Diversity, assortment, dissimilarity, variety: A Study of Diversity Measures using Low Level Features for Video Retrieval

Martin Halvey, P. Punitha, David Hannah, Robert Villa, Frank Hopfgartner, Anuj Goyal and Joemon M. Jose.

Department of Computing Science, Sir Alwyn Williams Building, 18 Lilybank Gardens, University of Glasgow, Glasgow, Scotland, G12 8QQ
{halvey, punitha, davidh, villar, hopfgarf, anuj, jj}@dcs.gla.ac.uk

Abstract. In this paper we present a number of methods for re-ranking video search results to introduce diversity into the set of search results. The usefulness of these approaches are evaluated in comparison with similarity based measures for the TRECVID 2007 collection and tasks. For the MAP of the search results we find that some of our approaches perform as well as similarity based methods. We also find that some of these results can improve the P@N values for some of the lower N values. The most successful of these approaches was then implemented in an interactive search system for the TRECVID 2008 interactive search tasks. The responses from the users indicate that they find the more diverse search results extremely useful.

1 Introduction

As a result of the improving capabilities and decreasing prices of current hardware systems, there are ever growing possibilities to store and manipulate images and video in a digital format. In addition to this, it is now feasible to view images and video online at home. Individuals now build their own digital libraries from materials created through digital cameras and camcorders, and use a number of systems to place this material on the web, as well as store them in their own personal collections. However the systems that currently exist to organise and retrieve these videos are not sufficient to deal with such large and rapidly growing volumes of video. There is an ever growing need to develop new methods and techniques to search for and retrieve relevant material in these vast oceans of data. Current state of the art systems rely on using annotations provided by users or on methods that use the low level features available in the videos. Neither of these methods is currently sufficient enough to overcome the problems associated with video search. Query by text is hindered by the lack of textual descriptions and annotations online [4] and in other collections [5]. Query by example suffers because of the difference between the low-level data representation of videos and the higher level concepts users associate with video, commonly known as the semantic gap [12]. Bridging the semantic gap is one of the most challenging research issues in multimedia information retrieval today. Current retrieval methods based on similarity or dissimilarity measures return a large number of very similar videos or images. Diversity ranked result lists will present the users with a wider range of options in their search results by presenting a diverse set of results that embody many possible interpretations of the users query.

In this paper we introduce a number of methods to promote diversity in video search results in an attempt to bridge this semantic gap. The intention is to also overcome some of the problems caused by the limited expressiveness afforded by query by example or query by text. In addition, by using low level features such as colour, edge histogram etc.; we are not restricted by the insufficient annotations that are attached to many images and videos [4]. The remainder of this paper is organised as follows: We will provide some related work in the following section. In Section 3 we will describe four novel approaches for providing diverse video search results. Section 4 will describe two evaluations of these approaches. The first looks at the effect of diversity measures on retrieved results and the second investigates some user interactions with diverse result lists. Finally we will provide a discussion of our work and conclusions.

2 Related Work

Some initial work has been carried out in the area of diversification in text retrieval. Zhang et al. [16] diversify search results in the context of Web search. They propose a novel ranking scheme named Affinity Ranking which re-ranks search results by what they call diversity and information richness. The TREC Novelty Track [14] aimed to encourage research in finding novel sentences in a set of relevant sentences. However, there have been very few studies of diversity measures for image and video search. Song et al. [15] acknowledge the need for diversity in search results for image retrieval. They propose a re-ranking method based on topic richness analysis to enrich topic coverage in retrieval results, while maintaining acceptable retrieval performance. More recently van Zwol et al. [18] propose a diversification model for image search results that is based on annotations associated with an image. The contribution of this work is two-fold. Firstly the diversity is a result of the retrieval model and not a post retrieval step. Secondly, they balance precision and diversity by estimating the query model from the distribution of tags which favours the dominant sense of the query. While this approach is shown to be useful, it suffers from the lack of annotations which is common for multimedia is shared online [4]. Although not on diversification, there has been some work carried out looking at the role of dissimilarity in image retrieval. [5, 8] look at a number of different methods for calculating dissimilarity. They evaluate the performance for a number of different measures for a number of different feature spaces for the Corel collection. Based on these evaluations, they identify a number of the best dissimilarity measures.

