

# An Empirical Framework to Evaluate Performance of Dissimilarity Metrics in Content-Based Image Retrieval Systems

Hetunandan M. Kamichetty, Pradeep Natarajan  
Dept. of Computer Science  
Indian Institute of Technology, Madras  
Chennai - 600036, India  
{khetu, npradeep}@peacock.iitm.ernet.in

Subrata Rakshit  
Center for Artificial Intelligence and Robotics  
Raj Bhavan Circle, High Grounds  
Bangalore - 560001, India.  
subrata@cair.res.in

**Abstract**—The dissimilarity metric used by a Content Based Image Retrieval (CBIR) system greatly affects its performance. Due to the large number of low level features used to represent each image in the database and the curse of dimensionality, only an empirical evaluation of the performance of metrics is possible. We present an experimental framework for comparing the performance of various distance metrics on a given feature set. This framework can be used to determine the best distance metric for any CBIR system. We use this framework to decide the best metric among thirteen candidate metrics for the CBIR system that we have developed.

**Index Terms**—Dissimilarity Metrics, Image Retrieval, CBIR, Image Match Measures.

## I. INTRODUCTION

WITH the expansion of the Internet there has been a rapid increase in the multimedia content available. Many of these databases contain huge collections of images of a generic nature. The problem of indexing and retrieving these images is extremely difficult and there is a need for developing automated indexing schemes. Two popular methods for querying by example in CBIR systems are querying-by-image and querying-by-sketch. Querying-by-sketch has many drawbacks [1] due to which query-by-image is generally preferred. In this method, the user's query is represented in the form of example images (exemplars) and the CBIR system retrieves images similar to these from the database. The results are evaluated by the user and if necessary, the process is repeated after addition to the exemplar sets. As an extension to the traditional example-based image retrieval systems, the user can specify negative exemplars (denoting images irrelevant to a query) along with the positive exemplars [1], [2].

Usually, the CBIR system searches the database for images "similar" to the positive exemplars and "dissimilar" to the negative exemplars. In general, in order to handle generic queries, a CBIR system will have a large set of low-level features to describe each image. Due to this, we do not have a clear picture of the "equi-potential" surfaces, *i.e.*, we are not sure how the images "identical" with respect to the features would "cluster" in the multi-dimensional feature space. Also, the options available for metrics is greatly dependent upon the kind of features in the system. Thus, choosing the best distance metric remains a largely empirical decision dependent on the actual CBIR system. While [3] and [4] provide methods to evaluate distance metrics for color features, they cannot be used for other kinds of features. [5] evaluates the

performance of the metrics in a system independent manner by using randomly selected image data to establish ground truth. However, there does not exist any general method for evaluating the distance metric for a given set of features. We present a framework for deciding the most effective metric for a given CBIR system. We use this framework to decide the most effective distance metric for the CBIR system developed at the Center for Artificial Intelligence and Robotics, which supports multiple image queries and relevance feedback using positive and negative exemplars.

We establish ground truth for different queries by manually selecting "good" and "bad" images. For each query, the distance of each of these images to each exemplar is then calculated. For each image, the set of distances to the exemplars is used to calculate its distance to the set of exemplars. This is done for both positive and negative exemplars. Thus, we end up with a distance to the set of positive exemplars and a distance to the set of negative exemplars for each image. This distance is then used by an image match measure to reclassify it. The accuracy in classification is our measure of the effectiveness of the metric. We repeat this procedure for different sizes of exemplar sets and evaluate the performance of the distance metrics.

## II. DISTANCE METRICS

In this section, we give a brief description of the metrics we used in our analysis. In what follows, we use the following notation:

$x_{ri}$  =  $i^{th}$  feature of  $r^{th}$  image  
 $\delta_{rs}$  = dissimilarity between  $r^{th}$  and  $s^{th}$  images  
 $d_r^+$  = distance of the  $r^{th}$  image to the set of positive exemplars  
 $d_r^-$  = distance of the  $r^{th}$  image to the set of negative exemplar.  
 $d_r$  represents either  $d_r^+$  or  $d_r^-$ .

