Properties of Optimally Weighted Data Fusion in CBMIR

Peter Wilkins
CLARITY: Centre for Sensor
Web Technologies
Dublin City University
Ireland
pwilkins@computing.dcu.ie

Alan F. Smeaton
CLARITY: Centre for Sensor
Web Technologies
Dublin City University
Ireland
alan.smeaton@dcu.ie

Paul Ferguson
CLARITY: Centre for Sensor
Web Technologies
Dublin City University
Ireland
pferguson@computing.dcu.ie

ABSTRACT

Content-Based Multimedia Information Retrieval (CBMIR) systems which leverage multiple retrieval experts (E_n) often employ a weighting scheme when combining expert results through data fusion. Typically however a query will comprise multiple query images (I_m) leading to potentially $N \times M$ weights to be assigned. Because of the large number of potential weights, existing approaches impose a hierarchy for data fusion, such as uniformly combining query image results from a single retrieval expert into a single list and then weighting the results of each expert. In this paper we will demonstrate that this approach is sub-optimal and leads to the poor state of CBMIR performance in benchmarking evaluations. We utilize an optimization method known as Coordinate Ascent to discover the optimal set of weights ($|E_n| \cdot |I_m|$) which demonstrates a dramatic difference between known results and the theoretical maximum. We find that imposing common combinatorial hierarchies for data fusion will half the optimal performance that can be achieved. By examining the optimal weight sets at the topic level, we observe that approximately 15% of the weights (from set $|E_n| \cdot |I_m|$) for any given query, are assigned 70%-82% of the total weight mass for that topic. Furthermore we discover that the ideal distribution of weights follows a log-normal distribution. We find that we can achieve up to 88% of the performance of fully optimized query using just these 15% of the weights. Our investigation was conducted on TRECVID evaluations 2003 to 2007 inclusive and ImageCLEFPhoto 2007, totalling 181 search topics optimized over a combined collection size of 661,213 images and 1,594 topic images.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Measurement, Experimentation

Keywords

Data Fusion, Multimedia Fusion, Content-Based

1. MOTIVATION

Content-based multimedia information retrieval (CBMIR) systems often combine multiple sources of evidence to answer an information need. These systems typically employ multiple 'retrieval experts' whose outputs are combined to create a response. This problem can be phrased as a *combination of experts* problem and is a case of *Data Fusion*. CBMIR is particularly characterized by the use of multiple 'noisy' signals such as the color of an image, or the textures it contains, and through combining multiple sources of noisy information, reasonable performance can be achieved [17]. However this makes the role of weighted data fusion paramount to the success of many CBMIR systems.

Given that the performance of any such retrieval system is dependant upon the optimal generation of weights, and the manner in which those weighted documents[1] are combined, it is essential that we know something about what the ideal distribution of weights for a query is, and the manner in which to combine these documents.

In order to gain significant insights into what the ideal form of weighting for CBMIR data fusion is, and indeed how ranked lists should be combined, we need to deviate from the traditional empirical model typically used to evaluate new algorithmic advances. Traditionally we have some form of training data, either topics, data or both, some proposed model which we want to test and some form of parameters which require tuning and a set of evaluation metrics and relevance assessments. Also included in this model is a test set from which final results will be reported. The common sequence of events is that a model is first optimised on training data, then the optimised model is used on the test data. The final result typically reported is the outcome of the evaluation metrics run on the output of the model on the test data, where presumably the model has not been overfitted.

The major problem with the established empirical model is that of evaluation, where what we want to determine is *not* the comparison of competing models and their associated performance, but rather given a maximally performing model, what parameters were associated with it? In data fusion tasks we will have a range of input sources of evidence, which we will then combine in some manner in order to compute a final response. The fundamental problem is that using the established empirical model, we could eval-

[1] The term 'document' will refer to any multimedia artifact.

uate two different fusion systems, and after executing both having first trained on the training collection we can make the observation that system 'a' outperforms system 'b' by 15%. On the surface this seems fine, system 'a' has achieved a good performance improvement over system 'b'. However, this 15% is a *relative* increase, it is only meaningful when comparing the two systems under observation, and far less important when we do not know when given a fixed set of inputs, what the maximum achievable performance is. For instance, if system 'a' scored a MAP of 0.2, but the theoretical maximum attainable given the same inputs was 0.8, then the relative importance of system 'a's 15% improvement is diminished as there is clearly room for greater improvement. However, if the maximum was determined to be 0.21, then that 15% improvement is very significant.

A better use when determining what the maximum performance is for a fixed set of inputs, is to study the properties of this maximally performing model. Rather than being primarily concerned with the maximum performance MAP value, to flip this around such that given we have a model which achieves excellent performance, what are the properties of this model that led to this performance. In order to achieve this it necessitates optimisation directly on the test data. We are in-fact not proposing any particular model for data fusion in this paper, but rather we have created and observed the optimal model for this retrieval problem, such that we can report on the properties of this model which future systems should seek to leverage.

The objective of this paper therefore is to study what are the variables which generate a maximally performing CBMIR data fusion system, and if there are any commonalities to these that can be discovered so as to inform the development of new data fusion algorithms and systems. The performance of these models is in themselves immaterial, it is these factors which we wish to identify and examine. In particular in this paper, we will be examining the impact of combinational hierarchies upon performance, and the distribution that an ideal form of weighting takes.