In the area of recommender systems diversity measures have been studied for many years. In the area of collaborative filtering, Karypis [8] evaluated a number of methods that promote serendipity and novelty. With respect to case based reasoning a number of different methods have been proposed to introduce diversity into a set of search results. Smyth and McClave [13] propose a bounded greedy algorithm to improve diversity while still maintaining target similarity [1]. This strategy incrementally builds up a set of diverse cases from a bounded set of cases similar to the target query. During each step the cases are re-ordered according to a "quality measure" which combines diversity and similarity. McSherry introduced the concept of similarity layers to promote diversity [10]. A set of cases ranked according to their similarity to the target query are partitioned into layers, such that all cases in a given layer have the same similarity to the target query. The retrieved cases are then chosen starting with the highest similarity cases continuing until k cases have been chosen. This approach introduces some diversity while maintaining high levels of similarity. Order based retrieval is an example of a method that is not designed to promote diversity but which does so by the nature of the method [2]. Order based retrieval constructs an ordering relation from the query and applies this relation to the case base of products returning the k items from the top of the ordering. The order relation is constructed from the composition of a set of canonical operators for constructing partial orders based on the feature types that make up the query. An empirical evaluation of the approach demonstrates the ability to enhance the diversity within the search results. Zeigler et al. [17] use topic diversification to balance and diversify personalised recommendation lists in order to reflect the user's complete spectrum of interests. This approach reduces average accuracy; however it improves user satisfaction with the system.

3 Diversity Re-Ranking Using Low Level Features

Random

Perhaps the simplest way to introduce diversity into a set of search results is to re-order those search results randomly. This reduces the chance of images that are visually similar being adjacent in the result set and is computationally simple. The problem with this approach is that it may result in an unacceptable drop in similarity or closeness to the original query.

Dissimilarity

Similar to work that has been carried out for case based reasoning [13] dissimilarity can also be used to introduce diversity into a set of image search results. For this approach we consider the diversity $D(I_n)$ within a set of search results, I_n, to be the average dissimilarity between the entire search results (see Equation 1).

$$D\ (I_n) = \frac{\sum_{k=1}^{n}\sum_{j=k}^{n}(1-Similarity(I_k,I_j))}{\frac{n}{2}(n-1)} \qquad (1)$$

Keeping Equation 1 in mind we use a bounded greedy algorithm [1] which allows us to improve the diversity in the result set while still maintaining the similarity to the original query. One of the main benefits of this approach is that we do not have to worry about how similarity is calculated. All that is needed is a similarity value, thus whether query by text or query by example is used we can still calculate a re-ranking. This strategy incrementally builds a new and more diverse set of results from the existing result set. During each step the images are re-ranked according to their "quality", with the highest "quality" image being added to the new result list R (see Equation 2). The Greedy Selection algorithm relies on a quality metric (see Equation 2) which combines the similarity between the query *t* and each image in the results set, with the dissimilarity between images in the result set with the re-ranked result set (see equation 3). The first image in the re-ranked result list is always the same as the original result list. For each subsequent iteration the image selected is the one with the highest combination of similarity to the original query and diversity relative to the re-ranked result list.

```
Greedy Selection
Define GreedySelection(t,C,k)
begin
    R = {}
    For i= 1 to k
            Sort C by quality Q(t,z,R) for each z in C
            R = R + First(C)
            C = C – First(C)
    End For
return R
end
```

Figure 1: Greedy selection algorithm for re-ranking a set of retrieved images, to improve diversity. C refers to the original result list, R is the re-ranked result list, z refers to an image in C.

$$Q(t,z,R) = Similarity(t,z)*RelDiversity(z,R) \qquad (2)$$

$$RelDiversity(z,R) = 0 \text{ if } R = \{\} \qquad (3)$$
$$= \frac{\sum_{i=1}^{m}(1 - Similarity(z,r_i))}{m}, \text{ otherwise}$$

There are alternative quality measures that can be used. Equation 4 defines quality as the harmonic mean of the similarity and relative diversity values. A weighted harmonic mean is related to the F-measure value in information retrieval. In the following section where we evaluate the different approaches to introducing diversity into our result list we use two versions of the greedy algorithm, one with Equation 2 and one with equation 4 for calculating the "quality"

$$Q(t,z,R) = 2/\left(\frac{1}{Similarity(t,z)} + \frac{1}{RelDiversity(z,R)}\right) \quad (4)$$