A dissimilarity measure has to satisfy the following properties:

- 1) Positivity:  $\delta_{rs} > 0$
- 2) Symmetry:  $\delta_{rs} = \delta_{rs}$
- 3) Identity:  $\delta_{rr} = 0$

A dissimilarity metric is a distance measure which satisfies the triangle inequality as well.  $\delta_{rs} < \delta_{rt} + \delta_{ts}$ . However, for our purpose, we do not make any distinction between the two and use them interchangeably for the rest of our discussion.

We choose a variety of metrics which model different types of "equi-potential" surfaces and empirically decide the one that is the best for our system.

### A. List of Distance metrics analyzed:

In the analysis of the effectiveness of various distance metrics, we have selected thirteen distance metrics. These can be approximately divided into five classes.

- 1) Metrics which normalize the distance component wise  
The following metrics were investigated in this class:

- a) Canberra measure

$$\delta_{rs} = \sum_i \frac{|x_{ri} - x_{si}|}{|x_{ri}| + |x_{si}|} \quad (1)$$

This metric assigns greater weight-age to large relative differences between corresponding components.

- b) Divergence measure

$$\delta_{rs} = \sum_i \frac{(x_{ri} - x_{si})^2}{(|x_{ri}| + |x_{si}|)^2} \quad (2)$$

- c) Wave-Hedges measure

$$\delta_{rs} = \sum_i \frac{|x_{ri} - x_{si}|}{\max(|x_{ri}|, |x_{si}|)} \quad (3)$$

This metric also assigns higher weight-age to large relative differences. However, if two vectors have a particular component differing by a small amount, then that component is emphasized more than in the case of the Canberra metric.

- 2) Metrics which normalize the distance on the whole  
The metrics which we analyzed in this class were as follows:

- a) Angular Separation

$$\delta_{rs} = 1 - \frac{\sum_i (x_{ri}x_{si})}{\sqrt{\sum_i (x_{ri}^2) \sum_i (x_{si}^2)}} \quad (4)$$

- b) Bray-Curtis separation

$$\delta_{rs} = \frac{\sum_i |x_{ri} - x_{si}|}{\sum_i (|x_{ri}| + |x_{si}|)} \quad (5)$$

- c) Soergel distance

$$\delta_{rs} = \frac{\sum_i |x_{ri} - x_{si}|}{\sum_i \max(x_{ri}, x_{si})} \quad (6)$$

- d) Correlation dissimilarity metric

$$\delta_{rs} = 1 - \frac{\sum_i (x_{ri} - \mu_r)(x_{si} - \mu_s)}{\sqrt{\sum_i (x_{ri} - \mu_r)^2 \sum_i (x_{si} - \mu_s)^2}} \quad (7)$$

- 3) Metrics which do not involve any normalization of the components

A very common family of metrics in this class is the *Minkowski* or  $\mathcal{L}_p$  distance defined by

$$\delta_{rs} = \left( \sum_i |x_{ri} - x_{si}|^p \right)^{\frac{1}{p}} \quad (8)$$

The higher the value of  $p$ , the greater the importance given to large differences. Thus,  $\mathcal{L}_1$  gives equal importance to all differences while  $\mathcal{L}_\infty$  takes into account only that component for which the difference is maximum. We analyzed the following:

- a) The City Block Distance ( $\mathcal{L}_1$ )
- b) The Euclidean Distance ( $\mathcal{L}_2$ )
- c) The Chebychev distance ( $\mathcal{L}_\infty$ )
- d) Weighted  $\mathcal{L}_2$  Distance

$$\delta_{rs} = \sqrt{\sum_i w_i (x_{ri} - x_{si})^2} \quad (9)$$

In our analysis, we used  $w_i = \frac{\sigma_i^r}{\sigma_i^s}$  where  $\sigma_i^r$  and  $\sigma_i^s$  were the standard deviations of the  $i^{th}$  feature among a set of randomly selected images and the set of positive exemplars respectively.

- 4) Entropy based measures

We investigated one measure in this class.

