OntoImage 2006

Workshop on Language Resources for Content-based Image Retrieval
during LREC 2006

Final Programme

Monday 22 May 2006
Magazzini del Cotone Conference Center,
GENOA - ITALY

14:30 – 14:45 Introduction (G Grefenstette)

14:45 – 16:30 First Oral session (20mn per speaker + 5mn for questions)
14:45 – 15:10 Allan Hanbury
Analysis of Keywords used in Image Understanding Tasks
15:10 – 15:35 Katerina Pastra
Image-Language Association: are we looking at the right features?
15:35 – 16:00 Christophe Millet, Gregory Grefenstette, Isabelle Bloch, Pierre-Alain Moellic, Patrick Hede
Automatically populating an image ontology and semantic color filtering
16:00 – 16:25 Mark Sanderson, Jian Tian, Paul Clough
Testing an automatic organisation of retrieved images into a hierarchy

16:30 – 17:00 Tea / Coffee break

17:00 – 18:40 Second Oral session (20mn per speaker + 5mn for questions)
17:00 – 17:25 Thierry Declerck, Manuel Alcantara
Semantic Analysis of Text Regions Surrounding Images in Web Documents
17:25 – 17:50 Diego Burgos, Leo Wanner
Using CBIR for Multilingual Terminology Glossary Compilation and Cross-Language Image Indexing
17:50 – 18:15 Michael Grubinger, Paul Clough, Henning Müller, Thomas Deselaers
The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems
18:15 – 18:40 Judith L. Klavans
CLiMB: Computational Linguistics for Metadata Building

18:40 – 19:00 Closing discussion (G Grefenstette)

OntoImage 2006

Workshop on Language Resources for Content-based Image Retrieval
during LREC 2006

Monday 22 May 2006
Magazzini del Cotone Conference Center,
GENOA - ITALY

Workshop Organisers and Program Committee

Gregory Grefenstette, CEA LIST, France
Mark Sanderson, University of Sheffield, UK
Françoise Preteux, INT, FRANCE

OntoImage 2006
Workshop on Language Resources for Content-based Image Retrieval
during LREC 2006

Monday 22 May 2006
Magazzini del Cotone Conference Center,
GENOA - ITALY

Table of Contents

Using CBIR for Multilingual Terminological Glossary Compilation and Cross-Language Image Indexing

Diego Burgos[1], **Leo Wanner**[2]

[1]Iulaterm Group
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra
La Rambla 30-32
08002 Barcelona
diego.burgos@upf.edu

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA)
Departament de Tecnologia, Universitat Pompeu Fabra
Passeig de Circumval·lació, 8
08003 Barcelona
leo.wanner@upf.edu

Abstract

In this paper, several strategies for cross-language image indexing and terminological glossary compilation are presented. The process starts form a source language indexed image. CBIR is proposed as a means to find similar images in target language documents in the web. The text surrounding the target matched image is chunked and the chunks are classified into concrete and abstract nouns by means of a discriminant analysis. The number of images retrieved by each chunk and the edit distance between each chunk and each image file name are taken as differentiating variables; a 74.4% rate of correctly classified labeled examples shows the adequacy of these variables. Nouns classified as concrete are used to retrieve images from the web and each retrieved image is compared with the image in the target document. When a positive matching occurs, the chunk used to retrieve the matched image is assigned as the index for the image in the target document and as the target language equivalent for the source image index. As the experiments are carried out in specialized domains, a systematic and recursive use of the approach is used to build terminological glossaries by storing images with their respective cross-language indices.

1. Introduction

Images (and, therefore, also Content-Based Image Retrieval, CBIR) play a primary role in specialized discourse. However, for an integral application of CBIR, comprehensive indexed image DBs and, as a consequence, comprehensive lists of suitable index terms are required. The availability of such lists and the availability of the material to index are language dependent. For instance, for English, considerably more resources are available than for Spanish. A study carried out by Burgos (forthcoming) with bilingual Spanish-English terminological dictionaries revealed that the average of retrieved Spanish documents per term from the web was dramatically lower (7,860) than the average of retrieved English documents (246,575). Obviously, one explanation is that the web search space for English is much larger than the search space for Spanish. However, another explanation is that Spanish terms found in traditional terminological dictionaries are not suitable for indexing since they occur with a low frequency in the corpus. More suitable index terms must be looked for!

In the present work, CBIR is proposed as a means for multilingual terminology retrieval from the web for the purpose of compiling a multilingual glossary and building up an image index. All experiments are done so far for English and Spanish.

2. Related Research

One of the major goals of CBIR is image indexing which aims at providing images with indices that describe objects clearly differentiated in the images; cf., for instance, parts of an engine. Some relevant work in this area has been done with respect to the segmentation of image regions that roughly correspond to objects (Carson et al., 2002; Barnard et al., 2003). Segmentation helps reducing the *semantic gap* (Chen et al., 2003; Tsai, 2003). Approaches that apply image retrieval directly to the web (Chang et al., 1997; Chen et al., 1999; Shen et al., 2000) are especially interesting to us since the present study is also carried out for the web. Moreover, these approaches propose HTML code as anchor to capture the semantics of images that could be used to build additional variables to improve the performance of the classification method proposed here. Indexing strategies have been mainly applied to general image collections. Yeh et al. (2004) report on a proposal to retrieve images of tourist sites from the web with a mobile phone whose end goal is similar to ours. The use of terminology for indexing specialized domain images in a bilingual or multilingual setting has not been discussed in previous literature.

3. BC Hypothesis

We assume language independent bimodal co-occurrence (BC) of images and their index terms in the corpus. This implies that (i) if a well chosen image index term occurs in a document of our corpus, it is likely that the corresponding image will also be available in the same document, and, vice versa, (ii) if an image occurs in a document of the corpus, the corresponding index term will also occur; see Figure 1.

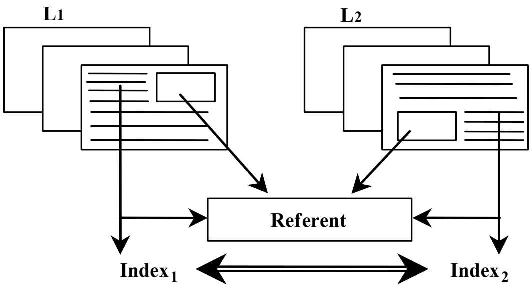


Figure 1: Representation of the BC-hypothesis

Preliminary empirical studies (carried out initially for English) buttress this BC-assumption. A total of 20 terms designating concrete entities by noun phrases[1] from the automotive engineering field were extracted from a recent issue of the *Automotive Engineering International Online*[2] journal's Tech Briefs section and used to retrieve documents from the web. The 20 terms to be included in this first sample were multiword expressions (MWE), basically noun phrases (NPs) of at least two tokens, whose referents were physical objects, i.e., spare parts or concrete devices belonging or related to the automotive engineering domain. The concrete nature of the terms' referents was confirmed by the definition or by the target language equivalent of each NP provided by a specialized dictionary. When the complete NP was not documented, the last modifier was removed and the remaining NP was searched again, and so on until it was found in the dictionary. For example, *supercharger drive pulley* was not found *as is* in the Routledge English Technical Dictionary, but *drive pulley* was. The intuition and knowledge of a Spanish native speaker (as in the case of the first author) was considered enough to determine that *polea conductora* is an object. When the intuition did not suffice, the definition prevailed over the equivalent:

Pulley: *A wheel-shaped, belt-driven device used to drive engine accessories*[3].

The BC-hypothesis was confirmed for 19 terms in 223 visited web sites, i.e., each of the 19 terms co-occurred in a document along with its respective image. The remaining term – *volume production engine* – did not confirm our hypothesis since it did not designate a concrete entity, as it initially seemed, but a general concept referring to the mass production of an engine. Certainly, the head noun *engine* would have confirmed the BC, but as its premodification makes it that general, it did not co-occur with any image. It was also observed that certain nouns that designate a group of constituents must be excluded from the study – although they could be considered concrete nouns. The entities designated by words such as *engine* or *system* tend to be so general that their boundaries often cannot be clearly determined or that their appearance cannot be accurately predicted.

Furthermore, in a bilingual setting, we assume that if in the source language corpus, an image of an object is available along with its index term, an image of the same object along with a term that denotes it (and that can thus serve as its index term) will be available as well in the target language. That is, in order to identify in the target language corpus the equivalent translation term of the source index term, we must (1) recognize that the two images represent the same object; (2) retrieve the term denoting this object in the target language corpus; see again Figure 1.

In order to prove the bilingual BC-hypothesis, a number of comparable (i.e., from the same domain) English and Spanish web sites on the automotive engineering field were collected and index terms and images were manually matched. Table 1 shows an example of two manually matched images taken from two different language websites which also serve to illustrate how cross-language equivalences between index terms can be established.

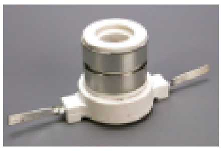
	Source (English)	**Target (Spanish)**
Image		
Index	Slip-Ring FD 3G 26.9 mm	Colector Ford 3G

Table 1. BC-hypothesis for indexing in a bilingual setting.

We prototypically implemented the above proposal and ran some preliminary experiments described below.

4. CBIR-Based Image indexing

For CBIR-based image indexing, we start from a source language indexed image. An internet segment in the target language is delimited as a corpus (= search space) and the images in this corpus are compared with the source language image using Imatch[4], a commercial software package with an embedded CBIR-module. When a positive image matching occurs according to a given threshold, the target language document containing the matched image is marked as a potential target index term location. Given that more noise results from a large search space, the size of the image database is usually one of the major concerns in CBIR-applications. In our work, we observed that the first problem to tackle is the appropriate definition of the web segment that will constitute the search space. The image DB-size and -quality will depend on this definition. Uniformity is more likely, for example, within the photographs of the same site than between the images of two or more sites. Likewise, there will be greater variance of image characteristics between the images of two different domains than within the images of the same domain, and so on.

Our proposal relies to a great extent upon the performance of CBIR-techniques for image matching automation. As a CBIR-module has not yet been developed for this study, yet, the CBIR-module settings of Imatch had to be adjusted in order to obtain good results. The most complex Imatch algorithm performs image matching based on color, texture and shape information contained in images. Current results were achieved using this algorithm. The observations made so far with respect to

[1] See (Quirk et al., 1985: 247) or (Bosque, 1999: 8-28, 45-51) with respect to the interpretation of the concept 'concrete noun'.

[2] Cf. http://www.sae.org/automag/, state January, 2006.

[3] Definition taken from http://www.autoglossary.com/.

[4] An evaluation version can be downloaded from http://www.photools.com/.

matching of images on the web suggest that other alternatives of CBIR must be considered, but that some positive matches in rather homogeneous search spaces provided enough target index term locations to pursue index candidate selection.

4.1. Index Candidate Selection

Once the indexing context (monolingual or bilingual) has been determined and the document has been located, the index candidate selection is carried out, ignoring abstract nouns from the text surrounding the image. Certainly, there are some ideal web layouts where the unique surrounding text within reasonable boundaries is the image's object name, that is, the index. In this case, a rather simple algorithm could extract the index. However, often considerable amounts of text must be parsed and concrete and abstract nouns must be disambiguated. In our study, this issue is being addressed as a classification problem where a set of NPs must be classified as concrete or abstract[5]. NPs classified as concrete make up the list of potential indices for the relevant image from which an index will be chosen by the index-image alignment process described below. The process of index candidate selection from the surrounding text consists of four phases: 1) Surrounding text chunking, 2) Chunks' cleaning, 3) Definition of variables for classification, and 4) Classification.

To distinguish NPs from VPs and other phrases, a chunker is used. Once all NPs have been chunked and extracted, some cleaning is done in order to prevent problems in the next phase of variable definition. The cleaning consists, first of all, in removing determiners at the beginning of the phrase; lemmatization (if appropriate); discarding NPs whose head noun (HN) is an acronym[6]; splitting Saxon possessives, and deleting proper nouns and numbers. Consider an example:

three development objectives ⇒ development objective
FSE's single direct injector ⇒ single direct injector

Obviously, some of the elements removed in the cleaning phase could be important for other purposes. However, for image indexing, their removal proved to be beneficiary.

Since concrete nouns do not present significant syntactic differences in comparison with abstract nouns, it is difficult to find linguistic variables that would be discriminatory enough to distinguish both types in the output provided by the chunker. Two alternative variables were analyzed: a) the number of images retrieved from the web by each NP, and b) the edit distance between the NP and the image file name (see below). It would be of great relevance to know whether the fact of being concrete or abstract could statistically differentiate the association or proximity of an NP to images in an index like maintained by Google. The first evaluations showed that sometimes even concrete nouns retrieved very *general* images and abstract nouns retrieved a good number of images too! As a consequence, a second variable was measured with the underlying assumption that if an image surrounding text does not contain the NP that led to the retrieval of the image, it is the image file name that should more closely designate the image's object, and, therefore, serve as and indicator of a concrete noun – provided that this name is not a simple number. Then, if the image file name matches the NP, the latter increases the probability of designating a concrete entity.

For the statistical analysis, 100 concrete nouns and 100 abstract nouns were selected according to the criteria mentioned in Section 3. For each of the selected NPs, one modifier was left in order to (i) avoid outliers in the values of the retrieved image frequency, (ii) assure a minimum of domain specificity in the image search and (iii) be coherent with the assumed average length of image file names. Thus, in the case of, e.g., the NP *powder-metal connecting rod*, instead of searching for images with the full NP (which would lead to the retrieval of 5 images), the search is performed with the shortened NP *connecting rod* (i.e., the first modifier *powder-metal* is removed). This leads to the retrieval of 5,940 images, instead of the retrieval of 923,000 images with the head of the NP, *rod*.

To measure the string distance between an NP and an image file name, the Levenshtein edit distance was used. The edit distance can be described as the minimum number of steps (substitutions, insertions or deletions) necessary to convert a word into another. The edit distance is 1 when there are transformations and 0 when no transformations are necessary. To analyze continuous values for this variable, the relative edit distance[7] was used to obtain values between 0 and 1. Negative values are assigned when the image file name is longer than the NP.

NP	Image file name	Edit distance
rear axle	rear axle	0
ignition coil	sparky	1
rear axle	stanley rear axle	-0.470588235
throttle valve	throttle	0
oil pan	oilpan	0.166666667
selector lever	image	0.8

Table 2. Some examples of the relative edit distance.

Table 2 shows some examples of the relative edit distance for some specific cases. Image file names were also cleaned so that underscores, numbers or symbols did not interfere in the measurement. As it can be noticed, spaces also count. If the file name is a substring of the NP, it is marked as a positive matching; if the file name contains at least one of the NP's characters, a positive score, although not the lowest, is also given. Each NP was compared with a maximum of 20 image names; a relative distance mean was established for each NP.

The tests of equality of group means proved a significant difference between the two measured variables, that is, image frequency and relative edit distance. 74.4% of originally grouped cases were correctly classified. A detailed analysis of the results shows that there is bigger variance within the values of concrete nouns than within abstract nouns. This suggests that another variable, may

[5] The experiments in this stage so far have been done for English.

[6] NPs with acronyms as HN are not included at this stage of the work since often do not reveal whether they designate concrete or abstract entities – which could hinder further validation.

[7] RD = number of transformation steps / possible maximum transformations.

be a linguistic one, might help improving the percentage of correctly classified cases of concrete nouns.

When a concrete noun is detected, it recovers the modifiers that had been removed for the phase of image retrieval and the complete NP is used in the index-image alignment stage.

4.2. Index-Image Alignment

In the previous section, a rather simplistic strategy was described to detect concrete nouns in the text surrounding an image in order to use them as index candidates for the image. The indexing process can be simplified if the image file name matches with any of the detected concrete nouns. For cases where such matching does not occur, the following procedure is proposed.

For target image indexing, i.e., image-index association, each NP classified as concrete is used to query Google for images. Each of the 20 first retrieved images is compared with the image to be indexed. When a positive image matching occurs, the original image is indexed with the NP that was used to retrieve from the web the image that yielded the positive image matching. Table 3 illustrates this procedure by an example. In the example, the images retrieved by *steering wheel* and *air filter* did not match with the original image, but one of the images retrieved by *cylinder head* did. Therefore the original image is indexed as *cylinder head*.

NP	**Google Images**	**Original image**	**Matching (+/-)**	**New index**
steering wheel	☼ → ☼ →	⇮	– –	–
cylinder head	◙ → ⇮ → OBJ →	⇮	– + –	⇮ cylinder head
air filter	◈ → ◈ →	⇮	– –	–

Table 3. Monolingual image-index alignment procedure.

The technique shows that image indices can be assigned taking into account usage, specificity and geographical variants. The fact of indexing the image with a term retrieved from its context assures that the index term is being used. Moreover, this technique tries to retrieve the appropriate degree of specificity that the index of a specific domain image is expected to present – which is often determined by the number of HN-modifiers of MWEs. Likewise, even for specialized discourse, indices should respond to geographical variants. This aspect can be controlled by specifying country domains.

5. Future Work

Given that not all process stages of the proposal presented in this paper have been completely integrated and automated, an overall evaluation has not been possible so far. Future work aims at implementing specific CBIR algorithms to be applied in specialized domains and integrated in modules for index candidate selection and index-image alignment. The goal is to be able to compile multilingual specialized glossaries after systematic and recursive exploration of well delimited web segments and storage of images with their respective cross-language indices. Likewise, some other variables to improve discrimination between concrete and abstract nouns will be researched. Even if linguistic specific features are hard to find in both groups, they are not completely discarded. Finally, further experiments will be carried out with other domains than automotive engineering.

6. Acknowledgements

This study is part of a wider research work being carried out by the first author within the framework of his PhD thesis at the IULA, Universitat Pompeu Fabra, Barcelona, Spain. It is supported by a grant from the Government of Catalonia according to resolution UNI/772/2003 of the Departament d'Universitats, Recerca i Societat de la Informació dated March 10th, 2003.

7. References

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M. (2001). Matching Words and Pictures. *Journal of Machine Learning Research*, 3, pp. 1107 - 1135.

Bosque, I. (1999). El nombre común. In Bosque, I., Demonte, V. (eds) *Gramática descriptiva de la lengua castellana*. Madrid: Espasa Calpe, pp. 3-75.

Burgos, D. (forthcoming). Concept and Usage-Based Approach for Highly Specialized Technical Term Translation. In Gotti, M., Sarcevic, S. (eds) Forthcoming 2006. *Insights into Specialized Translation*. Bern: Peter Lang.

Carson, C., Belongie, S., Greenspan, H., Malik, J. (2002). Blobworld: Image Segmentation Using Expectation-Maximisation and its Application to Image Querying. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(8), pp. 1026-1038.

Chang, S., Smith, J. R., Beigi, M., Benitez, A. (1997). Visual Information Retrieval from Large Distributed Online Repositories. *Communications of the ACM* 40(12). 63-71.

Chen, F., Gargi, U., Niles, L., Schutze, H. (1999). Multi-Modal Browsing of Images in Web Documents. *Document Recognition and Retrieval VI, Proceedings of SPIE* 3651, pp. 122-133.

Chen, Y., Wang, J. Krovetz, R. (2003). CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning. *IEEE Transactions on Image Processing*, Vol. 14 (8) pp. 1187-1201.

Quirk, R., Greenbaum, S., Leech, G. Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Routledge English Technical Dictionary. Copenhaguen: Routledge. 1998.

Shen H.T., Ooi B.C., Tan K.L. (2000). Giving Meanings to WWW Images. In: *Proceedings of the 8th ACM international conference on multimedia*, 30 October - 3 November 2000, Los Angeles, pp 39-48

Tsai, C. (2003). Stacked Generalisation: a Novel Solution to Bridge the Semantic Gap for Content-Based Image Retrieval. *Online Information Review*, Vol. 27 (6), pp. 442-445.

Yeh, T., Tollmar, K., Darrell, T. (2004). Searching the Web with Mobile Images for Location Recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'04), Vol. 2, pp. 76-81.

Semantic Analysis of Text Regions Surrounding Images in Web Documents

Thierry Declerck, Manuel Alcantara

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg,3
D-66123 Saarbrücken, Germany
declerck@dfki.de

Abstract

In this paper we present some on-going work and ideas on how to relate text-based semantics to images in web documents. We suggest the use of different levels of Natural Language Processing (NLP) to textual documents and speech transcripts associated to images for providing structured linguistic information that can be merged with available domain knowledge in order to generate additional semantic metadata for the images. An issue to be specifically addressed in the next future concerns the automation of the detection of relevant text/speech transcripts for a certain image (or video sequence). Beyond the time code approach, with its shortcomings, we expect from the discussion in this workshop on lexical characteristics of the language that can or should be used to describe image content an improvement of the approaches we are dealing with for the time being.

1. Introduction

We started our work within a past European project, Esperonto. The Esperonto project was dealing with annotation services for bridging the gap between the actual (html based) Web and the emerging Semantic Web. A smaller task of the project was dedicated to the investigation on how to automatically provide for semantic annotation for images present in a web page. A possible strategy we investigated was to provide for ontology-driven semantic annotation of the text surrounding an image in a web page.

This work is being continued and extended within a recently started Network of Excellence, called K-Space (Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content, http://kspace.qmul.net/), in which some Labs are specifically dedicated to the contribution of Human Language Technology (HLT) for the semantic indexing (and possibly retrieval) of multimedia content. K-Space, which will be described in more details in this paper, is offering a more integrated approach for multimedia semantics, aiming at a formal integration of low-level features extracted from multimedia material on the base of state-of-the-art audio-video analysis, and high-level features resulting from text analysis coupled with semantic web technologies.

2. Background: Multimedia Semantics

The topic of Multimedia Semantics has gained a lot of interest in recent years, and large funding agencies issued calls for R&D proposals on those topics. So for example a recent call of the European Commission, the 4th call of the 6th Framework, was dedicated to the merging of results gained from R&D projects on knowledge representation and cross-media content. The goal being in making the (semantic) descriptions of multimedia content re-usable on the base of a higher interoperability of media resources, which has been so far described mainly at the level of XML syntax, as can be seen with the MPEG-7 standard for encoding and describing multimedia content.

In the line of the recent developments in the fields of Semantic Web technologies, one approach consists in looking at ways for encoding so-called low-level features, as they can be extracted from audio-video material, into a high-level features organization as one can typically find in a (domain) ontology.

The EU co-funded project aceMedia is offering a very good example of such an approach. In this project, ontologies, which are typically describing knowledge as expressed in words, are extended in order to include the low-level visual features resulting from state-of-the-art audio-video analysis systems. For the description of low-level features, the project uses as its background the MPEG-7 standard, and proposes links from the MPEG-7 descriptors to high-level (domain) ontologies (see Athanasiadis, 2005). So in a sense no full integration is proposed here, but a linkage between the MPEG-7 description scheme and ontologies represented in a Semantic Web language, and interoperability of descriptions of audio-video material is indirectly realized.

Another closely related approach (see the papers by Jane Hunter) is proposing a reformulation of the semantic metadata of MPEG-7 descriptors in machine-understandable language (MPEG-7 Description Schema being only a machine-*readable* language) and use RDFs or OWL. This step is ensuring a better interoperability of semantic multimedia descriptions. But here the cross-media aspect is missing, since no textual analysis and/or speech transcripts are taken into account.

A new iniative, the K-Space European Network of Excellence, has started recently. This project is dealing with semantic inferences for semi-automatic annotation and retrieval of multimedia content. The aim is to narrow the gap between content descriptors that can be computed automatically by current machines and algorithms, and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media: the so-called *Semantic Gap*.

The project deals with a real integration of knowledge structures in ontologies and low-level descriptors for audio-video content, taking also into account knowledge that can be extracted from sources that are complementary to the audio/video stream, mainly speech transcripts and text surrounding images or textual metadata describing a video or images. The integration takes place at 2 levels: the level of knowledge representation, where features associated with various modalities (image, text/speech transcripts, audio) should be interrelated within conceptual classes in ontololgies (from domain-specific to general purpose ontologies), and the level of processing, where high-level semantic features should be integrating for guiding (and so possibly improving) the automatic analysis of audio-video material and the corresponding extraction of semantic features.

As such the K-Space activities are mostly dedicated to the analysis of multimedia and cross-media data and the feature extraction out of such data. Navigation, search and retrieval in the field of semantic cross-media archives are not primarily addressed.

An interesting project with respect to K-Space is MESH, which seems to build an application scenario on the base of the multimedia and cross-media knowledge structures discussed and proposed by K-Space and aceMedia. The domain of application is given by the News domain. The project will deal with the ontology-driven semantic integration of content features extracted from video, images, speech transcripts and text. Multi- and cross-media reasoning is an important issue here, insuring consistency and non-redundancy of the integrated cross-media features. A major issue will consist in proposing an appropriate syndication of the semantically encoded material for distribution to distinct (mobile) end-user hardware, also under consideration of personalization aspects. Supporting thus the distribution of relevant multi- and cross-media content.

The 2006 edition of TRECVid is offering an interesting development, since one of its tasks is addressing searching within a multimedia database, whereas interaction with the user is also foreseen. We can expect here that the user will input his/her queries in natural language, whereas the use of certain lexical items should guide he intelligent search in large archives containing cross-media material.

3. An integrative approach in the K-Space Network of Excellence

The projects mentioned above (and some others, not listed here for reason of place), are given us important information about methodologies and technologies for the “ontologization” of low-level audio-video features extracted from multimedia content. Here we describe in some more details the K-Space project and the activities related to the use and analysis of sources complementary to audio-video material. First we describe the foreseen ontology infrastructure, which will give the base for the integration of low-, mid- and high-level features extracted from audio-video and associated text/speech transcripts.

