Diversity in Image Retrieval: DCU at ImageCLEFPhoto 2008

Neil O'Hare[1], Peter Wilkins[1], Cathal Gurrin[1,2], Eamonn Newman[1], Gareth J.F. Jones[1], and Alan F. Smeaton[1,2]

[1] Centre For Digital Video Processing, Dublin City University, Ireland.
[2] CLARITY: Centre for Sensor Web Technologies.
nohare@computing.dcu.ie
http://www.cdvp.dcu.ie

Abstract. DCU participated in the ImageCLEF 2008 photo retrieval task, which aimed to evaluate diversity in Image Retrieval, submitting runs for both the English and Random language annotation conditions. Our approaches used text-based and image-based retrieval to give baseline runs, with the the highest-ranked images from these baseline runs clustered using K-Means clustering of the text annotations, with representative images from each cluster ranked for the final submission. For random language annotations, we compared results from translated runs with untranslated runs. Our results show that combining image and text outperforms text alone and image alone, both for general retrieval performance and for diversity. Our baseline image and text runs give our best overall balance between retrieval and diversity; indeed, our baseline text and image run was the 2nd best automatic run for ImageCLEF 2008 Photographic Retrieval task. We found that clustering consistently gives a large improvement in diversity performance over the baseline, unclustered results, while degrading retrieval performance. Pseudo relevance feedback consistently improved retrieval, but always at the cost of diversity. We also found that the diversity of untranslated random runs was quite close to that of translated random runs, indicating that for this dataset at least, if diversity is our main concern it may not be necessary to translate the image annotations.

Key words: Content-Based Image Retrieval, Data Fusion, Clustering

1 Introduction

The CLEF 2008 ImageCLEF photo retrieval task was concerned with evaluating diversity in image retrieval, as described by Arni et al [1]. For our participation in this task DCU used standard text retrieval, with and without pseudo relevance feedback, and content-based image retrieval (CBIR) approaches based on MPEG-7 low level visual features, and a combination of text retrieval and CBIR. K-Means clustering was run on the outputs from these retrieval approaches to create a more diverse set of images at the top of the result list. For cross-language information retrieval (i.e. random language runs), we classified documents as

English or German, and then translated German documents to English using machine translation. We also submitted runs that did not translate the random language documents, to explore whether it is necessary to translate non-English annotations in order to achieve diversity.

The remainder of this paper is organised as follows: Section 2 outlines the approaches that we used for both retrieval and clustering and details our submitted runs; Section 3 gives our results, along with some preliminary analysis of them, and finally Section 4 concludes the paper.

2 System Description

Our approach for the ImageCLEF photo retrieval task can be broken down into 3 main phases, which are described in more detail below.

- **Retrieval**. We first use text-based and image-based retrieval algorithms to create a traditional ranked list of images ordered by relevance to the query.
- **Clustering**. To improve the diversity of the results, the images towards the top of the result list are clustered, which will give us groups of similar images.
- **Cluster Representative selection and Final Ranking.** The clusters are then ranked in order of relevance to the query, and one representative image from each cluster is output to the final result list.

2.1 Retrieval

Since the topic set for CLEF 2008 consists of a subset of 39 of the 60 topics used in ImageCLEFPhoto 2006 and 2007, we used the remaining 21 topics as a training set for system development. We used the retrieval ground truth for these topics to guide development of our baseline retrieval systems. In the following subsections we outline our approaches used for text retrieval, image retrieval and combined text and image retrieval.

Text Retrieval We index the Title, Description, Notes and Location fields from the annotation of each photo, and use these for text-based retrieval. The location field is matched to a world gazetteer based on freely available resources[3], expanding the Town and Country location information to Town, State/County, Country and Continent. We construct text queries using the Title and Narr fields from the topics; since the Narr field often includes information about non-relevant documents, we remove any sentences containing the phrase 'not relevant' from this field. We use the BM25 ranking algorithm [9], as implemented in the Terrier search engine [8], for text retrieval. For pseudo relevance feedback (PRF) we use the diversion from randomness approach [8], using the top 10 terms from the top 3 documents for query expansion.

[3] http://nhd.usgs.gov/gnis.html, http://earth-info.nga.mil/gns/html/index.html

For random annotation language runs the annotation documents were processed using TextCat[4], an implementation of the text categorization algorithm proposed by Cavnar & Trenkle [3]. This uses an n-gram language model approach to language identification. After identifying the German documents, we translated them from German to English using Systran Version:3.0 Machine Translator[5]. The translated documents were then indexed by the search engine identically to the English documents.

