ISSN: 2321-7782 (Online)
Volume 4, Issue 4, April 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Similarity Measures through Histogram Based Image Retrival

O. Koteswara Rao[1]
Assoc Professor
St. Mary's Group Of Institutions Guntur
Chebrole(V&M),Guntur(Dist), Andhra Pradesh – India

G. Venkateswara Rao[2]
Assoc Professor
St. Mary's Group Of Institutions Guntur
Chebrole(V&M),Guntur(Dist), Andhra Pradesh – India

Abstract: Similarity measure is an essential consideration for a success of many methods. similarity measures are analyzed in the context of ordered histogram type data, such as gray-level histograms of digital images or color spectra. The performance of the studied similarity measures can be improved using a smoothing projection, called neighbor-bank projection. distance functions utilizing statistical properties of data, e.g., the Mahalanobis distance. The main idea behind the smoothing projection is a technique called as IHBM(Integrated Histogram Bin Matching).So in order to improve the similarity measures of ordered histograms ,the image should be first converted from RGB into HSV, and then calculate the Ordered Histogram values, from those we calculate the similarity measurements. These measurements are compared with different image values that are already stored in the database and finally the retrieved matched values are displayed as output result. The proposed projection method seems also to be applicable for dimensional reduction of histograms and to represent sparse data in a more tight form in the projection subspace.

Keywords: Ordinal histograms, Distance functions, Image Retrieval, Similarity Measures, Histogram based Image Retrieval.

I. INTRODUCTION

Many distance functions that can be used to measure similarities, distances, between most types of features. The two most common ones are the Euclidean and Mahalanobis istances. Our motivation is to study different distance functions for measuring similarity of ordered histograms. An ordered histogram is a histogram where adjacent bins contain related information. A priori information of ordered bins can be used to construct a more robust similarity measure by combining information from neighboring bins. There is no generic method for selecting a similarity measure or a distance function. However, a priori information and statistics of features can be used in selection or to establish a new measure (Aksoy and Haralick, 2001[10]; Hafner et al.,1995 [12]; Jin and Kurniawati, 2001; Mitra et al., 2002;Sebe et al., 2000). In practice, a similarity measure is often an underlying property of an algorithm ,and thus, the use of a measure is implicit. Still, the role and meaning of selecting a proper similarity measure in any algorithm should not be neglected. The accuracy of the most common distance functions, such as Euclidean, can be significantly improved if a priori information of the features is used. Our motivation is to study different distance functions for measuring similarity of ordered histograms. An ordered (also called ordinal, Cha and Srihari (2002))[11] histogram is a histogram where adjacent bins contain related information, for example a gray-level histogram or a color spectrum (wavelength distribution of light intensity).

II. MOTIVATIONS EXISTING METHODS

This section presents some of the popularly existing histogram similarity measures [1, 2, 3], namely, Histogram Intersection (HI), Histogram Euclidean Distance (HED) and Histogram Quadratic Distance Measures (HQDM).These following existing methods are studied in detail.

© 2016, IJARCSMS All Rights Reserved

2.1 Histogram Intersection (HI)

Histogram Intersection [2, 3] is for color image retrieval and to find known objects within images using color histograms,

$$D_{HI}(q,t) = \sum_{i=0}^{M-1} | h_q(i) - h_t(i) |$$

Where $D_{HI}(q,t)$ is the distance between query image q and target image t, and h_q and h_t are the color histograms of query and the target images respectively and m is the number of bins of histogram.

2.2 Histogram Euclidean Distance (HED)

The Euclidean distance [2, 3] is, as follows: given histograms h_q and h_t

$$D_{HED}(q,t) = (h_q - h_t)^T (h_q - h_t) = \sum_{i=0}^{M-1} (h_q(i) - h_t(i))^2$$

$D_{HED}(q,t)$ is the distance between query image q and target image t, and h_q and h_t are the color histograms of query and the target images respectively, moreover, M is the number of bins of histogram.

