Evidence Combination for Multi-Point Query Learning in Content-Based Image Retrieval

Jana Urban and Joemon M. Jose
University of Glasgow
Department of Computing Science
17 Lilybank Gardens, Glasgow, G12 8RZ, UK
{jana,jj}@dcs.gla.ac.uk

Abstract

In Multi-Point Query Learning a number of query representatives are selected based on the positive feedback samples. The similarity score to a multi-point query is obtained from merging the individual scores. In this paper, we investigate three different combination strategies and present a comparative evaluation of their performance. Results show that the performance of multi-point queries relies heavily on the right choice of settings for the fusion. Unlike previous results, suggesting that multi-point queries generally perform better than a single query representation, our evaluation results do not allow such an overall conclusion. Instead our study points to the type of queries for which query expansion is better suited than a single query, and vice versa.

1. Introduction

Relevance Feedback is a universally accepted means to improve Content-Based Image Retrieval (CBIR) systems. Many of the existing approaches (e.g. [6, 5]) are based on the geometric interpretation, in which an "ideal" query vector is constructed by moving it close to the positive (relevant) samples in feature space. While the ideal query was initially composed of a single average representation [6], it has been argued that relevant examples may form disjoint clusters, which are better captured by multiple query points [5, 4]. The retrieval algorithm for multi-point queries as proposed in [5] works in the following way. The cluster representatives, obtained from clustering all positive sample images, are chosen as the query points and issued to the retrieval system. The resulting scores from each query point are linearly combined to arrive at a single ranking of results.

In [5] the combination strategy of multi-point queries has not been studied extensively. To remedy this shortcoming, we compare three different list aggregation methods. We identify important parameters, such as the query size and the length of the lists, and evaluate their effect on the aggregation performance. Finally, we compare the multi-point approach to the single-point approach. In the following Section 2 we present the underlying techniques. Section 3 supplies the experimental details used to perform a simulated user-evaluation of the proposed fusion strategies for multi-point queries. The results and implications of this study are discussed in the remaining sections.

2. Multi-Point Query Learning

In our learning environment, we seek the best matching images (recommendations) for a selected group of images. The proposed group-based learning scheme involves (1) updating the system's matching parameters, (2) creating the multi-point query representation and computing a ranked list for each query point based on the learnt parameters, and (3) combining the individual result lists.

The parameter adaptation is achieved by the feature reweighting scheme described in [6]. The creation of multi-point queries for each group follows, whereby each query point represents one cluster of visually similar images in the group. The clusters are computed by an agglomerative hierarchical clustering algorithm, using Ward's minimum variance criterion [10]. The ideal number of clusters is automatically estimated using the method presented in [7]. The query points are the cluster centroids. For combining the result lists produced by each query point, we have considered three combination schemes.

1) The *Query Expansion (QEX)* scheme studied in this paper uses a simple linear combination of scores as originally proposed for multi-point queries in [5].

2) Inspired by the list aggregation problem in the web retrieval domain [1], we also consider an aggregation method purely based on ranks. In the *Voting Approach (VA)* each query representative is treated as a voter producing its own

Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04)
0-7695-2217-3/04 $20.00 © 2004 **IEEE**

individual ordering of candidates (images). The combined list is computed based on the *median rank aggregation* method proposed in [1]. It assumes a number of independent voters that rank a collection based on the similarity to a query. The aggregation rule then sorts the database objects with respect to their median of the ranks they receive from the voters. Their algorithm MEDRANK is very efficient and database friendly. The idea can be sketched as follows. Assume each voter produces a ranked list. From each list, access one element at a time, until a candidate is encountered in the majority of the lists, place this candidate as the top ranked of the final list. The second candidate will be placed second top, and so on. Continue until top k candidates are found, or there are no more candidates.

3) The *Dempster-Shafer (DS) Theory of Evidence Combination* is a powerful framework for the combination of results from various information sources, and has been extensively studied for IR purposes [3].

First, each information source (query point) is assigned an un-trust coefficient, β_j ($0 \geq \beta_j \geq 1$), which represents the uncertainty of the source of informal evidence. Initially, we use constant un-trust coefficients, i.e. $\beta_j = 1/L$, where L is the number of information sources (lists).

