzungen wurde an der HTW Berlin ein Konzept eines sehr leistungsfähigen Bildsuchsystems entwickelt, das schlagwort- und inhaltsbasierte Bildsuche kombiniert und diese jeweils in ihren gut funktionierenden Aspekten nutzt, ihre spezifischen Probleme jedoch vermeidet.

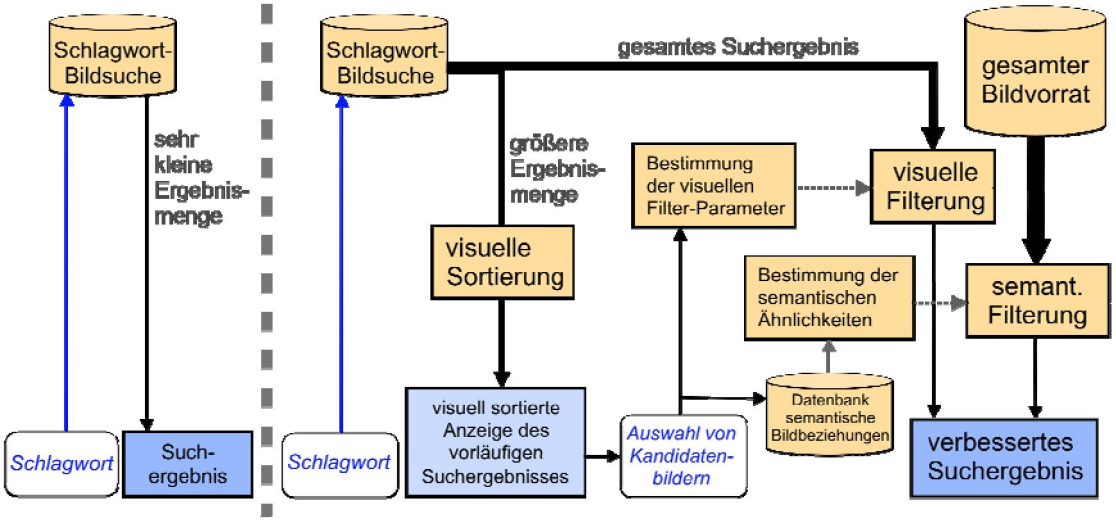

Abbildung 4: Links: ein klassisches schlagwortbasiertes Bildsuchsystem, rechts: das Blockschaltbild des vorgeschlagenen Bildsuchsystems

Abb. 4 (rechts) zeigt das Blockschaltbild des neuen Bildsuchsystems im Vergleich zu einem klassischen schlagwortbasierten Suchsystem (links). Zunächst erfolgt wie bei herkömmlichen Systemen eine Schlagwort-Anfrage, jedoch wird eine erheblich größere Menge von Ergebnisbildern geladen, die visuell sortiert werden, um sie erfassbar zu machen (Abb. 5).

Abbildung 5: 400 Eiffelturm-Ergebnisbilder einer Schlagwortsuche, links: Originalreihenfolge einer Google-Bildsuche, rechts: dieselben Bilder visuell sortiert

Hierdurch kann der Nutzer sehr leicht Bilder identifizieren, die seinem gewünschten Ergebnis nahe kommen und als Kandidatenbilder auswählen, anhand derer dann visuell und/oder semantisch ähnliche Bilder ermittelt werden. Das Besondere des neuen Systems besteht darin, dass es aus dem Nutzerverhalten die semantischen Zusammenhänge zwischen Bildern erlernt und somit bei sukzessiver Verwendung vieler Nutzer immer bessere Suchergebnisse liefert. Zentraler Punkt des neuen Systems ist die Erfassung der Nutzerinteraktion, die Modellierung eines semantischen Netzwerks und die Bestimmung der semantischen Bildähnlichkeiten. Für die Bildersuche war dieses Prinzip bisher nicht anwendbar, da es keine Möglichkeit gab, die extrem vielen Ergebnisbilder sinnvoll in Beziehung miteinander zu setzen. Bei dem vorgeschlagenen System wird dies möglich, da dem Nutzer eine sehr große Zahl von Ergebnisbildern gleichzeitig präsentiert werden kann. Typischerweise ist eine Menge von ca. 200 – 400 Bildern groß genug, um einerseits eine gute Übersicht über die möglichen Typen von Ergebnisbildern zu bieten, gleichzeitig besteht auch eine große Wahrscheinlichkeit, dass mehrere Bilder, die dem gewünschten Ergebnis nahekommen, angezeigt werden. Durch die visuell sortierte Darstellung lassen sich diese „Kandidatenbilder“ leicht identifizieren.