The Figure.1 shown below represents the Minkowski distance measures stated above.

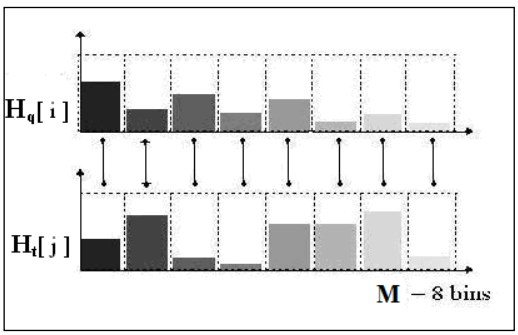

Figure 1. The Minkowski distance measures

2.3 Histogram Quadratic Distance Measures (HQDM)

A Histogram Quadratic Distance Measure is used in IBM QBIC system for color histogram based image retrieval [1 , 2 , 3]. In [3], it is reported that quadratic distance metric between color histograms provides more desirable results than"like-bins" that are only comparisons between color histograms. The quadratic form distance between histograms h_q and h_t given by

$$D_{HQDM}(q,t) = (h_q - h_t)^T A (h_q - h_t)$$

Where $D_{HQDM}(q,t)$ is the distance between query image q and target image t, and h_q and h_t are the color histograms of query and the target images respectively and $A = [a_{ij}]$ and a_{ij} denotes the similarity between image histograms with bins i and j. The Quadratic form metric is a true distance metric when $a_{ij} = a_{ji}$ and $a_{ii} = 1$.

The HQDM is computationally more expensive than the Minkowski form metrics since it computes the cross similarity between all histogram bins as shown in Figure.2.

© 2016, IJARCSMS All Rights Reserved

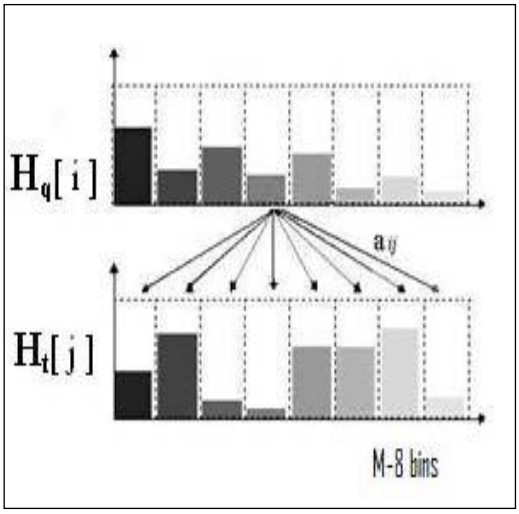

Figure 2. The quadratic distance measure

III. Similarity Measures for Ordered Histogram

In this section, a set of similarity measures is proposed for ordered histograms. An ordered histogram or distribution is a histogram where adjacent histogram bins contain related information. For example, in a gray-level histogram the neighbor dimensions represent pixel intensity values that are almost the same. In many natural smooth distributions, such as in color spectra, the same characteristics are present. Similarity measures for ordered histograms can be built upon common distance functions, but by utilizing a priori information by smoothing projections the results of similarity measures can be improved.

3.1. Distance functions

Most commonly used distance functions are shown in Fig. 3 for two feature vectors p= ($p_{0,} \ldots, p_{L-1}$) and q =($q_{0,} \ldots, q_{L-1}$). The most familiar distance functions are the Euclidean and Manhattan distances induced by L2 and L1 norms, respectively. They are both special cases of the Minkowsky distance (Lp). As the degree of the norm (p) increases, the weight of large differences between single attribute values increases. Both, the Euclidean and Manhattan, distances are calculated separately for each dimension, and thus, they are not good measures for similarity between two histograms, where attributes are correlated and ordered. For ordinal data, the cumulative Euclidean and land mover distances can be used. The cumulative Euclidean and land mover distances measure the spatial concentration of the values in the feature vector and the order of the feature attributes affects the value of the distance. Therefore, with ordered data these measures are likely to provide better results than the standard Euclidean and Manhattan distances.

