THE INFLUENCE OF THE SIMILARITY MEASURE TO RELEVANCE FEEDBACK

I. Mironică*, *B. Ionescu**,†, *C. Vertan**

*LAPI, University ”Politehnica” of Bucharest, 061071, Romania,
†LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944 France
{*imironica, bionescu, cvertan*}*@alpha.imag.pub.ro*

ABSTRACT

In this paper we discuss the influence of the similarity measure in the context of content-based image retrieval. To bridge inherent descriptor gap, we propose a new relevance feedback (RF) approach which uses a hierarchical clustering strategy. It has the advantage of performing on the initial set of retrieved images, instead of performing additional queries, as most approaches do. The images are clustered with respect to the positive/negative examples provided by the user, in a continuous manner, as user successively browses through the retrieved images. Comparative experimental tests conducted using state-of-the-art content descriptors and distance measures show that the proposed approach provides significant improvement in retrieval performance while the choice of the distance metric plays a decisive role on system performance.

Index Terms— content-based image retrieval, distance measures, relevance feedback, hierarchical clustering.

1. INTRODUCTION

During the last two decades, Content Based Image Retrieval (CBIR) established itself as a domain with high impact on application areas such as multimedia database systems [1]. The actual generation of CBIR systems focuses on attaining human-centered and inspired retrieval capabilities. However, the retrieval process still follows the classic feature-based mechanism. Images have to be summarized with content descriptors that aim to represent - as faithfully as possible - the underlaying semantic visual content. These descriptors are to be extracted for the entire data set, e.g. Internet, databases, etc., and stored accordingly for further use to matching user queries. The actual retrieval rely on defining the concept of similarity between these features, which is often described by means of some distance measures. Due to the subjectiveness of the process, the system typically provides the user with not only one response, but a ranked list of possible results by decreasing similarity with the query. The backbone of this mechanism are the *representative power* of the descriptors and the choice of the *similarity measure* [1] [2].

Existing image feature extraction techniques are now capable of providing good retrieval capabilities, approaching significantly the performance of high level text descriptors; a relevant example in this respect is the new Google Search by Image system. Current state-of-the-art descriptors range from classic MPEG-7 features [3], Accelerated Segment Test (FAST) to the popular Scale Invariant Feature Transform (SIFT) [10] and Speeded Up Robust Features (SURF) [9]. As for the similarity metric, some of the popular choices remain the Euclidean-based metrics [4] [5].

Regardless current state-of-the-art descriptor performance, CBIR systems are inherently limited by the gap between the real world and its projection captured by imaging devices and also by the gap between the knowledge automatically extracted from the recorded data and its actual semantic meaning. In addition to those, CBIR is affected by the subjectivity of human perception (different persons may perceive differently similar visual information). Current development of CBIR systems focus on bridging these paradigms [1] [2].

To this end, one of the adopted solutions was to take advantage directly of the human expertise in the retrieval process, known as Relevance Feedback (RF). For a certain retrieval query, user has to provide feedback by marking the results as relevant or non-relevant. Using this information, the system then automatically computes a better representation of the information needed and retrieval is further refined.

One of the earliest and most successful RF algorithms is the Rocchio algorithm [8]. It updates the query features by adjusting the position of the original query in the feature space according to the positive and negative examples and their associated importance factors. Another example is the Feature Relevance Estimation (FRE) approach [11], which assumes for a given query that a user may consider some specific features more important than others. Every feature is given an importance weight such that features with greater variance have lower importance than elements with smaller variations. More recently, machine learning techniques have been introduced to relevance feedback approaches. Some of the most successful techniques are using Support Vector Machines (SVM) [12], classification trees, such as Decision Trees [13], Random Forest [14]; or boosting techniques, such as AdaBoost [15], Nearest Neighbor [16] or Gradient

[†]Part of this work was supported under Financial Agreement POSDRU/89/1.5/S/62557 and POSDRU/89/1.5/S/64109.

Boosted Trees [17]. The relevance feedback problem can be formulated either as a two-class classification of the negative and positive samples or as a one-class classification problem (separating positive samples from negative samples).