We used 3 language conditions (English, translated random and untranslated random), with and without PRF, giving 6 distinct baseline text retrieval runs.

Image Retrieval For content-based image retrieval we make use of the following six global visual features defined in the MPEG-7 specification [7]:

- **Scalable Colour (SC)** is a Haar transform encoded colour histogram defined in the HSV colour space.
- **Colour Structure (CS)** represents an image by both the colour distribution (similar to a colour histogram) and the local spatial structure of the colour.
- **Colour Layout (CL)** is a compact descriptor that captures the spatial layout of the representative colours on a grid superimposed on an image.
- **Colour Moments (CM)** is similar to Colour Layout, it divides an image into 4x4 subimages and for each subimage the mean and the variance on each LUV colour space component is computed.
- **Edge Histogram (EH)** represents the spatial distribution of edges in an image, with edges categorized as vertical, horizontal, 45 degrees diagonal, 135 degrees diagonal or non-directional.
- **Homogeneous Texture (HT)** is a quantitative representation consisting of the mean energy and the energy deviation from a set of frequency channels.

To create a visual query we take the topic images and extract the six Query-Terms from each (i.e. a representation of the image by each of the six features above). For each Query-Term we query its associated retrieval expert (i.e. visual index and ranking function) to produce a ranked list. The ranking metric for each feature is as specified by MPEG-7 and is typically a variation on Euclidian distance. For our experiments we kept the top 1000 results per Query-Term. Each ranked list was then weighted and the results from all ranked lists are normalized using MinMax [5], then linearly combined using CombSUM [5].

We used a query-dependent weighting scheme for combining visual experts using an approach that requires no training data. This approach is based on the observation that if one was to plot the normalized scores of an expert against that of scores of other experts used for a particular query, then the expert whose scores showed the greatest initial change tends to be the best performer for that query. While we acknowledge this observation is not universal, it has been shown

[4] http://odur.let.rug.nl/vannoord/textcat
[5] http://www.systran.co.uk

empirically to improve retrieval performance [10]; we also used this technique for our participation in ImageCLEFPhoto 2007 [6].

So, if a topic has three query images for example, we will extract six features per image, resulting in the generation of 18 Query-Terms. Each of these is then queried against its respective retrieval expert to produce 18 ranked lists, each ranked list is then individually weighted and the lists linearly combined.

Combination of Image and Text Retrieval As with the combination of visual features, image and text results are combined using weighted CombSUM and MinMax normalisation [5]. Based on experiments on the set of 21 training topics we used global weights of 0.7 for text and 0.3 for image, as this outperformed the query-dependant weighting approach described in Section 2.1.

2.2 Clustering

The baseline retrieval results, whether text-based, image-based, or a combination of the two, are clustered into groups in an attempt to increase the diversity of the results. All of our clustering approaches use text information exclusively; we do not perform clustering on visual features. The topic description for the ImageCLEF Photo task in 2008 includes a 'cluster' tag, which defined what criteria would be used to create the ground truth for diversity evaluation [1]. To avoid confusion with the K-Means clustering algorithm that we use in this work, we will refer to this cluster tag as 'diversity criteria'. Since it was permitted to manually inspect this diversity criteria from the topic, we classified the diversity criteria into 3 categories: 'location', 'non-location' or 'general'. The 39 topics include 17 unique entries for this diversity criteria tag. After classifying them into the 3 categories, we use a different subset of the fields from the structured annotation as input into our text clustering algorithm, as follows:

- **Location:** Topics for which only the location field is used as input to the clustering algorithm, corresponding to the diversity criteria 'city', 'state', 'location', 'country', 'city national park' and 'venue'.
- **Non-location:** Topics for which the location field is ignored for clustering, corresponding to the diversity criteria 'animal', 'sport', 'bird', 'weather condition', 'vehicle type', 'composition' and 'group composition'.
- **General:** Topics for which all fields used for retrieval are also used for clustering: 'statue', 'venue', 'landmark', 'volcano' and 'tourist attraction'.

Apart from using a different subset of the annotation fields, each of these types is treated identically in our subsequent clustering. We also submitted runs that did not classify the diversity criteria and treated all topics the same. We use the K-Means clustering algorithm, as implemented in the Text Clustering Toolkit[6]. Using annotation fields from one of the 3 classes defined above, we take the top X documents from our baseline retrieval algorithms and cluster

[6] http://mlg.ucd.ie/content/view/20/

them using K-Means; we varied the parameter X in a number of runs, using values of 50, 100 and 150. We also varied k, the number of clusters, using 20, 30 and 40 clusters. An additional variant used the the Calinski-Harabasz index to automatically estimate the optimum number of clusters [2]. Since we are clustering a small number of documents (150 or less), the tf-idf weighting scheme may not have enough documents to calculate reliable inverse document frequency scores, so we use two separate approaches to term normalisation for clustering: term frequency (tf) and term frequency / inverse document frequency (tf-idf).