No doubt at this point after mention of optimizing directly on the test set, alarm bells are ringing in several readers minds, however we believe we have good justification for doing this. Whilst we have laboured the point as to why we are examining models optimized on the test data, we believe this is necessary as the approach is unorthodox, and inadequately justified can lead to the the results presented in this paper being dismissed out of hand.

The paper is organized as follows. Section 2 presents related work in data fusion and evaluation, with reference to CBMIR applications. Section 3 details our experimental environment including retrieval experts, data sets, and a brief discussion of the optimization technique. Section 4 presents the results of our optimization on TRECVID 2003-2007 and ImageCLEFPhoto 2007 and discusses observations on this data set. Section 5 tests these observations to determine if they can be exploited by existing data fusion approaches. We finish with our conclusions in Section 6.

2. RELATED WORK

Early data fusion research began with experimentation into the combination of different retrieval models, document representations and query representations [13, 6, 16]. Research by Belkin *et al.* [2, 3] noted that varying these different factors produced different sets of relevant documents, yet exhibited no major changes in performance metrics. Croft [5] notes that observations from these early studies suggested that it was beyond the capabilities of a single system to retrieve all the relevant documents for a given query. According to Croft this resulted in two streams of IR systems being developed, one stream was to create single models which can combine multiple *sources* of evidence such as the INQUERY system based on an inference network. The alternative stream is the development of systems which effectively combine the outputs of multiple searches from different retrieval models [7]. Croft notes for the task of multimedia retrieval, as different modalities are combined this requires the development of systems which combine the ranking from multiple subsystems (what we would term experts) [5].

Investigations into the behaviour of data fusion began with observations about low overlaps of the documents returned by different ranking models [13, 6, 16]. Belkin *et al.* [2] found "Different representations of the same query, or of the documents in the database, or different retrieval techniques for the same query, retrieve different sets of documents (both relevant and nonrelevant)". Lee [9] examined this research but contrasted it against the findings of Turtle *et al.* and Saracevic *et al.* [18, 16], where Turtle *et al.* in experiments combining probabilistic and Boolean retrieval results found that the relevant documents retrieved were shared by both approaches, whilst Saracevic & Kantor found that different query formulations found different documents, but that a document's odds of being judged as relevant increased monotonically as a document appeared in multiple result sets. Lee took these findings to formulate a new hypothesis for the effectiveness of data fusion: "different runs might retrieve similar sets of relevant documents but retrieve different sets of nonrelevant documents" [9].

Testing this hypothesis Lee introduced two evaluation metrics to measure the degree of overlap between relevant documents and nonrelevant documents, termed $R_{overlap}$ and $NR_{overlap}$. Lee [9] finds that the best result from data fusion was achieved when result sets were combined in which relevant documents had high overlap and low overlap for non-relevant documents. The work of Vogt *et al.* [19] confirms Lee's observations by conducting pairwise experiments combining 61 TREC submissions. Vogt and Cottrell term the relevant overlap as the 'Chorus Effect', that multiple retrieval systems return the same relevant documents.

Croft [5] interprets the findings of the work of Lee and of Vogt and Cottrell as being the result of combination of uncorrelated classifiers. Assuming that the retrieval systems being combined are good, that as the result lists being combined are truncated to 1000 results, and that for a given TREC query there are typically only 100-200 relevant documents, that most good systems will return within the 1000 results the 100-200 relevant documents, but as the 'classifiers' (search systems) are uncorrelated, they will return different sets of nonrelevant documents. Furthermore this emphasizes earlier observations that combinations of independent good search systems, produce gains in performance when fused [5].

The data fusion hypothesis of Lee was critically examined by both McCabe *et al.* [11] and Beitzel *et al.* [1]. Both conducted approaches where various system parameters were held constant whilst varying one aspect, such as the ranking model, stemming, stopping, relevance feedback etc. The work of McCabe *et al.* found that when systemic parameters are held constant, that the combination of vector, probabilistic and Boolean retrieval models did *not* improve performance of retrieval, contrary to previous accepted wisdom. This was further demonstrated by a lack of performance improvement when combining results from TREC-6, 7 and 8 queries which produced high overlaps in both $R_{overlap}$ and $N_{overlap}$, meaning that each of the approaches were returning very similar content. Nevertheless this work found that the overlap coefficients were a good predictor of the potential for performance improvement with data fusion, particularly when systems were combined with weights, such that a poor performing system could be discounted. The combination of a poor system with a good system, using weights where the good system was weighted highly, produced performance increases, lending support to the application of weights for expert combination [11].

Beitzel *et al.* [1] like McCabe also conducted experiments where system parameters are held constant to measure the impact of combination of different aspects of retrieval systems. The work of Beitzel *et al.* specifically examined the combination of "highly effective retrieval strategies". Assuming this, Beitzel *et al.* hypothesize that combination of highly effective systems through voting mechanisms like CombMNZ are more likely to harm performance, as the highly effective systems have already been optimized and will rank relevant documents highly, therefore the candidates for promotion up a ranked list are lower ranked common nonrelevant documents as the relevant documents are already highly ranked. They further hypothesize that as constants such as the query and stemming for each retrieval model are held constant, that different models will produce approximately the same set of documents for a query, only the relative ranks of these sets are likely to be different. For highly effective systems Beitzel *et al.* found that the combination of retrieval models (e.g. vector space and probabilistic) hurts performance, rather than helps, whilst the overlap coefficients defined by Lee [9] provide a poor indicator of potential for improvement through data fusion [1].