In this paper we propose a new relevance feedback approach which uses an adaptive agglomerative clustering strategy. The main advantages of the proposed hierarchical clustering relevance feedback (HCRF) approach are implementation simplicity and speed because it is computationally more efficient than other clustering techniques, such as SVMs [12]. Further, unlike most RF algorithms (e.g., FRE [11] and Rocchio [8]), it does not modify the query or the similarity. The remaining retrieved images are simply clustered according to class label.

Another main contribution of this work is in the study of the influence of the choice of the distance measure to the retrieval performance. We tested a broad variety of techniques which perform both in the pixel-domain and coefficient-domain: Euclidean, Manhattan (particular cases of the Minkovski distance); probabilistic divergence measures: Canberra and Bray-Curtis [6]; fidelity family metrics: Squared-Chored, Matusita and Bhattacharyya; squared L2 family: Pearson and Clark; intersection family: Cosine, Lorentzian, Soergel, Czekanowski, Motika, Ruzicka and Tanimoto [7]; Chi-Square distance used in machine learning and data clustering; and Shannon entropy family: Jefrey divergence and Dice [5]. The selection of these approaches was motivated by their appropriateness with the structure of content descriptors.

Experimental tests conducted on several standard image databases and using current state-of-the-art image descriptors show that the proposed RF achieves better retrieval performance than other consecrated approaches, while the choice of the distance metric plays an important role to this process.

The remainder of the paper is organized as follows: Section 2 depicts the algorithm of the proposed hierarchical relevance feedback approach; experimental validation is presented in Section 3, while Section 4 presents the conclusions and discusses future work.

2. PROPOSED RELEVANCE FEEDBACK

We propose an RF approach that is based on Hierarchical Clustering (HC). A typical agglomerative HC strategy starts by assigning one cluster to each object in the feature space. Then, similar clusters are progressively merged based on the evaluation of a specified distance measure. By repeating this process, HC produces a dendrogram of the objects, which may be useful for displaying data and discovering data relationships. This clustering mechanism can be very valuable in solving the RF problem by providing a mechanism to refine the relevant and non-relevant clusters in the query results. A hierarchical representation of the similarity between objects in the two relevance classes allows us to select an optimal level from the dendrogram which provides a better separation of the two than the initial retrieval.

The proposed hierarchical clustering relevance feedback is based on the general assumption that the image content descriptors provide sufficient representative power that, within the first window of retrieved images, there are at least some images relevant to the query that can be used as positive feedback. This can be ensured by adjusting the size of the initial feedback window. Also, in most cases, there is at least one non-relevant image that can be used as negative feedback. The algorithm comprises three steps: *retrieval*, *training*, and *updating*.

Retrieval. We provide an initial retrieval using a nearest-neighbor strategy. We return a ranked list of the N_{RV} images most similar to the query image using the distance between features. This constitutes the initial RF window. Then, the user provides feedback by marking relevant results, which triggers the actual HCRF mechanism.

Training. The first step of the RF algorithm consists of initializing the clusters. At this point, each cluster contains a single image from the initial RF window. Basically, we attempt to create two dendrograms, one for relevant and one for non-relevant images. For optimization reasons, we use a single global cluster similarity matrix for both dendrograms. To assess similarity, we compute the distance between cluster centroids (which, compared to the use of min, max, and average distances, provided the best results). Once we have determined the initial cluster similarity matrix, we attempt to merge progressively clusters from the same relevance class (according to user feedback) using a minimum distance criterion. The process is repeated until the number of remaining clusters becomes relevant to the image categories in the retrieved window (regulated by a threshold τ).

Updating. After finishing the training phase, we begin to classify the next images as relevant or non-relevant with respect to the previous clusters. A given image is classified as relevant or not relevant if it is within the minimum centroid distance to a cluster in the relevant or non-relevant image dendrogram.

Algorithm 1 summarizes the steps involved. The following notations were used: $N_{clusters}$ is the number of clusters, $sim[i][j]$ denotes the distance between clusters C_i and C_j (i.e., centroid distance; for assessing distance we investigate several strategies which are presented in Section 3), τ represents the minimum number of clusters which triggers the end of the training phase (set to a quarter of the number of images in a browsing window), τ_1 is the maximum number of searched images from the database (set to a quarter of the total number of images in the database), τ_2 is the maximum number of images that can be classified as positive (set to the size of the browsing window), TP is the number of images classified as relevant, and $current\_image$ is the index of the currently analyzed image.

