MMIS at ImageCLEF 2008: Experiments combining different evidence sources

Simon Overell[1], Ainhoa Llorente[2,3], Haiming Liu[2], Rui Hu[2], Adam Rae[2], Jianhan Zhu[2], Dawei Song[2] and Stefan Rüger[2,1]

[1]Multimedia & Information Systems
Department of Computing, Imperial College London, SW7 2AZ, UK

[2]Knowledge Media Institute
The Open University, Milton Keynes, MK7 6AA, UK

[3]INFOTECH Unit
ROBOTIKER-TECNALIA, Parque Tecnológico, Edificio 202
E-48170 Zamudio, Bizkaia, Spain

seo01@doc.ic.ac.uk, jianhanzhu@gmail.com and
{a.llorente, h.liu, r.hu, a.rae, d.song, s.rueger}@open.ac.uk

Abstract

This paper presents the work of the MMIS group at ImageCLEF 2008. The results for three tasks are presented: Visual Concept Detection Task (VCDT), ImageCLEFphoto and ImageCLEFwiki. We combine image annotations, CBIR, textual relevance and a geographic filter using our generic data fusion method. We also compare methods for BRF and clustering.

Our top performing method in the VCDT enhances supervised learning by modifying probabilities based on a matrix that shows how terms appear together. Although it occurred in the top quartile of submitted runs, the enhancement did not provide a statistically significant improvement.

In the ImageCLEFphoto task we demonstrate that evidence from image retrieval can provide a contribution to retrieval; however we are yet to find a way of combining text and image evidence in a way to provide an improvement over the baseline. Due to the relative performances of difference evidences in ImageCLEFwiki and our failure to improve over a baseline we conclude that text is the dominant feature in this collection.

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; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Content Based Image Retrieval, Geographic Retrieval, Data Fusion

1 Introduction

In this paper we describe the experiments of the MMIS group at ImageCLEF'08. We participated in three tasks: Visual Concept Detection Task (VCDT), ImageCLEFphoto and ImageCLEFwiki.

All experiments were performed in a single framework of independently testable and tuneable modules. The framework is described in detail in Section 2. The experiments conducted and individual runs submitted for each task are described in Sections 3, 4 and 5. Finally we present our conclusions in Section 6.

2 System

Our system framework is shown in Figure 1. The text elements of the corpus are indexed as a bag-of-words, analysed geographically and stored in a geographic index. Texture and colour features are extracted from images to form feature indexes, these features are further analysed to automatically annotate the images. This allows us to compare query images to our index using both semantic annotations and low-level features.

Blind Relevance Feedback (BRF) is employed across media types similarly to [9]. In training experiments we found the text results to have the highest precision of all individual media types. Because of this we use the top text results as feedback for the Image Retrieval engine to provide an additional Image BRF rank.

Our intermediate format is the standard TREC format. This allows us to evaluate and tune each module independently. The results of all the independent modules are combined in the data fusion module. The data fusion module combines both the *ranks* provided by the image and text query engine, and the *filters* provided by the geographic and annotation query engines. The difference between a rank and a filter is all elements in a filter are considered of equal relevance.

In the ImageCLEFphoto task we are evaluated on the novelty of our top results and provided with a subject which this novelty will be judged with respect to. We have clustered our top results using the geographic index, image annotations and text index.

Further details on the individual modules and tuning are described in the following sections.

2.1 Image Feature Extractor

Content-Based Image Retrieval (CBIR) provides a way to browse or search images from large image collections based on visual similarity. CBIR is normally performed by computing the dissimilarity between the object images and query images based on their multidimensional representations in content feature spaces, for example, colour, texture and structure. In this section, we are going to introduce the key issues of the image search applied to our three tasks.

2.1.1 Feature Extraction.

ImageCLEFphoto Task. Colour feature is the most commonly used visual feature e.g. HSV, RGB [14]. In [8] Colour feature HSV outperforms the texture feature Gabor and structure feature Konvolution on CBIR . Thus we use HSV in our CBIR system. HSV is a cylindrical colour space. Its representation appears to be more intuitive to humans than the hardware representation of RGB. The hue coordinate H is angular and represents the colour, the saturation S represents the pureness of the colour and is the radial distance, finally the brightness V is the vertical distance. We extract the HSV feature globally for every image in the query and test sets.

ImageCLEFwiki Task. We used the Gabor feature for the ImageCLEFwiki task. Gabor is a texture feature generated using Gabor wavelets [7]. Here we decompose each image into two scales and four directions.