Clustering

It has been argued that since result lists are ranked using some sort of similarity or dissimilarity measure that the images which resemble each other are grouped together and thus it should be easy to find this discrimination in the result list. However, many distance measures are non-metric, i.e. they do not fulfil the four properties of positivity, reflexivity, symmetry and triangle inequality, it is not the case that discrimination is easy [6]. Thus clustering based on low level features offers an obvious method to introduce diversity into a set of search results. Clustering allows similar images to be grouped together and then the result list can be re-organised accordingly to introduce more diversity into the result set. For our approach to achieve this outcome, an affinity matrix (also called a weight adjacency matrix) for the result list of images is computed using the appropriate similarity measure (detailed in the experimental section). We use single link for clustering as our goal is to find possible seed points from distinct videos in a sequential manner. Were we to use complete link it would connect images to already existing groups which we are not interested in. Also we do not predetermine a number of clusters. Instead the clustering stops when all points belong to a cluster and each cluster contains at least 10 videos. For single linkage clustering we use Kruskal's algorithm, which is a graph theoretic greedy approach and is generally employed to find the minimum spanning tree. From a graph theoretic point of view, n (the number of clusters) represents the number of spanning trees expected from the graph represented by the affinity matrix. Given this Affinity matrix A, representing the graph G(V, E) where V is the set of images and E a set of similarity weighted edges, between the images, we form $\{C_1, \mathrm{C}_2, \mathrm{C}_3, \ldots \mathrm{C}_\mathrm{n}\}$, such that $\bigcap\left(C_{i,} C_j\right) = \emptyset \; for \; i \neq j$ and $\bigcup_{i=1}^{n} C_i = V$. Each cluster C is then re-sorted based on the similarity of the representative image of the clusters. The representative image of a cluster is the image that is most similar to the members of the cluster, i.e. given a graph representation of images, a weight adjacency matrix, for the set of images making up the clusters, we pick the image of C_i such that $argmax_{j \in C_i}$ $\sum_{k \in C_i} A_{jk}$ as the representative image. To re-order the result list using these clusters, we pick one representative image from each cluster in sequence until all of the clusters have been exhausted, thus picking images from diverse groups, which in theory should propagate diversity though out the result list.

4 Experimental Evaluation

In order to evaluate the usefulness of diversity based measures for video retrieval, we carried out a number of evaluations. The first looked at the effect of diversity on automatically retrieved results. While it is important to introduce diversity in a result list, it is equally important that there is not a significant drop in the precision of the returned results. If our methods do not result in a loss in precision is it possible for diversity to boost the precision of the returned results. The second part of our evaluation concentrates on user reactions to our diversified video search results. We wish to discover if users find any difference between results ranked on similarity and those ranked based on diversity. If users do notice a difference, we want to determine whether user react positively or negatively to the diverse video search results.

4.1 Diversity for Automatic retrieval

Procedure

The evaluation of the effect of diversity measures was conducted using the TRECVID 2007 dataset. This corpus consists of 200 hours of television news videos, which is divided amongst 109 videos. These videos are segmented into short snippets or shots, the boundaries between shots are provided as part of the collection. From each shot we extract a representative keyframe; this resulted in a final dataset of 18131 keyframes. The TRECVID evaluation for 2007 had 24 topics. For each topic we were supplied with a number of example video shots and images, we used these examples as queries in order to retrieve search results. We used a number of visual features which had been extracted from each video shot and each example to calculate the similarity and dissimilarity measures. These features were Colour Histogram (CHD), Edge Histogram (EHD) and Homogenous Texture (HTD). For each feature we adopt two different fusion techniques for ranking the results. The first fusion technique is a simple similarity based measure, which we will refer to as EHD for edge histogram, CHD for colour histogram and HTD for homogenous texture in the following sections. The second fusion technique takes into account the significance of results derived for each image query ensuring that the results of all queries are equally important. To achieve this we employ a fixed threshold method to select similar target images for each query, this approach is similar to the fixed radius method in Chen et al [3]. Again for this fusion technique we will have three values which we will refer to as EHRR for edge histogram, CHRR for colour histogram and HTRR for homogenous texture in the following sections. Three different feature specific distance measures were employed depending on the low level features, Equation 5 was used for CHD and CHRR, Equation 6 was used for EHD and EHRR and Equation 7 was used for HTD and HTRR. Equation 6 is the most different of the measures. For each edge based vector there are 80 values, these 80 values are used to compute, five more global features, with 65 more values as the semi global features.