Second, we calculate the mass function for document d_i of information source j: $m_j(\{d_i\}) = S_{ij} \times (1 - \beta_j)$, where S_{ij} is the initial score of d_i from information source j. We have determined the score in two ways in this evaluation: score-based, and a rank-based one. The score-based S^s_{ij} is calculated as: $S^s_{ij} = \frac{d(q_{c_j}, d_i)}{\sum_{i=1}^{c} d(q_{c_j}, d_i)}$, where $d(q_{c_j}, d_i)$ is the distance between the cluster representative of the j-th cluster, q_{c_j}, and the i-th document, d_i, and c the number of items in the individual lists. While the rank-based S^r_{ij} is determined by: $S^r_{ij} = \frac{c-(r_{ij}-1))}{\sum_{i=1}^{c} i}$, where r_{ij} is the rank of d_i in the list produced by the information source j.

The final results are obtained by calculating an overall mass function for each document as a combination of the mass functions from the individual information sources and their un-trust coefficients. The details of the DS combination can be found in [3].

3. Experimental Setup

The evaluation compares the three fusion strategies QEX, VA, and DS for multi-point queries. DS_r and DS_s refer to the rank-based and score based combination, respectively. These strategies are based on combining the top c (list length or cutoff value) results from each query point, and returning the overall top k (recommendation size). Throughout the evaluation k is set to 10. The *average query point* AVG [6] is used as baseline.

Experiments are conducted on photo CD 7 of the Corel image collection containing 23796 images. Domain experts have organised the collection into 238 high-level semantic categories, from which we have selected 10, of 100 images each, for the evaluation ("aviation", "bob sledding", "flags", "minerals", "roses", "rock formations", "stamps", "tribal people", "volcano", "dolphins"). We use the category information as ground truth, that is images from the same category as the images in the query group are considered relevant. For each category 50 queries were randomly selected resulting in a total of 500 queries. We use 6 low-level colour (Average RGB (3), Colour Moments (9) [9]), texture (Co-occurrence (20), Autocorrelation (25), and Edge Frequency (25)) and shape (Invariant Moments (7) [2]) features (feature dimension in brackets).

The performance is measured in *precision* and *recall* [11]. We are primarily concerned with the quality of the recommendations, that is how many of the k returned images are relevant. The *precision* after the k-th image retrieved, $P(k)$, provides a good indication for this. $P(k)$ values are in the range $[0, 1]$ (corresponding to 0-100%). The *recall* value measures how many of the total available relevant images are returned. The recall level becomes an important performance measure when running the recommendation system over a number of feedback iterations.

4. Results Analysis

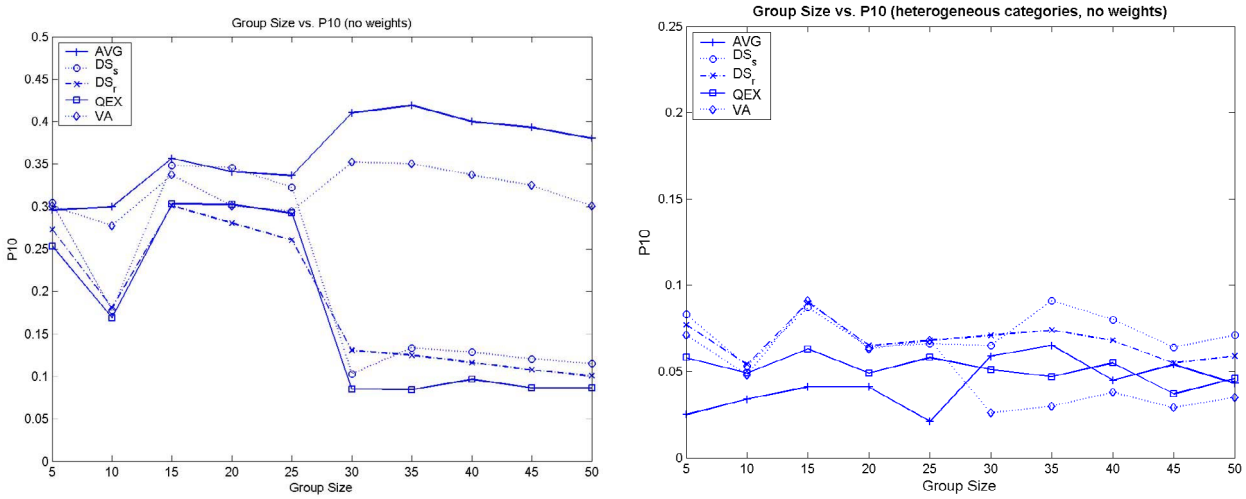
We have considered various axes of variation that can affect the performance of the mergers: the group size, the introduction of a weighting mechanism for weighing the query points, and the list length. In addition, we report the results of a relevance feedback run, where all of the three parameters above are fixed, to test the performance of the recommendation system "in action".

