Nutzung von Metainformation für die Navigation in (verteilten) Datenbeständen

Pascal Pein

1. Juni 2006

Motivation

„Content that cannot be easily found is like content that does not exist, [...]. The easier it becomes to produce content, the faster the amount of content grows and the more complex the problem of managing content gets.“

Fernando Pereira, Rob Koenen

„[..] users of information systems do not need more information, but information which better corresponds to their interests and needs, i.e. information of higher quality.“

Hartmut Wittig

Einordnung

Verteilte Systeme enthalten...

- viele
- unstrukturierte
- flüchtige
- gemeinsam nutzbare

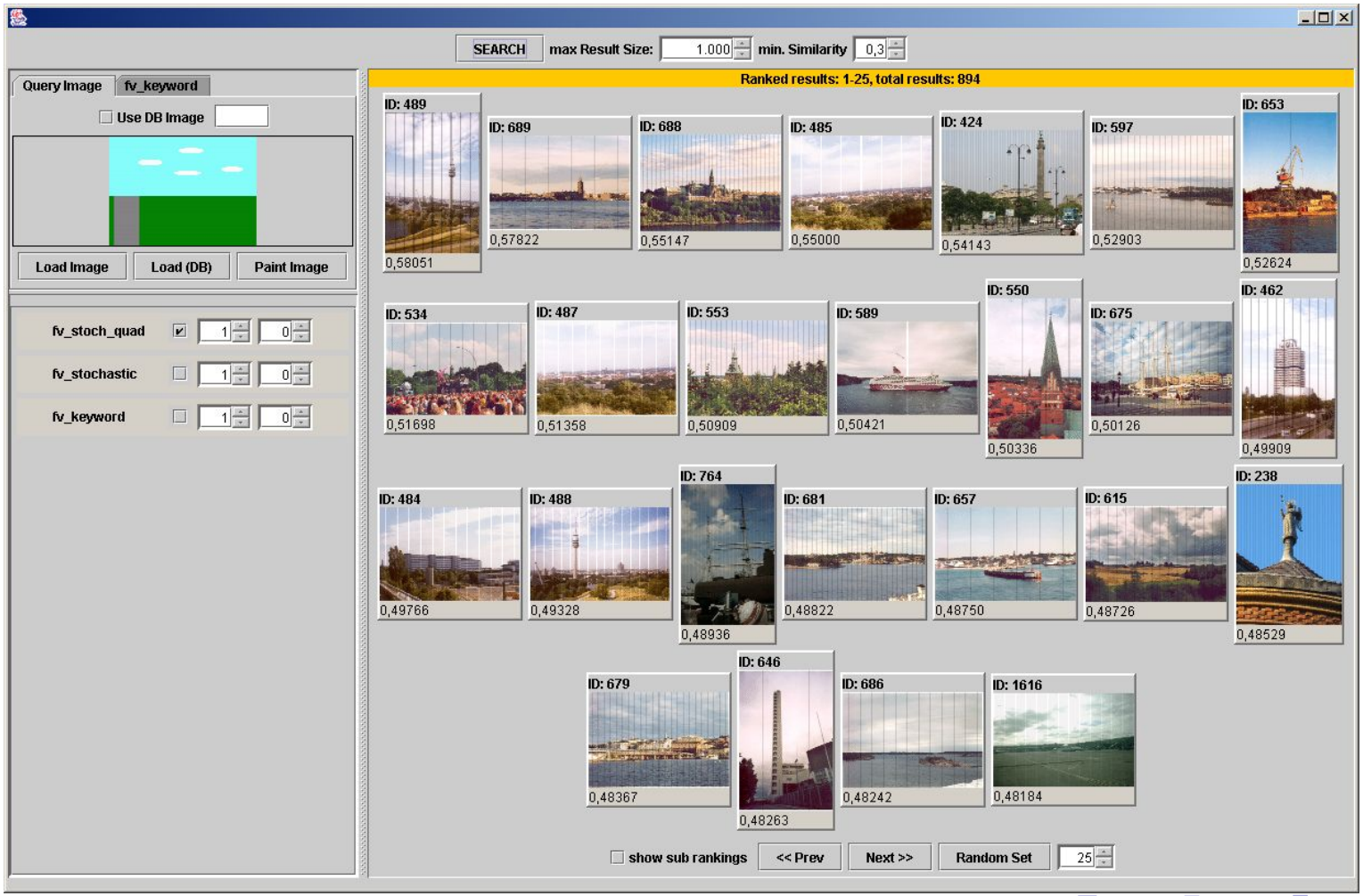
... Informationen

Wie kann man diese Informationen finden?