- a) Kullback-Leibler Divergence (KL Divergence)

$$\delta_{rs} = \sum_i \left( \frac{x_{ri}}{x_r} \log \left( \frac{x_{ri}/x_r}{x_{si}/x_s} \right) \right) \quad (10)$$

where  $x_r = \sum_i x_i$ . Strictly speaking, this is not a measure since it does not obey the symmetry property. However, it can be considered as kind of a distance between two probability distances [6].

- 5) We also investigated measures of the form  $f(m_1, m_2)$  where  $m_1$  and  $m_2$  are metrics themselves [7]. Specifically, we present a metric which is the product of the Angular distance and Euclidean distance. This metric was found to perform better than both of them.

## III. IMAGE SELECTION

### A. Distance to Cluster

The distance metrics discussed above provide the distance of an image to each exemplar. These individual distances have to be then used to obtain the distance from the image to the whole cluster. We analyzed two methods which perform this.

- 1) Nearest Neighbor approach: The distance to each cluster is the minimum of the distances of the image to the individual members of the set.

$$d_r = \min\{\delta_{rs}\} \quad (11)$$

This method has been widely used in literature [8]. It gives less emphasis to images that are equally far from all the exemplars but not near any one exemplar.

- 2) Harmonic mean approach: Here the distance to each cluster is defined as the harmonic mean of the distances to the individual members of the set.

$$d_r = \frac{n}{\sum_s \frac{1}{\delta_{rs}}} \quad (12)$$

where  $n$  is the number of exemplars. This method gives emphasis to images that represent the characteristics of all the images uniformly rather than emphasizing ‘‘closeness’’ to one particular exemplar.

### B. Image Match

Using  $d_r^+$  and  $d_r^-$ , we need to perform discrimination, *i.e.*, decide which group the image belongs to. The most intuitive measure for discrimination would be the difference

$$s_r = d_r^- - d_r^+ \quad (13)$$

where  $s_r$  is the image match measure. If  $s_r$  is positive, then the image belongs to the set of good images, otherwise it belongs to the set of bad images. This method is used by many contemporary systems when they use the nearest-neighbor approach. However, this method has a shortcoming - the positive exemplars are no longer guaranteed to be among the best images. Since the set of exemplars is decided by the user, it is certainly expected that the positive exemplars are among the best images. Therefore, we propose and investigate the properties of measures which guarantee this property.

- 1) Asymmetric Measure:

$$s_r = \frac{1}{d_r^+} - \frac{1}{d_r^-} \quad (14)$$

Here the distances are assumed to be positive with 0 corresponding to a perfect match with an exemplar. This method gives  $s_r$  values tending to  $\infty$  for positive exemplars and values tending to  $-\infty$  for negative exemplars.

- 2) Bounded Measure:

$$s_r = \frac{d_r^- - d_r^+}{d_r^- + d_r^+ - 2a} \quad (15)$$

This measure is a normalized version of the Difference measure discussed previously. Here distances are assumed to be in the range  $[a, \infty]$  where a distance of  $a$  corresponds to a match with an exemplar. When  $a = 0$ , this reduces to

$$s_r = \frac{(1/d_r^+) - (1/d_r^-)}{(1/d_r^+) + (1/d_r^-)} \quad (16)$$

Thus, for a positive exemplar  $s_r = 1$  and for a negative exemplar  $s_r = -1$ . The values obtained from this measure are in the range  $[-1, 1]$ . Here too, a greater value of  $s_k$  corresponds to a better image.

## IV. EVALUATION PROCESS

The performance of a metric for a given feature set depends on how well it discriminates between good samples and bad samples. One of the options would be to rank the performance of metrics using a figure of merit, such as

$$f = \left| \frac{\mu_+ - \mu_-}{\sigma_+ + \sigma_-} \right| \quad (17)$$

where  $\mu_+$  is the mean of the  $d_r^+$  of the good samples and  $\mu_-$  is the mean of  $d_r^-$  of the bad samples ( $\sigma_+$  and  $\sigma_-$  are defined similarly). Unfortunately such figures of merit suffer from the assumption of unimodality of the distributions (positive and negative) of the distances. This assumption clearly cannot be valid in all cases, more so in the case of negative exemplars. We propose an evaluation strategy based on a manual establishment of ground truth and analyze the performance of the metrics under varying conditions.

