Automatic Annotation of Images from the Practitioner Perspective

Peter G.B. Enser[1], Christine J. Sandom[1], and Paul H. Lewis[2]

[1] School of Computing, Mathematical and Information Sciences, University of Brighton
{p.g.b.enser, c.sandom}@bton.ac.uk
[2] Department of Electronics and Computer Science, University of Southampton
phl@ecs.soton.ac.uk

Abstract. This paper describes an ongoing project which seeks to contribute to a wider understanding of the realities of bridging the semantic gap in visual image retrieval. A comprehensive survey of the means by which real image retrieval transactions are realised is being undertaken. An image taxonomy has been developed, in order to provide a framework within which account may be taken of the plurality of image types, user needs and forms of textual metadata. Significant limitations exhibited by current automatic annotation techniques are discussed, and a possible way forward using ontologically supported automatic content annotation is briefly considered as a potential means of mitigating these limitations.

1 Introduction

The semantic gap is now a familiar feature of the landscape in visual image retrieval [1]. Its perception as “a huge barrier in front of researchers”[2] reflects an increasingly mature realisation of the limited functionality of content-based image retrieval (CBIR) techniques in realistic commercial and curatorial scenarios of image use.

The developing interest in bridging the semantic gap is a welcome response to the criticism directed at the visual image retrieval research community by, amongst others, Jőrgensen, who has expressed concern that “the emphasis in the computer science literature has been largely on what is computationally possible, and not on discovering whether essential generic visual primitives can in fact facilitate image retrieval in ‘real-world’ applications.” [3, p.197].

Image retrieval, like information retrieval generally, is a very long-standing form of transaction to which the human searcher brings a reasoning process in order to infer semantic content. This inferential reasoning process invokes personal experience, domain knowledge, cultural conditioning and collective memory in the decoding of knowledge recorded in the image. Among the practitioner community of picture researchers, librarians and archivists the traditional paradigm of image retrieval involves textual string matching between the client’s search request statement and the indexer’s inferred semantic content annotations embedded within the image collection metadata. Common variants of this paradigm engage the practitioner with oral requests and catalogue-embedded annotations within non-digitised collections of images.

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 497 – 506, 2005.

There are circumstances where the verbalisation of need for image material is a real challenge for the searcher (notably, where the need is for images which are abstractions of reality), but in general the user's preference to express his/her need for images in natural language is well understood. The challenge posed by manual, text-based indexing of image material is equally well understood and reflects the philosophical and practical challenges of translating visually encoded knowledge into a linguistic surrogate.

The nature of these challenges has been described in a number of general treatises on visual image indexing [4–6]. The visual image is an entropic message, upon which the human viewer's physiological and intellectual capacity to detect layers of meaning confers an inherent unpredictability of *retrieval utility*. To Shatford's observation [7] that "the delight and frustration of pictorial resources is that a picture can mean different things to different people" we add the observation that a picture can mean different things to the same person at different times, under different circumstances of need or when delivered by different presentation media.

The fact that the manual indexing process is time-consuming, costly, and may demand a high level of domain knowledge; that the appropriate level of indexing exhaustivity is indeterminate, and the choice of indexing terms is conditioned by contemporary language and prey to the subjectivity of the indexer, all contribute to the perception that "the inadequacy of text description is an obvious and very problematic issue" [2].

We cannot be surprised, therefore, that the development of automatic indexing techniques is perceived to be an attractive proposition. The semantic gap towards which such techniques have tended to lead the research community has given rise to increasing interest in the integration of CBIR techniques with traditional textual metadata as a possible means of achieving 'semantic' image retrieval.

2 Automatic Annotation of Images

A number of techniques have been reported which are designed to uncover the latent correlation between low-level visual features and high-level semantics [2,8-14]. Typically such approaches involve a training set of pre-annotated images and the identification of visual features in the image such as blobs or salient objects. One popular technique extends the "traditional" latent semantic analysis (LSA) approach for text by quantising low level image descriptors, treating them as "visual terms" and adding them in to a vector space representation for the associated text. LSA then uses singular value decomposition to reduce the dimensionality of the vector space model to a lower dimensional semantic space [12]. Vectors of visual terms from an un-annotated image can then be used to locate words associated with the same regions of the semantic space to provide the annotations required.

