
The Reality of the Semantic Gap in Image Retrieval

Tutorial held in conjunction with the

1st International Conference on Semantic and Digital Media Technologies

Athens, Greece

Wednesday 6th December 2006



**Presenters: Peter G.B. Enser, Christine J. Sandom,
Paul H. Lewis & Jonathon S. Hare**



The Reality of the Semantic Gap in Image Retrieval

The semantic gap is referred to frequently in papers on image retrieval or multimedia information handling. However, whilst many authors have been happy to make reference to it, few have attempted to characterize the gap in any detail. This tutorial will attempt to rectify this situation by characterizing the semantic gap in image retrieval rather more specifically than hitherto. It will summarise current attempts to begin to bridge the gap both through developments in content-based techniques, the application of semantic web and knowledge technologies and recent progress in auto image annotation. The tutorial will consist of presentations/demonstrations partly based on research in recent European and UK projects, and particularly on a project to investigate the semantic gap funded by the Arts and Humanities Research Council in the UK involving the four presenters.

This tutorial aims to provide valuable insights for those involved in research and development on image or multimedia retrieval and who wish to understand and address the concerns of real end-users and exploit recent research results in the field. In particular, the tutorial will provide practical insights into the problems associated with bridging the communication gap between the computer science/vision research community and the image management/practitioner community.

The first presentation will summarise research into the way picture searchers articulate real queries, how they are typically resolved through a combination of traditional metadata and the knowledge of the searcher. This section will include our own investigations into query categorization and image categorization and the identification of recurring semantic issues in image search such as significance of events, abstract and emotive concepts and unwanted features.

The second presentation, will explore the ways in which textual description of images is prescribed in theory and applied in practice within image collections, through the use of cataloguing schemas, metadata schemas, controlled vocabularies thesauri, etc. These approaches can be regarded as addressing issues at the semantic end of the semantic gap.

The third presentation will review progress in content-based image retrieval, automatic annotation and extraction of semantics in recent years and explore the types of query that they are able to address. The semantic gap between features that can be extracted directly from images and the semantics that the human searcher attaches to the visual information will be revisited and various staging posts across the gap will be identified such as raw data to features, features to objects, objects to labels and labels to semantics. The ways in which these sub gaps can be bridged in some instances will be discussed together with the substantial current interest in machine learning and auto annotation.

The final presentation will show how the application of ideas from cross language latent semantic indexing can be extended to build multimedia semantic spaces in which visual and conceptual “terms” are mapped to similar locations in the space. This means that images can be retrieved using either textual or visual descriptors whether or not they have collateral textual annotations.

Presenter Biographies:

Peter Enser, Computing, Mathematical and Information Sciences, University of Brighton, UK.

p.g.b.enser@bton.ac.uk

<http://www.cmis.brighton.ac.uk/Research/vir/VIR2.HTM>

Peter Enser is Professor of Information Science and Head of the Computing, Mathematical & Information Sciences (CMIS) Research Centre at the University of Brighton, U.K. He has had many years of engagement with research in the field of visual image retrieval, with a particular focus on user studies. He has directed a number of funded projects in this field, and his many publications and presentations on the topic have addressed international audiences within the library and archive management, cultural heritage and computer science communities.

Criss Sandom, Computing, Mathematical and Information Sciences, University of Brighton, UK.

c.sandom@bton.ac.uk

<http://www.cmis.brighton.ac.uk/Research/vir/VIR2.HTM>

Criss Sandom is a Research Officer in the School of Computing, Mathematical & Information Sciences at the University of Brighton, U.K. She has worked since 1999 on the VIRAMI and Semantic Gap research projects in image retrieval, under the direction of Peter Enser.

Paul Lewis, Electronics and Computer Science, University of Southampton, UK.

phl@ecs.soton.ac.uk

<http://www.ecs.soton.ac.uk/~phl>

Paul Lewis is a professor in the Intelligence, Agents, Multimedia Group within the School of Electronics and Computer Science at the University of Southampton. His main research interests are currently centred on the broad area of multimedia knowledge management. In particular he is addressing problems in image and video processing and analysis, multimedia annotation and semantic description of media. He is particularly involved with designing and developing novel facilities for multimedia information retrieval, navigation and browsing with applications in both the medical and the cultural heritage domains. His research is building on ideas from low-level media processing and knowledge management and emerging semantic web technologies.

Jonathon Hare, Electronics and Computer Science, University of Southampton, UK.

jsh2@ecs.soton.ac.uk

<http://www.ecs.soton.ac.uk/~jsh2>

Jonathon Hare is a research assistant in the Intelligence, Agents, Multimedia Group within the School of Electronics and Computer Science at the University of Southampton. His research interests lie in the area of multimedia information retrieval. In particular he is interested in investigating how content-based image retrieval and auto-annotation techniques can be integrated with semantic web technologies.

THE REALITY OF THE SEMANTIC GAP IN IMAGE RETRIEVAL

Peter G.B.Enser, Christine J.Sandom

School of Computing, Mathematical and Information Sciences,
University of Brighton, U.K.

Paul H. Lewis and Jonathon S. Hare

School of Electronics and Computer Science,
University of Southampton

TUTORIAL OUTLINE

Session 1 Peter Enser

Scoping the semantic content of images

Session 2 Christine Sandom

Textual subject description in practice

Session 3 Paul Lewis

Characterising the semantic gap

Session 4 Jonathon Hare

Multimodal searching and semantic spaces

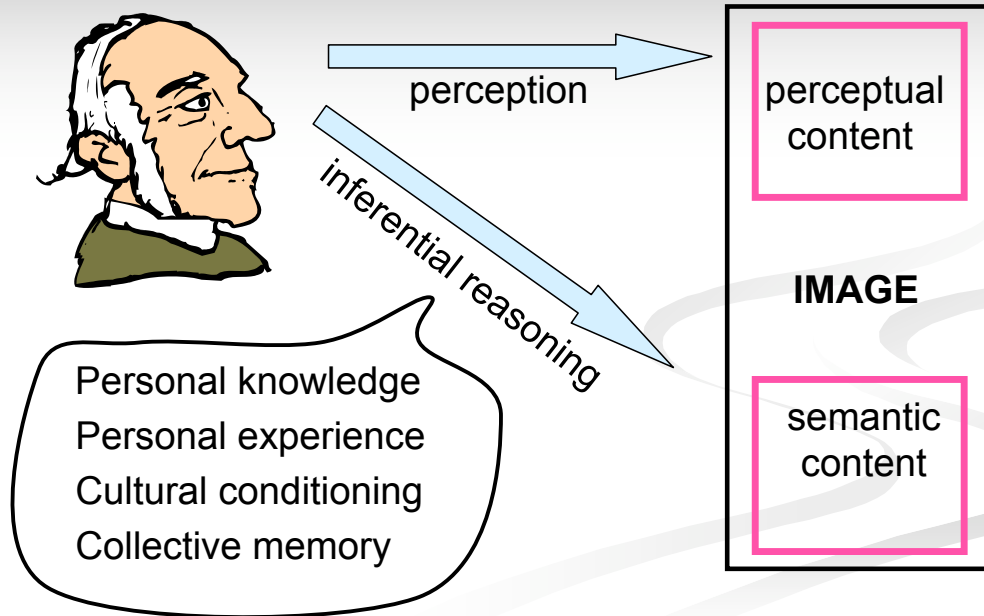
Scoping the semantic content of images

Peter Enser

Content

- **What do we understand by ‘the semantic content of images’?**
- **Characterising the semantic content of images**
facet structure
- **Seeking the semantic content of images**
faceted queries
- **The relationship between structured semantic content and visibility**

Image understanding from the human perspective



Semantic content
textual representation

Title
Caption / Description
Keywords
Summary
Shot list

Image retrieval: semantic paradigm

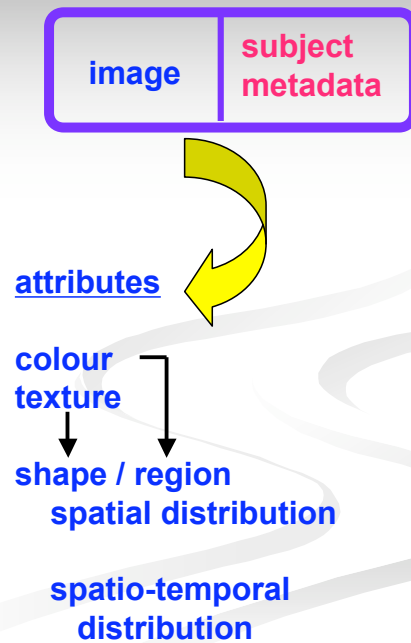
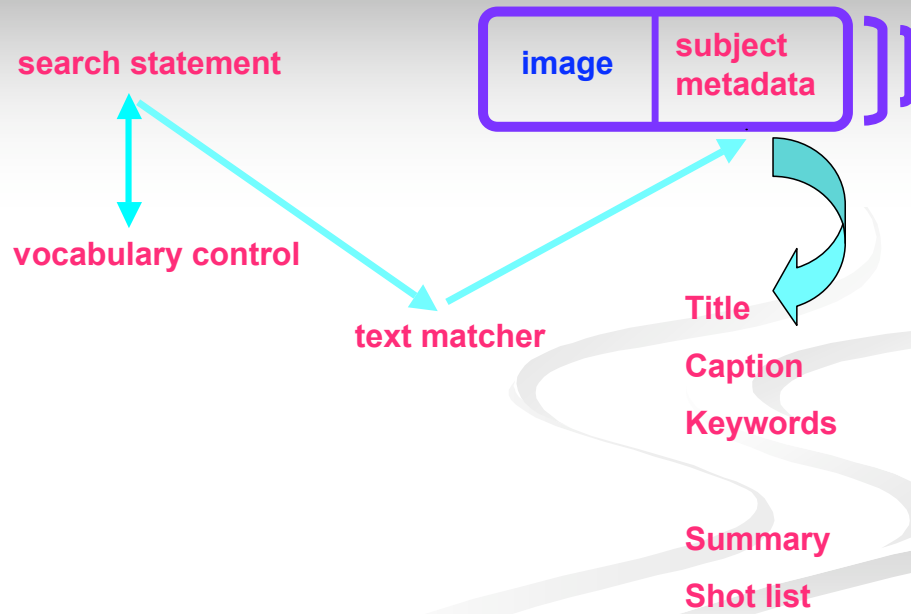
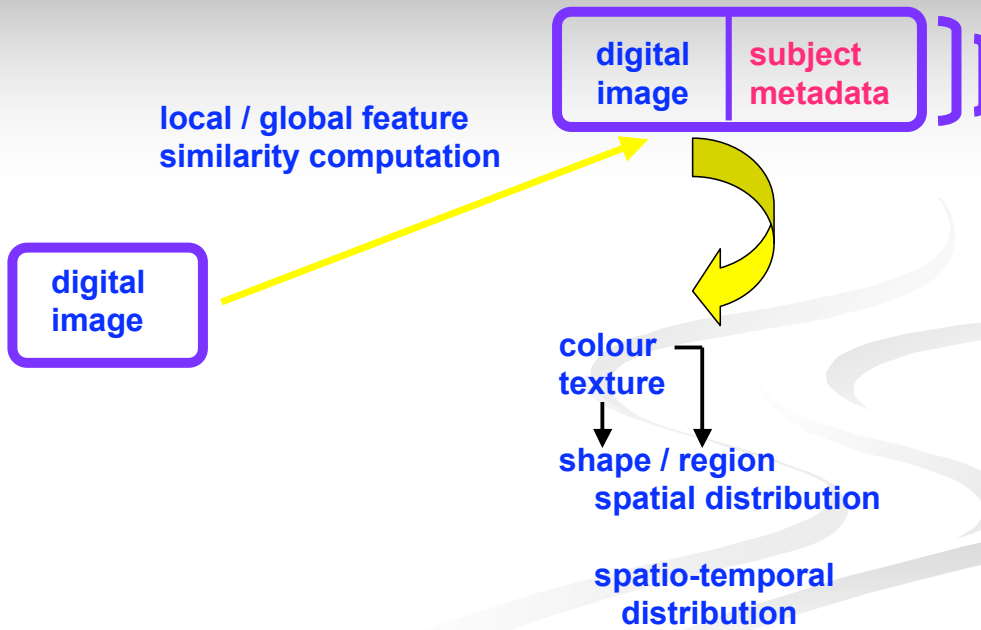


Image retrieval: content-based paradigm



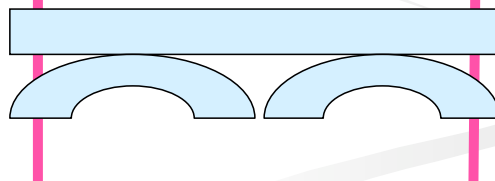
Pixel-encoded content
↓
Content-based Retrieval

Semantic gap

“... the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation”

(Smeulders *et al.*, 2000)

Semantic content
↓
Semantic Retrieval



Characterisation of the semantic content of images / queries

Eakins & Graham (1999); Greisdorf & O'Connor (2002)

Panofsky (1962)

Shatford (1986)

Enser & McGregor (1992); Armitage & Enser (1997)

Jaimes & Chang (2000); Jörgensen et al. (2001)

Specific

Generic

Abstract



Arts & Humanities
Research Council

Bridging the Semantic Gap in Visual Information Retrieval

MRG-AN6770/APN174290; 1/02/2004 – 31/01/2007

Peter G. B. Enser, Christine J. Sandom

School of Computing, Mathematical and Information Sciences,
University of Brighton

Paul H. Lewis, Jonathon S. Hare

School of Electronics and Computer Science, University of
Southampton

Characterisation of the semantic content of images

Shatford Layne (1994)

| | |
|----------------------|-----------------------|
| Object facet | I M A G E |
| Spatial facet | |
| Temporal facet | |
| Event/activity facet | |

The object facet

Generic Object Instance

blob interpreted at a basic level as man,
building, tree, vehicle, ...

The object facet

Generic Object Instance



Generic Object Class Hierarchy

blob interpreted at a basic level as man, building, tree, vehicle, ...

successively refined classification of an object employing knowledge-based inference drawn from visual attribute-values: man-in-uniform – policeman – traffic cop; residential dwelling - condominium; conifer – fir tree; ...

The object facet

Generic Object Instance



Generic Object Class Hierarchy



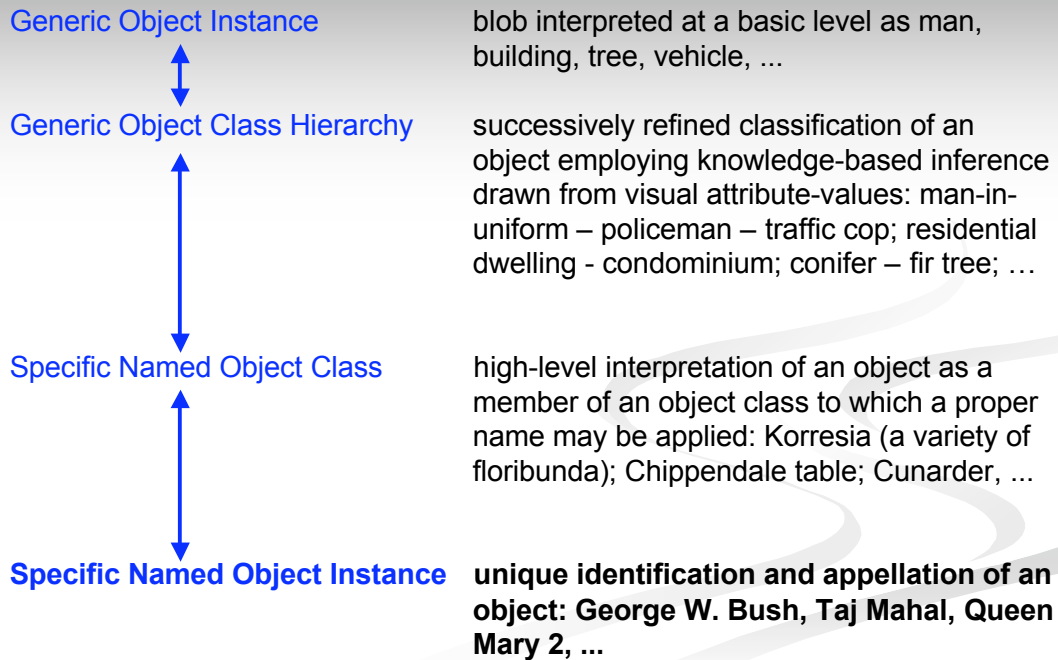
Specific Named Object Class

blob interpreted at a basic level as man, building, tree, vehicle, ...

successively refined classification of an object employing knowledge-based inference drawn from visual attribute-values: man-in-uniform – policeman – traffic cop; residential dwelling - condominium; conifer – fir tree; ...

high-level interpretation of an object as a member of an object class to which a proper name may be applied: Korresia (a variety of floribunda); Chippendale table; Cunarder, ...