### A. Evaluation Method Used

The metrics were evaluated by measuring their accuracy in discriminating between good and bad images for different queries. For each query, ground truth was established first. This was done by manually selecting  $M$  good and  $N$  bad images. Thus, the manual selection only distinguished between good and bad images and did not provide a complete ranking of the  $M + N$  images. The  $(M + N)$  images were then ranked in the increasing order of dissimilarity using the dissimilarity metrics. This discrimination was then compared to the manual discrimination by using a variant of the *Earth Mover’s Distance*.

Every image belonging to the set of bad images and ranked among the top  $M$  and every image belonging to the set of good images and ranked in the last  $N$  are cases of misclassification. A penalty of  $(M + 1 - i)$  was assigned in the former case and  $(i - M)$  in the latter, where  $i$  is the rank of image. The sum of all penalties is the penalty for a particular query. The average of the penalties for all queries is the earth mover’s distance for the query. This process was repeated for varying the number of exemplars.

For our evaluation, we chose  $M = 25$  and  $N = 50$  and established ground truth for 6 queries.  $N$  was chosen greater than  $M$  since in a given database of images, the number of bad images for a query is much larger than the number of good ones.

The different number of exemplars used were (1,1), (3,3), (5,5), (10,10) and (15,15) where in each  $(n_1, n_2)$  pair

- $n_1$  = number of positive exemplars and
- $n_2$  = number of negative exemplars.

### B. Database used for evaluation:

A total of 6 queries were used in the analysis. They are namely:

- 1) Black images
- 2) Cloud images
- 3) Rose images
- 4) Car images
- 5) Gold images