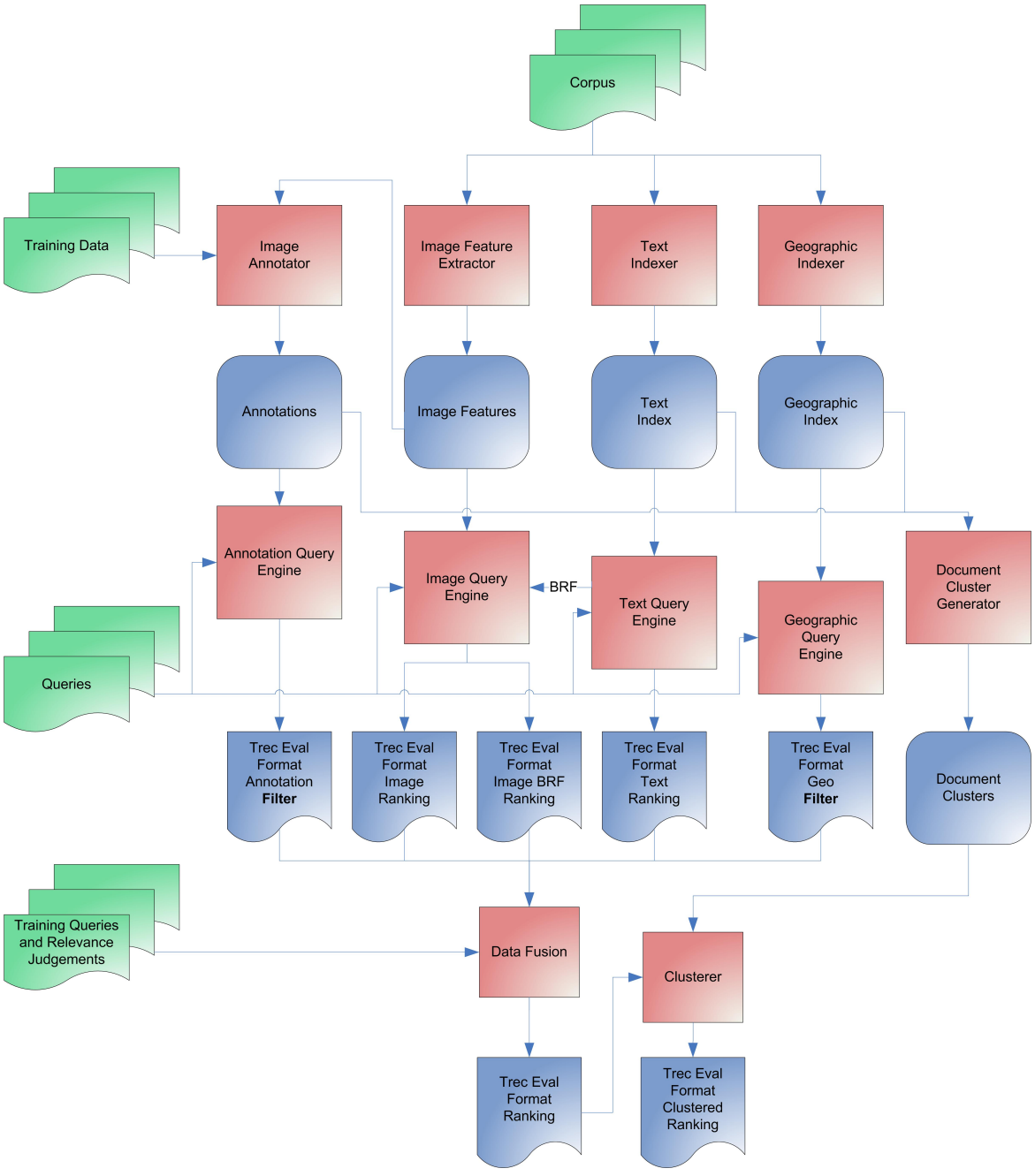

Figure 1: Our Application Framework

VCDT. The features used in the VCDT experiments are a combination of colour feature feature, CIELAB, and texture feature, Tamura.

CIE $L*a*b*$ (CIELAB) [4] is the most complete colour space specified by the International Commission on Illumination (CIE). Its three coordinates represent the lightness of the colour ($L*$), its position between red/magenta and green ($a*$) and its position between yellow and blue ($b*$).

The Tamura texture feature is computed using three main texture features called "contrast", "coarseness", and "directionality". Contrast aims to capture the dynamic range of grey levels in an image. Coarseness has a direct relationship to scale and repetition rates and it was considered by Tamura et al. [16] as the most fundamental texture feature and finally, directionality is a global property over a region.

The process for extracting each feature is as follows, each image is divided into nine equal rectangular tiles, the mean and second central moment feature per channel are calculated in each tile. The resulting feature vector is obtained after concatenating all the vectors extracted in each tile.

2.1.2 Dissimilarity Measure.

ImageCLEFphoto Task. The χ^2 statistic is a statistical measure that compares two objects in a distributed manner and basically assumes that the feature vector elements are samples. The dissimilarity measure is given by

$$d_{\chi^2}(A,B) = \sum_{i=1}^{n} \frac{(a_i - m_i)^2}{m_i}, \tag{1}$$

$$m_i = \frac{a_i + b_i}{2} \tag{2}$$

where $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_n)$ are the query vector and test object vector respectively. It measures the difference of the query vector (observed distribution) from the mean of both vectors (expected distribution)[20]. The χ^2 statistic was chosen as it was one of the consistently best performing dissimilarity measures in our former research [8].