Previous distance functions consider the attributes to be non-correlated. For cross-correlated attributes, statistical properties of the data set can be used to reduce the effect of the correlations. One distance function with such a statistical factor, the correlation matrix, is the Mahalanobis distance.

Calculation of the correlation matrix in the Mahalanobis distance needs quite much data, the exact amount depends on the length of the feature vectors and the variation between dimensions. If there is no enough data or no enough variation in the data, numerical calculation of the inverse of the covariance matrix may become an ill-posed problem. A more comprehensive study of the Mahalanobis distance with a limited sample set size can be found in (Takeshita et al., 1993)[13]. Another statistical method, the log likelihood ratio G, measures the degree that observed data fits to an expected distribution (Sokal and Rohlf, 1969)[14].

© 2016, IJARCSMS All Rights Reserved

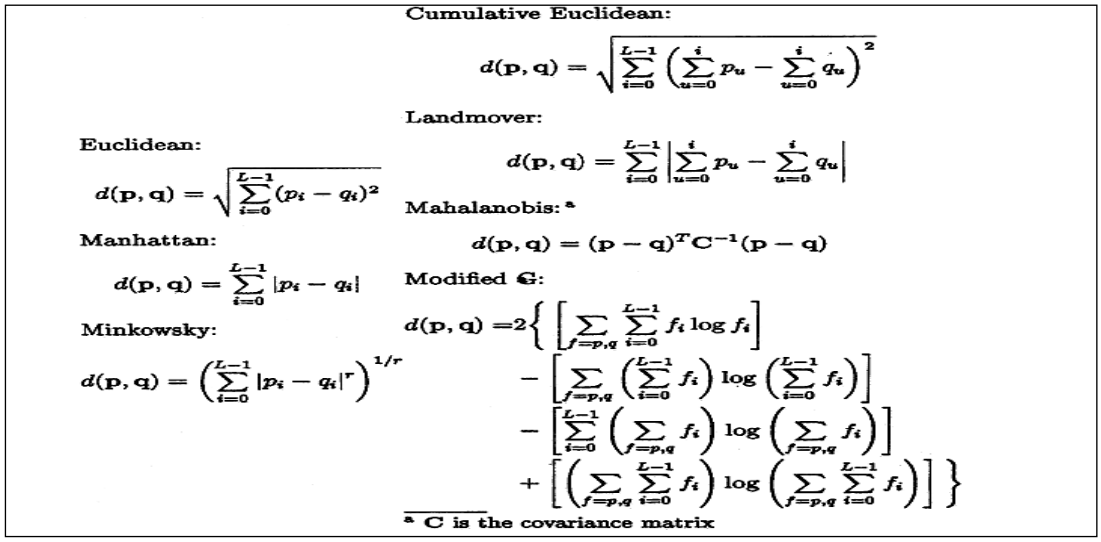

Fig. 3. Common distance functions

3.2. Smoothing projections

New similarity measures can be introduced based on the previous distance functions and a priori information concerning ordered histograms. Because in an ordered histogram the closely situated elements correlate more strongly than elements which are further apart, a feature vector can be projected to a smaller number of dimensions without significant loss of information. This kind of smoothing projection together with any distance function induces a new similarity measure.

In the case of statistical methods, such as the Mahalanobis distance, the statistical properties may be more evident in the projected space. A hypothesis is made that using a smoothing projection, the statistical properties of the samples are more evident and the similarity measure is improved (Kamarainen et al., 2001)[15].