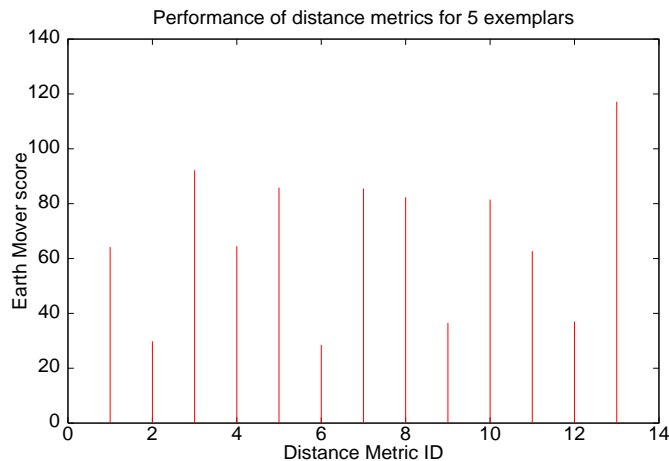


Fig. 1. Performance of metrics over all the features

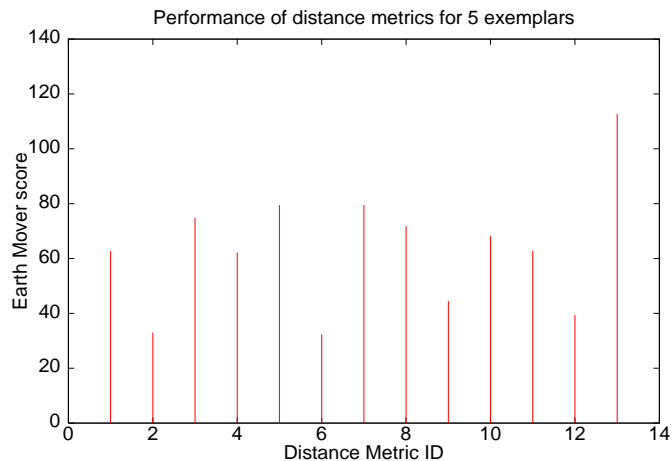


Fig. 3. Performance of metrics over local color histograms features

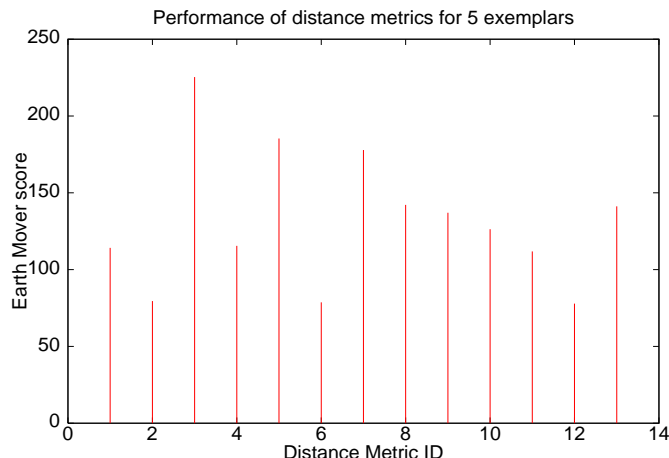


Fig. 2. Performance of metrics over global color histograms features

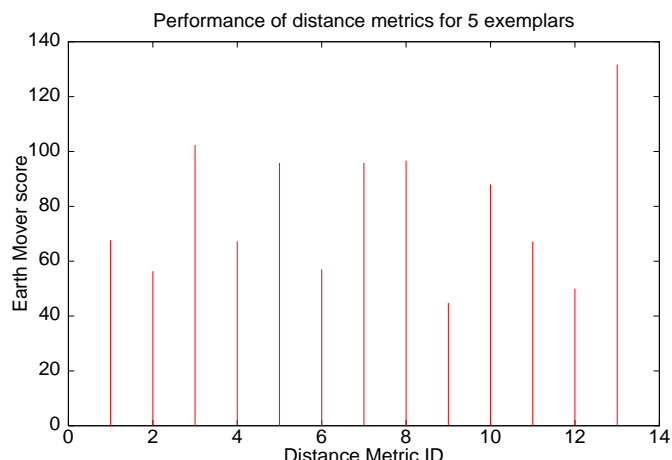


Fig. 4. Performance of metrics over co-occurrence matrix features

## 6) Boat images

The Black and Gold image sets contained images which were predominantly black and gold respectively. For these queries, similarity can be directly mapped onto the features (low level) while the similarity for the other queries are based on higher level abstractions and cannot be directly mapped. These images were selected from a database of 10000 images available to the CBIR system at CAIR.

### C. Set of Features used in the analysis:

The features we used consisted of the following feature sets. The reader is requested to refer [2] for further details.

- 1) Global Color Features: This consisted of 102 color histograms of the whole image
- 2) Local Color Features This consisted of 306 color histograms calculated over parts of the image.
- 3) Co-Occurrence Matrix: This set of features contains the elements of the co-occurrence matrix. Each element  $p(i, j)$  of the co-occurrence matrix is the probability that color  $i$  and color  $j$  occur in adjacent pixels.
- 4) The set consisting of all the above features.

For each feature set, the performance of the distance metrics was evaluated separately. This is because the performance of

TABLE I  
IDS USED FOR DISTANCE METRICS

Metric Name	Metric ID
Bray-Curtis	1
Canberra	2
Chebychev	3
City Block	4
Correlation	5
Divergence	6
Angular Separation	7
Euclid	8
KL Divergence	9
Combination (Magnitude * Angle)	10
Soergel	11
Wave-Hedges	12
Weighted Euclid	13

the metrics depend on the type of features. The performance of the distance metrics on the set containing all features was also measured.

