TRECVid 2007 Experiments at Dublin City University

Peter Wilkins, Tomasz Adamek, Gareth J.F. Jones, Noel E. O'Connor,
and Alan F. Smeaton,
Centre for Digital Video Processing & Adaptive Information Cluster
Dublin City University, Glasnevin, Dublin 9, Ireland
Alan.Smeaton@dcu.ie

Abstract

In this paper we describe our retrieval system and experiments performed for the automatic search task in TRECVid 2007. We submitted the following six automatic runs:

- *F_A_1_DCU-TextOnly6*: Baseline run using only ASR/MT text features.
- *F_A_1_DCU-ImgBaseline4*: Baseline visual expert only run, no ASR/MT used. Made use of query-time generation of retrieval expert coefficients for fusion.
- *F_A_2_DCU-ImgOnlyEnt5*: Automatic generation of retrieval expert coefficients for fusion at index time.
- *F_A_2_DCU-imgOnlyEntHigh3*: Combination of coefficient generation which combined the coefficients generated by the query-time approach, and the index-time approach, with greater weight given to the index-time coefficient.
- *F_A_2_DCU-imgOnlyEntAuto2*: As above, except that greater weight is given to the query-time coefficient that was generated.
- *F_A_2_DCU-autoMixed1*: Query-time expert coefficient generation that used both visual and text experts.

1 Introduction

TRECVid is an annual benchmarking activity for information retrieval tasks on collections of digital video [8] and for TRECVid 2007 Dublin City University (DCU) participated in the automatic search task. We submitted a total of six fully automatic runs. In our participation in TRECVid in 2006, we examined the use of semantic concepts in the automatic search process[3], this year however our automatic submission had made use of only low-level visual features and ASR/MT text. The emphasis of our submission this year was to further investigate methods for query-time coefficient generation for retrieval expert combination.

The rest of the paper is organized as follows. The retrieval experts used for these experiments, and the reference set of images used are described in Section 2. The techniques for query-time coefficient generation are described in Section 3, whilst experimental results are shown in Section 4. Preliminary conclusions are then presented in Section 5.

2 Retrieval Experts

A retrieval expert is what we refer to as some form of index which has an associated ranking function which allows that index to be queried and to return a ranked result set. Broadly speaking, we had two types of retrieval expert used for our automatic search experiments in TRECVid, a set of global visual features, and the ASR/MT text donated to TRECVid [2].

For our global visual features we needed a set of candidate keyframes to process. As no common keyframe set was released as part of the TRECVid 2007 collection, we extracted our own set of keyframes. Our keyframe selection strategy was to extract every second I-Frame from each shot. This gives us far more keyframes than the usual one-keyframe-per-shot which has been the norm in previous TRECVids and in fact gives us about 1 keyframe per second of video. For the remainder of this paper, we will refer to these images as K-Frames.

We extracted global low-level visual features from K-frames using several feature descriptors based on the MPEG-7 XM. These descriptors were implemented as part of the aceToolbox, a toolbox of low-level audio and visual analysis tools developed as part of our participation in the EU aceMedia project [1]. We made use of six different global visual descriptors namely Colour Layout, Colour Moments, Colour Structure, Homogenous Texture, Edge Histogram and Scalable Colour. A complete description of each of these descriptors can be found in [5].

Our text data was the ASR/MT text donated to TRECVid by the University of Twente [2]. This text was temporally arranged, so was aligned with the shot boundaries to create discrete text documents. This alignment was performed by ITI as part of their involvement in the K-Space TRECVid participation. The text was then indexed by Terrier [6], with retrieval results provided through a vector space model [7]. We did not do any advanced text processing such as story bound segmentation, or make use of Dutch text queries.

3 Result Fusion

Our automatic retrieval system for experiments this year is an extension of the system built for TRECVid 2006 [3] and makes use of our knowledge gained in query-time coefficient generation for retrieval experts [10].