These two results however give credence to the application of weighted data fusion to the task of CBMIR. Given that CBMIR is characterized by the combination of multiple poor retrieval experts [17], we are unlikely to be combining multiple experts that actually perform consistently well for any set of queries. Furthermore as the work of McCabe shows, weighted combination of poor retrieval experts can lead to significant performance improvements.

Within the multimedia research community, several fusion approaches have been investigated. Yan et al. propose the use of 'Query-Class' dependent weights [22], where a set of predefined query classes are assigned feature weights learned from the training data. This approach is extended by Kennedy et al. [8] to automatically discover query classes from training data. These approaches, however, typically weight entire features (retrieval experts).

Two previous investigations into the role of feature combination for multimedia retrieval have been completed, by McDonald and Smeaton [12] and by Yan and Hauptmann [21]. McDonald and Smeaton empirically compare combination approaches for score, rank and probability techniques. Their work evaluated these approaches by optimizing Mean Average Precision (MAP) on a training collection with multiple topics, then applying these generalized optimized parameters to a test set. Our work differs as our optimizations occur at the topic level (Average Precision), rather than the topic set level (MAP). Furthermore the fusion approaches detailed in [12] use a hierarchical approach which is likely to obscure the effect individual query images have on performance. Yan and Hauptmann [21] conduct experiments with TRECVID 2002 data to construct a theoretical framework for studying the upper bounds of combination functions. They found that linear forms of combination may be too restrictive for large numbers of experts to be combined effectively. However, like McDonald and Smeaton, this work examined combination at the expert level and as such did not delve down to the granularity of pairs $\langle I_i, E_j \rangle$.

Eval.	Type	Keyframes	Topic Images	Topics
TRECVID 03	Video	72,462	138	25
TRECVID 04	Video	48,818	160	24
TRECVID 05	Video	78,206	228	24
TRECVID 06	Video	146,497	169	24
TRECVID 07	Video	295,350	719	24
ImageClef 07	Photo	20,000	180	60

Table 1: Details of corpora used

3. EXPERIMENTAL SETUP

The task we explore is an ad-hoc search task, where a system is given an expression of an information need and is required to return as many relevant matches as possible. For our investigation we performed 'fully automatic' retrieval which processes a query with no human intervention.

We used six different multimedia corpora, five of which came from TRECVID [17] and one from ImageCLEF [4]. These two campaigns share similar objectives as both seek to promote research in content-based retrieval by utilizing common test collections and open, metrics-based evaluations. As previously noted, for all corpora we only consider the visual information provided. In the case of video, we use the extended keyframe set provided, meaning that in many cases we index more than one keyframe per shot. For query descriptions we make use of all visual data provided. In the case of videos used as part of the topic description, we sample keyframes from this video and add it to the topic image set. The six corpora we used are described in Table 1.

We make use of six global visual features defined in the MPEG-7 specification [10]: Scalable Color (SC), Color Structure (CS), Color Layout (CL), Color Moments (CM), Edge Histogram (EH) and Homogeneous Texture (HT). To compute an answer to a visual query, we take the topic images and we query them against each retrieval expert, producing for each pair $\langle I_i, E_j \rangle$ a ranked list of results. For our experiments we produced ranked lists of 1000 results per pair $\langle I_i, E_j \rangle$. Each ranked list is normalized using MinMax [7], then weighted and linearly combined using CombSUM [7]. We would note here that we deliberatly choose CombSUM over CombMNZ, as through the processes discussed in the paper, we have empirically shown that CombSUM with linear weighting offers superior performance to that of weighted CombMNZ (due to space constraints we cannot explore these results, see: [20]). The ranking metric for each expert is implemented as defined by the MPEG-7 standard, typically a variation on Euclidean distance.

The optimization method we use in this work is known as *Coordinate Ascent* (also known as *Alternating Variables Method*). It is a method which is able to optimize directly on Average Precision (AP), by randomly initializing the linear weights, then finding a local maxima based on AP. The method is then repeated multiple times so that a global maxima can be found. This approach has been used to good effect by Metzler and Croft, and a complete explanation of this method can be found in their work [14].

3.1 Hierarchical Combination Approaches

There are three basic levels of combination available to CBMIR designers whose systems utilize multiple retrieval experts and multiple query components: combination at the 'query' level [12], combination at the 'expert' level [22] and direct combination. Figure 1 illustrates these variations.

The elements for weighting can be formally defined as follows. A CBMIR search topic will contain multiple example visual images, Images $I = \{queryimage_i\ ...\ queryimage_n\}$ where $1 \leq i \leq n$. A CBMIR system will have at its disposal multiple retrieval experts, $E = \{expert_j,\ ...\ expert_m\}$ where $1 \leq j \leq m$. Therefore, we can define the pair $\langle I_i, E_j \rangle$, which is a unique coupling of every example query image to every visual retrieval expert. This will generate $n \times m$ pairs.

We can now further define the set of weights to be tuned at each of the three different levels of combination. For "Query" level combination, the results of every retrieval expert for a specific query image (I_i) are linearly combined with uniform weight into a single ranked list which represents a given image. The merged results lists for every image are then weighted and combined, meaning for this level we need to optimize n weights (i.e. $|I|$), giving Image Weights $IW = \{w_i\}$ where w is the weight and $\sum w_i = 1$.

