

The Eurovision St Andrews Photographic Collection (ESTA)

Paul Clough^α, Mark Sanderson^α and Norman Reid^β
(February 2003)

Department of Information Studies^α
University of Sheffield, UK.
p.d.clough,m.sanderson@sheffield.ac.uk

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

Abstract

This report describes the Eurovision image collection compiled for the Cross Language Evaluation Forum (CLEF) cross language retrieval in image collections (image CLEF) pilot experiment. The image collection consists of around 30,000 images from the photographic collection provided by St Andrews University Library. The construction and composition of this unique image collection are described in this report, together with the necessary information to use the image collection.

1. Introduction

The Eurovision St Andrews photographic collection (hereafter known as ESTA) forms the basis of a unique test collection used for a pilot experiment in the 2003 Cross Language Evaluation Forum (CLEF¹) called *Image CLEF*. 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 [1]. 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 interface² [2].

ESTA 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 St Andrews Library.

This report describes ESTA, the first stage in building a test collection for information retrieval research. This work is part of the Eurovision project [3].

2. Building the collection

To build the test collection, permission was granted by St Andrews University Library to trawl the web interface and download images to create a local version. The collection can be accessed in a variety of ways from a web interface, including a structured search and via a web page with a list of 999 pre-defined categories³ called the *index* page. The list of categories (Figure 1a) provides a way into the image collection as each link provides access to one or more images, suitable as a starting point for automatically trawling the collection. For example, the category “Abbeys and priories” links to the page shown in Figure 1b which in turn links to a page with a larger image and textual caption (Figure 1c).



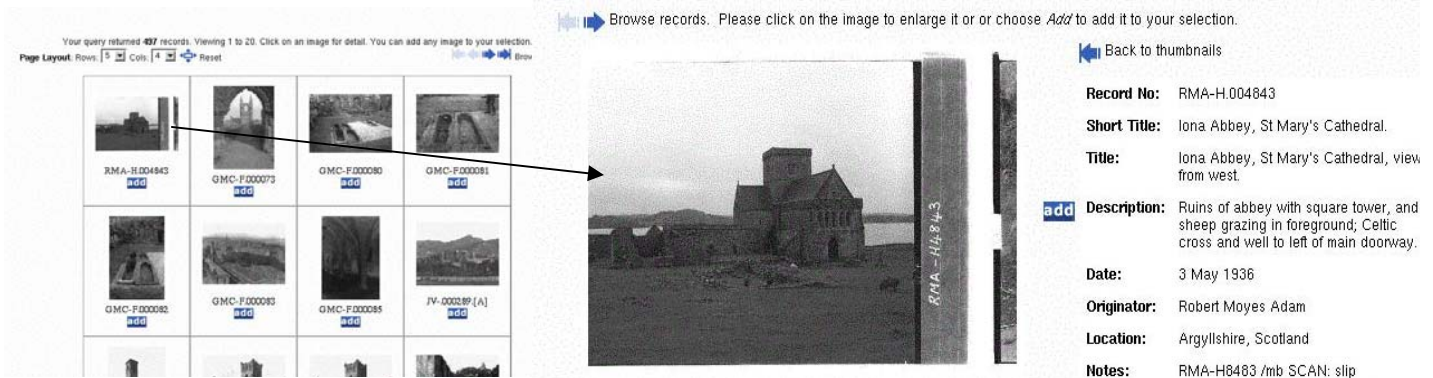
abbeyes & priories	503 items
Abers all views	724 items
aerial views	43 items
aerodromes	22 items
Ailsa Craig views	12 items
air force	175 items
airports	5 items
airships	1 items
airshows	72 items
Alberta all views	8 items
ambulance service	3 items
amphitheatres	7 items
Ang all views	22 items
angels	18 items
angling	44 items
Angus all views	908 items
animal skins	4 items
animal statuary	38 items

1a. The initial page of pre-defined categories

¹ CLEF2003: <http://clef.ici.pi.it:2002/>

² On-line access to the St Andrews collection is available from: <http://specialcollections.st-and.ac.uk>

³ See: <http://specialcollections.st-and.ac.uk/photo/controller/>



1b. Images for each category

1c. Image and caption

Figure 1: The web interface to the St Andrews collection

Some images are assigned to more than one index category because of cross-referencing which means that during a trawl of the site the same image may be found more than once. For example, image JV-.019165 is categorised with the category ID 1974 and 112. This is because it is a photograph by J.Valentine (category ID 112) of a scene in London (category ID 1974). 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 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.

think the title and description fields are likely to offer the most useful information for image retrieval, however the existence of these other fields also enables an IR system to implement structured searches.

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
Photographer	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 from 7 textual fields of the image captions

3. The contents of the collection

ESTA 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 8 fields (see, e.g. Figure 1c): (1) a title, (2) a short title, (3) a unique record number (unique with respect to the St Andrews collection), (4) a textual description of the image content, (5) the date when the photograph was taken (most frequently with 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. We

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 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.



2a. Larger version (368x234)



2b. Thumbnail version (120x76)

Figure 2: Example photographs illustrating actual image size

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 approaches to retrieval. Figures 2a and 2b show the most frequent image sizes: 368x234 for a large landscape image, and 120x76 for the corresponding thumbnail version. The thumbnail image was included because the Eurovision St Andrews collection was originally used to base an initial retrieval system upon where the thumbnails were used in the same way as the St Andrews web interface: to provide enough to show the user, but enable a page of images to be shown relatively quickly.

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 (rather oddly) that 24,267 (86.2%) are landscape and 13.8% portrait with an average size of 119x81 pixels.

