Image Retrieval Using Relevance Feedback Based on Mann-Whitney Test

Sanjoy K. Saha
CSE Department
Jadavpur University
Calcutta 700 032
sks@becs.ac.in

Amit K. Das
CST Department
B. E. College (DU)
Sibpore, Howrah 711 103
amit@becs.ac.in

Bhabatosh Chanda
ECS Unit
Indian Statistical Institute
Calcutta 700 035
chanda@isical.ac.in

Abstract

In CBIR system, relevance feedback is used to improve the retrieval performance. In this paper, a new Mann-Whitney test based method is presented to identify the features which can distinguish relevant and irrelevant images in a retrieved set. Then the discriminating features are given more emphasis to improve retrieval performance. Proposed scheme is implemented for two different similarity measure (Euclidean distance based and Human Perception based). The effectiveness of the proposed methodology is established through experiment.

1. Introduction

In a CBIR system, the feature extraction module computes features of various types like shape, texture and colour for the images. The retrieval module retrieves the images similar to the query image from the database using a similarity measure based on the features. But the importance of the features vary for different queries and applications. Therefore, to achieve better performance, different emphases have to be given to different features. Some systems like FIDS [1] ask the user to choose the features and their weights. As a typical user does not have the basic knowledge of the feature extraction, he is unable to use the system effectively. Hence, a CBIR system must be able to decide about the emphasis/weight of the features on its own through a simple interaction with the user and the concept of relevance feedback (RF) comes into picture.

Relevance feedback, originally developed in [2], is a learning mechanism to improve the effectiveness of information retrieval systems. For a given query, the CBIR system retrieves a set of images according to a predefined similarity measure. Then, user provides a feedback by marking the retrieved images as relevant to the query or not. Based on the feedback, the system takes action and retrieves a new set.

The classical RF schemes can be classified into two categories: query point movement (query refinement) and re-weighting (similarity measure refinement) [3, 4]. In query point movement method, it is tried to improve the estimate of ideal query point by moving it towards the relevant examples and away from bad ones. Rocchio's formula [3] is frequently used to improve the estimation iteratively. In [5], a composite query is created based on relevant and irrelevant images. Various systems like WebSEEk [6], Quicklook [7], iPURE [8], Drawsearch [9] have adopted the query refinement principle. In the re-weighting method, the weight of the feature that helps in retrieving the relevant images is enhanced and importance of the feature that hinders this process is reduced. Rui et al. [10] and Squire et al. [11] have proposed weight adjustment technique based on the variance of the feature values. Systems like ImageRover [12], RETIN [13] use re-weighting technique.

MARS [14] uses both – the query refinement and a modified version of re-weighting method. Ishikawa et al. [15] have proposed a new distance measure and allow for correlation between attributes along with the weight adjustment. In [16], the system uses a learning technique based on Bayesian approach. PicSOM [17] retrieves a set of images against a set of reference images by creating Tree-structured self-organizing maps(TS-SOM) corresponding to different features and combines them according to user's preference. Su et al. [18] has integrated keywords (semantics) and low level features in their feedback mechanism. In [19], a neural network based system is described where nonlinear relation between the features is updated by using radial basis function. Yoo et al. [20] has described a system that uses feature level (shape, texture etc.) weight and component level weight in similarity measure. In order to update the weights, the query is to be executed for each feature type and the combined similarity metric. Thus, it becomes computationally expensive.

A close study of past work indicates that re-weighting technique is widely used. But, most of the systems address how to update the weight without identifying the good features. In this paper, we present a RF scheme, which first identifies the useful features following a non-parametric statistical approach and then updates their weights. The paper is organised as follows. Section 2 elaborates the proposed scheme. Description of the experimental system and result are presented in section 3. Finally, it is concluded in section 4.

2. Proposed scheme

In the proposed scheme, distance (similarity) measure is refined by updating the emphasis of the useful features. The term *useful feature* stands for the feature capable of discriminating relevant and irrelevant images within the retrieved set. The most crucial issue is to identify the useful features. Once, it is done then question arises how to adjust the emphasis.