Dimensionality of a histogram is reduced by a linear projection to a subspace, called the neighbor-bank subspace. A set of discrete sampled $\cos^2$ functions can be used to form the neighbor-bank subspace. Histograms are projected on a set of cos2 functions (see Fig. 2). Let L be the length of an original histogram $p= (p_{0,} \ldots , p_{L-1})^T$. N be the number of banks indexed with k from 0 to N-1.

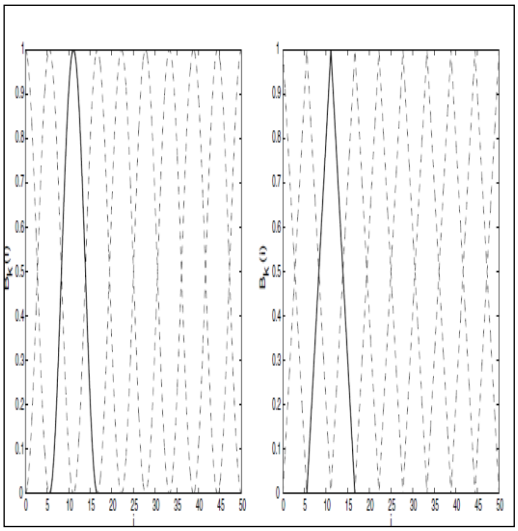

Fig. 4. N = 10 neighbor-banks of cos2 and triangle functions for discrete histograms of size L = 50 (the third neighbor-bank is highlighted).

© 2016, IJARCSMS All Rights Reserved

Then for B_K (i) the discrete neighbor-banks of $\cos^2$ function can be constructed as

$$B_k(i) = \begin{cases} \cos^2\left(\pi\left(\frac{i}{L}\frac{N-1}{2} + \frac{k}{2}\right)\right) & \text{if } \frac{L}{N-1}(k-1) \\ & \leqslant i \leqslant \frac{L}{N-1}(k+1), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

By constructing a transformation matrix

$$\boldsymbol{B} = \begin{pmatrix} B_0(0) & B_0(1) & \cdots & B_0(L-1) \\ B_1(0) & B_1(1) & \cdots & B_1(L-1) \\ \vdots & \vdots & \ddots & \vdots \\ B_{N-1}(0) & B_{N-1}(1) & \cdots & B_{N-1}(L-1) \end{pmatrix}, \quad (2)$$

The projection can be performed by matrix multiplication as

$$\boldsymbol{r} = \boldsymbol{B}\boldsymbol{p}, \quad (3)$$

Where r is the projection.

$$\mathbf{d(x, x)= 0;} \quad \mathbf{(4)}$$

$$\mathbf{d(x, y)= d(y, x)\ (commutativity)} \quad \mathbf{(5)}$$

$$\mathrm{d}(x,z) \leqslant \mathrm{d}(x,y) + \mathrm{d}(y,z) \quad \text{(triangle inequality)}. \quad (6)$$

Because $\mathrm{d}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is a metric in R, it satisfies (4)–(6), and in addition

$$\mathrm{d}(x,y) = 0 \Longleftrightarrow x = y. \quad (7)$$

Next, the properties of $\mathrm{d}(B\cdot, B\cdot)$ are inspected in P. The first condition (4) is satisfied for $\mathrm{d}(\boldsymbol{Bp}_1, \boldsymbol{Bp}_2)$ since

$$\mathrm{d}(\boldsymbol{Bp}, \boldsymbol{Bp}) = \mathrm{d}(\boldsymbol{r}, \boldsymbol{r}) = 0 \quad \text{by (7)}. \quad (8)$$

It should be noted that for a pseudo-metric it is also possible that $\mathrm{d}(\boldsymbol{Bp}_1, \boldsymbol{Bp}_2) = 0$ for $\boldsymbol{p}_1 \neq \boldsymbol{p}_2$. The second condition (5),

$$\begin{aligned} \mathrm{d}(\boldsymbol{Bp}_1, \boldsymbol{Bp}_2) &= \mathrm{d}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \mathrm{d}(\boldsymbol{r}_2, \boldsymbol{r}_1) \\ &= \mathrm{d}(\boldsymbol{Bp}_2, \boldsymbol{Bp}_1). \end{aligned} \quad (9)$$