Algorithm 1 Hierarchical Clustering Relevance Feedback.

```
$N_{clusters} \leftarrow N_{RV}$; $clusters \leftarrow \{C_1, C_2, ..., C_{N_{clusters}}\}$;
for $i = 1 \rightarrow N_{clusters}$ do
    for $j = i \rightarrow N_{clusters}$ do
        $compute\ sim[i][j]$;
        $sim[j][i] \leftarrow sim[i][j]$;
    end for
end for
while ($N_{clusters} \geq \tau$) do
    $\{min_i, min_j\} =$
    $argmin_{i,j}|_{C_i, C_j \in \{same\ relevance\ class\}}(sim[i][j])$;
    $N_{clusters} \leftarrow N_{clusters} - 1$;
    $C_{min} = C_{min_i} \cup C_{min_j}$;
    for $i = 1 \rightarrow N_{clusters}$ do
        $compute\ sim[i][min]$;
    end for
end while
$TP \leftarrow 0$; $current\_image \leftarrow N_{RV} + 1$;
while (($TP \leq \tau_1$) || ($current\_image < \tau_2$)) do
    for $i = 1 \rightarrow N_{clusters}$ do
        $compute\ sim[i][current\_image]$;
    end for
    if ($current\_image$ is classified as relevant) then
        $TP \leftarrow TP + 1$;
    end if
    $current\_image \leftarrow current\_image + 1$;
end while
```

3. EXPERIMENTAL RESULTS

The validation of the proposed relevance feedback approach was conducted on several standard image databases, namely: Microsoft Object Class Recognition[1] which sums up to 4300 images distributed into 23 categories (e.g. animals, people, airplanes, cars, etc.) and Caltech-101[2] which contains a total of 9146 images, split between 101 distinct objects (including faces, watches, ants, pianos, etc.) and a background category - for a total of 102 categories.

In what concerns the image content descriptors, we tested several state-of-the-art approaches from the existing literature which are known to be successfully employed to the CBIR task, namely: MPEG-7 image descriptors [3]: Color Histogram Descriptor, Color Layout Descriptor, Edge Histogram Descriptor and Color Structure Descriptors; classic color descriptors: Autocorrelogram, Color Coherence Vectors and Color Moments; and feature detectors: SURF, SIFT, Good Features to Track (GOOD), STAR, Accelerated Segment Test (FAST), Maximally Stable Extremal Regions (MSER) and Harris Detector available with the OpenCV library (Open Source Computer Vision[3]). Features were represented with a Bag-of-Visual-Words model.

To assess performance, we computed the overall Mean Average Precision (MAP) as the area under the uninterpolated precision-recall curve (see also $trec\_eval$ scoring tool[4] [2]). The evaluation consisted of systematically considering each image from the database as query image and retrieving the remainder of the database accordingly. Precision, recall and MAP are averaged over all retrieval experiments. Experiments were conducted for various retrieval browsing windows, N_{RV} ranging from 20 to 50 images. For brevity reasons, in the following we present only the most representative results which were obtain for $N_{RV} = 30$.

3.1. Retrieval experiment

In the first experiment we analyze the influence of the distance measure on the performance of a classic retrieval system. In this respect, we use the classic nearest neighbor retrieval step of the RF algorithm (see Section 2). Figure 1 presents the MAP obtained for the two data sets and the aforementioned features. Although the descriptors provide in average more or less comparable performance on same data set, results show that the distance measure plays a critical role.

In the case of the Microsoft data set which has lowest diversity of classes, the best results are obtained with Bhattacharyya using MPEG-7 descriptors (MAP of 57%) followed by Canberra and Clark using classic color descriptors (MAP of 55% and 54%, respectively) which is an improvement of around 18% above the average descriptor value. Results are significantly decreasing on the Caltech-101 data set which contains five times more categories. The highest accuracy is achieved again for the classic MPEG-7 descriptors using Bhattacharyya and Canberra distances (MAP of 23.4% and 23.2%, respectively). In this case, the improvement is of at least 5% above the average descriptor value. In what concerns the computation time, it should be considered the fact that Bhattacharyya is one of the most expensive solutions.