At an "Expert" level of combination we execute the opposite. For a specific expert (E_j) we query against it all query images for the topic, merging the results to produce for each expert a single ranked list. We then weight the result list of each expert, and combine to form our final ranked list. In this level, we are required to optimize m weights (i.e. $|E|$), formally Expert Weights $EW = \{w_j\}$ and $\sum w_j = 1$.

Finally we have the "direct" level of combination which specifies weights for every coupling of an example image and retrieval expert. That is, for every pair $\langle I_i, E_j \rangle$ we are required to set a weight. This will produce a weight set of size $n \times m$ (i.e. $|I| \cdot |E|$) where $IEW = \{w_{ij}\}$, where i refers to the query image, j refers to the retrieval expert and $\sum w_{ij} = 1$.

4. OPTIMIZATION RESULTS

We performed the optimization of CBMIR on TRECVID 2003 to 2007 and on ImageCLEFPhoto 2007. The results are presented in Table 2 (where TV is TRECVID and IC is ImageCLEFPhoto).

Eval.	TV03	TV04	TV05	TV06	TV07	IC07
MAP	0.122	0.108	0.141	0.056	0.130	0.216
Uniform	0.059	0.029	0.065	0.016	0.042	0.128
BR	N/A	N/A	0.126	0.087	0.087	0.189*

Table 2: Optimized Results compared to 'Best Reported' (BR). 'Uniform' represents using all pairs $\langle I_i, E_j \rangle$ with no weighting. *IC07 BR is visual only

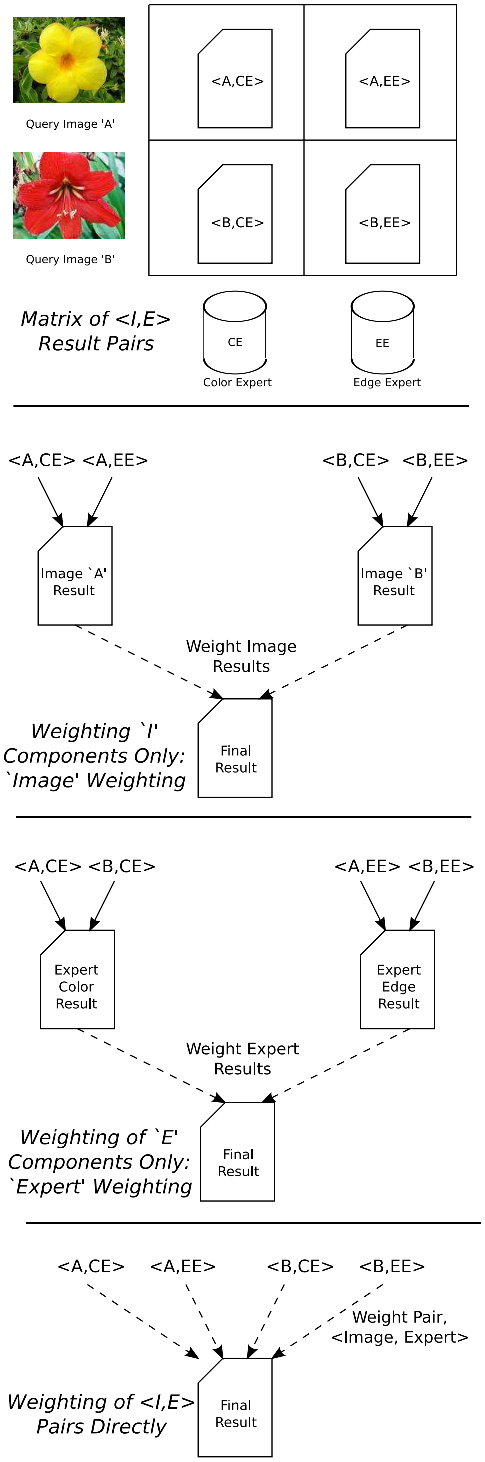

Figure 1: Levels of combination for a single search topic, with 2 retrieval experts (E) and 2 example query images (I), giving 4 ranked lists (pairs $\langle I_i, E_j \rangle$). Three levels are available, combination at the 'Query' Level, combination at the 'Expert' level and direct combination without any hierarchy.

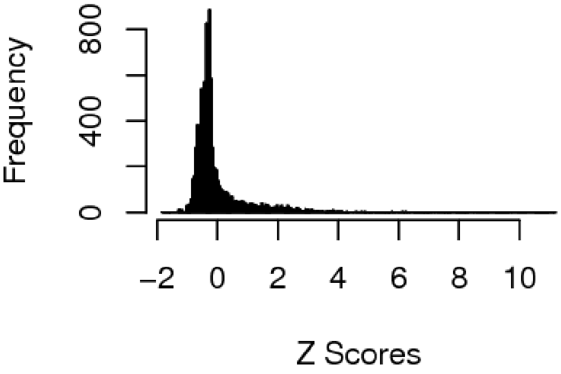

Figure 2: Standard Scores for assigned weights across all corpora

This optimization generated automatic retrieval runs which achieved excellent performance with the use of no semantic information or text. For comparison, the row 'BR' shows the best reported automatic system MAP from that year's activity. These figures are actually a bit startling, as the very high levels of performance achieved run contrary to expectations from previous experiments [17]. This is particularly apparent if we compare the optimized MAP to the 'Uniform' MAP. The 'Uniform' map demonstrates a retrieval run where all pairs $\langle I_i, E_j \rangle$ are equally weighted, i.e. there is no weighting at all. We can see that the optimal weights applied to pairs $\langle I_i, E_j \rangle$ can produce up to a 300% increase in performance. The impact of these figures is such that people question if such performance is achievable with low-level visual MPEG-7 features and no text, as it runs contrary to previous experimental knowledge.