ImageCLEFwiki Task. We use the City Block distance for the ImageCLEFwiki task, to compute the distance between a query image and each test image. The City Block distance belongs to the Minkowski family, which is given by:

$$d_p(A,B) = (\sum_{i=1}^{n} |a_i - b_i|^p)^{\frac{1}{p}}, \tag{3}$$

when the parameter p equals one.

2.1.3 Search Method.

ImageCLEFphoto Task. In the ImageCLEFphoto task, every query topic includes three independent example images a, b, c. The probability of an object image x being a relevant match for a query topic is determined by the joint probability of the relevance between the images a, b, c in the query topic and the object image x. According to probability theory, the joint result is given by

$$D(abc, x) = d(a, x) \times d(b, x) \times d(c, x), \tag{4}$$

where $D(abc, x)$ is the distance between a query topic including images a, b and c with an image x. $d(a, x), d(b, x)$ and $d(c, x)$ are distances between the three images a, b, c and the image x, respectively.

2.1.4 Blind Relevance Feedback.

We employed Blind Relevance Feedback (BRF) for CBIR in a similar fashion to [9]. Two different BRF method were applied in the ImageCLEFphoto task.

- The first BRF method takes the top seven results from the text retrieval results as new query examples, and the seven ranked result are combined by the search method introduced in Section 2.1.3.
- The second BRF method was employed in both the ImageCLEFphoto and ImageCLEFwiki tasks. The final ranked result of this BRF is the sum of the ranks of the top five examples from the text result with equal weights. The detail of this combining method are described in Section 2.6.

2.2 Image Annotator

The Image Annotator is the core part of the Visual Concept Detection Task (VCDT) which aims to create a model able to detect the presence or absence of 17 visual concepts in the images of the collection. The input is a training set of 1825 images that have already been annotated with words coming from a vocabulary of 17 visual concepts. The output is the annotations.

We use as a baseline for this module the framework developed by Yavlinsky et al. [19] who used global features together with a non-parametric density estimation.

The process can be described as follows. First, images are segmented into nine equal tiles, and then, low-level features are extracted. The features used to model the visual concept densities are a combination of colour CIELAB and texture Tamura, as explained in Section 2.1.

The next step is to extract the same feature information from an unseen picture in order to compare it with all the previously created models (one for each concept). The result of this comparison yields a probability value of each concept being present in each image.

Then we modify some of these probabilities using additional knowledge from the image context in order to improve the accuracy of the final annotations.

The context of the images is computed using a co-occurrence matrix where each cell represents the number of times two visual concepts appear together annotating an image of the training set.

The underlying idea of this algorithm is to detect incoherence between words with the help of this correlation matrix. Once incoherence between words has been detected, the probability of the word associated to the lowest probability will be lowered, as well as all the words which are semantically similar.

Among the many uses of the concept "semantic similarity," we refer to the definition by Miller and Charles [10] who consider it as the degree of contextual interchangeability or the degree to which one word can be replaced by another in a certain context. Consequently, two words are similar if they refer to entities that are likely to co-occur together like "mountains" and "vegetation", "beach" and "water", "buildings" and "road", etc. As shown in Figure 2, our vocabulary of 17 visual concepts adopt a hierarchical structure. In the first level we find two general concepts like "indoor" and "outdoor" which are mutually exclusive while in lower levels of the hierarchy we find more specific concepts that are subclasses of the previous ones. Some concepts can belong to more than one class, for instance, a "person" can be part of an "indoor" or "outdoor" scene but others are mutually exclusive, a scene can not represent "day" and "night" at the same time.

Thus, by modifying the probability values of some concepts, annotations are produced by selecting the concepts with the highest probability. The output of this module, the annotations, will be the input for the Annotation Query Engine and the Document Cluster Generator.

2.2.1 Annotation Query Engine

The input for this module is textual queries which are the concepts of our vocabulary and the annotations produced by the Image Annotator module. Given a query term, we represent the top

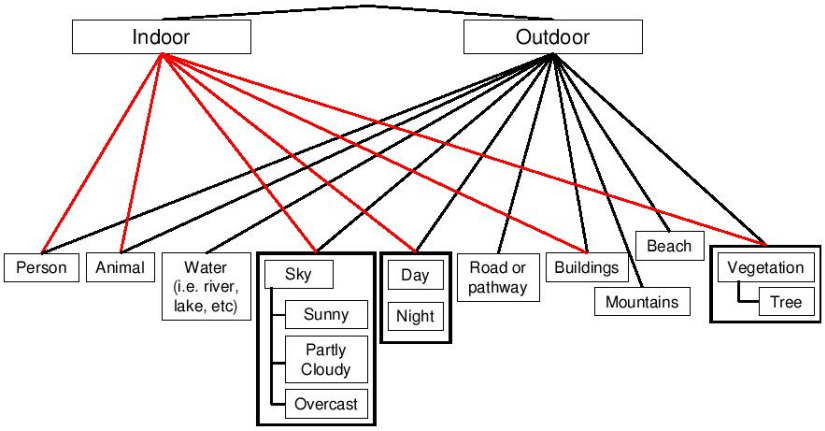

Figure 2: Hierarchy of the visual concepts

n images annotated by it following the standard TREC format. This constitutes one of the inputs for the Data Fusion module. Retrieval performance is evaluated with the mean-average precision (MAP) on the whole vocabulary of terms, which is the average precision, over all queries, at the ranks where recall changes (where relevant items occur).