An alternative approach tries to model directly the joint distributions of words and image features. There are various ways to do this and Barnard *et al* have described several [8]. One of these uses a technique called probabilistic latent semantic analysis, PLSA, which solves the same problem as LSA but has a more principled foundation. A comparison of PLSA and LSA approaches to image annotation is presented in [13]. From the practitioner perspective, however, these and many other annotation based approaches to bridging the semantic gap suffer from two important limitations, which are discussed below.

2.1 The Visibility Limitation

The indexing words drawn from the permitted vocabulary have to relate to *visible* entities within the image. However, studies of user need for image material, both still and moving, have revealed an important - because frequently-encountered - class of request which addresses the *significance* of a depicted object or scene [4,15-17]. Some examples of real requests obtained from these studies are shown in Figure 1, and an example of an image, the main property of which is significance, appears as Figure 2.

- WW1 - 'Cher Ami' (famous war homing pigeon)
- Prince Charles, first public engagement, as boy, aged 8 - first ever engagement
- West Ham v Bolton Wanderers - 1923 First Wembley cup final
- The first microscope
- Bannister breaking tape on 4 minute

Fig. 1. Image requests which address *significance*

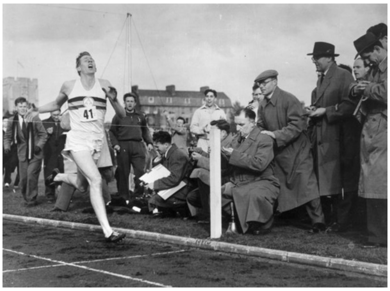

Fig. 2. A New Record
© Getty Images

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 which address significance necessarily involve the resolution of verbalised queries by matching operations conducted with textual metadata. Even if advances in automatic feature detection mitigate this constraint at some future point in time, there will have had to be a *seminal textual annotation* associated with the image to identify the significance of the depicted feature.

The issue of significance is a specific case of the more general property of *interpretatibility* of images. Figure 3 provides examples of real requests which seek a visualisation of conceptual material. This is a situation where either or both the indexer and searcher invoke an intellectualisation process which has no parallel in visual blobs or salient features. Figure 4 shows an image which the indexer has interpreted as a representation of anguish. We note in passing that there is no automatically detectable feature which enables the salient object to be interpreted as an actress.

Possibly the worst case scenario in this context occurs when image searchers specify unwanted features which must *not* be present in the retrieved image; Figure 5 shows some real-query examples. Provision is sometimes made in controlled keywor-

- Depictions of happiness
- Anguish
- Hell

Fig. 3. Image requests which seek visualisation of conceptual material

- Simon Mann before his incarceration in Zimbabwe, i.e. not in prison clothes.
- A viola d'amore, … not in performance
- The image should be of a young woman around 1870 or a little later, not too nicely dressed
- Grand prix racing, 1960-1965, Dramatic shots, but not crashing

Fig. 5. Image requests demonstrating unwanted features

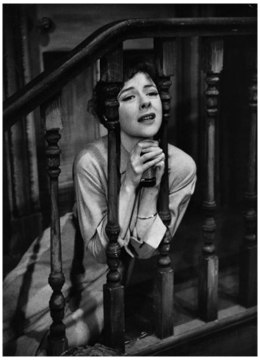

Fig. 4. Actress's Anguish
© Getty Images

ding schemes to indicate the absence of commonly visible features (e.g., 'no people', 'alone'), but this type of real-world need would seem to be at some remove from the present generation of automatic annotation techniques.

2.2 The Generic Object Limitation

Currently, experimentation with the automatic annotation of images has generally used small training sets of visible features and basic vocabularies ('sunrise', 'beach', 'horse', …). These features have been labelled 'pre-iconographic' [18], 'generic' [7] and 'perceptual' [3, p.206] by different analysts; they have the common property of visual stimuli which require a minimally-interpretive response from the viewer. However, studies of expressed need for image material have provided ample evidence that, in the context of institutional image collections, clients' requests very frequently reflect a desire to recover images of features *uniquely identified* by proper name [4, 15-17].