We show the comparison to the best reported runs in that year's evaluation, as it demonstrates the effectiveness of our optimization, producing retrieval runs which achieve excellent performance. The comparison highlights the maximum of what can be achieved with data fusion and global low-level visual features, particularly when compared against the top performing runs which made use of multiple evidence modalities including text and semantic information. We note that this comparison to published retrieval runs ('BR') is not a fair comparison as we optimized on the test data, however the intention of this work is to demonstrate the gains achievable with optimized weights, even when compared against retrieval runs that used high quality signals such as text.

We analyzed the optimal weight topic sets IEW generated for each topic and calculated the standard score (also known as Z-Score). The standard score allows us to express for any given pair $\langle I_i, E_j \rangle$ weight w_{ij} how far from the topic mean weight it is in terms of standard deviations. This provides us with a measure which can be used across topics reliably. Figure 2 is a histogram of the distribution of standard scores across all topics and corpora.

We can infer multiple insights from the presented distribution and measures of central tendency. Firstly, that whilst the distribution of weights has some properties of that of a normal distribution, such as a majority of the data points clustered around the mean and within the range $\pm 3\sigma$, there does exist a very definitive positive skew. Secondly, as part of this positive skew approximately 10%-11% of the weights were assigned values $> 1\sigma$. The implications of this are that overall the initial observations would suggest that a minority of the pairs $\langle I_i, E_j \rangle$ received the majority of a topic's weight.

Without other evidence there remains the possibility that the effect presented is a corpora-specific event and that the weights are indeed more normally distributed. To account for this we present in Figure 3 a corpora-specific plot of the weight distribution in the form of quantile-quantile (Q-Q) plot. In this figure, the x-axis represents a theoretical normal distribution of weights, whilst the y-axis is the actual weight which was assigned. The dashed line displays the trend line of the weights if they were normally distributed.

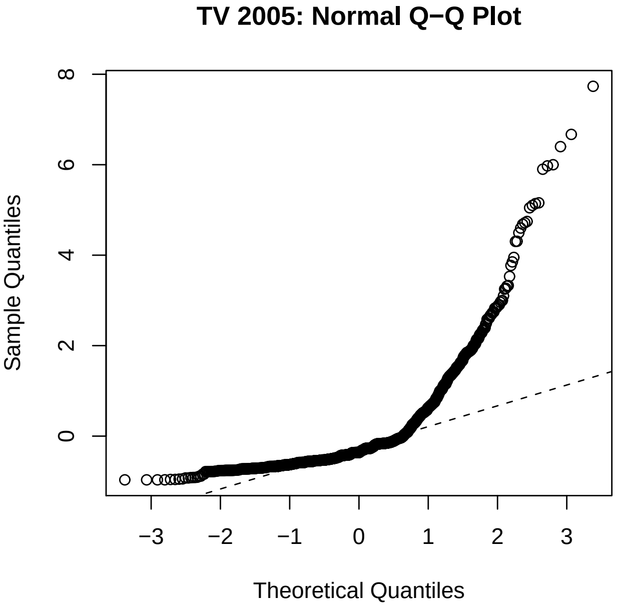

Figure 3: Weight Distribution for TRECVID 2005

Examining the Q-Q plot, we can see the same distributional pattern, as it demonstrates a significant departure from a normal distribution, particularly once the normalised weights values exceed 1σ. The pattern shown in the plot is similar to what would be expected if the distribution of the weights was log-normal, again we can also see demonstrated in each plot a positive skew. Whilst the data presented in this Figure is only for 2005, this pattern was repeated for all of our experimental corpora [20]. Based on this evidence this indicates that within topics, a minority of $\langle I_i, E_j \rangle$ weights are assigned a majority of the topic weight mass.

To explore this, we examined each corpus and its topics to determine where topic weight was assigned. For each topic we set a threshold of $+1\sigma$ and calculated the total amount of weight mass which was more than $+1\sigma$ from the mean weight, and what percentage of $\langle I_i, E_j \rangle$ were assigned these weights over this threshold, i.e. $w_{ij} > 1\sigma$. The results of our analysis are presented in Figure 4 which show two columns for each topic. The first column in blue (dark), represents the total amount of weight allocated in that topic which was $+1\sigma$ greater than the mean weight. The second column in yellow (light) represents how many of the Query-Terms value of w_{ij} was more than $+1\sigma$ from the mean. For example, the first graph in Figure 4 represents topics from TRECVID 2005. In topic '0149' we can observe a blue bar at 70%, and a yellow bar at 6%. This means that for topic '0149', 6% of the pairs $\langle I_i, E_j \rangle$ used for that topic were allocated 70% of the weight and the remaining 94% of $\langle I_i, E_j \rangle$ had only 30% of the topic weight.