Triangle inequality also holds, as for any $\boldsymbol{Bp}_1$, $\boldsymbol{Bp}_2$, $\boldsymbol{Bp}_3 \in R$ where $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3 \in P$,

$$\begin{aligned} \mathrm{d}(\boldsymbol{Bp}_1, \boldsymbol{Bp}_2) &= \mathrm{d}(\boldsymbol{r}_1, \boldsymbol{r}_2) \leqslant \mathrm{d}(\boldsymbol{r}_1, \boldsymbol{r}_3) + \mathrm{d}(\boldsymbol{r}_3, \boldsymbol{r}_2) \\ &= \mathrm{d}(\boldsymbol{Bp}_1, \boldsymbol{Bp}_3) + \mathrm{d}(\boldsymbol{Bp}_3, \boldsymbol{Bp}_2). \end{aligned} \quad (10)$$

After the projection of data to the subspace spanned by (B_0 (i), . . . , B_{N-1} (i)), any standard distance metric can be used. The advantage of using $\cos^2$ functions is that the sum over the banks is 1 over the whole interval of i (see Fig. 4). All attributes are thus equally weighted, although the property of equal weighting is not mandatory. A set of triangle functions also provides the same property of equal weighting. For triangle functions the neighbor-banks can be constructed from

$$B_k(i) = \begin{cases} 1 - \frac{N-1}{L}\left|i - k\frac{L}{N-1}\right| & \text{if } \frac{L}{N-1}(k-1) \\ & \leqslant i \leqslant \frac{L}{N-1}(k+1), \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

© 2016, IJARCSMS All Rights Reserved

Furthermore, in neural network classifiers, dimensionality of input must often be reduced, which may be performed using, for example, the PCA transform (Haykin, 1994). The neighbor bank projection could also be considered as a dimensional reduction method which can be used instead of the PCA transform. However, information loss in the dimensional reduction is small only when the a priori assumption of ordered histograms holds.

3.3. Histogram Similarity Example

Let us consider the three histograms on left in Fig. 5. The distances from histograms 2 and 3 to histogram 1a re calculated using the previously defined distance functions. Visually histogram 2 seems to be much more similar to histogram 1than to histogram 3. However, using the previously described distance metrics the similarity is not that obvious. The distances between the original histograms in Fig. 4 are shown in Table 1. In addition to the distances, the ratio between distances to histograms 2 and 3 is shown. The greater the ratio, the better the ability of the corresponding metric in discriminating the histograms. The Euclidean, Manhattan and G-statistics distance functions produce exactly the same distance to histograms 2 and 3. This is because they operate separately in different dimensions and the a priori information of neighbor correlations remains totally unused. On the other hand the cumulative Euclidean and landmover distances show a significant difference between the distances. They both measure the overall similarity in the shape of the histograms and thus histogram 2 is measured as being much closer to histogram 1. Next, the number of dimensions of the histograms is reduced by projecting them to the set of 10 $\cos^2$ functions shown in Fig. 3. The result can be seen on right in Fig. 4. The same distances calculated in the smoothing neighbor-bank subspace are also shown in Table 1. Now, the results for all distance metrics are as expected and histogram 1 is measured to be closer to histogram 2 than to histogram 3.

