DCU at ImageCLEFPhoto 2008

Neil O'Hare, Peter Wilkins, Cathal Gurrin, Eamonn Newman, Gareth J.F. Jones and Alan F. Smeaton.
Centre for Digital Video Processing, Dublin City University.
nohare@computing.dcu.ie

Abstract

DCU participated in the ImageCLEF 2008 photo retrieval task, submitting runs for both the English and Random language annotation conditions. Our approaches used text-based and image-based retrieval approaches to give baseline retrieval runs, with the highest-ranked images from these baseline runs clustered using K-Means clustering of the text annotations. Finally, each cluster was represented by its most relevant image and these images were ranked for the final submission. For random annotation language runs, we used TextCat[1] to identify German annotation documents, which were then translated into English using Systran Version:3.0 Machine Translator. We also compared results from these translated runs with untranslated runs. Our results showed that, as expected, runs that combine image and text outperform text alone and image alone. Our baseline image+text runs (i.e. without clustering) give our best MAP score, and these runs also outperformed the mean and median ImageCLEFPhoto submissions for CR@20. Clustering approaches consistently gave a large improvement in CR@20 over the baseline, unclustered results. Pseudo relevance feedback consistently improved MAP while also consistently decreasing CR@20. We also found that the performance of untranslated random runs was quite close to that of translated random runs for CR@20, indicating that we could achieve similar diversity in our results without translating the documents.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation

Keywords

Content-Based Image Retrieval, Data Fusion, Clustering

1 Introduction

For the CLEF2008 ImageCLEF photo retrieval task, our baseline retrieval system DCU used standard text retrieval, both 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 then run on the outputs from these retrieval approaches in order to promote a more diverse set of images towards the top of the result list in

[1] http://odur.let.rug.nl/ vannoord/TextCat/

keeping with this years version of the task, which is to promote diversity in image retrieval. For cross-language retrieval (i.e. random language runs) we used TextCat to classify documents as English or German, and then translated German documents to English using Systran. We also submitted runs based on indexing the random language documents without translating them, 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 that 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. Section 5 concludes the paper.

2 System Description

Our approach for the ImageCLEF photo retrieval task this year can be broken down into 3 main phases, as follows:

- **Retrieval**. We first run a 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 output 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.

Each of these is described in more detail below.

2.1 Retrieval

Since the topic set for 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 of topics for system development. Although we did not have ground truth for these topics for diversity, we did have ground truth for retrieval, so we could use 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.

2.1.1 Text Retrieval

For text retrieval we indexed the following field from the structured annotation of each photo: Title, Description, Notes and Location. The location field was matched to a world gazetteer built using data from the Geographic Names Information System [3] and the GEOnet Names Server [1]. This allows us to automatically expand the location information to Town, State/County, Country, Continent, instead of just the Town and Country provided. To formulating our queries, we made use of the title and narr fields from the topics. Since the narr field often includes information about non-relevant documents, we parsed the narr field to remove any sentences containing the phrase 'not relevant'. We perform text retrieval using the BM25 algorithm [12], as implemented in the Terrier search engine platform [11], using the following parameters: $k1 = 1.2, k3 = 8$ and $b = 0.75$. For runs that use pseudo relevance feedback, we used the diversion from randomness approach [11], using the top 10 terms from the 3 documents for query expansion.

For random annotation language runs, the annotation documents were processed using TextCat [4], which is an implementation of the text categorization algorithm proposed by Cavnar & Trenkle [7] [1]. TextCat uses a n-gram language model approach to language identification, i.e. a language is recognised through the identification of distinct n-grams which occur frequently in the language but seldom or not at all in other languages. After TextCat identified all the German sentences in the set of random language annotation, this content was translated from German to English using

Systran Version:3.0 Machine Translator [5]. The set of translations were then merged with the other items in the dataset (ie the English sentences) and passed to the next stage of our system.

For text retrieval we used 3 language conditions (english, translated random and untranslated random), each with and without PRF, giving 6 distinct baseline text retrieval runs.

2.1.2 Image Retrieval

For our visual retrieval in ImageCLFEPhoto we make use of six global visual features which are defined in the MPEG-7 specification [10]. The following low-level features were used:

- **Scalable Colour (SC):** derived from a colour histogram defined in the HSV colour space. It uses a Haar transform coefficient encoding, allowing scalable representation.
- **Colour Structure (CS):** based on colour histograms, represents an image by both the color distribution (similar to a color histogram) and the local spatial structure of the colour.
- **Colour Layout (CL):** compact descriptor which captures the spatial layout of the representative colours on a grid superimposed on an image.
- **Colour Moments (CM):** similar to Colour Layout, this descriptor divides an image into 4x4 subimages and for each subimage the mean and the variance on each LUV color space component is computed.
- **Edge Histogram (EH):** represents the spatial distribution of edges in an image, edges are categorized into five types: vertical, horizontal, 45 degrees diagonal, 135 degrees diagonal and non directional.
- **Homogeneous Texture (HT):** provides a quantitative representation using 62 numbers, consisting of the mean energy and the energy deviation from a set of frequency channels.