$$D^{CHD}(A,b) = \sum_{i=0}^{31} \left(\sqrt{CHD_A(i)} - \sqrt{CHD_B(i)}\right)^2 \qquad (5)$$

$$D^{EH}(A,B) = \sum_{i=0}^{79} |EHD_A(i) - EHD_B(i)| + 5\sum_{i=0}^{4} |EHD_A^g(i) - EHD_B^g(i)| + \sum_{i=0}^{64} |EHD_A^s(i) - EHD_B^s(i)| \qquad (6)$$

$$D^{HTD}(A,B) = \sum_{k=0}^{61} HTD_A(k) - HTD_B(k) \quad (7)$$

The similarity values based on the distance values obtained by the above equations are computed using a simple algebraic equation

$$im = e^{-(D(A,B))} \quad (8)$$

For each topic and using each visual feature we retrieved 1000 results based on similarity which form our baseline, each of these result list was then re-ranked using clustering (Clustering), the Greedy Algorithm with the product quality measure (greedy), the Greedy Algorithm with the f-measure quality measure (F-measure) and also using the random algorithm (Random).

Results

Since we were using the TRECVID 2007 collection and tasks, we were able to calculate precision and recall values for all of the tasks. As an example of the difficulty in TRECVID 2007, the best performing automatic run returned just over 2 relevant shots for each top 10 result set [11]. Figure 2 shows the mean average precision (MAP) for each approach and feature. It can be seen from Figure 2 that in all cases the similarity based measure has the highest MAP. The greedy algorithm using the product to calculate the quality is the next best in all but one of the cases, the difference between the two is negligible and not statistically significant (t=1.73 p=0.1438 for a t-test). However, looking at the average does not give an entirely true reflection of the performance as the results are highly variable depending on the tasks. The results were also analysed at an individual topic level for each algorithm and feature combination. It was found that for topics with an already high MAP that re-ranking the search results meant that the MAP was lower for the re-ranked results. However, where there was a low MAP for the similarity based measures, the re-ranking resulted in a higher MAP. In fact, across the 144 topics and feature combinations (24 topic and 6 features) the similarity based results had the highest MAP only 26 times, the Greedy Algorithm had the highest MAP 54 times (26 times for the product quality measure and 29 for the F-measure quality measure). In conclusion in scenarios where similarity ranking results in high MAP, diversity re-ranking can degrade the MAP slightly, but not in a significant way. In scenarios where the similarity measures result in a low MAP, the diversity based re-ranking can improve the MAP as well as potentially giving the user more options in their result list.

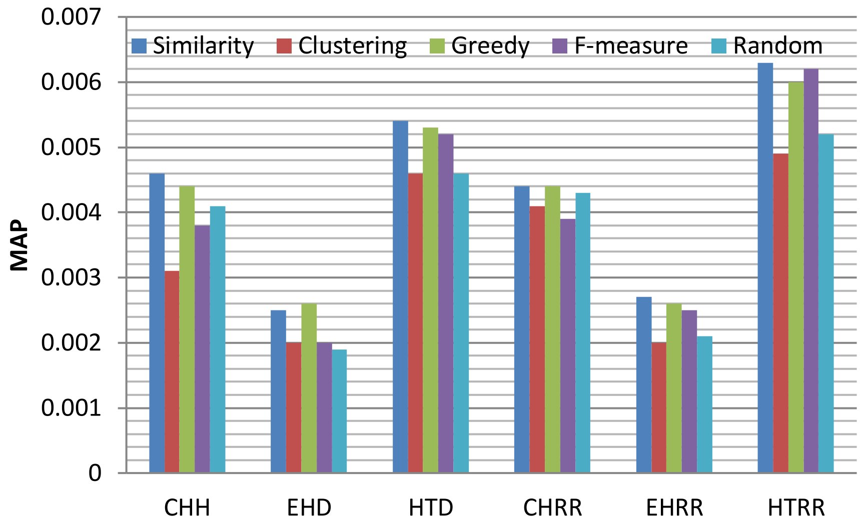

Figure 2: MAP for each algorithm and feature combination. MAP was calculated for each re-ranked result list.

