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.

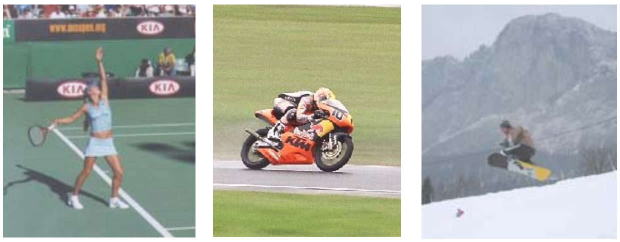


Figure 1: Examples for sports photos (Tennis, Motorcycling, Snowboarding)

[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/

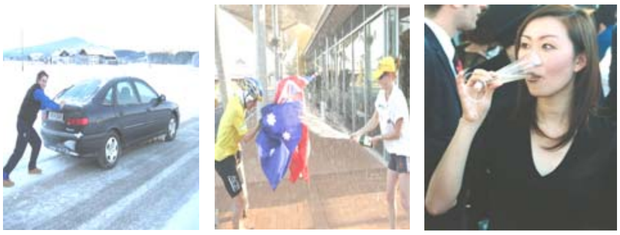


Figure 2: Examples for action pictures
(Pushing, Celebrating, Drinking)

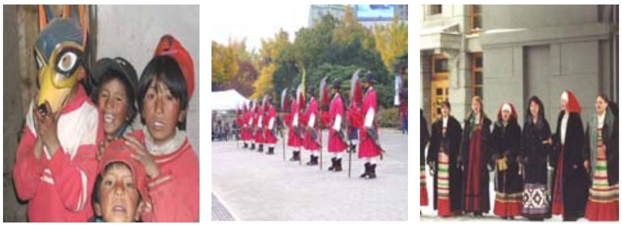


Figure 3: Examples for people shots
(Peruvian Children, Korean Guards, Russian Singers)

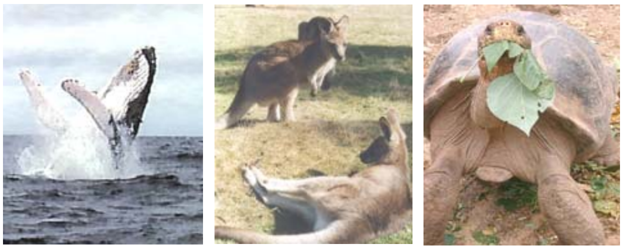


Figure 4: Examples for animal photos
(Humpback Whale, Kangaroos, Galapagos Giant Turtle)

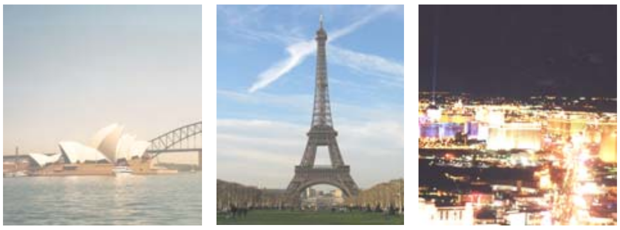


Figure 5: Examples for city pictures
(Sydney Opera House, The Eiffel Tower, Las Vegas Strip)

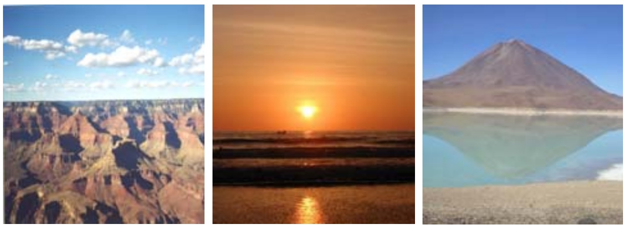


Figure 6: Examples for landscape shots
(Grand Canyon, Montañita Beach, Volcano Licancabur)

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.