Orient.	Width			Height		
	Mean (SD)	Med	Range	Mean (SD)	Med	Range
Large & Port.	253.3 (20.1)	236	116-306	339.6 (41.2)	343	193-397
Large & Land.	345.1 (34.5)	348	152-384	233.2 (18.8)	230	108-284
Small & Port.	119.7 (22.6)	120	62-120	80.0 (5.6)	80	36-92
Small & Land.	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

3.3. Colour variation

The majority of images in the St Andrews collection are monochrome or black and white, due to the nature of the collection being a collection of historic photographs. There are, however, a small proportion colour images from a range of eras which present a varying degrees of quality with which the images visually appear.



3a. Example colour images



3b. Example monochrome/B&W images

Figure 3: Example photographs illustrating various degrees of colour variation

Figure 3 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 Eurovision St Andrews collection, images were classified into two groups using k-means clustering based on the number of unique colours found within them⁴. In general, we find that more than approximately 20,000 colours suggest a 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 Eurovision St Andrews collection were taken prior to 1940. Figure 4 shows the distribution of images across dates illustrating a large cluster around 19-30 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 1832; the latest in 1992 (a range spanning 160 years). The mean date is 1920 (standard deviation is 26.2) and the median 1931.

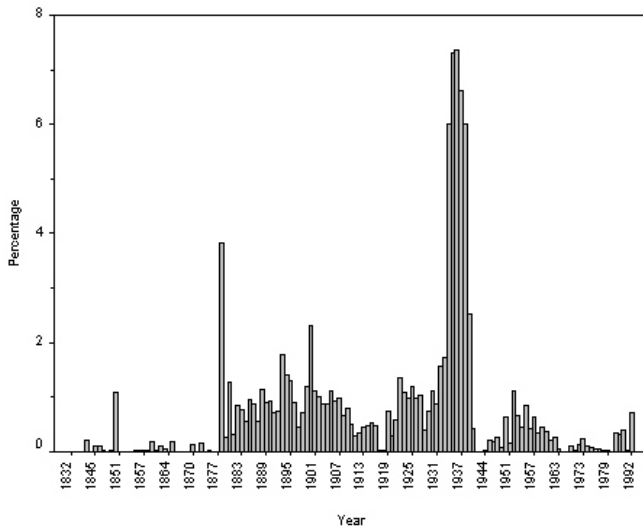


Figure 4: Distribution of photographs across years

3.5. Variation across 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

⁴ Colours were computed using PerlMagick, an OO interface to ImageMagick (<http://www.imagemagick.org>)

“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 4 categories (mean is 4.37, median 4.0 and standard deviation 1.631). The majority of images are assigned to 3 and 4 categories, with a smaller proportion assigned 1 or 2. The maximum number of categories assigned is 9. Figure 5 shows the variation of categories assigned to images.

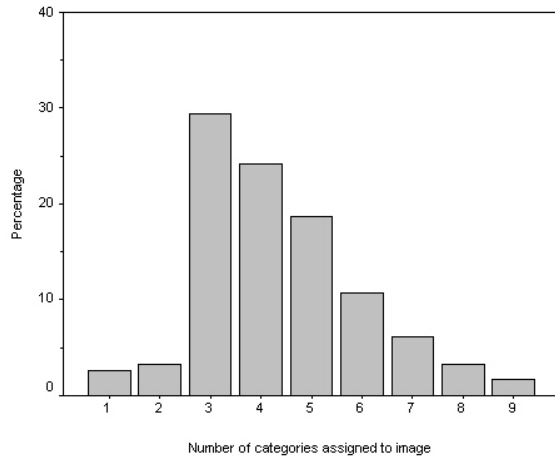


Figure 5: Distribution of number of categories assigned to each photograph

4. Distribution of the collection

The Eurovision St Andrews collection is distributed on 2 CDs containing the images, captions and documentation, in a manner similar to the TREC⁵ (Text REtrieval Conference) test collections. The monolingual and cross-language topics required for *Image CLEF* will be available in the near future. All files and directories that constitute the Eurovision St Andrews collection are prefixed with “stand03” to separate these from 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 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 on-line version of the St Andrews

⁵ TREC: <http://trec.nist.gov>

Record ID	JV-A.006906
Title	Arbroath. The Harbour and Beach from west, Harbour and Beach, Arbroath.
Short title	Pebble beach before sea wall round bay; town with houses, works chimneys, buildings and castellated signal tower with flag.
Description	Registered 24 June 1938
Date	J Valentine & Co
Originator	Angus, Scotland
Location	JV-A6906
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 6: Example caption and its image

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 6 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 on the CD 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 7.

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 Lemur⁶. 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.

```
<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 em-
bankment, 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</SMAL
L_IMG>
<LARGE_IMG>stand03_2096/stand03_10695_big.jpg</
LARGE_IMG>
</TEXT>
</DOC>

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

Figure 7: An example caption in TREC-format

4.3. Contents of the CDs

As well as the captions and images, the CDs contain a number of other “useful” files under the “docs” directory 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.
- **stand03_guide.pdf** – this document.

⁶ The CMU Lemur toolkit (<http://www.cs.cmu.edu/~lemur>).

4.4. Known problems with the collection

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 exists on the on-line version of the St Andrews collection.

5. 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). We would also like to thank St Andrews University Library (in particular Norman Reid) for allowing the University of Sheffield and CLEF2003 to access their photographic collection. We also thank St Andrews for granting us permission to use images from the on-line collection as the basis for ESTA.

6. 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 semi-structured textual description created manually for each image. 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.

7. References

- [1] Reid, N.H. (1999). "The photographic collections in St Andrews University Library", *Scottish Archives*, 1999 (vol. 5), pp83-90
- [2] 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, pp 106-119.
- [3] Sanderson, M. and 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-