The object facet



© Criss Sandom

| | |
|---|--|
| <p>Generic Object Instance</p> <p>Generic Object Class Hierarchy</p> <p>Specific Named Object Class</p> <p>Specific Named Object Instance</p> | <p>building, water, person, tree</p> <p>mausoleum, tomb, dome, minaret</p> <p>World Heritage Site</p> <p>Taj Mahal</p> |
|---|--|

Examples of requests which include the object facet

| | |
|---------------------------------------|---|
| Generic Object Instance | animals from movies |
| Generic Object Class Hierarchy | (very cute) dog guide dog; 'sniffer' dog |
| Specific Named Object Class | Norfolk Terrier Rottweiler |
| Specific Named Object Instance | the Shetland Sheepdog 'Champion Skye of Whytelaw' |

The spatial facet

| | |
|------------------------------------|---|
| Generic Location | the background to the image, the spatial context in which the object(s) within the image are placed; e.g. inside, outside, urban, countryside, field, lake, kitchen ... |
| Specific Location Hierarchy | successively refined geographical area, identified by proper name, e.g., Europe, Britain, England, London, Tottenham, Higham Road... |



Generic Location

outside

Specific Location Hierarchy

India, Uttar Pradesh, Agra

Examples of requests which include the spatial facet

| | |
|------------------------------------|---|
| Generic Location | ...various aspects of holiday making. Seaside, mountain or countryside holidays, holiday camps, caravans... |
| Specific Location Hierarchy | American air force in France Welsh mining c. 1930 Nancy Astor - if possible campaigning in Plymouth Ruins of Dresden after bombing Wigan unemployment New York buses |

The temporal facet

Generic Time natural periods, e.g., day, night, winter, summer...

periods of time expressed in other than standard temporal measurements, such as epochs or eras, e.g., Renaissance, Pleistocene, medieval, Victorian, ...

Specific Time time expressed in terms of standard quantifications: date, year or multiples thereof, e.g., twenty-first century, 1950s, 1896, September 2005, 12 June 2006, ...



Generic Time autumn, dusk, evening

Specific Time 1986

Examples of requests which include the temporal facet

| | |
|----------------------|---|
| Generic time | Houses of Parliament at night Frost fairs: ice on river in "olden days" Turkish troops in World War 1 |
| Specific time | Houses of Parliament, c. 1900 Churchill and Lord Halifax - walk to Parliament, March 28, 1938 1920s/30s/40s Really packed grounds or queuing at turnstiles of football grounds New York police cops. Must be in summer, 1950s-1960s, in caps |

The activity/event facet

| | |
|-----------------------------------|--|
| Generic Activity | gerunds associated with the object(s) in the image, e.g. running, bending, dancing. |
| Generic Event | a temporal and/or spatial relationship between a set of objects and a set of activities or actions which share a common purpose, e.g., basketball match, demonstration, wedding, ... |
| Specific Named Event Class | a type of event to which a proper name may be applied, e.g., Olympic Games, Rio Carnival, Papal Investiture, ... |
| Specific Event Instance | a unique occurrence of an event, e.g., 2006 Olympic Games, Investiture of Pope Benedict XVI, sinking of the 'Titanic', ... |



Ordination
© Getty Images

Archbishop of Canterbury Doctor Robert Runcie ordains the first women Deacons in the Church of England at Canterbury Cathedral. 27/02/87

Examples of requests which include the event/activity/significance facet

| | |
|-----------------------------------|--|
| Generic Activity | hop picking |
| Generic Event | registry office weddings |
| Specific Named Event Class | Early Olympic games and torch being lit at Olympia in Greece |
| Specific Event Instance | Stirling Moss winning Kentish 100 Trophy at Brands Hatch, 30 August 1968 |
| | 1967 Aberfan disaster |
| | West Ham v Bolton Wanderers - 1923 First Wembley cup final |
| | Bannister breaking tape on 4 minute |

Characterisation of the semantic content of images

| | |
|------------------------------|-----------------------|
| Object facet | I M A G E |
| Spatial facet | |
| Temporal facet | |
| Event/activity facet | |
| Topic | |
| Related concept/object class | |
| Abstract concept | |
| Context | |



Topic

Indian Architecture

Related concept/object class **Shah Jehan, Mumtaz Mahal, Islam**

Abstract concept **love, death, devotion, remembrance**

Context **built in memory of his wife Mumtaz Mahal, by Shah Jehan; completed 1648**

Examples of requests which include the abstract concept facet

History of **adolescence**

Industrial health

Anatomy, allegories of **life and death**

'**Society Life**'

The depiction of **vanity** in painting, the depiction of the female figure looking in the mirror, etc.

Victorian paintings on the subject - or incorporating the subject - of "**invention**"

The Antichrist, as 16th century minds might have perceived him.

Examples of requests which include affective content

Death, grief, mourning - 19th C British

... depictions of **happiness** – smiling, laughing, etc. together with more abstract representations in any period of art

Images which show a range of emotions, eg series of photographs showing someone in stages from **miserable** to **happy**

...**some irresponsible parents still purchase dogs for presents for children...** images ... to illustrate [this]

Very **cute** dog, preferably a Heinz variety

Stressed situations. People in business, rush hour traffic, tube/train travel, business man on phone, bank messengers, stock market, plane travel

Examples of requests which include features which must not be visible

Posters of agricultural scenes - but preferably **without any machinery or horses** in it - needs to be “timeless”

George V's coronation but **not procession or any royals**

J F Kennedy with woman - but **not Jacqueline or mother Rose**

Scoping the semantic content of images

Summary

The semantic content of images, as inferred by their viewers, and sought by their users, is multi-faceted and many-layered.

Representing and retrieving the richness of semantic content poses a very considerable challenge in metadata construction.

Textual subject description in practice

Christine Sandom

Content

- **Cataloguing and Classification of pictures:**
 - In theory**
 - In practice**
- **The Project's first phase: some results**
- **The Kennel Club case study**

Cataloguing pictures isn't just keywording

Non-subject metadata; includes:

- administrative metadata - to do with the management of the collection
- technical metadata - to do with the physical properties of the images
- Creator metadata – to do with the origination of the images

Subject metadata:

- Titles, captions, descriptions, keywords

Image requests may be for any of these data

Cataloguing schema and standards exist for picture collections, including:

- Visual Resources Association, VRA Core Categories, (<http://www.vraweb.org/vracore3.htm>)
- Dublin Core Element Set (<http://dublincore.org/>)
- Categories for the Description of Works of Art (CDWA) (http://www.getty.edu/research/conducting_research/standards/cdwa/)

But many picture collections have developed their own cataloguing systems – there is little uniformity in cataloguing practice.

Cataloguing (metadata) in practice

| | |
|------------------------|---|
| Title | View of the Taj Mahal, built by Emperor Shah Jahan (1592-1666), completed in 1643 (photo) |
| Additional Info | built in memory of his wife Mumtaz Mahal (d.1631); |
| Artist | Lahori, Ustad Ahmad (fl.1630-47) |
| Location | Agra, India |
| Century | C17th |
| Nationality | Persian |
| Classification: | INDIA & NEIGHBOURING COUNTRIES |
| Keywords: | mausoleum; minaret; minarets; islamic architecture; Indian; muslim; dome; domes; pool; reflection; India; Jehan |



Metadata from Bridgeman Art Library
<http://www.bridgeman.co.uk>

Cataloguing (metadata) in practice

| | |
|---------------------|--|
| Description: | Taj Maha, Agra, India, 17th century. Marble mausoleum built by Shah Jahan for his favourite wife, Mumtaz Mahal. Photograph. |
| Subjects: | |
| People | |
| Keywords: | 17th century; architectural; burial; burial chamber; century; colour; concept; death; decorative; dome; Islam; Islamic; Jahan, Shah; mausoleum; monument; Mughal empire; people; religion; religious; royal; royalty; seventeenth century; Shah Jahan; Taj Mahal; tomb |



Metadata from: Heritage Image Partnership
(Ann Ronan Picture Library)
<http://www.heritage-images.com/>

Subject metadata in practice – 6 images of the Taj Mahal

| | total | terms used |
|--------------------------------|-------|------------|
| Abstract Meaning/Mood | 12 | 12 |
| Generic Object Instance | 14 | 8 |
| Generic Object Class Hierarchy | 35 | 20 |
| Specific Named Object Instance | 8 | 3 * |
| Related concept | 18 | 17 |
| Adjectives | 19 | 15 |
| Generic Location | 1 | 1 |
| Specific Location Hierarchy | 9 | 5 |
| Generic time | 3 | 3 |
| Specific Time | 8 | 8 |
| Contextual, non-keyword | 16 | 15 |

* Taj Mahal; Taj; Mahal

Describing the content of images isn't just keywording

natural language – titles, captions, descriptions

controlled language – keywords, topics

thesauri / ontologies

subject heading lists

authority lists

classification schemes / taxonomies

Image description – natural and controlled language



Picture number: PERA000970

Title: Amy Johnson, British aviator, 12 May 1930.

Caption: In 1930 Johnson (1903-1941) became the first woman to fly solo from England to Australia, winning £10,000 from the 'Daily Mail' newspaper. Her plane was a De Havilland Gipsy Moth aircraft (nicknamed 'Jason'). In 1932, she set a record for the fastest solo flight from England to Capetown and broke that record four years later. In 1933, with her husband, James Mollison (1905-1959) she flew in a De Havilland biplane non-stop across the Atlantic in 39 hours. She joined the Air Transport Auxiliary as a pilot in WWII and died when her plane was lost over the Thames estuary.

Credit: NMPFT/Daily Herald Archive/Science & Society Picture Library

free text, natural language

Image and metadata from the Science and Society Picture Library

Image description – natural and controlled language



Picture number: PERA000970

Subject: PERSONALITIES

Controlled language terms

Title: Amy Johnson, British aviator, 12 May 1930.

Caption: In 1930 Johnson (1903-1941) became the first woman to fly solo from England to Australia, winning £10,000 from the 'Daily Mail' newspaper. Her plane was a De Havilland Gipsy Moth aircraft (nicknamed 'Jason'). In 1932, she set a record for the fastest solo flight from England to Capetown and broke that record four years later. In 1933, with her husband, James Mollison (1905-1959) she flew in a De Havilland biplane non-stop across the Atlantic in 39 hours. She joined the Air Transport Auxiliary as a pilot in WWII and died when her plane was lost over the Thames estuary.

Keywords:

[Personalities Johnson, Amy](#) [Woman](#) [Women](#) [Aviators](#) [Aviator](#) [on DH60 Moth](#) [Jason](#) [First](#) [Solo Flights](#) [Flying](#) [Pioneers](#) [De Havilland Gipsy Moth](#) [James Mollison](#) [Biplanes](#) [Air Transport Auxiliary](#) [Pilots](#) [Record Breakers](#) [Lost](#) [Mysteries](#) [Airplanes](#) [Aeroplanes](#) [Unattributed](#) [United Kingdom](#) [The 1930s \(1930-1938\)](#)

free text, natural language

Image and metadata from the Science and Society Picture Library

Example of thesaurus entry (TGMI)

Dogs

Used For: Puppies

Broader Term: Animals

Narrower Term: Bloodhounds; Bulldogs; Chow chows (Dogs); Collies; Dachshunds; Greyhounds; Hunting dogs; Irish wolfhounds; Poodles; Watchdogs; Working dogs

Related Term: Dog breeders; Dog licenses; Dog racing; Dog shows; Dog teams; Dog walking; Dogcatching; Dogs of war; Dogsledding; Fetch (Game); Kennels; Sled dog racing

Source: Library of Congress, (2006) *Thesaurus for Graphic Materials I: Subject Terms*
<http://www.loc.gov/rr/print/tgm1/>

Example of classification scheme: ICONCLASS

Classification scheme – a form of subject description in which each subject is represented by code or notation. A classification scheme devised specifically for images is ICONCLASS

“ICONCLASS is a subject specific international classification system for iconographic research and the documentation of images. ...

ICONCLASS is a collection of ready-made definitions of objects, persons, events, situations and abstract ideas that can be the subject of an image.”

From the ICONCLASS website: What is ICONCLASS.
<http://www.iconclass.nl>



ICONCLASS classification

Content of an image may be classified using one code ...

73C133



B. Gozzoli, Dance of Salome
(Image from CGFA <http://cgfa.sunsite.dk/>)

Composed as follows:

7 Bible

73 New Testament

73C public life of Christ: from his baptism until the Passion

73C1 story of John the Baptist (Matthew 3; Mark 1:4-11; Luke 3:1-22; John 1:19-34)

73C13 martyrdom and death of John the Baptist (Matthew 14:3-12; Mark 6:17-29)

73C133 Salome dancing during the banquet of Herod

ICONCLASS classification

Or many codes ...

Feast – 41C5

4

**Society, Civilization, Culture
material aspects of daily life**

41

nutrition, nourishment

41C

celebration meal, feast, banquet

41C5

**John the Baptist -
73A(JOHN THE
BAPTIST)**

7

Bible

73

New Testament

73A

**(scenes from the life of) John the Baptist and
Mary**

**73A(JOHN THE BAPTIST) series of scenes from the life
of John the Baptist**

**43C912 - woman
dancing alone**

4

**Society, Civilization, Culture
recreation, amusement**

43

sports, games and physical performances

43C

dancing

43C9

one person dancing alone

43C91

woman dancing alone

43C912



B. Gozzoli, Dance of Salome
(Image from CGFA
<http://cgfa.sunsite.dk/>)

Full ICONCASS classification (but more could be added)

73C133; 41C5; 73A(JOHN THE BAPTIST); 43C912



Arts & Humanities
Research Council

Bridging the Semantic Gap in Visual Information Retrieval

MRG-AN6770/APN174290; 1/02/2004 – 31/01/2007

Peter G. B. Enser, Christine J. Sandom

School of Computing, Mathematical and Information Sciences,
University of Brighton

Paul H. Lewis, Jonathon S. Hare

School of Electronics and Computer Science, University of
Southampton

Bridging the Semantic Gap in Visual Information Retrieval Project

Case study collections

| | |
|---|---------------------|
| Birmingham Central Library | Public Library |
| The Bridgeman Art Library | Commercial Library |
| BBC, News & Stills Archive | Private library |
| The Map Library, British Library | National Library |
| Centre for the Study of Cartoons and Caricature | Academic collection |
| East Sussex Record Office | County Archive |
| Edina (Hulton Getty) | Education Images |
| Guildhall Library Map and Print Collection | Public Library |
| Institution of Electrical Engineers | Private Library |
| The Kennel Club Picture Library | Private Library |
| National Monuments Record, English Heritage | National Collection |
| Photonica | Commercial |
| Royal Anthropological Institute | Academic collection |
| The Science and Society Picture Library | Academic collection |
| Wellcome Medical Photographic Library | Academic collection |
| West Sussex Record Office | County Archive |
| Witt Art Reference Library | Academic collection |

Bridging the Semantic Gap in Visual Information Retrieval Project

Information gathered

| | |
|--|-------------|
| Requests: | 492 |
| Images + metadata | 1058 |
| Number of collections | 17 |
| Collections with computer catalogues | 13 |
| Number with significant numbers of digitised images | 8 |

Bridging the Semantic Gap in Visual Information Retrieval Project

facet use in the collections

A typical catalogue record

Title: Isadora Duncan
Description: Isadora Duncan (1878 - 1927) performing a classical Greek dance.
Subject: Dance, Personality
Keywords: black & white, format portrait, clothing, costume, female, single, dancing, American, Isadora Duncan, dancer, Dancer, Dance
Date: circa 1905



Isadora Duncan
© Getty Images

| | |
|--------------------------------|--------------------------------|
| Topic | Dance, Personality |
| Generic Object Instance | clothing, female, single |
| Generic Object Class Hierarchy | costume, dancer |
| Specific Named Object Instance | Isadora Duncan |
| Adjectives | American |
| Generic Activity | dancing |
| Picture Attributes | black & white, format portrait |

Requests have facets too

the Trinity Church⁽¹⁾ in Paris⁽²⁾, 1930-1980⁽³⁾, where Messaien lived and was the organist⁽⁴⁾.

- (1) Specific Named Object Instance
- (2) Specific Location Hierarchy:
- (3) Specific Time
- (4) Contextual, non-keyword

picture of the Portugese court or royal family⁽¹⁾ in Lisbon⁽²⁾, in the 1720s⁽³⁾, showing in particular Princess Maria Barbara⁽⁴⁾.