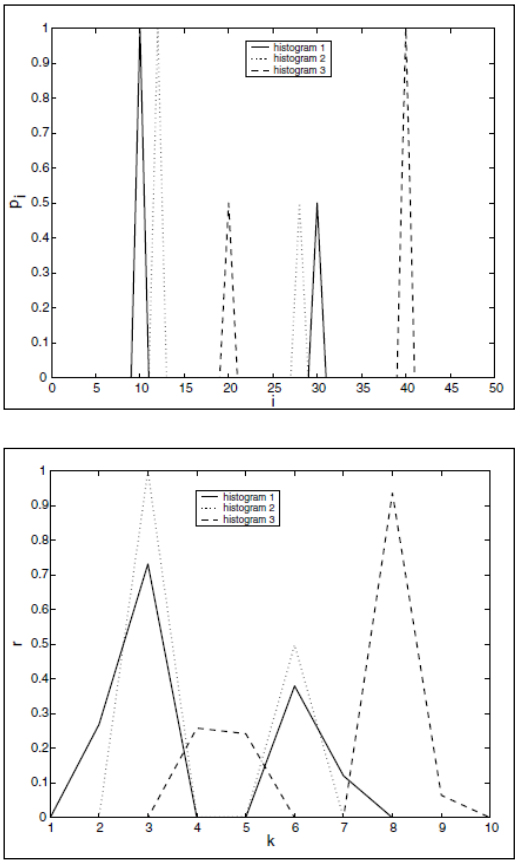

Fig. 4. Original values of example histograms 1, 2, and 3 of L = 50 discrete values and the projected histograms on the N = 10 neighbor-banks.

© 2016, IJARCSMS All Rights Reserved

Table 1
Distances from histogram 1 to histograms 2 and 3 for original data and projected data

Metric	Original			Projected		
	Histogram 2	Histogram 3	Ratio	Histogram 2	Histogram 3	Ratio
Euclidean	1.581	1.581	1.00	0.4152	1.331	3.21
Manhattan	3.000	3.000	1.00	0.7783	3.000	3.85
Landmover	2.000	16.67	8.34	0.2615	2.970	11.36
Cumulative Euclidean	1.111	10.00	9.00	0.0385	1.623	42.16
G-statistic	0.4405	0.4405	1.00	0.0314	0.3295	10.49

3.4 IHBM Smoothing Projection

The three main steps of the IHBM method is given below

1) Conversion of RGB space into HSV space for Quantization.

2) Compute the inter-bin distances matrix HISTd (Q, T) between all pairs of images.

Where HISTd (Q, T) satisfies the monge property as given in equation 1 , Q is query image and T is target image.

3) Computation of similarity measure using the proposed approach IHBM.

3.4.1 HSV Color Space

The determination of the optimum color space is an open problem, certain color spaces have been found to be well suited for the content-based query-by-color. The proposed method used HSV(Hue, Saturation and Value) Color space, because it is natural and is approximately perceptually uniform.

3.4.2 HSV Quantization

HSV Quantization gives 18 hues, 3 saturations, 3 values, and 4 gray levels, which results 166 bins [3 , 4] for each image. Then color histogram is computed for 166 bins, and then it is normalized.

3.5 Distance between histogram bins

To compute the distance between a bin pair, HISTd (Qi, Tj) is determined by the color characteristics of the histogram bins[4]. HISTd (Qi, Tj) can be computed a priori, independent of the Query image and target images. A Monge distance matrix $D_{Q,T}$ is computed from the HISTd (Qi, Tj) which is constant[5]. This distance matrix satisfied

Monge condition i.e. $m \times m$ matrices $D_{Q,T} = [d_{i,j}]$ which fulfill the so-called Monge property[6] given in equation 12.

$$d_{i,j} + d_{i+1,j+1} \leq d_{i,j+1} + d_{i+1,j} \quad \text{.....(12)}$$

Where 1 < I < m , 1 < j < m

Distance matrix DQ,T satisfies the discrete Monge condition. Then Hoffmann [5] pointed out that greedy approach gives an optimal solution.

3.5.1 Integrated Histogram Bin Matching (IHBM)

IHBM (Integrated Histogram Bin Matching), is a novel metric Similarity measure to compare the color feature of quantized images. The main idea of this, consists of modeling the comparison of color-quantized images as a Transportation problem

© 2016, IJARCSMS All Rights Reserved

[5,6,7,8].This model deals with the determination of a minimum-cost plan for transporting a commodity from a number of supply nodes to a number of demand nodes.