Feste Referenzen

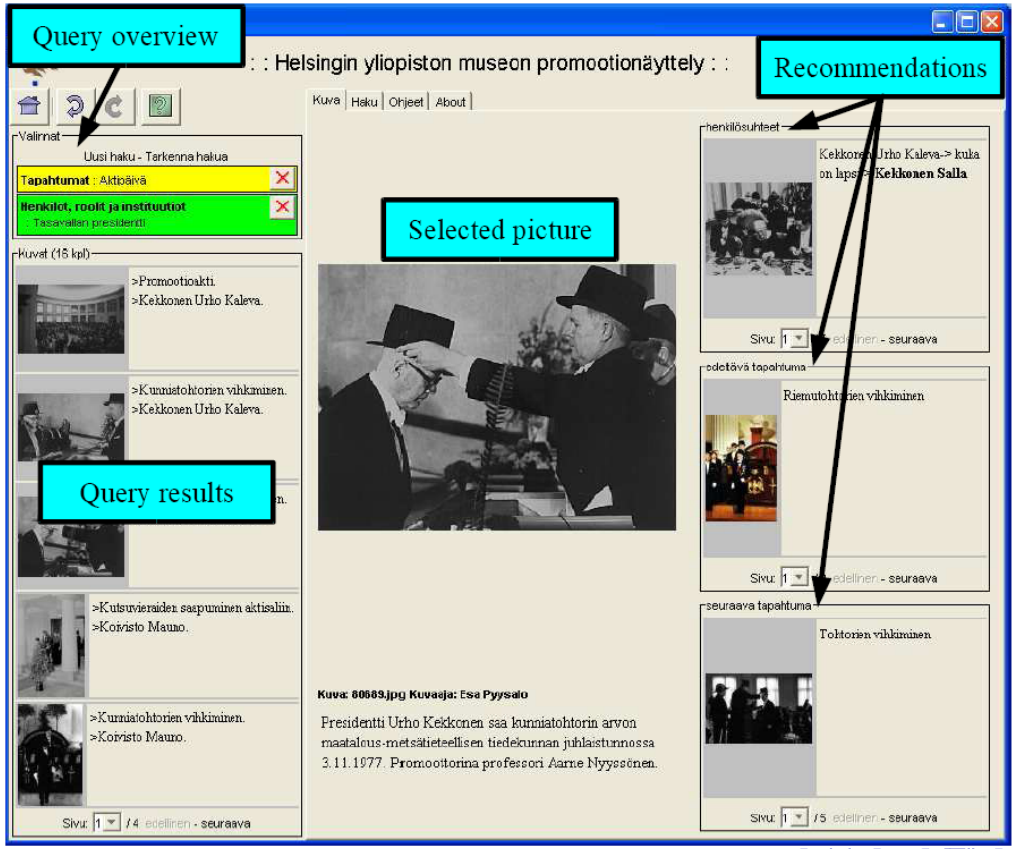
- Hypertext
  - Texte mit Hyperlinks (HTML) stellen Beziehungen untereinander her.
  - Navigation zwischen diesen Seiten ist möglich
- Hypermedia (im engeren Sinne)
  - Hypertextvariante mit Schwerpunkt auf multimediale Daten
  - Sprünge an beliebige Stellen innerhalb einer Audiodatei, eines Videos, ...
  - Synchronisation verschiedener Multimediadokumente