3.1. Development of a multimedia ontology infrastructure

The multimedia ontology infrastructure of K-Space will contain qualitative attributes of the semantic objects that can be detected in the multimedia material, e.g. color homogeneity, in the multimedia processing methods, e.g. color clustering, and in the numerical data or low-level features, e.g. color models. The ontology infrastructure will also contain the representation of the top-level structure of multimedia documents in order to facilitate a full-scale annotation of multimedia documents. R&D work will be dedicated to the specification and development of a multimedia content ontology supporting the representation of the structure of the content of multimedia documents. Work will also be dedicated to research on ontologies for low-level visual features, concentrating on a model for the concepts and properties that describe visual features of objects, especially the visualizations of still images and videos in terms of low-level features and media structure descriptions. Also, a prototype knowledge base will be designed to enable automatic object recognition in images and video sequences. Prototype instances will be assigned to classes and properties of the domain specific ontologies, containing low level features required for object identification.

Partners of K-Space dealing with textual analysis will integrate into this ontology infrastructure the typical features for text analysis, also proposing ontology classes at a higher-level, that supports the modeling of interrelated cross-media features (multimedia and text). We will base our work on the proposal made by (Buitelaar et. al 2005).

3.2. Use of Textual Information and Knowledge Bases for Semantic Feature Extraction from Audio Signal

In K-Space some work will be dedicated to the extension of state-of-the-art processing and analysis algorithms to handle high-level, conceptual representations of knowledge embedded in audio content based on reference ontologies and semantically annotated associated text (including speech transcripts, when the quality of the transcripts allows it).

K-Space will consider all types of audio sources ranging from speech to complex polyphonic music signals. The description schemes of the MPEG-7 standard define how audio signals can be described at different abstraction levels: from the lowest level primitives, such as temporal or audio spectrum centroids, spectrum flatness, spectrum spread, inharmonicity, etc., to the highest level, related to semantic information. Semantic information is related to textual information on audio such as titles of songs, singers’ names, composers’ names, duration of music excerpt, etc.

This textual information is often encoded using the text annotation tool of the Linguistic Description Scheme (LDS) of MPEG-7. An example of such a (manual) annotation related to a video sequence is given just below:

```
<VideoSegment id="shot1_13">
    <MediaTime>
        <MediaTimePoint>T00:01:40:11008F30000</MediaTimePoint>
        <MediaDuration>PT10S26326N30000F</MediaDuration>
    </MediaTime>
    <TextAnnotation confidence="0.500000">
        <FreeTextAnnotation>
    TRACKS STOPPED ROLLING NOSE AND FORMALLY FILED A HIGHWAY WITH EIGHT DAILY NEW YORK NEWSPAPERS WHERE THE VOID OF NEWSPAPERS THE VOID OF CUSTOMERS
        </FreeTextAnnotation>
    </TextAnnotation>
</VideoSegment>
```

Interesting to note here, is that the media time is also given, so that this can be used as a way to look for alignment of the low-level features and the high-level features that can be extracted from the text.

Our work will consist here in proposing a linguistic and semantic analysis of all the available free text annotations used in the semantic representantion of audio signal, and mapping this onto either the structured annotation scheme of LDS (specifying the "who", the "what", the "why", the "when" etc in an explicit way), or to provide for an ontology based semantic annotation (in term of instances of ontology classes).

We will also use TRECVid data, using aligned speech transcripts and video shots, and looks for ways to extracts high-level semantics from the transcripts (which are attached to the audio-video stream using also the LD scheme). For sure the quality of transripts is often bad, and here we will use robust NLP methods and limits ourself to the detection of basic textual chunks.

For improving the alignment of text/transcripts with the audio (or video) signal, we try to identify typical lexical items that link directly such text/transcripts to the signal ("here you can see" etc.).

3.3. Analysis of Complementary Textual Sources for adding Semantic Metadata to Multimedia Content

The human understanding of multimedia resources is often facilitated by usage of complementary sources. In order to simulate this attitude, K-Space will implement mining methods and tools for such complementary resources in order to reduce the semantic gap by deriving annotations from those sources, and so to reach a more complete annotation of (sequences of) images.

The project will address mining and analysis for semantic features extraction within two different types of resources:

- Mining and analysing primary resources: Analysis of the primary resources that are attached to the multimedia data, e.g. texts around pictures, subtitles of movies, etc.
- Mining and analysis of secondary and tertiary resources: Analysis of data and text related to the multimedia data under consideration, e.g. a programme guide for a TV broadcaster or a web site displaying similar pictures.

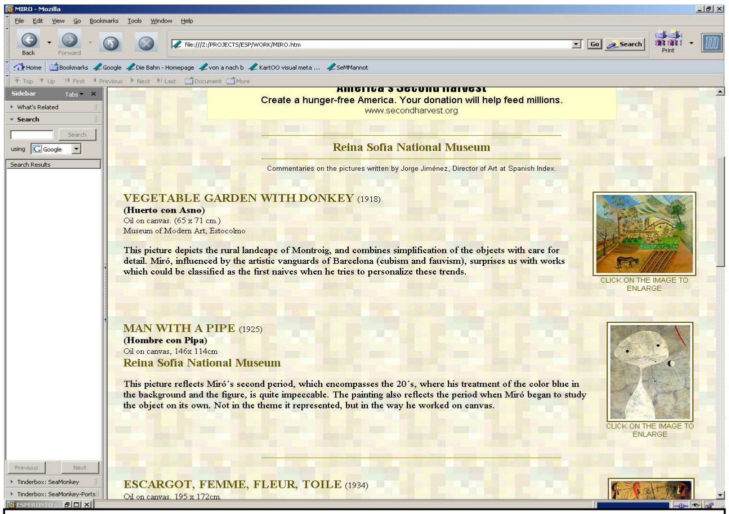
4. Linguistic Analysis of relevant Text Regions

We report on a first experiment made within the Esperonto project, where also a small ontology on artworks has been made availble to the project parners. In this ontology, typical terms were associated to every class (so for example the terms "surrealism" and "cubism" are associated to the class "artistic_movement".

In the Esperonto scenario, we first defined the possibly relevant text regions for the semantic annotation of the image (see below in Figure 1 the example of such an image, in a web page dedicated to the painter Miro, the first image being the base for our indexing prototype tool). We identified following text regions (in both the text and in the html code):

- Title of the document
- Caption text: „Click on the image to enlarge“ (a non relevant item, to be filtered by the tools, also on the base of lexical properties of the words).
- Content of the HTML „Alt“ tag: “'VEGETABLE GARDEN WITH DONKEY'”
- Content of the HTML „Src“ tag: *http://www.spanisharts.com/reinasofia/miro/burro_lt.jpg*
- Abstract text
- Running text

On the base of this, we wrote a tool that supports the manual selection of such textual regions, and send those to a linguistic processing engine. The linguistic processing engine has been augmented with metadata sepcifying the type of text to be processed (we expect for example the Title and the "Alt" text to consist mostly of phrases.)


MIRO - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward
file:///Z:/PROJECTS/ESP/WORK/MIRO.htm
Go Search
Print
Home Bookmarks Google Die Bahn - Homepage von a nach b KartOO visual meta ... SeMMannot
Top Up First Previous Next Last Document More
Sidebar Tabs
What's Related
Search
Search
using Google
Search Results
Previous Next
Tinderbox: SeaMonkey
Tinderbox: SeaMonkey-Ports

America's Second Harvest
Create a hunger-free America. Your donation will help feed millions.
www.secondharvest.org

Reina Sofía National Museum

Commentaries on the pictures written by Jorge Jiménez, Director of Art at Spanish Index.

VEGETABLE GARDEN WITH DONKEY (1918)
(Huerto con Asno)
Oil on canvas. (65 x 71 cm.)
Museum of Modern Art, Estocolmo

This picture depicts the rural landcape of Montroig, and combines simplification of the objects with care for detail. Miró, influenced by the artistic vanguards of Barcelona (cubism and fauvism), surprises us with works which could be classified as the first naives when he tries to personalize these trends.

CLICK ON THE IMAGE TO ENLARGE

MAN WITH A PIPE (1925)
(Hombre con Pipa)
Oil on canvas, 146x 114cm
Reina Sofía National Museum

This picture reflects Miró´s second period, which encompasses the 20´s, where his treatment of the color blue in the background and the figure, is quite impeccable. The painting also reflects the period when Miró began to study the object on its own. Not in the theme it represented, but in the way he worked on canvas.

CLICK ON THE IMAGE TO ENLARGE

ESCARGOT, FEMME, FLEUR, TOILE (1934)
Oil on canvas, 195 x 172cm.

Figure 1 Example of a web page with images of paintings. Various text regions are offering different kind of "metadata" to the

5. The Linguistic Analysis of the Various Text Regions

In the following lines, we show some of the (partial) results of the linguistic analysis, as applied to the various text segments. Our tools are delivering a dependency annotation:

- „Alt“ text: 'VEGETABLE GARDEN WITH DONKEY'

<NP HEAD="garden" PRE_MOD="vegetable" <POST_MOD CAT= "PP" HEAD="with" NP_COMP_HEAD="donkey"</POST_MOD></NP>

- Abstract/Running text: "…This picture depicts the rural landcape of Montroig …"

<SENT SUBJ="This picture" PRED="depicts OBJ="the rural lansdscape of Montroig"</SENT>

- Detailled annotation of the direct_object: <NP HEAD="landscape" PRE_MOD="rural" <POST_MOD CAT="PP" HEAD="of" NP_COMP_HEAD="Montroig"</POST_MOD> </NP>

6. The Semantic Annotation

On the base of a mapping between the linguistic dependency and the terms associated to the classes of the ontology (whereas we accomodated the classes of the ontology to be associated with patterns (for coping for example with date expressions), we could provide for a semantic annotation of the texts associated with the picture.

6.1. The (Toy) Art Ontology (schematized)

- Object > Artork > Painting [has_creator, has_name, has_subject, has_dimension, has_material, has_genre, has_date...]
- Person > Artist > Painter [has_name, has_birth_date, part_of_artistic_movement …]

6.2. The Instantiation of Classes

- Title: Vegetable garden with donkey
- Creator: Miro
- Date: 1918
- Genre: naïve (if correctly extracted by some reasoning on the linguistically and semantically annotated text)
- Subject: rural landscape of Montroig + garden and donkey (if the association between the title and the explanation given by the art expert can be grouped).
- Dimension: 65x71
- Material: Oil on canvas

6.3. Some remarks

This result was possible due to various facts. First, the system "knew" that the text was about art, and we assumed that the text is related to the picture. Second, we had an ad-hoc relation of terms to the concepts of the ontology (for example "Oil"). Third we had defined typical patterns realising some concepts (date, material etc.). But our focus was more on syntactic analysis (in fact dependency analysis). So the Subject of the sentence "This picture" together with the typical verb "depicts" and its DirectObject allowed here to "map" the whole DierctObject to the "subject" of the picutre (what the picture is about). The dependency analysis of the DirectObj allows us to further precise the topic of the picture: it is a rural (mod) Lanscape (head) of Montroig (post_nom_mod), thus introducing quite fine granularity in the indexing of the image.

The missing point here: there is no principled relation between the terms in the ontology and the results of the image analysis (in term of low-level features). We think here that a domain ontology taking into account the specific features for the multi-modal analysis components could help in establishing this relationship, not only at lexical level but also maybe at the syntactic level (the dependency relations in linguistic fragments of texts refering to images could give some hints about the distribution of objects in the picture).

But clearly one has to think first of a specific classification of lexical items in terms of possible indices of multimedia content, before looking a syntactic properties of text related to images.

7. Conclusions

We have described some approaches that take advantages of so-called complementary sources (text/transcripts) for automatically adding semantic metadata to image material. Till now we concentrated on the linguistic processing aspect, with a very small lexical base. Lexical consideration would allow to extend our approach and to really evaluate it. More principled lexical information would also support the automatic detection of text parts that are referring directly to the content of the image under consideration, and not to metadata related to this image (in which museum is the picture, wo made it etc.) or on topics not related to the image at all.

We will also have to think at principled ways for integrating the lexical knowledge into the multimedia infrastructure. At the beginning we would follow a similar approach that has been proposed for the integration of lexical information in domain specific ontologies, and proposed in the SmartWeb project.

8. Acknowledgments

"The research programm described in this paper is supported by the European Commission, contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space."

9. References

Buitelaar P., Sintek M., Kiesel M (2005).. Feature Representation for Cross-Lingual, Cross-Media Semantic Web Applications. In: *Proceedings of the ISWC 2005 Workshop "SemAnnot"*..

Athanasiadis T., Tzouvaras V., Petridis K., Precioso F., Avrithis Y. and Kompatsiaris Y. (2005). Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. *In proceedings of the ISWC 2005 Workshop "SemAnnot"*.

Jane Hunter: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Trans. Circuits Syst. Video Techn. 13(1): 49-58 (2003)

Jane Hunter: Adding Multimedia to the Semantic Web: Building an MPEG-7 ontology. SWWS 2001: 261-283

AceMedia project: http://www.acemedia.org/aceMedia

BUSMAN project: http://busman.elec.qmul.ac.uk/

Esperonto Project: http://www.esperonto.net

K-Space Project: http://kspace.qmul.net

SmartWeb Project: http://www.smartweb-projekt.de

TRECVid: http://www-nlpir.nist.gov/projects/trecvid/

The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems

Michael Grubinger[1], Paul Clough[2], Henning Müller[3] and Thomas Deselaers[4]

[1] School of Computer Science and Mathematics, Victoria University of Technology
PO Box 14428, Melbourne VIC 8001, Australia
michael.grubinger@research.vu.edu.au
[2] Department of Information Studies, Sheffield University
Western Bank, Sheffield, S1 4DP, UK
p.d.clough@sheffield.ac.uk
[3] Medical Informatics, University and Hospitals of Geneva
24, rue Micheli-du-Crest, 1211 Geneva 14, Switzerland
henning.mueller@sim.hcuge.ch
[4] Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, Aachen, Germany
deselaers@cs.rwth-aachen.de

Abstract

In this paper, we describe an image collection created for the CLEF cross-language image retrieval track (ImageCLEF). This image retrieval benchmark (referred to as the IAPR TC-12 Benchmark) has developed from an initiative started by the Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR). The collection consists of 20,000 images from a private photographic image collection. The construction and composition of the IAPR TC-12 Benchmark is described, including its associated text captions which are expressed in multiple languages, making the collection well-suited for evaluating the effectiveness of both text-based and visual retrieval methods. We also discuss the current and expected uses of the collection, including its use to benchmark and compare different image retrieval systems in ImageCLEF 2006.

1. Introduction

Standard datasets are vital for benchmarking the performance of information retrieval systems and allowing the comparison between different approaches or methods (Over et al., 2004; Müller et al., 2001; Narasimhalu et al., 1997; Smith, 1998). For example, initiatives such as TREC[1] (Text REtrieval Conference, Harman, 1996) and CLEF[2] (Cross-Language Evaluation Forum, Braschler & Peters, 2004) have provided the necessary resources to enable comparative evaluation of Information Retrieval (IR) systems. These initiatives have motivated and encouraged research and have clearly contributed to the advancement of information retrieval systems over the past years.

A core component of any benchmark is a set of documents (e.g. texts, images, sounds or videos) that are representative of a particular domain (Markkula et al., 2001). However, finding such resources for general use is often difficult, not least because of copyright issues which restrict the distribution and future accessibility of data. This is especially true of visual resources that are often more valuable than written texts and therefore subject to limited availability and access for the research community. For example, consider the Corbis Image Database[3] or Getty Images[4], large collections of images, but because of being commercial datasets they are generally inaccessible for research purposes. To evaluate aspects of visual information systems (e.g. automatic annotation, retrieval or pattern recognition), collections of visual objects that can be made available to the research community are required, e.g. the effort described in (Jörgensen, 2001) to create annotated databases for system evaluation, but the outcome of these efforts is still sparse.

1.1. Collections available for Evaluation

For a long time, the de–facto standard for image retrieval evaluation was the Corel Photo CDs. However, they are problematic: the CDs are expensive to obtain, are protected by copyright and legal restrictions on use and therefore difficult to distribute for large-scale evaluation, they have limited written metadata which makes them less suitable for evaluating methods of text-based image retrieval, and the CDs are currently unavailable to buy and therefore not available to researchers. It was also shown that subsets of this database can easily be tailored to show improvements (Müller, Marchand-Maillet & Pun, 2002).

An alternative database that is free of charge, not restricted by copyright restrictions, and previously used for evaluation is the collection built by the University of Washington[5]. It contains approximately 1,000 images, clustered by the location that images were taken from. Other databases are available for computer vision research, but rarely used for image retrieval[6] because they do not represent realistic retrieval data. The Benchathlon[7] created an evaluation resource, but without search tasks or ground truth. ALOI[8] (Amsterdam Library of Object Images) and LTU (LookThatUp) Technologies[9] have created large databases with colour images of small

[1] http://trec.nist.gov/
[2] http://www.clef-campaign.org/
[3] http://pro.corbis.com/
[4] http://www.gettyimages.com/
[5] http://www.cs.washington.edu/research/imagedatabase
[6] http://homepages.inf.ed.ac.uk/rbf/CVonline/CVentry.htm
[7] http://www.benchathlon.net/
[8] http://staff.science.uva.nl/~aloi/
[9] http://www.ltutech.com/

objects with varied viewing (and illumination) angles, but primarily designed for pure pattern recognition evaluation and less for information retrieval. There are a few royalty-free databases available in specialised domains like Casimage[10] and IRMA[11] for medical imaging, or the St. Andrews collection[12] that is copyrighted but was made available for retrieval evaluation of historic (mainly black and white) photographs. Many web pages actually make images available in large quantities and with copyright notices attached such as FlickR[13] or Morguefile[14]. Although many of these images are available without many copyright restrictions for simple use, it is often not allowed to redistribute them particularly not combined in large numbers. Intellectual property rights with respect to digital content (and particularly images) are currently not always clear.

The TRECVID (TREC video retrieval track, Smeaton et al., 2004) image collections have increasingly been used for image retrieval in the last two years as well. The key frames can indeed be used for image retrieval and object recognition, and the tasks created correspond well to simple journalists search tasks. As the videos also contain the speech of the video, multimodal retrieval evaluation is possible on these datasets as well.

The IAPR collection described in this paper is an example of another collection, specifically created with the following aims in mind: to provide a realistic collection of images suitable for a wide number of evaluation purposes, to provide images with associated written information representing typical textual metadata that can be used to explore the semantic gap between images and words, metadata expressed in multiple languages[15]. The goal is to provide a dataset that is free of charge and copyright restrictions and therefore available to the general research community. This paper describes the creation and composition of the IAPR TC-12 Benchmark and discusses how the collection is currently being used within ImageCLEF[16] for the evaluation of multilingual and multimodal image retrieval systems.

2. The Image Collection

At present, the IAPR TC-12 image collection consists of 20,000 images (plus 20,000 corresponding thumbnails) taken from locations around the world and comprising a varying cross-section of still natural images.

2.1. History of the IAPR benchmark

In 2000, the Technical Committee 12 (TC-12) of the International Association for Pattern Recognition (IAPR[17]) recognized the need for a standard benchmark for multimedia retrieval and began an effort to create a freely available database of images with associated annotations. This started by developing a set of recommendations and specifications of an image benchmark (Leung & Ip, 2000). Based on this criteria, a first version of a benchmark consisting of 1,000 multi-object colour images, 25 search requests (or queries), and a collection of performance measures was set up in 2002.

Developing a benchmark is an incremental and ongoing process. The IAPR TC-12 Benchmark was refined, improved and extended to 5,000 images in 2004, using a benchmark administration system (Grubinger & Leung, 2003). At the end of that year, an independent travel organisation (viventura[18]) provided access to around 10,000 of their images including multilingual annotations of varying quality in three languages (English, German, Spanish). This increased the total number of images in the benchmark to 15,000. Of course, a benchmark is not beneficial unless actually used by the research community. Therefore in 2005, discussions began for involving the IAPR TC-12 Benchmark as part of an image retrieval task in CLEF. ImageCLEF has begun using the collection and is expected to continue using it for future tasks (see Section 4). With 10,000 additional images from the travel organisation, the total number of available images rose to 25,000 images (Grubinger, Leung & Clough, 2005) but was soon reduced to 20,000 images annotated in three languages.

2.2. Origin and Selection of Images

The majority of the images are provided by viventura, an independent travel company that organizes adventure and language trips to South-America. At least one travel guide accompanies each tour and they maintain a daily online diary to record the adventures and places visited by the tourists (including at least one corresponding photo). Furthermore, the guides provide general photographs of each location, accommodation facilities and ongoing social projects. Not all of the images provided are suitable for a benchmark and must undergo a selection process (Grubinger & Leung, 2003). In total, 20,000 images were selected and added to the IAPR TC-12 Benchmark.

2.3. Example Images

The image collection includes pictures of a range of sports (Fig. 1) and actions (Fig. 2), photographs of people (Fig. 3), animals (Fig. 4), cities (Fig. 5), landscapes (Fig. 6) and many other aspects of contemporary life.

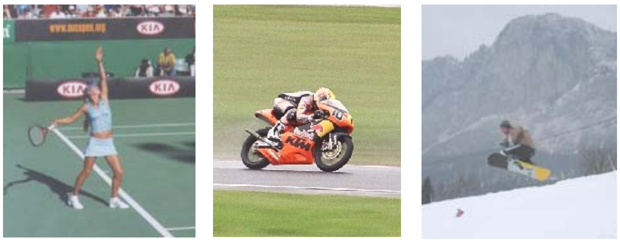


Figure 1: Examples for sports photos (Tennis, Motorcycling, Snowboarding)

[10] http://www.casimage.com/

[11] http://irma-project.org/

[12] http://www-library.st-andrews.ac.uk/

[13] http://www.flickr.com/

[14] http://morguefile.com/

[15] Considering annotations in multiple languages is an important aspect of text-based image retrieval as real-life collections such as FlickR are intrinsically multilingual.

[16] http://ir.shef.ac.uk/imageclef/

[17] http://www.iapr.org/

[18] http://www.viventura.de/

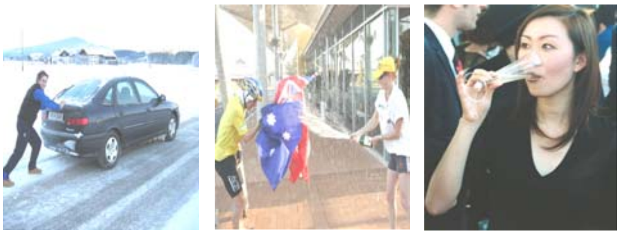


Figure 2: Examples for action pictures
(Pushing, Celebrating, Drinking)

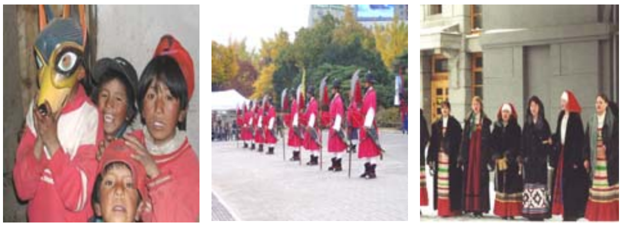


Figure 3: Examples for people shots
(Peruvian Children, Korean Guards, Russian Singers)

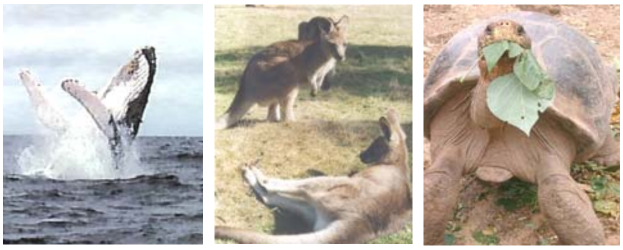


Figure 4: Examples for animal photos
(Humpback Whale, Kangaroos, Galapagos Giant Turtle)

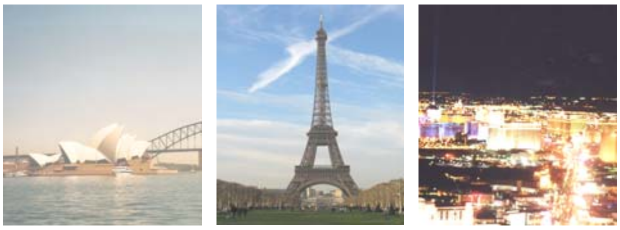


Figure 5: Examples for city pictures
(Sydney Opera House, The Eiffel Tower, Las Vegas Strip)

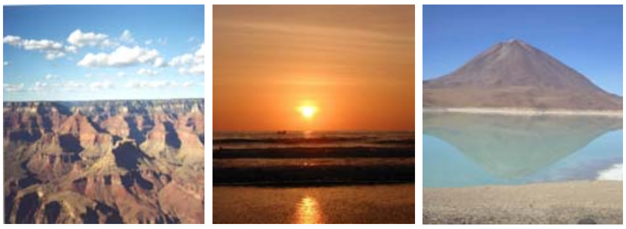


Figure 6: Examples for landscape shots
(Grand Canyon, Montañita Beach, Volcano Licancabur)

2.4. Diversity of the Image Collection

The IAPR TC-12 photographic collection contains many different images of similar visual content, but varying illumination, viewing angle and background. This is because most of the tours offered by the travel company are repeated on a regular basis and have fixed itineraries. Thus, the tours always visit the same tourist destinations where the guides usually take photos of tourists in varying poses (see Fig. 7) and/or of tourist attractions with varying viewing angles (Fig. 8), weather conditions (Fig. 9) or at different times of the day (Fig. 10). Hence, this makes the benchmark also well-suited for content-based retrieval tasks as it allows a range of prototypical searches to explore retrieval effectiveness with these varying settings.