Abbildung 6: Links: 300 erste Ergebnisbilder einer Suche nach dem Schlagwort „Tomato“.
Rechts: Suchergebnis nach Verfeinerung durch Vorgabe der zwei ausgewählten Beispielbilder (oben)

Indem der Nutzer zwei oder mehrere Kandidatenbilder auswählt (Abb. 6), gibt er dem Suchsystem unbewusst Feedback und drückt aus, dass diese Bilder – entsprechend seiner Suchintention – eine semantische Beziehung haben. Werden diese Beziehungen über viele Bildsuchen vieler Nutzer erfasst, so entsteht ein gewichteter Graph, anhand dessen die semantischen Beziehungen der Bilder modelliert werden können. Abb. 7 zeigt die Visualisierung eines Ausschnitts eines solchen Graphen. Anhand der ausgewählten Kandidatenbilder können dann semantische ähnliche Bilder bestimmt werden, indem diejenigen Bilder geliefert werden, die im semantischen Graphen die stärksten Verbindungen haben.

Eine Suche mit diesem neuen Bildsuchsystem führt dazu, dass sowohl visuell als auch semantisch ähnliche Bilder gefunden werden können. Gleichzeitig wird das entstehende semantische Netzwerk mit jedem Suchvorgang weiter angelernt, so dass die Suchergebnisse im Lauf der Zeit immer besser werden. Das Besondere des vorgeschlagenen Systems besteht darin, dass bei kontinuierlicher Nutzung eine automatische Vernetzung des gesamten Bildervorrats entsteht. Diese Vernetzung ist dabei völlig sprach- und schlagwortunabhängig, da die Beziehungen der Bilder ausschließlich über den Grad ihrer semantischen Verknüpfungen existieren. Dies bietet völlig neue Interaktionsformen mit den Bildern, da es hierdurch möglich wird, entlang der semantischen Beziehungen durch den Bildervorrat zu navigieren. Hierfür sind geeignete Navigationskonzepte zu entwickeln, wozu eine gleichermaßen übersichtliche als auch schnelle Darstellung der vielen Bilder notwendig ist.