To compute an answer for a visual query, we take the topic images and extract from each their six Query-Terms (i.e. a representation of the image by each of the six features previously detailed). 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 produced 1000 results per Query-Term. Each ranked list is then weighted and the results from all ranked lists are normalized using MinMax [8], then linearly combined using CombSUM [8].

The weighting scheme we used for combination of visual experts is a query-dependant weighting scheme for expert combination which requires no training data [13]. This approach is based on the observation that, if we were 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 exhibited the greatest initial change correlated with that expert being the best performer for that query. While we acknowledge this observation is not universal, it has been shown emperically to improve retrieval performance. This technique was also employed in DCU's participation in ImageCLEFPhoto 2007 [9].

For example, if our topic set has three query images, 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, then each ranked list is then individually weighted, using the aforementioned technique and linearly combined through data fusion.

2.1.3 Combination of Image and Text Retrieval

As with the combination of visual features, image and text results are combined by weighting each ranked list, normalizing the results using MinMax [8] and then linearly combined using CombSUM [8]. Our results on the set of 21 training topics showed that, for the text and image combination, global weights of 0.7 for text and 0.3 for image outperformed the query-dependant weighting approach described above for MAP, so we used global weights for combining text results with image results.

2.2 Clustering

The results of our baseline retrieval, whether text-based, image-based or a combination of the two, are then clustered to increase the diversity of the results. All of our clustering approaches use text information exclusively; we do not perform clustering on visual features. Since it was permitted in this task to inspect the cluster tag from the topic and create higher level cluster types, we classified the cluster tags into 3 categories: 'location', 'non-location' or 'general'. The 39 topics include 17 unique entries for the cluster tag. After classifying them into 3 categories we use a different subset of the fields from the structured annotation for clustering, as follows:

- **Location:** Topics for which only the location tag is used for clustering, corresponding to the cluster tags 'city', 'state', 'location', 'country', 'city national park' and 'venue'.
- **Non-location:** Topics for which the location tags is ignored for clustering, corresponding to the cluster tags 'animal', 'sport', 'bird', 'weather condition', 'vehicle type', 'composition' and 'group composition'.
- **General:** Topics for which all tags 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 cluster type is treated identically in our subsequent clustering. We also submitted runs that did not classify the cluster tag, and treated all topics the same.

For clustering, we employ K-Means clustering using the Text Clustering Toolkit, a toolkit for clustering text documents using a number of standard algorithms [2]. Using annotation fields from one of the 3 classes defined above, we take the top X documents from our baseline retrieval algorithms and cluster them using K-Means. Rather than choose one single value for X, we varied this parameter 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 [6].

Since we are clustering a small number of documents (ie. 150 or less), the tf-idf weighting scheme may not have enough documents to calculate reliable inverse document frequency scores. For this reason, we have used two separate approaches to term normalisation for cluster analysis: term frequency (tf) and term frequency / inverse document frequency (tf-idf).

2.3 Cluster Ranking and Cluster Representative Selection

The final step is to rank all clusters in order of relevance to the query, and then select a representative image for each cluster to output to the final ranked list. To rank clusters, we take the simple approach of using the maximum individual image score within the cluster as the overall cluster score. We also use the same maximum image as the cluster representative, and our final output is k images (i.e. the number of clusters), corresponding to the most relevant image from each cluster.

2.4 Description of Submitted Runs

For our submission to ImageCLEFPhoto 2008 we created 13 baseline retrieval runs as follows: 3 language conditions (english, translated and untranslated random) with and without pseudo relevance feedback for text only baselines; each of these 6 were 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 clusterin using a number of parameter variations, creating a number of different runs. The parameters were: X, the number of documents to cluster (50, 100 or 150); k, the number of clusters (20, 30, 40 or automatic using the Calinski-Harabasz index); term normalisation method (tf or tf-idf); cluster classification (classification used or classification not used). This gives a total of 48 variations of clustering for each baseline submission. Since we cluster the image-only baseline using each of the 3 language conditions, meaning we cluster 13 baselines

plus two addition 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, which shows our baseline unclustered results and the best clustered variation for each baseline. As one would expect our best results are given by combining text retrieval with image retrieval, with the best MAP of 0.354 and the best P20 of 0.476 given by the English text and image run using pseudo relevance feedback. The best result for Cluster Recall at 20 (CR@20) is given by the English text and image with clustering, with a score of 0.552. The clustered runs perform poorly for MAP, although this is not a surprise since we only rank the top k for these runs, giving a truncated result list which would be expected to perform poorly on MAP, which is dependent on the entire ranked list.

