Hierarchical Clustering Relevance Feedback for Content-Based Image Retrieval

Ionuţ Mironică[1], Bogdan Ionescu[1,2], Constantin Vertan[1]
[1] LAPI, University "Politehnica" of Bucharest, 061071, Romania
[2] LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944, France
{*imironica, bionescu, cvertan*}*@alpha.imag.pub.ro*

Abstract

In this paper we address the issue of relevance feedback in the context of content-based image retrieval. We propose a method that uses an hierarchical cluster representation of the relevant and non-relevant images in a query. The main advantage of this strategy is in performing on the initial set of the retrieved images (user feedback is provided only once for a small number of retrieved images) instead of performing additional queries as most approaches do. Experimental tests conducted on several standard image databases and using state-of-the-art content descriptors (e.g. MPEG-7, SURF) show that the proposed method provides a significant improvement in the retrieval performance, outperforming some other classic approaches.

1. Introduction

The actual challenge of the existing information retrieval systems is not in the accessibility of media, but in their capability of identifying and selecting only relevant information according to some user specifications. This issue became more critical due to the extent development of technology, e.g. portable multimedia terminals, wireless transmission protocols, imaging devices which basically unlimited the access to information everywhere. Information retrieval has become now a part of our daily social interaction.

The actual generation of content-based image retrieval systems (CBIR) focuses more and more on attaining human centered and inspired searching capabilities. However, the retrieval process still follows the classic feature-based mechanism. The basic idea behind CBIR is to compactly describe an image forming so a digital signature which best represent the underlaying visual contents. These descriptors are to be stored by the system and then used to match a user query image to the most resembling image within the data set (e.g. Internet, databases, etc.). This is carried out by employing some similarity criteria [1]. Due to the subjective nature of the process, the system typically provides the user with not only one response but a ranked list of possible choices to select from.

Despite the high variability of content descriptors used (color, shape, texture, features, statistical [2]) and of the classification techniques, CBIR systems are inherently limited by the gap between the real world and its representation through computer vision techniques. On one hand, we have a sensor gap, i.e. the discrepancy between the real world and its projection captured by imagining devices. On the other hand, there is a semantic gap between the knowledge automatically extracted from the recorded data and its semantic meaning (e.g. yellow pears have their own color and shape properties, but similar properties can be shared by other objects like a yellow car).

To make things worse, this can also benefit from the subjectivity of human perception. At some level, different persons may perceive differently similar visual information. Current development of CBIR systems focuses on overcoming these paradigms and narrow the gap between low-level image features and high-level semantic concepts.

One of the adopted solutions was to take advantage directly of the human expertise in the retrieval process, which is known to as Relevance Feedback or RF. A general RF scenario can be formulated thus: for a certain retrieval query, user gives his opinion by marking the results as relevant or non-relevant. Then the system automatically computes a better representation of the information needed based on this information and retrieval is further refined. Relevance feedback can go through one or more iterations of this sort. This basically improves the system response based on query related ground-truth.

In this paper we address these particular techniques with the objective of narrowing the descriptor semantic gap and thus improving the relevance of the retrieval results.

The remainder of the paper is organized as follows. Section 2 presents a state-of-the art of the literature and situates our work accordingly. Section 3 depicts the algorithm of the proposed hierarchical relevance feedback approach. Experimental validation is presented in Section 4 while Section 5 presents the conclusions and discusses future work.

978-1-4673-2369-7/12/$31.00 ©2012 IEEE

2. Previous work

One of the earliest and most successful RF algorithms is the Rocchio algorithm [3] [4]. The relevance feedback is based on a set of R relevant and N non-relevant documents selected from the current user relevance feedback window. Using this information, query features are updated 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 Relevance Feature Estimation (RFE) approach [5] which assumes, for a given query, that according to the user's subjective judgment some specific features may be more important than other features. Every feature will have an importance weight computed as $w_i = 1/\sigma$ where σ denotes the variance of relevant retrievals. Therefore, features with higher variance lead to lower importance factors than elements with reduced variation.

