RESEARCH INVENTY: International Journal of Engineering and Science
ISSN: 2278-4721, Vol. 1, Issue 1 (August 2012), PP 31-35
www.researchinventy.com

Effectiveness of Image (dis)similarity Algorithms on Content-Based Image Retrieval

Keneilwe Zuva[1], Tranos Zuva[2]
[1](*Department of Computer Science, University of Botswana, Gaborone, Botswana*)
[2](*Department of Computer Engineering, Tshwane University of Technology, Pretoria, South Africa*)

ABSTRACT
(Dis) similarity measure is a significant component of vector model. In content based image retrieval the compatibility of (dis)similarity measure and representation technique is very important for effective and efficient image retrieval. In order to find a suitable dis-similarity measure for a particular representation technique experimental comparison is needed. This paper highlights some of the (dis)similarity algorithms available in literature then compares Euclidean dissimilarity and cosine similarity on Kernel Density Feature Points Estimator (KDFPE) image representation technique. The retrieval results show that cosine similarity had slightly better retrieval rate than Euclidean dissimilarity.
Keywords - *Image Representation, image retrieval, cosine similarity, Euclidean dissimilarity*

I. INTRODUCTION

The enormous collection of digital images on personal computers, institutional computers and Internet necessitates the need to find a particular image or a collection of images of interest. This has motivated many researchers to find efficient, effective and accurate algorithms that are domain independent for representation, description and retrieval of images of interest. There have been many algorithms developed to represent, describe and retrieve images using their visual features such as shape, colour and texture [1], [2], [3], [4], [5-7]. The visual feature representation and description play an important role in image classification, recognition and retrieval. The content based image (dis)similarity measurement algorithms, if chosen correctly for a particular image representation technique, will definitely increase the efficiency and effectiveness retrieval of data of interest. In this paper we will discuss the (dis)similarity measurement algorithms of images represented using their visible features (shape, colour and texture). At the end compare the effectiveness of cosine similarity and Euclidean dissimilarity using KDFPE representation technique [7] on a shopping item images database.

(Dis) similarity Similarity (s) can be defined as the quantitative measurement that indicates the strength of relationship (closeness) between two image objects. Dissimilarity (d) is also a quantitative measurement that reflects the discrepancy (disorder, distance apart) between two image objects. We formalise the definition of (dis)similarity in definition 1.

Definition 1 **(Dis) Similarity** *(d/s)*
Let Y be a non-empty set and *s/d* be a function on a set Y, such that

$$d/s : YxY \to R,$$
$$where\ R\ is\ the\ set\ of\ real\ numbers$$

This function is called pair-wise Similarity/dissimilarity function. A (dis)similarity space is a pair (Y, *d (s)*) in which Y is a non-empty set and *d (s)* is a (dis)similarity on Y. It is possible to convert similarity value to dissimilarity value. The *d/s* function is bounded. There is a relationship that exists between similarity and dissimilarity that allows us to derive the similarity values from dissimilarity values. The relationship is given by

$$s_{ij} = 1 - d_{ij}$$
$$where\ d_{ij}\ is\ a\ normalized\ dissimilarity$$
$$value\ between\ objects\ i\ and\ j$$
$$s_{ij} \in [0,1] \tag{1}$$

Or

$$s_{ij} = 1 - 2d_{ij}$$
$$where\ d_{ij}\ is\ a\ normalized\ dissimilarity$$
$$value\ between\ objects\ i\ and\ j$$
$$s_{ij} \in [-1,1] \tag{2}$$

From the equations (1) and (2) we can have an equivalence relationship between dissimilarity and similarity measurements. This equivalent relationship is shown below

$$s_{ij} \geq s_{iz} \Leftrightarrow d_{ij} \leq d_{iz}, \forall i, j, z \in X \quad (3)$$

Table 1 summarises the interpretation of the values of similarity and dissimilarity.

TABLE 1
Interpretation of (dis)similarity values

Given two objects i and j using equation 1.1a		
	Similarity value	Dissimilarity value
Exact similar	1	0
Very different	0	1
Given two objects i and j using equation 1.1b		
Exact similar	-1	1
Very different	1	0
General interpretation	Higher value	Lower value
	Lower value	Higher value

The (dis)similarity measurement algorithms can be grouped into metric and non-metric. Metric is defined in definition 2.