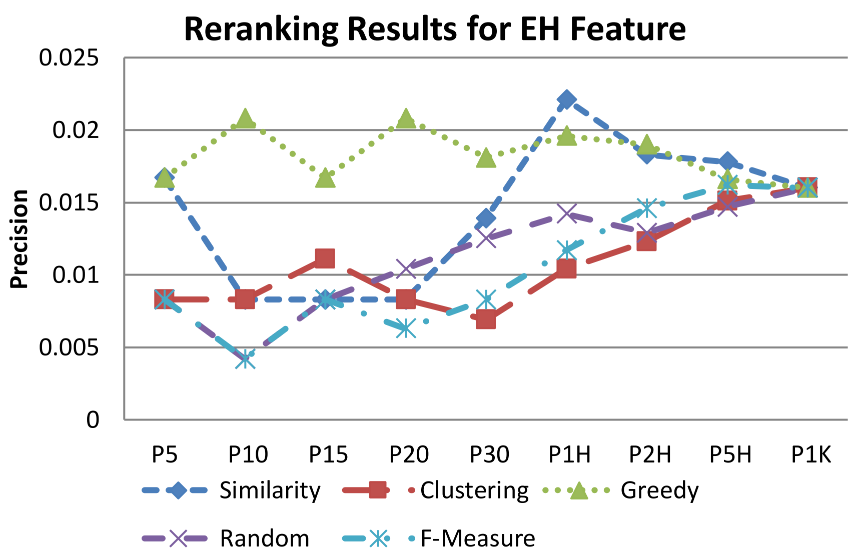

Figure 3: Average P@N values for all of the algorithms for the Edge Histogram Feature

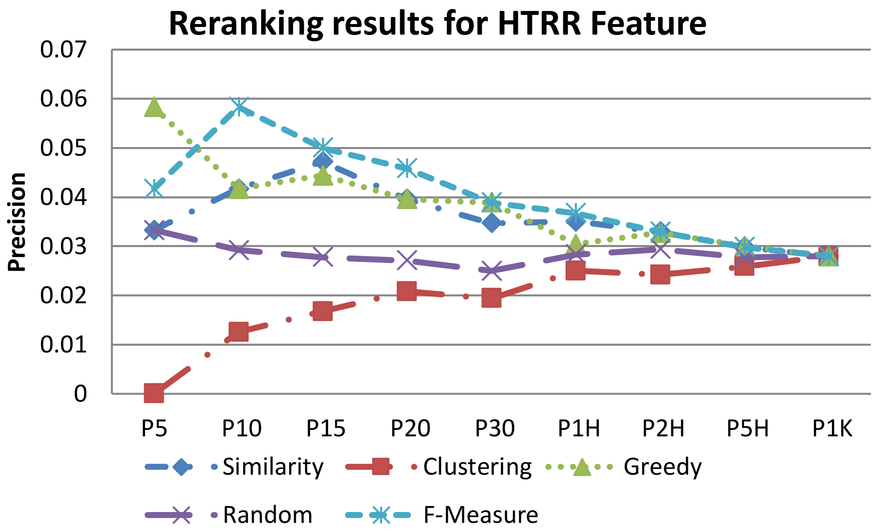

Figure 4: Average P@N values for all of the algorithms for the Homogenous Texture RR Feature

Figures 3 and 4 show the P@N for each feature for the edge histogram and homogenous texture features respectively for varying values of N. P@N is the ratio between the number of relevant documents in the first N retrieved documents and N. The P@N value focuses on the quality of the top results, with a lower consideration on the quality of the recall of the system. This gives us a more focused idea of the top ranked results. The results for P@N for 4 of the 6 features show that the similarity baseline has the highest precision values at the lower values of N. However for edge histogram (Figure 3) and homogenous texture (Figure 4) it was found that the Greedy Algorithm outperforms the similarity ranked results. In the case of EHD, CHRR and HTRR it was found that the differences between Similarity and Greedy were not statistically significant, the same was found for EHD and Cluster and EHRR and F-Measure. As with the MAP, it was found that many of the differences were topic dependent.