More recently, machine learning techniques found their application with relevance feedback approaches. Some of the most successful techniques are using Support Vector Machines [6], classification trees, e.g. Decision Trees [7], Random Forest [9] or boosting techniques, e.g. AdaBoost [8]. The relevance feedback problem can be formulated either as a two class classification of the negative and positive samples; or as an one class classification problem, i.e. separate positive samples by negative samples.

In this paper we propose a new approach which is based on employing a hierarchical agglomerative clustering strategy. The main advantage of the proposed scheme is first in the use of the set of primarily retrieved images instead of performing additional queries, like Rocchio and RFE approaches do. Also, the proposed approach allows faster implementation than other similar classification based implementations, like SVM, boosting or classification trees. The images are clustered with respect to the positive and negative examples provided by the user in a continuous manner, as the user successively browses through new sets of retrieved images. Experimental tests conducted on several standard image databases and using current state-of-the-art image descriptors (e.g. MPEG-7, SURF, color descriptors) prove that the proposed RF increases retrieval performance and outperforms other classic RF approaches.

3. Proposed relevance feedback approach

As previously stated, the proposed RF is based on a Hierarchical Agglomerative Clustering (HAC) approach. A general HAC partitioning strategy can be formulated thus: first, every document in the partitioning space is assigned to a new cluster. During each iteration, using a certain distance metric (e.g. average distance, minimal variance, Ward's distance), we successively search for the most similar clusters in the current partition. These clusters are then merged resulting in the decrease of the total number of clusters by one. By repeating the process, HAC will produce a dendrogram of the observations, which may be informative for data display and discovery of data relationship.

Applied to the retrieved images, this dendrogram based representation can be successfully exploited to our RF problem as it provides a multi-level cluster representation of the similarity of the images in the two classes: relevant and non-relevant. Refinement of the retrieval is performed by reassigning further retrieved images to these clusters. However, several preliminary hypothesis should be adopted. First, we should consider that the image content descriptors provide enough representative power such that within the first window of retrieved images there are at least some relevant images for the query that can be used as positive feedback. This can be assured by considering the right size of the window. Second, there should be at least one non-relevant image that can be used as negative feedback. We propose the following algorithm:

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 Euclidean 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 Euclidean 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.

Updating. After finalizing 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.

The method's algorithm is presented with Algorithm 1, where the following notations were adopted: N_{RV} is the number of images in the browsing window, $N_{clusters}$ is the number of clusters, $sim[i][j]$ denotes the distance between