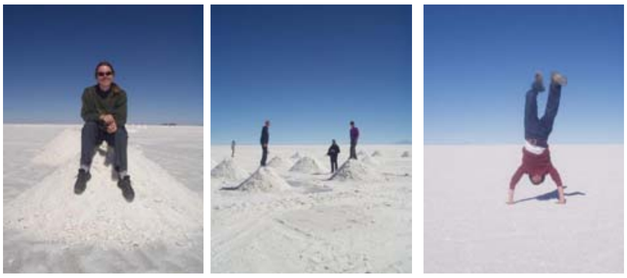


Figure 7: Tourists from three different tour groups at the Salt Lake of Uyuni in Bolivia

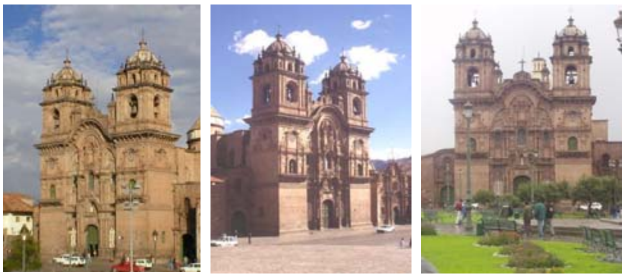


Figure 8: The Cathedral of Cuzco, Peru, in different viewing angles (right, left and front)

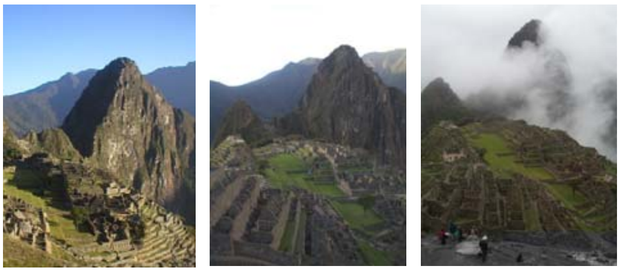


Figure 9: The Inca ruins of Machu Picchu in bright sunshine, on an overcast day and in foggy and rainy conditions

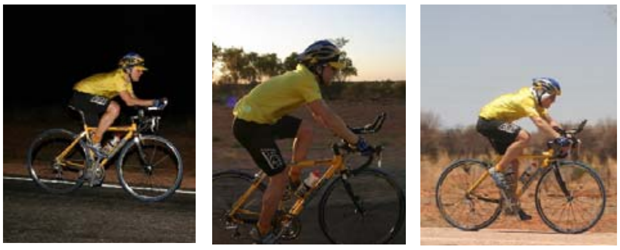


Figure 10: A cyclist riding a racing bike at night, in the morning and during the day

2.5. Image Statistics

This section provides information on a range of attributes which characterise the image collection (e.g. the size of images, image formats, and temporal and geographical extent of the collection).

2.5.1. Sizes of Images and the Collection

The photographs provided by the travel organisation exhibit the following differences based on the technology used to capture the images: photographs taken with digital cameras which have a 4:3 relation of width to height (96x72 pixels for thumbnails; 480x360 pixels for larger versions), and photographs taken with a non-digital (or traditional) camera which have been subsequently scanned and have a 3:2 relation of width to height (92x64 pixels for thumbnails; 480x320 pixels for larger versions).

Thumbnails require between 2 and 10 KB each (an average file size of 5.69 KB); the larger versions range from 20 to 200 KB (an average size of 85.25 KB), depending upon their content and colour composition. The total size of the image collection is 1.66 GB (and 111 MB for the corresponding thumbnails). All images are stored in the JPEG image format.

2.5.2. Temporal Range

Most photographs have been taken since 2001 and Fig. 11 shows the temporal distribution of images between 2001 and 2005. The earliest photo in the collection dates back to 2000; the most recent taken in July 2005. The mean date is June 2003, the standard deviation is 1.12 years and the median is January 2004.

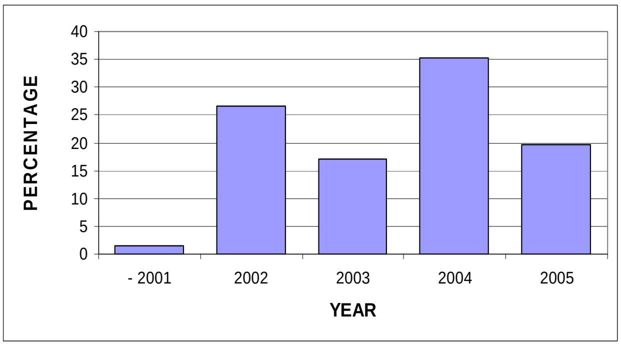


Figure 11: Temporal Range

2.5.3. Geographical Range

The IAPR TC-12 collection is spatially diverse, with pictures taken in more than 30 countries worldwide including Argentina, Australia, Austria, Bolivia, Brazil, Chile, Colombia, Ecuador, France, Germany, Greece, Guyana, Korea, Peru, Russia, Spain, Switzerland, Taiwan, Trinidad & Tobago, Uruguay, USA, and Venezuela. Fig. 12 shows the proportion of images taken in these countries (represented in their international three letter code[19]):

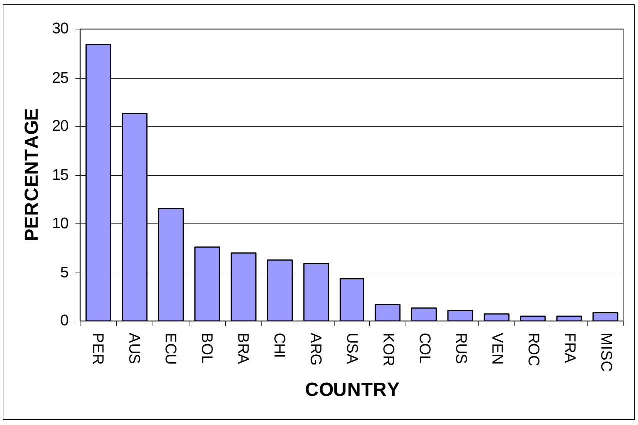


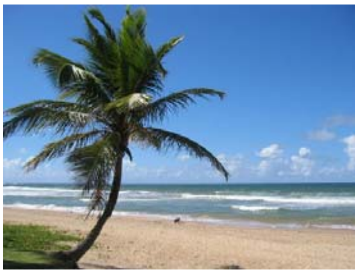
Figure 12: Variation across countries
(with more than 100 images)

Most of the images originate from Peru (28.4 %), followed by Australia (21.3 %) and Ecuador (11.6 %), reflecting the geographic location of contributors. The collection comprises a total of 11 countries contributing more than 1 % to the collection, and 14 countries with at least 100 images or 0.5% of the collection.

3. Image Annotations

3.1. Original Annotations

Tour guides are supposed to add a short caption for each image they include with their diaries. These captions include a title for the image, a short description, a location and date of creation. Most annotations are written in German as the travel company viventura targets the German-speaking market. However in some cases, guides also use Spanish, Portuguese or English.



Title: Praia do Flamengo
Description: Der Praia do Flamengo gilt als einer der schönsten Strände Brasiliens!
Location: Salvador, Brasilien
Date: 2. Oktober 2004

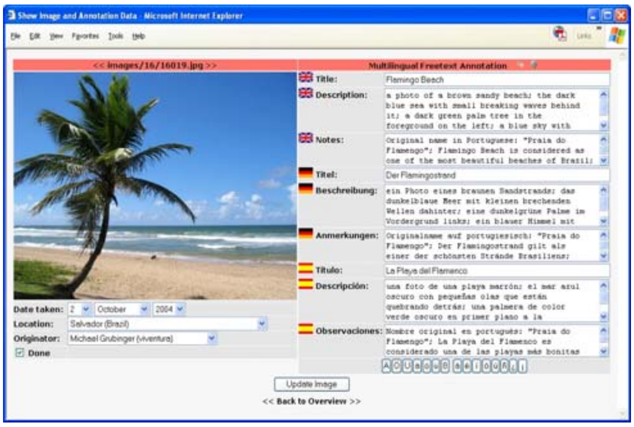
Figure 13: Example of an original annotation

Fig. 13 shows an example image with a mixed-language original annotation in Portuguese and German. The Portuguese title states briefly what the image is about (in this case the name of the beach "Flamingo Beach"); the description of the image is in German and provides further detail ("Flamingo Beach is considered as one of the most beautiful beaches of Brazil!"). Both location ("Salvador, Brazil") and the date ("October 2nd, 2004") are expressed in German language and form. Since most of the tour guides are local employees from South-America and therefore native Spanish or Portuguese speakers, the quality of the annotations (and also their detail) varies tremendously.

[19] Abbreviations of the International Olympic Committee

3.2. Revised Annotations

In order to provide a consistent set of annotations for benchmarking, the original annotations of images selected for inclusion in the IAPR TC-12 Benchmark have been manually checked, corrected and completed in compliance with slightly modified image annotation rules (Grubinger & Leung, 2003). These rules specify the use of the right terminology, annotation precision, cardinality, image settings and number of annotation sentences and also restrict the level of subjective interpretation.


Show Image and Annotation Data - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links

<< images/16/16019.jpg >>

Date taken: 2 October 2004
Location: Salvador (Brazil)
Originator: Michael Grubinger (viventura)
☑ Done

Multilingual Freetext Annotation

Title:	Flamingo Beach
Description:	a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with
Notes:	Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil;
Titel:	Der Flamingostrand
Beschreibung:	ein Photo eines braunen Sandstrands; das dunkelblaue Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne Palme im Vordergrund links; ein blauer Himmel mit
Anmerkungen:	Originalname auf portugiesisch: "Praia do Flamengo"; Der Flamingostrand gilt als einer der schönsten Strände Brasiliens;
Titulo:	La Playa del Flamenco
Descripción:	una foto de una playa marrón; el mar azul oscuro con pequeñas olas que están quebrando detrás; una palmera de color verde oscuro en primer plano a la
Observaciones:	Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas

Figure 14: Benchmark Administration System

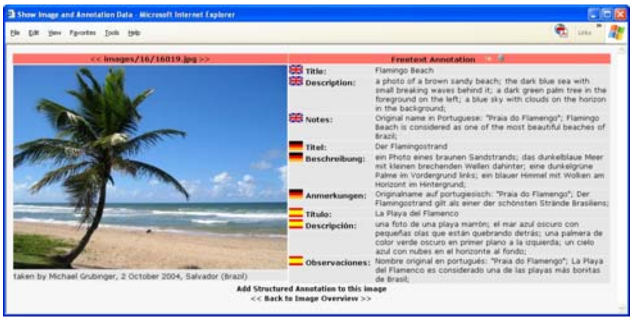
Fig. 14 shows a screenshot of a custom-built Benchmark Administration System used to carry out the revision process (see (Grubinger & Leung, 2004) for details of its specification, architecture and implementation). In particular, information provided about the location was checked and the image description divided into two separate fields: one part to describe visible information in the image; the other providing additional notes which are not part of visual content visible within the image. The original (German) annotations were corrected, missing text and notes from the images completed, and all annotations translated into English and Spanish.

3.3. Finalised Annotations

The final set of images and consistent data for the Benchmark associates each photograph with a semi-structured text caption consisting of the following seven fields:

- a unique identifier,
- a title,
- a free-text description of the semantic (and visual) contents of the image,
- notes for additional information,
- the name of the photographer,
- fields describing where and when the photograph was taken.

These annotations are stored in a MySQL database and managed by the Benchmark Administration System. Fig. 15 shows a complete annotation for an example image.


Show Image and Annotation Data - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links

<< images/16/16019.jpg >>

taken by Michael Grubinger, 2 October 2004, Salvador (Brazil)

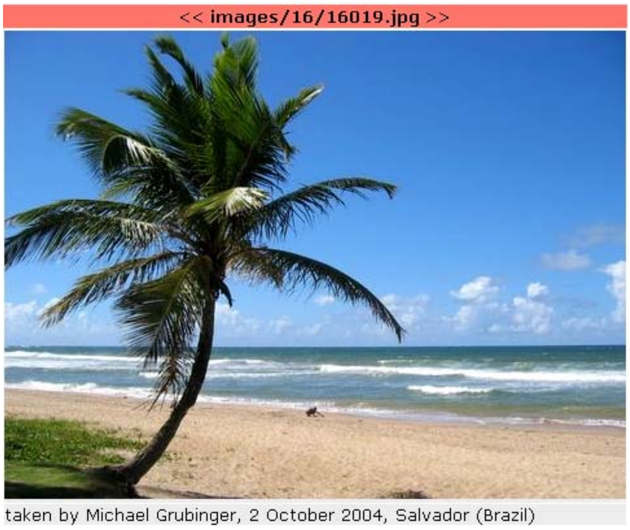
Freetext Annotation

Title:	Flamingo Beach
Description:	a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;
Notes:	Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil;
Titel:	Der Flamingostrand
Beschreibung:	ein Photo eines braunen Sandstrands; das dunkelblaue Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne Palme im Vordergrund links; ein blauer Himmel mit Wolken am Horizont im Hintergrund;
Anmerkungen:	Originalname auf portugiesisch: "Praia do Flamengo"; Der Flamingostrand gilt als einer der schönsten Strände Brasiliens;
Título:	La Playa del Flamenco
Descripción:	una foto de una playa marrón; el mar azul oscuro con pequeñas olas que están quebrando detrás; una palmera de color verde oscuro en primer plano a la izquierda; un cielo azul con nubes en el horizonte al fondo;
Observaciones:	Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas de Brasil;

Add Structured Annotation to this image
<< Back to Image Overview >>

Figure 15: Complete Annotation for Image 16019

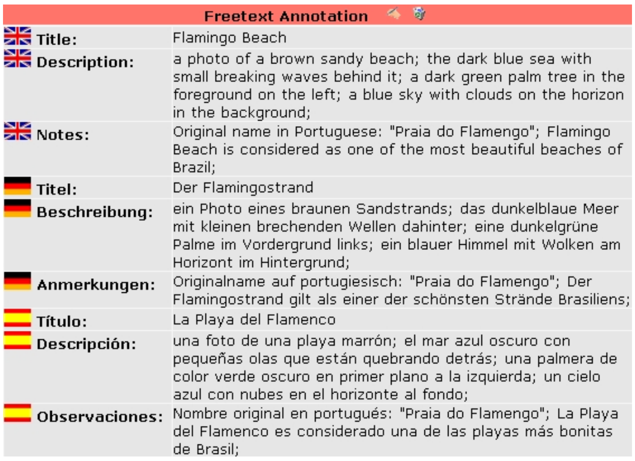
The information on the screen is divided into two parts: the left (see Fig. 16) displays the image, its unique identifier (see Section 3.3.1) and part of the image meta-data: the photographer, the location (see Section 3.3.5) and the date (Section 3.3.6).


<< images/16/16019.jpg >>

taken by Michael Grubinger, 2 October 2004, Salvador (Brazil)

Figure 16: The left half of the annotation: image meta-data

The right part of the screen (see Fig. 17) contains multi-lingual free-text annotations of the title (Section 3.3.2), the image description (Section 3.3.3) and the notes (Section 3.3.4).


Freetext Annotation

Title:	Flamingo Beach
Description:	a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;
Notes:	Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil;
Titel:	Der Flamingostrand
Beschreibung:	ein Photo eines braunen Sandstrands; das dunkelblaue Meer mit kleinen brechenden Wellen dahinter; eine dunkelgrüne Palme im Vordergrund links; ein blauer Himmel mit Wolken am Horizont im Hintergrund;
Anmerkungen:	Originalname auf portugiesisch: "Praia do Flamengo"; Der Flamingostrand gilt als einer der schönsten Strände Brasiliens;
Título:	La Playa del Flamenco
Descripción:	una foto de una playa marrón; el mar azul oscuro con pequeñas olas que están quebrando detrás; una palmera de color verde oscuro en primer plano a la izquierda; un cielo azul con nubes en el horizonte al fondo;
Observaciones:	Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas de Brasil;

Figure 17: The right half of the annotation: multi-lingual free-text annotations in English, German and Spanish

These free-text annotations (and also the location and date information) are currently available in three languages, with the German and English versions in a release status and the Spanish version currently being verified. The German version uses Austrian vocabulary and spelling because the annotation creator is Austrian. Australian vocabulary and spelling (almost equivalent to British English) for the English version is used because the annotation process was undergone in Melbourne, Australia. The author did, in cases of doubt, ask local native speakers for translations or vocabulary.

3.3.1. Unique Image Identifiers

Each image is assigned a unique identifier. For instance, the unique identifier of the example in Figure 15 is “16019”, which determines the filename of the image (“16019.jpg”) and of the annotation files (“16019.eng” for English, “16019.ger” for German and “16019.spa” for Spanish).

3.3.2. Title

The title field contains a short statement describing what the image is about. This can include proper names like “Flamingo Beach”, general noun phrases like “cyclist at night”, or a combination of both such as “llamas at Machu Picchu”. The title can also be a short sentence such as “Max is surfing in Torquay”.

This title field is equivalent to descriptive annotations found in many personal photographic collections (i.e. annotations that typical users might add to their own photographs). In most cases the title field is not very different to the original annotations. The average length of the title field for English is 5.35 words, with a standard deviation of 2.37 words. The shortest title consists of one word; the longest consisting of 17 words. Table 1 displays statistics for different versions of the titles.

Number of Words	German	English	Spanish
Average	4.85	5.35	5.97
standard deviation	2.10	2.37	2.68
Minimum	1	1	1
Median	5	5	6
Maximum	14	17	19

Table 1: Word statistics for the title field.

German titles are on average shorter in length (and Spanish titles longer) than the English titles. This does not necessarily mean that the Spanish titles are more complex than the German ones; it is more likely due to the fact that composite nouns that can be described in one word in German (e.g. “Flamingostrand”) are often expressed by two words in English (“Flamingo Beach”), whereas Spanish requires three words (“Playa del Flamenco”).

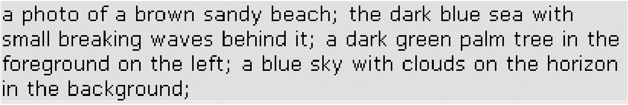
3.3.3. Description

The description field contains a semantic description of the image contents, or in other words, it describes in short sentences and noun phrases (terminated by semi-colons) what can be recognized in an image without any prior information or extra knowledge. Keywords alone are not used as they are not very precise due to the lack of syntax (Tam & Leung, 2001) and studies show that users tend to create short narratives to describe images when unconstrained from a retrieval task (Jörgensen, 1996; O'Connor B., O'Connor M. & Abbas, 1999).

Number of Words	English	German	Spanish
average	23.06	18.92	N/A
standard deviation	10.35	8.48	N/A
minimum	2	2	N/A
median	22	18	N/A
maximum	85	74	N/A

Table 2: Word statistics for the description field.

The average length of the description field is 23.06 words (with a standard deviation of 10.35 words). The shortest description comprises two words; the longest is 85 words, with a median of 22 words (see Table 2). Again, the German descriptions use fewer words than the English version (see section 3.3.2).



a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;

Figure 18: the description field of image 16019

Number of Annotation Sentences. Obviously, there is no limit to how semantically rich one could make the description of an image. Most of the annotations have between one and five more or less complex annotation sentences (Fig. 18, for instance, has four). In many annotations, two or more of these sentences are conjunct (and), hence, a statistic evaluation of the number of sentences is not representative for the annotations.

Sentence Order. The semantic descriptions of the image follow a certain priority pattern: The first sentence(s) describe(s) the most obvious semantic information (like “a photo of a brown sandy beach”). The latter sentences are used to describe the surroundings or settings of an image, like smaller objects or background information (“a blue sky with clouds on the horizon in the background”).

Linguistic Patterns. Many of these annotation sentences or noun phrases follow one of the main linguistic patterns P (or a more different combination based on these) shown in Table 3.

Pattern P	Example
S	a red rose
S–V	a boy is singing
S–TA	a boy at night
S–PA	a boy in a garden
S–PA–TA	a boy in a garden at night
S–V–TA	a boy is singing at night
S–V–PA	a boy is singing in a garden
S–V–PA–TA	a boy is singing in a garden at night
S–V–O	a girl is kissing a boy
S–V–O–TA	a girl is kissing a boy at night
S–V–O–PA	a girl is kissing a boy in a garden
S–V–O–PA–TA	a girl is kissing a boy in a garden at night

Table 3: Linguistic Pattern of Descriptions.

Any of these patterns P mentioned in Table 3 are also used for background and foreground information and can be further specified as to where they lie within the image (see Table 4):

Pattern	Example
P–PA	P on the left
P–BG	P in the background
P–FG	P in the foreground
P–BG–PA	P in the background on the right
P–FG–PA	P in the foreground on the left

Table 4: Linguistic Pattern of the Descriptions.

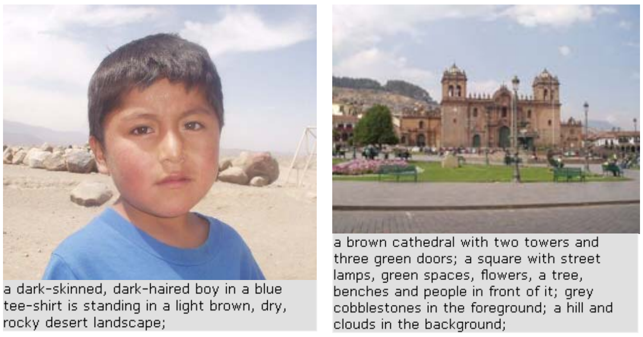
Table 5 provides an overview and a description of the symbols used in Tables 3 and 4.

Symbol	Description
S	subjects (with or without adjectives)
V	verbs (with or without adverbs)
O	objects (with or without adjectives)
PA	place adjunct(s) with place preposition
TA	time adjunct(s) with time preposition
P	any pattern or combination of patterns described in Table 3
FG	in the foreground
BG	in the background

Table 5: Symbols.

Appropriate Tense. Annotations describe actions or situations in images at certain times. The grammatically correct tenses, therefore, are the *present continuous tense* in English, the *Präsens* in German and *estar + gerundio* in Spanish. The auxiliary verbs for English (be) and Spanish (estar) are omitted in some annotations.

Adjectives. As with the number of annotation sentences, there is obviously no limit how detailed each object could be described by the use of adjectives. In general, the fewer objects there are in the image, the more adjectives are used to describe such an object and vice versa (Fig. 19).



a dark-skinned, dark-haired boy in a blue tee-shirt is standing in a light brown, dry, rocky desert landscape;

a brown cathedral with two towers and three green doors; a square with street lamps, green spaces, flowers, a tree, benches and people in front of it; grey cobblestones in the foreground; a hill and clouds in the background;

Figures 19: Examples for the use of adjectives

Use of Colour Attributes: Most of the annotation nouns have received at least one colour attribute if the pattern was not too complicated. However, the use of colour attributes for nouns in image annotations is not as trivial as it might seem. The colour value of a pixel is usually stored using 24 bits in the RGB colour space which means that there are more than 16 million possible colour values for each pixel. Although the perceptual ability of humans allows a much lower level of granularity for the visual differentiation of colour, there exist an immense number of colour names for ever so slightly different shades, saturations or intensities of colours (see Coloria[20] for a very impressive list and representation of many colour names in several languages).

Consequently, the more colour names are used in annotations, the smaller the difference between the colour names and therefore the harder it will be to provide a consistent use of colour attributes among all the annotations. This is further made difficult by the fact that one and the same colour can appear to be different in many images due to different surrounding colours.

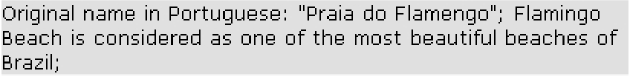
It is also known (Berlin & Kai, 1969) that significant differences exist between naming colours in different languages and cultures. For example, a kind of sea green, called “aoi” in Japanese, in English is generally regarded as a shade of "green", while in Japanese what an English speaker would identify as “green” can be regarded as a different shade of the kind of “sea green”.

A study by Berlin and Kay (1969) has shown that there are substantial regularities in naming colours across many languages. In the study, a concept of the following basic colour terms has been identified: black, grey, white, pink, red, orange, yellow, green, blue, purple and brown. All other colours are considered to be variants of these basic colours.

Due to these reasons, colour attributes are just using the aforementioned eleven basic colour terms. Variations in intensity are expressed by adding the labels *light* and *dark* (like “a *dark* green palm tree”). The suffix *–ish* is used if the colour is similar to one of the base colours (“a *greenish* palm tree”). Objects with a colour between two basic colour terms are described with a combination of the two (like “a *yellowish-orange* drink”).

3.3.4. Notes

This field contains additional free-text information about images such as background information and these fields do not follow any underlying patterns or annotation rules.



Original name in Portuguese: "Praia do Flamengo"; Flamingo Beach is considered as one of the most beautiful beaches of Brazil;

Figure 20: the notes field of image 16019

This can include information like original names in other languages (Fig. 20), historical information, eventual results of sports events (Fig. 21) or any other description that is not visible in the image and requires prior or deeper knowledge of the image contents.

[20] http://www.coloria.net/bonus/colornames.htm

The National Palace was built in Versaillan style from 1960 to 1992;
Vodafone Arena; Nikolay Davydenko (RUS) vs Tim Henman (GBR) 6-4 6-2 6-2;

Figures 21: Examples for historical and sports events

Not all images have note fields. In fact, just 10.3 % of the images hold additional, non-visible information, with an average length of 11.88 words per notes field and a standard deviation of 7.99. The longest notes field contains 55 words, the shortest just one, with a median of eleven words (see Table 6).

Number of Words	English	German	Spanish
average	11.88	10.84	N/A
standard deviation	7.99	7.26	N/A
minimum	1	1	N/A
median	11	9	N/A
maximum	53	59	N/A

Table 6: Word statistics for the notes field.

3.3.5. Locations

The location field describes the place where the image has been taken and is divided into two parts: (1) the exact location (e.g. Salvador) and (2) the country where this location belongs to (e.g. Brazil). Some images (2.35 %) only have country information in cases where the exact location in that country could not be verified.

Location names are stored in three languages. The question of whether place names are to be translated or not is a special challenge in se as there is no general answer for this question. While most countries do have their own version in each of the three languages like “Brazil” (English), “Brasilien” (German) and “Brasil” (Spanish), there is no pattern as to whether, for example city, names are translated or not. In many cases it is true that the more unknown a place is, the less likely it will be translated into a foreign language. However, this rule of thumb is not always applicable. Consider the places Rome and Buenos Aires for example, both big and famous cities: the Argentine capital is the same in all the three languages (“Buenos Aires”), whereas the Italian capital has a different version in each of the languages: “Rome” in English, “Rom” in German and “Roma” in Spanish. Hence, since there is no general rule, each location or place had to be checked individually whether there is an official translation or not, no matter how big or famous the location.