4.1. Variations of Group Size

In the first run the group size (# query images) is varied from 5 to 50 in steps of 5. The list length, c, is set to 100. Figure 1 shows the results. The graphs in Figure 1(a) reflect a "scissor trend", where all methods start at approximately equal performance for a group size of 5, but then dramatically diverge from a group size of 25.[1] AVG and VA continually increase performance with growing group size, while all other multi-point query methods tend to worsen.

Analysing the individual categories, we could identify two classes, namely *homogeneous* and *heterogeneous* categories. Homogeneous categories contain visually similar

1 The large jump in performance at the group size 25 is influenced by the automatic choice of the number of clusters. The average number of clusters increases steeply, while the average cluster size decreases at this point.



(a) All categories (b) Heterogeneous categories

Figure 1. P(10) vs. group size, average over all and heterogeneous categories.

images and are well distinguishable from other categories (e.g. "roses"), while heterogeneous categories contain visually less similar images and/or are not easily distinguishable from other categories (e.g. "tribal people"). Our sample categories contained 5 of each.

The graph for the homogeneous categories shows the same trend as the overall graph only about 20% higher precision on average, and is therefore omitted. Figure 1(b) shows the results for the heterogeneous categories. From these results we see that AVG is best suited for homogeneous categories, where one can assume an "ideal" query representation to describe it. In heterogenous categories, which are not necessarily described best by a single representation, the multi-point queries succeed in a slight increase in performance.

4.2. Introduction of Cluster/List Weights

In the next run we have introduced a weighting scheme for the multi-point queries similar to the one proposed in MARS [5]. Each query point is associated with a weight proportional to the cluster size it represents, i.e. $w_i = \frac{m_i}{M}$, where m_i is the number of images in cluster i, and M the total number of images in the group.

In VA, the weights influence the ranking in two ways. First, the lists are sorted in descending order of their weights, as this algorithm is sensitive to the sequence in which they are processed in the merging process. Second, the scores each list gives its candidates will be weighed. Formally, to incorporate the query-point weights, w_i, determined above, each list, l_i (where $1 \leq i \leq L$ and L the number of voters), is able to score its candidates by its weight. The overall score of a candidate c, $s(c)$, is accumulated: $s(c) = \sum_{i=1}^{l} w_i$, where $l \leq L$. The majority criterion from above, which states that a candidate is carried forward to the final list if it is seen in more than half of the lists, is fulfilled, if $s(c) > 0.5$.

	QEX	DS_s	DS_r	VA	AVG
weighted	15.9%	17.5%	**21.1%**	**34.5**	36.3
non-w.	**17.6%**	**21.0%**	18.8%	31.8	36.3

Table 1. Average P(10) (in %) for weighted and non-weighted versions.

In all the other methods, the inverse of weights are used, since the lists are sorted by distance or rank values. Thus $w'_j = \frac{1/w_j}{\sum_{\iota=1}^{L} w'_\iota}$. In QEX, this results in a weighted linear combination of distance scores from the individual lists, i.e. $s(c) = \sum_{j=1}^{L} w'_j s_j(c)$, where $s(c)$ is c's overall score and $s_j(c)$ its score in the j-th list. In DS, the weights are used to derive the un-trust coefficient, i.e. $\beta_j = (1 - w'_j)$.

Table 1 contrasts the weighted and non-weighted performance for each method averaged over the various group sizes. The two rank-based methods, VA and DS_r, perform slightly better if list weights are introduced in the merging process. On the contrary, the performance of QEX and DS_s drops by 1.7% and 3.5% points, respectively.