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
    $\{min_i, min_j\} = argmin_{i,j}|_{C_i, C_j \in \{same\ relev.\}}(sim[i][j])$;
    $N_{clusters} \leftarrow N_{clusters} - 1$; $C_{min} = C_{min_i} \cup C_{min_j}$;
    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 $((TP \leq \tau_1) \parallel (current\_image < \tau_2))$ 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
```

clusters C_i and C_j (i.e., centroid distance), τ represents the minimum number of clusters which triggers the end of the training phase (presented later), 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.

The proposed algorithm involves the choice of several parameters. An important parameter of the hierarchical agglomerative clustering algorithms is the distance measure. This choice will influence the shape of the clusters in the feature space and in consequence the way clusters are being merged. As previously stated, we tested several strategies which are described with the experimental results section.

Another important aspect is the minimum number of clusters (τ) to be considered relevant for the current image categories and which triggers the end of the training phase. We estimate τ in an adaptive manner using the "elbow criterion" [11]. The minimum number of clusters is determined at the point where adding another cluster doesn't add sufficient information. By plotting the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. We choose the minimum number of clusters at this point.

The main advantages of the proposed HCRF approach are implementation simplicity and speed because it is computationally more efficient than other clustering techniques, such as SVMs [6] (which also motivated the choice of HC as clustering method). Further, unlike most RF algorithms (e.g., FRE [5] and Rocchio [4]), it does not modify the query or the similarity. As previously presented, the remaining retrieved images are simply clustered according to class label.

4. Experimental results

The validation of the proposed relevance feedback approach was conducted on several standard image databases, namely: Microsoft Object Class Recognition [16] which sums up to 4300 images distributed into 23 categories (e.g. animals, people, airplanes, cars, etc.) and Caltech-101 [12] 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 test 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 [13]: Color Histogram Descriptor, Color Layout Descriptor, Edge Histogram Descriptor and Color Structure Descriptors; color descriptors: Autocorrelogram, Color Coherence Vectors and Color Moments; and Speeded Up Robust Features - SURF descriptors [15] (represented through a Bag-of-Visual-Words model [14]).

The user feedback is to be automatically simulated from the known class membership of each image (ground truth is provided with the databases). Considering the experimental nature of the validation process, e.g. repetition of tests for different parameter settings, tests conducted for all the images in the database, this approach allows a fast and extensive simulation which could not be achieved involving real interaction, first of all due to the time constraints.