2.1. Identification of useful features

Useful features are identified using Mann-Whitney test. In a two-sample situation where two samples are taken from different populations, Mann-Whitney test is used to determine whether the null hypothesis that the two populations are identical can be rejected or not.

Let, $X_1, X_2, \ldots, X_n$ be random samples of size n from population 1 and $Y_1, Y_2, \ldots, Y_m$ be the the random samples of size m from population 2. Mann-Whitney test determines whether X and Y come from the same population or not. It proceeds as follows [21]. X and Y are combined to form a single ordered sample and ranks 1 to $n + m$ are assigned to the observations from smallest to largest. In case of a tie ($i.e.$ if the sample values are equal), average of the ranks that would have been assigned in case of no ties are assigned. Based on the ranks, a test statistic is generated to check the null hypothesis. If the value of the test statistic falls within the critical region then the null hypothesis is rejected. Otherwise, it is accepted.

In CBIR systems, a set of images are retrieved according to a distance measure. Then, feedback is taken from the user to identify the relevant and irrelevant outcome. For the time being, let us consider only j^{th} feature and $X_i = dist(Q_j, f_{ij})$ where, Q_j is the j^{th} feature of the query image and f_{ij} is the j^{th} feature of the i^{th} relevant image retrieved so far. Similarly, $Y_i = dist(Q_j, f'_{ij})$ where f'_{ij} is the j^{th} feature of i^{th} irrelevant image. Thus, X_is and Y_is form the different random samples. Then, Mann-Whitney test is applied to judge the discriminating power of the j^{th} feature. Let $F(x)$ and $G(x)$ be the distribution function corresponding to X and Y respectively. The null hypothesis, H_0 and alternate hypothesis, H_1 may be stated as follows:

H_0: j^{th} feature cannot discriminate X and Y(X and Y come from same population) $i.e.$,

$$F(x) = G(x) \text{ for all x.}$$

H_1: j^{th} feature can discriminate X and Y(X and Y come from different population) $i.e.$,

$$F(x) \neq G(x) \text{ for some x.}$$

It becomes a two tailed test. Because, H_0 is rejected for any of the two cases: $F(x) < G(x)$ and $F(x) > G(x)$.

Physically, it can be interpreted that a useful feature can separate the two sets and X may be followed by Y or Y may be followed by X in the combined ordered list. Thus, if H_0 is rejected then j^{th} feature is taken to be as a useful feature. The steps are as follows:

1. Combine X and Y to form a single sample of size N, where $N = n + m$.
2. Arrange them in the ascending order
3. Assign rank starting from 1. If required, resolve ties.
4. Compute test statistic T as follows.

$$T = \frac{\sum_{i=1}^{n} R(X_i) - n \times \frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

   where, $R(X_i)$ denotes rank assigned to X_i and $\sum R_i^2$ denotes sum of the squares of the ranks of all X and Y.

5. If the value of T falls within critical region then H_0 is rejected and the j^{th} feature is considered as a useful one else not.

The critical region depends on the level of significance α which denotes maximum probability of rejecting a true H_0. If T is less than its $\alpha/2$ quantile or greater than its $1 - \alpha/2$ quantile then H_0 is rejected. In our experiment distribution of T is assumed to be normal and α is taken 0.1. If the concerned feature discriminates and places the relevant images at the beginning of the combined ordered list, then T will fall within the lower critical region. On the other hand, if the concerned feature discriminates and places the relevant images at the end of the same list then T will fall within the upper critical region.

It may be noted that, the proposed work proceeds only if the retrieved set contains both – relevant and irrelevant images. Otherwise, samples from two different populations will not be available and no feedback mechanism can be adopted.

2.2. Adjustment of the emphasis of features

Adjustment of the emphasis of feature is closely related with the distance/similarity measure adopted by the system.

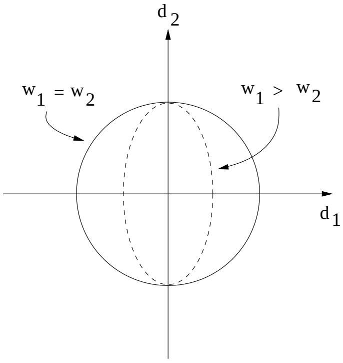

Figure 1. Variation of search space with weights of the features.