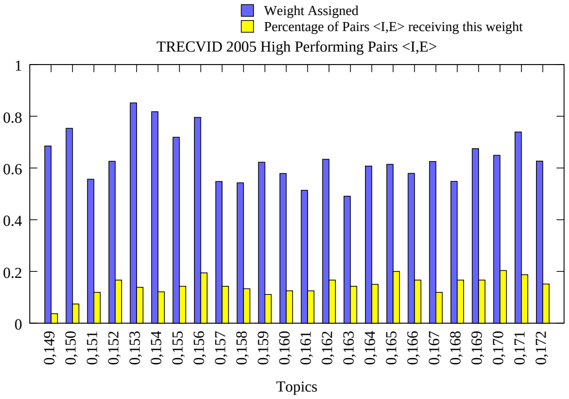

Figure 4: Evaluation Campaign Weights

We can see from this graph that for all topics across TRECVID 2005, achieving the maximum average precision possible is dependant upon specific pairs $\langle I_i, E_j \rangle$ being allocated the bulk of the weight for that topic, rather than specific experts EW being correctly weighted. Whilst we only show TRECVID 2005 here, the patterns expressed in this graph are replicated for all evaluations examined.

The possibility exists that we are inadvertently seeing one retrieval expert for a topic performing strongly, and thus all pairs $\langle I_i, E_j \rangle$ which utilize that expert are up-weighted. We examined the distribution of each of the six experts within the set of highly weighted $\langle I_i, E_j \rangle$ to determine if there was a bias towards any particular expert, shown in Table 3.

CL	CM	CS	SC	EH	HT
15%	17%	13%	13%	24%	18%

Table 3: Distribution of Retrieval Experts in $\langle I_i, E_j \rangle$ with $w_{ij} > 1\sigma$

There is a slight bias towards EH and to a lesser extent the HT experts, however as there are only two texture but four color experts, this bias can be accounted for. The data presented in Figure 4 and Table 3 shows that highly weighted $\langle I_i, E_j \rangle$ are distributed across different experts.

We observe that the key to maximizing AP is to correctly identify salient pairs $\langle I_i, E_j \rangle$ and ensure that these are highly weighted, rather than weighting the overall performance of any given retrieval expert. To test this observation, we devise a series of experiments that utilize only highly weighted pairs $\langle I_i, E_j \rangle$ to see if we still achieve good performance. A highly-weighed pair $\langle I_i, E_j \rangle$ is a pair whose weight w_{ij} is greater than $+1\sigma$ of the mean weight for that topic.

5. EXPERIMENTS

To test our observations we devised three experiments in order to (1) determine to what extent the highly-weighted pairs $\langle I_i, E_j \rangle$ impact upon performance; (2) to determine if the weighting of these pairs needs to be exact or if merely identification is enough; and finally, (3) to determine the impact the remainder of the pairs $\langle I_i, E_j \rangle$ which do not have much weight allocated to them have upon performance. As a comparison we have also included two optimizations, "Query" level and "Expert" level optimizations. These represent the best performance achievable if we utilize existing data fusion methods (such as Query-Class, single feature machine learning [22][15]), and allow us to determine if our suggested strategies of targeted weighting of pairs $\langle I_i, E_j \rangle$ rather than expert level weighting offers improvement.

- **(1σ) 1σ**: For each topic, only use highly-weighted pairs $\langle I_i, E_j \rangle$ (i.e. pairs $\langle I_i, E_j \rangle$ whose assigned value from optimization was $+1\sigma$ for the mean weight). The value of w_{ij} will be the value determined during optimization (Section 4). This test will examine the impact of precisely weighted high-performing pairs $\langle I_i, E_j \rangle$. It can be thought of as a high-precision experiment as for each topic we will be using only 5%-20% of the available ranked lists for that topic.

- **(1σU) 1σ Uniform**: Using only the highly-weighed pairs $\langle I_i, E_j \rangle$, assign each a uniform weight. This will examine if just the identification of high-performing pairs $\langle I_i, E_j \rangle$ is sufficient to yield performance increases, specifically determining if accurate weighting of pairs is required, or if they can be assigned a binary weight [0,1]. As the task of determining the optimal set w_{ij} is realistically only viable post-experiment, this experiment tests if realistic fusion approaches can be developed, as it does not require perfect weights, only identification of likely high performing pairs $\langle I_i, E_j \rangle$.

- **(1σU-T) 1σ & Tail**: We extend experiment 1σ, by taking the remaining weight mass that isn't assigned to high-performing pairs and allocate it uniformly amongst the remaining pairs in IEW. This experiment complements the previous, we assign a large weight to the high-performing pairs, whilst a low weight to the remainder. As the high-performing pairs constitute only 5%-20% of available pairs for a topic, this experiment is testing the impact of recall, i.e. can we include the remainder of the data without accurate weighting so as to increase our recall.

- **Expert Optimized**: We implement the "Expert" level of combination as described in Section 3.1 and as is implemented by several data fusion approaches. Here we utilize the optimization approach as described in Section 4 so as to determine the near-optimal set of weights EW for "Expert" level combination, i.e. we optimize weights w_j. This demonstrates the best performance that can be expected using the same query images and experts as the previous experiments if we impose a combination hierarchy at the "Expert" level.

- **Query Optimized**: This experiment is as for "Expert" optimized, except that the weight set we are optimizing is IE, i.e. weights w_i, and demonstrates the best performance that can be achieved if we combine at the "Query" level.