3.3.6. Dates

The date field contains the date when the image was taken, with each of the languages having its own version and format: German (e.g. "2 Oktober 2004"), English (e.g. "2 October, 2004") and Spanish (e.g. "2 de octubre de 2004");

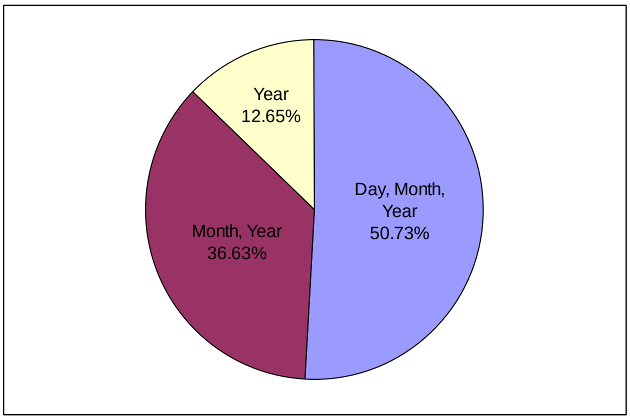


Figure 22: Percentages of the time granularity levels

There are three different time granularity levels: 51 % of the images have a complete date (day, month, year), 37 % contain have month and year, and 12 % of the annotation just state the year (see Fig. 22).

3.4. Generated Annotations

Annotations are stored in a database which is also managed by a benchmark administration system that allows the specification of parameters according to which different subsets of the image collection can be generated. Fig. 23 shows an example of an annotation format generated for ImageCLEF.

```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>Flamingo Beach</TITLE>
<DESCRIPTION> a photo of a brown sandy beach;
the dark blue sea with small breaking waves
behind it; a dark green palm tree in the
foreground on the left; a blue sky with clouds
on the horizon in the background;
</DESCRIPTION>
<NOTES> Original name in Portuguese: "Praia
do Flamengo"; Flamingo Beach is considered as
one of the most beautiful beaches of Brazil;
</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>2 October 2002</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figure 23: The generated English annotation file

Since the annotations are saved in three languages, one of these parameters is the annotation language. The annotation files can, at this stage, be generated in three different languages (and it is also possible to randomly select the annotation language). Figures 24 and 25 show the German and Spanish equivalents to the English annotation in Fig. 23.