At the time of the Partition the nodes are divided into two sets m and n, where nodes in m are supply nodes and nodes in n are demand nodes, and for each arc (i, j), i is in m and j is in n. Let Z denote total transportation cost, let x_{ij} denote the no. of units shipped from supply node i to demand node j, and cij denote the cost of shipping a unit shipped from supply node i to demand node j. The general form of the Transportation problem is then

$$\text{Minimize} \quad Z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$
$$\sum_{j=1}^{n} x_{ij} = s_i \quad \text{for all } i = 1,2,\ldots, m$$
$$\sum_{i=1}^{m} x_{ij} = d_j \quad \text{for all } j = 1,2,\ldots, n$$
$$x_{ij} \geq 0 \quad \text{for all } i \text{ and } j$$

Where s_i denotes Supply Constraints and d_j denotes demand Constraints

For matching histogram bins of two images, the closest histogram bin pair is considered first. If the bins are of the same size then the two most similar bins are matched otherwise a partial match occurs. This process is repeated until all the histogram bins are matched completely. After matching histogram bins, the similarity measure is computed as a weighted sum of the similarity between histogram bin pairs, with weights determined by the matching scheme. This is known as **Integrated Histogram Bin Matching (IHBM),** which emphasizes the integration of histogram bins in the retrieval process. The Figure.5 represents the similarity measure mechanism of the proposed IHBM approach with 8 bins.

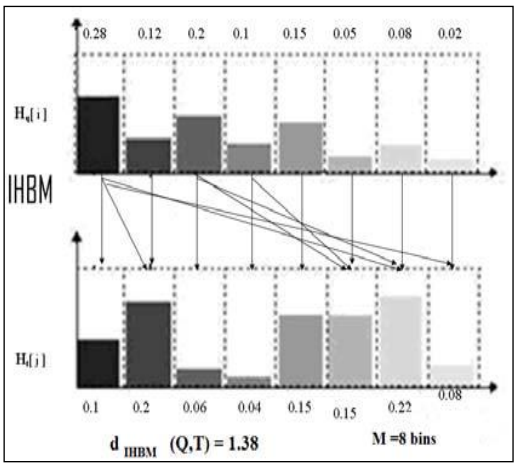

Fig. 5. IHBM approach

3.5.2 IHBM Algorithm

BEGIN

1 $D_{Q,T}$ = HISTd (Q, T)

2 Detect and compute the Monge sequence

$$d\,i,j + d\,i+1, j+1 \leq d\,i+1, j + d\,i, j+1$$

3 for each pair of histogram Bins Qi Є Q and T_j Є T

4 Qi .status = 0

5 Tj .status = 0

6 sort out the computed distances DQ,T in non-decreasing order

© 2016, IJARCSMS All Rights Reserved

7 DIHBM = 0

8 for each distance DQ,T in non- decreasing order

9 if Qi . status = Tj . status = 0

10 if Qi . size < Tj . size

11 w = Qi . size

12 Tj . size = Tj . size – w

13 Qi . status = 1

14 else

15 w = Tj. size

16 Qi . size = Qi . size - w

17 Tj . status = 1

18 if Qi . size = 0 then Qi . status = 1

19 DIHBM = DIHBM + w × DQ,T

20 END

IV. Proposed System Architecture

The system architecture for improvement of Histogram based image retrieval is shown in figure 6.