Once again, the resolution of requests such as those shown in Figure 6 calls for textual metadata. No matter how sophisticated automatic visual feature analysis may become in future there will, again, have to be a defining *seminal textual annotation* somewhere.

When considered from the aspect of a newly-presented, unannotated image, the application of indexing terms which provide unique identification of visible entities

- Abraham Lincoln standing – to show he was taller than others
- Ivatt Class 4MT 2-6-0 of the LMS 3000 Class (43000 under BR ownership). … the engine in a freshly-outshopped state at the Derby works.
- Churchill and Lord Halifax - walk to Parliament, March 28, 1938
- Rialto Cinema, the Strand, London

Fig. 6. Image requests which require identification

may also invest the image with the property of significance. Allocation of the annotation 'Roger Bannister' to Figure 2 illustrates the point.

3 The Bridging the Semantic Gap in Visual Information Retrieval Project

From these considerations flow the aims and objectives of a project entitled Bridging the Semantic Gap in Visual Information Retrieval, funded by the Arts and Humanities Research Council in the United Kingdom, the aims of which are:

- to develop a fully informed view of the semantic gap in visual information retrieval research, and an appreciation of approaches to bridging it.
- to create, for the benefit of the research community, a test collection of digital images which reflects the plurality of different user communities .
- to investigate the extent to which existing metadata standards enable the integration to take place.

This is a significant undertaking, but one which is specifically designed to take account of the needs and interests of both the practitioner and research communities in image retrieval. The paucity of shared perceptions and vocabulary between these communities was first noted by the late Tony Cawkell [19], but remains a serious problem today [3 p.274, 20] and must be thought detrimental to the full exploitation of visual knowledge asset management which the digital age invites.

3.1 A Taxonomy of Images

The work undertaken thus far has been framed by a still image taxonomy shown in Figure 7.

Table 1 contains the definitions which have been used in the taxonomy, together with examples of image types represented by each leaf node.

The attempt has been made to identify collections of each type of image, and, from each collection, to sample requests and the metadata associated with those images deemed relevant to each such request. One interesting observation which has arisen from this aspect of the project is the incidence of image use which does not depend on the existence of organised collections of such images. Various kinds of professional

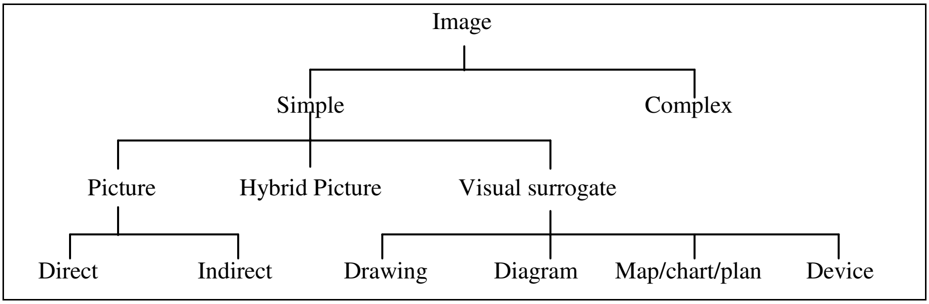

Fig. 7. A taxonomy of still images

practice use image material, where the images are recovered as adjuncts of other, uniquely identified records. Since the attempt is not being made to recover images on the basis of some attribute value within a collection of such images, however, this type of image use does not represent an image retrieval problem. Medical applications of **Indirect Picture**s is a case in point, such images most often being attached to a particular patient's record, and only ever retrieved with respect to that record. Building plans, similarly, are much more likely to be retrieved in recognition of their specific association with a uniquely identified structure, rather than in recognition of their belonging to a species of **Drawing**. In such cases an image retrieval scenario in the generally understood sense is only likely to be encountered if collections of such images are maintained for education and training purposes.