To assess the retrieval performance, we use several measures. First, we compute classical precision and recall. Precision is the fraction of retrieved documents that are relevant to the search (measure of false positives) and recall is the fraction of the documents that are relevant to the query that are successfully retrieved (measure of false negatives). The system retrieval response is assessed with the precision-recall curves which plots the precision for all the recall rates that can be obtained according to the current image class population. Second, to provide a global measure of performance we determine the overall Mean Average Precision - MAP as the area under the uninterpolated precision-recall curve (see also trec_eval scoring tool at `http://trec.nist.gov/trec_eval/` [2]). The

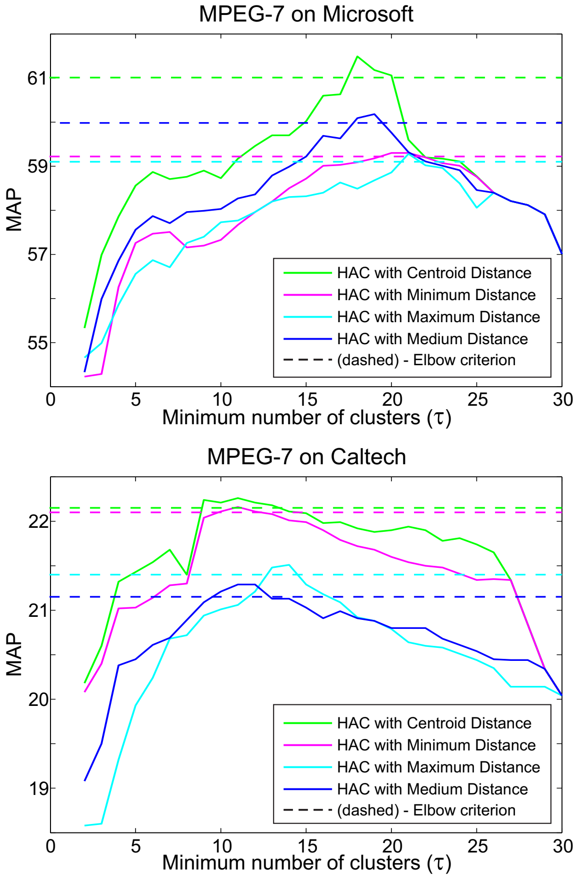

Figure 1. MAP against the minimum number of clusters (see Algorithm 1).

evaluation consists 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 browsing windows, ranging from 20 to 50 images. For brevity reasons, in the following we shall present only the best results which were obtain in the case of using 30 image windows.

In the first experiment we analyze the influence of the parameter τ (the minimum number of clusters which trigger the end of the relevance feedback training phase) and of the distance measure on the relevance feedback performance. Figure 1 plots the overall MAP obtained with the proposed RF in two situations, first linearly varying the number of clusters (depicted with solid line) and second using the "elbow criterion" (depicted with dashed line; we exemplify in the case of the MPEG-7 descriptors). One may observe that the adaptive choice of τ provide results very close to the highest MAP which can be achieved, but having a lower computational complexity than varying the number of clusters till the rich of the maximum MAP (which for large image sets is totaly inefficient). In terms of distance, we selected the centroid distance as it provides the greatest improvement in performance.

In the following, we compare our approach against other validated algorithms from the literature, namely: the Rocchio algorithm [3], Relevance Feature Estimation (RFE) [5], Support Vector Machines (SVM) [6], Decision Trees (TREE) [7], AdaBoost (BOOST) [8] and Random Forests (RF) [9]. Figure 2 presents the precision-recall curves after one iteration of feedback for different descriptor categories. Globally, all RF strategies provide significant improvement in retrieval performance compared to the retrieval without RF (see the dashed black line in Figure 2). Better performance is naturally obtained when targeting a more reduced number of image categories, i.e. on Microsoft data (only 23, compared to 102 in the case of Caltech-101).

