The Eurovision St Andrews Collection of Photographs

Paul Clough, Mark Sanderson

Department of Information Studies
University of Sheffield, UK.
[p.d.clough|m.sanderson]@sheffield.ac.uk

Norman Reid

University of St Andrews Library
University of St Andrews, UK.
nhr1@st-and.ac.uk

Abstract

This report describes the Eurovision image collection compiled for the ImageCLEF (Cross Language Evaluation Forum) evaluation exercise. The image collection consists of around 30,000 photographs from the collection provided by the University of St Andrews Library. The construction and composition of this unique image collection are described, together with the necessary information to obtain and use the image collection.

1 Introduction

St Andrews University Library holds one of the largest and most important collections of historic photography in Scotland exceeding over 300,000 photographs in size from a number of well-known Scottish photographers and photographic companies (Reid, 1999a). A cross-section of 30,000 images from the main collection has been part of a large-scale digitisation project to enable public access to the collection via a web interface[1] (Reid, 1999b).

The collection contains both colour and black-and-white photographs (the majority being B&W) taken by Scottish photographers or Scottish photography companies. Each image is accompanied by a textual caption which describes the content of the photograph, as well as other information considered useful by the St Andrews Library.

This report describes *St Andrews Collection* (SAC) – a core collection of the ImageCLEF evaluation exercise from 2002-2005 (Clough et al, 2005) – by first outlining the means of gathering the collection, followed by a description of the collection's content. Details of its transformation into a test collection are provided next, before the paper concludes with a description of how SAC is distributed.

2 Building the collection

To build the test collection, permission was granted by the University of St Andrews Library to 'scrape' its web site and download images to create a local version of the collection. The site presents the collection in a variety of ways: full text search; or browsing a list of 999 pre-defined index terms organised alphabetically and hierarchically via a categories page. The list of categories acts as a suitable starting point for gathering the collection.

Some images are assigned to more than one index category which means that during a trawl of the site, the same images may be found more than once. To build a local version of the St Andrews collection, all images were first downloaded via links from the index page, before a filtering stage was used to remove duplicate images but record the categories assigned to each image. St Andrews granted permission to download a thumbnail image, a larger version and the textual caption associated with each photograph. Further information about the download process can be obtained from the first author.

3 The contents of the collection

The SAC consists of 28,133 thumbnail images (around 120x76 pixels), larger versions of these images (around 368x234 pixels), and associated captions, giving a total of 84,399 files in the main body of the collection. In this section, we provide more information on a number of characteristics of the collection and in particular a breakdown of image distribution across these characteristics.

3.1 Captions

Each photograph has a caption which consists of the following eight fields (see, Figure 2): (1) a unique record number, (2) a short title, (3) a full title, (4) a textual description of the image content, (5) the date when the photograph was taken (most frequently with

Figure 1: The initial page of pre-defined index terms

[1] On-line access to the St Andrews collection, now offering an increased selection images, is available from http://specialcollections.st-and.ac.uk. Example screenshots are shown in Figure 2. As of 2005, the collection holds of 50,000 images. Both the in-house and the online software are provided by iBase Media Services Ltd.