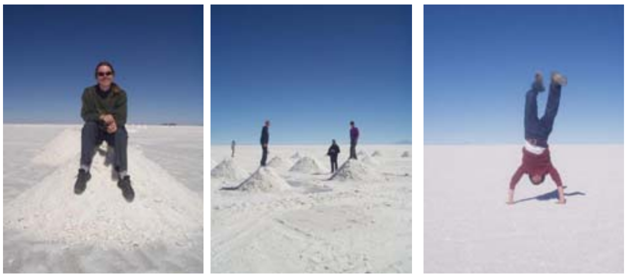


Figure 7: Tourists from three different tour groups at the Salt Lake of Uyuni in Bolivia

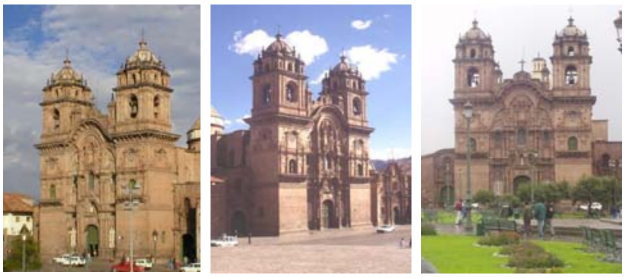


Figure 8: The Cathedral of Cuzco, Peru, in different viewing angles (right, left and front)

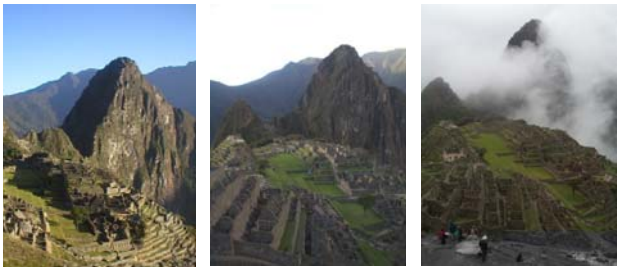


Figure 9: The Inca ruins of Machu Picchu in bright sunshine, on an overcast day and in foggy and rainy conditions

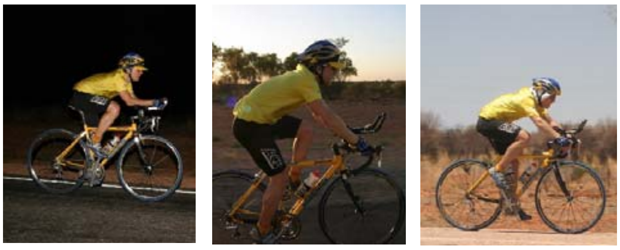


Figure 10: A cyclist riding a racing bike at night, in the morning and during the day

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.

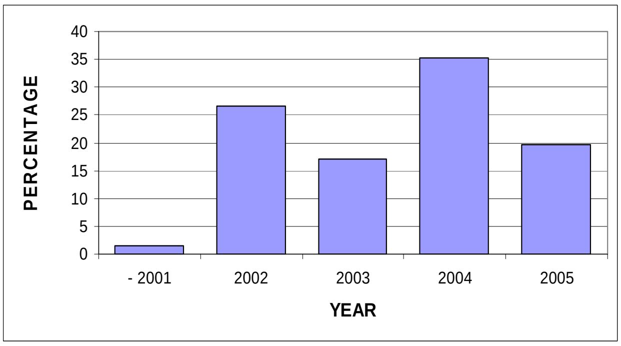


Figure 11: Temporal Range

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]):