However, the proposed hierarchical clustering RF tends to provide better retrieval performance in all cases (see solid black line in Figure 2). The highest increase in system performance is obtained using MPEG-7 descriptors on Microsoft database, increase of MAP from 30.21% (without RF) to 64.52%. At the other end, the smallest increase in performance is obtained for Caltech-101 database using SURF descriptors, namely increase of MAP from 10.90% (without RF) to only 18.44%. This is mainly due to the high diversity of classes for which the SURF descriptors provide little representative power compared to more general descriptors (MAP without RF is up to only 10.90%). In consequence, there are not sufficient positive feedback examples for the relevance feedback algorithms to work with.

In Figure 3 we present the variation of MAP with respect to the number of relevance feedback sessions for the best performance descriptor set on each database, namely color descriptors with Caltech-101 and MPEG-7 descriptors with Microsoft database. One may observe that the retrieval performance increases with each new feedback session. The best performance is still obtained with the proposed hierarchical RF, followed by Relevance Feature Estimation (RFE). For instance, at 4 feedback sessions, the largest increase of performance on Microsoft database is from MAP 30.21% (without RF) to 84.71% while on Caltech-101 is from 10.66% (without RF) to 55.78%. Compared to the RFE, the proposed hierarchical clustering RF provides an increase of MAP up to 6% on Microsoft database and 3% on Caltech-101, respectively (see dark red bar in Figure 3).

In the end, we present in Figure 4 several retrieval examples obtained without feedback and with the proposed RF algorithm (we exemplify using the color descriptors). The first image is the query image and the retrieval ranking is from left to right and top to bottom (for visualization purposes we limit the presentation to only the first 15 im-

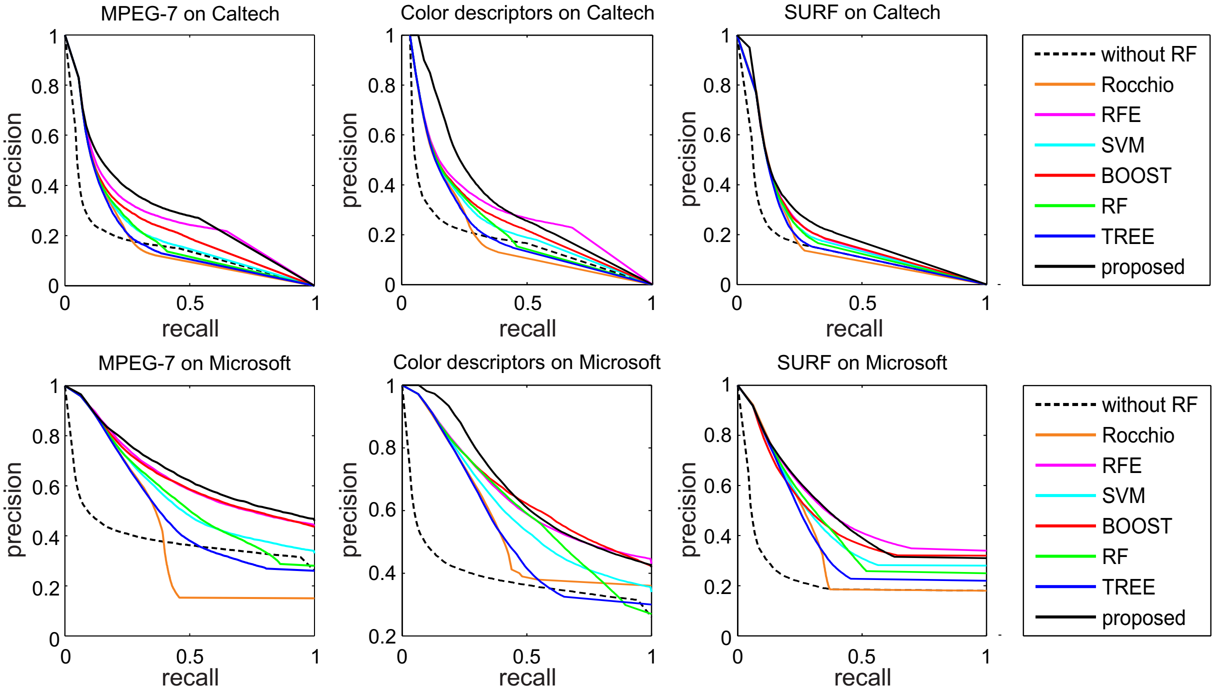

Figure 2. Precision-recall curves for different content descriptors and test databases.

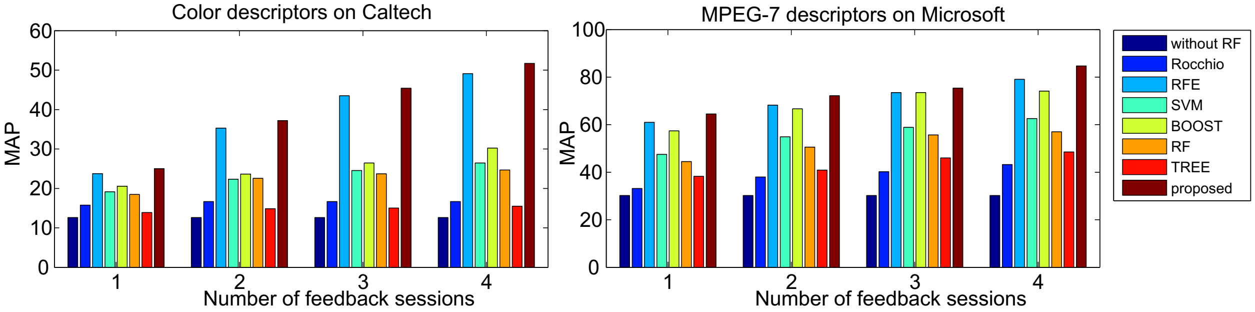

Figure 3. Mean Average Precision (MAP) variation with the number of feedback sessions.

ages). First example on Microsoft database shows that even though the initial response contain only three positive documents, after only one RF session all the returned images are correct. The second example presents a more difficult situation when the descriptors are not up to the query. However, even in this case, the use of RF doubles the retrieval performance (e.g. from 5 correct images up to 11 after RF).

5. Conclusions

We addressed relevance feedback techniques in the context of image retrieval and discussed a new approach which takes advantage of an hierarchical aggregative clustering scheme of the query results. The adaptation of the hierarchical clustering by the use of a cluster aggregation stopping criterion proves a good and reliable replacement to the heuristic, database-dependent choice of the fixed number of final clusters in the any of the relevant/non-relevant category partitions.

