H-METRIC: CHARACTERIZING IMAGE DATASETS VIA HOMOGENIZATION BASED ON KNN-QUERIES

Welington M. da Silva[1], Jose F. Rodrigues Jr.[2], Agma J. M. Traina[2], and Sergio F. da Silva[2]

[1]Universidade Federal de Sao Carlos - Campus Sorocaba - Rodovia Joao Leme dos Santos, Km 110 - 18052-780 - SP-264 - Sorocaba, SP, Brazil
[2]Inst. de Ciencias Matematicas e de Computacao - Universidade de Sao Paulo - CP 668 - 13560-970 Sao Carlos, SP, Brazil
[*]Email: junio@icmc.usp.br

ABSTRACT

Precision-Recall is one of the main metrics for evaluating content-based image retrieval techniques. However, it does not provide an ample perception of the properties of an image dataset immersed in a metric space. In this work, we describe an alternative metric named H-Metric, which is determined along a sequence of controlled modifications in the image dataset. The process is named homogenization and works by altering the homogeneity characteristics of the classes of the images. The result is a process that measures how hard it is to deal with a set of images in respect to content-based retrieval, offering support in the task of analyzing configurations of distance functions and of features extractors.

Keywords: Content-based image retrieval, Metric spaces, Precision-Recall

1 INTRODUCTION

Content-based data retrieval is one of the most important techniques for non-dimensional data indexing, such as images, sounds, and video. Such techniques are based on the concept of metric space, a mathematical abstraction that allows data of any kind to be embedded in a space within which it can be queried. Metric spaces are based on features extraction, distance functions, and metric access methods, factors that may determine different configurations for the data indexing. The possible definitions of metric spaces are numerous, which leads to different solutions for the same problem. For this reason, we need a method to measure the effectiveness of the different solutions that are proposed. The more usual way of measuring such efficiency is the metric called Precision and Recall (Baeza-Yates, 1999).

The use of Precision-Recall occurs in settings where content-based data retrieval techniques are designed and proposed to the research community, which must verify the efficiency of what is proposed - including features extractors, distance functions, and metric access methods. Under these circumstances, Precision-Recall provides parameters for the evaluation of results.

In addition to the methodologies involved in the definition of a metric space, the main factor that influences the calculation of the Precision-Recall metric is the dataset over which the metric indexing is done. In the case of images, sets in which the images have a high degree of homogeneity tend to have better outcomes in Precision-Recall plots; while in sets where the images are more visually heterogeneous, the tendency is to have plots showing a lower efficiency. In other words, different sets of images represent different challenges when the goal is retrieval of images satisfying visual similarity. It can be said that some sets are “more difficult” than others.

Despite the fact that the set of images has a great impact over the results of the Precision-Recall metric and over the actual design of the indexing system, there are no metrics to inform the researcher about how challenging a given set of images is or even to state a comparative perspective between different sets of images. This fact may hamper the search for techniques of content-based image retrieval, leading to the following problems:

- lack of a numerical concise reference with respect to the set of images, preventing a full assessment of the results of the Precision-Recall;
- absence of a clear criterion for choosing sets of test images;
- impossibility of finding equivalence among different sets of images;
- additional burden on the choice of sets of images, which must be checked individually;
- complexity in the description of the characteristics of a given set of images in scientific vehicles.

In the context of these problems, this paper describes the methodology *H-Metric*. The proposed metric is an approach to describe the properties of a set of images within the problem of data recovery supported by metric indexing. H-Metric is based on the Precision-Recall metric; it proceeds by analyzing the spectrum of this measure over a sequence of controlled modifications of the characteristics of an image dataset. The modifications are based on the concept of homogeneity, that is, how clustered are the classes of images in the metric space? The result of the metric corresponds to the point of convergence where the controlled modifications of the classes produce a Precision-Recall plot close to ideal. This work focuses on the domain of images; however, the proposed measure applies to any field that can be metrically indexed.