2.3 Text Indexer

Our text retrieval system is based on Apache Lucene [15]. Text fields are pre-processed by a customised analyser similar to Lucene's default analyser: text is split at white space into tokens, the tokens are then converted to lower case, stop words discarded and stemmed with the "Snowball Stemmer". The processed tokens are held in Lucene's inverted index.

We use only the English meta-data and queries (monolingual retrieval). In ImageCLEFphoto both the title and location fields are searched as text. We do not use the notes field as in previous training experiments we have found this gives worse results.

2.4 Geographic Indexer

We process the text fields with Sheffield University's General Architecture for Text Engineering (GATE) [1]. The bundled Information Extraction Engine, ANNIE, performs named entity recognition, extracting named entities and tagging them as locations. Our disambiguation system matches these placenames to unique locations in the Getty Thesaurus of Geographical Names (TGN) [5].

In ImageCLEFwiki we compare two different disambiguation methods, both based on a geographic co-occurrence model mined from Wikipedia [13]. The first method (MR), builds a default gazetteer based on statistics from Wikipedia on which locations are *most referred* to by each placename. It is not context aware and disambiguates every placename with the same name to the same location. For example every reference to *Cambridge* will be matched to Cambridge, Massachusettes regardless of context.

The second method (Neighbourhoods), builds *neighbourhoods* of trigger words from Wikipedia. Depending which trigger words occur in the vicinity of an ambiguous placename dictates which location it will be disambiguated as. For example if *Oxford* occurs in the context of *Cambridge*, Cambridge will be disambiguated as Cambridge, Cambridgeshire as Oxford is a trigger word of Cambridge, Cambridgeshire.

As the ImageCLEFphoto corpus contains minimal referrences to ambiguous placenames we only use the MR method.

To query our geographic index we extract locations from the query and based on the topological data contained in the TGN return a filter of all the documents mentioning either the query locations or locations within the query locations. For example a query location of the "United States" will return all the documents mentioning the United States (or synonyms such as "USA", "the States" etc.) and all documents mentioning states, counties, cities and towns within the United States.

2.5 Document Cluster

We only employ clustering in ImageCLEFphoto. We propose a simple method of re-ordering the top of our rank based on document annotations. We consider three sources of annotations: Automated annotations assigned to images, words matched to WordNet and locations extracted from text (described in Section 2.4). WordNet is a freely available semantic lexicon of 155,287 words mapping to 117,659 semantic senses [17]. In our experiments we compare two sources of annotations: automated image annotations (Image clustering) and words matched to WordNet (WordNet clustering).

In Image clustering all the images have been annotated with concepts coming from the Corel ontology. This ontology was created using SUMO (Suggested Upper Merged Ontology) [12] and enriching it with a taxonomy of animals created by the University of Michigan [11]. After that, the ontology was populated with the vocabulary of 374 terms used for annotating the Corel dataset. Among many categories, we can find animal, vehicle and weather.

For example, if the cluster topic is "animal" we will split the ranked list of results into sub ranked lists, one corresponding to every type of animal and an additional uncategorised rank. These ranks will then be merged to form a new rank, where all the documents at rank 1 in a sub list appear first, followed by the documents at rank 2, followed by rank 3 etc. The documents of equal rank in the sublists are ranked amongst themselves based on their rank in the original list. This way the document at rank 1 in the original list remains at rank 1 in the re-ranked list. We only maximise the diversity of the top 20 documents, after the 20th document the other documents maintain their ordering in the original list.

Similarly in WordNet clustering we build clusters of sub-categories of animal, bird, sport, vehicle and weather. We match these to the bag-of-words for each document contained in the text index. For example if the cluster topic is sport we will have a sub-ranked list containing every document that mentions "tennis". If for image clustering the cluster topic is not animal , vehicle or weather, or for WordNet clustering the topic is not animal, bird, sport, vehicle or weather, we default to location clustering. In this case we have a sublist for every different location a document is annotated with.

2.6 Data Fusion

With multiple sources of evidence being provided by the different query engines, a method of combining these data was required. Each query engine produced an output rank of data set images ordered with respect to their relevance to the input query. Not all engines gave values for the entire data set—some gave ranks of relevant sub-sets of the main data set.

We adopted a group consensus function based on the Borda Count method [6, 2] as it was simple to implement and fair to the input data, but it was extended by introducing per-rank scaling parameters.