- (1) Generic Object Class Hierarchy
- (2) Specific Location Hierarchy
- (3) Specific Time
- (4) Specific Named Object Instance

I need some posters⁽¹⁾ of agricultural scenes⁽²⁾ - but preferably without any machinery⁽³⁾ or horses⁽³⁾ in it - needs to be "timeless"⁽⁴⁾

- (1) Picture attribute
- (2) Topic
- (3) Generic Object Class Hierarchy [things not wanted]
- (4) Abstract Meaning/Mood

Bridging the Semantic Gap in Visual Information Retrieval Project Distribution of facet types in the image and request collections

| | Images Number | Images % | Requests Number | Requests % |
|--------------------------------|------------------|-------------|--------------------|---------------|
| Abstract Meaning/Mood | 122 | 12 | 15 | 3 |
| Adjectives | 284 | 27 | 33 | 7 |
| Contextual, non-keyword | 221 | 26 | 37 | 8 |
| Generic Activity | 252 | 24 | 34 | 7 |
| Generic Event | 105 | 10 | 19 | 4 |
| Generic Location | 261 | 25 | 17 | 4 |
| Generic Object Class Hierarchy | 637 | 60 | 89 | 18 |
| Generic Object Instance | 223 | 21 | 23 | 5 |
| Generic time | 231 | 22 | 19 | 4 |
| Related concept | 177 | 17 | | |
| Specific Event Instance | 17 | 2 | 17 | 4 |
| Specific Location Hierarchy: | 197 | 19 | 60 | 12 |
| Specific Named Event Class | 18 | 2 | 7 | 12 |
| Specific Named Object Class | 154 | 15 | 56 | 1 |
| Specific Named Object Instance | 237 | 22 | 111 | 23 |
| Specific Time | 170 | 16 | 66 | 14 |
| Topic | 407 | 38 | 40 | 8 |

Bridging the Semantic Gap in Visual Information Retrieval Project

Requests – Facet analysis summary– non-subject facets

| | numbers | % |
|--------------------|---------|----|
| Picture attributes | 111 | 23 |
| Creator | 108 | 22 |
| Title | 81 | 17 |
| Id number | 67 | 14 |
| Date | 11 | 2 |
| Picture example | 6 | 1 |
| Embedded Text | 1 | <1 |
| Publication | 2 | <1 |

Requests aren't always for subjects – many include reference numbers or creators

Both natural and controlled language are important for image retrieval, as is the mediation of the picture librarian

In the Bridging the Semantic Gap in Visual Information Retrieval Project

Of the requests received by collections that keyworded their images:

- 72% were subjects requests
- 16% were non-subject requests (e.g. for image reference numbers, creators, etc)
- 12% were a mix of subject and non-subject requests.

Both natural and controlled language are important for image retrieval, as is the mediation of the picture librarian

In the Bridging the Semantic Gap in Visual Information Retrieval Project

Of the images retrieved in response to 'subject' or mixed requests:

- **17% had all their search terms in controlled language**
- **30% had all their search terms in natural language**
- **53% had their search terms fully or partially in a combination of natural and controlled language**

Both natural and controlled language are important for image retrieval, as is the mediation of the picture librarian

In the Bridging the Semantic Gap in Visual Information Retrieval Project

Of the images retrieved in response to 'subject' requests:

- **18% either had none of the request's search terms in their metadata or had no subject metadata at all**
- **29% needed the search term(s) to be modified in order to match their subject metadata**

The Kennel Club

The first major case study

Small image collection: around 60,000 images, of which some 7,000 digitised, and 3,000 have descriptive metadata

Single domain: all dog and dog-show related



©Kennel club

The Kennel Club – descriptive terms – problems encountered in natural language indexing

| | |
|----------------------|---|
| Spelling, etc | Dalmation Puppys Best of Show (s/b Best in Show) |
| Ambiguities | Chocolate (noun or adjective) |
| Synonyms | Alsatian; German Shepherd Dog Behind; bottom; rump; rear |
| Acronyms | BSD / Belgian Shepherd Dog BIS / Best in Show |

Answer – create a thesaurus from the KC metadata, specifying preferred terms and relationships, to enable the description process to be formalised, and the retrieval operation to be Improved.

An extract from the KC Thesaurus

Sad

BT: Appearance

Path: Sad > Appearance > Dog Attributes > Dogs

Safe and Sound

BT: Demonstrations

Used for: **SAS**

Path: Safe and Sound > Demonstrations > Events > Dog Shows

Saluki

BT: Hounds

Path: Saluki > Hounds > Dogs

Samoyed

BT: Pastoral Dogs

Path: Samoyed > Pastoral Dogs > Dogs

Which was then converted to an ontology, but that's Paul's story.....

Characterising the Gap

Paul Lewis

The Challenge for Image Retrieval

- Representing and retrieving the richness of semantic content poses a very considerable challenge in metadata construction - **for humans**
- Developing systems to extract rich semantic descriptions from low level image features **automatically** (using content analysis, prior knowledge, machine learning etc) poses a monumental challenge and in the general case is far from being achieved

Queries and Images

- **Query representation**
 - typically a textual expression of the required semantics
 - in reality query by example is an uncommon paradigm
- **Image representation**
 - a 2-D array of pixels
 - only colour and/or brightness of each pixel is explicit
- **To extract the semantics** we need
 - transformations between representations
 - a large injection of prior knowledge
 - ie image understanding

What is the Semantic Gap in Image Retrieval?

The gap between information extractable automatically from the visual data and the interpretation a user may have for the same data

...typically between low level features and the image semantics

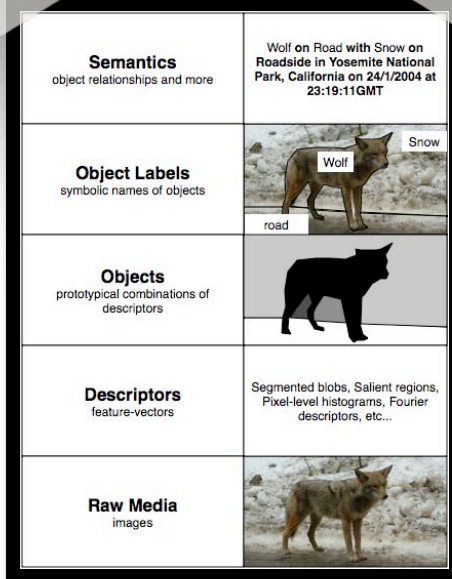
Representations for Image Understanding

LOW LEVEL

- Raw image - pixel level
- Low level descriptors (edge, colour, texture etc)
- Segmented regions, Salient regions
- Region descriptors (shape, colour, texture etc)
- Individual objects (region groupings?)
- Object labels
- Relations between labeled objects
- Scene descriptions
- Full semantics

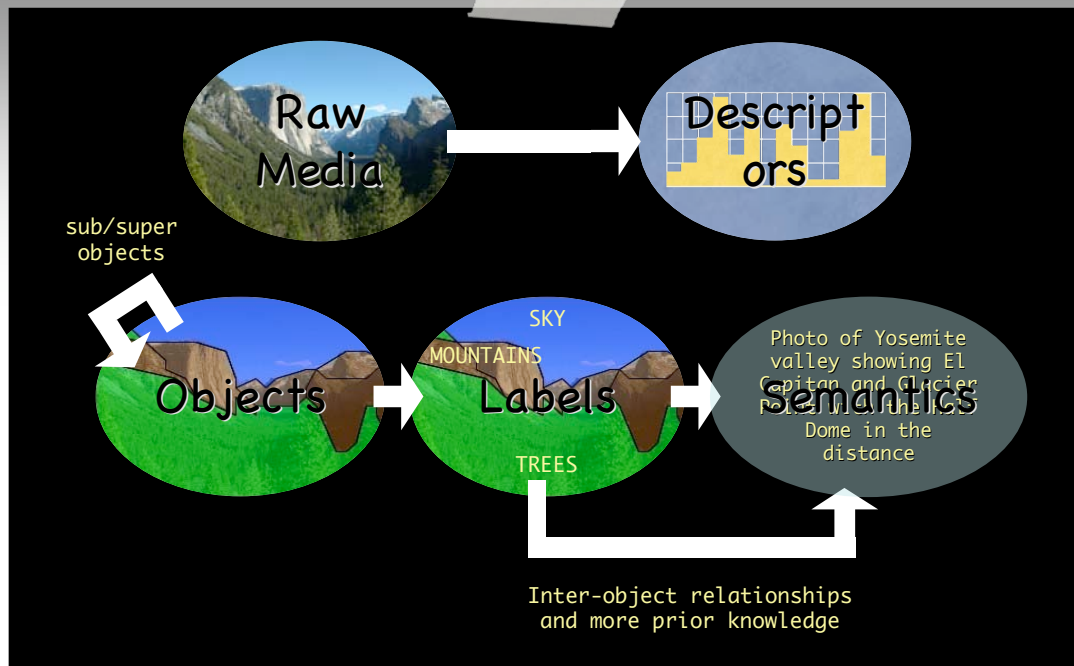
HIGH LEVEL

These representations are effectively staging posts
across the semantic gap
Prior knowledge is required for every transition



Characterising the Gap

A hierarchy of levels between media and semantics



Characterising the Gap

Of course, its not that simple...

Analysing the Gap

- Instructive to break the gap into two parts...

Analysing the Gap

from descriptors to labels



- Most current research into bridging the semantic gap is actually trying to bridge the gap between descriptors and (object) labels

The object facet

Generic Object Instance



Generic Object Class Hierarchy



Specific Named Object Class



Specific Named Object Instance

blob interpreted at a basic level as man, woman, building, tree, vehicle, ...

successively refined classification of an object employing knowledge-based inference drawn from visual attribute-values: man-in-uniform – policeman – traffic cop; residential dwelling - condominium; conifer, ...

high-level interpretation of an object as a member of an object class to which a proper name may be applied: Korresia (a variety of floribunda); Chippendale table; Cunarder, ...

unique identification and appellation of an object: George W. Bush, Taj Mahal, Queen Mary 2, ...

Increasing requirement for specific prior knowledge

Analysing the Gap

from labels to semantics



- However, user queries are typically formulated in terms of semantics

Remember the other Facets of Semantics

- The Spatial Facet
 - Generic Location
 - Specific Location Hierarchy
- The Temporal Facet
 - Generic Time
 - Specific Time
- The Activity/Event Facet
 - Generic Activities and Events
 - Specific Named Event Class
 - Specific Named Event Activity
- And... topics, related concepts, and context

Bridging the Gap Involves Transforming Representations and Injecting a lot of Prior Knowledge

- **At the bottom end of the gap** Most CBIR research has been here, bridging between raw media and low level descriptors representing image content in the QBE paradigm
- **Larger Bridges** -Some approaches (eg object detection/recognition and image annotation) go from raw media to objects and labels - very little work on extracting higher level semantics directly
- **At the top end of the gap** Use of knowledge structuring techniques to represent and reason over high level semantics (ontologies, description logics)

At The Bottom End of The Gap CBIR Research

- Active research area for at least 20 years
- Some progress in general CBIR
- Better progress in specific domains
- Still no significant CBIR in major web search engines
- Basic idea -use image content extracted from pixel data, as opposed to metadata, to assist retrieval
- Motivation - labour intensive nature and limitations in coverage of textual annotations

At the Bottom End of the Gap Query By Example

- Most popular CBIR paradigm -
 - find me an image similar this one
 - find me part of an image (a sub-image) like this one
 - find me a sub-image like this sub-image
- What might searchers mean by “similar”?
 - similar in every respect
 - similar colours, textures, shapes, objects
 - similar subject area, people, moods, seasons, [semantics](#)

The QBE Approach

- Extract image descriptors (feature vectors) from the images in the collection
- Extract same descriptor(s) from the query image
- Calculate similarity between query descriptors and descriptors for the image collection (typically based on distance in some feature space)
- Return best n matches from image collection in decreasing order of similarity

Descriptors

- Wide variety - some defined in MPEG 7 standard
- Are they trying to describe the whole image specific objects or aspects of the image?
- Global descriptors versus local descriptors
- Regions from segmentation versus salient regions from interest points
- Colour based descriptors have been the most popular
- Colour histograms, colour coherence vectors, colour layout
- Edge histograms/curvature scale space for shape, wavelets and more for texture
- Invariance properties: Colour invariants, Rotation, Scale, Translation invariance, Scale Invariant Feature Transform (SIFT)

CBIR System Survey (2001)

Remco C. Veltkamp, Mirela Tanase

<http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/cbir-survey.html>

- Compare 42 - mainly research systems - a few commercial ones

| | | | |
|-----------------------|-------------|-----------|--------------------|
| ADL | Excalibur | MIR | TODAI |
| AltaVista Photofinder | FIDS | NETRA | VIR Image Engine |
| Amore | FIR | Photobook | VisualSeek |
| ASSERT | FOCUS | Picasso | VP Image Retrieval |
| BDLP | ImageFinder | PicHunter | WebSEEK |
| CANDID | ImageMiner | PicToSeek | WebSeer |
| Blobworld | ImageRetro | QBIC | Wise |
| CANDID | ImageRover | Quicklook | |
| C-Bird | ImageScape | SIMBA | |
| Chabot | Jacob | SQUID | |
| CBVQ | LCPD | Surfimage | |
| DrawSeearch | MARS | SYNAPSE | |

Features Compared

Colour

global histogram
 Correlation histogram
 Average Colour Vector
 Colour Coherence vector
 Fixed subimage Histogram
 Region Histogram
 Dominant Colour
 Eigen Image

Texture

Wavelet
 Atomic texture features
 Random fields
 Local binary patterns
 Edge statistics

Shape


Template matching
 Fourier descriptors
 Elastic models
 Curvature scale space
 Elementary descriptors

Additional aspects

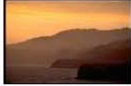




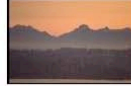
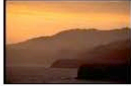










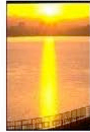
Keywords
 Layout
 Face detection

Example QBE Using Global Colour

Query Image:



ColourHistogram

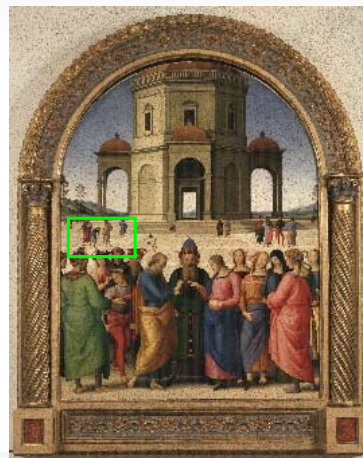
| | | | | | |
|---|---|---|---|---|---|
|  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  32 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |
|  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |
|  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |  192 x 128 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |  128 x 192 pixels - 0 Bytes - jpeg ©2006 Corel |

Example QBE Using CCV and Sub Image Matching



“Where does this image fragment come from?”

Multiscale - CCV (MCCV)



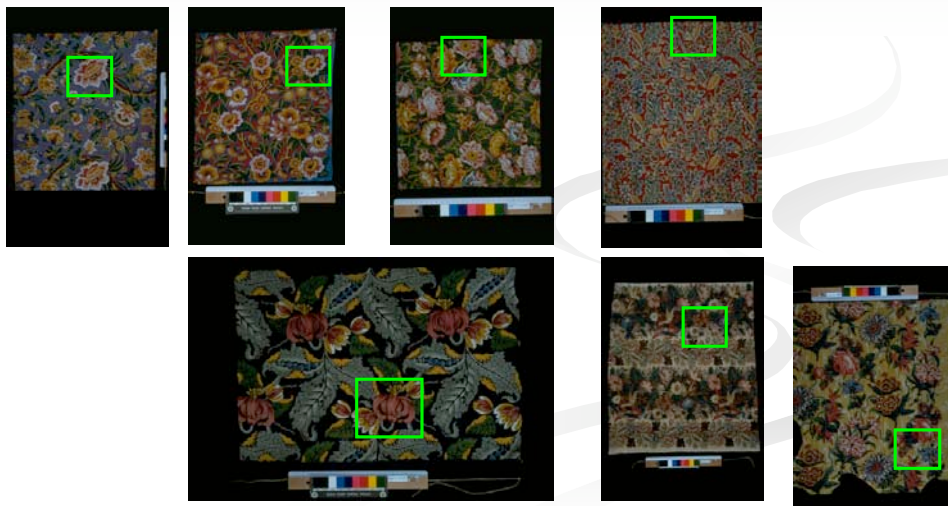
QBE Using Wavelet Based Descriptor and Sub-image Matching

- Query Image

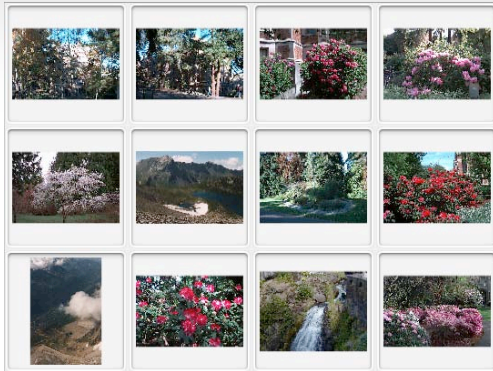


Best Matches

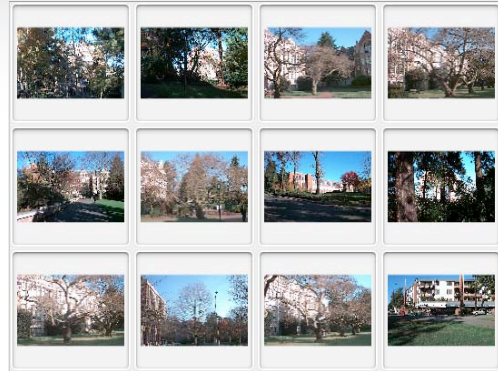
Retrieved results start from top-left to bottom right.