2 RELATED WORK

When using a set of images in metric indexing experiments, the properties of the data are the main aspect related to the results presented at the Precision-Recall plot. It may happen, for example, that a given configuration (features extractor and distance function) shows excellent results for a given set of images; however, this set would present good results for a large number of other settings. This happens in sets of images whose extracted features define well-defined clusters, a circumstance where the determination of an accurate indexing is not a tough challenge. In other cases, a researcher may develop new techniques and, by testing her/his methodology, she/he observes unsatisfactory results in the form of Precision-Recall curves inferiorly bended. However, the set of images that was used would present bad results for a large number of metric indexing configurations. This happens with datasets that, in a metric space, show no homogeneity (well-defined groups) with respect to their classes. In fact, according to Reeker (2001), the Precision-Recall is biased because it is a superficial metric, as it is necessary to count on alternative metrics that work in greater depth to measure not only the metric indexing setting but also the trend embedded in the dataset.

In the cases described in the preceding paragraph, there is a drawback with regard to the process of research and development. First, there may be non-efficient techniques that appear to have more potential than what they actually do (Powers, 2007). Second, promising techniques can be abandoned early on the basis of dataset configurations that are not expected to be treatable in the context of that particular technique. In both cases, the process of improving a given configuration metric is harmed because the researcher has not received adequate parameters to evaluate her/his methodology. The way it is used, Precision-Recall is often a magnitude without reference; it is not known exactly whether it refers to difficult or to easy problems. Thus, this work fits into the line advocated by Berger (1985), who states that the validity of an experiment should be conditional to the knowledge of how the expected outcome (success) is, *a priori*, present in the test data.

3 CONCEPTS

3.1 Features extraction

The first task required for non-dimensional data indexing is to transform the data elements into an appropriate format, consisting of dimensions - numerical features. In the field of images, this step requires a process called *features extraction*, which is mainly oriented to characteristics of color, shape, and texture. That is, one must convert the images into a numerical representation corresponding to a vector $x = \{x_0, x_1,..., x_{n-1}\}$ of n representative numbers intrinsic to the original data. Classic examples are the color histogram (Felipe, 2005; Sheshadri, 2006) and the coefficients obtained using the Fourier transform (Zhang, 2001). Thus, throughout this

text, the term *features extraction* refers to the general formula f:D → □ where *D* is the data domain and □ ⊂ R^n is the *n*-dimensional space of characteristics.

3.2 Distance Functions

After the extraction of features, it is necessary to use a similarity measure or distance function, which measures the similarity between vectors of numbers extracted from data objects. The simplest way of doing this is to consider each numerical characteristic as a coordinate of an *n*-dimensional space and to calculate the Euclidean distance between the vectors. Other examples of measures are the City Block and the Minkowisk family of distances (Aggarwal, 2001). The use of different distance measures allows not only a variety of scopes in metric spaces but also the numerical weighting of specific dimensions, adding semantic interest in the data recovery.

3.3 Metric Spaces

Once a features extractor and a distance function are defined, one can establish a metric space. A metric space is a pair M = <□, δ>, where □ is the field of data being indexed and δ:□ *x* □ → $\mathbb{R}^+$ is a function that associates a distance to any pair o_i, o_j ∈ □. Given three elements o_i, o_j, and o_k ∈ □, the pair $M = <□, \delta( )>$ defines a metric space where the $\delta( )$ satisfies the following axioms:

- Symmetry: $\delta(o_i, o_j) = \delta(o_j, o_i)$;
- Non negativity: $\delta(o_i, o_i) = 0$ and $0 < \delta(o_i, o_j) < \infty$, if $o_i \neq o_j$;
- Triangular inequality: $\delta(o_i, o_j) \leq \delta(o_i, o_k) + \delta(o_k, o_j)$.

In this context, δ( ), the metric distance function, is responsible for calculating the similarity between the domain objects. The more similar the objects are, the lower the calculated value is, as well as the more dissimilar the objects are, the higher the calculated value. Thus, the data retrieval operations (queries) become intuitive in metric spaces based on the concept of similarity.