Inhaltsbasierte Bildersuche



Semantische Suche

Ontogator (Hyvönen) - Ontologien
TopicSEEK - TopicMaps



Kombinierte Suche und Navigation in verteilten Systemen

- Nutzung verschiedener Aspekte/Blickwinkel
  - Einschränkung des Suchraums
  - Filterung irrelevanter Daten
  - eigentlicher Typ der gesuchten Daten (Text, Bild, ...) zweitrangig
- Verteilte Suche über mehrere Systeme
  - Hyperlinks vergleichbar mit HTML
  - P2P Suchanfragen an Nachbarknoten (z.B. YaCy)
  - Verschmelzung der einzelnen Ergebnismengen

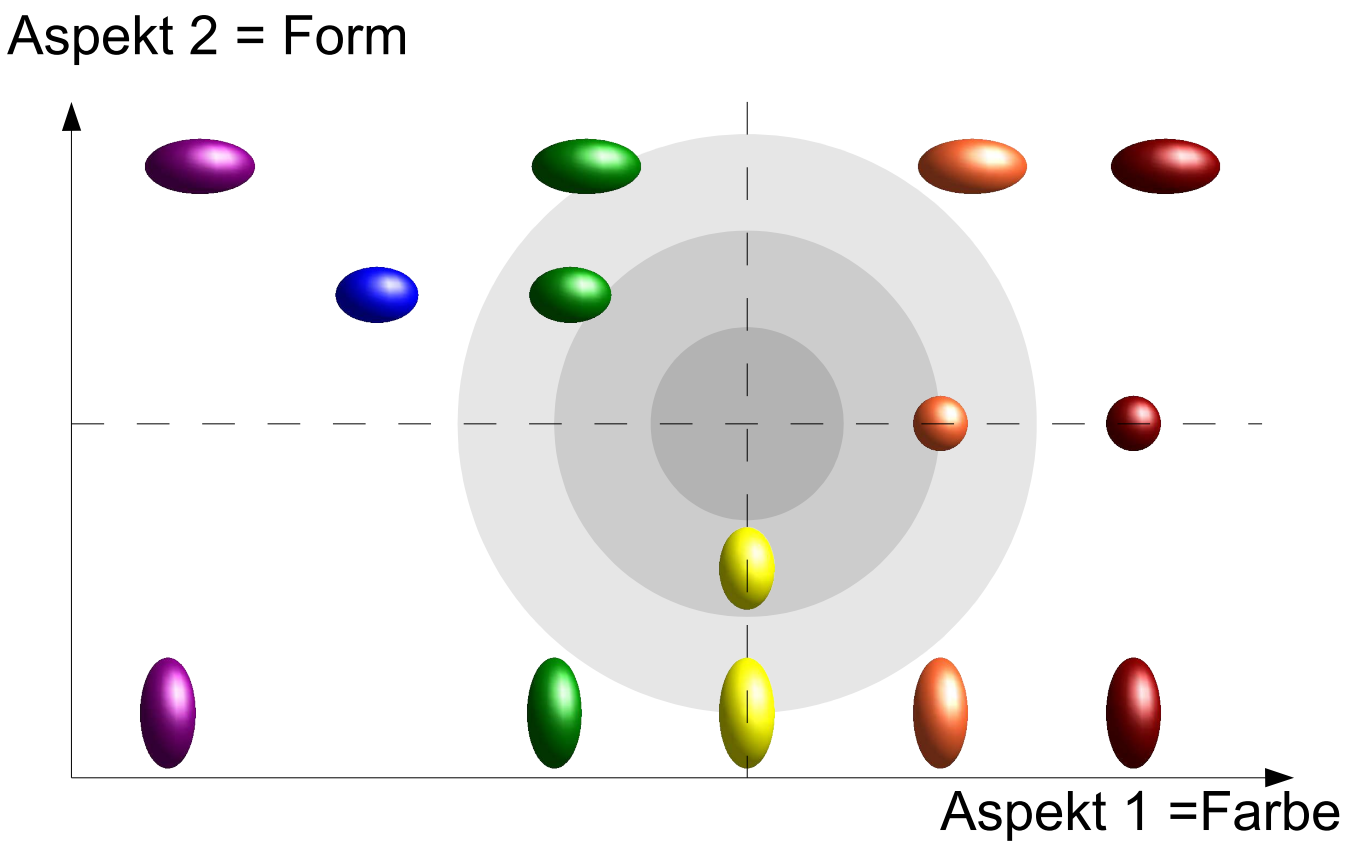
Was sind ähnliche Informationen?