Global Colour Histogram Compared with Salient Region Matching



Global RGB Histogram



DoG Peaks RGB Histograms

Larger Bridges across the Gap Images to objects

- Increasing awareness of the need to work with semantics in image and video retrieval
- Movement from QBE to text driven modes
- Increasing number of papers on “so called” semantic retrieval
- Typically identify and label objects
- Face detection and recognition is prime example
- Growth in research on automatic image annotation

Computer Vision

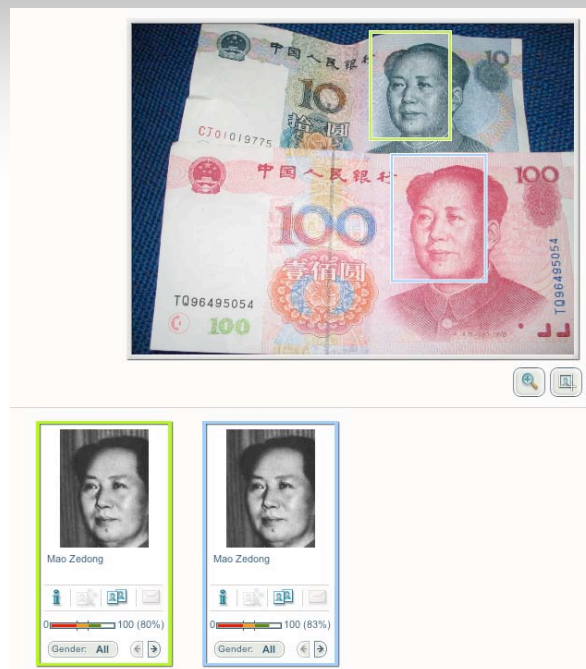
The ideal solution?

- Active research area for longer than CBIR
- Still no general solution
- Emphasis has shifted to motion
- Research overlaps with CBIR research (image descriptors, matching, segmentation, recognition, learning)
- Like CBIR -good progress in specific domains
- Industrial “machine vision” (robot vision/autonomous vehicle guidance/inspection)
- Model based vision/ Specific object recognition (eg faces)
- Ideas feed into CBIR research

Example of Face Detection



Example of Face Recognition






Images to Labels: Auto-annotation (attacking the gap from below)

- Takes us further across the semantic gap
- Basic idea - use training set of annotated images to learn relationship between image features and annotations
- Use learned relationships to predict annotations for un-annotated images
- Global approaches: - Learning without segmentation
- Local approaches:- Learning with segments or salient regions
- Many techniques:
 - Typically involve clustering descriptors as visterms
 - Co-occurrence of keywords and visterms
 - Machine translation
 - Cross-media relevance model
 - Probabilistic methods
 - Latent-spaces
 - Simple classifiers using low level features

Example of Global (Nearest Neighbour) Annotation

| | | | |
|--------------|---|---|---|
| good |  |  |  |
| Original | arctic, fox, snow | buildings, clothes, shops, street | buildings, sculpture, street, tree |
| Predicted NN | arctic, fox, head, snow | costume, street, village, buildings, people | building, sky, street, tree |
| bad |  |  |  |
| Original | people, pool, swimmers, water | buildings, church, town, tree | frozen, ice, snow |
| Predicted NN | lake, mountain, scotland, water | military, sky, tree | mountain, scotland, snow, winter |

Example of Region Based Annotations

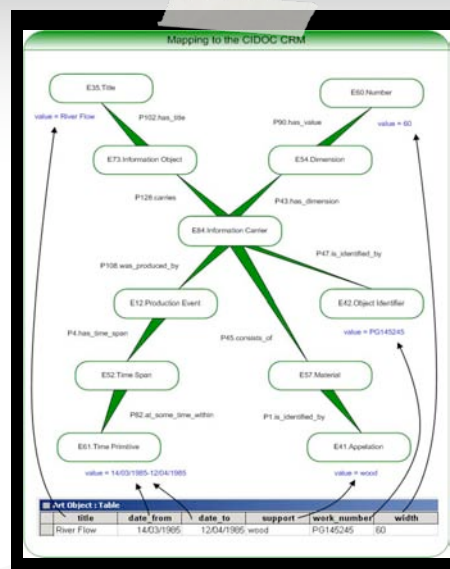
| | | | |
|---------------------------------|---|---|---|
| Images |  |  |  |
| Methods | | | |
| True Annotations | Tree, Bush, Sidewalk | Temple, Sky | Flower, Bush, Tree, Sidewalk, Building |
| Empirical Annotations | Tree, Building, People, Bush, Grass | Tree, Building, People, Bush, Grass | Tree, Building, People, Bush, Grass |
| Vector-Space Annotations | Tree, Bush | Tree, Building, Grass, Sidewalk, Pole, People, Clear Sky | Flower, Bush, Tree, Building, Partially Cloudy Sky |
| LSI Annotations | Tree, Bush, Grass, Sidewalk | Steps, Wall | Flower, Bush, Tree, Ground |
| Region-based CMRM Annotations | Tree, Flower, Building, Bush, Overcast sky | Tree, Building, People, Clear sky, Cloudy sky | Tree, Building, Bush, Flower, People |
| Saliency-based CMRM Annotations | Tree, Cloudy sky, Bush, Overcast sky, Post | Clear sky, Rock, Snow, Tree, Building | Tree, Bush, Flower, Ground, Building |

At the Top End of the Gap

- Increasing use of ontologies in connection with image annotation to provide structured domain knowledge
- A popular knowledge representation scheme
- Impetus from semantic web activity
- A shared conceptualisation of a domain
- Can structure and enhance the semantics of the image and its content
- Spatial relations between features in the image can help us infer relations between objects in the real world
- Ontologies can help us try to capture high level knowledge by modeling relations between concepts in the real world
- For image retrieval it is useful to consider the ontology in two parts:
content ontologies and **context** ontologies

Context Ontology The SCULPTEUR project

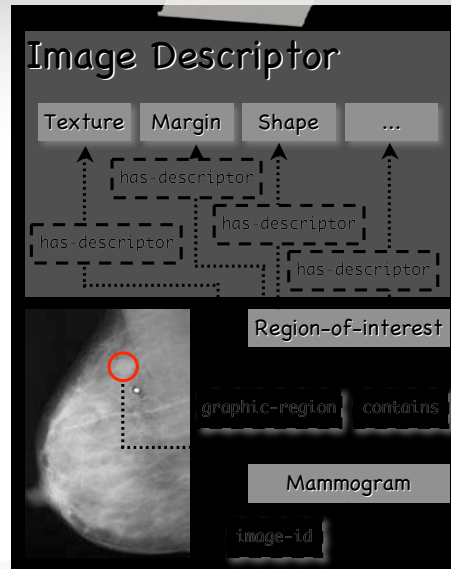
- SCULPTEUR was a large three year European project
- Finished last year
- Aimed to develop multimedia handling facilities for museums
- The CIDOC CRM was used to model contextual information about art objects
- Provided interoperability between museum collections



Content Ontology

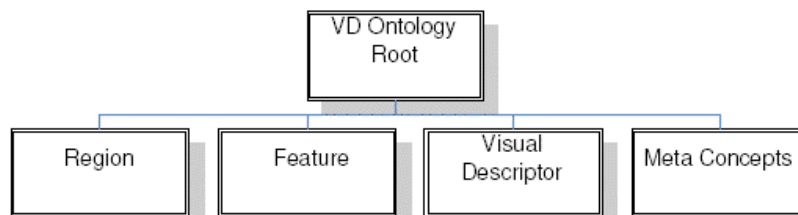
The MIAKT project

- 2 MIAKT was a two year UK (EPSRC) funded project formed from the AKT and MIAS IRCs
 - 3 Aimed to develop software to support the breast cancer screening process
- 4 Used (manual) ontology-mediated image annotation to mark-up the image content



Project at NTUA (National technical University of Athens)

- Use MPEG-7 (XML schema) to define visual descriptors
- Visual descriptor Ontology specifies how descriptors are defined through relations with their components
- Descriptors related to higher level concepts through inference rules defined in description logic
- System provides semantic segmentation and links labels with existing ontologies



(Some) Other Projects and Systems

- aceMedia <http://www.acemedia.org/>
- eCHASE <http://www.echase.org/>
(Uses semantic web technologies for cultural heritage MM)
- MUSCLE <http://www.muscle-noe.org/>
- SCHEMA <http://www.schema-ist.org/>
- PASCAL <http://www.pascal-network.org>
- BOEMIE <http://www.boemie.org/>
- K-Space <http://kspace.qmul.net/>
- The Knowledge Web <http://knowledgeweb.semanticweb.org/>
- MARVEL <http://www.research.ibm.com/marvel/details.html>
- MediaMill <http://www.mediamill.nl/>

How is the bridging of the gap progressing?

- In some domains- ok
- In general - not well
- Better descriptors, machine learning, ontology based knowledge structuring may be a way forward

----- but also see the next session

Multimodal Searching and Semantic Spaces

...or how to find images of Dalmatians when there is no metadata

Jonathon Hare

Contents

- Introduction
- The problem: un-annotated images
- A brief introduction to textual information techniques: Bags-of-words, the Vector-space model and Latent Semantic Indexing
- The image-to-semantics pipeline
 - Modelling image content
 - Visual terms and the bag-of-words for image content
- Semantic-Spaces
 - Problems associated with auto-annotation w.r.t search
 - Underlying theory and justification
 - A simple example
 - Experimental results
 - K9 Search demo
- Wrap-up

Introduction

- The previous sessions have described the issues associated with image retrieval from the practitioner perspective -- a problem that has become known as the 'semantic gap' in image retrieval.
- In the previous session, Criss described a number of techniques for improving the ability to search image collections using thesauri and ontological techniques.
 - Unfortunately, these techniques require the image collection to be extensively annotated.
- This final session aims to explore how the use of novel computational and mathematical techniques can be used to help improve content-based multimedia search by enabling textual search of un-annotated imagery.

Un-annotated Imagery

- Manually constructing metadata in order to index images is expensive.
 - Estimates of US\$50-\$100 per image for the level of detail involved in a typical Getty archival image (keywords, caption, title, description, dates, times, events).
 - Every day, the number of images is increasing.
 - In many domains, manually indexing everything is an impossible task!

Un-annotated Imagery - Example

- Kennel club image collection.
 - relatively small (~60,000 images)
 - ~7000 of those digitised.
 - ~3000 of those have subject metadata (mostly keywords), remainder have no information.
 - Each year, after the Crufts dog show they expect to receive an additional 4000-8000 (digital) images with no metadata other than date/time (and only then if the camera is set-up correctly)

A brief introduction to (textual) information retrieval techniques

- “Information retrieval (IR) is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand alone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data.” --wikipedia
- Traditionally in IR, documents represented as sets/bags of terms.
 - Bag: allows multiple instances of the same term.
 - terms normally equate to words in the document.
 - Document structure ignored.
 - Commonly used terms (stop words) often ignored.

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

| Indexed Term | Document 1 | Document 2 |
|--------------|------------|------------|
| | aid | 0 |
| all | 0 | 1 |
| back | 1 | 0 |
| brown | 1 | 0 |
| come | 0 | 1 |
| dog | 1 | 0 |
| fox | 1 | 0 |
| good | 0 | 1 |
| jump | 1 | 0 |
| lazy | 1 | 0 |
| men | 0 | 1 |
| now | 0 | 1 |
| over | 1 | 0 |
| party | 0 | 1 |
| quick | 1 | 0 |
| their | 0 | 1 |
| time | 0 | 1 |

Stop Word List

| |
|-----|
| for |
| is |
| of |
| 's |
| the |
| to |

courtesy of Phillip Resnik

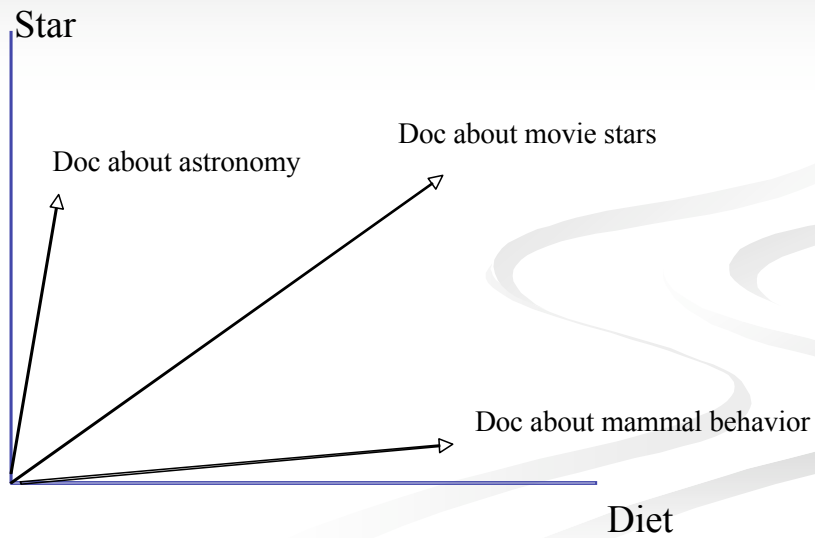
Vector-space model

- A collection of n documents with t distinct terms can be represented by a (sparse) matrix.
 - Documents are row vectors of the matrix.
 - Elements, w , are a function of the number of times term i occurs in document j .

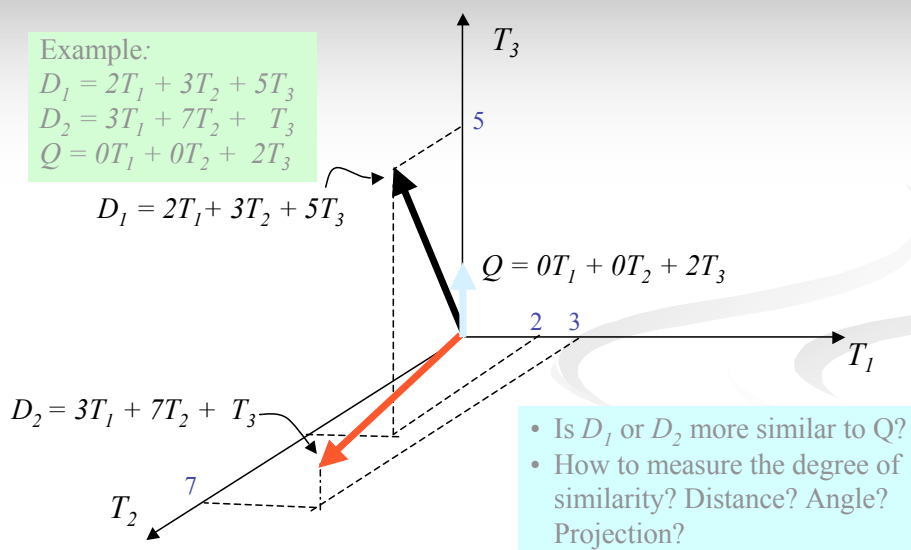
$$\begin{pmatrix}
 & T_1 & T_2 & \dots & T_t \\
 D_1 & w_{11} & w_{21} & \dots & w_{t1} \\
 D_2 & w_{12} & w_{22} & \dots & w_{t2} \\
 \vdots & \vdots & \vdots & & \vdots \\
 \vdots & \vdots & \vdots & & \vdots \\
 D_n & w_{1n} & w_{2n} & \dots & w_{tn}
 \end{pmatrix}$$

- A query can also be represented as a vector like a document.

Documents as Vectors



Geometric Interpretation



Assumption: Documents that are “close together” in space are similar in meaning.

Cosine Similarity

- Most popular measure of similarity between the query and document vectors is the cosine similarity measure.

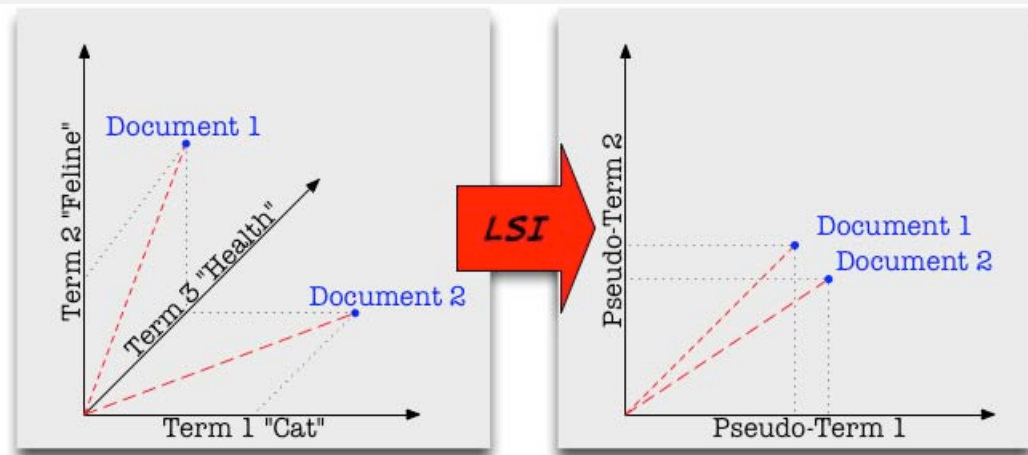
$$\frac{Q \cdot D}{|Q|^{1/2} \times |D|^{1/2}}$$

- ...the cosine of angle between the vectors.
 - Value of 1.0 means document and query are identical.
 - Distance metric easily obtained by taking \cos^{-1} (similarity).

Latent Semantic Indexing/Analysis

- A problem with the Vector-space model is that it relies on lexical matching.
 - A search for 'automobile' would not find documents about 'cars'.
 - Doesn't deal with synonymy (multiple terms with the same meaning), or polysemy (words with multiple meanings).
- Latent Semantic Indexing is a technique that attempts to take advantage of implicit higher-order structure in the term-document matrix in order to circumvent this problem.
 - Works by reducing the dimensionality of the vector-space in order to bring similar terms and documents closer together.