In the current work we have taken up two different similarity measures for discussion – Euclidean distance based measure and a human perception based similarity measure[22].

Euclidean distance is a widely used metric for CBIR systems. Let, an image is described by M features. Then, the distance between two images can be expressed as $\sum_{j=1}^{M} w_j d_j$ where, d_j denotes Euclidean distance between them with respect to j^{th} feature and w_j is the weight assigned to the feature.

In the proposed scheme, w_j is modified only if j^{th} feature is useful one. To explain the strategy for adjustment of weights of the features, let us consider a system that relies on two features only, say, f_1 and f_2. Difference in feature values between the query image and the database image are d_1 and d_2. With $w_1 = w_2$, the search space corresponding to Euclidean distance is a circle(as seen in figure 1 with solid line). Now suppose f_1 is a useful feature such that the test statistic of d_1 lies in the lower critical region. That means, f_1 can discriminate between relevant and irrelevant images, and the d_1 of the relevant images are, in general, less than d_1 of the irrelevant images. By making $w_1 > w_2$, search space is changed to ellipse(as seen in figure 1 with dashed line) and thereby discarding irrelevant images as much as possible from the retrieved set. Similarly, if f_1 is a useful feature and the test statistic of d_1 lies in the upper critical region then d_1 of relevant images are, in general, greater than d_1 of irrelevant images. Hence, by making $w_1 < w_2$, more relevant images can be included in the retrieved set. Thus by increasing the weight of the useful feature with lower test statistic, we try to exclude the irrelevant images from the retrieved set. On the other hand, by decreasing the weight of the useful feature with higher test statistic, we try to include the relevant images in the retrieved set.

Once images are retrieved, feedback is taken from the user and useful features are identified. Finally, weight adjustment is done according to the following steps:

1. Initialize all w_j to 1.
2. For each j^{th} useful feature where test statistic falls within lower critical region, set w_j as follows.

   $$w_j = w_j + \sigma_x^2$$

   where, σ_x^2 is the variance of X.
3. For each j^{th} useful feature where test statistic falls within upper critical region, set w_j as follows.

   $$\text{if } w_j > \sigma_x^2 \text{ then } w_j = w_j - \sigma_x^2$$

   where, σ_x^2 is the variance of X.
4. Repeat step 2 and 3 for successive iteration.

In [22], a human perception based similarity measure has been proposed. It says that two similar images may not match with respect to all the features. Based on this observation, it declares that two images are similar if at least K out of M features match. In this scheme, real valued feature vector is converted into a tag consisting of characters. Mapping the real value to character tag is done by dividing the feature space into a number of buckets through percentile ranking. For each feature, a character tag is assigned depending on the bucket into which it is mapped [22]. With respect to a feature, two images match if the corresponding characters in the tag lie within a tolerance range.

As in case of Euclidean distance based measure, here also useful features are identified following the same technique. But, the adjustment of emphasis of a feature is to be addressed in a slightly different manner. In this method, whether or not an image would be retrieved is decided by the count of matched features with the query image. Hence, updation of emphasis of feature must have a direct impact on feature matching. So that, irrelevant images are excluded and relevant ones are included by deploying the user feedback. It can be achieved by changing the match tolerance for the useful features. The basic principle is similar to the Euclidean distance based search. When similar images lie in the close vicinity of the query image in terms of the useful features, test statistic falls within lower critical region. In that case tolerance is reduced to restrict the inclusion of irrelevant images. The situation is reverse for the useful features with test statistic falling in the upper critical region. In that case, the similar images are lying in the distant buckets. Thus, to increase the possibility of inclusion of similar images the match tolerance is increased. The steps are as follows:

1. Initialize the tolerance for all features to t.

2. For all j^{th} useful features with test statistic in lower critical region
   set, $tolerance_j$ = $tolerance_j$ - 1
   If $tolerance_j <$ MIN then $tolerance_j =$ MIN

3. For all j^{th} useful features with test statistic in upper critical region
   set, $tolerance_j$ = $tolerance_j$ + 1
   If $tolerance_j >$ MAX then $tolerance_j$ = MAX

4. Repeat step 2 and 3 for successive iteration.

MIN and MAX denote minimum and maximum possible tolerance value. In our experiment, we have considered t as 2, MIN as 0 and MAX as B-1 where B is number of buckets in the feature space.