## V. RESULTS

### A. Performance of different Metrics

The following plots show the Average Earth Mover distances for different metrics with different sets of features

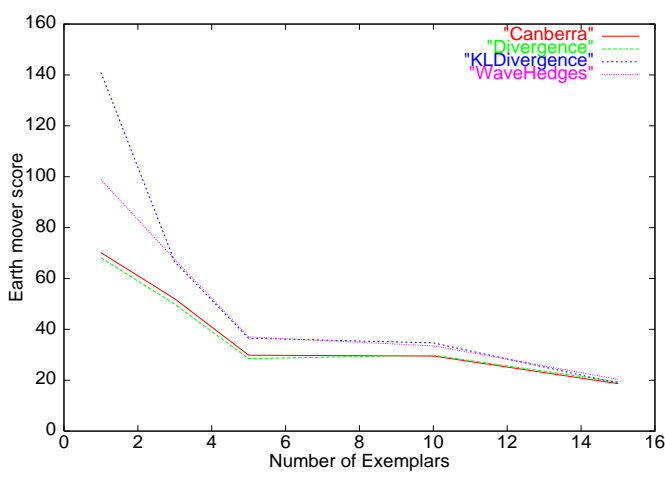


Fig. 5. Harmonic Bounded approach

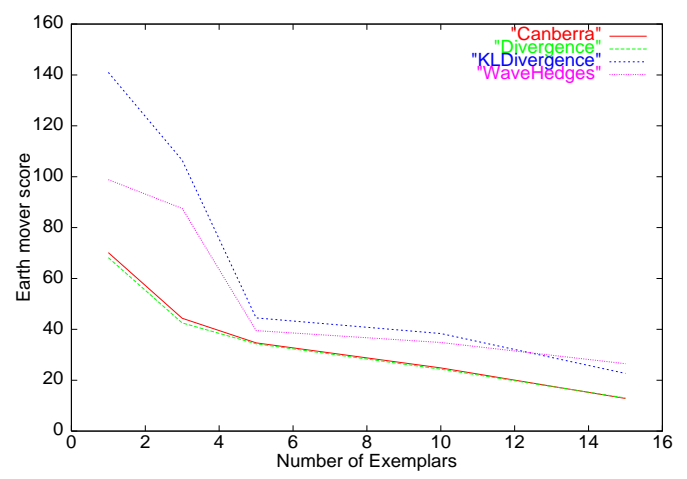


Fig. 7. Nearest Neighbor Bounded approach

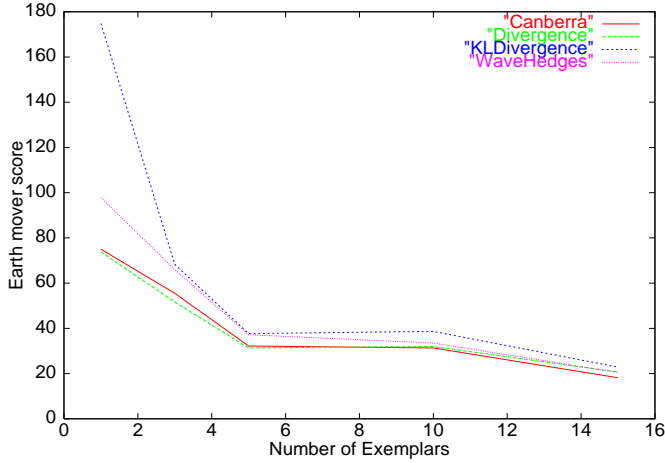


Fig. 6. Harmonic Asymmetric approach

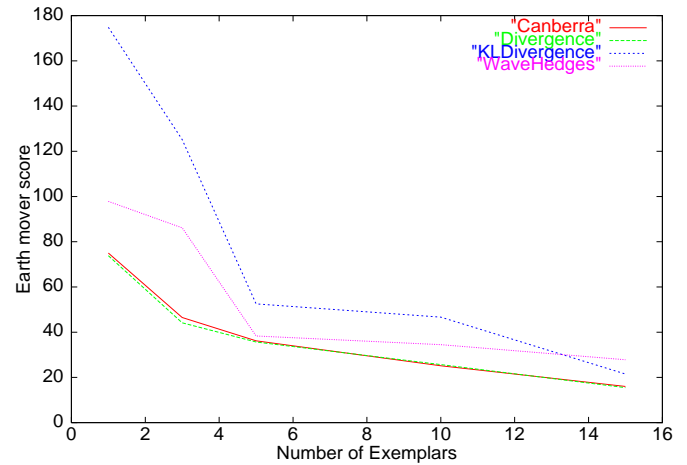


Fig. 8. Nearest Neighbor Asymmetric approach

with 5 positive and negative exemplars each. The Earthmover distances shown in Figs 1-4 were calculated using the *Bounded* measure with distance from exemplars estimated using the *Harmonic mean* approach. For convenience the different metrics are denoted by numbers. In the plots in this section the following convention is used for denoting metrics.