```
<DOC>
<DOCNO>annotations/16/16019.ger</DOCNO>
<TITLE>Der Flamingostrand</TITLE>
<DESCRIPTION> ein Photo eines braunen
Sandstrands; das dunkelblaue Meer mit kleinen
brechenden Wellen dahinter; eine dunkelgrüne
Palme im Vordergrund links; ein blauer Himmel
mit Wolken am Horizont im Hintergrund;
</DESCRIPTION>
<NOTES> Originalname auf portugiesisch:
"Praia do Flamengo"; Der Flamingostrand gilt
als einer der schönsten Strände Brasiliens;
</NOTES>
<LOCATION>Salvador, Brasilien</LOCATION>
<DATE>2 Oktober 2002</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figure 24: The generated German annotation file

```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>La Playa del Flamenco</TITLE>
<DESCRIPTION> una foto de una playa marrón;
el mar azul oscuro con pequeñas olas que están
quebrando detrás; una palmera de color verde
oscuro en primer plano a la izquierda; un
cielo azul con nubes en el horizonte al fondo;
</DESCRIPTION>
<NOTES>Nombre original en portugués: "Praia do
Flamengo"; La Playa del Flamenco es
considerado una de las playas más bonitas de
Brasil; </NOTES>
<LOCATION>Salvador, Brasil</LOCATION>
<DATE>2 de octubre de 2002</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figure 25: The generated Spanish annotation file

Other parameters of the flexible annotation generation module of the Benchmark Administration System include (1) a range of annotation formats, (2) the level of annotation quality by suppressing the generation of certain fields, (3) varying levels of location information and (4) the introduction of spelling mistakes.

4. IAPR TC-12 Benchmark at ImageCLEF

The IAPR TC-12 Benchmark will be used for an ad-hoc image retrieval task at ImageCLEF, the text and/or content-based image retrieval track of CLEF from 2006 onwards.

4.1. Introduction to ImageCLEF

ImageCLEF conducts evaluation of cross-language image retrieval and is run as part of the CLEF campaign. The ImageCLEF retrieval benchmark has previously run in 2003 with the aim of evaluating image retrieval from English document collection with queries in a variety of languages. ImageCLEF 2004 added a visual retrieval task on a medical image collection and increased the participation from the visual retrieval community. ImageCLEF 2005 (Clough et al, 2005) provided tasks for system-centred evaluation of retrieval systems in two domains: historic photographs and medical images. These domains offer realistic scenarios in which to test the performance of image retrieval systems and offer various challenges and problems to participants. One purely visual task was offered on the automatic annotation of medical images. An interactive image retrieval tasks was also offered.

The ImageCLEF benchmark aims to evaluate image retrieval from multilingual document collections and a major goal is to investigate the effectiveness of multimodal retrieval (visual image features and textual description combined). ImageCLEF has already seen participation from both academic and commercial research groups worldwide from communities including the following: Cross-Language Information Retrieval (CLIR), Content-Based Image Retrieval (CBIR), medical information retrieval and user interaction. Campaigns such as CLEF and TREC have proven invaluable in providing standardised resources for comparative evaluation for a wide range of retrieval tasks and ImageCLEF aims to provide the research community with similar resources for image retrieval.

4.2. ImageCLEF 2006

ImageCLEF has been provided with a subset of the IAPR TC-12 Benchmark for its upcoming evaluation event (ImageCLEF 2006[21]) for a task concerning the ad-hoc retrieval of images from photographic image collections (called ImageCLEFphoto). Participants are provided with the full collection of 20,000 images; however they will not receive the complete set of annotations, but a range from complete annotations to no annotation at all. Data will be provided in English and German in order to enable the evaluation of multilingual text-based retrieval systems. In addition to the existing text and/or content based cross-language image retrieval task, ImageCLEF will also use the IAPR TC-12 Benchmark in an extra task for content-based image retrieval.

Other tasks offered in ImageCLEF 2006 include:

- an interactive retrieval evaluation using a database provided by FlickR;
- a medical image retrieval task with a database in three languages and varied annotation;
- a medical automatic annotation task (or image classification).
- a non-medical image annotation task (object recognition).

4.3. ImageCLEF 2007 and onwards

ImageCLEF has also expressed interest in having just one text annotation file with a randomly selected language for each image for ImageCLEF 2007, making full use of the benchmark's parametric nature.

[21] http://ir.shef.ac.uk/imageclef/2006/

Based on the discussions at the ImageCLEF workshop, the exact format of the benchmark will be decided as the most important goal is to include the research community into the task development process.

5. Conclusion

Publicly available benchmark efforts are an important part of research fields that are growing up. The goal is to ease for researchers the effort of evaluation of their algorithms and to provide a platform for information exchange and discussions among researchers. Sometimes these efforts are even done on a national level (ImageEval[22], France) to supply active researchers with a common evaluation structure for their algorithms. If benchmarks are well made according to the needs of researchers, the participation will follow.

An important part of the benchmark is the dataset and this is certainly no exception in the case of visual information systems. The benefits of the collection described in this paper are:

- high-quality colour photographs;
- pictures from a range of subjects and settings;
- high-quality multilingual text annotations which together make the collection suitable to evaluate a range of tasks;
- no copyright restrictions enabling the collection to be used in general by the research community.

It is recognised that benchmarks are not static as the field of visual information search might (and will) develop, mature and/or even change. Consequently, benchmarks will have to evolve and be augmented with additional features or characteristics depending on the researchers needs, and the IAPR TC-12 Benchmark will be no exception here. Apart from the planned completion of annotations in Spanish, and a possible extension to other annotation languages like French, Italian or Portuguese, the addition of several different annotation formats following a structured annotation defined in MPEG-7, an ontology-based keyword annotation (Hanbury, 2006) or even non-text annotations like an audio annotation are viable.

The method of generating various types of visual information might produce different characteristics in the future, and databases might have to be searched in different ways accordingly. Hence, benchmarks with several different component sets geared to different requirements will be necessary, and the parametric IAPR TC-12 Benchmark has taken a significant step towards that goal.

The IAPR TC-12 collection is also targeting an important market, that of personal picture collections. While desktop search for text is becoming a common utility, the search in private picture collections is still awaiting easy-to-use tools. With the large majority of pictures now taken in digital form, this is a field that is very likely to develop, creating a need for well-performing tools. ImageCLEFphoto can be a first test for such algorithms to prove their performance for real-world use.

6. References

Berlin B. & Kay P. (1969). Basic Color Terms: Their Universality and Evolution. *University of California Press.*

Braschler, M & Peters, C (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval* 7(1-2): pp. 7 - 31.

Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J. & Hersh, W. (2006). The CLEF 2005 Cross-Language Image Retrieval Track. *In Proceedings of the Cross Language Evaluation Forum 2005,* Springer Lecture Notes in Computer Science - to appear.

Grubinger, M. & Leung, C. (2003). A Benchmark for Performance Calibration in Visual Information Search. In *Proceedings of The 2003 International Conference on Visual Information Systems (VIS 2003),* Miami, FL, USA, pp. 414 – 419.

Grubinger, M. & Leung, C. (2004). Incremental Benchmark Development and Administration. In *Proceedings of The Tenth International Conference on Distributed Multimedia Systems (DMS'2004), Workshop on Visual Information Systems (VIS 2004),* San Francisco, CA, USA, pp. 328 – 333.

Grubinger, M., Leung, C. & Clough, P. (2005). The IAPR Benchmark for Assessing Image Retrieval Performance in Cross Language Evaluation Tasks. In *Proceedings of MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Vienna, Austria, pp. 33 - 50.

Hanbury, A. (2006). Analysis of Keywords in Image Understanding Tasks. In *Proceedings of the OntoImage workshop at the International Conference on Language REsources and Evaluation (LREC) – to appear.*

Harman, D. (1996). Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference (TREC-4),* Gaithersburg, MD, USA.

Jörgensen, C. (1996). The applicability of existing classification systems to image attributes: A selected review. *Knowledge Organisation and Change,* 5, pp. 189 – 197.

Jörgensen, C. (2001). Towards an image test bed for benchmarking image indexing and retrieval systems. In *Proceedings of the International Workshop on Multimedia Content–Based Indexing and Retrieval,* Rocquencourt, France.

Leung, C. & Ip, H. (2000). Benchmarking for Content-Based Visual Information Search. In *Proceedings of the Fourth International Conference on Visual Information Systems (VISUAL'2000),* Lyon, France: Springer Verlag, pp. 442 – 456.

Markkula, M., Tico, M., Sepponen, B., Nirkkonen, K., Sormunen, E. (2001). A Test Collection for the Evaluation of Content-Based Image Retrieval Algorithms—A User and Task-Based Approach, *Information Retrieval* 4(3-4), pp. 275 – 293.

Müller, H., Müller, W., Squire, DM., Marchand-Maillet, S. & Pun, T. (2001), Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals, *Pattern Recognition Letters (Special Issue on Image and Video Indexing), 22(5).* H. Bunke and X. Jiang Eds. pp. 593 - 601

Müller, H., Marchand-Maillet, S., Pun, T. (2002). The truth about Corel – evaluation in image retrieval. In

[22] http://www.imageval.org/

Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002), Springer Lecture Notes in Computer Science (LNCS 2383) London, England, pp. 38-49.

Narasimhalu, AD., Kankanhalli, MS. & Wu, J. (1997). Benchmarking Multimedia Databases, In *Multimedia Tools and Applications* 4, pp. 423 - 429.

O'Connor, B., O'Connor, M., Abbas, J. User Reactions as Access Mechanism: An Exploration Based on Captions for Images. *Journal of the American Society For Information Science*, 50(8), pp 681-697.

Over, P., Leung, C., Ip, H. & Grubinger, M. (2004). Multimedia Retrieval Benchmarks. *Digital Multimedia on Demand, IEEE Multimedia April-June 2004*, pp. 80 - 84.

Smeaton, AF., Kraaij, W. & Over, P. (2004). The TREC VIDeo Retrieval Evaluation (TRECVID): A Case Study and Status Report. In *Proceedings of RIAO 2004*, pp .

Smith, JR. (1998). Image Retrieval Evaluation *IEEE Workshop on Content-based Access of Image and Video Libraries*, Santa Barbara, California, USA, pp 112-113.

Tam, A. & Leung, C. (2001). Structured Natural-Language Descriptions for Semantic Content Retrieval of Visual Materials. *In Journal of the American Society for Information Science and Technology*, 52(11), pp. 930 – 937.

Analysis of Keywords used in Image Understanding Tasks

Allan Hanbury

Pattern Recognition and Image Processing Group (PRIP)
Institute of Computer-Aided Automation
Favoritenstraße 9/1832, A-1040 Vienna, Austria
hanbury@prip.tuwien.ac.at

Abstract

In the field of computer vision, automated image annotation and object recognition are currently important research topics. It is hoped that these will lead to improved general image understanding which can be usefully applied in Content-based Image Retrieval. In this paper, an analysis of the keywords that have been used in automated image and video annotation research and evaluation campaigns is presented. The outcome of this analysis is a list of 525 keywords divided into 15 categories. Given that this list is collected from existing image annotations, it could be used to check the applicability of ontologies describing entities which are portrayable in images.

1. Introduction

The usual reason to annotate data (i.e. add metadata to it) is to simplify access to it. This is particularly important for the semantic web. The metadata added to documents or images allow for more effective searches. The problem with adding metadata manually is that it is an extremely labour-intensive and time-consuming task. In the field of computer vision, automated image annotation and object recognition are currently important research topics (Barnard et al., 2003; Carbonetto et al., 2004; Csurka et al., 2004; Li and Wang, 2003; Winn et al., 2005). This automatic generation of image metadata should allow image searches and Content-Based Image Retrieval (CBIR) to be more effective. For example, an image database could be annotated offline by running a keyword annotation algorithm. Every image containing a cup would then have the keyword “cup” associated with it. If a user wishes to find images of a specific cup in this database, he/she would select a region containing the target cup from an image. An object recognition algorithm could then categorise the selected region as a cup and a text search could be carried out to find all images in the database with an associated keyword “cup”. This would significantly reduce the number of images in which it would be necessary to attempt to recognise the specific cup selected by the user.

To measure progress towards successfully carrying out this task, evaluation of algorithms which can automatically extract this sort of metadata is required. For successful evaluation of these algorithms, reliable ground truth is necessary. This ground truth should be a semantically rich description of the objects in an image (Leung and Ip, 2000). There is obviously almost no limit to how semantically rich one could make the description of an image. Indeed, for manual annotation of such documents destined to aid in online searching for them, semantic richness is an advantage. For images, one can create complex ontologies allowing the specification of objects and actions. For example, Schreiber et al. (2001) create such an ontology for annotating photographs of apes. One can specify the type of ape, how old it is and what it is doing. Nevertheless, it should be borne in mind that the automated content description and annotation algorithms being developed cannot yet be expected to perform at the same level as a human annotator. The current state-of-the-art in automated annotation tends to operate at an extremely low level — for example, there is still no algorithm that can make an error-free distinction between images of cities and images of landscapes, or which can make an error-free decision as to the presence or absence of human faces in an image.

Evaluating the abilities of current algorithms requires a rather low level of annotation. Even though different modalities of annotation exist, such as description using keywords, annotations based on ontologies and free text description, the majority of these annotations are done by assigning keywords to images. For object recognition tasks, controlled vocabularies are often used, with the vocabulary being defined by the capabilities of the object recognition algorithm used (Winn et al., 2005). In applications which aim to do a more general image labelling using a larger number of keywords, the vocabulary is often uncontrolled, as in (Li and Wang, 2003). For example, the TRECVID 2005 high-level feature detection task tested automatic detection of only 10 concepts. The IBM MARVEL Multimedia Search Engine[1] extracts only six concepts in the online image retrieval demo version[2] (face, human, indoor, outdoor, sky, nature). Carbonetto et al. (2004) use a vocabulary of at most 55 keywords. The largest number of keywords have been used by Li and Wang (2003), who assigned 433.

A good way of collecting keywords which would be useful in an ontology describing images is to analyse the vocabularies used in the ground truth of image annotation and object recognition tasks. In this way, one can find out which words are important in applications and which words correspond to objects which can be detected using current state-of-the-art image understanding algorithms. After an overview of some approaches to collecting manual image annotations (Section 2.), we analyse the annotations which have been used in image and video understanding publications and evaluation campaigns in Section 3. The list of collected keywords is at the end of the paper in Section 6.

[1] http://www.research.ibm.com/marvel

[2] http://www.alphaworks.ibm.com/tech/marvel

2. Manual annotation collection methods

The manual annotation of images is a very labour-intensive and time-consuming task. Various systems to simplify the collection of image annotations or to receive input from a large number of people have been set up.

An interesting experiment is taking place on the *Gimp-Savvy Community-Indexed Photo Archive* website[3]. This archive contains more then 27 000 free photos and images, and the users of the site are requested to annotate the images using keywords which they are free to choose (tips on choosing keywords are made available[4]). That this "free annotation by all" approach has not been totally successful can be seen by the extremely large number of "junk" keywords on the master list[5] as well as the over-annotation (assignment of too many keywords) of many of the images.

On the *Flickr*[6] photo archive, people who upload photos may also assign keywords to them. These are then used to search for images. Other users may add comments to the images. There is no standardised keyword list, so this database represents a good example of the annotation practice of amateur photographers on their own images.

An innovative approach to collecting annotations of images by keywords has been developed by Ahn and Dabbish (2004). In their ESP game[7], they aim to make the annotation of images enjoyable. Players access the ESP game server and are paired randomly. They have no way of communicating with each other. Pairs of players are shown 15 images during the game, with the aim being for both players to type in the same keyword for an image so as to advance to the next. This is an intelligent way of avoiding the problem of "junk" keywords, as the pairs of players verify the keywords. Keywords which are typed often for an image are added to a "taboo" list shown for that image, and can no longer be entered as keywords by the players. The keywords entered correspond to the whole image, although the authors have discussed implementing, for example, a "shooting game", where the players have to click on the requested object. The Peekaboom game[8] from the same research group is of this type. An image search engine based on the keywords collected from the ESP game for about 30 000 images is accessible on the web[9].

An online annotation application aimed at collecting keywords for image regions is the LabelMe tool[10]. Here the user clicks the vertices of a polygon around an object and then enters a keyword describing the object. As the vocabulary is not controlled, multiple keywords and misspelled keywords often occur, as can be seen by examining the keyword statistics on the webpage[11]. This problem is solved by a verification step by the database administrators. At present[12], there are 101 verified keywords, the majority of which are shown in Table 2. The incentive to annotate the images is that the annotator may then download the latest annotations.

3. Analysis of Keywords used in Annotation Experiments

In this section we analyse the keywords that have been used in image annotation, categorisation and object recognition experiments and evaluation campaigns. To begin, a brief discussion on the difference between annotation and categorisation is presented in Section 3.1. Some methods currently used for collecting manual annotations of images are listed in Section 2. We then present an analysis of the keywords that have been used in image annotation experiments. The analysis was carried out in two steps. The first step consisted of creating a list combining all the keywords used in the experiments, datasets and evaluations considered and removing the unsuitable words (Section 3.2.). The second step was the categorisation of keywords (Section 3.3.). From a practical point of view, it is useful if the keywords are sorted into categories. When one is annotating images, this simplifies the choice of a word from the keyword list — one can select the category that the image belongs to in order to reduce the choice of keywords. The result of this analysis is a list of 525 keywords assembled from various sources and divided into 15 categories.

3.1. Annotation and Categorization

There are two approaches to associating textual information with images described in the literature: *annotation* and *categorisation*. In annotation, keywords or detailed text descriptions are associated with an image, whereas in categorisation, each image is assigned to one of a number of predefined categories (Chen and Wang, 2004). This can range from more general two category classification, such as *indoor/outdoor* (Szummer and Picard, 1998) or *city/landscape* (Vailaya et al., 2001) to more specific categories such as *African people and villages*, *Dinosaurs*, *Fashion* and *Battle ships* (Chen and Wang, 2004). Categorisation can be used as an initial step in image understanding in order to guide further processing of the image. For example, in (Wang et al., 2001) a categorisation into textured/non-textured and graph/photograph classes is done as a pre-processing step. *Recognition* is concerned with the identification of particular object instances. Recognition would distinguish between images of two structurally distinct cups, while categorisation would place them in the same class (Csurka et al., 2004). Recognition also has its uses in annotation, for example in the recognition of family members in the automatic annotation of family photos.

Categorisation can be considered as annotation in which one must choose from a fixed number of keywords (the categories) and one is limited to assigning one keyword to each image. The discussion of annotation and categorisation is therefore combined in this section.

[3] http://gimp-savvy.com/PHOTO-ARCHIVE/

[4] http://gimp-savvy.com/PHOTO-ARCHIVE/tips_on_indexing.html

[5] http://gimp-savvy.com/cgi-bin/masterkeys.cgi

[6] http://www.flickr.com

[7] http://www.espgame.org

[8] http://www.peekaboom.org/

[9] http://www.captcha.net/esp-search.html

[10] http://people.csail.mit.edu/brussell/research/LabelMe/intro.html

[11] 400 keywords on the 29th of July 2005.

[12] 27 July 2005

3.2. Overview of Visual Keywords

We present a collection of groups of keywords which have already been used for testing automated image annotation algorithms or in automated image and video annotation evaluation campaigns.

The 10 features which were tested in the TRECVID 2005 high-level feature detection task are described in Table 1. All 40 news concepts defined for TRECVID 2005 are available for download[13] (they are part of the LSCOM creation task (Hauptmann, 2004)).

Two categorisation tasks are part of the ImagEVAL[14] campaign: for the general image description task, the hierarchically organised global image categories shown in Figure 1 will be tested. There is also an object detection task, although the list of objects to be tested has not been finalised yet. The examples given are car, tree, chair, Eiffel Tower and American Flag.

The PASCAL Visual Object Classes Challenge 2005 consisted of classification and detection tasks for four objects: motorbikes, bicycles, people and cars. However, in the database collection set up as part of this challenge[15], five databases are provided with standardised ground truth object annotations. The keyword list arising from this standardisation is shown in Table 2.

As part of the EU LAVA project[16], a database consisting of 10 categories of images was made available[17]. These categories are: bikes, boats, books, cars, chairs, flowers, phones, roadsigns, shoes and soft toys.

Chen and Wang (2004) classified images into 20 categories: African people and villages, Beach, Historical buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and glaciers, Food, Dogs, Lizards, Fashion, Sunsets, Cars, Waterfalls, Antiques, Battle ships, Skiing and Deserts.

Two databases have been released by Microsoft Research in Cambridge[18]. The "Database of thousands of weakly labelled, high-res images" contains images divided into the following 23 categories: aeroplanes, cows, sheep, benches and chairs, bicycles, birds, buildings, cars, chimneys, clouds, doors, flowers, forks, knives, spoons, leaves, countryside scenes, office scenes, urban scenes, signs, trees, windows, miscellaneous. Some of these are divided into sub-classes, such as different views of cars. The "Pixel-wise labelled image database" contains 591 images in which regions are manually labelled using the following 23 labels: building, grass, tree, cow, horse, sheep, sky, mountain, aeroplane, water, face, car, bicycle, flower, sign, bird, book, chair, road, cat, dog, body, boat. The majority of the images are roughly segmented, although accurate segmentations of some of the images are available.

It is, of course, possible to greatly extend the number of categories if one is recognising specific objects, such as in the Caltech 101 category database[19] (Fei-Fei et al., 2004), which contains images of objects in the categories shown in Table 3.

If one restricts oneself to such specific categories, it is obviously possible to create many thousands. A set of 16 broader categories has been defined for the 15 200 images in the CEA-CLIC database (Moëllic et al., 2005). These are shown in Table 4.

A number of papers on automatic image or image region annotation have also been published. The following three all use parts of the Corel image database along with keywords usually extracted from the annotations accompanying the Corel images. The 55 keywords used by Carbonetto et al. (2004) are given in Table 5. Li and Wang (2003) used the largest number of keywords. They defined 600 categories of image, and to each category assigned on average 3.6 keywords. Each of the 100 images in each category was then assigned the same keywords associated with the category. For example, all images in the "Paris/France" category were assigned the keywords "Paris, European, historical building, beach, landscape, water", the images in the "Lion" category were assigned the keywords "lion, animal, wildlife, grass" and the images in the "eagle" category were assigned the keywords "wildlife, eagle, sky, bird". Barnard et al. (2003) used 323 keywords. These lists are not reproduced in this paper due to lack of space, but can be seen in (Hanbury, 2006).

3.3. Analysis of Visual Keywords

The aim of this analysis is to create a list of keywords which reflect the current interest in automated image annotation with keywords. These keywords could then serve as an initial controlled vocabulary for re-annotating the image collections used in previous experiments and for annotating new image collections.

3.3.1. Creation of a combined keyword list

The first step of the analysis consisted of creating a list combining all the keywords and categories used in the experiments, datasets and evaluations covered in Section 3.2. We then removed words which were considered to be unsuitable. These include place names, such as "Australia", "Boston" and "New Zealand", which, even for a human, are very difficult to assign to images for which one has no supplementary information. Confusing keywords, such as "history" and "north", and keywords requiring too high a level of a priori semantic information, such as "landmark" and "rare animal" were also removed. We have not yet collected statistics on how often a single keyword appears in different lists.

3.3.2. Categorisation of keywords

From a practical point of view, it is useful if the keywords are sorted into categories. When one is annotating images, this simplifies the choice of a word from the keyword list —

[13] http://www-nlpir.nist.gov/projects/tv2005/LSCOMlite_NKKCSOH.pdf

[14] http://www.imageval.org

[15] http://www.pascal-network.org/challenges/VOC/

[16] http://www.l-a-v-a.org

[17] ftp://ftp.xrce.xerox.com/pub/ftp-ipc/

[18] Downloadable here: http://www.research.microsoft.com/vision/cambridge/recognition/default.htm. Version 1 of the pixel-wise labelled image database has been ignored here, as it forms a subset of version 2.

[19] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

Keywords	Segment contains video of ...
People walking/running	more than one person walking or running
Explosion or fire	an explosion or fire
Map	a map
US flag	a US flag
Building exterior	the exterior of a building
Waterscape/waterfront	a waterscape or waterfront
Mountain	a mountain or mountain range with slope(s) visible
Prisoner	a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.
Sports	any sport in action
Car	an automobile

Table 1: The 10 features which were tested in the TRECVID 2005 high-level feature detection task.

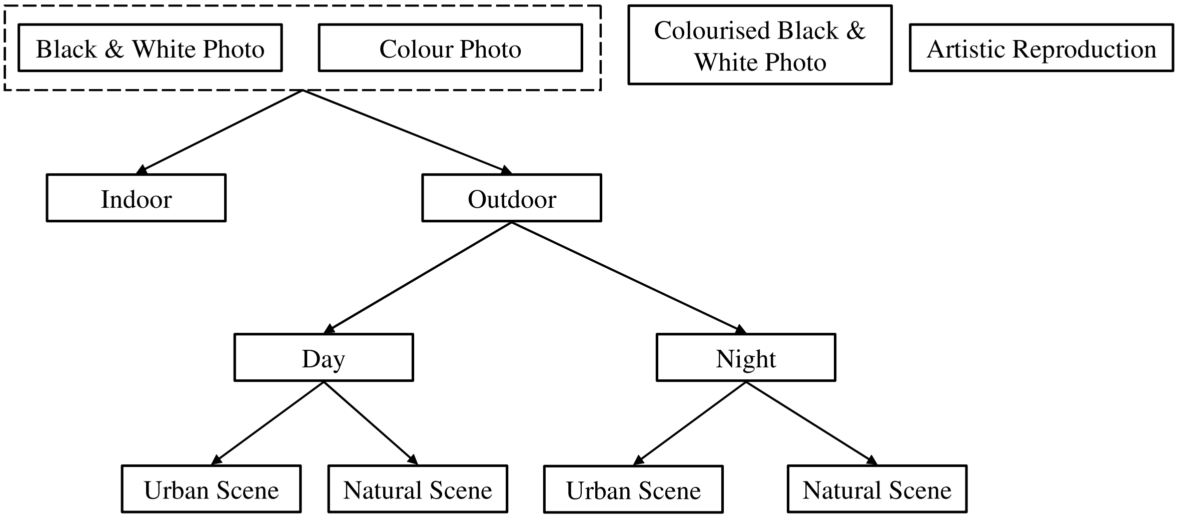


Figure 1: The hierarchy of keywords used in the global image characteristics task of ImagEVAL.

one can select the category that the image belongs to in order to reduce the choice of keywords. The 16 categories of the CEA-CLIC database (Moëllic et al., 2005), with some minor changes, turn out to be well-suited to grouping the combined list of keywords.The changes are:

- the fusion of the “Architecture” and “City” categories to form an “Architecture / City” category. This was done as it is often difficult for an annotator to decide between these two categories.
- the addition of an “Abstract / Global” category to contains words such as “female” and “exterior”.
- the removal of the “Mathematics” category, which has no members in the list of keywords collected.
- the removal of the “linguistic” category, as this is an image category and not a keyword category.
- the addition of the “Anatomy and Medicine” category, which at present includes one keyword, but can be expanded later.

The list of categories and their descriptions are in Table 6. We assigned each of the keywords in the combined list to at least one category. A few keywords were assigned to two categories, for example, “grass” appears in the “Texture” and “Nature and Landscapes” categories. A table showing the keywords assigned to each category is given in Section 6. A histogram of the number of keywords per category is shown in Figure 2.

One can see from this histogram that the categories “Objects”, “Nature and Landscapes” and “Zoology” contain the most keywords, which could be an indicator that these categories have received the most attention in past research on automated image annotation and categorisation. This could be because of the image databases used — the Corel databases, for example, appear to contain a high proportion of natural and animal images. The man-made objects appear to be more prevalent in the databases designed for object categorisation experiments.

