# IAM@ImageCLEFphoto 2009: Experiments on Maximising Diversity using Image Features

Jonathon S. Hare, David P. Dupplaw, Paul H. Lewis
Intelligence Agents Multimedia Group
School of Electronics and Computer Science, University of Southampton, Southampton, UK
{jsh2|dpd|phl}@ecs.soton.ac.uk

**Abstract**

This paper describes the diversity enabled retrieval system constructed at Southampton for the ImageCLEFphoto 2009 task. The retrieval system used Terrier as the underlying textual indexing and retrieval system, and combined it with a technique for re-ranking the results by maximising the visual dissimilarity of retrieved images. The results show that our visual re-ranking methods does indeed work at increasing the diversity in the top results, however, at the same time it causes a slight drop in precision. The text-based approach designed for handling the 'part 1 topics' of the task is also shown to perform very well.

### Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

### General Terms

Image search, Diversity, Measurement, Performance, Experimentation

### Keywords

Image Content Analysis, Data Fusion, Content-based Image Retrieval

## 1 Introduction

The 2009 ImageCLEF photo retrieval task [9] aimed to promote diversity in image search. The was performed using a set of nearly 500,000 captioned images provided by the Belga photo agency. The task incorporated two separate query types. The part 1 topics were described as a main topic (i.e. 'David Beckham'), together with a set of clusters or sub-topics (i.e. 'Manchester United', 'Real Madrid', etc.). The part 1 topics also included detailed information about what might be expected in the results of a search for each of the clusters. The part 2 topics provided a single topic with no context. Both part 1 and 2 topics included example images which could be used for content-based search or classification.

In the 2009 ImageCLEF photo retrieval task, Southampton's baseline system used standard text retrieval techniques for the part 2 topics. The baseline handling of the part 1 topics augmented the standard text search with multiple sub-queries (one per cluster) followed by a merge phase in order to build a complete ranking for the topic. On top of the baseline system we developed a re-ranking procedure for the results lists that leveraged visual features extracted from the images and attempted to re-order the list such that the first images in the list were highly visually dissimilar.

# 2 Methodology

The overall methodology for tackling the task involved building a baseline retrieval system using only the textual captions, and then augmenting the search results generated by the baseline system with information extracted from the actual content of the images in order to promote a diverse spectrum of different images near the top of the ranked search results. Each of the different aspects of this methodology is described below.

## 2.1 Text-based Baseline System

The Terrier text retrieval system [8], developed at the University of Glasgow, was used as the underlying text search technology for our submissions. In particular, we adapted Terrier to index the image captions, and modified the search algorithm based on the particular query formulations in the task.

### 2.1.1 Indexing of Captions

The caption tokeniser was configured to tokenise all of the text in the input record after the document identifier. The tokeniser took any non-alphanumeric character, or run of non-alphanumeric characters, as being a token separator. Each token was converted to lowercase, and tokens with more than four digits or three consecutive instances of the same letter were rejected. Tokens matching the standard Terrier list of stopwords were also discarded. For experimentation, we build two separated indices; one with the Porter Stemmer [10] feature enabled, and another without.

### 2.1.2 Retrieval using the Indexed Captions

For retrieval, Terrier was configured to use the standard TF-IDF weighting model (based on Robertson's definition of TF and the standard Sparck-Jones IDF definition [3]). Within the photo retrieval task, there were two sets of queries or topics. Each query in the first set contained a title, together with a number of explicit *clusters*, which themselves included a title ("clusterTitle"), description ("clusterDesc") and sample images. The second set of queries only contained a single title together with three sample images relevant to that title. The retrieval methods we used were necessarily different for the two sets of query modalities, however, the search and retrieval process is completely automatically driven by the provided topic files.

**Part 1 topics.** This year, we only considered the cluster titles for building our search. Because the cluster titles also contained the terms used in the overall topic title, for the purposes of reporting it is assumed that that the topic title is used as well. For each of the topics, a three stage process was used to generate the results:

1. Convert each "clusterTitle" into a Terrier query such that all words with a '-' **must not** appear anywhere in the captions of the returned images, and **all** the remaining words **must** appear in the captions of the result documents.