3. Experimental results

In our experiment we have used a collection of around 2000 images of various types like airplane, car, flower, animal and fish obtained from Corel database. Each image contains only one dominant object. Features are computed for the dominant object. Hence, a fast segmentation technique as described in [23] is used to extract the dominant object. Then, various shape, texture and colour features are computed. The feature vector is of 48 dimension of which 23 are shape features, 18 features denote texture and remaining 7 represent the color.

Projection method is an interesting technique for extraction of shape information. In our system, petal projection based various shape features are computed [23]. In this scheme, an object is divided into a number of petals where a petal is an angular strip area originating from the centre of gravity. Area of the object lying within a petal is taken as the projection along it. Finally, an 8 dimensional feature vector is obtained. Based on it, circularity, symmetricity, aspect ratio and concavity are computed. To supplement these features, another set of simple but effective measures of circularity, symmetricity etc. [23] are used in our system.

We have used a 15×15 texture co-occurrence matrix [24] to describe the texture feature. In order to compute the matrix, the intensity component of the colour image is divided into blocks of size 2×2 pixels. Then grey level pattern of the block is converted to a binary pattern by thresholding at the average value of the intensities. Decimal equivalent of the binary string formed from this pattern arranged in raster order gives the texture value. Thus we get the quantized image whose height and width are half of that of the original image and the pixel values range from 0 to 14. Finally, a grey level co-occurence matrix of size 15×15 is computed from this image. Based on this matrix, features like moments, energy, entropy, homogeneity etc. are computed.

In order to compute the colour feature, a hue histogram is formed based on HSV model. It is then smoothened and normalized. For each of the six major colours (red, yellow, green, $\ldots$, magenta) and grayness, index of fuzziness is computed as it has been outlined in [24].

To study the performance of proposed feedback scheme, each database image is used as the query image and an exhaustive search in the database is carried on. Queries are executed once using the Euclidean distance based similarity measure and again using the perception based similarity measure. In the latter case, each feature space is divided into 10 buckets and value of K is taken as 30. As the database is already groundtruthed, after retrieving a set of images they are automatically marked as relevant or irrelevant. Thus, feedback is automatically obtained. Hence, to prepare the recall precision graph, all the images retrieved to achieve a particular recall participate in feedback mechanism. Figure 2 show the recall precision graphs for Euclidean distance based search after third iteration of relevance feedback. measure. It is clear from the graphs that use of proposed scheme improves the performance. Moreover, increase in the time overhead for adopting the relevance feedback mechanism is very low.

Muller et al. [25] has mentioned that, from the perspective of a user, top order retrievals are of major interest. Secondly, in case of retrieval using perception based similarity measure, as it is quite likely that similar images may spread over multiple buckets, achievement of high recall is quite difficult. Hence, performance is studied based on top order retrievals. It may be noted that, for top order retrievals feedback can be directly taken from the user. Table 1 and table 2 show that proposed schemes are highly successful.

Table 1. Retrieval using Euclidean Distance based Similarity Measure (All figures indicate % match).

	No Relevance Feedback	Relevance Feedback		
		Iteration1	Iteration2	Iteration3
P(10)	76.16	77.91	79.61	81.40
P(20)	70.87	74.50	76.03	78.48
P(30)	68.05	69.89	71.38	72.63

4. Conclusion

In this paper we have presented two methods for upgrading emphasis of features to improve the performance of image retrieval system. Mann-Whitney test can identify the useful features which are capable of discriminating relevant and irrelevant images within a retrieved set. For Euclidean