It can be noted that some distance measures can be less adapted to the structure of the descriptors, an example are Bhattacharyya and Canberra which performed significantly worst with Bag-of-Visual-Words representation of feature descriptors (see SURF, SIFT, Harris and GOOD in Figure 1). Another interesting result is that the classic Euclidean distance, despite its popularity, proves to have poor discriminant power in most of the cases.

3.2. Relevance feedback experiment

Form the previous experiment it can be seen that the retrieval performance is relatively low with either of the methods. In this experiment we test the advantage of using relevance feedback. We compare the proposed HCRF approach against other validated methods from the literature: Rocchio algorithm [8], Relevance Feature Estimation (RFE) [11], Support Vector Machines (SVM) [12], Decision Trees (TREE) [13], AdaBoost (BOOST) [15], Random Forest [14], Gradient Boosted Trees (GBT) [17] and Nearest Neighbor (NN) [16]. As for the previous experiment, we assess similarity using

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition.

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101.

[3] http://opencv.willowgarage.com/wiki.

[4] http://trec.nist.gov/trec_eval.

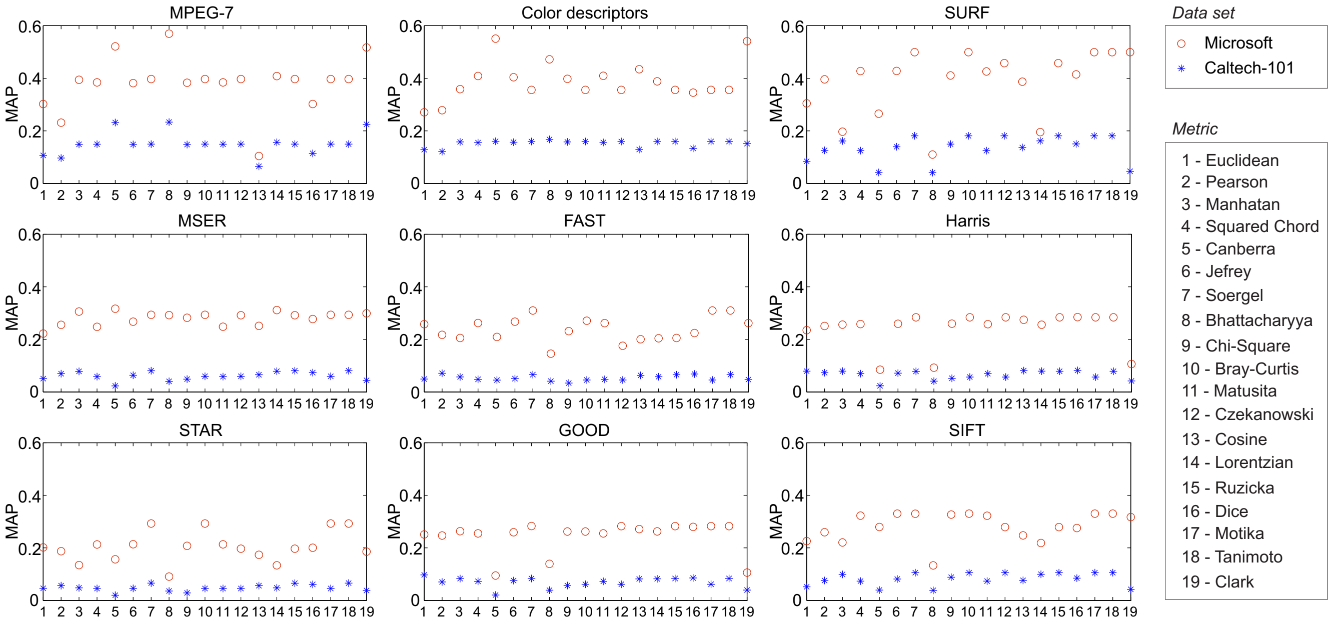

Fig. 1. Mean Average Precision for retrieval using various descriptor set - metric combinations.