2.3 Cluster Representative Selection and Final Ranking

Finally, we rank all clusters in order of relevance and select a representative image for each cluster for the final ranked list. We use the maximum individual image ranking score within the cluster as the overall cluster score, using the same maximum image as the cluster representative, and our final output is k images, corresponding to the most relevant image from each cluster.

2.4 Submitted Runs

We created 13 baseline retrieval runs as follows: 3 language conditions (English, translated and untranslated random) with and without pseudo relevance feedback; each of these 6 text-only baselines was combined with image retrieval to give 6 text-image baselines; additionally, we had 1 image-only baseline. These 13 baseline runs were used as input into clustering using a number of parameter variations, creating a number of different runs. The parameters were: X, the number of documents to cluster; k, the number of clusters; term normalisation method; diversity criteria classification. This gives a total of 48 variations of clustering for each baseline submission. We cluster the image-only baseline using each of the 3 language conditions, meaning we cluster 13 baselines plus two additional language variants for the image baseline, we have $15x48 = 720$ clustered runs and 13 baseline runs, giving a total of 733 runs submitted.

3 Results

Our results are summarised in Table 1. This shows our baseline unclustered text and text-image results along with the best clustered variation for each baseline. As one would expect, runs that combine text and image retrieval always give the best performance. It is noteworthy, however, that there is no tradeoff involved: general retrieval (measured by MAP or P@20) and diversity (CR@20) are both improved simultaneously by combining text and image. For English language retrieval with PRF but without clustering, for example, P@20 is improved from 0.405 to 0.476, and CR@20 is improved from 0.348 to 0.454, by combining text retrieval with image retrieval. Similar improvements can be observed for all comparable configurations when text retrieval and image retrieval are combined. This makes intuitive sense because these different modalities will retrieve

Language	Translated	Modality	Clustered	PRF	MAP	P@20	CR@20
English	-	Txt	No	No	0.312	0.376	0.407
English	-	Txt	No	Yes	**0.351**	**0.405**	0.348
English	-	Txt	Yes	No	0.070	0.232	**0.514**
English	-	Txt	Yes	Yes	0.092	0.294	0.50
English	-	TxtImg	No	No	0.352	0.463	0.455
English	-	TxtImg	No	Yes	**0.354**	**0.476**	0.454
English	-	TxtImg	Yes	No	0.095	0.265	**0.552**
English	-	TxtImg	Yes	Yes	0.097	0.262	0.525
Random	Yes	Txt	No	No	0.258	0.339	0.406
Random	Yes	Txt	No	Yes	**0.279**	**0.345**	0.353
Random	Yes	Txt	Yes	No	0.081	0.246	**0.472**
Random	Yes	Txt	Yes	Yes	0.073	0.231	0.464
Random	No	Txt	No	No	0.169	0.283	0.404
Random	No	Txt	No	Yes	**0.173**	**0.289**	0.381
Random	No	Txt	Yes	No	0.053	0.214	**0.488**
Random	No	Txt	Yes	Yes	0.059	0.209	0.473
Random	Yes	TxtImg	No	No	0.309	0.440	0.467
Random	Yes	TxtImg	No	Yes	**0.309**	**0.442**	0.453
Random	Yes	TxtImg	Yes	No	0.1063	0.332	**0.536**
Random	Yes	TxtImg	Yes	Yes	0.101	0.283	0.513
Random	No	TxtImg	No	No	**0.225**	**0.381**	0.455
Random	No	TxtImg	No	Yes	0.222	0.372	0.400
Random	No	TxtImg	Yes	No	0.081	0.264	**0.518**
Random	No	TxtImg	Yes	Yes	0.077	0.247	0.491

Table 1. DCU Results for ImageCLEFPhoto 2008.

different relevant documents, and so combining them will improve both retrieval and diversity. Since there is no tradeoff here, it suggests that one very effective way to improve diversity is to use evidence from independent modalities, and we expect that we could further improve our results by using automatically extracted visual concepts such as those extracted as part of the ImageCLEF Visual Concept Detection Task [4].

Using K-Means Clustering gives a large improvement in diversity, but this comes at the cost of degraded retrieval performance. Our best CR@20 score of 0.552 on English language text and image with clustering and without PRF, for example, gives a 21% improvement over the unclustered equivalent, but P@20 for this run falls by 43%% to 0.265; a similar tradeoff can be seen with all comparable clustered and unclustered runs.