Table 1. Definitions and examples of each type of image within the taxonomy

Image	a two-dimensional visual artefact.
Simple Image	an undifferentiated image.
Complex Image	an image which comprises a set of simple images.
Picture	a scenic or otherwise integrated assembly of visual features.
Hybrid Picture	a picture with integral text; e.g. posters and other advertisements, cartoons.
Visual surrogate	non-scenic, definitional visual artefact.
Direct Picture	a picture, the features of which can be captured and/or viewed within the human visible spectrum; e.g. Photographs, works of art.
Indirect Picture	a picture, the features of which must be captured and/or viewed by means of equipment which extends viewability beyond the human visible spectrum; e.g. X-rays, ultrasound scans, MRI scans
Drawing	an accurate representation (possibly to scale) of an object; e.g. engineering and architectural drawings
Diagram	a representation of the form, function or workings of an object or process, which may be encountered in different formats and applications, and may incorporate textual or other symbolic data; e.g. anatomical diagrams, circuit diagrams
Map/chart/plan	a representation (possibly to scale) of spatial data; e.g. Geographic and geological maps, marine and astronomical charts, weather charts.
Device	a symbol or set of symbols which uniquely identifies an entity; e.g. trademark, logo, emblem, fingerprint, coat of arms.

The project has already furnished a better-informed view than that which has been available heretofore about those users of image material who seek some intervention by library/archive/curatorial staff in the resolution of their needs. Our investigation suggests a continuum of usage, with four foci corresponding with general public, 'infotainment' publishing, academic publishing and professional practice. Experience to date shows that the largest proportion of users are associated with the middle two foci, mainly because the largest number of requests come from the publishing sector, in the widest sense.

We are conscious, of course, that these observations relate to image retrieval transactions which address image collections which have no presence on the visible web.

These institutionalised collections, huge in volume, preserve a nation's visual cultural heritage, and are central to the commerce in copyrighted visual images; they represent the real business of image retrieval transactions. The informal use of web-based image resources to which access is provided by standard search engines remains outside our purview.

To date, 14 organisations have been collaborating in the project by providing records of requests and metadata. A small sample of the requests within the project test collection, segmented by class of image, is presented in Table 2.

Table 2. Examples of request, segmented by image type

Direct Picture	Bannister breaking tape on 4 minute
Indirect Picture	Human HeLa cancer cells cytokinesis
Hybrid Picture	An LMS railway poster circa 1930. Advertising New Brighton and Wallasey. Woman on high diving board
Drawing	Trevithick's tram engine, December 1803.
Diagram	The adverse health effects of space travel, specifically long periods of zero gravity … weakening of the heart
Map/Chart/Plan	Map of central London before 1940, specifically where Red Cross Street Barbican is
Device	CRESTS: Southern Railway

The subject metadata associated with example images retrieved in response to each request in Table 2 is shown in Table 3 below.

Table 3. Examples of subject metadata

Direct Picture: Bannister breaking tape on 4 minute [21]

Title	A New Record Date : 6th May 1954
Description	Roger Bannister about to cross the tape at the end of his record breaking mile run at Iffley Road, Oxford. He was the first person to run the mile in under four minutes, with a time of 3 minutes 59.4 seconds.
Subject	Sport, Personality, Feats & Achievements
Keywords	black & white, format landscape, Europe, Britain, England, clothing, sportswear, male, group, running, British, English, Roger Bannister, Athletics, Middle Distance, Mile, finish line, excitement

Indirect Picture: Human HeLa cancer cells cytokinesis [22]

Title	Cells interacting to cause immune response
Description	Immune system in action. Different cell types in the spleen interacting to cause a specific immune response.
Keywords	Immunology, B Cells, White Blood Cells, Immunisation, Cell Interactions, Cytokines, Affinity Maturation, Cell Membranes.

Hybrid Picture: An LMS railway poster circa 1930. Advertising New Brighton and Wallasey. Woman on high diving board [23]

Title	'New Brighton and Wallasey', LMS poster, 1923-1947.
Caption	London Midland & Scottish Railway poster. Artwork by Septimus E Scott.
Keywords	New Brighton; Wallasey; London Midland & Scottish Railway; swimming pools; woman; women; swimming costumes; bathing costumes; swimsuits; diving boards; beaches; crowds; tourism; holidays; resorts; summer; sea; seaside; coast; holiday-makers; tourists; leisure; Social; recreation; railway poster; railway; poster; posters; poster art; graphic design; graphics; design; advertisements; ads; advertising

Drawing: Trevithick's tram engine, December 1803 [23]