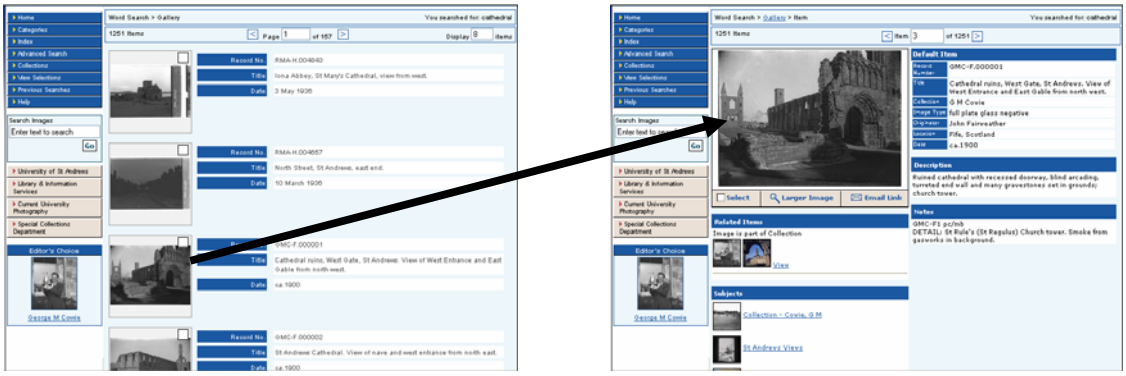

Figure 3: The web interface to the St Andrews collection

the day, month and year), (6) the originator, i.e. the name of an individual or company to which the photograph is attributed, (7) the location of the photograph (e.g. the county and the country), and (8) a line for notes to offer additional information about the photograph.

Table 1 provides a breakdown of each textual field of the image caption (ignoring the record ID field). The %null indicates the number of captions which do not have an entry for that field (marked as null or left blank), the mean number of words shows the *geometric mean* rather than the arithmetic because the former is less affected by outlying values. For the mean, standard deviation, minimum and maximum field values, we compute these only across the fields which are *non-null*.

Also from Table 1, we notice that almost all captions have a title of some kind, but only around 81% of captions have a description field, hence the reason for including all fields in the collection upon which retrieval can be based. On average the description field is one sentence of around 15 words, although it can range from 1 (very infrequently) to 27 words. We observe that in most cases the description field is a grammatical sentence which may be of importance if using natural language processing on this field.

Figure 2: Example photographs illustrating actual image sizes, 120x76 pixels and 368x234

Field	%null	Number of words			
		Mean	Std Dev	Min	Max
Title	0.6	5.44	4.36	1	24
Short title	0.2	3.23	1.12	1	8
Description	19.0	15.08	4.38	1	27
Date	0.8	2.56	0.99	1	6
Originator	0.1	3.54	0.63	1	6
Location	0.3	2.13	0.67	1	7
Notes	0.1	6.58	22.09	1	421

Table 1: Statistics of the seven text fields

3.2 Image sizes

Photographs in the Eurovision St Andrews collection have not been modified in any way from the originals. We have found that not all images are exactly the same size and there exists some degree of variation which may or may not affect approaches incorporating content-based retrieval. Figure 3 shows the most frequent image sizes: 368x234 for a large landscape image, and 120x76 for the corresponding thumbnail version.

Table 2 provides a breakdown of image sizes across all 28,133 images. For larger images, we find 24,223 (86.1%) are landscape with a median width of 345 pixels, and height of 233 pixels. The remaining 13.9% of images are portrait with an average width of 235 pixels, and height of 340 pixels. For the thumbnail images, we find that 24,267 (86.2%) are landscape and 13.8% portrait with an average size of 119x81 pixels. (It is our understanding the discrepancy may result from some reference images having been cropped to remove unwanted background.)

Size & orientation	Width Mean (SD)	Med	Range	Height Mean (SD)	Med	Range
Large portrait	253.3 (20.1)	236	116-306	339.6 (41.2)	343	193-397
Large landscape	345.1 (34.5)	348	152-384	233.2 (18.8)	230	108- 284
Small portrait	119.7 (22.6)	120	62-120	80.0 (5.6)	80	36-92
Small landscape	81.7 (5.9)	80	48-90	119 (4.4)	120	64-120

Table 2: Variation in image sizes for both the large and thumbnail (small) image versions