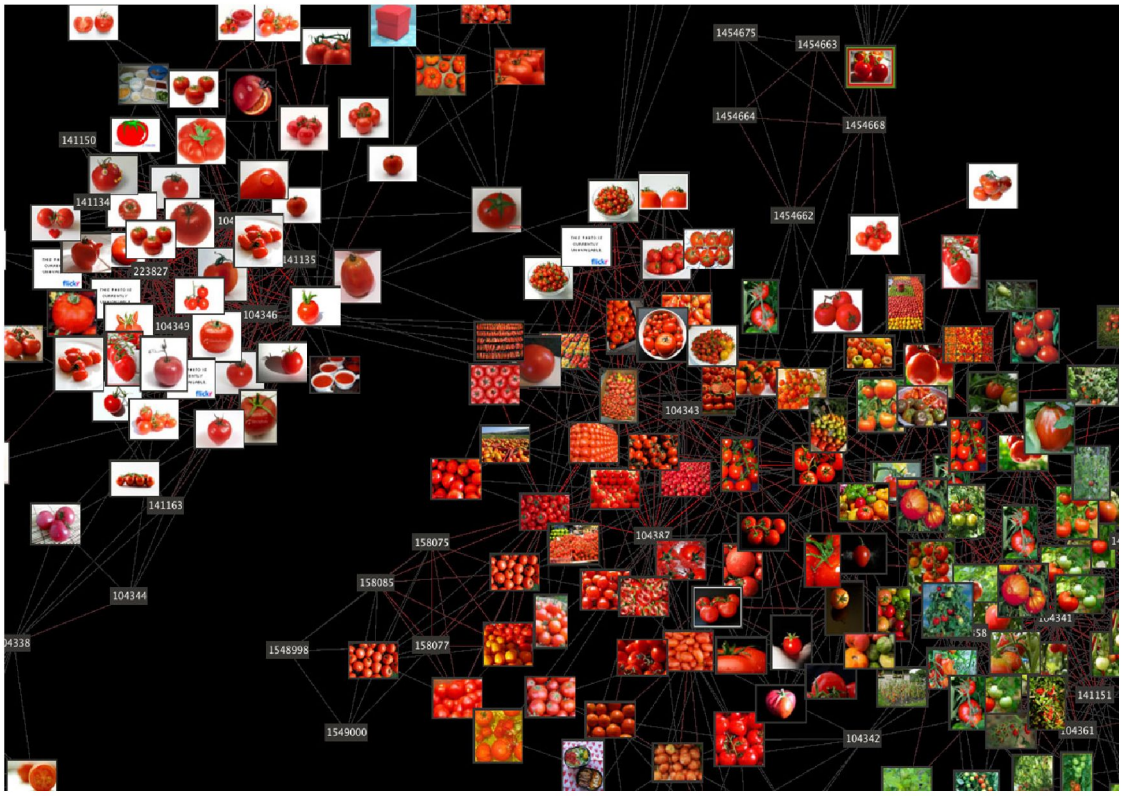

Abbildung 7: Ausschnitt eines semantischen Graphen (zweidimensional dargestellt)

[1] A.W.M. Smeulders et al., “Content-Based Image Retrieval at the End of the Early Years,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, 2000, pp. 1349- 1380.

[2] Ritendra Datta, Jia Li, James Ze Wang: Content-based image retrieval: approaches and trends of the new age. Multimedia Information Retrieval 2005: 253-262

[3] Liu Ying, Zhang Deng-Sheng, Lu Guojun, Ma Wei-Ying, A survey of content-based image retrieval with high-level semantics, PR(40), No. 1, 2007, pp. 262-282.

[4] V. Mezaris, I. Kompatsiaris, M.G. Strintzis, An ontology approach to object-based image retrieval, Proceedings of the ICIP, vol. II, 2003, pp. 511–514.

[5] R. Shi, H. Feng, T.-S. Chua, C.-H. Lee, An adaptive image content representation and segmentation approach to automatic image annotation, International Conference on Image and Video Retrieval (CIVR), 2004, pp. 545–554.

[6] E. Chang, S. Tong, SVMactive-support vector machine active learning for image retrieval, Proceedings of the ACM International Multimedia Conference, October 2001, pp. 107–118.

[7] David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[8| Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, Relevance feedback: a power tool for interactive content-based image retrieval, IEEE Trans. Circuits Video Technol. 8 (5) (1998) 644–655

[9] Y. Rui, T.S. Huang, Optimizing learning in image retrieval, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 2000, pp. 1236–1243.

[10] X.S. Zhu, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia System 8 (6) (2003) 536–544.

[11] Liu Wenyin, Susan Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, Brent Field: Semi-Automatic Image Annotation (2001)

[12] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, Pinar Duygulu: GCap: Graph-based Automatic Image Captioning (2004)

[13] Wei Wang, Yimin Wu, Aidong Zhang: SemView: A Semantic-sensitive Distributed Image Retrieval System, National Conference on Digital Government Research (2003)

[14] X. S. Zhou, T. S. Huang, “Unifying Keywords and Visual Contents in Image Retrieval”, IEEE Multimedia, April-June Issue, 2002.

[15] Y. L. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, “A unified framework for semantics and feature based relevance feedback in image retrieval systems,” in Proceedings of 8th ACM International Conference on Multimedia (MM '00), pp. 31–37

[16] J. Han, K.N. Ngan, M. Li, H.J. Zhang, “A Memory Learning Framework for Effective Image Retrieval”, IEEE Transactions on Image Processing, Vol. 14, No. 4, April 2005