Latent Semantic Indexing: Dimensionality Reduction



Latent Semantic Indexing: How does it work?

- Definition: **Rank** of a matrix is the number of **linearly independent** rows or columns of that matrix.
- Some intuition: Given a large enough corpus, many terms or documents in a given term-document matrix will be **linearly dependent**.
 - i.e. the "Cat" and "Feline" term columns will be *approximately* linearly dependent on each other!
- ...So, all we need do is estimate a lower-rank version of the original term-document matrix.

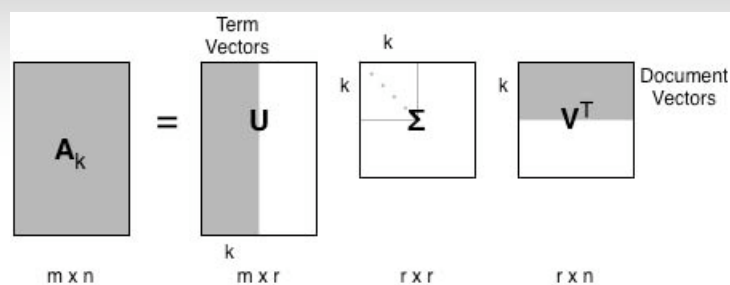
Latent Semantic Indexing: Mathematical details

- LSI uses a matrix decomposition called the Singular Value Decomposition (SVD).
 - SVD decomposes a matrix A into three separate matrices U , Σ , and V , such that:

$$A = U\Sigma V^T$$

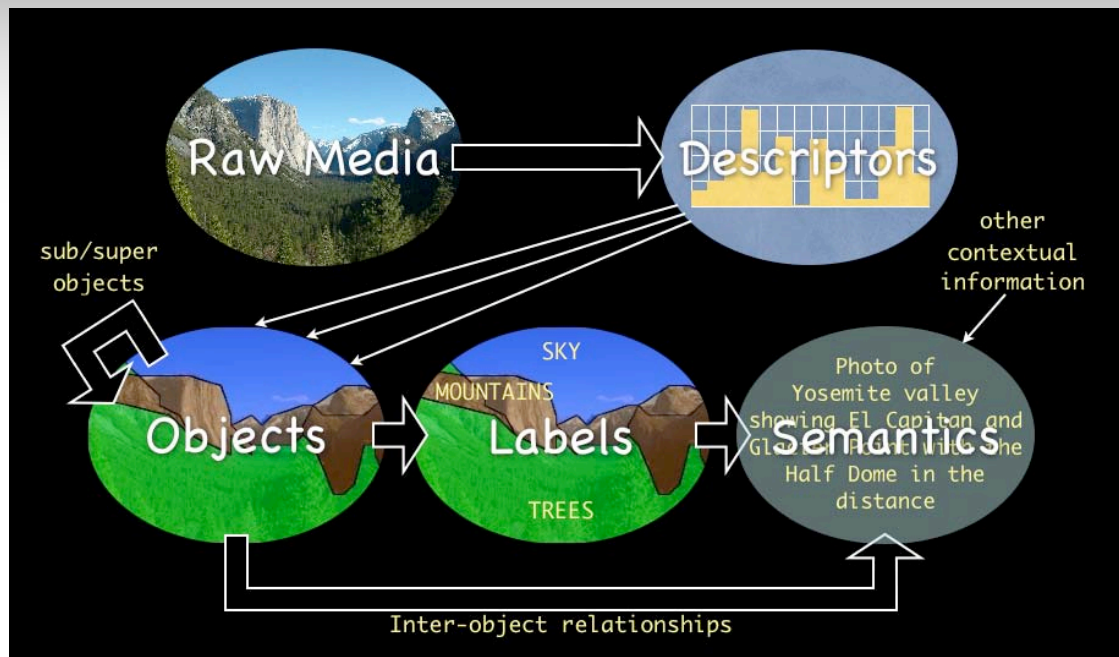
- The decomposition has some nice properties, in particular, Σ is a diagonal matrix containing the so-called singular values of A in monotonically decreasing order. The decomposition also separates the documents and terms into vector separate spaces (U and V respectively, both to each other by Σ).
- It can be shown that by selecting the K largest singular values and corresponding left and right vectors it is possible to estimate the rank K version of A with minimum error (in the least-squares sense).

Latent Semantic Indexing: Mathematical Details



Note: In practice it isn't necessary to reconstruct A_k . The term vectors in U and document vectors in V can be used directly.

Typical Information Flow from Raw Media to Semantics

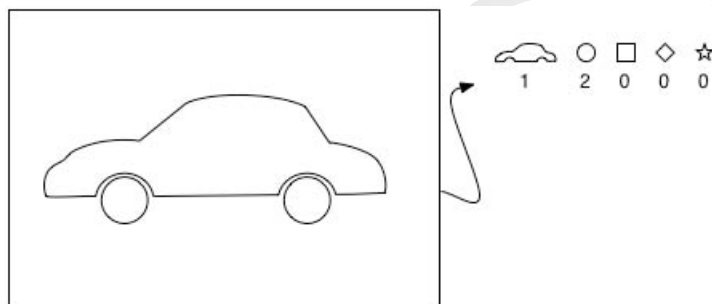


Modelling visual information

- In order to model the visual content of an image we can generate and extract descriptors or **feature-vectors**.
- Feature-vectors can describe many differing aspects of the image content.
 - Low level features:
 - Fourier transforms, wavelet decomposition, texture histograms, colour histograms, shape primitives, filter primitives, etc.
 - Higher-level features:
 - Faces, objects, etc.

Visual Term Representations

- The text indexing approaches described earlier use a bag-of-terms approach, whereby the documents are split into a vector counting the occurrences of the individual components.
 - It is possible to represent purely visual information in the same way, using feature vectors.
 - Some feature-vectors which are continuous may need to be quantised into discrete visual terms.

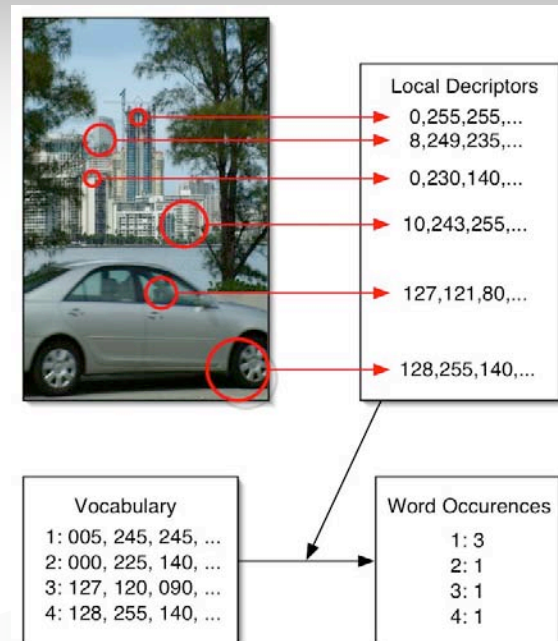


Global Colour Visual Terms

- A common way of indexing the global colours used in an image is the colour histogram.
 - The each bin of the histogram counts the number of pixels of the colour range represented by that bin.
 - The colour histogram can thus be used directly as a term occurrence vector in which each bin is represented as a visual term.

DoG/SIFT Visual Terms

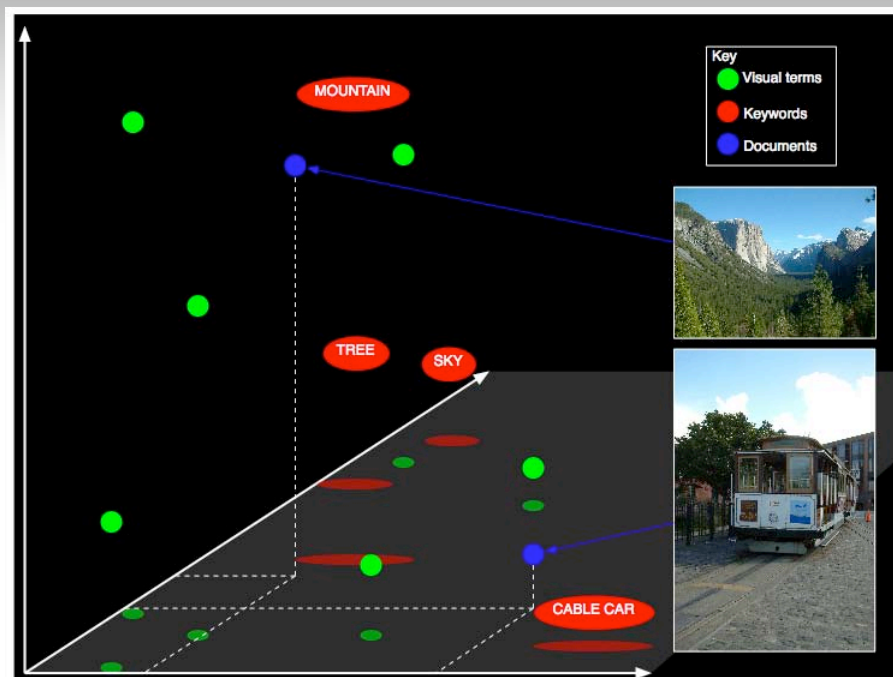
- Features based on Lowe's difference-of-Gaussian region detector and SIFT feature vector.
- A vocabulary of exemplar feature-vectors is learnt by applying k-means clustering to a training set of features.
- Feature-vectors can then be quantised to discrete visual terms by finding the closest exemplar in the vocabulary.



Semantic Spaces

- Idea based on a technique from text retrieval known as cross-language latent semantic indexing.
 - Extension of standard latent semantic indexing into multiple languages.
- Combines image content indexed as 'visual' terms with textual terms in the same term-occurrence vector.
 - Basically, just a big vector-space where the documents are images (or any kind of media) and terms from multiple languages/vocabularies/modalities are used.

Conceptual diagram of semantic space



Creating a semantic space using linear algebra *Overview*

- Technique consists of two stages.
 - Training stage:
 - Associations between image annotations/keywords and image features are learnt using annotated training images.
 - **In addition inter-keyword and inter-visual-term relations are discovered.**
 - Propagation stage:
 - The associations are applied to a set of un-annotated images.
 - The result is a searchable 'semantic space'.

Creating a semantic space using linear algebra

Mathematical Overview: Training

- Construct a *fully-observed* (multilingual) term-document matrix from the training images \mathbf{O} .
 - Combine annotation terms and visual terms into cross-domain word occurrence vectors and stack into matrix.
- Factorise \mathbf{O} in to a term matrix \mathbf{T} and document matrix \mathbf{D} such that $\mathbf{O} \approx \mathbf{T}\mathbf{D}$.
 - The factorisation is approximate because we want to *filter-out noise* in the \mathbf{O} matrix.

Creating a semantic space using linear algebra

Hang on! What do we mean by noise?

- Consider the following noise-free term document matrix with two documents and two terms. The first term is the keyword “Red”, and the second represents a visual term - the RGB triple {255, 0, 0}.

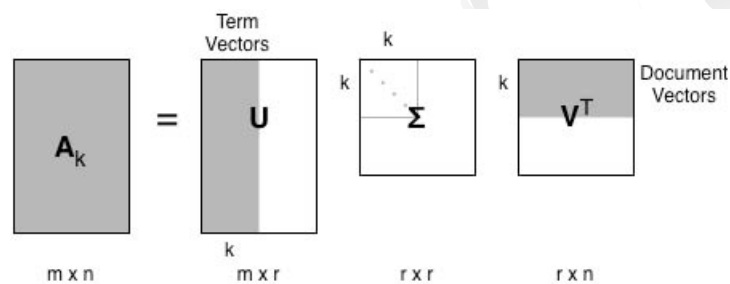
| | Doc. 1 | Doc. 2 |
|-------------|--------|--------|
| “Red” | 1 | 1 |
| {255, 0, 0} | 1 | 1 |

- The matrix has rank 1, indicating that the terms “Red” and “{255, 0, 0}” are equivalent (as are Doc. 1 and Doc. 2).
- Of course, in reality the matrix is not this clean, and so some form of filtering needs to be performed to match-up the relationships.

Creating a semantic space using linear algebra

So how do we filter and perform the factorisation?

- Simple!
 - The solution to the filtering and factorisation problem is simple: Use the Singular Value Decomposition and throw away all but the K most important singular values.
 - Just like in LSI
 - also closely related to Tomasi-Kanade factorisation for structure-from-motion.



Creating a semantic space using linear algebra

But hang-on, doesn't SVD give you 3 matrices?

- SVD gives you three matrices; $P=U\Sigma V^T$
 - However, we can take the K -subspace first two and multiply them together to get the T matrix.
 - $T = U_k \Sigma_k$
 - The K -subspace of the third matrix forms D .
 - $D = V_k^T$

Creating a semantic space using linear algebra

Mathematical Overview: Training II

$$\mathbf{O} \approx \mathbf{T}\mathbf{D}$$

- The **T** matrix tells us how terms are inter-related - each term forms a point in a vector-space, with the coordinates given by the relevant row of **T**.
- The **D** matrix tells us how documents (images) are inter-related - each term forms a point in a vector-space, with the coordinates given by the relevant column of **D**.
- Together, the **T** and **D** matrices capture the relationships between the terms and documents.

Creating a semantic space using linear algebra

Mathematical Overview: Propagation

- In order to propagate the associations learnt in the training stage to un-annotated images, the **T** matrix can be used to project a *partially observed* term-document matrix, **P**, into the space defined by **T** thus creating a new document matrix **Δ**.

$$\mathbf{P} = \mathbf{T}\mathbf{\Delta}$$

- It can be shown that to solve for **Δ**, it is simply a matter of pre-multiplying the transpose of **T** with **P**.

$$\mathbf{\Delta} = \mathbf{T}^T \mathbf{P}$$

Creating a semantic space using linear algebra

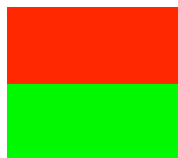
Mathematical Overview: Searching

- We now have two matrices:
 - T which contains the coordinates of each term is a *semantic space*,
 - and Δ , which contains the coordinates of each un-annotated image/document **in the same semantic space**.
 - So, in essence, we have a vector-space of terms and documents which can be searched using standard techniques.
 - i.e. ranking documents using cosine similarity to a query term.

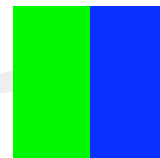
Simple Example

Training I

Annotated Training Images



Red, Green



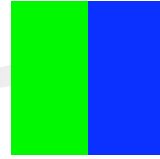
Green, Blue

Simple Example Training II

Simple feature-vector/visterms representing dominant primary colours



$\{255,0,0\}, \{0,255,0\}$



$\{0,255,0\}, \{0,0,255\}$

Simple Example Training III

Construct Term-Document Matrix $O_{(\text{TRAIN})}$:

| | | | |
|------------------------|-----------------|-----------------|---------------|
| | <i>Image #1</i> | <i>Image #2</i> | |
| $O_{(\text{TRAIN})} =$ | 1 | 0 | Red |
| | 1 | 1 | Green |
| | 0 | 1 | Blue |
| | 1 | 0 | $\{255,0,0\}$ |
| | 1 | 1 | $\{0,255,0\}$ |
| | 0 | 1 | $\{0,0,255\}$ |

Simple Example Training IV

Perform SVD, etc to calculate term (T) and document (D) matrices

$$T = \begin{bmatrix} -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \\ -0.289 & 0.500 \\ -0.577 & 0.000 \\ -0.289 & -0.500 \end{bmatrix}$$

$$D = \begin{bmatrix} -1.735 & -1.735 \\ 1.000 & -1.000 \end{bmatrix}$$

Simple Example Projection I

Un-annotated Testing Images



Simple Example Projection II

Simple feature-vector/visterms representing dominant primary colours



{255,0,0}



{0,255,0}



{0,0,255}

Simple Example Projection III

Construct Term-Document Matrix $O_{(\text{TEST})}$:

$$O_{(\text{TEST})} = \begin{array}{c} \text{Image \#1} \\ \text{Image \#2} \\ \text{Image \#3} \end{array} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{array}{l} \text{Red} \\ \text{Green} \\ \text{Blue} \\ \{255,0,0\} \\ \{0,255,0\} \\ \{0,0,255\} \end{array}$$

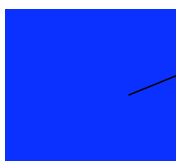
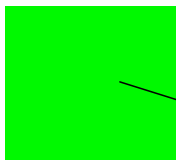
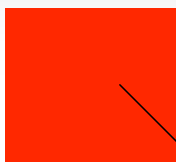
Simple Example Projection IV

Project $O_{(\text{TEST})}$ into the semantic space

$$D_{(\text{TEST})} = \begin{bmatrix} -0.289 & -0.577 & -0.289 \\ 0.500 & 0.000 & -0.500 \end{bmatrix}$$