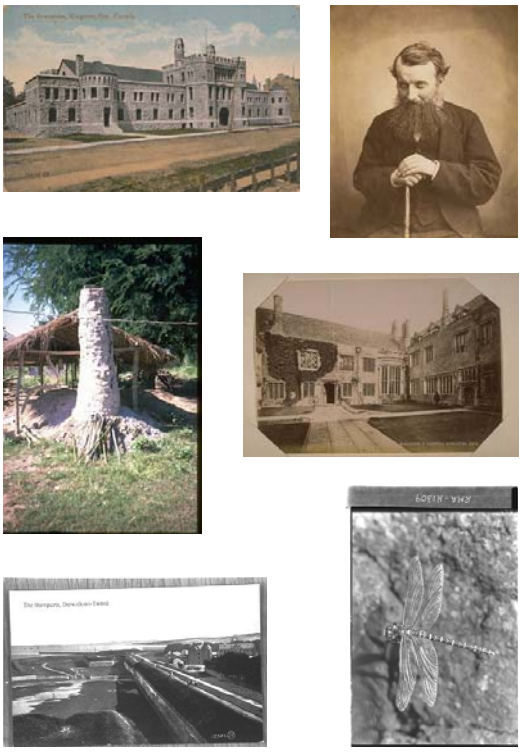

Figure 4: Example photographs illustrating various degrees of colour variation

3.3 Colour variation

The majority of images in the St Andrews collection are monochrome or black and white, due to the historic nature of the collection. There is, however, a small proportion of colour images. Figure 4 provides exemplars from the St Andrews collection demonstrating the range of image colour variation commonly found. To quantify the proportion of colour versus monochrome images in the St Andrews collection, images were classified into two groups using k-means clustering based on the number of unique colours found within them[2]. In general, we find colour image (including older colour images); anything below this we generally find to be monochrome or black and white. Clustering on this basis, we find 11% of images are grouped and represent colour images; the remaining 89% of images are monochrome or black and white.

3.4 Variation across date

The majority of photographs in the St Andrews collection were taken prior to 1940. Figure 5 shows the distribution of images across dates illustrating a large cluster around 1930 to 1940. We computed this by selecting the four digit year from the date part of the caption resulting in 27,723 year values. The earliest photograph was taken in 1839; the latest in 1992 (a range spanning 160 years). The mean date is 1920 (standard deviation is 26.2) and the median 1931.

3.5 Categories

Each image in the St Andrews collection has been assigned to one or more descriptive categories. Some categories are fairly general, e.g. “airports”, “airships”, “flowers”, “beach scenes” and “breweries”. However, other categories are more specific, e.g. “cattle – oxen”, “Collection – J. Valentine and Co.”, “dress – uniforms – paramilitary”, “Fife urban views” and “golf ball manufacture”. On average, each image is assigned to four categories (mean is 4.37, median 4.0 and standard deviation 1.631). The majority of images are assigned to three and four categories, with a smaller proportion assigned one or two. The maximum number of categories assigned is nine. Figure 6 shows the variation of categories assigned to images.

3.6 Collection content

An analysis of query logs of searches made of the collection revealed that topics typically searched for were as follows

- Social history, e.g. old towns and villages, children at play and work.
- Environmental concerns, e.g. landscapes and wild plants.
- History of photography, e.g. particular photographers.
- Architecture, e.g. specific or general places or buildings.
- Golf, e.g. individual golfers or tournaments.
- Events, e.g. historic, war related.
- Transport, e.g. general or specific roads, bridges etc.
- Ships and shipping, e.g. particular vessels or fishermen.

[2] Colours were computed using PerlMagick, an OO interface to ImageMagick: www.imagemagick.org