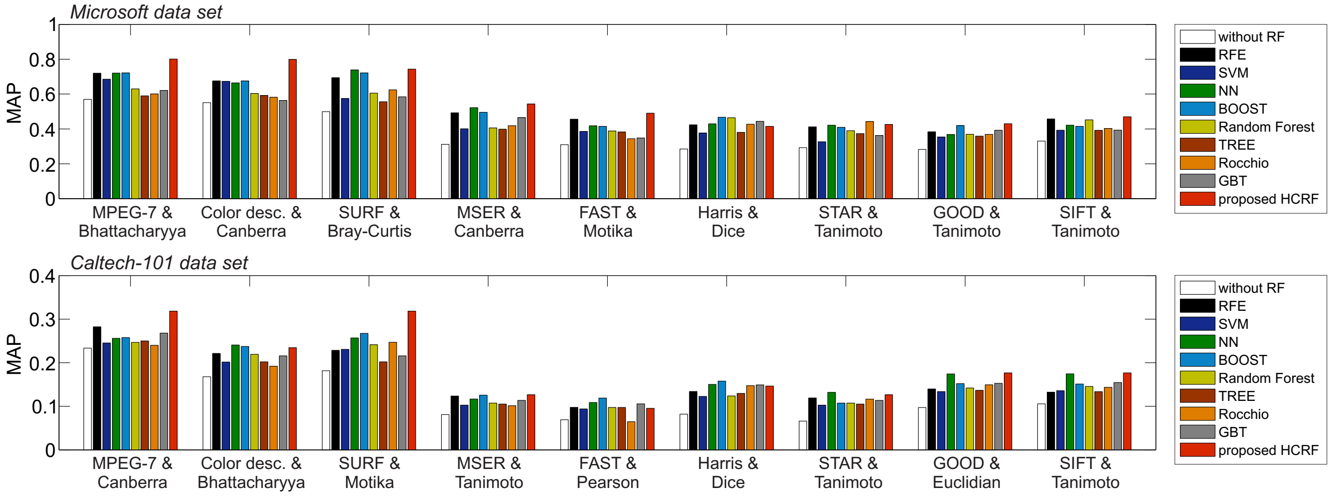

Fig. 2. Mean Average Precision for retrieval with relevance feedback using various descriptors.

various distance measures. Each experiment is conducted using only one feedback iteration. Some of the results are presented in Figure 2. For brevity reasons, we depict only the results obtained with the distance measure providing the highest performance.

From the RF point of view, globally, all RF strategies provide significant improvement in retrieval performance compared to the retrieval without RF. Better performance is naturally obtained when targeting a more reduced number of image categories. For instance, on the Micr̈osof data set (23 classes) RF MAP is up to 80% (proposed HCRF), compared to only 57% without RF (improvement of 23%). On Caltech-101 (102 classes) the highest MAP is 32% (proposed HCRF) compared to 23% without RF (improvement of 9%).

The proposed HCRF tends to provide better retrieval performance in most of the cases. Table 1 summarizes some of these results. For the Microsoft data set, the highest increase in performance is achieved for MPEG-7 descriptors, namely 8% compared to BOOST; while for Caltech-101, the highest increase is of 5% for SURF compared also to BOOST. Less accurate results are obtained for descriptors such as FAST, STAR or MSER due to their limited discriminant power for

Table 1. Improvement achieved by the proposed HCRF.

Microsoft data set			
descriptor	*1st MAP*	*2nd MAP*	*3rd MAP*
MPEG-7	HCRF - 80%	BOOST - 72%	NN - 72%
Color desc.	HCRF - 80%	RFE - 68%	BOOST - 68%
Caltech-101 data set			
descriptor	*1st MAP*	*2nd MAP*	*3rd MAP*
MPEG-7	HCRF - 32%	RFE - 28%	GBT - 27%
SURF	HCRF - 32%	BOOST - 27%	NN - 26%

this particular task.