From the graphs, it is clear that Canberra, Divergence and Wave-Hedges outperform all the others consistently. KL Divergence outperforms these three metrics in the feature set containing the co-occurrence values. This is to be expected since the co-occurrence values are probabilities and KL Divergence was originally formulated for measuring the distance between two probability distributions.

### B. Performance of Harmonic mean versus Nearest Neighbor approach

It was noticed that *Harmonic mean* approach performed better for most metrics. The Earth mover score decreased rapidly as the number of exemplars increased for *Harmonic mean* as compared to the *Nearest Neighbor* approach. This variation has been shown for 4 metrics for the set containing all features in Figs 5-8. It was also observed that both *Asymmetric* and *Bounded* image match measures performed similarly.

### C. Variation in metric performance with number of exemplars

The variation in performance of the top four metrics, with number of exemplars given, is shown in the Figs 9-12. For simplicity, only the Earthmover distances calculated using the *Bounded* measure with distance from exemplars estimated using the *Harmonic mean* approach are shown. It was noticed that, as expected the performance of the metrics improved as the number of exemplars increased. This was observed in all the four feature sets. Canberra and Divergence perform consistently well, while Wave Hedges is a close third. KL Divergence has the best performance for the co-occurrence features when the number of exemplars is greater than 5.

## VI. CONCLUSION

We have presented a procedure for analyzing and determining a suitable metric for Example-Based Image Retrieval Systems. The results of applying our procedure to decide the best metric among thirteen candidate metrics for the Content-Based Image Retrieval system being developed at the Center