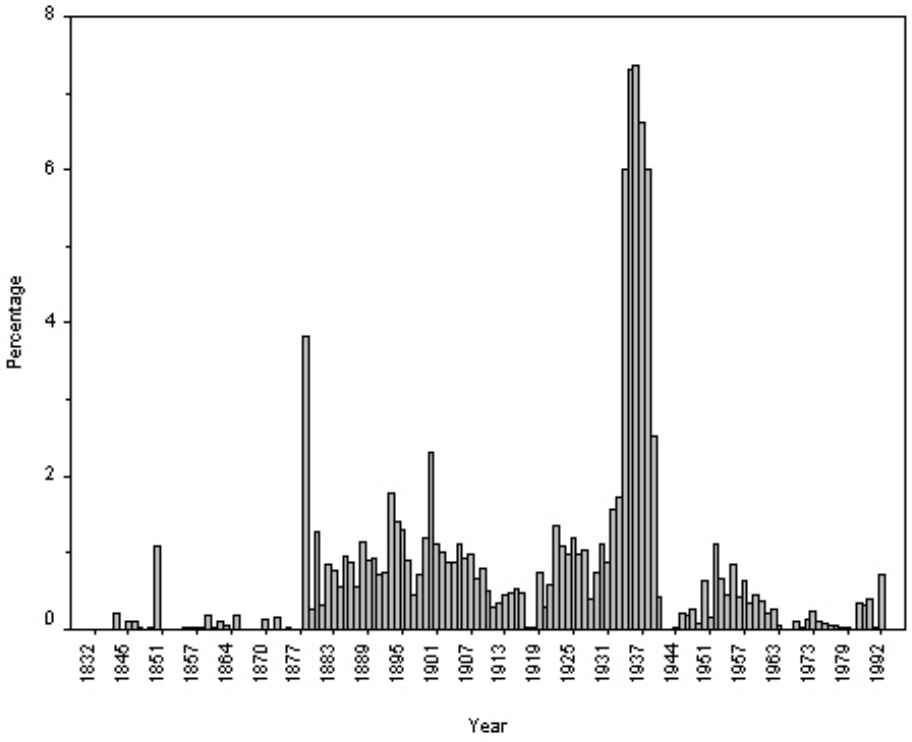

Figure 5: Distribution of photographs across years

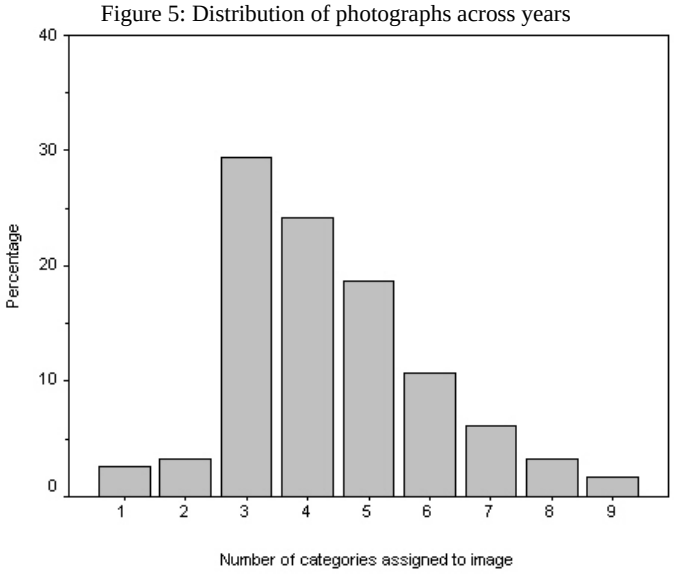

Figure 6: Distribution of numbers of categories assigned to each photograph

4 Distribution of the collection

Access to the St Andrews collection is mediated through CLEF: the Cross Language Evaluation Forum[3] where monolingual and cross-language topics as well as relevance judgements are also available; details are outlined in ImageCLEF overview papers (Clough et al, 2005). All files and directories that constitute the St Andrews collection are prefixed with "stand03" in anticipation of future releases of this collection (e.g. when more images have digitised). Rather than store all images and captions in one directory, they are grouped into 565 directories of variable size (this facilitates easier image browsing using standard file managers). No significance should be given to which images are stored in which

[3] http://www.clef-campaign.org/

Record ID	JV-A.006906
Title	Arbroath. The Harbour and Beach from west,
Short title	Harbour and Beach, Arbroath.
Description	Pebble beach before sea wall round bay; town with houses, works chimneys, buildings and castellated signal tower with flag.
Date	Registered 24 June 1938
Originator	J Valentine & Co
Location	Angus, Scotland
Notes	jf/pc/mbDETAIL: Children paddling by rock outcrops in water. Lifeboat station, with lifeboat moored at slipway beside harbour pier. Spire on wooded hill. ADD: The Signal Tower is now a local history museum, which features the story of the Bell Rock Lighthouse.
Categories	[signal towers], [flags & banners], [harbours], [beach scenes], [beach scenes], [lifeboat service], [beacons & lighthouses], [Angus all views], [Forfars all views], [Collection - J Valentine & Co]

Figure 7: Example caption and its image

directories, the number of images per directory, and the directory names themselves.