The results were also analysed at an individual topic level for each algorithm and feature combination. It was found that for topics with an already high P@N that re-ranking the results resulted in the P@N being lower for the re-ranked results. However where there was a low P@N for the similarity based measures, the re-ranking resulted in a higher P@N. What was especially interesting was that for topics where the P@N was equal or close to zero, the random algorithm was one of the top performers as it did not depend on the similarity values like the other algorithms. In order to identify potential trends in the data we selected P@20 as representative; we feel this is reasonable as users can only examine a small number of results and 20 is small enough to assume that users may examine up to 20 results, but large enough to allow the different re-ranking algorithms to have an effect. Across the 144 topics and feature combinations it was found in the majority of cases that the re-ranking resulted in a similar or identical P@20 values as the similarity based measures (86 times), the similarity based measures had higher P@20 only values 17 times. The greedy algorithm performed the best with it resulting in the highest P@20 values 24 times (15 for product based quality, 9 for F-Measure based quality).

In conclusion it has been shown that the diversity based re-ranking of the results can result in comparable precision and recall to similarity based measures. In some cases the dissimilarity results in superior results in comparison with the similarity based measures. In the following section we will outline a user evaluation to gauge user reaction to diversity based results.

4.2 Diversity for Interactive Retrieval

Procedure

In order to evaluate the potential usefulness of the diversity based re-ranking of video search results we carried out a user centric evaluation. This evaluation was carried out as part of our TRECVID 2008 interactive search run. Each user carried out 12 of the 24 search topics using 3 different interfaces, 4 topics per interface. Order of topic and system was rotated to reduce any learning affects that might occur. Each interface presented different results, the first presented as a grid of keyframes retrieved based on similarity (Grid), the second presented as a grid of keyframes retrieved based on similarity and re-ranked using the greedy algorithm with the product quality measure (GridRR), and the final method presented an entire video starting at the relevant keyframe where relevance is judged based on similarity (Video). In post search task questionnaires we solicited subjects' opinions on the videos that were returned in the result list by the system. . The following 7-point Likert scales were used. The scales were, "The search results were" "Predictable/Unpredictable", "Clear/Confusing", "Narrow/Broad", "Relevant/Not Relevant", "Inappropriate/Appropriate", "Complete/Incomplete" and "Surprising/Expected". The most positive response for each system is shown in bold.

Results

Differential	Grid	GridRR	Video
Predictable	3.9529	3.7812	**4.1974**
Clear	**4.2978**	4.287	4.0497
Broad	4.2554	3.8448	**4.3408**
Relevant	4.1126	4.0354	**4.254**
Appropriate	4.1758	**4.2486**	3.8143
Complete	3.8102	**3.9973**	3.697
Expected	**4.2549**	4.2199	3.9194

Table 1: User reaction to the search results (Higher=Better)

From the results in Table 1 it appears that participants have a mixed reaction to the re-ranked based results. We applied two-way analysis of variance (ANOVA) to each differential across all 3 systems and the 24 topics to test the significance of these results. It was found that the differences in how broad (F=4.71, p=0.0096) and appropriate (F=3.36, p=0.0359) the users found the results was system dependent. When examining the other results the trend is that the users found the results ranked based on diversity to be the most appropriate and complete. It can be seen that the other two systems produced results which users expected and predicted; this result can also be viewed as encouraging as we are presenting new and unexpected results to the users of the system that presents diversity re-ranked results. However, the users of the three systems found the results of the Video system to be the broadest and most relevant. This is hardly surprising as this system presents entire videos as results thus showing large portions of the collection. Overall these results demonstrate that the users feel that they have seen the most appropriate and complete results when viewing re-ranked results, and in addition they are seeing results that they had not considered.