Data were split into two categories which were treated differently; ranks and filters. Ranks were processed without adjustment and directly compared to each other during the combination process. Filters were used to filter out non-relevant results from an intermediate combined rank stage. This was used, for example, with the output of the geographic query engine, where it was judged more appropriate to 'filter in' explicitly determined relevant results than to treat it like a rank.

The final output of the data fusion algorithm was the combined result of the multiple input ranks and filters.

Run	CBIR	BRF	Text	Geographic
ImgTxt	0.3	-	0.7	-
ImgTxtBrf	0.25	0.1	0.65	-
TxtGeoMR	-	-	1.0	3.0
TxtGeoNEI	-	-	1.0	3.0
ImgTxt	0.0	-	1.0	-
ImgBrfWeighting	0.4	0.6	-	-
ImgBrfWeightingIterative	0.4	0.6	-	-
TxtGeo	-	-	1	42
ImgTxtGeo	0.9	-	0.1	44

Table 1: The weights and penalisation values for the various ranks used to produce the final runs of the system. The upper part of the table shows the parameters for the ImageCLEFwiki Task, whereas the lower shows the parameters for the ImageCLEFphoto Task.

The rank weights described in Stage 2 below were crucial for an effective combination process. We used a convex parameter vector W where the individual weights w_i were restrained by:

$$\sum_{i=1}^{n} w_i = 1 \tag{5}$$

where n is the number of pure ranks to combine.

This function then defined our parameter space within which we had to search to find the more appropriate weights for the rank combination process. We performed a brute force search of the entire parameter space to find the optimum weights. We used past years' data for training and optimum sets of weights were considered those that gave the maximum mean average precision (MAP). The final weights are given in Table 1 for the two tasks that used rank combination; the ImageCLEFwiki and the ImageCLEFphoto tasks.

The parameter values used by the filter stage were also important. Unlike the pure rank weights that were used to multiply an entire rank, the filters were used by penalising those values that were not present in the filter. These penalisation weights p_i, making up weight vector P, were subject to the same constraints and were derived in a similar way to the pure rank weights by exhaustively searching all possible values up to a limit which was defined as when the variation in the parameter yielded no change in the final rank's MAP value greater than 0.0001.

The overall process is described here as a five stage algorithm:

1. Read in rank data
   Data was provided by the query engines in ordered ranked lists with those elements that were judged to be most relevant by the engine appearing first, with each rank denoted R_i. Each of the m elements was assigned a value based on its position in the list, so the first element was given '1', and increased consecutively down the list until the last element had a value of m.

2. Multiply by rank weights
   Each query engine output list was scaled by multiplying every rank value in the list by its parameter value w_i.

3. Sum ranks
   The scaled ranks were then combined by producing a new rank $R\prime$ where every element r in each R_i was replaced by value $r\prime$. Stages 2 and 3 can be summarised thus:

$$r\prime = \sum_{i=1}^{m} w_i r_i \tag{6}$$

   where m is the number of ranks to combine and w_i is an element of the weight vector W.

4. Filter rank
   The newly combined intermediary rank $R\prime$ was then subjected to the output of each filter engine to produce rank $R\prime\prime$. For each element in the rank $R\prime$, any element that was not found in the filter data had its rank value penalised by that rank's parameter value as described by filter function $f(r)$. This pushed less relevant elements further down the ordered rank.

$$f(r\prime_i) = \begin{cases} r\prime_i & r\prime_i \text{ present in filter} \\ r\prime_i p_i & r\prime_i \text{ not present in filter} \end{cases} \tag{7}$$

5. Sort rank
   The filtered rank $R\prime\prime$ was then sorted to ensure that the rank values were in ascending order. The sorted rank is then the output of the combination stage of the overall system.

3 VCDT

3.1 Experiments

The objective of the Visual Concept Detection Task (VCDT) is to detect the presence or absence of 17 visual concepts in the 1000 images that constitute the test set. In addition to that, some confidence scores are provided once an object is detected. The higher the value the greater the confidence of the presence of the object in the image.

For the VCDT task, we submitted four different algorithms, all of them correspond to automatic runs dealing with visual information. The second run uses statistical information about visual concepts co-occurring together in addition to the visual information, and the final run is a combination of the other three.

3.1.1 Automated Image Annotation Algorithm

This algorithm corresponds to the work carried out by Yavlinsky et al. [19]. Their method is based on a supervised learning model that uses a Bayesian approach together with image analysis techniques. The algorithm exploits simple global features together with robust non-parametric density estimation using the technique of kernel smoothing in order to estimate the probability of the words belonging to a vocabulary being present in each one of the pictures of the test set. This algorithm was previously tested with the Corel dataset and the Getty collection.

3.1.2 Enhanced Automated Image Annotation Algorithm

This second algorithm is described in detail in Section 2.2. The submitted run is based on an enhanced version of the algorithm described in the previous section. The input is the annotations achieved by the algorithm developed by Yavlinsky et al. together with a matrix that represents the probabilities of all the words of the vocabulary being present in the images. This algorithm was also tested on the Corel collection of 5,000 pictures and a vocabulary of 374 words obtaining statistical significant results (5%).