4.1 Images and captions

Images and their corresponding caption in the collection are given the same unique ID which corresponds to their filename as used to access the images in the on-line version of the St Andrews collection, e.g. “stand03_17750”. Captions are given the file extension “.txt”, thumbnail images the extension “.jpg” and large images the extension “_big.jpg”. For example, image “stand03 _17750” would be stored as “stand03_17750.txt”, “stand03_17750.jpg” and “stand03_17750_big. jpg” respectively.

Figure 7 shows an example semi-structured caption with its image. All captions are stored in plain text format with each line containing one field. The last line (or field) contains the categories assigned to the image, the text for each category surrounded by “[ ]” and multiple categories comma (i.e. [category1], …, [categoryN]).

4.2 The captions in TREC-format

As well as plain text captions, a single text file is also included in the release containing all captions encapsulated in an SGML format compatible with existing TREC collections. This text file (called “stand03_captions.trec”) contains the captions in annotated as shown in Figure 8.

The <DOCNO> tag contains a unique document reference identifier, in this case the pathname of the image. The rest of the caption fields are not annotated individually, except for the title indicated by the <HEADLINE> tag, the record identifier indicated by the <RECORD_ID> tag, and the categories indicated by the <CATEGORIES> tag. The thumbnail and large images are demarcated by the <SMALL_IMG> and <LARGE_ IMG> tags respectively.

To check the mark-up, the captions have been parsed with two TREC-compatible parsers, one our own parser, the other the TREC parser which comes with

```
<DOC>
<DOCNO>stand03_2096/stand03_10695.txt</DOCNO>
<HEADLINE>Departed glories - Falls of Cruachan Station above Loch Awe on the Oban line.</HEADLINE>
<TEXT>
<RECORD_ID>HMBR-.000273</RECORD_ID>
Falls of Cruachan Station.
Sheltie dog by single track railway below embankment, with wooden ticket office, and signals;
gnarled trees lining banks.
ca.1990
Hamish Macmillan Brown
Argyllshire, Scotland
HMBR-273 pc/ADD: The photographer's pet Shetland collie dog, 'Storm'.
<CATEGORIES>[tigers],[Fife all views], [gamekeepers],[identified male],[dress -
national],[dogs]</CATEGORIES>
<SMALL_IMG>stand03_2096/stand03_10695.jpg</SMALL_IMG>
<LARGE_IMG>stand03_2096/stand03_10695_big.jpg</LARGE_IMG>
</TEXT>
</DOC>

<DOC>
<DOCNO>stand03_2095/stand03_35.txt</DOCNO>
```

Figure 8: An example caption in TREC-format

Lemur[4]. In both cases, tags which are not recognised are ignored and text within these tags treated as part of the document itself and treated as valid output from the parser.

4.3 Other files released

As well as the captions and images, the collection contains a number of other "useful" files, which include the following:

- stand03_bigimages.txt – a list of all large images including their pathname.
- stand03_thumbnails.txt – a list of all thumbnails including their pathname.
- stand03_captions.txt – a list of all captions including their pathname.
- stand03_captions.trec – all captions in TREC format.

4.4 Known problems

One problem we are aware of in the collection is derived from the St Andrews web site. In a very small number of cases, we have found examples of large and small images which are not the same photograph. This problem also existed on the original on-line version of the St Andrews collection, and was the result of errors arising from data migration between earlier software versions. It has since been corrected in the current version of the on-line software and the error no longer occurs on the collection web site.