Title	Trevithick's tram engine, December 1803.
Caption	Drawing believed to have been made by John Llewellyn of Pen-y-darran. Found by FP Smith in 1862 and given by him to William Menelaus. Richard Trevithick (1771-1833) was the first to use high pressured steam to drive an engine. Until 1800, the weakness of existing boilers had restricted all engines to being atmospheric ones. Trevithick set about making a cylindrical boiler which could withstand steam at higher pressures. This new engine was well suited to driving vehicles. In 1804, Trevithick was responsible for the first successful railway locomotive.
Keywords	Trevithick, Richard; Drawings; Pen-Y-Darran; Wales; Llewellyn, John; Smith, F P; Menelaus, William; locomotives; tram engines

Diagram: the adverse health effects of space travel, specifically long periods of zero gravity … weakening of the heart [22]

Title	Heart block
Description:	Heart block Colour artwork of cut-away heart, showing right and left ventricles with diagrammatic representation of a right bundle block, usually caused by strain on the right ventricle as in pulmonary hypertension
ICD code	426.9

Map/Chart/Plan: Map of Central London pre 1940, specifically where Red Cross Street Barbican is [24].

Title	Stanfords Library Map of London and its suburbs/ Edward Stanford, 6 Charing Cross Road
Notes	Extent: Crouch End – Canning Town – Mitcham – Hammersmith. Title in t. border. Imprint and scale in b. border. Hungerford and Lambeth bridges shown as intended. Exhibition buildings shown in Kensington.

Device: CRESTS: London, Brighton and South Coast Railway [23]

Title	Coat of arms of the Southern Railway on a hexagonal panel, 1823-1947.
Caption	The coat of arms of the Southern Railway features a dragon and a horse on either side of a shield.
Keywords	SR; Southern Railways; horses; dragons; shields; coat of arms; railways; railway coat of arms

4 Conclusion

The richness of the manual annotations shown in Table 3 clearly indicates the necessity of enhancing the functionality of current automatic annotation techniques if there is to be any possibility of the semantic gap being bridged in real-world applications.

One approach which shows promise in this regard employs the sharable ontology concept of the semantic web. Rather than just providing associations between image features and semantic labels, an appropriate ontology can make explicit the relationships between the labels and concepts with which they are associated. Several investigators are now exploring the idea of using ontologies for enriched media description [25-30]. An example application is the SCULPTEUR project [28,29], in which integrated content, metadata and concept based image retrieval facilities have been developed for a number of major European museums using an ontology to expose the knowledge in the multimedia collections.

Implementation of ontologically-supported content annotation represents a considerable challenge in the elicitation and representation of domain knowledge. Nevertheless, the authors perceive ontologically annotated image sets to be a means by which possibilities may be tested for enhanced image retrieval performance.

Central to that endeavour is our perception that, typically, experimentation in image indexing and retrieval has taken a highly selective view of the community of users of image collections, and that future work needs to be much better informed about the nature of information need in the visual realm. To this end, a detailed survey of the image retrieval landscape is underway, framed by a taxonomy which seeks to represent the plurality of image types, user needs and forms of textual metadata by which real image retrieval transactions are realised.

Acknowledgements

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. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (12), (2000) 1349-1380
2. Zhao, R. & Grosky, W.I.: Bridging the semantic gap in image retrieval. In: Shih, T.K.(Ed.) Distributed multimedia databases: techniques & applications. Idea Group Publishing, Hershey, PA (2002) 14-36
3. Jőrgensen, C.: Image retrieval: theory and research. The Scarecrow Press, Lanham, MA and Oxford (2003)
4. Enser, P.G.B.: Pictorial information retrieval. (Progress in Documentation). Journal of Documentation 51(2), (1995) 126-170
5. Rasmussen, E.M.: Indexing images. In: Williams, M.E. (ed.), Annual Review of Information Science 32. Information Today (ASIS), Information Today, Medford, New Jersey (1997) 169-196
6. Sandore, B. (ed.): Progress in visual information access and retrieval. Library Trends, 48(2) (1999) 283-524
7. Shatford, S.: Analysing the subject of a picture; a theoretical approach. Cataloging & Classification Quarterly 6(3) (1986) 39-62
8. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., Jordan, M. I.: Matching Words and Pictures. Journal of Machine Learning Research 3(6) 1107-1135
9. Jeon, J., Lavrenko, V., & Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM Press, New York, NY (2003) 119-126 <http://ciir.cs.umass.edu/pubfiles/mm-41.pdf>
10. Fan, J., Hangzai Luo, Y.G. & Xu, G.: Automatic image annotation by using concept-sensitive salient objects for image content representation. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM Press, New York, NY (2004) 361-368