Simple Example Querying

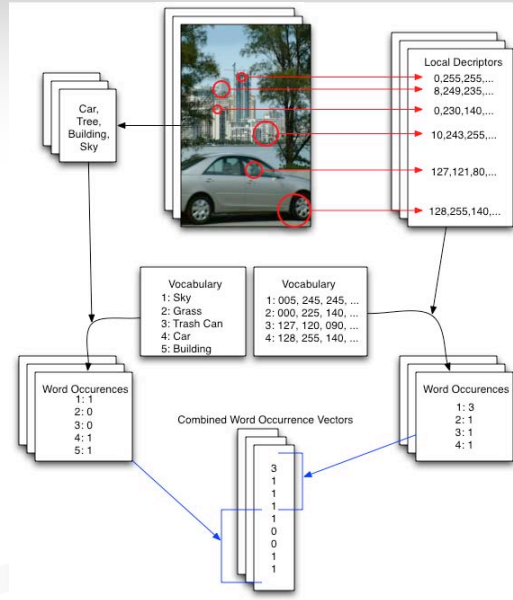
Its now possible to query the projected document space by keyword!



| Image | Cosine similarity with query: | | |
|-------|-------------------------------|-------|------|
| | Red | Green | Blue |
| 1 | 1.0 | 0.5 | -0.5 |
| 2 | 0.5 | 1.0 | 0.5 |
| 3 | -0.5 | 0.5 | 1.0 |

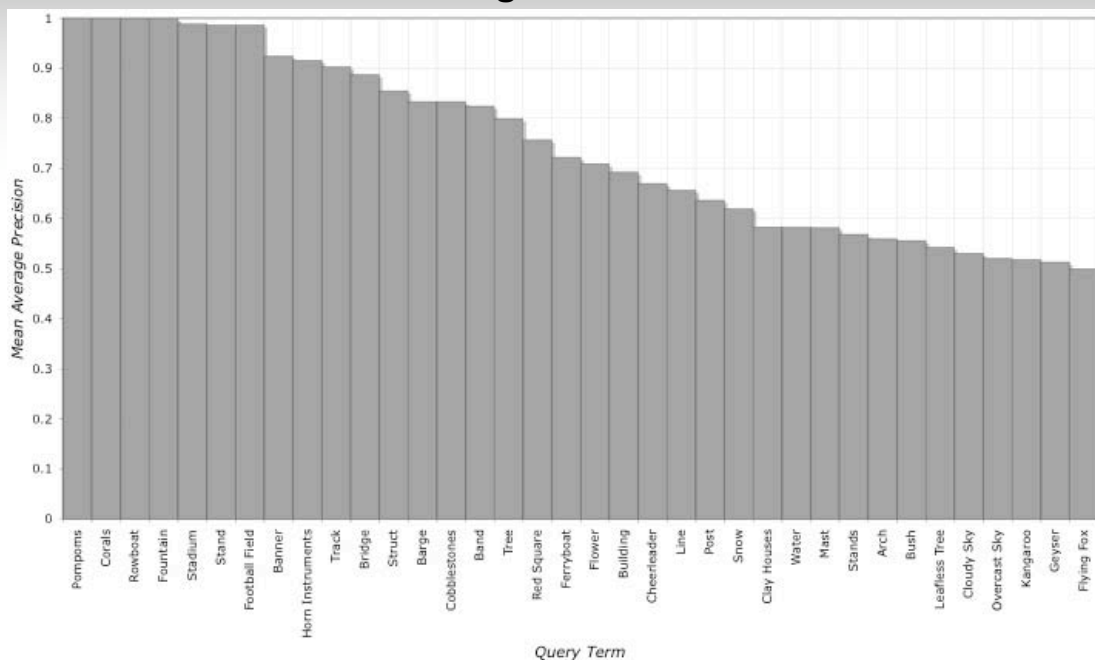
Experimental Results - Washington dataset + DoG/SIFT Visual Terms

- Washington data-set split randomly into two halves (training/testing).
- Quantised SIFT features used (3000 term vocabulary).
- Each keyword tested for retrieving relevant images and precision/recall recorded.
- K value optimised by trying to maximise MAP at the same time as keeping K as small as possible.



Experimental Results - Washington dataset + DoG/SIFT Visual Terms

Average Precision

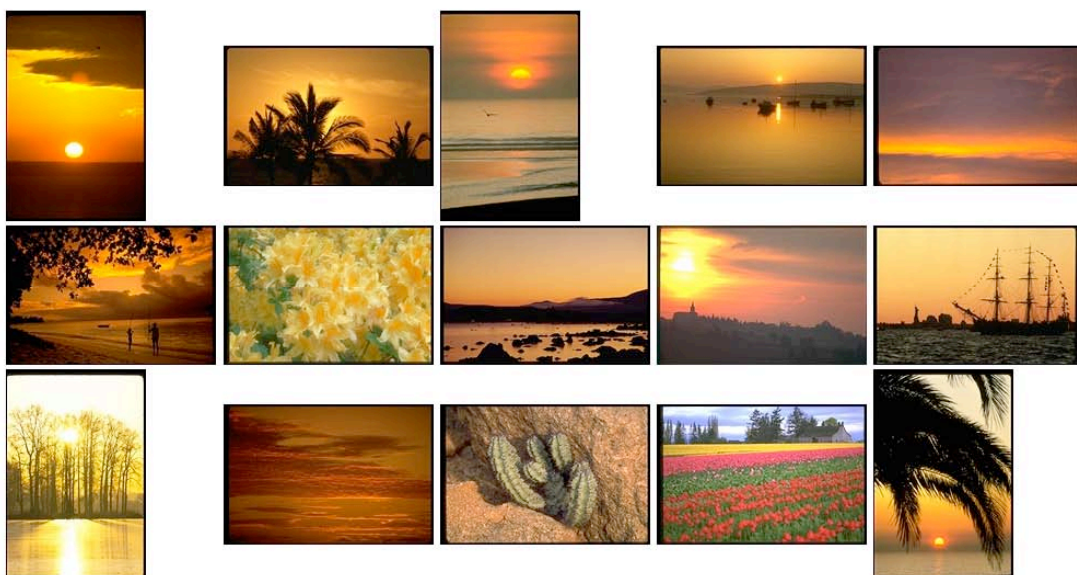


Experimental Results - Corel dataset + Global Colour Visual Terms

- RGB Histograms used as visual terms (each bin representing a single term).
- Standard collection: 500 test images, 4500 training images.
- Results quite impressive ~ comparable with Machine Translation auto-annotation technique (but remember we are using much simpler image features).
 - Works well for query keywords that are easily associated with a particular set of colours,
 - but not so well for the other keywords.

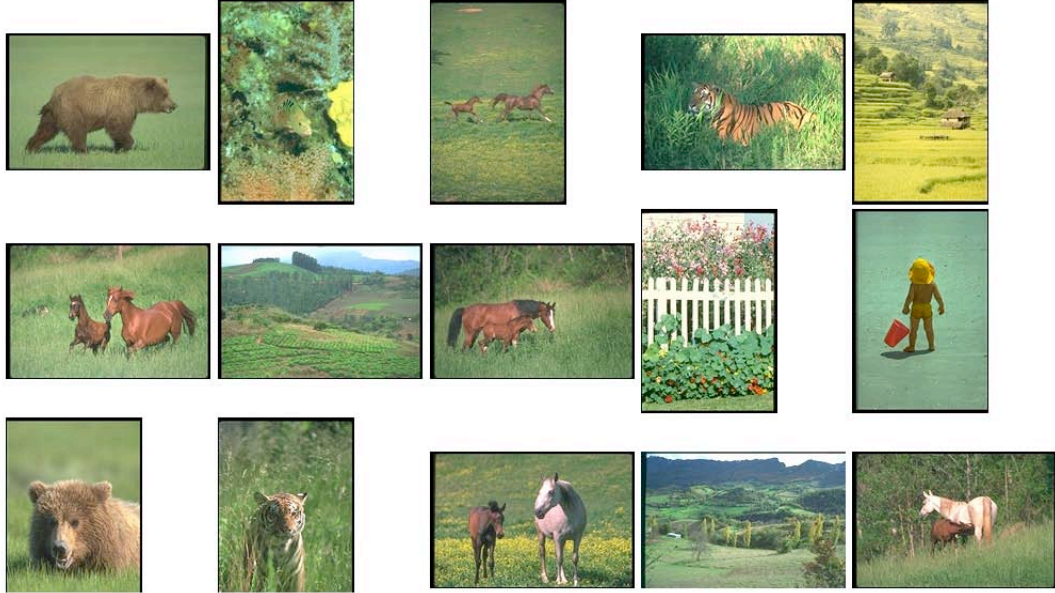
Experimental Results - Corel dataset + Global Colour Visual Terms

Query for 'sun'



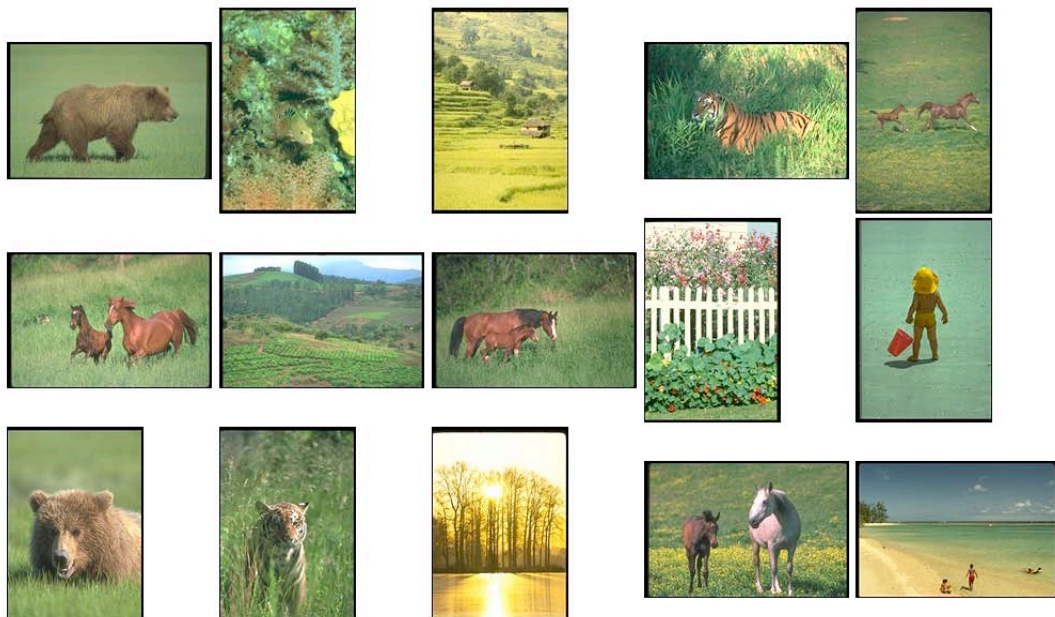
**Experimental Results - Corel dataset + Global Colour
Visual Terms**

Query for 'horse'



**Experimental Results - Corel dataset + Global Colour
Visual Terms**

Query for 'foals'



Real-world example - K9 Search

- Images of dog related activities from the Kennel Club.
 - About 3000 annotated images (noisy keywords).
 - ~3600 unannotated images.
 - Images indexed with quantised DoG/SIFT features.
 - 3000 term vocabulary, trained on Washington data-set.
 - Naively applied the factorisation technique, without any cleaning of the keywords.
 - **Demo...**

Mind the Gap: Another look at the problem of the semantic gap in image retrieval

Jonathon S. Hare^a, Paul H. Lewis^a, Peter G. B. Enser^b and Christine J. Sandom^b

^aSchool of Electronics and Computer Science, University of Southampton, UK;

^bSchool of Computing, Mathematical and Information Sciences, University of Brighton, UK

ABSTRACT

This paper attempts to review and characterise the problem of the semantic gap in image retrieval and the attempts being made to bridge it. In particular, we draw from our own experience in user queries, automatic annotation and ontological techniques. The first section of the paper describes a characterisation of the semantic gap as a hierarchy between the raw media and full semantic understanding of the media's content. The second section discusses real users' queries with respect to the semantic gap. The final sections of the paper describe our own experience in attempting to bridge the semantic gap. In particular we discuss our work on auto-annotation and semantic-space models of image retrieval in order to bridge the gap from the bottom up, and the use of ontologies, which capture more semantics than keyword object labels alone, as a technique for bridging the gap from the top down.

Keywords: Semantic Gap, Image Retrieval, Automatic Annotation, Ontologies, Cross Language Latent Semantic Indexing

1. INTRODUCTION

At the present time, many of the papers on image retrieval make reference to the problem of the semantic gap. There is a growing awareness in the community of many of the limitations of current retrieval technology and the incompatibility between queries formulated by searchers and the facilities that have been implemented so far in image retrieval systems. Whether in papers by researchers of content based techniques who believe they may be providing a bridge to the semantics or by professional searchers frustrated by the inability of systems to accommodate their queries, the semantic gap appears as a recurring issue in their endeavours.

In a review of the early years of content-based retrieval, Smeulders *et al*¹ define the semantic gap as the “lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation”. At the end of the survey the authors conclude that: “A critical point in the advancement of content-based retrieval is the semantic gap, where the meaning of an image is rarely self-evident. ...The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics.”

Smeulders *et al* also mention another gap of relevance to content based retrieval, the sensory gap, which they define as “the gap between the object in the world and the information in a (computational) description derived from a recording of that scene”. Although this is an important issue, we will confine ourselves in this paper to the problem of the semantic gap.

Our aim in this paper is to try and characterise the gap rather more clearly and explore what is and is not being done to bridge it. We begin in Section 2 by defining the gap more carefully to aid later discussion and suggest that it can be divided usefully into a series of smaller gaps between definable representations. In Section 3 we look at queries and their categorisation in order to show how an awareness of the requirements of real searchers can sharpen an understanding of the limiting effects of the gap. In sections 4 and 5 we present some of our own gap bridging work and summarise that of others. In particular, in Section 4, we describe some work on image annotation which attempts to build bridges between low level features and higher level “object” labels: i.e. tackling the gap from the bottom upwards. In Section 5 we argue that ontologies and ideas from emerging

Further author information: E-mail: {jsh02r | phl}@ecs.soton.ac.uk, {p.g.b.enser | c.sandom}@bton.ac.uk

semantic web technology can help to represent and integrate higher-level knowledge about images, potentially capturing more of the semantics than a set of “object” labels alone. In Section 6 we draw some brief conclusions and outline future work.

2. CHARACTERISING THE GAP

The semantic gap manifests itself as a computational problem in image retrieval. The representations one can compute from raw image data cannot be readily transformed to high-level representations of the semantics that the images convey and in which users typically prefer to articulate their queries. It may be useful to look at the series of representations between and including the two extremes. At the lowest level of representation are the raw media, which in this particular case refers to raw images but our analysis is quite general. Content-based retrieval algorithms typically extract feature vectors, or in MPEG 7 parlance, descriptors and these constitute the second level. They may be simple colour histograms, texture statistics or more sophisticated feature vectors developed for content based tasks and may represent parts of an image or the whole image. At a higher level there are representations of “objects” which may be prototype combinations of feature vectors or some other more explicit representation. Once identified, these objects may be given symbolic labels, ideally the names of the objects. This is a simplification as labels may be general or specific e.g. a mountain or Mount Everest. Even where it is possible, labelling all the objects in an image does not typically capture all the semantics. The relationships between the objects as depicted in the image, and the variety of connotations invoked, the implied relationship with the world at large, implied actions, and the broader context, all contribute to the rich high level full semantic representation of the image. The hierarchy of levels between the raw media and full semantics is illustrated in Figure 1.

Needless to say, this is a gross simplification. For example, the objects may have components, with their own labels. But this simple notation is sufficient to enable us to characterise the gap.

The first thing to observe is that the characteristics of the gap vary from one problem to another. There are (rather rare) situations involving simple images where it is possible to pass computationally from the raw image through descriptors to extraction of objects, labels and any required semantics fully automatically. An example might be a robot vision system that can identify parts on a conveyer belt and capture all relevant semantics to use the captured images effectively. But in general the semantic gap starts at the descriptors and goes all the way up to the semantics. In some situations it is possible to extract objects and assign labels but a gap may remain between the labels and the semantics. That is, we may be able to identify the names of the objects in an image but the meaning or significance of the image remains unknown. Our system may be capable of identifying that there are people and buildings in the image but is not able to recognise that this is a demonstration involving police and students. In some cases the required semantics in a query may be expressed directly as a set of object labels but more often the expressed semantics in the query are at a higher level than simply object label lists.

It may be instructive to see the gap in two major sections, the gap between the descriptors and object labels and the gap between the labelled objects and the full semantics.

Two important observations are that firstly, as we will see later, user queries are typically formulated in terms of the semantics and secondly, much of the interesting work which is attempting to bridge the semantic gap automatically is tackling the gap between descriptors and labels and not that between the labels and the full semantics.

The problem of the gap presents itself particularly because, although many image analysis researchers would like queries to be formulated in terms of the descriptors or using the query by example paradigm which can often be reduced to the problem of descriptor matching, most genuine users of image collections formulate their queries at the other side of the gap in terms of the semantics or at best in terms of labels. A number of studies have tried to characterise queries in some formal way and in the next section we review this work as a significant activity, which is taking place to understand the requirements at one side of the gap.