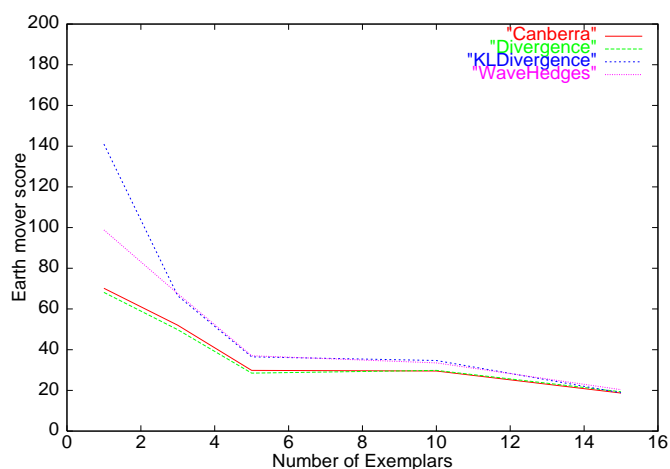


Fig. 9. Variation in performance over the whole set of features

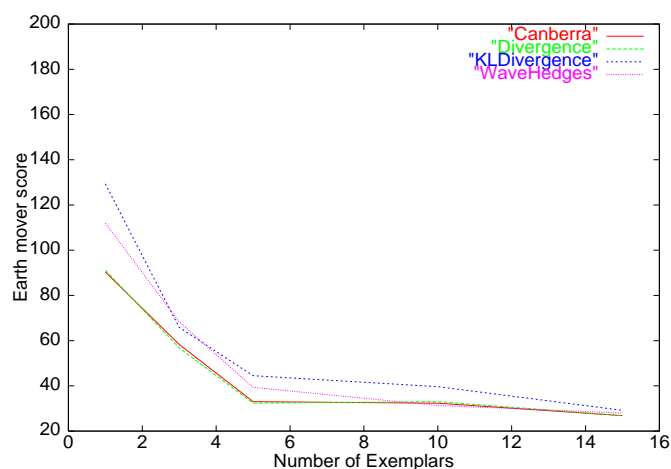


Fig. 11. Variation in performance over local color histograms features

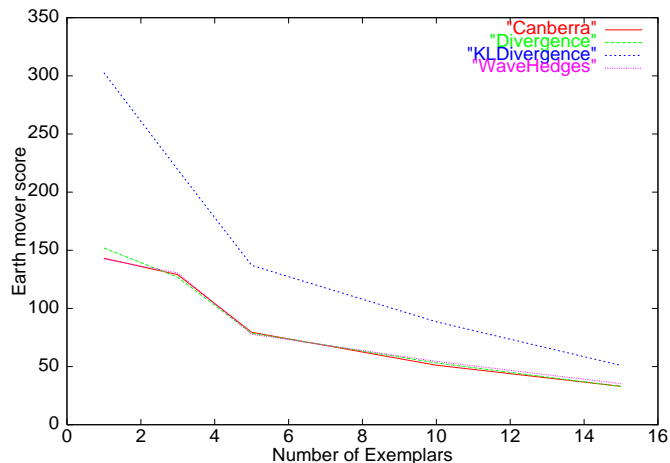


Fig. 10. Variation in performance over global color histograms features

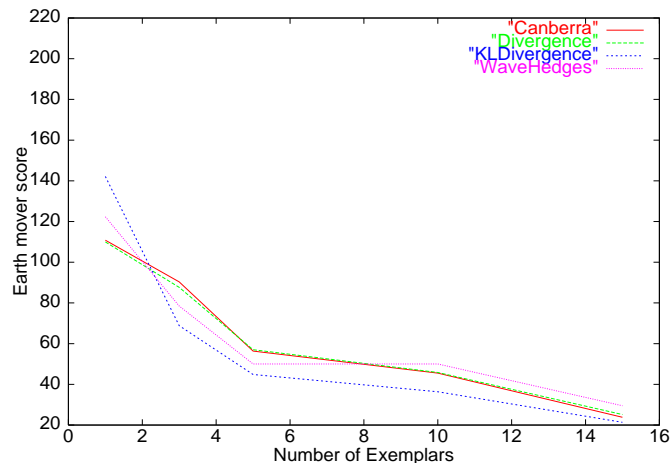


Fig. 12. Variation in performance over co-occurrence matrix features

for Artificial Intelligence and Robotics have been presented. The results indicate that for color features, Canberra and WaveHedges (which involved a component-wise normalization) performed better than others and KL Divergence performed best for color co-occurrence features. We have found that using the harmonic mean of distances to the exemplars to perform discrimination is more effective than the nearest-neighbor approach as the error in discrimination decreases rapidly with increase in the number of exemplars in the former.

## VII. ACKNOWLEDGMENTS

This work was sponsored by CAIR, Bangalore. The authors wish to thank Director, CAIR, for his support. Hetunandan and Pradeep worked under the aegis of the Summer Student Project Training Program (May-July 2002)

## REFERENCES

- [1] J. Assfalg, A. D. Bimbo, and P. Pala, "Image retrieval by positive and negative examples," in *Proc. of ICPR*, 2000.
- [2] C.V.Jawahar, P.J.Narayanan, and S. Rakshit, "A flexible scheme for representation, matching and retrieval of images," in *Proc. of ICVGIP*. Allied Publishers Ltd., 2000, pp. 271–277.
- [3] D. Androustos, K. N. Plataniotis, and A. N. Venetsanopoulos, "Distance measures for color image retrieval," in *Proc. of ICIP*, 1998.

- [4] M. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval for Image and Video Database V. SPIE Proc.*, 1995, pp. 381–392.
- [5] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi, "Empirical evaluation of dissimilarity metrics for color and texture," in *Proc. of ICCV*, 1999.
- [6] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.
- [7] D. Androustos, K. N. Plataniotis, and A. N. Venetsanopoulos, "A novel vector-based approach to color image retrieval using a vector angular-based distance measure," *Computer Vision and Image Understanding*, vol. 75, no. 1, pp. 46–58, July/August 1999.
- [8] H. Kosch and S. Atnafu, "Processing a multimedia join through the method of nearest neighbor search," *Information Processing Letters*, vol. 82, no. 5, pp. 269–276, 2002.