3.1.3 Dissimilarity Measures Algorithm

The third algorithm follows a simple approach based on dissimilarity measures. Given one test image, we compute its global distance (Cityblock distance) to the mean of the training images which share one common keyword. The smaller the distance value the higher the probability for a test image to be annotated by the keyword of that category. Our submitted result is based on the combination of two results from two single feature spaces. One is the CIELAB colour feature. For each pixel, we compute the CIELAB colour values. Within each tile the mean and the second moment are computed for each channel. The other is the Tamura texture feature. The combination here is a simple additive combination of the probability for each test image for each category.

Algorithm	EER	AUC
Enhanced automated image annotation	0.284425	0.779423
Automated image annotation	0.288186	0.776546
Combined algorithm	0.318990	0.736880
Dissimilarity measures algorithm	0.410521	0.625017

Table 2: Comparative results of MMIS group

3.1.4 Combined Algorithm

This submitted run is based on the combination of all the other runs submitted by this team, in addition to one extra combined result formed from the output of a feature extraction algorithm for the Tamura and CIELAB features. This was a simple additive combination which, through testing was shown to be useful when used in combination with the other algorithm outputs. By testing the output of each individual algorithm on subsets of the training data they were produced with, a rough indication of performance of the algorithm per concept was derived. These figures then allowed an automated system to pick for each concept the best performing algorithm's output and combine it into a new result set. The individual algorithms' outputs were not adjusted or scaled in any way (other than the pre-combined result set mentioned above).

In our testing routines based on splitting the available training data into training and evaluation sets, the combination performed marginally better than the individual component algorithm outputs. This was due to selecting those algorithms which performed better at certain concepts and classes of concepts. Further work will be carried out to more robustly take advantage of concept classification when combining algorithm results.

3.2 Evaluations and Results

The evaluation metric followed by the ImageCLEF organisation is based on ROC curves. Initially, a receiver operating characteristic (ROC) curve was used in signal detection theory to plot the sensitivity versus (1 - specificity) for a binary classifier as its discrimination threshold is varied. Later on, ROC curves [3] were applied to information retrieval in order to represent the fraction of true positives (TP) against the fraction of false positives (FP) in a binary classifier. The Equal Error Rate (EER) is the error rate at the threshold where FP=FN. The area under the ROC curve, AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The results obtained by the four algorithms developed by our group are represented in Table 2. Our best result corresponds to the "Enhanced Automated Image Annotation" algorithm as seen in Figure 3.

3.3 Analysis

Our best algorithm was previously tested with the Corel dataset, a collection of 5,000 images and 374 terms obtaining a statistically significant (5%) improvement over the baseline approach followed by Yavlinksy et al. [19]. However, with the IAPR collection and the vocabulary of 17 terms used in VCDT, the results although better were not statistically significant.

4 ImageCLEFphoto

4.1 Experiments

In this section, we describe the nine submitted runs.

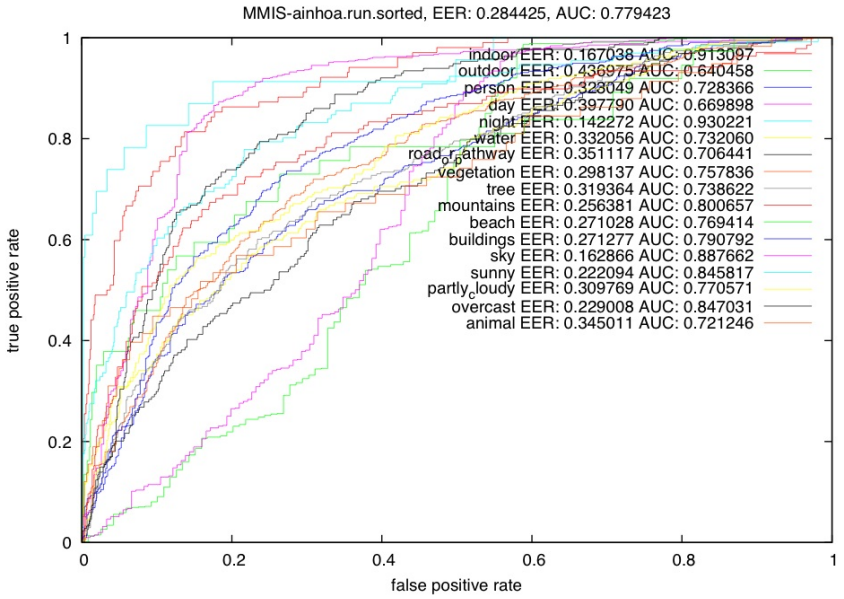

Figure 3: ROC curves for our best annotation algorithm

Txt. Our text only run uses Lucene as a base with TF•IDF term weighting and comparing queries to documents using the vector space model. Stemming is performed using the snow ball stemmer. Stop words are removed and text is lowercased and stripped of diacritics.