From the distance point of view, results show that there is no general preference for a certain distance metric. As expected, the choice of distance is dependent on the type of content descriptors. Nevertheless, Canberra and Bhattacharyya distances prove to be more reliable for use with classic numeric content descriptors, such as MPEG-7 and color descriptors, while Tanimoto provided better performance for Bag-of-Visual-Words approaches.

4. CONCLUSIONS

We have brought into discursion the influence of the distance measure to the performance of image retrieval. We have proposed a relevance feedback approach that uses the hierarchical clustering of the query results. Experimental testing performed on several standard databases using state-of-the-art descriptors and distance measures, show the advantage of the proposed approach (performance increase is up to 23% in terms of mean average precision). Although descriptors provided more or less comparable retrieval results, the choice of the distance measure proves to be highly critical for the performance. Distances such as Canberra and Bhattacharyya proved to be more reliable for use with classic numeric descriptors, such as MPEG-7 and color descriptors, while metrics such as Tanimoto provided better performance for Bag-of-Visual-Words approaches. Future work will mainly involve considering the constraints of large-scale indexing.

5. REFERENCES

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, ”Content-based Image Retrieval at the End of the Early years”, IEEE Trans. on PAMI, 22(12), 2000.

[2] A. F. Smeaton, P. Over, W. Kraaij, ”High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements”, Multimedia Content Analysis Theory and Applications, pp. 151-174, 2009.

[3] J. M. Martinez: ”Standards - MPEG-7 Overview of MPEG-7 Description Tools”, IEEE MultiMedia, 9(3), pp. 83-93, 2002.

[4] R.O. Duda, P.E. Hart, D.G. Stork ”Pattern Classification”, Wiley-Interscience; ISBN-978-0-471-05669-0, 2000.

[5] E. Deza, M.M. Deza ”Dictionary of Distances”, Elsevier Science, 1st edition, ISBN-13: 978-0-444-52087-6, 2006.

[6] M. Hatzigiorgaki, A. N. Skodras, ”Compressed Domain Image Retrieval: A Comparative Study of Similarity Metrics”, SPIE Visual Communications and Image Processing, vol. 5150, 2003.

[7] S.-H. Cha, ”Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions”, Int. Journal of Mathematical Models and Methods in Applied Sciences, 1(4), pp. 300-307, 2007.

[8] N. V. Nguyen, J.-M. Ogier, S. Tabbone, A. Boucher, ”Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR”, International Conference on Machine Learning and Pattern Recognition, 2009.

[9] H. Bay, A. Ess, T. Tuytelaars, L. J. V. Gool, ”SURF: Speeded up Robust Features”, Computer Vision and Image Understanding, 110(3), pp. 34635, 2008.

[10] D. Lowe. ”Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 60(2), 91-110, February 23, 2005.

[11] Y. Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman, ”Relevance Feedback: a Power Tool for Interactive Content-Based Image Retrieval”, IEEE Trans. on Circuits and Video Technology, 8(5), pp. 644-655, 1998.

[12] S. Liang, Z. Sun, ”Sketch Retrieval and Relevance Feedback with Biased SVM Classification”, Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

[13] S.D. MacArthur, C.E. Brodley, C.-R. Shyu, ”Interactive Content-Based Image Retrieval Using Relevance Feedback”, Computer Vision and Image Understanding, 88(2), pp. 55-75, 2002.

[14] Y. Wu, A. Zhang, ”Interactive Pattern Analysis for Relevance Feedback in Multimedia Information Retrieval”, Journal on Multimedia Systems, 10(1), pp. 41-55, 2004.

[15] S.H. Huang, Q.J Wu, S.H. Lu, ”Improved AdaBoost-Based Image Retrieval with Relevance Feedback via Paired Feature Learning”, ACM Multimedia Systems, 12(1), pp. 14-26, 2006.

[16] G. Giacinto, ”A Nearest-Neighbor Approach to Relevance Feedback in Content-Based Image Retrieval”, ACM Int. Conf. on Image and Video Retrieval, 2007.

[17] J. Ye, J. Chow, J. Chen, Z. Zheng, ”Stochastic Gradient Boosted Distributed Decision Trees”, ACM Conference on Information and Knowledge Management, 2009.