One of the major differences between our 2006 and 2007 submission is the order in which we fused various experts. Figure 1 illustrates our 2006 methodology. In this setup, we fused experts according to some pre-defined semantics. For instance, in the illustrated example, our query consists of two example query images and a text query. We have available a colour expert and an edge expert as well as the text expert. Last year we first fused together the results from each query example image into a single result list for that image. We then fused the result lists for each image into a single image result list. Finally this was merged with the text expert results. At each aggregation step we applied our coefficient generation techniques, such that each result set that was being combined had some weight associated with it. Using the 2006 example we had three main aggregation steps, once for single image results, once for the merged single image results into a single all image result, and finally when combined with the text.

This approach was modified significantly in 2007. Rather than having a series of aggregations, we instead treat each expert equally. Figure 2 illustrates the new approach. As per

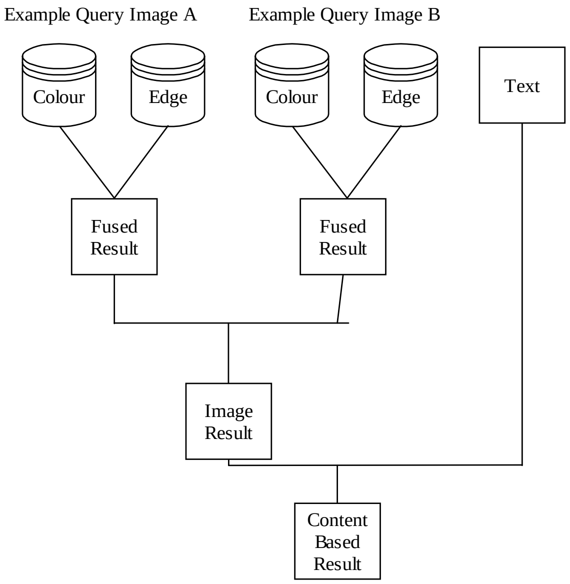

Figure 1: 2006 Fusion Framework

the 2006 example, we have two example images, with two visual experts and a text expert. Instead of a multi-step aggregation, we now have a single aggregation step where all experts are combined at the same time, meaning in the example we combine two colour expert results, two edge expert results and a text expert result. Our techniques for coefficient generation are able to handle either the 2006 or 2007 approaches, as our technique can combine any arbitrary number of result sets into a single set, providing some weighting coefficient to each.

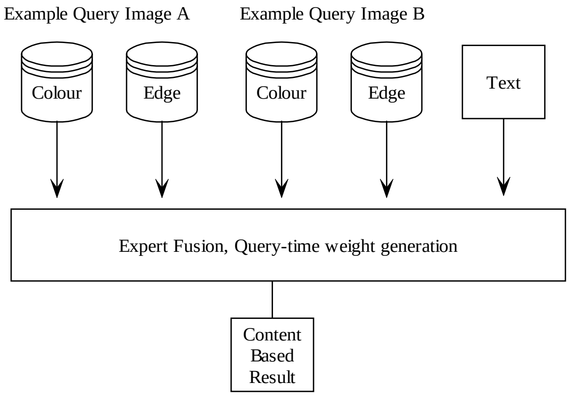

Figure 2: 2007 Fusion Framework

3.1 Coefficient generation

For the fusion of multiple sources of information we require weights/coefficients. The weights we employed for this task were dynamically generated at query-time and reflect the degree to which we believe one source of information will provide better performance as opposed to the other sources we have. This process is described in [10] for retrieval. We have also used this approach for the combination of classifier outputs, described in [9]. We will now provide a brief overview of this approach.

Our dynamic weighting function, which given a set of retrieval expert result sets is designed to infer relative performance of these experts (e.g. expert A will perform better than expert B) and weight accordingly. It is not design to infer absolute performance of a expert (e.g. expert A will achieve average precision for this topic greater than 0.5).

The central thesis of the our approach detailed here is that by examining the distributions of the scores generated from a retrieval expert for a query, that it is possible to infer relative performance of one expert against another.

We have previously observed a correlation in the search domain where if a feature undergoes a rapid change in its normalized scores, then that feature is likely to perform better than a feature which undergoes a more gradual transition in normalized scores. Examples of this can be found in [10] and in the classification domain in [9].