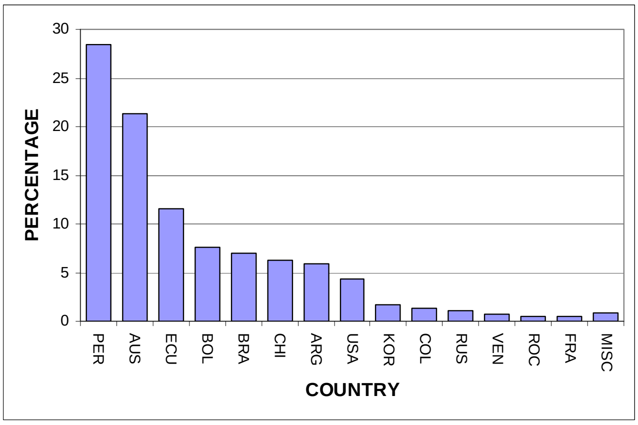


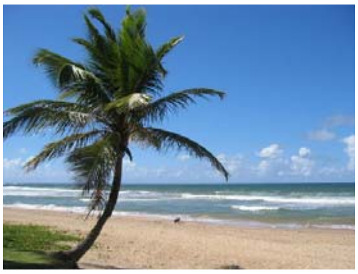
Figure 12: Variation across countries (with more than 100 images)

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

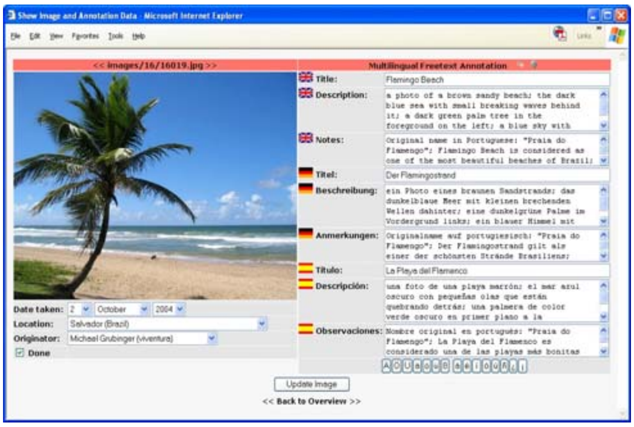
Figure 13: Example of an original annotation

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
Update Image
<< Back to Overview >>

<< images/16/16019.jpg >>

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;

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

Observaciones: Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas

Date taken: 2 October 2004

Location: Salvador (Brazil)

Originator: Michael Grubinger (viventura)

☑ Done

Figure 14: Benchmark Administration System

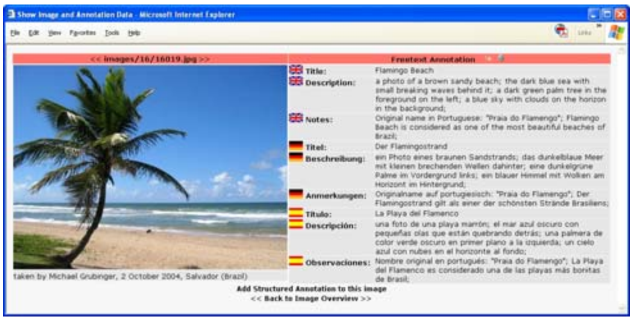
Fig. 14 shows a screenshot of a custom-built Benchmark Administration System used to carry out the revision process (see (Grubinger & Leung, 2004) for details of its specification, architecture and implementation). In particular, information provided about the location was checked and the image description 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
Add Structured Annotation to this image
<< Back to Image Overview >>

<< images/16/16019.jpg >>

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

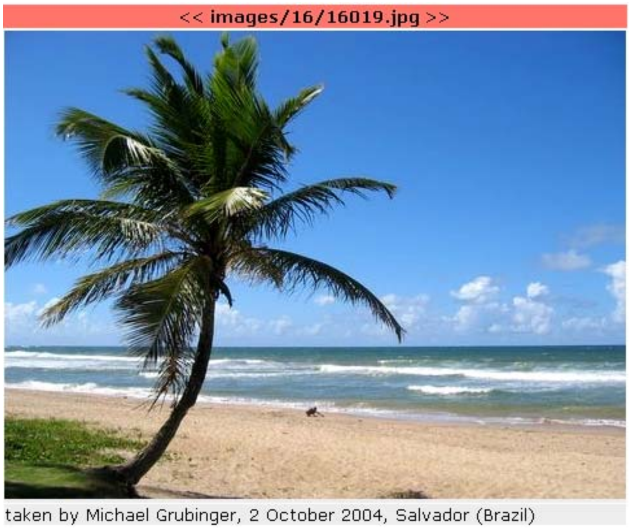
Descripción: una foto de una playa marrón; el mar azul oscuro con pequeñas olas que están quebrando detrás; una palmera de color verde oscuro en primer plano a la izquierda; un cielo azul con nubes en el horizonte al fondo;