5 St Andrews collection at ImageCLEF

The St Andrews collection has been used for the past three years at ImageCLEF[5], the cross-language image retrieval task (Clough, Müller, Hersh, Deselaers, Lehmann, Grubinger, 2005; Clough, Müller, Sanderson, 2005; Clough and Sanderson, 2003). This is part of a wider cross-language evaluation campaign called CLEF[6], the Cross Language Evaluation Forum (Peters & Braschler, 2001). The St Andrews collection has been used for bilingual ad-hoc retrieval where queries typical to this kind of historic collection have been generated in English and translated into languages including a range of Indo-European, Asian and Romance languages. The St Andrews collection is typical of many photographic collections found in cultural heritage where domain specialists (e.g. historians or librarians) annotate images with specific attributes such as the name of the photographer, the date and description of the image for archival purposes. This contrasts with less structured collections such as the Web, shared photographic collections such as Flickr[7] and personal photographs.

Year	No. topics	Total relevant	No. participants	No. runs
2003	50	2271	4	45
2004	25	829	12	190
2005	28	1916	11	349

Table 3: summary of the ImageCLEF bilingual ad-hoc task 2003-2005

As a retrieval task, cross-language image retrieval encompasses two main research areas: (1) image retrieval, and (2) cross-language information retrieval (CLIR). The St Andrews collection is particularly challenging for visual-based image retrieval techniques due, in part, to variety in the images' composition and predominately non-colour appearance. For CLIR, challenges include: captions which are short in length increasing the likelihood of vocabulary mismatch, captions with text not directly associated with the visual content of an image (e.g.

[4] The CMU Lemur toolkit, www.cs.cmu.edu/~lemur.
[5] http://ir.shef.ac.uk/imageclef/
[6] http://www.clef-campaign.org/
[7] http://www.flickr.com/

expressing something in the background), and the use of colloquial and domain-specific language in the caption (i.e. British English).
Table **3** summarises the use of the St Andrews collection in ImageCLEF.

6 Acknowledgements

The authors would like to acknowledge the UK Engineering and Physical Sciences Research Council for financial support for the Eurovision project (GR/R56778/01, Sanderson and Clough, 2002). We also thank the University of St Andrews Library for granting us permission to use its images as the basis for SAC.

7 Summary

The Eurovision St Andrews collection is a unique collection of around 30,000 photographs with significant historic value. The collection is well-suited to image retrieval via captions because each image in the collection is accompanied by a manually-created semi-structured textual description. The collection offers a unique opportunity for information retrieval researchers to create a realistic test collection for monolingual and cross-language image caption retrieval of reasonable size and with a wide variety of content. The additional category information would lend itself well to other areas of content-based image analysis, e.g. image classification. As part of CLEF, this test collection will offer researchers worldwide the opportunity to experiment and further investigate methods of image retrieval.

8 References

Clough, P., Müller, H., Hersh, W., Deselaers, T., Lehmann, T. and Grubinger, M. (2005), Overview of the 2005 cross-language image retrieval track (ImageCLEF), In the *working notes of the CLEF workshop, Vienna, Austria, 21-23 September 2005*

Clough, P., Müller, H. and Sanderson, M. (2005), The CLEF 2004 Cross Language Image Retrieval Track, In Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Eds (Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B.), *Lecture Notes in Computer Science* (LNCS), Springer, Heidelberg, Germany, Volume 3491/2005, 597-613.

Clough, P. and Sanderson, M. (2003), The CLEF 2003 cross language image retrieval track, In *Proceedings of Cross Language Evaluation Forum (CLEF) 2003*, Trondheim, Norway.

Reid, N.H. (1999). The photographic collections in St Andrews University Library, *Scottish Archives*, vol. 5, 83-90

Reid, N.H. (1999). Photographic archives: Aberdeen, Dundee and St Andrews, *Making information available in digital format: perspectives from practitioners*, ed T. Coppock, Edinburgh, The Stationery Office, 1999, 106-119.

Peters, C., Braschler, M (2001). Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067-1072, 2001.

Sanderson, M., Clough, P. (2002). Eurovision – an image-based CLIR system, Workshop held at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, *Workshop 1: Cross-language Information Retrieval: A Research Roadmap*, Finland, August 15th 2002, 56-59.