5 Conclusions

In this paper we have presented a number of novel methods for promoting diversity in a set of video search results. These methods were evaluated with respect to their impact on search results for the TRECVID 2007 collection and tasks. We measured MAP and P@N values for similarity based retrieval in comparison with our diversity re-ranking methods. There are a number of conclusions that can be drawn from this evaluation. It was found that our average dissimilarity approach does not reduce the MAP of the results, in contrast with the random and clustering methods. In fact the P@N values are improved for two of the six features that we used for the average dissimilarity approach. Another interesting finding was that when P@N values are low that random diversity re-ranking can improve precision values, as this approach does not rely on using similarity values to calculate a ranking. Following on from this we implemented an average dissimilarity re-ranking approach in a system for a TRECVID 2008 interactive search run. We elicited user responses to the search results presented by three different interfaces. It was found that the users feel that the diversity ranked results presented the most appropriate and complete results. In addition, they feel that when viewing diversity re-ranked results that they see results which they would not have considered otherwise. In conclusion, our approaches for introducing diversity into video search results has highlighted the promise of this approach in alleviating the major problems that users have while searching for multimedia, thus presenting an important tool which provides a work around to the semantic gap [12] and other problems associated with video search.

Acknowledgements

This research work was partially supported by the EC under contracts: K-Space (FP6-027026), SEMEDIA (FP6-045032) and Salero (FP6-027122).

References

1. K. Bradley and B. Smyth. Improving Recommendation Diversity. In Proceedings of the 12th National Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland, 75-84, 2001.
2. D. Bridge. Diverse Product Recommendations Using an Expressive Language for Case Retrieval. In Proceedings of the 6th European Conference on Case-Based Reasoning, Springer-Verlag, 43-57, 2002.
3. Y. Chen, J.Z. Wang, and K. Robert, FRM CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning, IEEE transactions on Image Processing, Vol.14, No.8, 2005
4. M. Halvey and M.T. Keane: Analysis of online video search and sharing. In Proceedings of Hypertext 2007, 217-226, 2007.
5. Hopfgartner, F. Understanding Video Retrieval. VDM Verlag, 2007.
6. R. Hu, S. Rueger, D. Song, H. Liu and Z. Huang. Dissimilarity Measures for Content-based Image Retrieval. In Proceedings of IEEE International Conference on Multimedia & Expo, 2008.
7. D. W. Jacobs, D. Weinshall, and Y. Gdalyahu, Classification with nonmetric distances: image retrieval and class representation, IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 6, 583–600, 2000
8. G. Karypis. Evaluation of Item-Based Top-N Recommendation Algorithms. In Proceedings of the 10th International Conference of Information and Knowledge Management (CIKM), 247-254, 2001.
9. H. Liu, D. Song, S. Rüger, R. Hu, and V. Uren. Comparing Dissimilarity Measures for Content-based Image Retrieval. In Proceedings of the 2008 Asia Information Retrieval Symposium (AIRS'2008), LNCS, Harbin, China, 2008.
10. D. McSherry. Diversity Conscious Retrieval. In Proceedings of the 6th European Conference on Case Based Reasoning, Springer-Verlag, 219-233, 2002.
11. P. Over, G. Awad, W. Kraaij, and A. F. Smeaton. TRECVID 2007—Overview. In Proceedings of TRECVID 2007.
12. A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12): 1349-1380, 2000.
13. B. Smyth and P. McClave. Similarity v's Diversity. In Proceedings of the 3rd International Conference on Case-Based Reasoning, Springer-Verlag, 347-361, 2001.
14. I. Soboroff. Overview of the TREC 2004 novelty track. In Proceedings of the 13th Text Retrieval Conference (TREC), 2004.
15. K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, 707-710, 2006.
16. B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen and W.-Y. Ma. Improving Web Search Results Using Affinity Graph. In Proceedings of the 28th International SIGIR Conference on Research and Development in Information Retrieval, 504-511, 2005.
17. C.-N. Ziegler, S.M. McNee, J.A. Konstan and G. Lausen. Improving Recommendation Lists through topic Diversification. In Proceedings of the 14th International Conference in the World Wide Web, 22-32, 2005.
18. R. van Zwol, V. Murdock, L. Garcia Pueyo and G. Ramirez. Diversifying Image Search with User Generated Content. In Proceedings of 1st ACM international conference on Multimedia Information Retrieval, 2008.