Observaciones: Nombre original en portugués: "Praia do Flamengo"; La Playa del Flamenco es considerado una de las playas más bonitas de Brasil;

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

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).

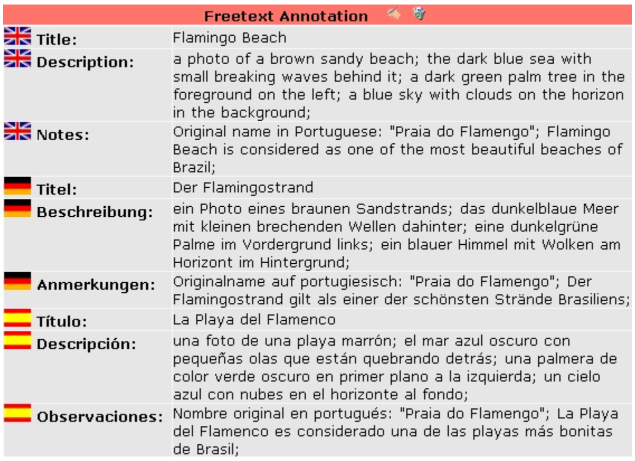


<< images/16/16019.jpg >>

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”).

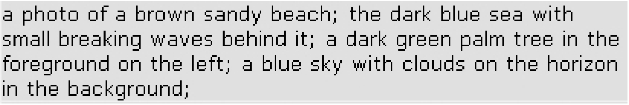
3.3.3. Description

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.

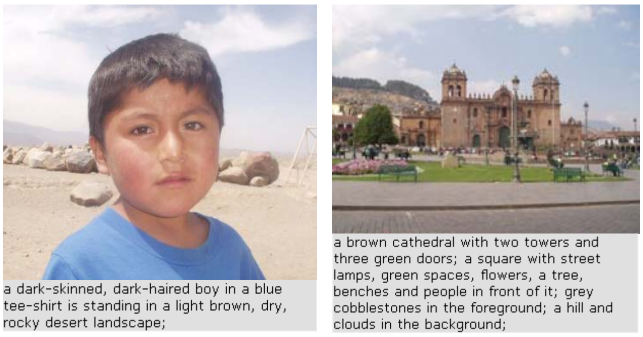
Table 5 provides an overview and a description of the symbols used in Tables 3 and 4.

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.

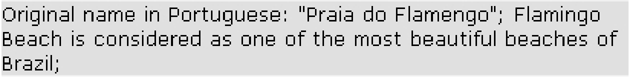
It is also known (Berlin & Kai, 1969) that significant differences exist between naming colours in different languages and cultures. For example, a kind of sea green, called “aoi” in Japanese, in English is generally regarded as a shade of "green", while in Japanese what an English speaker would identify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

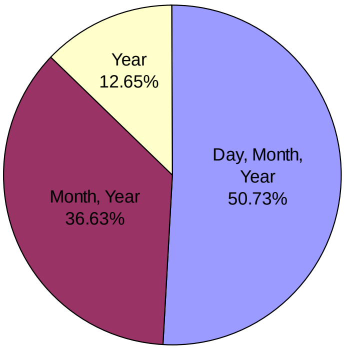


Figure 22: Percentages of the time granularity levels

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.