Definition 2 (**Dis) similarity Metric (Frechet)**
Let X be a non-empty set. A metric on X is a function *d* of $X \; x \; X$ into $[0, \infty)$, that satisfies the following conditions:

i. $d(x, y) \geq 0, \forall x, y \in X$ Reflexivity

ii. $d(x, y) = 0, \textit{if and only if } x = y,$ Non-negativity

iii. $d(x, y) = d(y, x), \forall x, y \in X,$ Symmetry

iv. $d(x, y) \leq d(x, y) + d(z, y), \forall x, y, z \in X.$ Triangle inequality

A metric space is a pair (X, *d*) in which X is a non-empty set and *d* is a metric on X. Observation from the definition is that the metric is not bounded. In our case we need a bounded metric, thus we will have an upper bound transforming it into bounded metric.

Non-metric (dis)similarity algorithms do not fulfil at least one metric conditions. Depending on which metric condition(s) the non-metric (dis)similarity algorithm does not fulfil a distinguishing term is used as shown in Table 2 [8].

TABLE 2
Non-metric Classification

Metric Condition Fulfilled	**Metric Condition not Fulfilled**	**Distinguishing Term**
Reflexivity, Non-negativity, Symmetry	Triangle Inequality	Semi-Metric (Non-Metric)
Non-negativity, Symmetry, Triangle Inequality	Reflexivity	Pseudo-Metric (Non-Metric)
Reflexivity, Non-negativity, Triangle Inequality	Symmetry	Quasi-Metric (Non-Metric)
Reflexivity, Symmetry, Triangle Inequality	Non-negativity	? (Non-Metric)
None	Reflexivity, Non-negativity, Symmetry, Triangle Inequality	Full-Non-Metric

II. PERFORMANCE EVALUATION

These (dis)similarity algorithms have been used effectively to retrieve images of interest successfully [9],[10],[11]. What makes an algorithm perfect for a certain image database is the contribution it has to the effectiveness and efficiency of content based image retrieval system. Effectiveness of retrieval is usually measured by precision and recall [12].

$$precision = \frac{A}{A + C} \quad (4)$$

$$recall = \frac{A}{N} \quad (5)$$

$$effectiveness = \begin{cases} \frac{A}{N} \; if \; T \rangle N \\ \frac{A}{T} \; if \; T \leq N \\ \frac{A}{A + C} \; precision \end{cases} \quad (6)$$

Where A is the number of relevant image objects retrieved and C is the number of not relevant image objects retrieved, N total relevant images in database and T is the number of relevant images that the user requires from the database.

Efficiency of retrieval is the speed of retrieval [8]. Metric and non-metric (dis)similarity algorithms compete equally well in the effectiveness of image retrieval. Non-metric lags behind in the efficiency of retrieval. This is because the indexing of databases is skewed in favour of metric (dis)similarity algorithms. It must be noted that an effective retrieval system is useless in large databases if it is not efficient. Next sections we are going to look at some metric and non-metric (dis)similarity algorithms.

III. METRIC (DIS) SIMILARITY (D/S) ALGORITHMS

Metric (D/S) algorithms exhibited high degree of effective and efficient retrieval of images of interest from a very large image database. Many researchers used metric S/D algorithms showed high precision and recall retrieval results [6],[13],[4]. The metric conditions could be used to index the image database for high efficient retrieval [8]. The following are some of the mostly used metric D/S algorithms:

1. Minkowski Family

$$L_p \ (p \geq 1 where \ p = 1, 2, 3.....\infty)$$

$$d = \sqrt[p]{\sum_{i=1}^{n} | x_i - y_i |^p} \tag{7}$$

Within this family very few have been used in image retrieval and they are Euclidean L_2, City block L_1 (taxicab norm, Manhattan) and Chebyshev L_∞ dissimilarity formulas. The formulas are given in equations 8 to 10 below:

Euclidean L_2

$$d = \sqrt{\sum | x_i - y_i |^2} \tag{8}$$

City block L_1(taxicab norm, Manhattan)

$$d = \sum_{i=1}^{n} | x_i - y_i | \tag{9}$$

Chebyshev L_∞

$$d = \max_i | x_i - y_i | \tag{10}$$

V. NON-METRIC (DIS) SIMILARITY ALGORITHMS

Non-metric D/S algorithms have been used and produced high degree of effective and efficient retrieval results from very large databases. This is in part due to the fact that researcher created weak metric (dis)similarity algorithms from these non-metric algorithms [14]. We are going to look at some of the non-metric S/D algorithms.