Ähnlichkeit kann auf verschiedene Weise definiert werden:

- Menschliche Wahrnehmung (subjektiv)
- Mathematische Distanzen (objektiv)
- Zugehörigkeit zu einer bestimmten Kategorie/Menge
- ...

Mehrdimensionale Suchräume

Beispiel: Suche nach einer gelben Kugel



Beispiele

- Schlüsselwörter
- Semantik
- Kategorien
- Erstellungsdatum/Zeitpunkt
- Histogramme (Bilder)
- Formen (Bilder)

3-Schichten-Modell für Bilder (J.P. Eakins, M.E. Graham):

Layer 1	**primitives:** **color, texture, geometric shapes and spatial distribution**
Layer 2	**entities:** **object instances or individuals**
Layer 3	**abstract entities:** **activities, states, ...**

Erfassung

- automatisch
  - \+ Für große Datemengen geeignet
  - \+ Objektiv
  - – Nur für wenige Aspekte („Primitive“) nutzbar
- manuell
  - \+ Flexibel
  - \+ Hohe Qualität möglich (Spezialisten)
  - \+ Erfassung abstrakter Merkmale
  - – Hoher Arbeitsaufwand
  - – Subjektiv

Beschreibungsstandards

- textuell
  - Standard Generalized Markup Language (SGML)
  - Extensible Markup Language (XML)
  - Resource Description Framework (RDF)
- audiovisuell
  - Moving Picture Expert Group / Multimedia Content Description Interface (MPEG-7)

Vorgehensweisen

- Navigation (gezielt)
  - „Entlanghangeln“ an vorhandenen Strukturen, um an die gewünschte Information zu kommen
  - Voraussetzung ist eine stimmige Aufbereitung/Katalogisierung der Daten
  - Benutzer kann vorher grob abschätzen, welche Ergebnisse er erhalten wird
- Suche (unscharf)
  - Herausfiltern weniger bestimmter Elemente aus einer schwer überschaubaren Menge an Daten
  - Gerade bei mangelhaft aufbereiteten Daten wichtig
  - Ergebnisqualität nur schwer vorherbestimmbar

Mögliche Anfragen an ein System

1. Filter (hart)
   - Kategorie
   - Zeitraum
2. Beispiele als Anfrage (unscharf)
   - Suchbild
   - Textdatei
3. Metainformationen (gemischt)
   - Schlüsselwörter
   - Suchstring