Our hypothesis that the reason this correlation exists is that the rapid initial change in score of an expert can be seen as an indicator of 'interesting-ness' or confidence that the expert has in it's initial rankings. That is that the expert has made definite decisions about the ranking by having greater distances between the scores. Conversely, an expert which exhibits a more gradual change in the initial scores could be thought of as having found many results that are very similar, and as such has not been able to differentiate these to a large degree from the set.

This thinking is derived from observations made by Lee [4] where he states that fusion appears to work because "different runs might retrieve similar sets of relevant documents but retrieve different sets of non-relevant documents". We note that the observed correlations are not universal and there are instances in which the correlation does not hold. The investigation into the causes of this correlation are an open research question which we are currently pursuing.

If we assume that this correlation exists in the TRECVid 2007 corpus we can leverage it to generate query-time expert coefficients. In order to combine these sources of evidence, we first normalize the scores using MinMax normalization (Equation 1).

$$Norm_{score(x)} = \frac{Score_x - Score_{min}}{Score_{max} - Score_{min}} \tag{1}$$

Next we calculate the average change in score for a given set size of a feature, which we refer to as the Mean Average Distance (MAD) 2.

$$MAD = \frac{\sum_{n=1}^{N}(score(n) - score(n+1))}{N-1} \tag{2}$$

A direct comparison of this average score distance between features would not necessarily work as it may not account for differences in scoring metrics that are used, or by the natural distribution of a feature amongst its scores. In order to make a cross-comparable feature value we computed a ratio of MAD of a top subset of that feature, versus a larger set of that feature. In this series of experiments we used examined the top subset size of 5% of the feature over

95% of the feature. The value this produces we refer to as a Similarity Cluster (SC) value. This is formally defined in (3).

$$SC = \frac{MAD(subset)}{MAD(largerset)} \tag{3}$$

Once we have this value for a given feature, we can then use it to generate a weight for that feature, as shown by 4.

$$Feature\ Weight = \frac{Feature\ SC\ Score}{\Sigma All\ SC\ Scores} \tag{4}$$

This formula will generate larger weights for features which exhibit larger values of SC, in turn meaning that these features underwent the greater initial score change and according to our observations these features are likely to be the better performing features and should be weighted accordingly.

The approach just described was used for our runs *F_A_1_DCU-ImgBaseline4* and *F_A_2_DCU-autoMixed1*. Our image only baseline made use of just global visual features, whilst our mixed submission also included the text expert results.

We produced a variation on this approach which sought to approximate the the query-time coefficient generation approach at indexing time, which with the exception of our text only run, was used in our remaining submissions.

3.2 Index-Time Coefficient Generation for Single Images

If we continue with our assumptions defined in the previous section, that a rapid change in an expert's initial ranked scores indicates that the expert is more confident in its selections, we can approximate this at indexing time.

To achieve this, for each visual expert we first calculate the average image. For every image in the corpus, we then calculate the distance (using that expert's distance metric) from the average image. We are left then with a score for every image in the corpus which indicates its distance to the average image. The Z-Score (gaussian normalization) is then calculated for each distance score. We next take the absolute values of the Z-Scores, sort these from highest to lowest, then transform the values to the range [0..1].

The motivation for the Z-Score normalization, is that images with a high Z-Score are those which appear at the tails of a Gaussian distribution. This means that these images have a greater distance from the average document, and therefore these images can do a better job at producing ranked lists which exhibit a greater change in initial score, than those images which have a lower Z-Score. Images with a lower Z-Score will have many images which share a similar distance to the average document, and thus a query with these images will more likely produce a result set which displays a gradual change in score. This thinking is predicated on the assumption that a rapid change in score from a retrieval expert is more likely to perform better than an expert which has a gradual decline.

The remaining three runs we submitted make some use of the aforementioned method.

- *F_A_2_DCU-ImgOnlyEnt5*: This run makes use only of the index time scores generated.