While PRF leads to consistently better retrieval performance in terms of MAP and P@20, it also consistently harms diversity. Runs without feedback consistently perform better for CR@20, and this pattern can be observed both in clustered and unclustered runs. This result is not particularly surprising as PRF uses that top retrieved images to expand the query, meaning that the results will be dominated by images similar to these.

Comparing Random language runs with English language runs, the best text and image Random runs in terms of diversity perform quite close to the English runs, achieving a CR@20 score of 0.536, only a 3% decrease from best English score at 0.552, although this run is 7% worse than English for P@20. Comparing English text-only runs with Random text-only runs, the best Random runs for diversity are 5% worse for CR@20 and 15% worse for CR@20. The fact that this difference is much smaller for text and image retrieval shows that image retrieval can is particularly helpful for cross-lingual retrieval, where it can help to close the gap between mono-lingual retrieval and cross-lingual retrieval. Comparing the best overall (ie. F1-measure, P@20 and CR@20) translated Random runs, again the text and image unclustered run, at 0.4531 performs within 3% of the best English run for this measure (0.4647).

Our untranslated runs show that by essentially discarding 50% of the documents in the collection (although, for the text and image runs, some of these 'discarded' documents may be recovered if their image score is high enough), we can still maintain a similar level of diversity with a best CR@20 score of 0.518 for the clustered run, only 3% below the score achieved if we translate the annotation documents. The untranslated runs perform much more poorly for P@20 and MAP, but for scenarios where we consider diversity to be our main concern this result suggests that it is not necessary to translate non-English documents, particularly when we are combining text retrieval with image retrieval. It is unclear whether this is an effect of this particular test collection or if this conclusion would be valid in other scenarios.

Comparing our results with those of other participants [1], DCU had the 2nd best automatic run for the English language condition, with an F1-measure (P@20 and CR@20) score of 0.4647. This run is only 0.0003 behind the best automatic run, submitted by Xerox Research Centre Europe, a difference small enough to suggest there is no clear difference between our best run and the best overall run for the task in 2008. In fact, our system performs better in terms of diversity than the Xerox system (0.4542 CR@20 compared with 0.4262) and worse in terms of retrieval (0.5115 P@20 compared with 0.4756), so we would argue that our system would be preferable if the focus is on diversity. Due to the small number of submissions from other groups for the Random language condition, it is not possible to fruitfully compare our approaches with other participants [1].

4 Conclusions

Our paritipation in ImageCLEF Photo 2008 has allowed us to draw a number of conclusions about diversity in image retrieval. PRF improves performance for standard retrieval measures, but this comes at the cost of less diversity. Clustering the results of the baseline retrieval algorithms gives a large improvement in diversity, while harming retrieval performance. Combining image with text retrieval gives a large improvement in diversity and retrieval over text alone, and in our experiments this was the most effective way of improving diversity. For

cross-lingual information retrieval we have shown that it is possible to maintain diversity in our results without translating the German annotations into English; our cross-lingual runs have also shown that using image retrieval in combination with text retrieval narrows the gap in performance between cross-lingual and mono-lingual information retrieval.

Acknowledgements

This work is supported by Science Foundation Ireland under grant number 07/CE/I1147.

References

1. T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In C. Peters, D. Giampiccol, . Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. J. F. Jones, and M. Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).
2. T. Calinski and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistica*, 3:1–27, 1974.
3. W. B. Cavnar and J. M. Trenkle. N-Gram-Based Text Categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV, 11-13 April 1994. UNLV Publications/Reprographic.
4. T. Deselaers and A. Hanbury. The Visual Concept Detection Task in ImageCLEF 2008. In C. Peters, D. Giampiccol, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. J. F. Jones, and M. Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008 (printed in 2009).
5. E. A. Fox and J. A. Shaw. Combination of Multiple Searches. In *Proceedings of the Third Text REtreival Conference (TREC-1994)*, pages 243–252, Gaithersburg, MD, 1994.
6. A. Jarvelin, P. Wilkins, T. Adamek, E. Airio, G. J. F. Jones, A. F. Smeaton, and E. Sormunen. DCU and UTA at ImageCLEFPhoto 2007. In *ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop*, Budapest, Hungary, 2007.
7. B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley, 2002.
8. I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. *Novatica/UPGRADE Special Issue on Web Information Access*, 2007.
9. S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, MD, 1995.
10. P. Wilkins, P. Ferguson, and A. F. Smeaton. Using Score Distributions for Query-time Fusion in Multimedia Retrieval. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Santa Barbara, CA, 2006.