2. Query the index using each of the generated queries from the cluster titles in turn and store the results.

3. Merge the results lists in a round-robin fashion, ignoring the scores assigned by Terrier (i.e. just using ranked position). The top ranked results from each of the sub-queries will come first, followed by the second most relevant images, and so on. Duplicate images (i.e. those retrieved from more than one of the searches) are also filtered, so only the higher ranking example is retained. At this point we also gave each image an arbitrary score based on its position in the sub-query search; in our implementation all of the top-ranked images from the sub-queries scored 4000 and the second ranking images scored 3999, etc.

**Part 2 topics.** The second set of queries was processed in a much simpler manner; basically, results were generated by feeding the title into a Terrier query (marking all terms in the title field as required, as in the part 1 topics). No attempt to improving diversity by further analysis of the textual information was made for these topics.

## 2.2 Enhancing Result-set Diversity with Image Features

We hypothesise that the use of image features could give a large boost to the diversity of a result set. In particular, in our approach we developed a technique that re-ranks a list of search results by maximising the visual dissimilarity of the top-ranking images.

### 2.2.1 Image Features

The image feature used in the submitted experiments was a visual-term representation based on quantised SIFT features extracted from a multiscale difference-of-Gaussian pyramid [4]. The features were quantised to a vocabulary of 3125 terms [11, 2]. The codebook for the vector quantiser was learnt using a hierarchical K-means algorithm [6] (5 levels with 5 clusters per node); Due to time constraints, we used a pre-existing codebook that we had previously trained on the 5000 training images from ImageCLEF 2009 photo annotation task [7]. We also generated similar features using the MSER algorithm [5] coupled with quantised SIFT and Colour-SIFT [1] features, however, again for time reasons these were not included in the submitted results.

### 2.2.2 Re-ranking Algorithm

In order to try and improve the diversity of the result set, we propose a technique that incorporates visual information into the ranking. The proposed algorithm works by maximising the distance between the set of already re-ranked images, $R$, and an image from the ranked list of images retrieved from the text-based search, $I$. The distance between the feature-vectors of a set of images, $R$, and the feature-vector from a single image, $f_q$, is calculated using Equation 1. The function $d(.,.)$ in Equation 1 can be any distance function that compares two vectors; in this work we chose to use the Euclidean distance.

$$D(f_q, R) = \prod_{f_r \in R} d(f_q, f_r) \tag{1}$$

Algorithm 1 shows the steps taken to re-order the results from a text-based search using image feature-vectors. The output is a list containing all of the input images, but in a different order. Note that the algorithm treats the first input image as a special key, and that that image will also appear in the rank 1 position in the output.

**Algorithm 1**: Re-ranking by maximising visual dissimilarity.

**input** : A ranked list of $n$ images, $I = I_1...I_n$, from the text-based search
**output**: A re-ranked list of $n$ images, $R = R_1...R_n$

```
begin
    Construct an empty list R
    Add the first element of I, I_1, to R, and remove it from I
    while I is not empty do
        Find I_x from I such that D(I_x, R) (from Equation 1) is maximised
        Add I_x to R and remove it from I
    return R
end
```

**Application to Part 1 Topics.** In our experiments, we applied the visual re-ranking procedure individually to each of the results sets formed from the sub-searches created from the cluster titles. The re-ranked sub-result-sets were then merged as with the text-based search for part 1 topics.

**Application to Part 2 Topics.** Re-ranking of the results was simply a matter of applying the algorithm to the result set formed through the part 2 topics text search.

# 3 Experiments, Results and Discussion

We submitted four runs to the task organisers. The run parameters were configured by controlling the application of the Porter Stemmer in the text indexing and retrieval stage, and applying, or not applying, the visual re-ranking. The run titles and their configurations can be seen in Table 1.

| Diversity Re-ranking Configuration | Terrier Configuration | |
|---|---|---|
| | Porter Stemmer | No Stemming |
| None | SOTON1_T_CT_TXT | SOTON2_T_CT_TXT |
| Visual Features | SOTON1_T_CT_TXT_IMG | SOTON2_T_CT_TXT_IMG |

Table 1: Matrix showing the configurations of our submitted experimental runs. The mean, median, min and max are calculated over all the submitted results from each of the participants.