4. Conclusion

We analyse the keywords which have been used to annotate images in a number of image retrieval publications and evaluation campaigns. A significant contribution is the creation of a combined keyword list based on these keywords. From this analysis one can see that the main automated annotation effort has been directed at images of everyday objects; nature and landscapes; and animals (zoology). As

aeroplaneSide	apple	background	bicycle	bicycleSide
bookshelf	bookshelfFrontal	bookshelfPart	bookshelfSide	bookshelfWhole
bottle	building	buildingPart	buildingRegion	buildingWhole
can	car	carFrontal	carPart	carRear
carSide	cd	chair	chairPart	chairWhole
coffeemachine	coffeemachinePart	coffeemachineWhole	cog	cow
cowSide	cpu	desk	deskFrontal	deskPark
deskPart	deskWhole	donotenterSign	door	doorFrontal
doorSide	face	filecabinet	firehydrant	freezer
frontalWindow	head	keyboard	keyboardPart	keyboardRotated
light	motorbike	motorbikeSide	mouse	mousepad
mug	onewaySign	paperCup	parkingMeter	person
personSitting	personStanding	personWalking	poster	posterClutter
pot	printer	projector	roadRegion	screen
screenFrontal	screenPart	screenWhole	shelves	sink
sky	skyRegion	sofa	sofaPart	sofaWhole
speaker	steps	stopSign	street	streetSign
streetlight	tableLamp	telephone	torso	trafficlight
trafficlightSide	trash	trashWhole	tree	treePart
treeRegion	treeWhole	walksideRegion	wallClock	watercooler
window				

Table 2: The keywords in the PASCAL Object Recognition Database Collection (the prefix “PAS” has been removed from each keyword).

Faces	Faces easy	Leopards	Motorbikes	accordion	airplanes
anchor	ant	barrel	bass	beaver	binocular
bonsai	brain	brontosaurus	buddha	butterfly	camera
cannon	car side	ceiling fan	cellphone	chair	chandelier
cougar body	cougar face	crab	crayfish	crocodile	crocodile head
cup	dalmatian	dollar bill	dolphin	dragonfly	electric guitar
elephant	emu	euphonium	ewer	ferry	flamingo
flamingo head	garfield	gerenuk	gramophone	grand piano	hawksbill
headphone	hedgehog	helicopter	ibis	inline skate	joshua tree
kangaroo	ketch	lamp	laptop	llama	lobster
lotus	mandolin	mayfly	menorah	metronome	minaret
nautilus	octopus	okapi	pagoda	panda	pigeon
pizza	platypus	pyramid	revolver	rhino	rooster
saxophone	schooner	scissors	scorpion	seahorse	snoopy
soccer ball	stapler	starfish	stegosaurus	stop sign	strawberry
sunflower	tick	trilobite	umbrella	watch	water lilly
wheelchair	wildcat	windsor chair	wrench	yin yang	

Table 3: The 101 categories used by Fei-Fei et al. (Fei-Fei et al., 2004).

these keywords were extracted from annotations of existing image datasets, they should be well-suited to a more precise re-annotation of these same datasets. For the same reason, they are also suited to verify the applicability of newly developed image ontologies intended to represent portrayable entities and objects.

A disadvantage is that while the keywords in this list certainly correspond well to the images used in image annotation experiments so far, there is no guarantee that these images are representative of all possible electronic images. It would therefore be useful to compare this collection of keywords to an ontology constructed in a more rigorous way, such as the ontology of portrayable objects based on WordNet (Zinger et al., 2005). This should provide a useful link between possible portrayable objects and those that are often found in images, or that are of interest to image understanding researchers.

5. References

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proc. ACM CHI*, pages 319–326.

Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Category	Description
Food	Images of food, and meals.
Architecture	Images of architecture, architectural details, castles, churches, Asian temples.
Arts	Paintings, sculptures, stained glass, engravings.
Botanic	Various plants, trees, flowers.
Linguistic	Images containing text areas.
Mathematics	Fractals.
Music	Images of musical instruments.
Objects	Images representing everyday objects such as coins, scissors, etc.
Nature & Landscapes	Landscapes, valley, hills, deserts, etc.
Society	Images with people.
Sports & Games	Stadiums, items from games and sports.
Symbols	Iconic symbols, roadsigns, national flags (real and synthetic images)
Technical	Images involving transportation, robotics, computer science.
Textures	Rock, sky, grass, wall, sand, etc.
City	Buildings, roads, streets, etc.
Zoology	Images of animals (mammals, reptiles, bird, fish).

Table 4: The 16 categories in the CEA-CLIC image database and their descriptions (Moëllic et al., 2005).

airplane	astronaut	atm	bear	beluga	bill	bird
boat	building	cheetah	church	cloud	coin	coral
cow	crab	dolphin	earth	elephant	fish	flag
flowers	fox	goat	grass	ground	hand	horse
house	lion	log	map	mountain	mountains	person
pilot	polarbear	rabbit	road	rock	sand	sheep
shuttle	sky	snow	space	tiger	tracks	train
trees	trunk	water	whale	wolf	zebra	

Table 5: The 55 keywords used by Carbonetto et al. (Carbonetto et al., 2004).

#	Category	Description
0	Abstract / Global	Words which describe the whole image or which are applicable to more than one class of objects.
1	Food	Food and meals.
2	Architecture / City	Architecture, architectural details, castles, churches, Asian temples, buildings, roads, streets, etc.
3	Arts	Paintings, sculptures, stained glass, engravings.
4	Botanic	Plants, trees, flowers.
5	Objects	Everyday objects such as coins, scissors, etc.
6	Nature & Landscapes	Landscapes, valley, hills, deserts, etc.
7	Society	People, groups of people, activities undertaken by society (celebrations, parades, war, etc.).
8	Sports & Games	Stadiums, items from games and sports.
9	Symbols	Iconic symbols, roadsigns, national flags
10	Technical	Transportation, robotics, computer science.
11	Textures	Words which describe a texture.
12	Zoology	Animals (mammals, reptiles, birds, fish).
13	Anatomy and Medicine	Biological organs, anatomical diagrams, etc.
14	Music	Musical instruments.

Table 6: The 15 categories of the combined keyword list and their descriptions. The first column contains a category number.

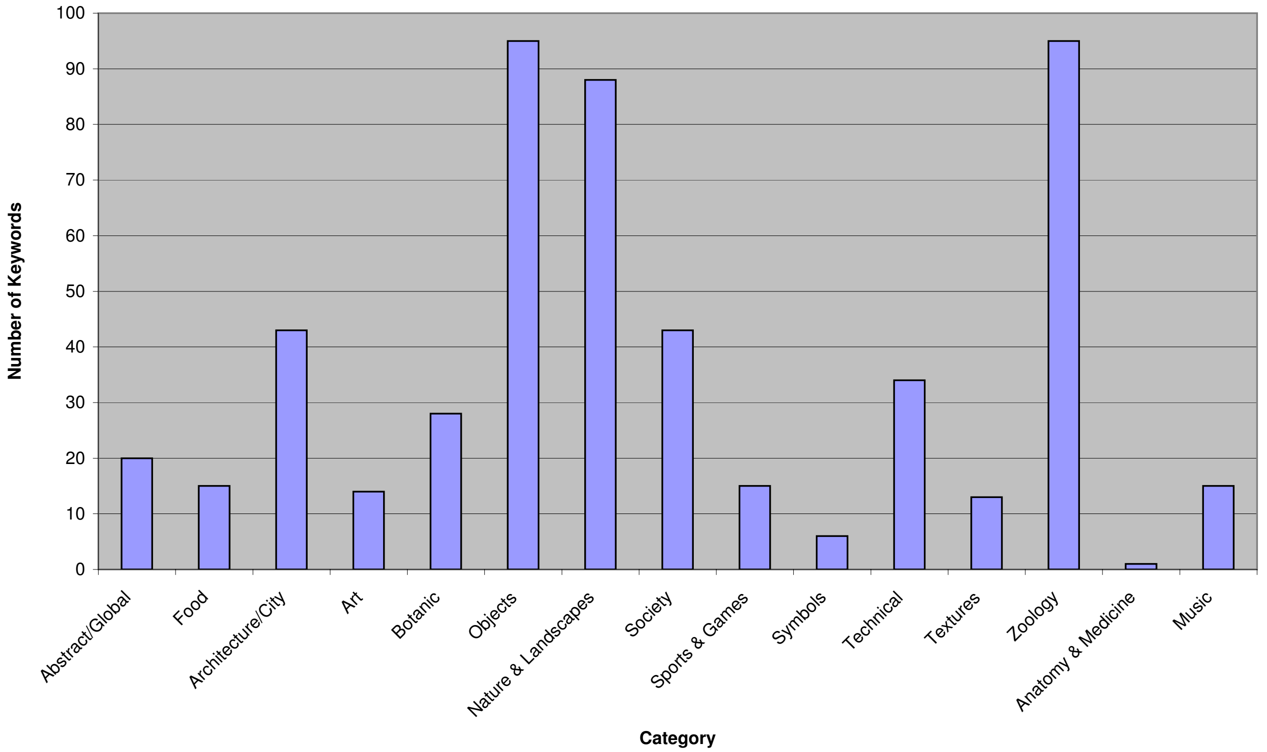


Figure 2: The number of keywords in each category.

Peter Carbonetto, Nando de Freitas, and Kobus Barnard. 2004. A statistical model for general contextual object recognition. In *Proceedings of the ECCV 2004, Part I*, pages 350–362.

Yixin Chen and James Z. Wang. 2004. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939.

Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision (at ECCV)*.

L. Fei-Fei, R. Fergus, and P. Perona. 2004. Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In *Proceedings of the Workshop on Generative-Model Based Vision*, June.

Allan Hanbury. 2006. Review of image annotation for the evaluation of computer vision algorithms. Technical Report PRIP-TR-102, PRIP, TU Wien, January.

Alexander G. Hauptmann. 2004. Towards a large scale concept ontology for broadcast video. In *Proceedings of the Third Intl. Conf on Image and Video Retrieval*, pages 674–675.

Clement H. C. Leung and Horace Ho-Shing Ip. 2000. Benchmarking for content-based visual information search. In *Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 442–456.

Jia Li and James Z. Wang. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088.

Pierre-Alain Moëllic, Patrick Hède, Gregory Grefenstette, and Christophe Millet. 2005. Evaluating content based image retrieval techniques with the one million images clic testbed. In *Proceedings of the Second World Enformatika Congress, WEC'05*, pages 171–174.

A. Th. (Guus) Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. 2001. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74.

M. Szummer and R. W. Picard. 1998. Indoor-outdoor image classification. In *Proc. IEEE International Workshop on Content-based Access of Image and Video Databases*, pages 42–51.

A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130.

James Z. Wang, Jia Li, and Gio Wiederhold. 2001. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

J. Winn, A. Criminisi, and T. Minka. 2005. Object categorization by learned universal visual dictionary. In *Proceedings of the International Conference on Computer Vision(ICCV)*.

S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, and P.-A. Moëllic. 2005. Extracting an ontology of portrayable objects from WordNet. In *Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pages 17–23, Vienna, Austria, September.

6. Combined Keyword List

The following table lists the combined keyword list. It is a simple two-level hierarchy, with 15 headings at the top level (in bold). Note that some words are repeated under more than one heading.

Abstract / Global				
background	black	black_and_white	blue	color
exterior	female	fractal	green	group
indoor	interior	male	nature	orange
outdoor	pattern	red	shadow	yellow

Food				
apple	cuisine	dessert	drink	feast
food	fruit	grapes	herb_spice	orange
pizza	pumpkin	strawberry	vegetable	wine

Architecture / City				
arch	architecture	building	castle	chimney
church	city	college	column	courtyard
dock	fountain	harbor	historical_building	hotel
house	hut	industry	kitchen	market
minaret	monument	mosque	museum	office
pagoda	palace	park	pillar	restaurant
roof	ruin	shop	skyline	stairs
statue	street	studio	temple	tower
town	village	window		

Art Objects				
art	carving	decoration	design	drawing
graffiti	mosaic	mural	painting	photo
poster	sculpture	statue	still_life	

Botanic				
apple	bonsai	botany	branch	bush
cactus	flower	foliage	fungus	grapes
leaf	lichen	log	moss	mushroom
orchid	palm	perenial	petal	plant
pumpkin	rose	seed	strawberry	sunflower
tree	tulip	water_lily		

Objects (man-made everyday)				
anchor	antique	atm	balloon	barbecue
barrel	bath	bead	bench	bicycle
binoculars	book	bookshelf	bottle	camera
can	candy	card	cd	cellphone
chair	clock	cloth	coffee_machine	cog
coin	cup	currency	decoration	desk
dish	dogsled	doll	door	dress

Easter_egg	fabric	fan	fence	file_cabinet
fire_hydrant	firearm	firework	flag	floor
freezer	furniture	glass	gun	hat
headphones	horn	jewelry	keyboard	lamp
light	map	marble	mask	medicine
money	mousepad	mug	paper	paper_cup
parking_meter	pill	pot	printer	projector
relic	scissors	screen	shelves	shoe
sink	sofa	speaker	sponge	stamp
stapler	table	telephone	textile	tool
toy	traffic_light	trash	umbrella	wall
watch	watercooler	wheelchair	wood	wrench

Nature and Landscapes				
agriculture	autumn	barnyard	bay	beach
canyon	cave	cliff	cloud	coast
coral	crop	crystal	dawn	desert
dune	dusk	earth	farm	field
flowerbed	forest	frost	frozen	garden
gem	glacier	grass	ground	hill
ice	iceberg	island	lake	landscape
maritime	meadow	mountain	night	ocean
pastoral	path	peak	plain	planet
polar	pyramid	rapids	reef	reflection
river	road	rock	ruin	runway
rural	sail	sand	shell	shore
shrine	sky	smoke	snow	space
spring	star	steam	stone	sub_sea
summer	sun	sunset	surf	tree
tropical	tundra	valley	vegetation	vineyard
volcano	wall	water	waterfall	wave
wind	winter	woodland		

Society				
astronaut	baby	ballet	barbecue	battle
builder	business	child	Christmas	costume
couple	diver	face	fashion	festival
fight	glamour	graffiti	guard	hand
head	holiday	home	hunter	leisure
man	model	occupation	parade	person
pilot	pomp_and_pageantry	religion	royal	sacred
science	travel	tribal	war	woman
work	worship	youth		

Sports and Games				
fitness	football	game	golf	kungfu
play	polo	race	rafting	recreation
rodeo	ski	sport	tennis	wind_surfer

Symbols

public_sign	road_sign	sign_do_not_enter	sign_stop	sign_oneway
sign_yield				

Technical				
aeroplane	aviation	balloon	battle_ship	boat
bridge	bus	cannon	canoe	car
communication	engine	ferry	helicopter	highway
jet	lighthouse	locomotive	machine	military
molecule	motorcycle	pathology	railroad	road
runway	sailboat	ship	space_shuttle	street
tallship	train	transportation	vehicle	

Textures				
fabric	fire	glass	grass	ground
ice	marble	sand	skin	stone
textile	texture	wood		

Zoology				
anemone	angelfish	animal	ant	antelope
antlers	bear	beaver	beetle	bird
bobcat	bull	butterfly	camel	caribou
cat	caterpillar	cheetah	coral	cougar
cow	coyote	crab	crayfish	crocodile
cub	deer	dinosaur	dog	dolphin
dragonfly	eagle	elephant	elk	feline
fish	flamingo	foal	fowl	fox
giraffe	goat	hawk	hedgehog	herd
hippopotamus	horn	horse	iguana	insect
jaguar	kangaroo	kitten	leopard	lion
lizard	llama	lobster	lynx	mammal
moth	mouse	nest	ocean_animal	octopus
owl	panda	penguin	pet	pigeon
polar_bear	predator	primate	rabbit	reptile
rhinoceros	rodent	rooster	scorpion	seahorse
seal	sheep	skin	snake	sponge
squirrel	starfish	tiger	turtle	whale
wildcat	wildlife	wolf	young_animal	zebra

Anatomy and Medicine				
brain				

Musical Instruments				
accordion	cello	double_bass	electric_guitar	guitar
horn	mandolin	piano	piano_grand	saxophone
trombone	trumpet	tuba	viola	violin

Automatically populating an image ontology and semantic color filtering

Christophe Millet[*][†], Gregory Grefenstette[*], Isabelle Bloch[†], Pierre-Alain Moëllic[*], Patrick Hède[*]

[*]CEA/LIST/LIC2M
18 Route du Panorama, 92265 Fontenay aux Roses, France
{milletc, grefenstetteg, moellicp, hedep}@zoe.cea.fr

[†] GET-ENST - Dept TSI - CNRS UMR 5141 LTCI
Paris, France
isabelle.bloch@enst.fr

Abstract

In this paper, we propose to improve our previous work on automatically filling an image ontology via clustering using images from the web. This work showed how we can automatically create and populate an image ontology using the WordNet textual ontology as a basis, pruning it to keep only portrayable objects, and clustering to get representative image clusters for each object. The improvements are of two kinds: first we are trying to automatically locate the objects in images so that the image features become independent of the context. The second improvement is a new method to semantically sort clusters using colors: the most probable colors for an object are learnt automatically using textual web queries, and then the clusters are sorted according to these colors. The results show that the segmentation improves the quality of the clusters, and that meaningful colors are often guessed, thus displaying pertinent clusters on top, and bad clusters at the bottom.

1. Introduction

Since available annotated image databases or ontologies are still only a few and are far from representing every object in the world, we are working on automatically constructing an image ontology using a textual ontology on the one hand, and the Internet as a huge but incompletely and inaccurately annotated image database on the other hand. Such approaches have been first proposed by (Cai et al., 2004) and (Wang et al., 2004).

(Wang et al., 2004) developed a method to automatically use web images for image retrieval. An attention map is used to find the object in an image, and the text surrounding the image is matched to the region level instead of the image level. Then, regions are clustered and each cluster is annotated using the text-region matching. Results are promising and can be improved with query expansion. (Cai et al., 2004) proposed to cluster images from the web using three kinds of representation: textual information extracted from the text and links appearing around the image in the web pages, visual features, and a graph linking the regions of the image. The application given is to show web image search results grouped into clusters instead of giving a list that mixes different topics. However, no work has been done to try to semantically sort clusters by relevance.

In this paper, we propose to improve our previous work (Zinger et al., 2006) on automatically filling an image ontology via clustering, first by trying to automatically locate the objects in images, and then by proposing a method to semantically sort clusters using colors. The skeleton of the image ontology is built using a textual ontology as a basis: WordNet[1].

Not all words are picturable objects, so this ontology has to be pruned before we try to fill its nodes with images. The next step is to get the images from the Web, try to isolate the object in these images so that it becomes independent of the context, and cluster them into coherent groups. In order to reduce the noise in images returned from the Web using textual queries, we can refine the query adding the category of the desired object. This will be described in Section 2.

Then, we would like to sort the obtained clusters in order to try to have the most relevant images first, and optionally to eliminate clusters that do not contain the expected object. We propose to apply a semantic color filtering. The idea is to give more importance to the images containing the probable colors of an object. For example, if we are querying for images of bananas, we are expecting to see yellow images first. A list of possible colors of an object is retrieved automatically from the web. We have also developed a matching between the name of colors and the HSV values of a pixel allowing us to compare the colors contained in an image with the possible colors of the object it is supposed to be depicting. This is explained in Section 3.

Eventually, we will discuss our results in Section 4.

2. Obtaining image clusters from the Web

2.1. Pruning Wordnet

The objects we are interested in are picturable objects. Some words such as *happiness* or *employment* are concepts that cannot really be pictured, so we have to prune the WordNet ontology in order to keep only the picturable objects. These objects are mostly the ones that can be found as being hyponyms of the node *physical objects* which has two definitions in WordNet: *a tangible and visible entity* and *an entity that can cast a shadow*. However, some of these hyponyms have to be removed manually because the WordNet ontology contains some inconsistencies. For example, *tree of knowledge* appears as a kind of *tree* which is an hyponym of *physical objects*. Once this pruning is completed, from the original 117097 nouns contained in the WordNet ontology, about 24000 leaves candidates for images are left.

[1] http://wordnet.princeton.edu/

2.2. Using the right set of keywords

Now that we have the skeleton of our ontology, we would like to populate the ontology with images from the web. In order to retrieve images from the web, we use text queries, such as Google or Yahoo! image search engines, where the name of the pictures and the text surrounding the pictures in the web pages have been used as a textual indexing. For some requests, we notice that the amount of noise can be quite important, and furthermore, we would like to disambiguate the query to obtain images representing only one object: asking for jaguar on an image search engine returns a mix of animals and cars because the word *jaguar* is polysemic. Here, the ontological information extracted from WordNet helps to obtain more accurate images. Adding an upper node of the ontology in the text queries allows disambiguating the query, and gives better results even for words that are not ambiguous. For the jaguar example, we will have two separate queries: *jaguar car* and *jaguar animal*. The precision is increased, but the recall is decreased: Google Image Search returns 3 750 images for *jaguar animal* and 40 100 images for *jaguar car*, which is to be compared with the 553 000 images returned for *jaguar* most of which are either animals or cars: we only obtain a tenth of the images.

2.3. Segmentation

Since we want to construct an ontology that can be used for learning, we are interested in images where the object we are looking for is big enough for image processing (the more pixels the better), but small enough to be entirely contained in the image: we do not want to add part of objects in the ontology, we want to add pictures of the whole object. Furthermore, we would like to index only the object of interest, without taking the context into account: a blue car on green grass, and a blue car on a gray road should be recognized as the same object.

We are making the following three hypotheses on the images:

- there is only one object in the image,
- the object is centered,
- its surface is greater than 5% of the image surface.

The method proposed here is to automatically segment the image and keep only the central object. The following steps are accomplished: the image is segmented into 20 regions using a waterfall segmentation algorithm (Marcotegui and Beucher, 2005), the regions touching the edges of the image are discarded, and the other regions are merged together. The largest connected region is considered as the object and used for further processing. Only the images where an object larger than 5% of the image in surface are kept.

2.4. Clusterisation

These segmented images are then clustered with the shared nearest neighbor method (Ertz et al., 2001), using texture and color features (Cheng and Chen, 2003). The shared nearest neighbor clustering algorithm is an unsupervised algorithm mostly used in text processing which tries to group together the images that have the same nearest neighbors. The texture features used are a 512bins local edge pattern histogram, and the color features are a 64bins color histogram. Clusters containing less than 8 images are discarded.

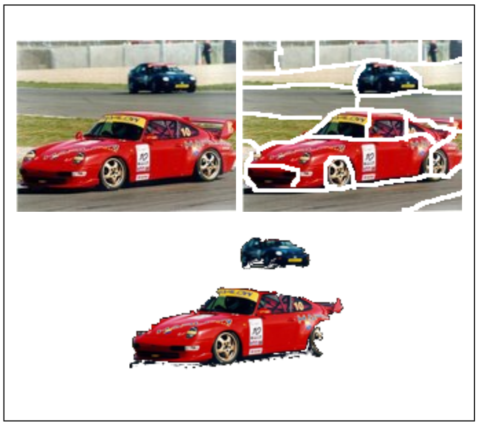


Figure 1: Automatic segmentation of a car image. The top left image is the original image. The top right image is the result of the segmentation in 20 regions. After removing the regions touching the edges of the image and merging the other regions we obtain the bottom image. There are two connected regions in this image, and we will keep the largest one corresponding to the red car. Here, the second connected region is also a car, but it is often noise.

We have noticed that the segmentation step improves the quality of the clusters for most queries, mostly because it makes it independent of the context. We will show some examples in Section 4.

3. Color sorting

3.1. Obtaining the colors of images

The HSV (hue, saturation, value) color space is used, as it is more semantic than RGB, and therefore makes it easier to deduce the color name of a pixel. Each component of the HSV space has been scaled between 0 and 255. A negative hue is assigned to pixels with a low saturation ($S < 20$) meaning that the pixel is achromatic.

Since we are computing statistics over an image, the definition of the color do not need to be accurate on each pixel. Being accurate would mean using fuzzy logic where there is a fronteer between two colors. The correspondance between HSV and the color name presented here has been designed to be simple and fast to compute. Only 11 colors are considered: black, blue, brown, green, grey, orange, pink, purple, red, white, yellow. More complicated and accurate methods can be designed but our simple method proved to be sufficient for our purpose.

The main criteria used to name the color of a pixel from its HSV values are explicited in Table 1. Brown and orange ($14 < hue < 29$) are the hardest colors to distinguish. We propose the following rule: given the two points $B : (S = 184, V = 65)$ and $O : (S = 255, V = 125)$ and the L1 distance in the (S, V) plane, a pixel whose hue is in the range 14-29 is considered orange if it is closer to O than to

Hue	Color
< 0	black/grey/white
$0 - 14$	red
$14 - 29$	orange/brown
$29 - 60$	yellow/green/brown
$60 - 113$	green
$113 - 205$	blue
$205 - 235$	purple
$235 - 242$	pink
$242 - 255$	red

Hue< 0	
Saturation	Color
$0 - 82$	black
$82 - 179$	grey
$179 - 255$	white

$29 <$ Hue < 60	
Saturation, Value	Color
$S > 80, V \geq 110$	yellow
$S > 80, V < 110$	green
$S \leq 80$	brown

Table 1: Getting the color from the HSV space

B, and brown otherwise. These thresholds were choosen experimentally from the observation of many images. It works well when the color of a pixel is obvious, that is when everybody would agree on the same color for that pixel. We do not deal with the fronteers of colors where the name of the color is subjective and can vary for different observers.

3.2. Obtaining the colors of objects

The colors of objects can be obtained from a huge text corpus, and we propose to use the web to do so. The idea is to study if the objects and the color often appear together or not in the corpus. We have experimented two methods to get the color of an object. For example, let us imagine that we want to get the color of a banana.

The first one is to ask *"yellow banana"* on a web text query where yellow can be any color, and get the number of pages returned. The second way is by asking *"banana is yellow"*. Then again, the category of the object can be used to reduce the noise, so instead of the examples given above, we can ask *"yellow banana" fruit* and *"banana is yellow" fruit*.

We use 14 color words for web querying: black, blue, brown, gray, green, grey, orange, pink, purple, red, rose, tan, white, yellow. This is more than the 11 colors used in image color description, but some colors are merged together: gray and grey are synonyms, brown/tan and rose/pink are also considered as synonyms. For these colors, the corresponding number of results are summed up giving a the number of occurrences $N(C|object)$ of color C for a given object.

In Tables 2 and 3, we show the top five colors returned for banana using Google Search, and the number of results in parentheses. Yellow and green (in that order) are the two main colors we expect to get, and this is what is returned by method 2.

"color banana"	"color banana" fruit
blue (201000)	orange (72300)
green (140000)	green (35300)
orange (134000)	yellow (26600)
yellow (109000)	red (21900)
red (66200)	blue (11500)

Table 2: Colors returned for "banana" using Google Search and method 1

"banana is color"	"banana is color" fruit
yellow (594)	yellow (288)
green (217)	green (51)
purple (107)	black (24)
black (94)	brown (21)
white (93)	blue (16)

Table 3: Colors returned for "banana" using Google Search and method 2

The banana example is representative of what we observed in general for other objects: method 2 provides more accurate results, but less answers than method 1. However, method 1 can be disturbed with proper nouns. For example, "blue banana" is the name of several websites, and "white house" will return a lot of results. Phrases will have the same influence: "blue whale" will give whales as mostly blue, and "white chocolate" will have more hits than "black chocolate" or "brown chocolate". Also, in the specific example of banana, in "orange banana", orange can be a noun (the fruit) instead of an adjective (the color).

These three issues do not arise using method 2. However, sometimes method 2 does not return any color, as for example with the word "passerine" (a type of bird), and in that case, the method 1 can be of help.

3.3. Giving a score to the cluster

The probable colors for an object, and the histogram of colors H_{img} of the images img in each cluster are compared to assign a score to each cluster.

For each image img, the score of the image is the sum over all colors C of the number of pixels $H_{img}(C)$ that have the color C in img, multiplied by the number of occurrences $N(C|object)$ of the color C for the studied object. This score is normalized by the number of pixels of the image.

For each cluster $clust$, the score S_{cl} of the cluster $clust$ is the mean score of the images it contains:

$$S_{cl} = \frac{1}{size(clust)} \sum_{img \in clust} \sum_{C} \frac{H_{img}(C) * N(C|object)}{surface(img)}$$

4. Results

4.1. Segmentation improves clusters quality

Figures 2 and 3 show an example of the differences we can have between clusters without or with segmentation. The query was made using the word "porsche" and the category "car" on Google Image Search. We downloaded

800 images. For the experiment without segmentation, about 460 have been clustered in 14 clusters. For the second experiment, about 700 images were left after segmentation, 500 of which have been clustered in 16 clusters. Here, we are showing 3 of these clusters for each experiment (only 8 images per cluster are displayed) to illustrate the advantage of using the segmentation.

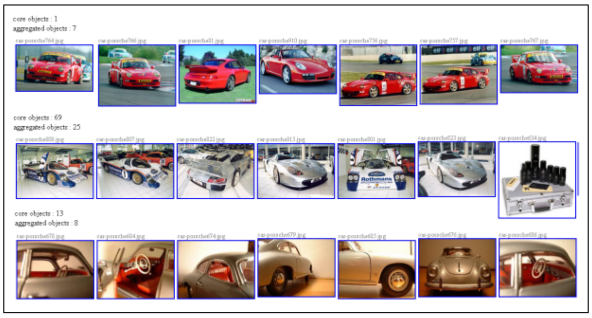


Figure 2: Results without segmentation. The first, second and third clusters contain 8, 94 and 21 images respectively.

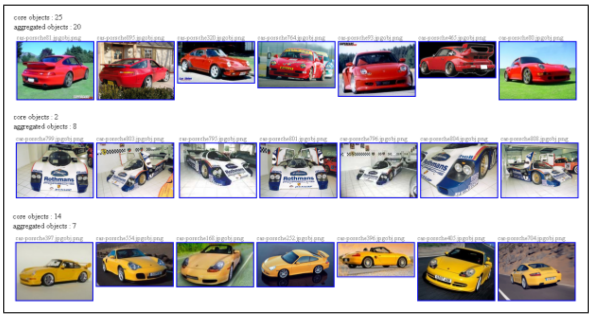


Figure 3: Results with segmentation. The first, second and third clusters contain 45, 10 and 21 images respectively. The full images are shown so that we can notice it is independent of the context.

In Figure 2, the red car cluster depends on the context. The second cluster mixes several car types, and the third one is composed of objects that are not entirely contained in the image. In Figure 3, we see the improvements on the red car cluster which becomes independent of the context, and thus contains more images. At the same time, the grey car cluster has been split and is now more consistent. The yellow car cluster is new, and could not be formed without the segmentation because of the context again. Another cluster has disappeared because the segmentation removes the images which do not contain a centered object.

4.2. Sorted clusters

We are presenting here results of sorted clusters, using the automatic segmentation and the second method for guessing the color of an object. Since up to 500 images can be clustered for a query, we cannot show all the clusters for each query, therefore we decided to show only the first five clusters for several queries.
Anyway, our aim here was: given the name of an object, we want to obtain images of that object which could be further used to build a database for learning. Sorting clusters allows us to decide which are the good clusters to keep, and which are the bad clusters to discard. In this application, having a good precision means having relevant images in the first clusters. Having a good recall means not discarding good images, that is, not having good images in the last clusters. We do not want to have as many images as possible, but we do want to keep only relevant images. Thus, what is important is the precision regardless of the recall.
In Figure 4, the first five clusters obtained for the query *banana fruit* are displayed. The first three clusters contain mostly bananas, some of which have been badly segmented. The other two clusters are not as good, so, only the first three clusters should be included in our database for learning, which gives 64 images of bananas.

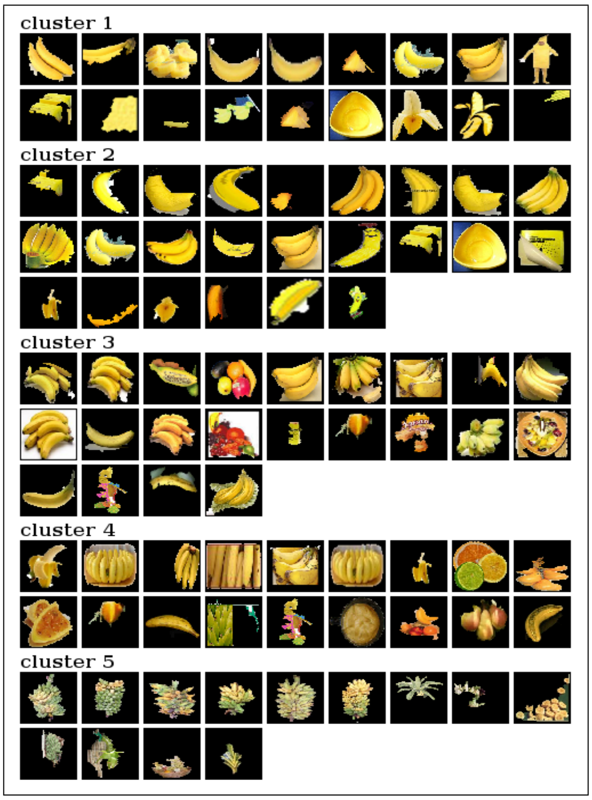


Figure 4: The first 5 clusters out of 17 for the query *banana fruit* are given here. The top ranked clusters are the one containing the more yellow images. The second color for banana is green, which justifies the presence of cluster 5 in that position.

This is really a low number if we consider that we asked for 1000 images on the Internet (Google and Yahoo! image search engines do not allow to retrieve more than 1000 images). After downloading (some links are broken), and segmenting (segmentation discards some images), we still had 566 images, 403 of which have been clustered.
At least two ways could be used to get more images. They both are about asking more queries to the web image search engine. The first one would be to ask the query in multiple languages, using automatic translation and then grouping the clusters together. This method can multiply the number