For each experiment we include the minimum and maximum achieved for that corpus. The minimum is a 'Uniform' run, where all pairs $\langle I_i, E_j \rangle$ are equally weighted, demonstrating the performance achieved if no weighting scheme at all is employed. The maximum is the fully optimized result as shown in Section 4, demonstrating the best performance that can be achieved. These two figures provide a lower and upper bound for data fusion performance comparisons, allowing us to make decisions using absolute observations with regard to the bounds, rather than relative observations by comparing only to existing data fusion approaches.

Our results are presented in Figure 5. Each table presents the minimum (Uniform All), maximum (All Optimized) and results of the 5 experiments using MAP, recall and P10. For every experiment's MAP, we show in brackets how close that approach came to achieving the optimal performance. The MAP of each of the experiments, along with the maximum MAP, is graphed in Figure 6. For each of our 5 runs we ran significance tests (partial randomization) with ρ 0.05. For the TRECVID benchmarks we found no significant difference between the 'Query' and 'Expert' levels of hierarchical combination, indicating that if hierarchical combination is employed and optimally weighted, there is no difference in between them. However for ImageCLEF 'Expert' was significantly different. For benchmarks TRECVID 2003-2006, all runs using highly weighted (1σ) pairs performed significantly better than the hierarchical combination approaches. For TRECVID 2007, only run 1σ was significantly different.

The graph presents a clear stratification of the results, particularly for benchmarks TRECVID 2003 - 2006. We can clearly see the very large discrepancy in performance between the hierarchical fusion approaches (at the bottom of the graph) versus the targeted weighting approaches in the middle. This separation illustrates the performance gains achievable by moving away from hierarchical combinations. Of exception is TRECVID 2007 and ImageCLEF 2007, where there is less of a difference in performance. These two benchmarks exhibit the greatest ratio of topic images to collection images – in the case of ImageCLEF one topic image for every 112 collection images. This indicates that recall plays a more prominent role in these evaluations, and that the selection of highly-weighted pairs may have been too restrictive to provide adequate topic coverage. This is reinforced by the run 1σU-T, which included all pairs $\langle I_i, E_j\rangle$: it performed the best even though it used non-specific weights.

The run 1σ highlights that, using a subset of pairs $\langle I_i, E_j\rangle$ from IEW, very good performance can be achieved despite a reduction in potential recall by not using all pairs. Far more encouraging is the performance of runs 1σU and 1σU-T. Whilst run 1σ had value as an illustrative run, it is hard to conceptualize a data fusion algorithm that would create the exact optimal weights for these pairs. However, as runs 1σU and 1σU-T did not use the optimal weights, but rather only identified what the high-performing pairs 1σU and 1σU-T were (essentially a binary weighting), and still achieved excellent performance, it provides a clear direction for development of data fusion algorithms. These runs demonstrate that if methods can be developed to identify pairs $\langle I_i, E_j\rangle$ that are likely to be highly weighted, then exact weighting is not required to obtain performance superior to that of methods which employ hierarchies.

6. CONCLUSIONS

In this paper we have demonstrated that the application of a data fusion hierarchy severely limits the performance that a CBMIR retrieval run can possibly achieve. We propose that rather than weighting combinations at the "Expert" or "Query" level, data fusion algorithms will achieve far greater performance by optimizing specific instances of an example query image and retrieval expert. Furthermore, we have demonstrated through our optimization process, that the ideal distribution of weights for data fusion in CBMIR is that of a log-normal distribution. Our observations are robust, as they occur within a fixed frame of reference, i.e. the lower and upper bounds achievable with data fusion, such that we determined the absolute effectiveness of particular approaches without having to make relative comparisons. Of practical concern however is how these observations may be interpreted to further aid CBMIR performance, and what methods may be used to weight at the direct level. Potential avenues for exploration involve looking for some form of correlation between documents which attract a large weight and content-analysis techniques such as entropy measures.

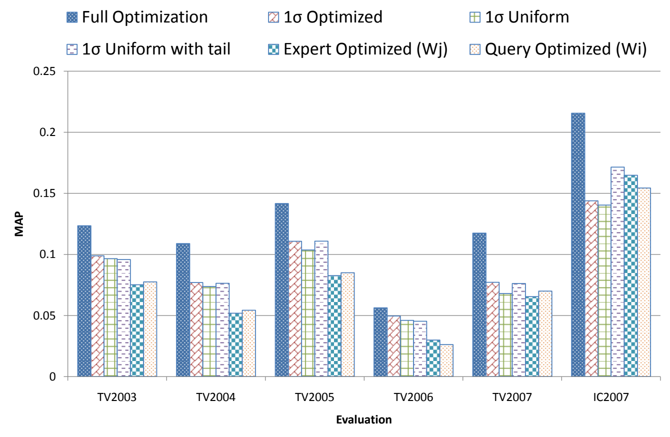

Figure 6: MAP Values for Data Fusion experiments

Acknowledgements

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

7. REFERENCES

[1] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, O. Frieder, and N. Goharian. Fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology*, 55(10):859–868, 2004.

[2] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. Effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, pages 339–346, Pittsburgh, PA, USA, 1993.

[3] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448, 1995.

[4] P. Clough, M. Grubinger, A. Hanbury, and H. Müller. Overview of the imageclef 2007 photographic retrieval task. In *Proceedings of the CLEF 2007 Workshop*, LNCS, Budapest, Hungary, 2008.