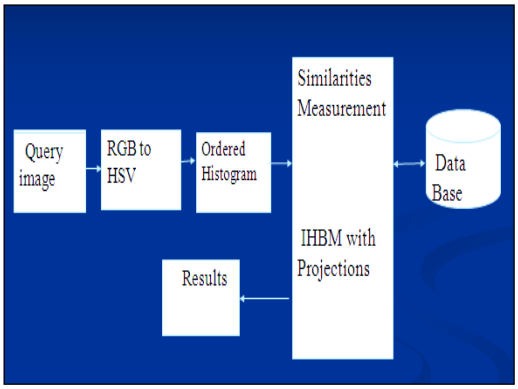

Fig. 6. Improvement of Histogram based Image Retrieval

V. Conclusion

In this paper, the properties of various distance functions were examined primarily in the context of ordered histogram type data. A new smoothing projection, the neighbor-bank projection, IHBM Smoothing was also introduced. The smoothing projection seems to improve the accuracy of some of the studied distance functions and to have advantages when combined with methods utilizing the statistical properties of the data, such as the Mahalanob distance and the Bayesian classifier. Our new smoothing technique like IHBM is experimented on 1000 color images and the experimental results with the help of tables and graphs clearly indicate the proposed method IHBM is more accurate and efficient than the three existing methods i.e. HI, HED and HQDM. The proposed method is proved as metric, which satisfies non-negativity, commutative and triangle inequality properties.

© 2016, IJARCSMS All Rights Reserved

References

1. Hafner, J., Sawhney, H.S., Equitz, W., Flickner, M., Niblack,W., 1995. Efficient color histogram indexing for quadratic form distance functions. IEEE Trans. Pattern Anal.Machine Intel. 17 (7), 729–736.Jou, F.D., 2003.
2. J.R.Smith and S.-F Chang,Integrated spatial and feature image query, Multimedia Syst. 7(2) ,129-140 (1999).
3. J. R. Smith. Integrated Spatial and Feature Image Systems: Retrieval, Analysis and compression . PhD thesis, Columbia University, New York,NY, February 1997.
4. John R. Smith and Shih-Fu Chang,"Tools and Techniques for Color Image Retrieval" Columbia University Department of Electrical Engineering and Center for Telecommunications Research New York, N.Y. 10027.
5. Rainer E. Burkard *, Bettina Klinz Riidiger Rudolf " Perspective of Monge properties in optimization"- Discrete Applied Mathematics 70 (1996) 95-161.
6. J. Li, J.Z. Wang, G. Wiederhold, IRM: integrated region matching for image retrieval, in:Proceedings of the Eighth ACM Multimedia Conference, 2000, pp. 147–156.
7. Yossi Rubner, Carlo Tomasi and Leonidas J. Guibas, The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision 40(2), 99–121, 2000.
8. Hamdi A. Taha. Operations Research. Prentice Hall, 1982.
9. svetlozar T.rachev "Mass transportation problems-vol1:theory", ,springer.
10. Aksoy, S., Haralick, R.M., 2001. Feature normalization and likelihood-based similarity measure for image retrieval. Pattern Recognition Lett. 22, 563–582.
11. Cha, S.-H., Srihari, S.N., 2002. On measuring the distance between histograms. Pattern Recognition 35, 1355–1370.
12. Hafner, J., Sawhney, H., Equitz, W., Flickner, M., Niblack, W., 1995. Efficient color histogram indexing for quadratic form distance functions. IEEE Trans. Pattern Anal. Machine Intell. 17 (7), 729–736.
13. Takeshita, T., Nozawa, S., Kimura, F., 1993. On the bias of Mahalanobis distance due to limited sample size effect. In: Proc. Second Internat. Conf. Document Anal. Recognition, pp. 171–174.
14. Sokal, R.R., Rohlf, F.J., 1969. Biometry. W.H. Freeman, New York.
15. Kamarainen, J.-K., Kyrki, V., K€alvi€ainen, H., June 2001. Similarity measures for ordered histograms. In: Proc. 12th Scand. Conf. Image Anal. SCIA2001, Bergen, Norway, pp. 699–705.

© 2016, IJARCSMS All Rights Reserved