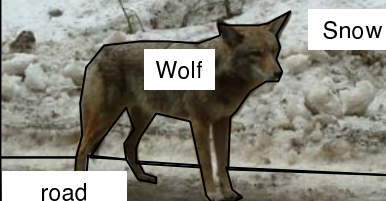
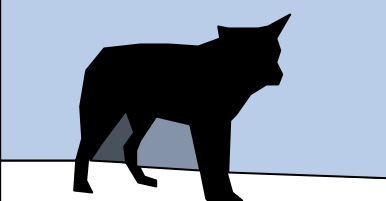

| | |
|---|---|
| <p style="text-align: center;">Semantics <i>object relationships and more</i></p> | <p style="text-align: center;">Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11GMT</p> |
| <p style="text-align: center;">Object Labels <i>symbolic names of objects</i></p> |  |
| <p style="text-align: center;">Objects <i>prototypical combinations of descriptors</i></p> |  |
| <p style="text-align: center;">Descriptors <i>feature-vectors</i></p> | <p style="text-align: center;">Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...</p> |
| <p style="text-align: center;">Raw Media <i>images</i></p> |  |

Figure 1. The Semantic Gap: Hierarchy of levels between the raw media and full semantics.

3. USERS' QUERIES SHOULD BE THE DRIVER

The hallmark of a good image retrieval system is its ability to respond to queries posed by searchers, presented in the desired way. There has been a tendency for much image retrieval research to ignore the issue of user queries and to concentrate on content-based techniques. In spite of this, some investigators have analysed and characterised image queries, providing valuable insights for retrieval system design and highlighting rather starkly the problem of the semantic gap.

One of the earliest investigations of user queries was undertaken by Enser and McGregor² who categorised requests in terms of unique/non-unique features, cross-classified by refinement/non-refinement whereby a request is qualified by the addition of temporal, spatial, affective, technical or other facets. Such facets generally serve to locate a query at the high-level, full semantic end of the representation spectrum

Further studies,^{3,4} analysed user requests using a tool which recognised the multi-layering of semantic content in both still and moving documentary imagery. This multi-layering has been described in different ways. The art historian Panofsky, working with creative images, identified 'pre-iconographic', 'iconographic' and 'iconologic' levels of expression,⁵ which Shatford's generalisation in terms of generic, specific and abstract

| | |
|-------------|---|
| Title | Roomy Fridge |
| Date | circa 1952 |
| Description | An English Electric 76A Refrigerator with an internal storage capacity of 7.6 cubic feet, a substantial increase on the standard model. |
| Subject | Domestic Life |
| Keywords | black & white, format landscape, Europe, Britain, England, appliance, kitchen appliance, food, drink, single, female, bending |

Table 1. Metadata used for resolving the request of the query ‘A photo of a 1950s fridge’.



Figure 2. Roomy Fridge ©Getty Images

content, respectively, made amenable to general purpose documentary images.⁶ Shatford is more particularly associated with the of-ness and about-ness of image content, the former corresponding with the denotational properties, the latter with connotational properties of visual imagery. Such an approach resonates with the perceptual and interpretive layers of meaning postulated by Jørgensen⁷ and with recent classification of queries postulated by Hollink *et al.*⁸

Eakins & Graham⁹ offer an alternative three level classification of queries based on primitive features, derived (sometimes known as logical) features and abstract attributes, the latter involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. In our experiences within the realm of real user needs for visual imagery, both still and moving, the incidence of requests based on primitive features is very rare indeed.

Within the particular context of archival imagery, a large proportion of queries typically seek uniquely defined objects; e.g. ‘HMS Volunteer’; ‘Balshagary School (Glasgow)’; ‘Marie Curie’.^{2,4} A study of archival moving image requests³ generated a similar finding, with 68% of the requests including at least one uniquely defined facet; e.g. ‘Stirling Moss winning Kentish 100 Trophy at Brands Hatch, 30 August 1968’. Depiction of an event such as this, necessarily invokes the full semantic level because any event is a temporal interpretative relationship between objects. Similarly, it can be argued that the attaching of a label to a place invokes full semantics because a place has to be interpreted as a spatial relationship between objects. In all such cases, detailed textual metadata is necessary in order to represent and recover the full semantic content.

The essential nature of textual metadata is emphasised, furthermore, by the frequent occurrence of requests that address issues of identification, interpretation and significance of depicted features within still images.^{10,11}

For example, a request for ‘A photo of a 1950s fridge’ was resolved using the metadata in Table 1.¹² The corresponding image is shown in Figure 2.

Within the metadata reference is made to a specific manufacturer and model of the depicted object, whilst enabling requests at the more generic levels of ‘refrigerator’ or ‘fridge’ and ‘kitchen appliance’ to be satisfied. Furthermore, the process of identification often involves context, recognition of which would seem to invoke high-level cognitive analysis supported by domain and tacit knowledge (*viz* ‘Domestic Life’ in the above example).

In general, contextual anchorage is an important role played by textual annotation within the image metadata. The request for a 1950s fridge is an example of query ‘refinement’ or qualification, moreover, which needs textual annotation for its resolution.

A yet more pressing need for supporting textual metadata occurs when the significance of some visual feature is at issue. Studies of user need have revealed that significance is an important - because frequently encountered - class of request. The problem here is that significance is a non-visible attribute, which can only be anchored to an image by means of some explanatory text. Significance frequently takes the form of the first or last occasion when some visible feature occurred in time, or the first/only/last instantiation of some physical object. Clearly, significance has no counterpart in low-level features of an image. Image retrieval operations that address significance, necessarily involve the resolution of verbalised queries by matching operations conducted with textual metadata.

When the requester’s focus of interest lies with the abstract or affective content of the image, wanting images of ‘suffering’ or ‘happiness’, for example, appropriate textual cues within the metadata will help to condition our interpretation of the image.

An even more challenging scenario in this context occurs when image searchers specify features that must not be present in the retrieved image; e.g. ‘George V’s coronation but not procession or any royals’. Provision is sometimes made in controlled keywording schemes to indicate the absence of commonly visible features (e.g., ‘no people’, ‘alone’).

The above examples combine to indicate the scale of the challenge faced in trying to overcome the constraints innate within current automatic image indexing and retrieval techniques on their ability to recover appropriate images in response to real expressions of need.

4. IMAGE ANNOTATION AND SEMANTIC SPACES: ATTACKING THE GAP FROM BELOW

By developing systems to automatically annotate image content, we can attempt to identify symbolic labels to apply to the image, or parts of the image. Auto-annotation attempts to bridge the gap between descriptors and symbolic labels by learning which combinations of descriptors represent objects, and what the labels of the objects should be.

The first attempt at automatic annotation was perhaps the work of Mori *et al.*,¹³ which attempted to apply a co-occurrence model to keywords and low-level features of rectangular image regions. The current techniques for auto-annotation generally fall into two categories; those that first segment images into regions, or ‘blobs’ and those that take a more scene-orientated approach, using global information. The segmentation approach has recently been pursued by a number of researchers. Duygulu *et al.*¹⁴ proposed a method by which a machine translation model was applied to translate between keyword annotations and a discrete vocabulary of clustered ‘blobs’. The data-set proposed by Duygulu *et al.*¹⁴ has become a popular benchmark of annotation systems in the literature. Jeon *et al.*¹⁵ improved on the results of Duygulu *et al.*¹⁴ by recasting the problem as cross-lingual information retrieval and applying the Cross-Media Relevance Model (CMRM) to the annotation task. Jeon *et al.*¹⁵ also showed that better (ranked) retrieval results could be obtained by using probabilistic annotation, rather than *hard* annotation. Lavrenko *et al.*¹⁶ used the Continuous-space Relevance Model (CRM) to build continuous probability density functions to describe the process of generating blob features. The CRM model was shown to outperform the CMRM model significantly. Metzler and Manmatha¹⁷ propose an inference network approach to link regions and their annotations; unseen images can be annotated by propagating belief through the network to the nodes representing keywords.

The models by Monay and Gatica-Perez,¹⁸ Feng *et al.*¹⁹ and Jeon and Manmatha²⁰ use rectangular regions rather than blobs. Monay and Gatica-Perez¹⁸ investigates Latent Space models of annotation using Latent Semantic Analysis and Probabilistic Latent Semantic Analysis, Feng *et al.*¹⁹ use a multiple Bernoulli distribution to model the relationship between the blocks and keywords, whilst Jeon and Manmatha²⁰ use a machine translation approach based on Maximum Entropy. Blei and Jordan²¹ describe an extension to Latent Dirichlet Allocation²² which assumes a mixture of latent factors is used to generate keywords and blob features. This approach is extended to multi-modal data in the article by Barnard *et al.*²³

Oliva and Torralba^{24,25} explored a scene oriented approach to annotation in which they showed that basic scene annotations, such as ‘buildings’ and ‘street’ could be applied using relevant low-level global filters. Hare and Lewis²⁶ showed how vector-space representations of image content, created from local descriptors of salient regions within an image,^{27–29} could be used for auto-annotation by propagating semantics from similar images. Yavlinsky *et al*³⁰ explored the possibility of using simple global features together with robust non-parametric density estimation using the technique of ‘kernel smoothing’. The results shown by Yavlinsky *et al*³⁰ were comparable with the inference network¹⁷ and CRM.¹⁶ Notably, Yavlinsky *et al* showed that the Corel data-set proposed by Duygulu *et al*¹⁴ could be annotated remarkably well by just using global colour information.

Most of the auto-annotation approaches described above perform annotations in a *hard* manner; that is, they explicitly apply some number of annotations to an image. A *hard* auto-annotator can cause problems in retrieval because it may inadvertently annotate with a similar, but wrong label; for example, labelling an image of a horse with “foal”. Jeon *et al*¹⁵ first noted that this was the case when they compared the retrieval results from a fixed-length hard annotator with a probabilistic annotator. Duygulu *et al*¹⁴ attempt to get around this problem by creating clusters of keywords with similar meaning.

Our current approach to auto-annotation³¹ is different; Instead of applying *hard* annotations, we have developed an approach in which annotation is performed implicitly in a *soft* manner. The premise behind our approach is simple; a semantic-space of documents (images) and terms (keywords) is created using a linear algebraic technique. Similar documents and/or terms within this semantic-space share similar positions within the space. For example, given sufficient training data, this allows a search for “horse” to return images of both horses and foals because the terms “horse” and “foal” share similar locations within the semantic space. The following subsections describe the approach in brief, and illustrate the performance with results using the Corel data-set proposed by Duygulu *et al*.

4.1. Building a semantic-space: Using linear algebra to associate images and terms

Latent Semantic Indexing is a technique originally developed for textual information retrieval. Berry *et al*³² described how Latent Semantic Indexing can be used for cross-language retrieval because it ignores both syntax and explicit semantics in the documents being indexed. In particular, Berry *et al* cite the work of Landauer and Littman³³ who demonstrate a system based on LSI for performing text searching on a set of French and English documents where the queries could be in either French or English (or conceivably both), and the system would return documents in both languages which corresponded to the query. The work of Landauer and Littman negates the need for explicit translations of all the English documents into French; instead, the system was trained on a set of English documents and versions of the documents translated into French, and through a process called ‘folding-in, the remaining English documents were indexed without the need for explicit translations. This idea has become known as *Cross-Language Latent Semantic Indexing* (CL-LSI).

Monay and Gatica-Perez¹⁸ attempted to use straight LSI (without ‘folding-in’) with simple cross-domain vectors for auto-annotation. They first created a training matrix of cross-domain vectors and applied LSI. By querying the left-hand subspace they were able to rank an un-annotated query document against each annotation term in order to assess likely annotations to apply to the image. Our approach, described below, is different because we do not explicitly annotate images, but rather just place them in a semantic-space which can be queried by keyword.

Our idea is based on a generalisation of CL-LSI. In general any document (be it text, image, or even video) can be described by a series of observations made about its content. We refer to each of these observations as terms. In order to create a semantic-space for searching images, we first create a ‘training’ matrix of terms and documents that describe observations about a set of annotated training images; these observations consist of low-level descriptors and observations of which keywords occur in each of the images. This training term-document matrix then has LSI applied to it. The final stage in building the semantic-space is to ‘fold-in’ the corpus of un-annotated images, using purely the visual observations. The result of this process is two matrices; one representing the coordinates of the terms in the semantic space, and the other representing the coordinates of documents in the space. Similarity of terms and documents can be assessed by calculating the angle between the respective coordinate vectors.

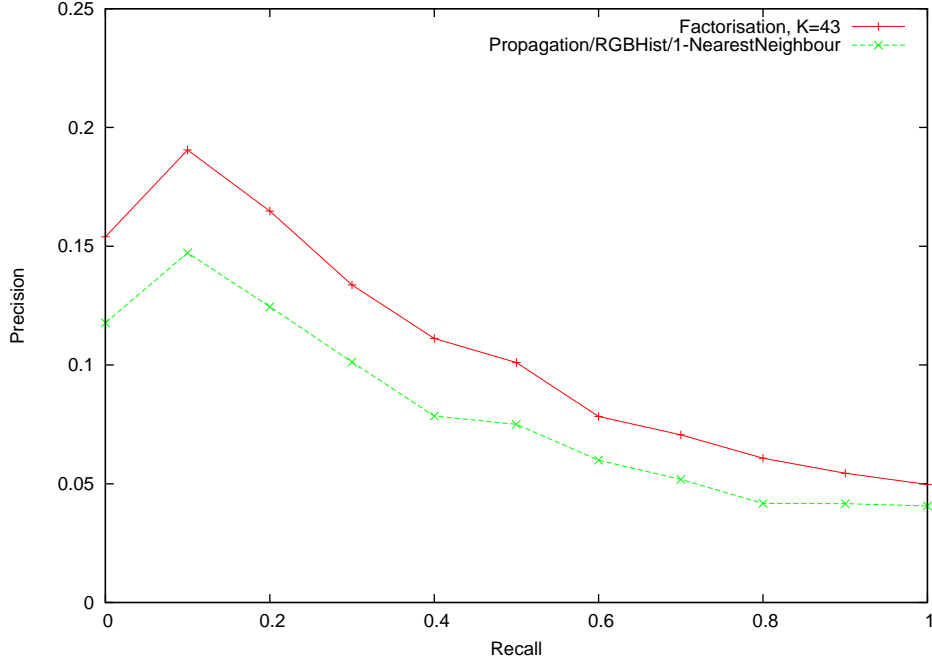


Figure 3. Average Precision-Recall plots for the Corel data-set using RGB-Histogram descriptors for both the CL-LSI and vector-space propagation algorithms.

4.2. Experiments with the Corel data-set

In order to demonstrate the approach described above, we have experimented using the training set of 4500 images and test set of 500 images described by Duygulu *et al.*¹⁴ The visual observations have been kept simple in order to demonstrate the power of the approach; each observation term is a bin from a 64-bin global RGB histogram of the image in question. Because all of the images in the data-set have ground truth annotations, it is possible to automatically assess the performance of the retrieval. By splitting the data-sets into a training set and testing set, it is possible to attempt retrieval for each of the annotation terms and mark test images as relevant if they contained the query term in their annotations. Results from using this technique are presented against results using the ‘hard’ annotations from the semantic propagation technique.²⁶

The overall average precision-recall curves of the CL-LSI and Vector-Space Propagation approaches are shown in Figure 3. As before, the CL-LSI approach outperforms the propagation approach. Whilst the overall averaged precision-recall curve doesn’t achieve a very high recall and falls off fairly rapidly, as before, this isn’t indicative of all the queries; some query terms perform much better than others. Figure 4 shows a histogram of the R-Precision for the best query terms. Figure 5 shows precision-recall curves for some queries with *good* performance.

Ideally, we would like to be able to perform a direct comparison between our CL-LSI method and the results of the statistical machine-translation model presented by Duygulu *et al.*¹⁴ which has become a benchmark against which many auto-annotation systems have been tested. Duygulu *et al* present their precision and recall values as single points for each query, based on the number of times the query term was predicted throughout the whole test set. In order to compare results it should be fair to compare the precision of the two methods at the recall given in Duygulu2002 *et al*’s results. Table 2 summarises the results over the 15 *best* queries found by Duygulu *et al*’s¹⁴ system (base results), corresponding to recall values greater than 0.4.