Legend	TRECVID 2003			TRECVID 2004			TRECVID 2005		
Run	**MAP**	**Recall**	**P10**	**MAP**	**Recall**	**P10**	**MAP**	**Recall**	**P10**
Uniform All	0.0593	0.2375	0.1080	0.0288	0.1440	0.1000	0.0646	0.1140	0.2333
Expert Level	0.0752 (61%)	0.2653	0.1960	0.0519 (47%)	0.1684	0.2000	0.0827 (59%)	0.1344	0.3625
Query Level	0.0776 (63%)	0.2729	0.2280	0.0543 (50%)	0.2018	0.1870	0.0850 (60%)	0.1456	0.3458
1σ Uniform	0.0966 (79%)	0.2786	0.2920	0.0738 (68%)	0.2268	0.2870	0.1037 (74%)	0.1484	0.4917
1σ Uniform & Tail	0.0958 (78%)	0.2828	0.2720	0.0764 (70%)	0.2251	0.2783	0.1109 (79%)	0.1513	0.4958
1σ	0.0989 (80%)	0.2805	0.3080	0.0770 (71%)	0.2246	0.3087	0.1108 (79%)	0.1574	0.5542
All Optimized	0.1224	0.3027	0.3760	0.1084	0.2318	0.3913	0.1407	0.1725	0.6583

Legend	TRECVID 2006			TRECVID 2007			ImageCLEF 2007		
Run	**MAP**	**Recall**	**P10**	**MAP**	**Recall**	**P10**	**MAP**	**Recall**	**P10**
Uniform All	0.0164	0.0926	0.0792	0.0422	0.2007	0.1417	0.1283	0.4095	0.3467
Expert Level	0.0299 (53%)	0.1138	0.2208	0.0655 (56%)	0.2500	0.2583	0.1648 (76%)	0.4133	0.4733
Query Level	0.0262 (47%)	0.1150	0.1625	0.0700 (60%)	0.2614	0.2750	0.1544 (71%)	0.4148	0.4450
1σ Uniform	0.0460 (82%)	0.1332	0.3583	0.0680 (58%)	0.2840	0.3458	0.1404 (65%)	0.3809	0.4150
1σ Uniform & Tail	0.0453 (80%)	0.1379	0.3417	0.0762 (65%)	0.2859	0.3583	0.1715 (80%)	0.4128	0.4567
1σ	0.0496 (88%)	0.1393	0.4167	0.0772 (66%)	0.2615	0.3750	0.1439 (68%)	0.3814	0.4450
All Optimized	0.0563	0.1493	0.4875	0.1175	0.3121	0.5083	0.2156	0.4379	0.5900

Figure 5: Experimental Results, all corpora

[5] W. B. Croft. Combining approaches to information retrieval. *Advances in Information Retrieval*, pages 1–36, 2000.

[6] P. Das-Gupta and J. Katzer. A study of the overlap among document representations. *SIGIR Forum*, 17(4):106–114, 1983.

[7] E. A. Fox and J. A. Shaw. Combination of Multiple Searches. In *Proceedings of the 3rd Text REtrieval Conference (TREC-2)*, Gaithersburg, MD, USA, 1994.

[8] L. S. Kennedy, A. P. Natsev, and S.-F. Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05)*, pages 882–891, Singapore, Singapore, 2005.

[9] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pages 267–276, Philadelphia, Pennsylvania, USA, 1997.

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

[11] M. McCabe, A. Chowdhury, D. Grossman, and O. Frieder. System fusion for improving performance in information retrieval systems. In *Proceedings of International Conference on Information Technology: Coding and Computing (ITCC 2001)*, Las Vegas, NV, USA, 2001.

[12] K. McDonald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Proceedings of the 4th ACM international Conference on Image and Video Retrieval (CIVR '05)*, Dublin, Ireland, 2005.

[13] M. McGill, M. Koll, and T. Noreault. An evaluation of factors affecting document ranking by information retrieval systems. Technical Report NSF-IST-78-10454 to the National Science Foundation (USA), Syracuse University, 1979.

[14] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[15] A. P. Natsev, M. R. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05)*, pages 598–607, Singapore, Singapore, 2005.

[16] T. Saracevic and P. Kantor. A study of information seeking and retrieving, iii: Searchers, searches, overlap. *Journal of the American Society for Information Science and Technology (JASIST)*, 39:177–196, 1988.

[17] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVid. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia information retrieval (MIR 2006)*, 2006.

[18] H. Turtle and W. Croft. Evaluation of an Inference Network-based Retrieval Model. *ACM Transactions on Informaion Systems*, 9(3):187–222, 1991.

[19] C. C. Vogt and G. W. Cottrell. Fusion Via a Linear Combination of Scores. *Information Retrieval*, 1(3):151–173, 1999.

[20] P. Wilkins. *An Investigation Into Weighted Data Fusion for Content-Based Multimedia Information Retrieval.* PhD thesis, Dublin City University, Glasnevin, Dublin, Ireland, September 2009.

[21] R. Yan and A. G. Hauptmann. The combination limit in multimedia retrieval. In *Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA '03)*, pages 339–342, Berkeley, CA, USA, 2003.

[22] R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 324–331, Seattle, Washington, USA, 2006.