Img. This run is a pure CBIR. The dissimilarity value between an image example and an object image is computed by χ^2 Statistics measure based on their HSV colour feature space. Probability theory is adapted for combining the three dissimilarity values of the image examples in each topic with an object image. The final result is the Mean Average Precision (MAP) of 39 query topics.

ImgBrfWeightingIterative. This run is a combination of image evidence and text evidence. BRF is employed in this run. The top seven examples from ranked text results of iteration one were taken as new image examples for each query topic. The same search method as with "Img" is used in this run. The final result is the combination of the result ranks of this run and the ranks of "Img" using weight 0.6 and 0.4.

ImgBrfWeighting. The differences between this run and "ImgBrfWeightingIterative" is that the BRF rank for this run was produced by summing the ranks of the first five results of the text engine's results when the original query was used as input.

ImgTxt / TxtGeo / ImgTxtGeo. These three runs are the combination of image evidence and text evidence, text evidence and geographic evidence, image evidence and text evidence and geographic evidence, respectively. These evidences were merged by the data fusion process as described in Section 2.6. The ranks' and filters' weights are derived through exhaustive search of their parameter spaces. The final combination is made with optimal weights (see Table 1).

ImgTxtGeo-ImageCluster. This run combines the text, image and geographic data as with the ImgTxtGeo run and then clusters the result using the Image Clustering method described in Section 2.5.

Run ID	MAP	GMAP	CR20
Txt	0.0923	0.008	0.1926
Img	0.0256	0.0071	0.0449
ImgBrfWeightingIterative	0.0326	0.0089	0.2286
ImgBrfWeighting	0.0217	0.0058	0.1894
ImgTxt	0.0715	0.0241	0.2231
TxtGeo	0.0696	0.0025	0.2061
ImgTxtGeo	0.078	0.0291	0.2516
ImgTxtGeo-ImageCluster	0.0465	0.0221	0.2388
ImgTxtGeo-MetaCluster	0.0461	0.0258	0.2503

Table 3: Results of submitted runs of ImageCLEFphoto task

ImgTxtGeo-MetaCluster. This run combines the text, image and geographic data as with the ImgTxtGeo run and then clusters the result using the WordNet Clustering method described in Section 2.5.

4.2 Results

Table 3 shows the result of nine runs submitted for the ImageCLEFphoto task, from which the following observations can be made. Firstly, with respect to MAP, the BRF method which multiplied seven ranks outperforms the BRF method summing five ranks. Secondly, with respect to the Geometric Mean Average Precision (GMAP), combining text and image data outperforms plain text data. As GMAP emphasises performance in the worst case, this shows that image retrieval definitely has a contribution to the combined result; the reason for the lower MAP is simply because we have not determined the best way to combine the two evidences yet.

Unfortunately there was no significant difference between either of our clustering methods and the inclusion of geographic information actually gave us worse results than our text baseline so we can draw no further conclusions.

5 ImageCLEFwiki

The ImageCLEFwiki task aims to investigate retrieval approaches in the context of a larger scale and heterogeneous collection of images. The dataset of this task is a wikipedia image collection, which contains 151,518 images created and employed by the INEX Multimedia track in 2006-2007 [18]. 75 topics are considered.

5.1 Experiments

In this section, we describe the six submitted runs.

SimpleText. This run is our baseline – pure text based search. The detailed description can be found in Section 2.3.

TextGeoNoContext. This run is a combination of text evidence and geographic evidence. It uses the MR method or placename disambiguation described in Section 2.4, which is not context aware. The geographic filter and text rank were combined by multiplying the rank of each document in the text rank not appearing in the geographic filter by a penalisation value of 3.0 (detailed in Section 2.6).

Run ID	MAP	P@5	P@10	R-prec.	Bpref
SimpleText	0.1918	0.3707	0.3240	0.2362	0.2086
TextGeoNoContext	0.1896	0.3760	0.3280	0.2358	0.2052
TextGeoContext	0.1896	0.3760	0.3280	0.2357	0.2052
Image	0.0037	0.0187	0.0147	0.0108	0.0086
ImageText	0.1225	0.2293	0.2213	0.1718	0.1371
ImageTextBRF	0.1225	0.2293	0.2213	0.1718	0.1371

Table 4: Results of runs of ImageCLEFwiki task

TextGeoContext. This run is a combination of text evidence and geographic evidence. It uses the context aware Neighbourhood method or placename disambiguation described in Section 2.4. Text and geographic evidence are combined in the same way as the TextGeoNoContext run.

Image. This run is a pure content based image search. Gabor texture features were extracted from the test collection and used to form a high dimensional space. The query images were compared to the corpus images using the Cityblock distance.

ImageText. This run is a combination of text evidence and image evidence. The two were combined using a convex combination of ranks (Section 2.6) using weights 0.3 and 0.7.