1. Pearson Dissimilarity Family

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X - \bar{X}}{\sigma_x} \right) \left(\frac{Y - \bar{Y}}{\sigma_y} \right) \tag{11}$$

where r , $\frac{X - \bar{X}}{\sigma_x}$, $\bar{X}$, σ_x-are Pearson correlation coefficient, the standard score, mean and standard deviation respectively.

Pearson dissimilarity measure algorithms are given as

$$d = 1 - r, \ where \ d \in [0, 2] \tag{12}$$

$$d = 1 - | r |, \ where \ d \in [0, 1] \tag{13}$$

There are other (dis)similarity algorithms that use correlation, some of them are Spearman rank correlation, Kendall's τ, Uncentred correlation[15].

2. **Shannon Entropy Family**

In this family of (dis)similarity algorithms are Kullback-Leibler, Jeffreys/J divergence, Jensen-Shannon and Jensen difference just to mention a few, are some of the non-metric algorithms that have been used in image retrieval systems [15]. The formulas are given in equations 14 to 17 below.

Kullback-Leibler

$$d = \sum_{i=1}^{n} x_i \ln \frac{x_i}{y_i} \tag{14}$$

Jeffreys/J divergence

$$d = \sum_{i=1}^{n} (x_i - y_i) \ln \frac{x_i}{y_i} \tag{15}$$

Jensen-Shannon

$$d = \frac{1}{2} \left[\begin{array}{l} \sum_{i=1}^{n} x_i \ln \left(\frac{2x_i}{x_i + y_i} \right) + \\ \sum_{i=1}^{n} y_i \ln \left(\frac{2y_i}{x_i + y_i} \right) \end{array} \right] \tag{16}$$

Jensen difference

$$d = \sum_{i=1}^{n} \left[\frac{x_i \ln x_i + y_i \ln y_i}{2} - \left(\frac{x_i + y_i}{2} \right) \ln \left(\frac{x_i + y_i}{2} \right) \right] \quad (17)$$

3. Inner Product Family

The inner product family (dis)similarity measurement include the inner product explicitly in their formulas. In this family we are going to look at only three formulas, that is inner product, harmonic mean and cosine. We are interested in cosine since it is a normalised inner product which allows for physical comparison of (dis)similarity measurements of images. The formulas of the inner product family members are given in equations 18 to 20.

Inner Product

$$d = \sum_{i=1}^{n} x_i y_i \quad (18)$$

Harmonic Mean

$$d = 2\sum \frac{x_i y_i}{x_i + y_i} \quad (19)$$

Cosine

$$d = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i}\sqrt{\sum_{i=1}^{n} y_i}} \quad (20)$$

VI. EXPERIMENTATION

The main objective of the experimentation is to compare the effectiveness of cosine similarity and Euclidean dissimilarity methods. We chose the two dis-similarity methods because the mostly used in image retrieval and that they belong to distinct categories. An image database of at least 200 shop items shapes is created. Some of the image objects are rotated at 90, 180 and 270 degrees. The images that are rotated were not rotated lossless, meaning degradation of the image object occurred during rotation. The image objects were of different dimensions M x N or N x N where M and N are real numbers when they are brought to the system. The images that are used only have one image object with a homogeneous background. The image object shape of grid dimension 45 x 45 is segmented using the Chan and Vese active contour without edge [16]. All images are converted to gray scale images. They are then represented using KDFPE. Each image was used as a query and the retrieval rate was measured using the Bull's Eye Performance (BEP), recall and precision performance. Matlab 7.6 was used to implement the system. Example of classes of shapes experimented with are given in Figure 1. In each class there are at least ten elements with some items rotated and scaled.

Fig. 1 Examples of classes of items in database

VII. RESULTS

Figure 2 shows the images considered similar to the query image on the top left of the figure. It can be seen that the query image is part of the retrieved images, indicating that it belongs to the database. In the sample of retrieval in Figure 2, cosine has a 90% precision while Euclidean has a 80% precision.