Relevance Feedback

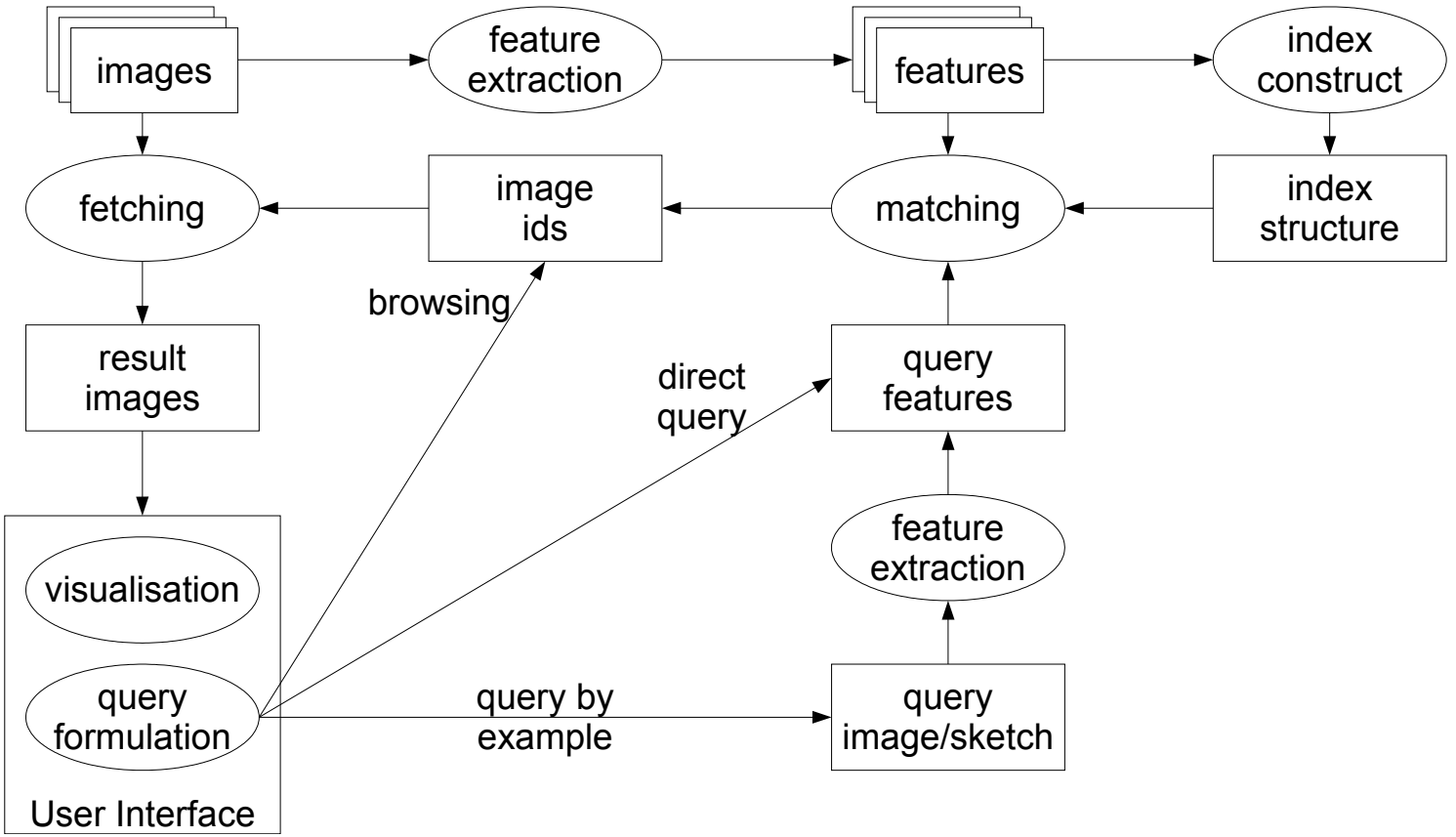
Wie lernt die Suchengine, welche Ergebnisse eine gute Qualität haben, mit der der Nutzer zufrieden ist?

- Relevance Feedback: direkte Rückmeldung während einer Suche
- Wikiprinzip: Nutzer können Annotation jederzeit frei editieren/erweitern

Dies gilt insbesondere für Aspekte, die nicht maschinell extrahiert werden können.