of images by the number of considered languages. The other way is to use more accurate queries, which would be here the different species of bananas. For the precise example of banana, we would have to use Latin: bananas are *Musa*, and subspecies are for example *Musa acuminata* and *Musa balbisiana* [2]. The obtained images are then considered as *bananas*, since we do not want to be that accurate in our database, and since too accurate queries will return fewer answers, and the clustering does not work with too few images. This second method may only double the number of images.

The algorithm works well in general with objects that have mostly one color, such as *swan animal* (Figure 5).

Disambiguation works well, as can be seen for *jaguar* on Figures 6 (*jaguar car*) and 7 (*jaguar animal*): animals and cars are not mixed in clusters.

The *jaguar car* query also shows that the clustering sorting will work for objects that be of any color, as are man-made objects in general. But we will lose some possible colors of objects. For example, jaguar cars can be blue, but the first blue jaguar car cluster is in 14th position. Thus, for man-made objects, we should explicitly ask for a certain color when retrieving images of an object: since the object can take many colors, people tend to specify it in their annotation, contrary to objects that have mainly only one color, such as fruits or animals.

Some objects are textured with many colors, and for these objects, the algorithm will not perform well. This happens for example with the jaguar (Figure 7), often described as orange-yellow colored, rosette textured. Since black jaguars also exist, they will alter the results. The probable colors found for *jaguar animal* are black (10), orange (4), blue (3), tan (3) and yellow (3): the black cluster appear first. A cluster with orange-yellow jaguars appears in fourth position.

5. Conclusion

In this paper, we have designed a system which, given the name of an object, is able to download images from the web that are likely to illustrate that object. It then automatically segments the images in order to isolate the object from the context. Since results from the web are very noisy, clustering is used to group similar images together, and reject single images. Not all clusters are relevant, therefore we proposed a method to semantically sort these clusters: the probable colors for an object are guessed automatically from the Web, and the clusters are sorted according to these colors.

Further work will be achieved to see if we can find a threshold on cluster scores to separate good clusters from bad clusters. It would also be interesting to test this automatically generated database on real applications such as object recognition and measure its performances.

6. References

Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. 2004. Hierarchical clustering of www image search results using visual. In *Proceedings of the 12th*

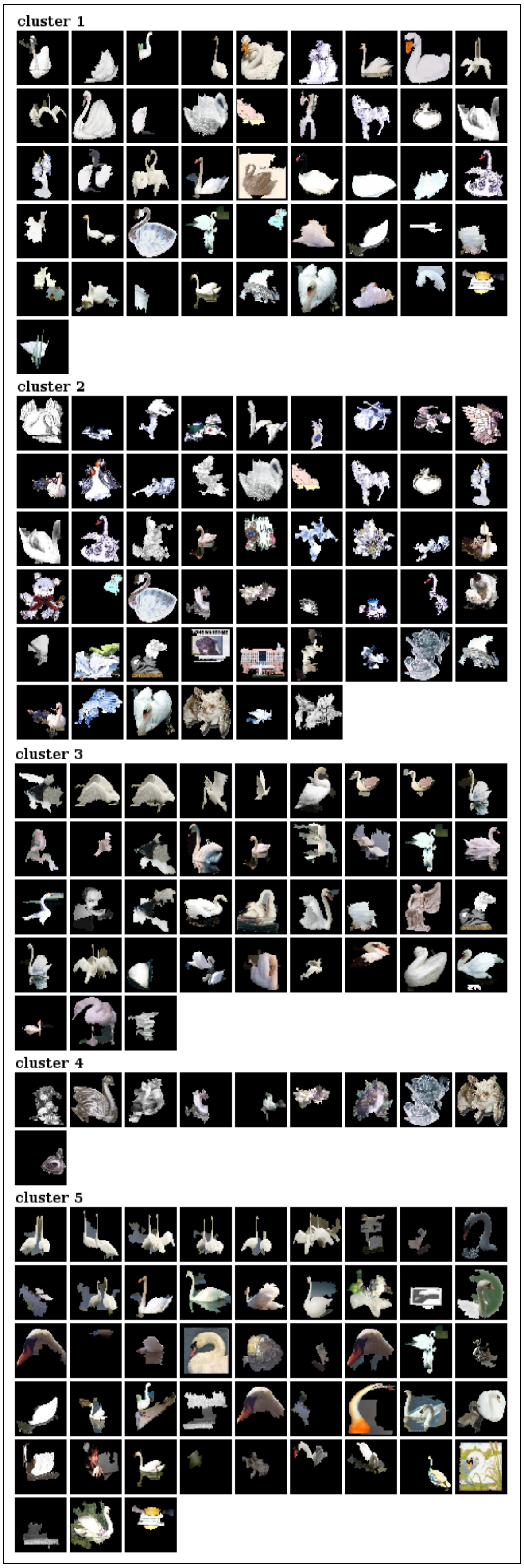


Figure 5: The first 5 clusters for the query *swan animal*. The clusters do not contain many white objects are not about swan and do not appear here

[2] Tens of species of bananas exist and are listed on Wikipedia.

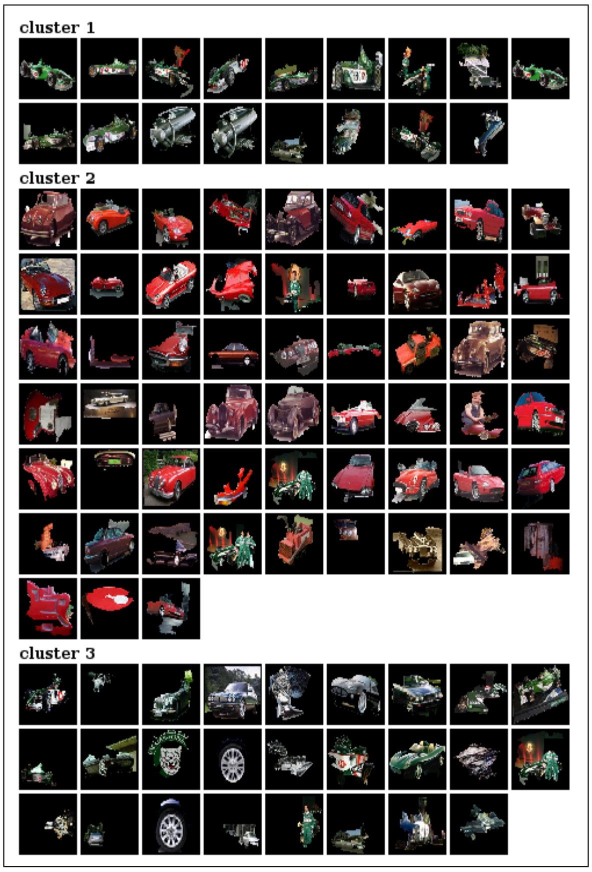


Figure 6: The first 3 clusters out of 16 for the query *jaguar car*. The 3 most probable colors are in that order: green (cluster 1 shows green F1 cars), red (cluster 2) and black.

annual ACM international conference on Multimedia, pages 952–959, New York, USA.

Ya-Chun Cheng and Shu-Yuan Chen. 2003. Image classification using color, texture and regions. *Image Vision Comput.*, 21(9):759–776.

Levent Ertz, Michael Steinbach, and Vipin Kumar. 2001. Finding topics in collections of documents: A shared nearest neighbor approach. In *Text Mine '01, Workshop on Text Mining, First SIAM International Conference on Data Mining*, Chicago, Illinois.

Beatriz Marcotegui and Serge Beucher. 2005. Fast implementation of waterfall based on graphs. In C. Ronse, L. Najman, and E. Decencière, editors, *Mathematical Morphology: 40 Years On*, volume 30 of *Computational Imaging and Vision*, pages 177–186. Springer-Verlag, Dordrecht.

Xin-Jing Wang, Wei-Ying Ma, and Xing Li. 2004. Data-driven approach for bridging the cognitive gap in image retrieval. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME 2004)*, pages 2231–2234, Taipei, Taiwan, June.

Svetlana Zinger, Christophe Millet, Benoit Mathieu, Gregory Grefenstette, Patrick Hède, and Pierre-Alain Moëllic. 2006. Clustering and semantically filtering web images to create a largescale image ontology. In *Proceedings of the IS&T/SPIE 18th Symposium Electronic Imaging 2006*, San Jose California, USA, January.

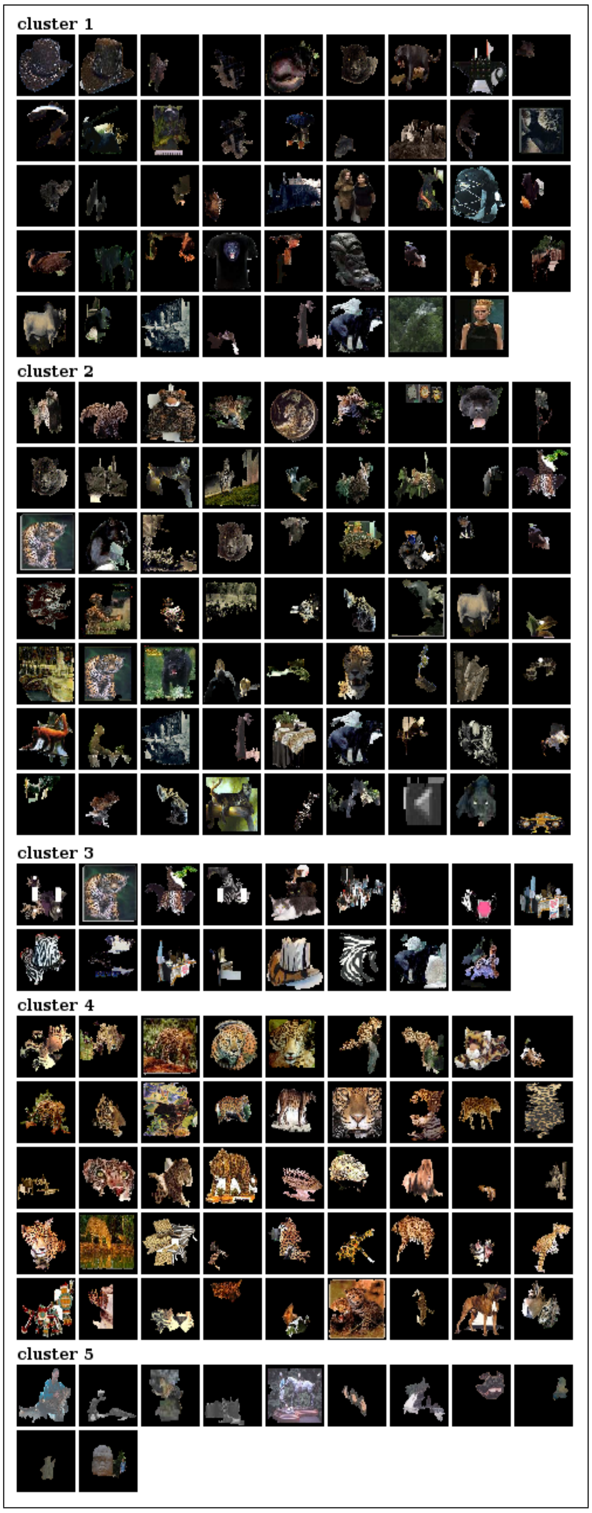


Figure 7: The first 5 clusters out of 9 for the query *jaguar animal*. The black clusters appear first. The 4th cluster looks better, but the segmentation is not satisfactory.

Image-Language Association: are we looking at the right features?

Katerina Pastra

Institute for Language and Speech Processing
Artemidos 6 and Epidavrou, Maroussi, 151-25, Greece
kpastra@ilsp.gr

Abstract

The ever growing popularity and availability of multimedia information has rendered automatic image-language association essential in a number of multimedia integration applications. Bridging the gap between the two media requires an appropriate feature-set for describing their common reference; one that will be both distinctive of the entities referred too and feasible to extract automatically from visual media. In this paper, we suggest an alternative –to current approaches- feature set, which has been used in OntoVis, a domain model for a prototype that describes three-dimensional (3D) indoor scenes. We argue that it is worth employing this feature-set in a larger scale for image-language association and investigating the feasibility of doing so and of detecting such features automatically even beyond 3D visual data, in 2D images.

1. Introduction

Internet Protocol Television, image and video-blogs and image and video-search engines are just a few of the latest technology-trends which become more and more popular, rendering digital multimedia content pervasive. Within such a context, the need for intelligent tools for efficient access to multimedia content has boosted research efforts and interest in automatic image-language association. The research issue is not new, of course; it spans a number of decades and a wide range of application areas, from Winograd's SHRDLU system in 1972, which verbalized visual changes in a 2D blocks scene, to medium translation systems (e.g. automatic sports commentators) and to conversational robots of the new millennium (cf. a review in Pastra and Wilks 2004 and Pastra 2005, ch.3).

Association in these multimedia prototypes took mainly the form of correlating visual information and accompanying text/speech or translating one modality into another (i.e. verbalizing visual information or visualizing linguistic information). In most cases, the systems made use of *a priori* known vision-language associations or used simple inference-mechanisms on small-scale association resources, resorting at the same time to either manually abstracted visual information or just working with miniworlds/blocksworlds. Lack of scalability and heavy human intervention for the task was among the most significant criticisms (cf. Pastra and Wilks 2004).

In this paper, we focus on the features used for describing/detecting common visual and linguistic references to entities in real-world scenes. We first look into the limitations of the features used in state of the art image-language association mechanisms, and then present three different types of features used in OntoVis, a feature augmented ontology for logic-based verbalization (description) of 3D indoor scenes in the VLEMA prototype (Pastra 2006). We discuss the possibilities of scaling the use of the suggested feature set and of detecting it automatically in 2D visual data.

2. Image-Language Association approaches

In the last few years, image-language association mechanisms as such are being developed for automatic image/keyframe annotation, with the vision of being, at some point, mature enough for being embedded in multimedia prototypes and mainly in indexing and retrieval prototypes. The approaches are either probabilistic (Barnard 2003, Wachsmuth et al. 2003) or logic-based (Dasiopoulou et al. 2004, Pastra 2005, ch. 5.). Learning approaches require properly annotated training corpora (Lin et al. 2003, Everingham et al. 2005) for learning the associations between images/image regions represented in feature-value vectors and corresponding textual labels, while symbolic logic approaches rely on feature-augmented ontologies (Dasiopoulou et al. 2004, Simou et al. 2005). Srikanth et al. (2005) report also on the use of both training corpora and ontologies for achieving automatic image annotation.

In all these cases, the features used for describing image content are low-level ones, such as shape, colour, texture, position (2D coordinates of image region), size (portion of image covered by image region), i.e. features used by image analysis components for automatic object detection. Justification of this choice is obvious: the need for relying on such features for automating object detection within image-language association tasks. However, is it a coincidence that all these approaches are exemplified in mini-worlds (e.g. soccer games, where the identification of the ball and the playground is quite straight-forward through shape and colour descriptors)? How distinctive could such features be for more complex objects, such as e.g. furniture (which comes in many different shapes, colours, textures) scenes (i.e. configurations of objects) and therefore, how scalable could the corresponding image-language association mechanisms be? Actually, what kind of object features/properties could one use, so that:

- they are distinctive of object classes (i.e. they allow differentiation among a large number of object types), and
- their values can be detected and used by an image analysis module for automatic object segmentation?

These questions lead back to an old problem in cognitive linguistics, that of the use of features for conceptual representation (Lakoff 1987; Barsalou and Hale, 1993); in meaning analysis/decomposition, feature-based methods define finite sets of conditions or attributes which determine the reference of a word. As pointed out in criticisms of feature-based representations, no set of features can fully represent an entity, cf. ch. 7 in (Lakoff, 1987). However, from within many possible abstractions/features a certain feature-set can be more or less successful in fixing the reference of a concept.

3. The OntoVis suggestion

OntoVis is a domain model (domain ontology with corresponding knowledge-base) for interior scenes. It has stemmed out of OntoCrime, a domain ontology for indoor and outdoor scenes (Pastra et al., 2003), built through priming with the Common Data Model of the UK Police Information Technology Organisation (PITO), the latter being an attempt to standardize the wording used in all tasks that involve police forces. OntoVis includes the part of OntoCrime which refers to indoor scenes, augmented with properties for a number of entities that one can find in sitting-rooms in particular. The ontology is implemented in the form of a directed acyclic graph (DAG) through the use of the XI Knowledge Representation Language (Gaizauskas and Humphreys, 1996).

The same ProLog-based representation language is used for the OntoVis knowledge-base. The object-property assertions in the latter form a kind of “object-profiles” at the “basic-level” of categorization (Rosch 1978, Lakoff 1987), which cover for each object all following types of features/properties:

- *physical structure*:

the number of parts into which an object is expected to be decomposed in different dimensions, e.g. a sofa is always decomposed into more than one parts along its X dimension (each one corresponding to a seat) as opposed to a chair.

- *visually verifiable functionality*:

visual characteristics an object may have which are related to its function e.g. whether an object has a surface on which things can be placed/fixed, and

- *interrelations*:

these refer mainly to (allowable) spatial configurations of objects and object parts (e.g. whether an object could be *on* the floor or not), the dimension according to which size comparisons would be meaningful etc.

Here is an example of the property profiles of two quite similar objects, both of which belong to the same class, that of “*furniture*”:

props(sofa(X),[has_xclusters_moreThan(X,1)]).
props(sofa(X),[has_yclusters_equalMoreThan(X,2)]).
props(sofa(X),[has_ yclusters_equalLessThan(X,4)]).
props(sofa(X),[has_ zclusters_equalMoreThan(X,2)]).
props(sofa(X),[has_zclusters_equalLessThan(X,3)]).
props(sofa(X),[on_floor(X,yes)]).
props(sofa(X),[has_surface(X,yes)]).
props(sofa(X),[size(X,XCLUSTERS)]).

Table 1: part of the "sofa" object profile

props(chair(X),[has_xclusters (X,1)]).
props(chair(X),[has_ yclusters_equalMoreThan(X,2)]).
props(chair(X),[has_ yclusters_equalLessThan(X,4)]).
props(chair(X),[has_zclusters_equalMoreThan(X,2)]).
props(chair(X),[has_zclusters_equalLessThan(X,3)]).
props(chair(X),[on_floor(X,yes)]).
props(chair(X),[has_surface(X,yes)]).
Props(chair(X),[size(X,XCLUSTER_YValue,TableYDIM_UpperConstraint)]).

Table 2: part of the "chair" object profile

Looking at tables 1 and 2, one realises that the two objects (sofas and chairs) are similar in most of their properties; both of them intersect with the floor[1], they have a surface on which other objects may be placed, and they can structurally be decomposed into 2 or 3 parts in their Z dimension (these being the back, the seat+legs part that touches the floor, and optionally the arms, if there are any). Similarly, they can be decomposed into 2-4 parts along their Y dimension (back, seat, arms, and legs, the last two are optionally present). However they differ in their decomposition along their X dimension: a sofa has always more than one X-parts (more than one seats), while a chair may have only one seat. Size is a variable (changeable) property for sofas, and it is actually determined by the number of seats that the object has, while size for chairs makes normally sense only in terms of the height of the chair (e.g. short chairs for children, tall stool-like chairs etc.).

An “armchair” has the same object profile with a chair, apart from the fact that it will always have three or four Y-clusters (back, seat, arms and optionally legs), and always three Z-clusters (back, seat, arms), *i.e.*, arms are not optional, they must be present. Furthermore, an armchair’s relative size does not make sense to be expressed in terms of its height; it is so for chairs, because they are expected to “co-locate” with tables/bars (the height of which may vary considerably), and the “chair’s” height is constrained by a table’s height. Table 3 presents part of the object profile of “tables”:

props(table(X),[has_xclusters(X,1)]).
props(table(X),[has_yclusters(X,2)]).
props(table(X),[has_zclusters(X,1)]).
props(table(X),[on_floor(X,yes)]).
props(table(X),[has_surface(X,yes)]).
props(table(X),[size(X,YDIM,XDIM, Relative_to_Room_YXDIM)]).

Table 3: part of the "table" object profile

A table has a surface (table-top) which can be identified along its X dimension, it has two yclusters (table-top and legs) and one zcluster (the whole table). Its length and height are relative to the corresponding dimensions of the room it is found in.

As seen in the above examples, there is a whole network of interrelations between objects in OntoVis, the detection and identification of each of which contributes

[1] The floor, as well as other room-parts/walls, is defined, in its turn, as the one-dimensional object (surface) with the lowest Y-values in an indoor-scene.

to the detection and identification of the other. The profiles include also assertions regarding the object parts objects are being formed of (and which are not included in the above tables due to space restrictions). For example, a sofa consists of a back, more than one seats and optionally legs and arms; these object parts are themselves defined in a similar way, using the property types suggested above.

There are many arguments in favour of the suggested feature selection for object naming in the literature. In particular, the need for defining objects through their physical structure and their functionality/purpose has been argued by many researchers, such as Minsky (1986). Structural properties were described by Minsky as ones which do not change "capriciously", while functional ones capture intentional aspects of the objects and both are important when defining visual objects. On the other hand, Landau and Jackendoff have explicitly argued that spatial representations imply properties of the objects involved (1993); for example, an "on" relation between two objects requires that the reference object is one with a surface or line boundary on which the figure object is located.

While not panacea, the suggested feature set could assist scaling image-language associations beyond mini-worlds, and actually allow for:

- going beyond differences in the appearance of similar objects (e.g. different styles of sofas) naming these objects in the same way, and

- generalizing over viewpoint differences e.g. identifying a sofa as such even when seen from the side (rather than *en face*)

These are generalizations that current image-language association algorithms cannot do easily (or at all). Identifying objects which differ in appearance as ones of the same type is something that cannot be achieved even with a very large amount of training data (cf. e.g. the visual ontology by Zinger et al. 2005), if a similar example is not present in the training data. Similarly, current approaches cannot deal with viewpoint differences in the appearance of an object and there is an almost infinite number of different images of the same object which may result from differences in the viewpoint (viewing angle and distance) from which the object is seen in a complex scene.

4. Using OntoVis

While the effectiveness of each feature type individually has been argued in the literature, their use in conjunction and their incorporation in a domain model has not been attempted before. Actually, in the case of OntoVis, the feature set has been determined by the visual data itself, and the need to perform automatic object naming within an application scenario that goes from vision to language. OntoVis was created for the development of VLEMA, a system that attempts to test the extend to which one may currently "emancipate" a vision-language integration prototype, in order for the prototype to work with *real visual scenes*, to *analyze its visual data automatically*, and have *inference mechanisms for scalable vision and language association* abilities.

The VLEMA prototype works with automatically reconstructed in 3D images of sitting rooms. It includes a module that performs object segmentation in 3D space by extracting physical structure-related information (clusters of faces forming part of an object in each dimension) to detect objects and/or object parts. An object-naming module refines this detection results by either naming a candidate object or/and suggesting the clustering of candidate object parts into one object which it also names. The module relies on OntoVis for drawing inferences for object naming; the inference mechanisms take advantage of the rich visual information that can be extracted in the 3D space (i.e., 3D coordinates of the candidate objects, relative information on their spatial interrelations, size etc., as well as lack of occlusion, registration, viewpoint problems etc.) to check whether the property assertions in the OntoVis object profiles actually hold (cf. ch. 5 in Pastra 2005a and Pastra 2006). This means that the specific feature set suggested in the previous section stemmed out of a prototype that worked on 3D visual data, and it actually includes features that can be more easily identified in 3D space.

In Computer Vision, research on the automatic 3D reconstruction of real indoor and outdoor visual scenes, as well as on the automatic transformation of 2D images into 3D worlds points to optimistic prospects of taking advantage of the rich information one could extract in 3D space, in real-world application scenarios rather than merely in manually built virtual worlds. While OntoVis was used in such a real-world setting and it was applied on visual data that had been reconstructed in 3D automatically, these reconstruction mechanisms and the ones that transform 2D into 3D are still immature. The question then becomes, whether the OntoVis suggestion could be applied to 2D images, on which the vast majority of state of the art vision-language association mechanisms run.

While this is an issue that should be thoroughly explored with computer vision experts, there is some first evidence that automatic techniques for detecting such (or a reduced version of) visual information in 2D images exist. For example, there are methods for detecting spatial relations between objects in 2D images (cf. e.g. the work by Regier and Carlson, 2001), and there is also research on identifying object structure/parts in 2D images and associating them with textual labels (cf. Wachsmuth et al. 2003).

5. Future Plans

Currently, OntoVis includes "object profiles" for twenty basic-level objects (with their corresponding parts); our plans for the immediate future are to extend this resource to concrete-objects of indoor and outdoor scenes and test their discriminative power in a corpus of manually-constructed virtual reality scenes. Mechanisms for detecting the specified object features in these scenes automatically for object naming purposes will also be applied, as an extension to the work done in the VLEMA prototype.

Given the advantages that could be gained, we believe that it is also worth investigating the possibility of using the suggested feature set in state of the art image-language association mechanisms for 2D images; it is towards this direction that we tend to head our research efforts towards.

6. Conclusions

In this paper we presented a feature-set for the representation of real world objects and scenes, within tasks that attempt to bridge the gap between low-level visual information and high-level (conceptual) linguistic descriptions of entities. The suggestion has been implemented in OntoVis, a domain model for building-interior scenes; the suggested features have been detected automatically in 3D visual data and have been used for the verbalization of this data. We argue that the feature set could be an alternative or complimentary one to feature sets used in state of art image-language association mechanisms and would like to invoke cooperation and collaboration towards this direction of research.

7. References

Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D., Jordan M. (2003), "*Matching words and pictures*", in Machine Learning Research, 3:1107-1135.

Dasiopoulou S., Papastathis V., Mezaris V., Kompatsiaris I., Strintzis M. (2004), "An ontology framework for knowledge-assisted semantic video analysis and annotation", in Proceedings of the International workshop on Knowledge markup and semantic annotation, International Semantic Web Conference.

Everingham M., Van Gool L., Williams C. and A. Zisserman (2005), "*PASCAL Visual Object Classes Challenge Results*", Technical Report, PASCAL Network of Excellence, http://www.pascal-network.org/challenges/VOC/voc/results_050405.pdf

Gaizauskas, R. and K. Humphreys, (1996), "*XI: A Simple Prolog-based Language for Cross-Classiffication and Inheritance*", In *Proceedings of the 7th International Conference in Artifficial Intelligence: Methodology, Systems, Applications*, pages 86-95.

Jackendoff, R. (1987). "*On beyond Zebra: the relation of linguistic and visual information*", *Cognition*, 20:89-114.

Lakoff, G., (1987). "*Women, Fire, and Dangerous Things*". The University of Chicago Press.

Landau, B. and R. Jackendoff (1993) "*What" and "Where" in spatial language and cognition*", *Behavioural and Brain Sciences*, 16:217-265.

Lin C., Tseng B. and J. Smith (2003), "*Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets*", in online Proceedings of TRECVID 2003.

Minsky, M. (1986). *The Society of Mind*. Simon and Schuster Inc.

Pastra K. (2006), "*An alternative suggestion for vision-language integration in intelligent agents*", in Proceedings of the International Hellenic Artificial Intelligence Conference, Athens, Greece.

Pastra K. (2005), "*Vision-Language Integration: a Double-Grounding Case*", PhD thesis, Department of Computer Science, University of Sheffield.

Pastra K. and Y. Wilks (2004), "*Vision-Language Integration in AI: a reality check*", in Proceedings of the 16th European Conference on Artificial Intelligence (ECAI), pp. 937-941, Valencia, Spain.

Pastra, K., H. Saggion, and Y. Wilks, (2003), "Intelligent indexing of crime-scene photographs", IEEE Intelligent Systems, 18(1):55-61.

Regier, T. and L. Carlson, (2001), "*Grounding spatial language in perception: An empirical and computational investigation*". *Journal of ExperimentalPsychology*, 130(2):273-298.

Simou N., Tzouvaras V., Avrithis Y., Stamou G., Kollias S. (2005), "*A visual descriptor ontology for multimedia reasoning*", in Proceedings of the Workshop on Image analysis for Multimedia Interactive Services (WIAMIS).

Srikanth M., Varner J., Bowden M., Moldovan D. (2005), "*Exploiting ontologies for automatic image annotation*", in Proceedings of SIGIR.

Wachsmuth S., Stevenson S., Dickinson S. (2003), "*Towards a framework for learning structured shape models from text-annotated images*", in Proceedings of the HLT-NAACL workshop on Learning word meaning from non-linguistic data.

Zinger S., Millet C., Mathieu B., Grefenstette G., Hede P., Moellic P. (2005), "*Extracting an ontology of portrayable objects from WorNet*", in Proceedings of the MUSCLE/Image CLEF workshop on Image and Video Retrieval Evaluation.

Testing an automatic organisation of retrieved images into a hierarchy

Mark Sanderson, Jian Tian, Paul Clough

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello St,
Sheffield, S1 4DP, UK
(m.sanderson|p.d.clough)@shef.ac.uk

Abstract

Image retrieval is of growing interest to both search engines and academic researchers with increased focus on both content-based and caption-based approaches. Image search, however, is different from document retrieval: users often search a broader set of retrieved images than they would examine returned web pages in a search engine. In this paper, we focus on a concept hierarchy generation approach developed by Sanderson and Croft in 1999, which was used to organise retrieved images in a hierarchy automatically generated from image captions. Thirty participants were recruited for the study. Each of them conducted two different kinds of searching tasks within the system. Results indicated that the user retrieval performance in both interfaces of system is similar. However, the majority of users preferred to use the concept hierarchy to complete their searching tasks and they were satisfied with using the hierarchical menu to organize retrieved results, because the menu appeared to provide a useful summary to help users look through the image results.

1. Introduction

One process that users must perform when information seeking is to examine and interpret the search results. In most Information Retrieval (IR) systems, results are ranked in order of relevance to the query. However, if many search results are returned it can be difficult for the user to examine them all. In addition, reliably providing an intuitive summary of the search results is an obvious benefit to any user of an IR system. Hearst (1999) discusses various interface techniques for summarising results to make the document set more understandable to the user. These include: visualising the relationship of documents to the query, providing collection overviews and highlighting potential relationships between documents.

A variety of *clustering* techniques have been developed in IR to group documents. This can help users to browse through the search results, obtain an overview of their main topics/themes and help to limit the number of documents searched or browsed in order to find relevant documents (i.e. limit exploration to only those clusters likely to contain relevant documents). Two common variations are: (1) to group documents by associated terms (i.e. a set of words or phrases define a cluster and membership is based on its containing a sufficient fraction of a cluster's terms), and (2) to assign documents to pre-defined thematic categories (manually or automatically). Scatter/Gather (Cutting et al, 1992) and the Vivisimo[1] metasearch engine are an example of the former and Yahoo! Categories an example of the latter.

Organizing a set of documents automatically based upon a set of categories (or concepts) derived from the documents themselves is an obviously appealing goal for IR systems: it requires little or no manual intervention (e.g. deciding on thematic categories) and like unsupervised classification, depends on natural divisions in the data rather than pre-assigned categories (i.e. requiring no training data). In this paper we make use of such an approach for organizing search results called concept hierarchies (Sanderson & Croft, 1999; Sanderson & Lawrie, 2000). This simple method of automatically associating terms extracted from a document set has been successfully used to help users searching and browsing for documents (Joho, Sanderson, Beaulieu, 2004). In this simple method, words and noun phrases (called concepts) are extracted from passages of the top *n* documents and organized hierarchically based on document frequency and a statistical relation called subsumption.

Given the simplicity of this method and its success for document retrieval, in this paper we apply concept hierarchies to textual metadata associated with images for image retrieval and user test the resulting system. There are many instances of when images are associated with some kind of text semantically related to the image (i.e. metadata or captions). For example, collections such as historic or stock-photographic archives, medical databases, art/history collections, personal photographs (e.g. Flickr.com) and the Web (e.g. Yahoo! Images). Retrieval from these collections is typically supported by text-based searching which has shown to be an effective method of searching images (Markkula & Sormunen, 2000). To enhance such systems, various approaches have been explored to organize search results based on either textual and visual features (or a combination of both). A summary of related work is provided in section 2. In practice, given the proliferation of textual metadata, investigating methods to exploit this text (e.g. for organizing results) is beneficial.

The paper is ordered as follows: in section 3 we describe how we used concept hierarchies as a method for presenting image search results by displaying extracted concepts within a hierarchical structure. We describe the methodology and results of two user experiments to test the system and finally conclude.

2. Related Work

For image retrieval, clustering methods have been used to organize search results by grouping the top *n* ranked images into similar and dissimilar classes. Typically this is based on visual similarity and the cluster closest to the query or a representative image from each cluster can then be used to present the user with very different images enabling more effective user feedback. For example Park et al. (2005) take the top 120 images and cluster these using hierarchical agglomerative clustering methods