## 3.1 Preliminary Results Analysis

Table 2 shows a summary of Southampton's results calculated over all the topics from both parts 1 and 2. All of the results are significantly above the mean scores calculated from averaging the results of all runs submitted by all participants. However, the MAP scores in particular are still significantly below the maximum achieved by other participants.

The results show that the application of the visual re-ranking algorithm causes a drop in precision, which is however traded off by an increase in the cluster recall, implying a higher diversity in the ranked results. Turning off the Porter stemmer gives a slight boost in precision.

We hypothesise that turning off the stemmer has this effect because a number of the queries contained names of entities which would be adversely affected by the stemmer, and produce many irrelevant images in the result set. As an example, consider the topic which contained the term "fortis". Using the Porter stemmer, this is indexed as "forti". Unfortunately, the number "forty" is also indexed as "forti", and so searching for "fortis" will include many irrelevant results from images that had nothing to do with "fortis", but did have "forty" in their captions.

Results from only the part 1 topics are shown in Table 3. The table shows that the run entitled 'SOTON2_T_CT_TXT' achieved the highest F-measure (which measures the combined precision and cluster recall after 10 retrieved documents) and P10 scores from all the submitted results. Again, the effect of the visual re-ranking is to increase cluster recall, but decrease precision. Turning off the stemmer gives a slight precision boost (in the first few retrieved documents) and also boosts cluster recall slightly.

In the part 2 topic results shown in Table 4, the trend of increased cluster recall with the addition of the visual re-ranking holds true. The same is true of the stemming; disabling stemming increases precision slightly. One difference with the part 1 results however, is that the F-measure statistic reports slightly better results with stemming enabled.

| Run Name | S | V | MAP | F-measure | P@10 | P@20 | CR@10 | CR@20 |
|---|---|---|---|---|---|---|---|---|
| *Mean* | | | *0.294* | *0.585* | *0.655* | *0.455* | *0.547* | *0.623* |
| *Median* | | | *0.330* | *0.629* | *0.754* | *0.506* | *0.557* | *0.641* |
| *Min* | | | *0.003* | *0.095* | *0.068* | *0.019* | *0.158* | *0.206* |
| *Max* | | | *0.506* | *0.809* | *0.848* | *0.691* | *0.824* | *0.862* |
| SOTON1_T_CT_TXT | y | n | 0.372 | 0.720 | 0.802 | 0.648 | 0.653 | 0.718 |
| SOTON1_T_CT_TXT_IMG | y | y | 0.332 | 0.716 | 0.720 | 0.567 | 0.711 | 0.770 |
| SOTON2_T_CT_TXT | n | n | 0.379 | 0.729 | 0.824 | 0.649 | 0.654 | 0.711 |
| SOTON2_T_CT_TXT_IMG | n | y | 0.339 | 0.727 | 0.746 | 0.566 | 0.745 | 0.767 |

Table 2: Summary of results over all part 1 and part 2 queries.

| Run Name | S | V | MAP | F-measure | P@10 | P@20 | CR@10 | CR@20 |
|---|---|---|---|---|---|---|---|---|
| *Mean* | | | *0.297* | *0.600* | *0.677* | *0.448* | *0.558* | *0.640* |
| *Median* | | | *0.347* | *0.639* | *0.768* | *0.533* | *0.574* | *0.646* |
| *Min* | | | *0.001* | *0.033* | *0.028* | *0.010* | *0.041* | *0.072* |
| *Max* | | | *0.513* | *0.818* | *0.868* | *0.658* | *0.829* | *0.877* |
| SOTON1_T_CT_TXT | y | n | 0.361 | 0.784 | 0.824 | 0.603 | 0.747 | 0.810 |
| SOTON1_T_CT_TXT_IMG | y | y | 0.322 | 0.776 | 0.760 | 0.535 | 0.793 | 0.832 |
| SOTON2_T_CT_TXT | n | n | 0.371 | 0.818 | 0.868 | 0.601 | 0.773 | 0.820 |
| SOTON2_T_CT_TXT_IMG | n | y | 0.333 | 0.805 | 0.804 | 0.528 | 0.806 | 0.842 |