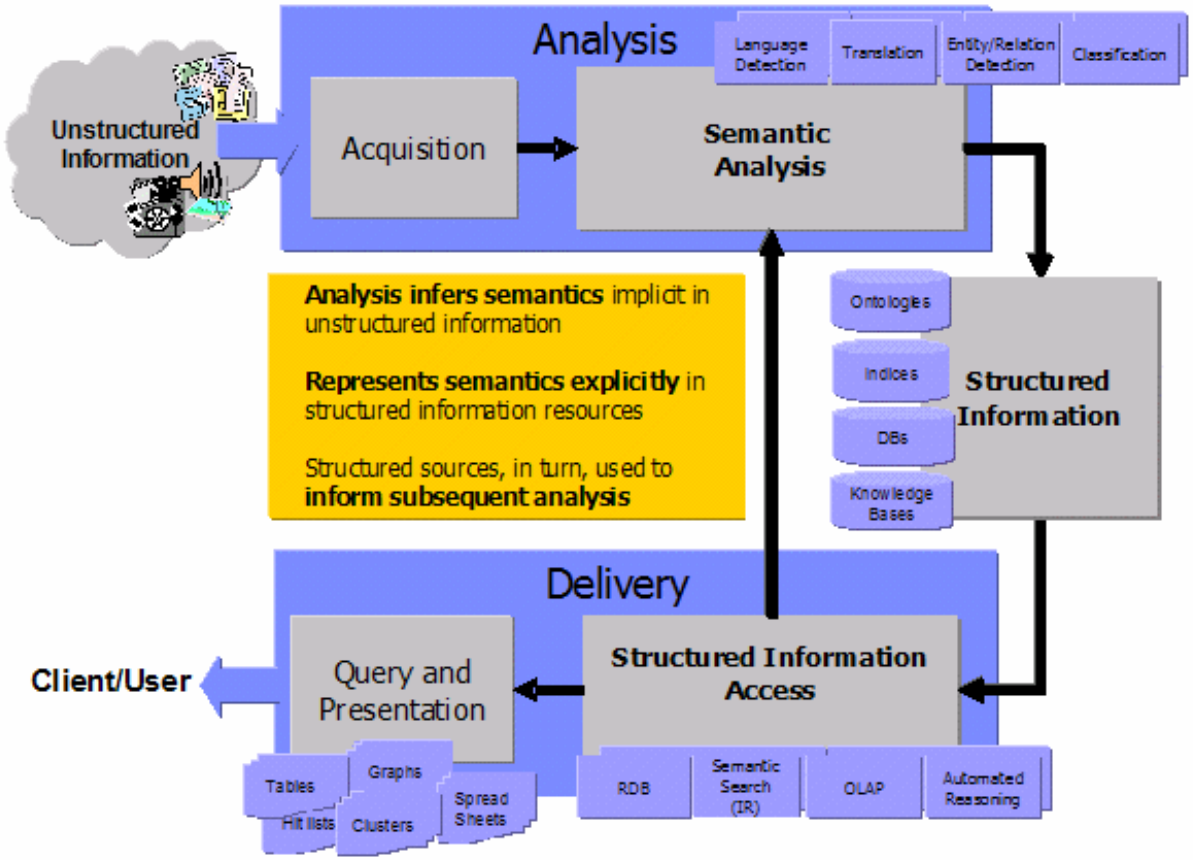
Standard CBIR-Architektur (Veltkamp, Tanase)

„Content Based Image Retrieval“



UIMA-Architektur (IBM Research)

„Unstructured Information Management Architecture“



Vision

- Flexible Navigation im verteilten Datenbestand
- Anschauliche Präsentation der Suchergebnisse
- Unabhängigkeit von eigentlichen Datentypen
- Verteilte, iterative, spontane Annotation (Wiki)
- Relevance Feedback
- Einbindung von Multimediadaten, wie z.B. Videos mit Einsprungstellen

Masterprojekt

Konzepte/Machbarkeitsstudie:

- Ähnlichkeitssuche über extrahierte Aspekte
- Semantische Beziehungen über Topicmaps
  - -> Kombination dieser Anfragen
- Erfassung von Meta/Indexdaten
- Manuelle Erweiterung/Verfeinerung der Daten
- Feste Verknüfungen zwischen Dokumenten
- Nutzung einer standardisierten Beschreibungssprache (z.B. MPEG-7)

Zusammenfassung

„The same digital technology that lowers the thresholds for producing and publishing content can also help in analyzing and classifying it, in extracting and manipulating features for specific applications and in searching and discovering content.“

Fernando Pereira, Rob Koenen

Weiterführende Literatur I


B. S. Manjunath, Philippe Salembier, Thomas Sikora
Introduction to MPEG-7
John Wiley & Sons Ltd., West Sussex, England, 2002,
ISBN 0-471-48678-7


Hartmut Wittig
Intelligent Media Agents
Vieweg & Sohn Verlagsgesellschaft mbH,
Braunschweig/Wiesbaden, ISBN 3-528-05706-8


J.P. Eakins, M.E. Graham
Content-based Image Retrieval. A Report to the JISC Technology Applications Programme
University of Northumbria at Newcastle, 1999

Weiterführende Literatur II


Remco C. Veltkamp, Mirela Tanase
Content-Based Image Retrieval Systems: A Survey
Department of Computing Science, Utrecht University, 2002


Eero Hyvönen, Samppa Saarela, Kim Viljanen
Intelligent Image Retrieval and Browsing Using Semantic Web Techniques - A Case Study
Helsinki Institute for Information Technology (HIIT) / University of Helsinki, 2003

Die letzte Seite

Vielen Dank für die Aufmerksamkeit