[1] `http://vivisimo.com`

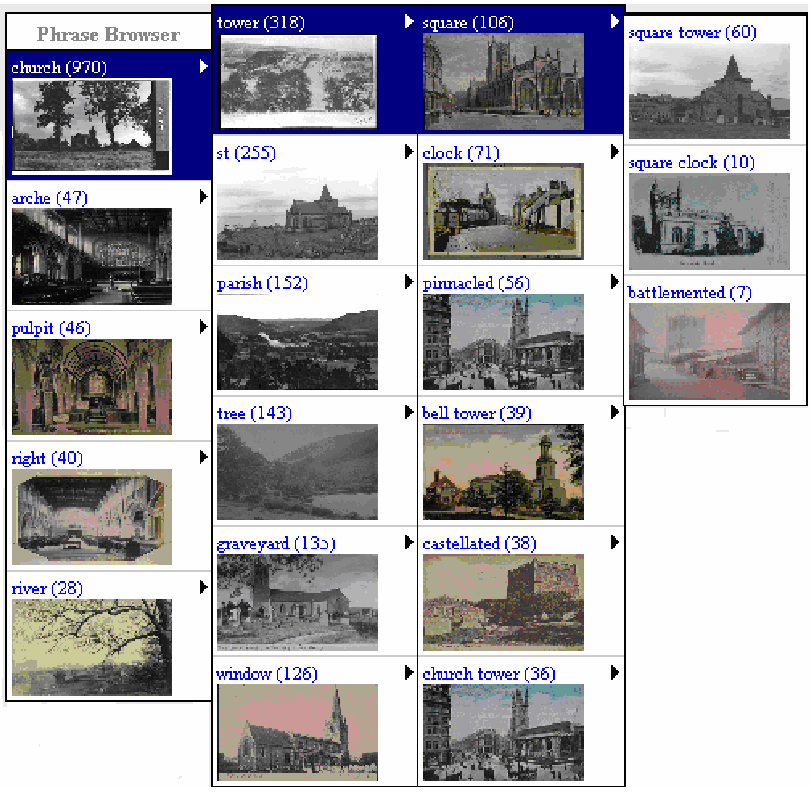


Figure 1: Example fragment from generated menu for the query "church"

(HACM). Clusters are then ranked based on the distance of the cluster from the query. The effect is to group together visually similar images in the results.

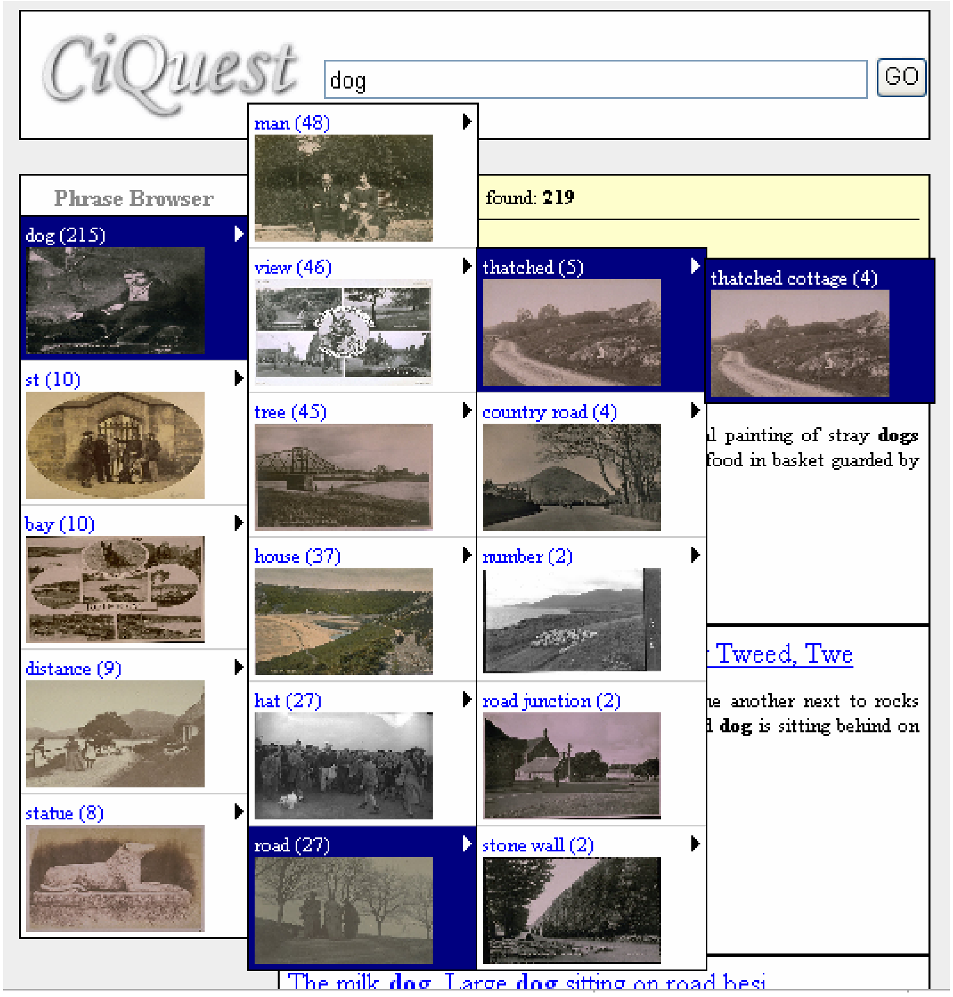
Other approaches have combined both visual and textual information to cluster sets of images into multiple topics. For example, Cai et al. (2004) use visual, textual and link information to cluster Web image search results into different types of semantic clusters. Barnard and Forsyth (2001) organize image collections using a statistical model which integrates semantic information provided by associated text and visual features provided by image features. During a training phase, they train a generative hierarchical model to learn semantic relationships between low-level visual features and words. The resulting hierarchical model associates segments of an image (known as *blobs*) with words and clusters these into groups which can then be used to browse the image collection.

Approaches using only semantic information derived associated text have also been used to organize search results and to aid browsing. For example, Yee, et al. (2003) describe Flamenco, a text-based image retrieval system in which users are able to drill-down results along conceptual dimensions provided by hierarchically faceted metadata. Categories are automatically derived from Wordnet synsets based on texts associated with the images, but assignment of those categories to the images is then manual. Finally, Rodden et al. (2001) performed usability studies to determine whether organization by visual similarity is actually useful. Interestingly, their results suggest that images organized by category/subject labels or were more understandable to users that those grouped by visual features.

3. Building Concept Hierarchies

The approach of building a concept hierarchy proposed by Sanderson and Croft (1999) aims to automatically produce, from a set of documents, a concept hierarchy similar to manually created hierarchies such as the Yahoo! categories. The main difference being that concepts are in fact words and phrases (referred to as *terms*) found within the given set of documents and not categories defined manually. In their method of building concept hierarchies, word and noun phrases (called concepts) are extracted from retrieved documents and used to generate a hierarchy. Concepts are associated based on the set of documents indexed by the two concepts: the more documents two terms share, the more similar they are. However, concept hierarchies go beyond simple grouping of terms by discovering whether concepts are also related hierarchically. Document frequency and a statistical relation called subsumption is used to generate a hierarchy by detecting whether a parent term refers to a related, but more general concept than its children (i.e. whether the parent's concept subsumed the child's). Using document frequency (DF) to determine the semantic specificity of concepts is commonly used for weighting terms in IR based on Inverse Document Frequency (IDF).

With subsumption, concept C_i is said to subsume concept C_j when a set of documents in which C_j occurs is a subset of the documents in which C_i occurs. Or more formally, when the following conditions are held: $P(C_i|C_j) \geq 0.8$ and $P(C_j|C_i) < 1$. The assumption is that C_i is likely to be more general than C_j because, first, the former appears more frequently than the latter [13], and second, the former subsumes a large part of C_j's document set. Also they are likely to be related since they co-occur frequently within documents. The results can be visualised



l painting of stray **dogs**
food in basket guarded by

Tweed, Twe

e another next to rocks
l **dog** is sitting behind on

The milk **dog** Large **dog** sitting on road besi

Figure 2: Example of the menu interface

using cascading menus where more general terms are placed at a higher level followed by related but more specific terms (Figure 1).

Sanderson and Croft analysed a random sample of parent-child relations and found that approximately 50% of the subsumption relationships within the concept hierarchies were of interest and that the parent was judged to be more general than the child. In particular, 49% of children were judged to reflect an aspect of the parent (a holonymic relation), e.g. actor is an aspect (or part) of a movie, 23% judged as a type of the parent (a hypernymic relation), e.g. a poodle is a type of dog, 8% judged to be the same as the parent, 1% as opposite to the parent, and 19% to be an unknown relation. We discuss relations commonly found using image captions in section 5. In summary, to generate a concept hierarchy for image browsing, the following steps are followed after an initial retrieval:

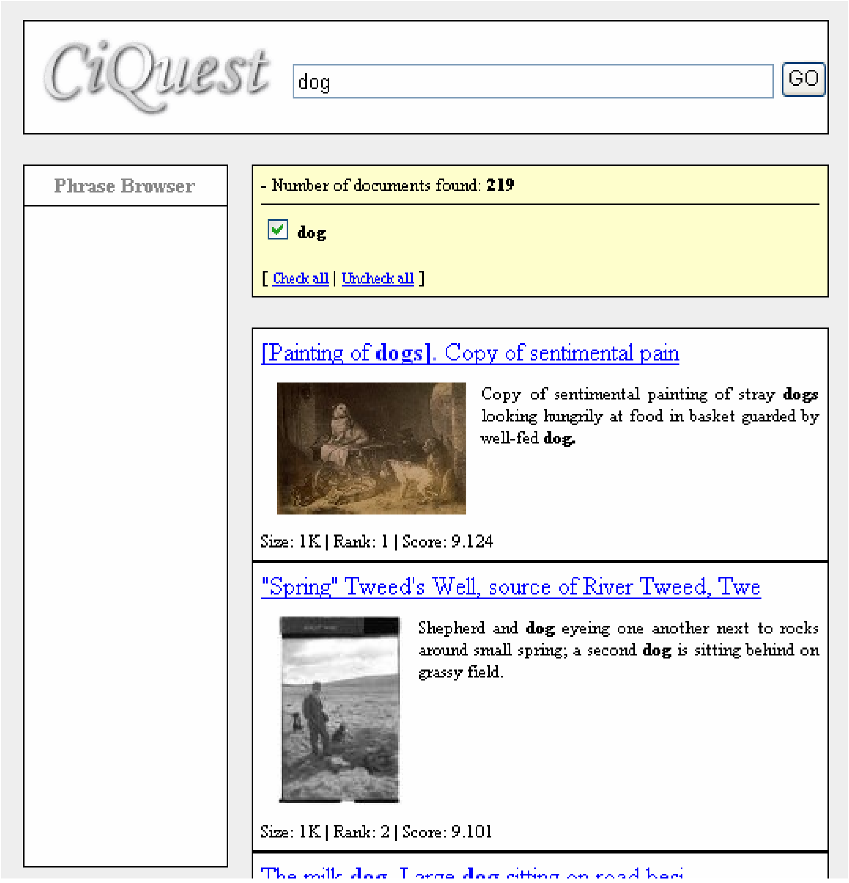
1. Extract concepts (words and noun phrases) from up to the top *n* image captions.
2. Compare each concept with every other concept and test for subsumption relationships.
3. Order concepts hierarchically based on DF scores (general to specific) and subsumption relation (concepts with no parent – no other concept subsumes - are top-level concepts).
4. Randomly select an image from the cluster to represent the cluster visually and create the menu.

For our image retrieval prototype, we used a version of the CiQuest system created to investigate user interaction with a standard textual document collection (Bernard & Forsyth, 2001). The system uses a probabilistic retrieval model based on the BM25 weighting function (Robertson et al 1995) to perform initial retrieval. A DHTML menu is generated dynamically representing the concept hierarchy, enabling users to interact with and browse the search results (Figure 1). The number in parenthesis is document frequency. A number of parameters can be adjusted in the prototype including:

1. *menu_depth*: maximum depth of menu;
2. *menu_height*: maximum height of menu;
3. *top_n*: number of documents to extract concepts from.

4. Experimental methodology

The current study is primarily concerned with evaluating the utility of the concept hierarchy menus to organise retrieved results and observe user interaction with the concept hierarchy menu based on a user-oriented task. To elaborate:



- Number of documents found: **219**

☑ **dog**

[Check all | Uncheck all]

[Painting of **dogs**]. Copy of sentimental pain

Copy of sentimental painting of stray **dogs** looking hungrily at food in basket guarded by well-fed **dog**.

Size: 1K | Rank: 1 | Score: 9.124

"Spring" Tweed's Well, source of River Tweed, Twe

Shepherd and **dog** eyeing one another next to rocks around small spring; a second **dog** is sitting behind on grassy field.

Size: 1K | Rank: 2 | Score: 9.101

The milk **dog**. Large **dog** sitting on road besi

Figure 3: Example of the list interface

- evaluate the usability of concept hierarchy menus used in image retrieval from a user's perspective;
- obtain participants' perceptions of using concept hierarchy menus to group image retrieval results;
- gather participant's general impressions of menu interface (see Figure 2), compared with traditional list interface (see Figure 3); and
- analyse participants' searching behaviour with the concept hierarchy menu in image retrieval system.

4.1. Test Image Collection

The dataset used consisted 28,133 historic photographs from the library at St Andrews University[2]. All images are accompanied by a caption consisting of 8 distinct fields (short title, long title, description, location, date, photographer, notes and topic categories) which can be used individually or collectively to facilitate image retrieval. The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English and contain colloquial expressions and historical terms. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are black and white, although colour images are also present. The dataset has been used for previous image retrieval experiments, the most notable being the ImageCLEF evaluation[3] campaign for cross-language image retrieval, see Clough, Mueller, and Sanderson (2005).

The methodology of the study was by means of conducting usability tests, including, task records, observation notes, pre- & post-session questionnaire and post-search interviews in order to get the perception of the participants. In the user test, each participant will be presented with two different version of the CiQuest interface and be asked to perform two user tasks on each.

4.2. Participants

A total of 30 participants were recruited for doing the user test. The majority of the participants (23) were graduate students of the Department of Information studies, University of Sheffield, and the rest were from other Departments of University. They consisted of 14 females and 16 males. The age of the participants ranges from 20 to 31 with an average of 25. All participated in the study as volunteers.

4.3. Experimental Tasks

Task one was designed as real life retrieval task, participants were required to search for images about a pre-specific topic using the CiQuest system with its different interfaces. In task two, participants were shown three photos taken from the St Andrews historic photographic collection and were required to find them using the CiQuest system with two different interfaces respectively. This task in real life can be described as,

[2] http://specialcollections.st-and.ac.uk/

[3] http://ir.shef.ac.uk/imageclef/

users trying to search for a specific image they have in mind; however, they do not know the exact keyword information to find it, so they need to describe the image by themselves. This task could be used to measure usability of experimental system, focusing on the effectiveness and efficiency.

In order to minimize order effects, users were shown either the menu interface first, or the list.

5. Results and Analysis

The results and analysis of current study are presented as follows.

5.1. Task One

In task one, each participant needed to work with both interfaces. Participants were asked to find 15 photos using CiQuest that were relevant to pre-designed topics. Based on their actual searching performance, participants were required to answer questions to evaluate the two different interfaces of the system. The participants were asked to work through 5 queries each. Results are presented in Table 1.

Mean score for task one	Menu	List
Av. number of pages user browsed	5	8
Av. number of queries type into system	1.6	3

Table 1: Mean score of five topics

As can be seen, in the list interface, users browsed more pages and entered more queries than when using the menu system. When participants use the list interface to search for photographs, they type the initial query into system and then at least examined one page of returned results to judge whether or not they need to reformulate their initial query. Based on author observation during the test, the majority participants were noted to browse at least two pages of results before they changed their query. So, if they change queries frequently, they must spend a lot times to view results. Therefore, in general the number of queries is proportional to the number of result pages.

When using the menu interface, the majority of participants spent time with the terms chosen for the menu as opposed to submitting a new query or going to view results page by page. The majority of participants used the menu interface usually to browse the first page of retrieved results in response to their initial query at first. Then if they could not find the relevant images they required, they prefer to view the concept menu before they went to the next page. They try to find appropriate terms on the menu to limit their initial retrieved results, and then they click term to browse associated results. If they could not find the photos, they went back to concept menu and tried other terms.

5.1.1. Questionnaire

Participants' general impressions of the two interfaces were gathered. Participants indicated on how easy or hard it was to find relevant images and how confident they were when locating images. The average time spent on completing this task was also shown in the table below.

As Table 2 shows, participants using the list interface spent more time on searching than using the menu interface a probable consequence of needing to enter more queries to complete their task. From observation of participants interaction with concept hierarchy menu, we can found the automatically generated concept hierarchy menu really helped users to narrow their result set down.

Task 1	Menu	List
Av. Time to complete task (min.)	10.2	12.4
How easy to judge relevance	4.0	3.2
How confident in judgements	4.1	3.8
Satisfied with the results	4.1	3.8

Table 2: Mean score of five topics

Also according to the table, the majority of participants thought it was easier to judge relevant images using the menu interface. The next question showed on the table was designed to evaluate how confident participants were with their relevant image choice. The mean score of using menu interface was 4.1, which slightly higher than mean score 3.8 of using list interface.

With information gained from the results of the experiments in Task one, we moved onto the second Task.

5.2. Task two

In task two, each participant again tested both the list and menu interfaces, with the aim of locating a "known item" image in the collection. All participants were asked to locate 3 images: half searched the menu interface first (referred to here as the menu group) and the other half used the list interface first (the list group). Results of the experiment are shown in Table 3

Task 2	Menu	List
Av. Time taken to find image	3.0	4.0
Av. number of result pages user browsed before finding the image	9.7	13.3
Av. number of queries	3.7	7.0
Success retrieval rate	91%	78%

Table 3: Mean score for task 2

As can be seen as with task one the average number of pages viewed and queries entered was smaller for the hierarchy interface than it was for the list, also (as before) the time users took to find the image on the menu system was shorter. What is more striking is the success rate of users in locating their known image: users were noticeably more successful in finding their target image with the menu system than they were for the list system. This result indicates that the concept hierarchy menu could provide some useful clues to help participants to find images. The concept hierarchy menu can improve retrieval effectiveness.

5.2.1. User behaviours

According to notes taken while observing users, the majority of participants in the menu group spent a lot of time browsing the menu. They seemed to prefer to view all parts of the menu, in order to find some similar images. They were particularly pleased when the required image was found with this strategy. Participants appeared to prefer searching through the menu than to re-formulate their query. It would appear that building a simple term hierarchy coupled with presenting that hierarchy in a quick browsing form is liked by users

6. Study findings

We analyzed the qualitative and quantitative results about they experimental system. By combining all results, some findings can be detected in this study.

The overall research aim of this study was to establish if the image retrieved results organized by automatically generated concept hierarchy menu is usable from the user perspective.

According to the task one result, image retrieval performance using menu interface was slightly better than using list interface. Although there was no significant difference between them, the results illustrated that the automatically generated hierarchy menu does support the image retrieval process. The concept hierarchy menu could group the image retrieved results by specific term related to the participants' initial query, in order to narrow the number of results returned to the screen. Based on the observation note, when participants used the menu interface, majority of them prefer to browse concept hierarchy menu choosing appropriate term instead of changing query or viewing a large number of results page by page. According to the evaluation questionnaire, the results illustrated that participants using menu interface were more satisfied with their task results than using list interface.

Secondly, from previous discussion of task two, although it was shown that there was no significant difference in retrieval performance between menu group and list group, using concept hierarchy menu can be seen as benefit to image retrieval process. The terms displayed on the concept hierarchy menu provided some useful clue for user to improve the successful rate on finding photos. Browsing concept hierarchy menu could be seen as providing an alternative choice for user to successfully find image, especially when participants' queries did not work.

Finally, based on the results of evaluation questionnaire, the majority of participants thought the menu interface is not as easy to use as list interface. However, the menu interface is easy to learn to use. All participants were never used the experimental system before. After the training session, they can easily learn to use it to complete two search tasks. Therefore, the learnability of the menu interface can be seen as acceptability. In addition, majority of participants gave the positive remark on concept hierarchy menu used in image retrieval. The satisfaction rate in menu interface was slightly higher than list interface. The majority participants were satisfied with using concept hierarchy menu to organize the retrieved results. They also mentioned that they prefer to use menu interface to retrieve image in the future.

However, some participants had a number of negative opinions in using menu interface. For example, two participants who favoured list interface mentioned that some terms displayed on the menu totally make them feel confused; they have no idea why these terms could be generated. Other participants also stated that some terms make them to the wrong path, result in waste a lot time and may sidetrack their original thought.

7. Conclusions

Overall the participants' impression of the experimental system CiQuest as image retrieval system was encouraging. They were satisfied with the search results and retrieval performance. Although both interfaces of experimental system had the similar capability to retrieve relevant images in response to users' query, majority participants prefer to use menu interface to organize their retrieved results in current study. Participants indicated that concept hierarchy menu could provide an intuitive preview for large numbers of retrieved results that gave them a better idea of the topics of image retrieved. So they can effectively narrow a lot returned retrieved results by choosing specific relevant topic, in order to avoid wasting so many time on browsing large numbers of results page by page. Participants also prefer to consider browsing concept hierarchy menu as an alternative way to help them successfully and effectively retrieve images, especially when their queries did not work well.

8. Acknowledgements

The work in this paper was part of a student Masters dissertation project conducted at the University of Sheffield and was also supported in part by the EU's BRICKS project IST-2002-2.3.1.12.

9. References

Bernard, K. and Forsyth, D. (2001) Learning the Semantics of Words and Pictures. In: *Proceedings of the Intentional Conference on Computer Vision*, vol 2, pp. 408-415.

Cai, D., He, Xiaofei., Li, Zhiwei., Ma, W-Y., and Wei, J-R. (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 952-959.

Clough, P., Mueller, H. and Sanderson, M. (2005), The CLEF 2004 Cross Language Image Retrieval Track, In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B. (Eds.) *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, Lecture Notes in Computer Science, Springer, to appear.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. (1992) Scatter/gather. A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*

Hearst, M. (1999). User Interfaces and Visualization. In: Baeza-Yates, R. & Ribeiro-Neto, B. (eds.), *Modern Information Retrieval*, pp. 257-323. New York: ACM Press.

Joho, H., Sanderson, M., and Beaulieu, M. (2004) A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool. In: McDonald, S. & Tait, J. (eds), *Advances in Information Retrieval, 26th European Conference on Information Retrieval*, pp. 42-56.

Markkula, M. and Sormunen, E. (2000) End-use searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval*, 1, pp. 259-285.

Park, G., Baek, Y., and Lee, H-K. (2005) Re-ranking algorithm using post-retrieval clustering for content-based image retrieval, *Information Processing and Management*, 41(2), pp. 177-194.

Rodden, K., Basalaj, W., Sinclair, D., and Wood, K. (2001) Does Organisation by Similarity Assist Image

Browsing?, In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 190-197.

Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M. & Payne, A. (1995). Okapi at TREC-4. In: Harman, D.K. (ed.), *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD. pp. 73-97.

Sanderson, M. and Croft, B. (1999) Deriving concept hierarchies from text In: *Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval*, pp. 206-213.

Sanderson, M. and Lawrie, D. (2000) Building, Testing, and Applying Concept Hierarchies In: W. Bruce Croft, (ed.), *Advances in Information Retrieval: Recent Research from the CIIR*, Kluwer Academic Publishers, pp. 235-266.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28 (1), 11-21.

Yee, K-P., Swearingen, K., and Hearst, M. (2003) Faceted metadata for image search and browsing. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401-408.

(extended abstract)

CLiMB: Computational Linguistics for Metadata Building

Judith L. Klavans, Ph. D.
University of Maryland
jklavans@umd.edu

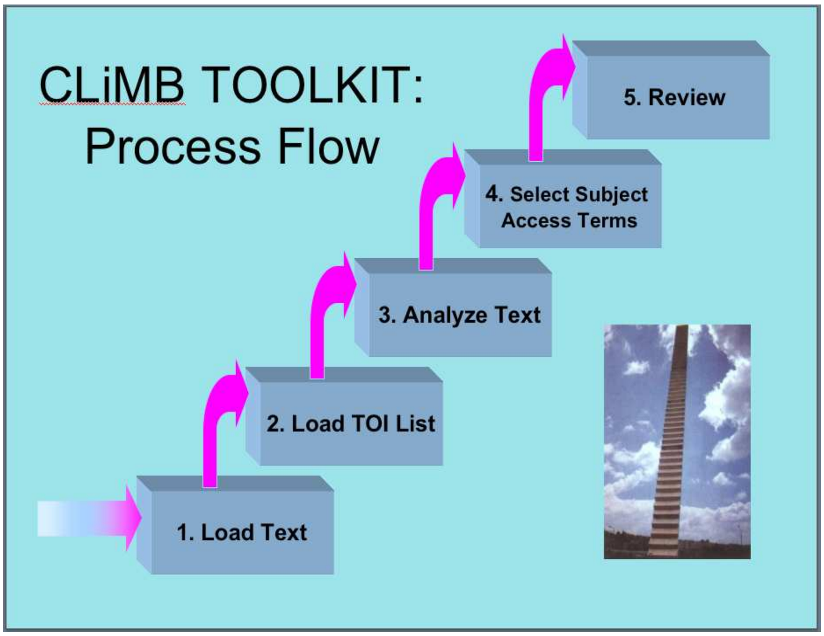
The goal of the **C**omputational **Li**nguistics for **M**etadata **B**uilding[i] (CLiMB) project is to discover to what extent and under which circumstances automatic techniques can be used to extract descriptive subject-oriented metadata from scholarly and authoritative texts associated with image collections. Although manual cataloging is an established field, what is novel about the CLiMB approach is the notion that high-precision cataloging might be accomplished automatically. The achievement of CLiMB's goals will not only address the cataloging bottleneck that arises as the volume of available data increases with new technology; it will also benefit end-users by providing a set of tools to aid in the access of information across collections and vocabularies. As a research project, CLiMB has created a cataloger's platform in which to determine whether such an approach is indeed useful, and to what extent results can be incorporated into existing metadata schema, e.g. VRA, MARC, Dublin Core.

The initial CLiMB Toolkit was fully implemented and evaluated at Columbia University as a Web-based application, operated from within a standard browser. This paper will describe the results of that implementation and the uses of the prototype toolkit. CLiMB employs text sources that are tightly-coupled with a digital image collection to automatically extract descriptive metadata from those texts – in effect, making the writings of specialist scholars useful to enrich the catalog entry. In many cases, researchers have already described aspects of selected images in contexts such as scholarly monographs and subject specific encyclopedias. The challenge is to identify the *meaningful* facts (or metadata) in the written material and distinguish them from among the thousands of other words that make up the text in its primary form.

Ordinarily, descriptive metadata (in the form of catalog records and indexes) are compiled manually, a process that is slow, expensive, and often tailored to the purpose of a given collection. Our goal as a research project is to explore the potential for employing computational linguistic techniques to alleviate some of the obstacles that prevent wide access to digital collections. In the short run, by enhancing the identification of descriptive metadata through the use of automatic procedures, the CLiMB project has enabled the selection of candidate terms for review by catalogers. These candidate terms are extracted from written and tightly-coupled material associated with images in digital collections.

The Climb Architecture

Figure One shows the CLiMB process flow. The text to be loaded in Step One refers to the selected text for processing by CLiMB. This text must, of course be in electronically-readable form, and must be either free of copyright, or have the permissions arranged in advance with the CLiMB team. Although a user will never see, nor be able to recreate the source text, the cataloger must be able to explore the context of a selected term, thus bringing in the questions of rights and permissions. For Step Two, the TOI (target object identifier) list refers to predefined named entities, provided by the minimal catalog record created upon the initial intake of an image collection



CLiMB TOOLKIT: Process Flow

1. Load Text

2. Load TOI List

3. Analyze Text

4. Select Subject Access Terms

5. Review

Figure 1: CLiMB Toolkit Process

Step Three will be the focus of the longer version of this paper, since the LREC audience is likely to be more interested in NLP methodologies as in issues such as rights and permissions, although all topics are essential for the success of the project as a whole. In the CLiMB-1 toolkit, a segmentation algorithm was implemented based on Kan, Klavans, and McKeown (1998). This step permits association between a text unit and a related image; for texts such as a catalog raisonne, no such step is generally required since text entries tend to be short and closely tied to a single image. The segmentation-association step is one where future research will enable more accuracy in linking the relevant text section with images and is a part of the association process flow which provide some key research areas in the future. A standard tagger (we have used the Mitre public toolkit, and we have experimented with other taggers), along with a named entity recognizer is applied to the segmented text. Lookup in the Art and Architecture Thesaurus (AAT) is performed. At this point, the user (a cataloger in this implementation of the Toolkit), is shown potential metadata for selection and feedback. The cataloger can, at Step Four, select subject access terms to be loaded into the catalog record. Step Five, review, permits the user to alter, delete, or insert any changes before the final load.

CLiMB includes an extensive evaluation component, including formative evaluations with a wide variety of user types and iterative evaluation with cataloger specialists. These results, to be reported in Passonneau et al. (2006) will lead to the ability to assess the usefulness of CLiMB metadata once included in image search platforms. To our knowledge, CLiMB is engaged in a novel approach to issues of automatic metadata extraction from selected authoritative texts combined with thesauri and other authority lists to assign weights to potential terms for use in image access. Thus, running studies with catalogers requires some initial training to familiarize them with the types of information, and types of error, they are likely to encounter. By enlisting catalogers to judge the output, we can then collect additional feedback for the ultimate application of new techniques.

CLiMB Achievements

Although the focus of this paper will be on the NLP components of CLiMB, at the service of the application, we will also discuss three aspects of the project that are required for success. The first is the identification of collections appropriate for us by CLiMB. We have developed selection criteria guidelines for us in the project (e.g. rights and permissions, digital format of text, etc). Secondly, we will discuss some of the issues in creating a toolkit that are not necessarily NLP-centric, e.g. client-server architecture points, implementation issues, and system-dependent usability issues such as speed. Thirdly, we will review the various types of evaluation that we have considered, including formative, iterative, and summative. We have explored utility across several sets of users (ranging from our own computer science graduate students to highly trained catalogers), and have observations about the toolkit that apply either to specific user groups or to applications.

Future NLP Research in CLiMB

In our next phase, we will explore methods to add ranked terms for selection, using relations with high-quality domain specific thesauri, such as the AAT At the moment, no disambiguation is performed. We will explore the applicability of using machine-learning techniques over data collected from experiments with CLiMB output and catalogers. We will compare combinations of taggers and chunkers to find the optimal requirements for this application We will expand our texts to those having a less tightly-coupled relationship to an image, in order to push our techniques beyond tidy data, to the more intractable (such as the Web.) We will test utility with catologers ranging from the most naïve (namely our information studies graduate students) to more sophisticated (from our new CLiMB-2 partners). We will load two, and possibly three, collections into the CLiMB platform, to test with image searchers, not just with catalogers. Among our objectives are the creation of a set of client-side downloadable tools to enhance access by labeling descriptive metadata for review by experts as well as, ultimately, to enable sophisticated automatic analysis procedures for the wider digital library community.

Very Selected and Incomplete References (full set with final paper)

www.umiacs.umd.edu/~climb

Alembic Sentence **Tagger** (**MITRE**)
Brill **Tagger** (Eric Brill).
NP-Chunk

Davis, Peter T., David K. Elson, Judith L. Klavans. Methods for Precise Named Entity Matching in Digital Collections. Third ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2003

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown (1998). "Linear segmentation and segment relevance." *Sixth Annual Workshop on Very Large Corpora* (WVLC6). Montreal, Quebec, Canada, 1998.

Moëllic, Pierre-Alain, Patrick Hède, Gregory Grefenstette, Christophe Millet, "Evaluating Content Based Image Retrieval Techniques with the One Million Images CLIC TestBed", Proceedings of the Second World Enformatika Congress, WEC'05, February 25-27, 2005, Istanbul, Turkey, pp 171-174.

Passonneau, et al. (2006) Evaluation of CLiMB . LREC.

Town C. and D. Sinclair. Language-based querying of image collections on the basis of an extensible ontology. IVC, 22(3):251--267, March 2004

[i] CLiMB-1 was funded by the Andrew W. Mellon Foundation to Columbia University in the City of New York., 2002-2004. CLiMB-2 was funded by the Mellon Foundation to the University of Maryland, College of Information Studies, 2005-2007.

Author Index

OntoImage 2006
Workshop on Language Resources for Content-based Image Retrieval
during LREC 2006

Monday 22 May 2006
Magazzini del Cotone Conference Center,
GENOA - ITALY