We can see that, while pseudo relevance feedback leads to consistently better retrieval performance in terms of MAP and P@20, it also decreases diversity. Runs without feedback consistently perform better for CR@20, and this pattern can be observed both in clustered and unclustered runs. Combining image and text retrieval also gives a large improvement in diversity: for English language multimodal clustered runs, for example, the best performance for CR@20 is improved from 0.514 to 0.552, an improvement of 7% (for unclustered English runs, the improvement is 12%). Since image retrieval and text retrieval naturally retrieve different relevant documents, it is not a surprise that combining them gives a large improvement in CR@20. Combining image and text retrieval also gives a large improvement in P@20, although it only gives a modest improvement in MAP. The unclustered English text and image runs also show that it is possible to achieve good CR@20 score without using clustering: the CR@20 score of 0.455 for this run is 29% above the mean (0.455 compared with 0.353), without using any clustering. This unclustered run give our most consistent performance across all evaluation measures, performing above the mean for MAP, P@20 and CR@20: in fact, all our unclustered English and unclustered translated Random runs achieve this, and our untranslated random runs beat the mean when combined with image retrieval.

Comparing random language runs with English runs, the best random runs perform quite close to the English runs in terms of diversity, achieving a CR@20 score of 0.536, only 3% below the best English score. Our untranslated runs also show that, by effectively 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 score of 0.518 for the clustered run, only 3% below the score achieved if we translate the annotation documents. It is an open question where this is an effect of this particular test collection or whether in real world scenarios there would be a higher correlation between clusters and document languages.

As expected, the runs that use image retrieval as the baseline retrieval perform quite poorly, although the CR@20 scores that they achieve are not very far below the median, and at the moment it is not possible to compare them to image-only runs from other groups.

4 Conclusions

In this paper we presented the approaches we used for the ImageCLEF 2008 Ad-hoc photographic retrieval task, along with a preliminary analysis of our results. This preliminary analysis showed that pseudo relevance feedback, while improving performance for standard retrieval measures, harms performances when it comes to diversity. Combining image with text retrieval gives a large improvement in diversity even when the improvement in overall retrieval (as measured by MAP) is modest. Clustering the results of the baseline retrieval algorithms gives a large improvement in diversity, while unsurprisingly harming overall retrieval performance. For the random language condition (where half of the documents are in English and half in German) we have shown that it

Language	Translated	Modality (Retrieval)	Clustered	PRF	MAP	P@20	CR@20
		Mean			*0.219*	*0.320*	*0.353*
		Median			*0.219*	*0.320*	*0.352*
		Min			*0.033*	*0.123*	*0.178*
		Max			*0.429*	*0.696*	*0.680*
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
-	-	Img	No	No	0.107	0.240	0.320
English	-	ImgTxt(Img)	Yes	No	0.052	0.171	0.326
Random	Yes	ImgTxt(Img)	Yes	No	0.047	0.162	0.327
Random	No	ImgTxt(Img)	Yes	No	0.071	0.212	0.331

Table 1: DCU Results for ImageCLEFPhoto 2008, compared to the Mean, Median, Min and Max for all submissions. For each clustered run, the best submission for that modality and language pairing is shown. Scores in bold represent the best result for that language, translated, modality triple for a given evaluation measure. For the modality field, a separate value in brackets indicates the modality for the baseline retrieval: for example, ImgTxt(Img) means the the baseline retrieval value image-based, but since text was used for clustering the overall modality was ImgTxt.

is possible to maintain diversity in our results without translating the German annotations into English.

This paper represents a preliminary analysis of our results, and we plan to analyse our results in more detail. In particular, we plan to look in detail at how the clustering parameters that we varied in our submissions affect results. Also, we have not yet analysed the results from the runs that do not classify the clustering approach used, instead clustering for all topics using the entire annotation instead of just, say, the location tag. We plan to compare these unclassified results to the results presented here, to examine the level of diversity performance that we can achieve if we have no information about the clustering criteria. We would also like to conduct some topic by topic, particularly in order to distinguish those topics which rely on location-based clustering from those that do not.

References

[1] http://earth-info.nga.mil/gns/html/index.html.

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

[3] http://nhd.usgs.gov/gnis.html.

[4] http://odur.let.rug.nl/~vannoord/textcat/.

[5] http://www.systran.co.uk/.

[6] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistica*, 3:1–27, 1974.

[7] 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.

[8] Edward A. Fox and Joseph A. Shaw. Combination of Multiple Searches. In *Proceedings of the Third Text REtreival Conference (TREC-1994)*, pages 243–252, Gaithersburg, MD, 1994.

[9] Anni Jarvelin, Peter Wilkins, Tomasz Adamek, Eija Airio, Gareth Jones, Alan F. Smeaton, and Eero Sormunen. Dcu and uta at imageclefphoto 2007. In *ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop*, Budapest, Hungary, 2007.

[10] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley, 2002.

[11] Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. *Novatica/UPGRADE Special Issue on Web Information Access*, 2007.

[12] Stephen E. Robertson, Steve 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.

[13] Peter Wilkins, Paul Ferguson, and Alan F. Smeaton. Using Score Distributions for Querytime Fusion in Multimedia Retrieval. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Santa Barbara, CA, 2006.