Tested on several standard databases (e.g. Microsoft Object Class Recognition and Caltech-101) and using several descriptor approaches (MPEG-7, color descriptors and SURF) the proposed approach drastically improve the retrieval performance, outperforming some other existing approaches, e.g. Rocchio, RFE, SVM, etc. Future improvements will mainly consist of fine tuning and adapt the method to address a higher diversity of image categories (e.g. Caltech-256, use of the Internet).

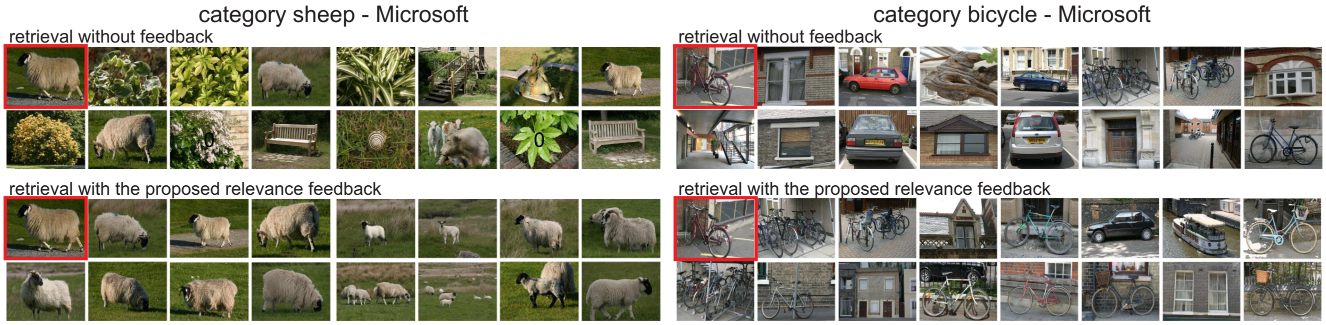
category sheep - Microsoft
retrieval without feedback
retrieval with the proposed relevance feedback
category bicycle - Microsoft
retrieval without feedback
retrieval with the proposed relevance feedback

Figure 4. Image retrieval without user feedback and with the proposed hierarchical RF.

Acknowledgments

Part of this work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013, Financial Agreement `POSDRU/89/1.5/S/62557` and `POSDRU/89/1.5/S/64109`.

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, vol. 22, no. 12, pp. 1349 - 1380, 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. Rocchio: ”Relevance Feedback in Information Retrieval”, in The Smart Retrieval System Experiments in Automatic Document Processing, G. Salton (Ed.), Prentice Hall, Englewood Cliffs NJ, pp. 313-323, 1971.

[4] 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.

[5] 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.

[6] S. Liang, Z. Sun, ”Sketch retrieval and relevance feedback with biased SVM classification,” Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

[7] S.D. MacArthur, C.E. Brodley, C.-R. Shyu, ”Interactive Content-Based Image Retrieval Using Relevance Feedback” , 12(1), 14-26. Computer Vision and Image Understanding 88, 55-75, 2002.

[8] 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), 14-26, 2006.

[9] Y. Wu, A. Zhang, ”Interactive pattern analysis for relevance feedback in multime information retrieval,” ACM Journal on Multimedia Systems, 10(1), pp. 41-55, 2004.

[10] W. J. Krzanowski, ”Principles of Multivariate Analysis: A User’s Perspective”, Clarendon Press, Oxford, 1993.

[11] R. Mojena ”Hierarchical grouping methods and stopping rules: An evaluation”, The Computer Journal 20 (4): 359-363, 1977.

[12] L. Fei-Fei, R. Fergus, P. Perona. ”Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories.” IEEE. CVPR, Workshop on Generative-Model Based Vision, 2004.

[13] J. M. Martinez: ”Standards - MPEG-7 Overview of MPEG-7 description tools, part 2”. IEEE MultiMedia, vol. 9, no. 3, 83-93, 2002.

[14] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, ”Evaluating bag-of-visual-words representations in scene classification,” International Workshop on Multimedia Information Retrieval, pp. 197-206, 2007.

[15] H. Bay, A. Ess, T. Tuytelaars, L. J. V. Gool, ”SURF: Speeded up robust features,” Computer Vision and Image Understanding, 110(3), pp. 346-359, 2008.

[16] Microsoft Object Class Recognition dataset, `http://research.microsoft.com/en-us/projects/objectclassrecognition/`.