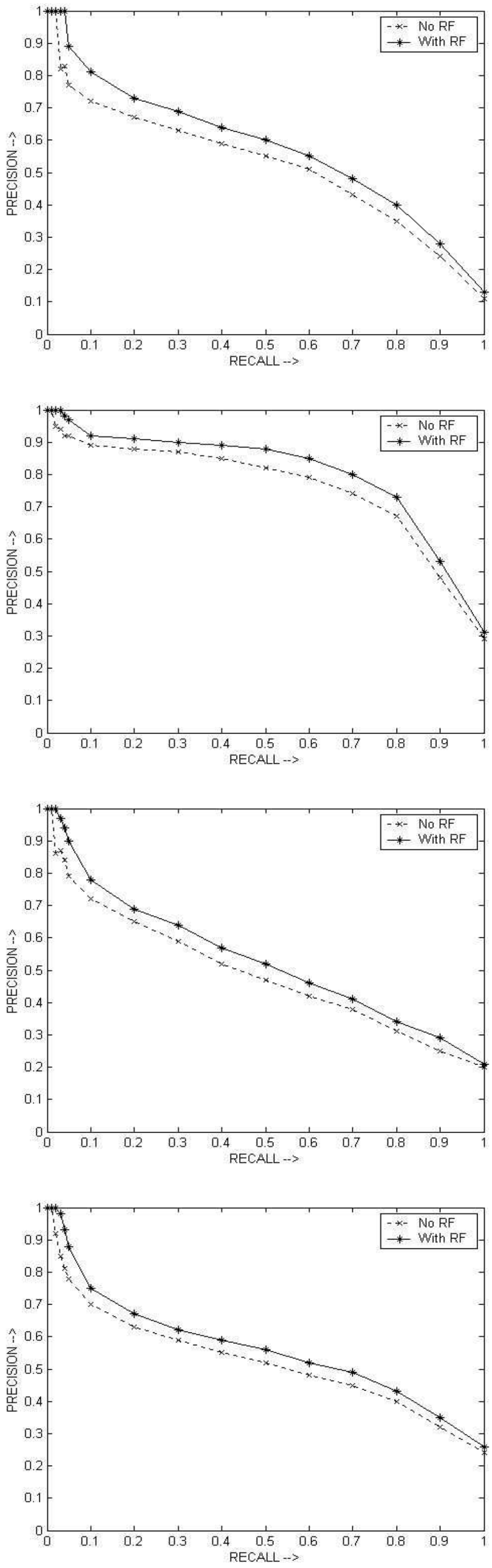

Figure 2. Recall-precision graph for different classes; they are (from top to bottom) Airplane, Car, Fish and Overall database.

Table 2. Retrieval using Perception based Similarity Measure (All figures indicate % match).

	No Relevance Feedback	Relevance Feedback		
		Iteration1	Iteration2	Iteration3
P(10)	81.10	87.39	89.32	91.17
P(20)	76.39	82.39	84.85	86.63
P(30)	73.15	78.61	81.34	83.20

distance based search, a weight upgradation scheme is proposed based on the variance of the features of the retrieved images. Following the similar principle, a tolerance updation scheme is proposed for human perception based similarity search. In both the cases, the schemes are found to be successful and effectiveness is established in the result.

References

[1] A. P. Berman and L. G. Shapiro, “A flexible image database system for content-based retrieval,” *Computer Vision and Image Understanding*, vol. 75, pp. 175–195, 1999.

[2] H. R. Turtle and W. B. Croft, “A comparison of text retrieval models,” *The Computer Journal*, vol. 35(3), pp. 279–290, 1982.

[3] J. J. Rocchio, “Relevance feedback in information retrieval,” in *The SMART Retrieval System: Experiments in Automatic Document Processing(G. Salton eds)*, pp. 313–323, Prentice Hall, 1971.

[4] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval for Image and Video Databases*. McGraw-Hill Book Company, 1983.

[5] J. Huang, S. R. Kumar, and M. Mitra, “Combining supervised learning with color correlogram for content-based retrieval,” in *5th ACM Intl. Multimedia Confernce*, pp. 325–334, 1997.

[6] J. R. Smith, *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Graduate School of Arts and Sciences, Columbia University, February 1997.