4.3. Variations of Cutoff Value

We have limited the length of the individual lists being merged, c, to $k \leq c \leq N$ (N: total # images), for computational and retrieval performance reasons. To determine the influence of c, we have varied it from 10 to 1000. The group

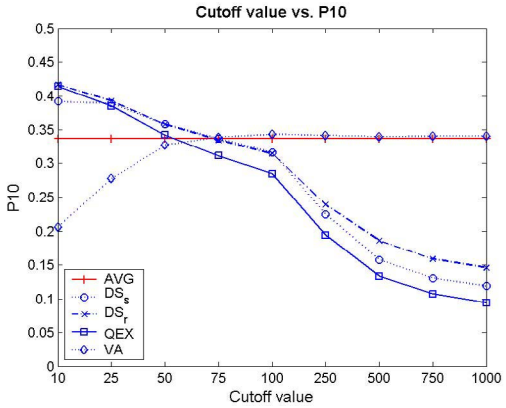


Figure 2. P(10) vs. cutoff value.

size is set to 25 in this run. The graph in Figure 2 plots c versus the $P(10)$ performance. As the graph shows, the performance of QEX and DS is best at $c = 10$, decreasing rapidly with a growing c. The curve for VA exhibits the opposite behaviour, increasing steadily up to a peak at $c = 100$.

4.4. Performance on Relevance Feedback

In the interactive scenario, user interaction is simulated by starting with groups of 3 randomly chosen images from a given category and performing relevance feedback from the top 10 results. In each feedback iteration the simulated user adds all relevant images to the current group. A query run terminates, when no more relevant images can be found.

From the previous runs, we determined the optimal settings for each fusion method. The list cutoff value is set to 100 for VA, and 10 for all other fusion methods. Further, the rank-based methods (VA and DS_r) incorporate a weighting of the query points, while in the score-based methods (QEX and DS_s) lists are combined without weights.

The average number of relevant images found after convergence are 13.8% for QEX, 11.5% for DS_s, 12.2% for DS_r, 29.4%, for VA, and 29.8% for AVG. Overall, AVG outperforms every multi-point query strategy. While VA's performance is almost as good as the baseline, all other strategies perform significantly worse.

5. Discussion and Conclusions

The evaluation has confirmed that list combination is an intricate topic, and previous results of superior performance of multi-point queries over single point queries in general as reported by [5] or [4] should be used cautiously. Factors, such as the cluster algorithm, the list cutoff value, the weighting of lists etc. can have a detrimental impact if not applied carefully.

Overall, multi-point queries can provide a benefit over a single group representative, but only if a suitable combination strategy is employed. A simple linear combination of the raw scores is sensitive to noise, especially when the number of lists becomes large and the lists are very different from each other. In this case, computing the average of scores acts like a smoothing operation. Kim et al. have already observed that this form of query expansion creates a large contour covering all query points [4]. On the other hand, VA has exhibited stable performance and is the only fusion method with comparable performance to AVG under various settings.

In general, multi-point queries perform better than a single point query in heterogeneous groups, where the images will indeed form multiple distinct clusters. On the contrary, a single query point is sufficient to describe homogeneous groups. In addition, from a sufficiently large group size a single query representation can be employed in any case.

References

[1] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 301–312, 2003.

[2] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8:179–187, Feb. 1962.

[3] J. M. Jose. *An Integrated Approach for Multimedia Information Retrieval*. PhD thesis, The Robert Gordon University, Aberdeen, Apr. 1998.

[4] D.-H. Kim and C.-W. Chung. Qcluster: relevance feedback using adaptive clustering for content-based image retrieval. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 599–610, 2003.

[5] K. Porkaew, K. Chakrabarti, and S. Mehrotra. Query refinement for multimedia similarity retrieval in MARS. In *Proc. of the ACM Int. Conf. on Multimedia*, pages 235–238, Orlando, Florida, 1999.

[6] Y. Rui and T. S. Huang. Optimizing learning in image retrieval. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 236–245, Los Alamitos, June 2000.

[7] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. Technical Report CS-2003-18, Florida Institute of Technology, 2003.

[8] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Brooks and Cole Publishing, 2nd edition, 1998.

[9] M. Stricker and M. Orengo. Similarity of color images. In *Proc. of the SPIE: Storage and Retrieval for Image and Video Databases*, volume 2420, pages 381–392, Feb. 1995.

[10] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[11] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, London, 2nd edition, Jan. 1979.