ImageTextBRF. This run is a combination of text evidence and image evidence. The same text and image retrieval system are used as in the previous section. We use the top five results from the text retrieval results as query images for our BRF system. The three were combined using a convex combination of ranks using weights 0.25, 0.1 and 0.65 for Image, BRF and Text relevance respectively.

5.2 Results

Table 4 shows the results of our six runs.

From the results we can see that our baseline (the "SimpleText" run) outperformed all the other methods based on the majority of performance measures. While the pure content based image search run gives us the worst performance. There were negligable differences whether we used context based placename disambiguation or not. Similarly only negligable differences were seen whether blind relevance feedback was employed or not.

6 Conclusions

Our conclusions from ImageCLEF'08 are limited as none of our experiments achieved a statistically significant improvement over a baseline. Discussions of the results are provided below.

VCDT. The enhanced automated image annotation method performed well appearing in the top quartile of all methods submitted, however it failed to provide significant improvement over the automated image annotation method. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm and another in the nature of the vocabulary itself, where instead of incoherence we have mutually exclusive terms and almost no semantically similar terms.

ImageCLEFphoto. As mentioned in Section 4.2, despite the combination of Text and Image evidence performing worse than the text baseline with respect to MAP, the superior GMAP shows that Image evidence is improving performance in the worst case. In future work we would like

to explore further data fusion methods and find a way to take full advantage of this additional evidence without undermining text retrieval where it is performing well.

ImageCLEFwiki. Our text baseline outperformed all other retrieval methods. From this we conclude that text is by far the dominant feature on retrieval for this heterogeneous collection. The fact that blind relevance feedback made negligible positive or negative difference to image retrieval combined with the very low retrieval results for image retrieval alone, implies simple features can contribute little to retrieval on this collection. Geographic retrieval offered some improvement in some measures (p@10 and R-prec.), but not statistically significant. We can only conclude that a geographically aware system *could* provide some improvement, but due to the short length of the documents, context based placename disambiguation will be unlikely to provide a significant improvement.

References

[1] H Cunningham, D Maynard, V Tablan, C Ursu, and K Bontcheva. Developing language processing components with GATE. Technical report, University of Sheffield, 2001.

[2] M Van Erp and L Schomaker. Variants of the Borda Count method for combining ranked classifier hypotheses. In B. Zhang, D. Ding, and L. Zhang, editors, *International Workshop on Frontiers in Handwriting Recognition*, 2000.

[3] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[4] A Hanbury and J Serra. Mathematical morphology in the CIELAB space. *Image Analysis & Stereology*, 21:201–206, 2002.

[5] P Harping. *User's Guide to the TGN Data Releases*. The Getty Vocabulary Program, 2.0 edition, 2000.

[6] T K Ho, J J Hull, and S N Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence*, 16(1):66–75, 1994.

[7] P Howarth and S Rüger. Robust texture features for still-image retrieval. *Vision, Image and Signal Processing*, 6(152 (6)):868–874, 2005.

[8] H Liu, D Song, S Rüger, R Hu, and V Uren. Comparing dissimilarity measures for content-based image retrieval. In *Asian Information Retrieval Symposium*, pages 44–50, 2008.

[9] N Maillot, J Chevallet, V Valea, and J H Lim. IPAL inter-media pseudo-relevance feedback approach to imageCLEF 2006 photo retrieval. In *CLEF 2006 Workshop, Working notes*, 2006.

[10] G A Miller and W G Charles. Contextual correlates of semantic similarity. *Journal of Language and Cognitive Processes*, 6:1–28, 1991.

[11] P Myers, R Espinosa, C S Parr, T Jones, G S Hammond, and T A Dewey. The animal diversity web (online). http://animaldiversity.org.

[12] I Niles and A Pease. Towards a standard upper ontology. In *International Conference on Formal Ontology in Information Systems*, pages 2–9, New York, NY, USA, 2001. ACM Press.

[13] S Overell and S Rüger. Geographic co-occurrence as a tool for GIR. In *CIKM Workshop on Geographic Information Retrieval*, 2007.

[14] M J Pickering and S Rüger. Evaluation of key frame based retrieval techniques for video. *Computer Vision and Image Understanding*, 92(2):217–235, 2003.

[15] Apache Lucene Project. http://lucene.apache.org/java/docs/. Accessed 1 August 2007, 2007.

[16] H Tamura, T Mori, and T Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics*, 8(6):460–473, 1978.

[17] Princeton University. WordNet, online lexical database. http://www.cogsci.princeton.edu/∼wn/.

[18] T Westerveld and R van Zwol. The inex 2006 multimedia track. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *Advances in XML Information Retrieval: INEX 2006*. Springer-Verlag, 2007.

[19] A Yavlinsky, E Schofield, and S Rüger. Automated image annotation using global features and robust non-parametric density estimation. In *International ACM Conference on Image and Video Retrieval*, pages 507–517, 2005.

[20] D Zhang and G Lu. Evaluation of similarity measurement for image retrieval. In *International Conference on Neural Networks & Signal Processing*, pages 928–931, 2003.