Table 3: Summary of results for all part 1 queries.

| Run Name | S | V | MAP | F-measure | P@10 | P@20 | CR@10 | CR@20 |
|---|---|---|---|---|---|---|---|---|
| *Mean* | | | *0.288* | *0.569* | *0.632* | *0.460* | *0.542* | *0.614* |
| *Median* | | | *0.307* | *0.639* | *0.740* | *0.496* | *0.574* | *0.642* |
| *Min* | | | *0.001* | *0.102* | *0.072* | *0.017* | *0.096* | *0.115* |
| *Max* | | | *0.530* | *0.819* | *0.836* | *0.724* | *0.819* | *0.876* |
| SOTON1_T_CT_TXT | y | n | 0.383 | 0.652 | 0.780 | 0.693 | 0.560 | 0.626 |
| SOTON1_T_CT_TXT_IMG | y | y | 0.342 | 0.653 | 0.680 | 0.600 | 0.629 | 0.708 |
| SOTON2_T_CT_TXT | n | n | 0.387 | 0.635 | 0.780 | 0.696 | 0.594 | 0.602 |
| SOTON2_T_CT_TXT_IMG | n | y | 0.345 | 0.648 | 0.688 | 0.604 | 0.613 | 0.692 |

Table 4: Summary of results for all part 2 queries.

# 4 Discussion, Conclusions and Future Possibilities

The multiple sub-query and merge approach for the part 1 topics clearly works very well. The approach taken for the part 2 topics suffers from a lack of precision in the retrieved result sets from the text retrieval.

The visual re-ranking algorithm described in section 1 has been shown to work as planned through the increased cluster recall scores it is able to produce; however, at the same time it causes a drop in F-measure because precision at the top-end of the result list also drops. One possible remedy to this problem would be to improve the precision of baseline text retrieval system so that fewer irrelevant images get passed into the re-ranking algorithm. Another possible approach would be to incorporate the retrieval score of the text-retrieval phase into the re-ranking so that images that were predicted to be more relevant still appear higher in the final result list.

Turning off the Porter stemmer gives a small boost in performance. We need to do some more analysis of the results, however, we hypothesise that the reasons are attributable to the problem of stemming named entities as described earlier. A future modification to the textual indexing and query processors might be to incorporate natural language processing (NLP) techniques to automatically detect named entities and not stem them, whilst still using stemming for other words.

In the experiments described in this paper we only used a single form of visual feature. It would be interesting to repeat the experiments in the future using a broader spectrum of visual features, and to also look at combining various features. There are also possibilities for using the sample images that were provided as part of the topic specification to help drive the search using both content-based techniques, and perhaps query expansion or automatic relevance feedback using information in the captions belonging to those images.

# Acknowledgements

The authors wish to thank the European Union, which supported this work under the Seventh Framework project LivingKnowledge (IST-FP7-231126) and the LiveMemories project, graciously funded by the Autonomous Province of Trento (Italy).

# References

[1] Gertjan J. Burghouts and Jan-Mark Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48 – 62, 2009.

[2] Jonathon S. Hare and Paul H. Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In Wee Kheng Leow, Michael S. Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn, and Erwin M. Bakker, editors, *CIVR*, volume 3568 of *LNCS*, pages 540–549. Springer, 2005.

[3] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[4] David Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, January 2004.

[5] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and A. David Marshall, editors, *BMVC*. British Machine Vision Association, 2002.

[6] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *In CVPR*, pages 2161–2168, 2006.

[7] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[8] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

[9] Monica Paramita, Mark Sanderson, and Paul Clough. Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[10] M F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[11] J Sivic and A Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, October 2003.