[7] G. Ciocca, I. Gagliardi, and R. Schettini, “Quicklook2: An integrated multimedia system,” *International Journal of Visual Languages and Computing, Special issue on Querying Multiple Data Sources Vol 12 (SCI 5417)*, pp. 81–103, 2001.

[8] G. Aggarwal, P. Dubey, S. Ghosal, A. Kulshreshtha, and A. Sarkar, "ipure: Perceptual and user-friendly retrieval of images," in *Proceedings of IEEE Conference on Multimedia and Exposition (ICME 2000)*, vol. 2, (New York, USA), pp. 693–696, July 2000.

[9] E. D. Sciascio, G. Mingolla, and M. Mongiello, "Content-based image retrieval over the web using query by sketch and relevance feedback," in *Visual Information and Information Systems, Proceedings of the Third International Conference VISUAL '99*, (Amsterdam, The Netherlands, June 1999, Lecture Notes in Computer Science 1614, Springer), pp. 123–130, 1999.

[10] Y. Rui, T. S. Haung, S. Mehrotra, and M. Ortega, "Relevance feedback: A power tool in interactive cotent-based image retrieval," *IEEE Tran. on Circuits and Systems for Video Technology, Special issue on interactive Multimedia Systems for the internet*, vol. 8(5), pp. 644–655, Sept. 1998.

[11] D. M. Squire, W. Muller, H. Muller, and T. Pun, "Content-based query of image databases: inspirations from text retrieval," *PRL*, vol. 21, pp. 1193–1198, 2000.

[12] S. Sclaroff, L. Taycher, and M. L. Cascia, "Imagerover: A content-based image browser for the world wide web," in *IEEE Workshop on content-based Access of Image and Video Libraries*, (San Juan, Puerto Rico), pp. 2–9, 1997.

[13] J. Fournier, M. Cord, and S. Philipp-Foliguet, "Retin: A content-based image indexing and retrieval system," *Pattern Analysis and Applications*, vol. 4, pp. 153–173, 2001.

[14] T. P. Minka and R. W. Picard, "Interactive learning using a society of models," *PR*, vol. 30(4), 1997.

[15] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Query databases through multiple examples," in *24th VLDB Conference*, (New York), 1998.

[16] I. J. Cox, M. L. Miller, T. P. Minka, T. Papathomas, and P. N. Yianilos, "The bayesian image retrieval system, pichunter: Theory, implementation and psychophysical experiments," *IEEE Transactions on Image Processing*, vol. 9(1), pp. 20–37, 2000.

[17] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja, "Picsom - content-based image retrieval with self-organizing maps," *PRL*, vol. 21, pp. 1199–1207, 2000.

[18] Z. Su and H. Zhang, "Relevance feedback in cbir," in *VDB6 Conf.*, pp. 21–35, 2002.

[19] H. K. Lee and S. I. Yoo, "Nonlinear combining of heterogeneous features in content-based image retrieval," *Intl. Journal of Computer Research*, vol. 11(3), 2002.

[20] H.-W. Yoo, D.-S. Jang, S.-H. Jung, J.-H. Park, and K.-S. Song, "Visual information retrieval system via content-based approach," *PR*, vol. 35, pp. 749–769, 2002.

[21] W. J. Conover, *Practical nonparametric statistics, 3rd edition*. New York: John Wiley and sons, 1999.

[22] S. K. Saha, A. K. Das, and B. Chanda, "An efficient search technique for cbir systems," in *3rd Intl. workshop on content based multimedia indexing*, (France), Sept., 2003.

[23] S. K. Saha, A. K. Das, and B. Chanda, "Graytone image retrieval using shape feature based on petal projection," in *ICAPR 2003*, (India), pp. 252–256, 2003.

[24] S. K. Saha, A. K. Das, and B. Chanda, "Cbir using perception based texture and colour measures," in *ICPR 2004*, (Cambridge,UK), Aug., 2004.

[25] H. Muller, W. Muller, S. Marchand-Mallet, T. Pun, and D. M. Squire, "Automated benchmarking in content-based image retrieval," in *ICME 2001*, (Tokyo, Japan), pp. 22–25, 2001.