System	MAP
Colour Layout	0.0109
Colour Moments	0.0173
Colour Structure	0.0065
Edge Histogram	0.0192
Homogenous Texture	0.0179
Scalable Colour	0.0050
Text (ASR/MT)	0.0023

Table 1: 2007 Single Expert Search Results

- *F_A_2_DCU-imgOnlyEntHigh3*: In this run, we combined the coefficients generated by our query-time techniques and our index time techniques. The motivation for this combination is that if our query-time technique produces a coefficient that is off-base, then it can be dampened by being combined with the index time coefficient which is calculated off the average image. For this particular run, we increased the value of the index-time coefficient by 50% when combining the two sets of weights, thus giving greater importance to the index-time coefficient.
- *F_A_2_DCU-imgOnlyEntAuto2*: Similar to the previous run, except that instead of increasing the value of the index-time coefficient by 50%, we decrease it's value by 50%, thereby giving greater importance to the query-time generated coefficient.

4 Results

Here we will highlight the results we achieved in our 2007 experiments. As the major objective of our work was the investigation of automatic coefficient generation techniques which were leveraged off the distribution of scores from retrieval experts, we will first present the results that each individual expert obtained in our results.

We can see from Table 1 that our best performing single feature was Edge Histogram, with a MAP of 0.0192. Text was our worst performing feature, (which is also run *F_A_1_DCU-TextOnly6*) which is not surprising given that we performed no advanced text retrieval techniques such as story bound detection for this expert.

Our submitted run results can be observed in Table 2.

We can make preliminary observations about these results. The first is that all of our fusion attempts produced large improvements over any single retrieval expert that was used. Secondly, our image-only run, which used our standard query-time coefficient generation technique was our best performer, further emphasizing the applicability of this approach to different collections.

System	MAP
F_A_2_DCU-autoMixed1	0.0289
F_A_2_DCU-imgOnlyEntAuto2	0.0292
F_A_2_DCU-imgOnlyEntHigh3	0.0373
F_A_1_DCU-ImgBaseline4	0.0420
F_A_2_DCU-ImgOnlyEnt5	0.0369
F_A_1_DCU-TextOnly6	0.0023

Table 2: 2007 Fused Expert Search Results

We had mixed results with our index-time fusion approaches, with both the index only coefficient run, and index-time weighted high run achieving good performance.

All our runs were image only expert runs with the exception of our mixed modality run *F_A_2_DCU-autoMixed1*. The relative low performance of this run requires further investigation. However we have noted that several participants achieved much stronger text only performance than we did, and we would be eager to investigate the use of those text experts with our visual experts.

5 Conclusions

In this paper we have presented the DCU results for automatic search for TRECVid in 2007. Our aim was to further investigate our approaches for coefficient generation for retrieval expert combination. Our runs demonstrated that the approaches we have used so far are successful in combining multiple retrieval experts and achieving performance which greatly exceeds that of any single expert. Future work now includes a greater investigation into why our techniques are working, and an examination of the effect that various collections bring to retrieval expert performance.

Acknowledgements

This work was partially supported by Science Foundation Ireland under grant 03/IN.3/I361 and by the European Commission under contract FP6-027026 (K-Space). We are grateful to the AceMedia project (FP6-001765) which provided us with output from the AceToolbox image analysis tooklit.

References

[1] The AceMedia Project, available at http://www.acemedia.org.

[2] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the second international conference on Semantics And digital Media Technologies (SAMT)*, Lecture Notes in Computer Science, Berlin, December 2007. Springer Verlag.

[3] M. Koskela, P. Wilkins, T. Adamek, A. F. Smeaton, and N. E. O'Connor. TRECVid 2006 Experiments at Dublin City University. In *TRECVid 2006 – Text REtrieval Conference, TRECVID Workshop, Gaithersburg, Md., 13-14 November 2006*, 2006.

[4] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[5] N. O'Connor, E. Cooke, H. le Borgne, M. Blighe, and T. Adamek. The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

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

[7] G. Salton. *Automatic Text Processing.* Addison–Wesley, 1989.

[8] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[9] P. Wilkins, T. Adamek, N. O'Connor, and A. F. Smeaton. Inexpensive fusion methods for enhancing feature detection. *Signal Processing: Image Communication, Special Issue on Content-Based Multimedia Indexing and Retrieval*, 22(7-8):635–650, 2007.

[10] P. Wilkins, P. Ferguson, and A. F. Smeaton. Using score distributions for query-time fusion in multimedia retrieval. In *MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006.