Fig. 2. Ten retrieval results of cosine on the left and Euclidean on the right

The result in Figure 3 shows that cosine is better in retrieving images from the database. It can be appreciated that the difference is not very substantial. The Bull's Eye Performance (BEP) of Euclidean is 92.60% while cosine is 93.05%.

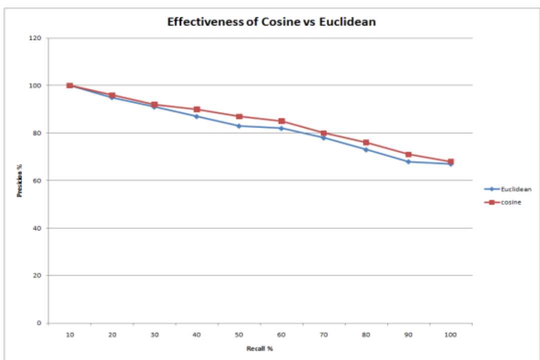

Fig 3. Comparison of cosine and Euclidean using Recall-Precision Chart

VIII. CONCLUSION

From the results it can be concluded that cosine similarity is more compatible with KDFPE representation method than Euclidean dissimilarity. If we are only looking at the effectiveness of a method then cosine similarity would be used together with the KDFPE technique. Since the efficiency of the system need to be taken into consideration then the trade-off between effectiveness and efficiency needs to be calculated to make a final decision. In general both method shows high retrieval rate. Euclidean dissimilarity brings metric properties which enables efficiency of retrieval. Cosine similarity has an advantage of naturally normalized similarity. So before any dis-similarity method is used experiments to ascertain the compatibility with the representation method must be done.

REFERENCES

[1] M. X. Ribeiro, *et al.*, "Statistical Association Rules and Relevance Feedback: Power Allies to Improve the Retrieval of Medical Images," *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, 2006.

[2] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition Society*, vol. 37, pp. 1-19, 2004.

[3] Y. Li and L. Guan, "An effective shape descriptor for the retrieval of natural image collections," in. *Proceedings of the IEEE CCECE/CCGEI*, Ottawa, 2006, pp. 1960-1963.

[4] X. Zheng, *et al.*, "Perceptual shape-based natural image representation and retrieval," in *Proceedings of the IEEE International Conference on Semantic Computing*, 2007, pp. 622-629.

[5] Y. Mingqiang, *et al.*, "A survey of shape feature extraction techniques," in *Pattern Recognition*, 2008, pp. 43-90.

[6] D. C. Tran and K. Ono, "Content-based image retrieval: Object representation by the Density of feature Points," pp. 213-218, 2000.

[7] T. Zuva, *et al.*, "Object Shape Representation by Kernel Density Feature Points Estimator," in *First International Workshop on Signal and Image Processing (SIP)*, Bangalore, India, 2012, pp. 209-216.

[8] T. Skopal and B. Bustos, "On Nonmetric Similarity Search Problems in Complex Domains," *ACM Journal Name*, vol. V, pp. 1-56, January 2010.

[9] S. Antani, *et al.*, "Evaluation of shape similarity measurement methods for spine X-ray images," *J. Vis. Commun. Image R. (Elsevier)*, vol. 15, pp. 285-302, 2004.

[10] E. G. M. Petrakis and C. Faloutsos, "Similarity Searching in Medical Image Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, May/June 1997.

[11] Z. Stejic, *et al.*, "Genetic algorthm-based relevance feedback for image retrieval using local similarity patterns," *Information Processing and Management*, vol. 39, pp. 1-23, January 30 2003.

[12] X. Zheng, *et al.*, "Perceptual shape-based natural image representation and retrieval " in *International Conference on Semantic Computing*, 2007.

[13] D. Zhang and G. Lu, "Generic Fourier Descriptor for Shape-based Image Retrieval," in *IEEE Transactions on multimedia*, 2002, pp. 425-428.

[14] K. L. Clarkson, "Nearest-Neighbor Searching and Metric Space Dimensions," in *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, ed Cambridge: MIT Press, Cambridge, MA, 2005.

[15] S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, pp. 300-307, 2007.

[16] T. F. Chan and L. A. Vese, "Active Contours Without Edges," *IEEE*, vol. 10, pp. 266-277, 2001.