Table 2 shows that nine of the fifteen queries had better precision for the same value of recall with the CL-LSI algorithm. This higher precision at the same recall can be interpreted as saying that more relevant images are retrieved with the CL-LSI algorithm for the same number of images retrieved as with the machine learning approach. This result even holds for Duygulu *et al*’s slightly improved *retrained* result set. This implies, somewhat surprisingly, that even by just using the rather simple RGB Histogram to form the visual observations,

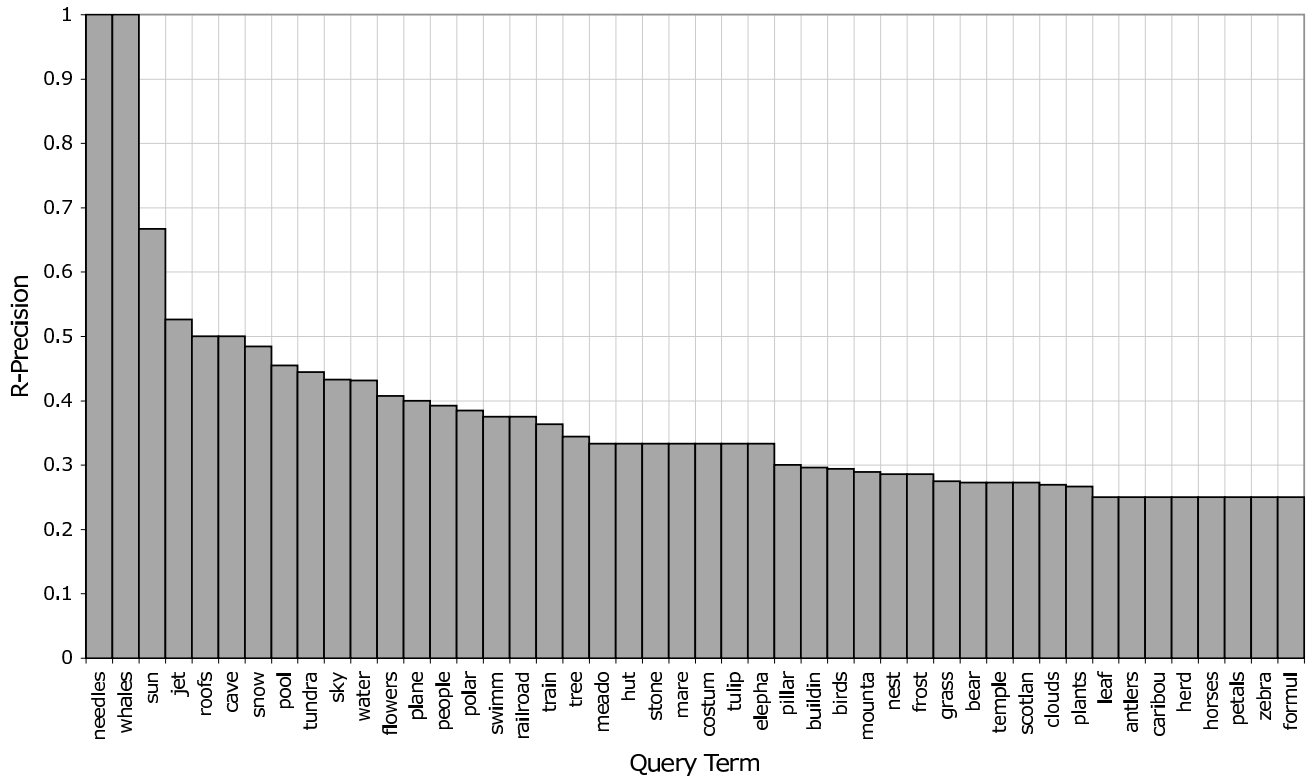


Figure 4. R-Precision of all queries with an R-Precision of 0.25 or above, in decreasing order.

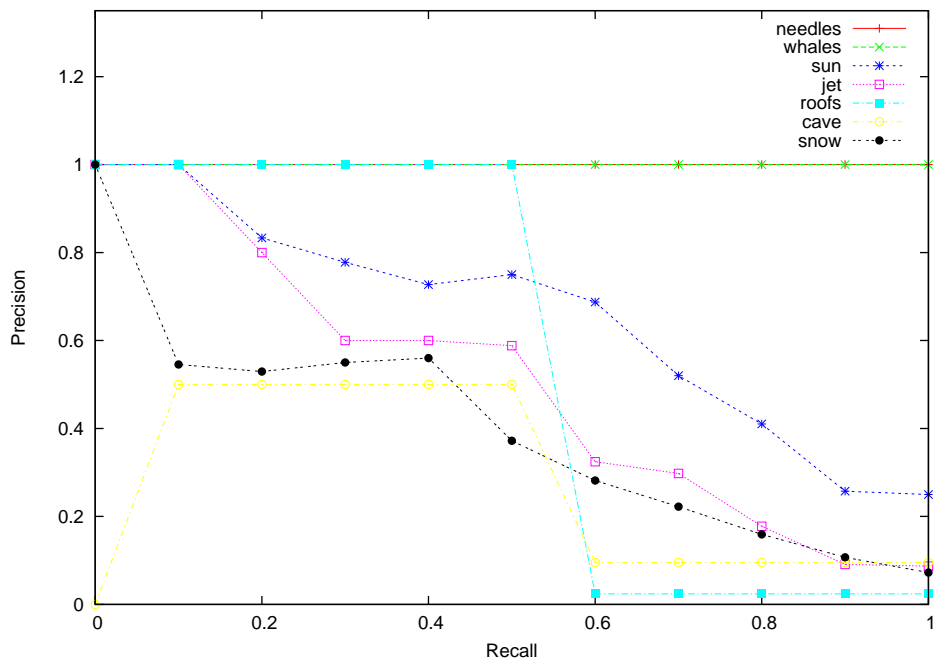


Figure 5. Precision-Recall curves for the top seven Corel queries.

| Query Word | Recall | Precision | |
|------------|--------|---|--------------------------------|
| | | Machine Translation Base Results, th=0 | CL-LSI, RGB Histogram, K=43 |
| petals | 0.50 | 1.00 | 0.13 |
| sky | 0.83 | 0.34 | 0.35 |
| flowers | 0.67 | 0.21 | 0.26 |
| horses | 0.58 | 0.27 | 0.24 |
| foals | 0.56 | 0.29 | 0.17 |
| mare | 0.78 | 0.23 | 0.19 |
| tree | 0.77 | 0.20 | 0.24 |
| people | 0.74 | 0.22 | 0.29 |
| water | 0.74 | 0.24 | 0.34 |
| sun | 0.70 | 0.28 | 0.52 |
| bear | 0.59 | 0.20 | 0.11 |
| stone | 0.48 | 0.18 | 0.22 |
| buildings | 0.48 | 0.17 | 0.25 |
| snow | 0.48 | 0.17 | 0.54 |

Table 2. Comparison of precision values for equal values of recall between Duygulu *et al.*'s machine translation model and the CL-LSI approach.

the CL-LSI approach performs better than the machine translation approach for a number, of queries. This, however does say something about the relative simplicity of the Corel dataset.³⁰ Because not all of the top performing results (c.f. Figure 4) from the CL-LSI approach are reflected in the *best* results from the machine translation approach, it follows that the CL-LSI approach may actually perform better on a majority of *good* queries compared to the machine translation model. Of course, whilst the CL-LSI approach may outperform the machine translation approach in terms of raw retrieval performance, it doesn't have the capability of applying keywords to individual segmented image regions that the translation model does.

5. ONTOLOGIES: ATTACKING THE GAP FROM ABOVE

Although automatic image annotation techniques can take us some way across the semantic gap and may enable us to reach the label representation of Section 2, above, as we have shown in Section 3, even a very full set of image labels falls far short of the richness required to represent the full semantics required to describe most images. How might such semantics be represented? The artificial intelligence community has developed many knowledge representation schemes over the years, but recently, the use of ontologies is seen as an increasingly popular way of representing high-level knowledge about application domains. Part of the reason for this increasing interest is the role which ontologies are playing in the emerging semantic web technologies aimed at making web based information understandable by software systems as well as by humans. An ontology is a *shared conceptualisation of a domain* and typically consists of a comprehensive set of concept classes, relationships between them, and instance information showing how the classes are populated in the application domain.

Once knowledge from documents is represented richly in this way several new capabilities are facilitated. First and foremost at least some of the semantics is made explicit and allows queries to be formulated in terms of concepts and their relationship. It is possible to reason over the knowledge domain via the ontology using reasoning software. The ontology can provide a platform for interoperability between systems and a versatile vehicle for browsing and navigating around the document collection.

Although most published work on the use of ontologies has been concerned with textual information, there is increasing interest and research into the use of ontologies with multimedia collections. Some early work on semantic description of images using ontologies as a tool for annotating and searching images more intelligently was described by Schreiber *et al.*³⁴ More recently his team have extended the approach³⁵ and also shown how spatial information could be included in the annotations semi-automatically.³⁶ Jaimes, Tseng and Smith described a semi-automatic approach to the construction of ontologies for semantic description of videos, using

associated text in the construction³⁷ and several authors have described efforts to move the MPEG-7 description of multimedia information closer to ontology languages such as RDF and OWL.^{38,39} Currently, the aceMedia Project⁴⁰ is developing a knowledge infrastructure for multimedia analysis, which incorporates a visual description ontology and a multimedia structure ontology.

It is useful to consider ontologies for semantic description of multimedia in two parts, one describing the multimedia content i.e. capturing knowledge about objects and their relationships in the image for example and the other part capturing wider contextual knowledge about the multimedia object, how it was formed, by whom it was created etc.

In the MIAKT project^{41,42} we integrated image annotation tools for region delineation, feature extraction and image analysis with an ontology to capture the semantics associated with the various imaging modalities associated with the breast screening process. The aim of the project was to demonstrate enhanced support at the semantic level for decision making which needs to draw on low level features and their descriptions as well as the related case notes. It also provides a platform for reasoning about new cases on the basis of the semantically integrated set of (multimedia) case histories. By contrast, in the Sculpteur project⁴³ we mapped museum multimedia object metadata (as opposed to image content) to an ontology based on the CIDOC Conceptual Reference Model in order to provide semantic level navigation and retrieval which could be combined with content based techniques which were also developed in the project.

6. CONCLUSIONS AND FUTURE WORK

In Section 3 we saw how the majority of queries by searchers are presented at the semantic level and in Section 4 we explored image annotation which attempts to bridge part of the gap from below; from the descriptors to the object labels. The use of ontologies as a way of capturing the semantics of multimedia data was explored briefly in Section 5 and if annotations (labels) can be linked automatically into ontology based representations of the semantics, a tentative bridge across the semantic gap begins to emerge. However, current descriptors are inadequate and current annotations and ontologies are far from rich. But on the positive side, multimedia retrieval research is tackling the semantic issue. Eventually approaches to annotation will be coupled with software to discover spatial and other relations between objects in images and more of the semantics will be integrated into the ontological representation automatically to provide a richer platform for the support of semantic level query mechanisms.

In the ‘Bridging the Semantic Gap’ project, funded in the UK by the Arts and Humanities Research Council, we are exploring how well test-bed ontologies, combined with content-based techniques and annotation can meet the real needs of image searchers in limited domains.

ACKNOWLEDGMENTS

The ‘Bridging the semantic gap in visual information retrieval’ project is funded by the Arts and Humanities Research Council (MRG-AN6770/APN17429), whose support together with that of our various contributors, is gratefully acknowledged.

REFERENCES

1. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), pp. 1349–1380, 2000.
2. P. G. B. Enser and C. G. McGregor, “Analysis of visual information retrieval queries,” in *British Library Research and Development Report*, (6104), p. 55, British Library, London, 1992.
3. C. J. Sandom and P. G. B. Enser, “Virami - visual information retrieval for archival moving imagery,” in *Library and Information Commission Report 129*, p. 159, *Re:source: The Council for Museums, Archives and Libraries*, 2002.
4. L. H. Armitage and P. G. B. Enser, “Analysis of user need in image archives,” *Journal of Information Sciences* **23**(4), pp. 287–299, 1997.
5. E. Panofsky, *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*, Harper & Rowe, 1962.

6. S. Shatford, "Analyzing the subject of a picture: A theoretical approach," *Cataloguing & Classification Quarterly* **5**(3), pp. 39–61, 1986.
7. C. Jörgensen, *Image Retrieval: Theory and Research : Theory and Research*, Scarecrow Press, Lanham, MA and Oxford, July 2003.
8. L. Hollink, A. T. Schreiber, B. J. Wielinga, and M. Worring, "Classification of user image descriptions," *Int. J. Hum.-Comput. Stud.* **61**(5), pp. 601–626, 2004.
9. J. Eakins and M. Graham, "Content-based image retrieval," Tech. Rep. JTAP-039, JISC, 2000.
10. P. G. B. Enser, C. J. Sandom, and P. H. Lewis, "Surveying the reality of semantic image retrieval," in *8th International Conference on Visual Information Systems, VISUAL2005*, (Amsterdam, Netherlands), July 2005.
11. P. G. B. Enser, C. J. Sandom, and P. H. Lewis, "Automatic annotation of images from the practitioner perspective.," in Leow *et al.*,⁴⁴ pp. 497–506.
12. Edina, "Education image gallery." <http://edina.ac.uk/eig>.
13. Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, 1999.
14. P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pp. 97–112, Springer-Verlag, (London, UK), 2002.
15. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 119–126, ACM Press, (New York, NY, USA), 2003.
16. V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, eds., MIT Press, Cambridge, MA, 2004.
17. D. Metzler and R. Manmatha, "An inference network approach to image retrieval.," in Enser *et al.*,⁴⁵ pp. 42–50.
18. F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pp. 275–278, ACM Press, 2003.
19. S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation.," in *CVPR (2)*, pp. 1002–1009, 2004.
20. J. Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation.," in Enser *et al.*,⁴⁵ pp. 24–32.
21. D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, ACM Press, (New York, NY, USA), 2003.
22. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.* **3**, pp. 993–1022, 2003.
23. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.* **3**, pp. 1107–1135, 2003.
24. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision* **42**(3), pp. 145–175, 2001.
25. A. Oliva and A. B. Torralba, "Scene-centered description from spatial envelope properties," in *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pp. 263–272, Springer-Verlag, (London, UK), 2002.
26. J. S. Hare and P. H. Lewis, "Saliency-based models of image content and their application to auto-annotation by semantic propagation," in *Proceedings of the Second European Semantic Web Conference (ESWC2005)*, (Heraklion, Crete), May 2005.
27. J. S. Hare and P. H. Lewis, "Salient regions for query by image content.," in Enser *et al.*,⁴⁵ pp. 317–325.

28. J. S. Hare and P. H. Lewis, "On image retrieval using salient regions with vector-spaces and latent semantics.," in Leow *et al.*,⁴⁴ pp. 540–549.
29. J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, pp. 1470–1477, October 2003.
30. A. Yavlinsky, E. Schofield, and S. Rüger, "Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation," in *Proceedings of the 4th International Conference on Image and Video Retrieval*, D. Polani, B. Browning, A. Bonarini, and K. Yoshida, eds., *Lecture Notes in Computer Science* **3568**, pp. 507–517, Springer-Verlag, (Singapore), July 2005.
31. J. S. Hare, *Saliency for Image Description and Retrieval*. PhD thesis, University of Southampton, 2005.
32. M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," Tech. Rep. UT-CS-94-270, University of Tennessee, 1994.
33. T. K. Landauer and M. L. Littman, "Fully automatic cross-language document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pp. 31–38, (UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada), October 1990.
34. A. T. G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, "Ontology-based photo annotation," *IEEE Intelligent Systems* **16**(3), pp. 66–74, 2001.
35. L. Hollink, A. T. Schreiber, B. Wielemaker, and B. Wielinga, "Semantic annotation of image collections," in *In Proceedings of the KCAP'03 Workshop on Knowledge Markup and Semantic Annotation*, (Florida, USA), October 2003.
36. L. Hollink, G. Nguyen, A. T. G. Schreiber, J. Wielemaker, B. J. Wielinga, and M. Worring, "Adding spatial semantics to image annotations," in *4th International Workshop on Knowledge Markup and Semantic Annotation at ISWC'04*, 2004.
37. A. Jaimes, B. L. Tseng, and J. R. Smith, "Modal keywords, ontologies, and reasoning for video understanding.," in *CIVR*, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. S. Zhou, eds., *Lecture Notes in Computer Science* **2728**, pp. 248–259, Springer, 2003.
38. J. Hunter, "Adding multimedia to the semantic web: Building an mpeg-7 ontology.," in *SWWS*, I. F. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, eds., pp. 261–283, 2001.
39. C. Tsinaraki, P. Polydoros, N. Moumoutzis, and S. Christodoulakis, "Coupling owl with mpeg-7 and tv-anytime for domain-specific multimedia information integration and retrieval," in *Proceedings of RIAO 2004*, (Avignon, France), April 2004.
40. I. Kompatsiaris, Y. Avrithis, P. Hobson, and M. Strinzis, "Integrating knowledge, semantics and content for user-centred intelligent media services: the acemedia project," in *Proceedings of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '04)*, (Lisboa, Portugal), April 2004.
41. B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt, "Ontology-based medical image annotation with description logics," in *Proceedings of The 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 77–82, IEEE Computer Society Press, 2003.
42. D. Dupplaw, S. Dasmahapatra, B. Hu, P. H. Lewis, and N. Shadbolt, "Multimedia Distributed Knowledge Management in MIAKT," in *Knowledge Markup and Semantic Annotation, 3rd International Semantic Web Conference*, S. Handshuh and T. Declerck, eds., pp. 81–90, (Hiroshima, Japan), 2004.
43. M. J. Addis, K. Martinez, P. H. Lewis, J. Stevenson, and F. Giorgini, "New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web," in *Proceedings of Museums and the Web 2005*, J. Trant and D. Bearman, eds., (Vancouver, Canada), 2005.
44. W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds., *Image and Video Retrieval, 4th International Conference, CIVR 2005, Singapore, July 20-22, 2005, Proceedings, Lecture Notes in Computer Science* **3568**, Springer, 2005.
45. P. G. B. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, eds., *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings, Lecture Notes in Computer Science* **3115**, Springer, 2004.