11. Lavrenko, V., Manmatha, R., & Jeon, J.: A model for learning the semantics of pictures. In: Seventeenth Annual Conference on Neural Information Processing Systems (2003)
12. Zhao, R. & Grosky, W.I.: From Features to Semantics: Some Preliminary Results, IEEE International Conference on Multimedia and Expo, New York, New York, (2000) <http://www.cs.sunysb.edu/~rzhao/publications/ICME00.pdf>
13. Monay, F. & Gatica-Perez, D.: On image auto-annotation with latent space models. ACM Multimedia (2003) 275-278
14. Kosinov, S. & Marchand-Maillet, S.: Hierarchical ensemble learning for multimedia categorisation and autoannotation. In: Proceedings IEEE Machine Learning for Signal Processing workshop (MLSP), Sao Luis, Brazil (2004)
15. Enser, P.G.B.: Query Analysis in a Visual Information Retrieval Context. Journal of Document and Text Management, 1(1) (1993) 25-52
16. Armitage, L.H, and Enser, P.G.B.: Analysis of user need in image archives. Journal of Information Science 23(4) (1997) 287-299
17. Enser, P. & Sandom, C.: Retrieval of Archival Moving Imagery - CBIR Outside the Frame? In: Lew, M.S.; Sebe, N., Eakins, J.P. (eds.): Image And Video Retrieval. International Conference, CIVR 2002, London, UK (2002) Proceedings. Springer, Berlin, (2002) 202-214
18. Panofsky, E.: Meaning in the visual arts. Doubleday Anchor Books, Garden City, NY (1955)
19. Cawkell, A.E.: Selected aspects of image processing and management: review and future prospects. Journal of Information Science 18(3) (1992) 179-192
20. Enser, P.: Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. Journal of Information Science 26(4) (2000) 199-210
21. Edina: Education Image Gallery <http://edina.ac.uk/eig/>
22. Wellcome Trust: Medical Photographic Library <http://medphoto.wellcome.ac.uk>
23. Science & Society Picture Library.< http://www.scienceandsociety.co.uk>
24. Corporation of London: Talisweb. <http://librarycatalogue.cityoflondon.gov.uk:8001/>
25. Town, C. & Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. Image and Vision Computing 22(3) (2003) 251-267
26. Jaimes, A. & Smith, J.R.: Semi-automatic, Data-driven Construction of Multimedia Ontologies. In: Proceedings of the IEEE International Conference on Multimedia and Expo (2003) <http://mia.ece.uic.edu/~papers/MediaBot/pdf00002.pdf>
27. Hollink, L., Schreiber, A. Th., Wielemaker, J. & Wielinga, B.: Semantic Annotation of Image Collections. In: Proceedings of the KCAP'03 Workshop on Knowledge Capture and Semantic Annotation, Florida, (2003) < http://www.cs.vu.nl/~guus/papers/Hollink03b.pdf>
28. Goodall, S., Lewis, P.H., Martinez, K., Sinclair, P.A.S., Giorgini, F., Addis, M.J. Laharnier, C., & Stevenson, J.: Knowledge-based exploration of multimedia museum collections. In: Proceedings of the European workshop on the integration of knowledge semantics and digital media technology, London, (2004) 415-422
29. Addis, M., Boniface, M., Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K. & Stevenson, A.: SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling. In: Proceedings of the International Semantic Web conference. (ISWC 2003) (Lecture Notes in Computer Science Vol. 2870) Springer (2003) 582 -596
30. Hu, B., Dasmahapatra, S., Lewis, P. & Shadbolt, N.: Ontology-based Medical Image Annotation with Description Logics. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (in press), Sacramento, CA, USA. 2003