3.4 K-Nearest Neighbors Query

On a metric space, it becomes possible to perform similarity queries. In such queries, given an element of interest - the query center, we want to retrieve the elements of the set of images that have smaller distances (higher similarity) to this element. The two basic similarity queries are the nearest neighbor query and the range query. The nearest neighbor query is set out below:

Definition (Nearest Neighbor Query): given a query object o_q represented by its vector $f(o_q)$ and given domain D, the nearest neighbor refers to the element o_n defined as $NNQuery(o_q) = \{o_n \in D \mid \forall\ o_i \in D, \delta(f(o_q), f(o_n)) \leq \delta(f(o_q), f(o_i))\}$. In the real world, it might translate: "find the image in *D* which is more similar to the photo of the pope".

The extrapolation of this definition for *k* nearest neighbors, k ≥ 1, is straightforwardly given by $KNNQuery(o_q, k)$, which produces an ordered list of elements in which the $(n-1)$-th element is closer to o_q than the *n*-th element, $2 \leq n \leq k$.

3.5 Precision-Recall

The Precision-Recall calculation is done using pre-classified datasets. In such sets, it is possible to perform queries whose results can be verified by the examination of the classes of the returned objects and by comparing them with the known classes of the given set. The Precision-Recall metric stems from the fact that when a query is performed, the retrieved information may or may not be relevant to a given element of interest - according to the criteria established for the given domain. This relevance is what determines the pre-classification of data; as such, what one wants to assess is the satisfaction of the expected relevance during queries.

Precision-Recall can be formalized as follows. For a given query centered on an element of interest o_q, consider: the data domain *D*, *R* the set of relevant (expected) elements, *A* the set of elements returned by the query, and R_A the set of relevant elements that were actually retrieved.

$$P = \frac{R_A}{A} \tag{1}$$

$$R = \frac{R_A}{R} \tag{2}$$

In words, given the context of a query, precision is defined as the ratio of the number of relevant elements that were retrieved ($|R_A|$) to the total number of elements returned in the query ($|A|$) - Eq. 1. Recall is defined as the ratio of the number of relevant elements that were retrieved ($|R_A|$) to the number of known relevant items ($|R|$) - Eq. 2. The Precision-Recall is measured along the spectrum of the quantity of relevant elements returned in a given query - considering a set *A* with cardinality big enough for one to obtain Recall 1 (100% of the relevant elements) and Precision $|R|/|A|$. The Precision-Recall calculus, usually, is performed over multiple *k*-nearest neighbors queries. As such, one makes a query for each element of a significant subset of the pre-classified items, each query considering $k=|A|=|D|$ elements; the series of Precision-Recall calculations is then aggregated by arithmetic mean.

4 H-METRIC

In the context presented in Sections 1 and 2, the proposed methodology is the use of Precision-Recall not only on the original configuration of a set of images but also extrapolating its principle of operation over a sequence of controlled modifications of the set of classes.

Following the practice of Precision Recall experiments, our methodology demands a pre-classified test set with the intent of observing how the data retrieval techniques behave within a controlled test set. Initially, one must perform a series of nearest neighbors queries - considering the centroid of each class - and calculating the Precision-Recall the traditional way. During the queries, we proceed by identifying the retrieved elements that have reduced the performance of the metric retrieval system; in this first iteration, we consider only the $t = 1$ element closer to the center. Then, for the elements whose class does not correspond to the class of the center, we redefine their classes so that they will not reduce the Precision-Recall measure in further iterations. That is, the *t*-th closest element to the center of each query has its class redefined to conform to the class of the query center; in case the class already matches the class of the query center, nothing is done. After this iteration, the performance of the Precision-Recall increases. Next we consider larger values of $t = \{2, 3,...\}$ and proceed with the redefinition of the classes of the elements closer to the center. At each iteration, the performance of the Precision-Recall increases until, at a certain time, for $t < |D|$, its curve approaches the ideal.

The process of redefining the classes of the nearest neighbors, as described in the preceding paragraph, we call *homogenization process*. This designation is due to the definition of better defined - more homogeneous - classes within the metric space, what artificially increases the performance of the Precision-Recall. The idea, therefore, is to make the challenge of image retrieval increasingly "easier" and to monitor the performance achieved at each step.

Formally: given a query center o_q and an image domain *D*, then, the *t*-homogeneity of a dataset *I* refers the redefinition of the classes of the elements of sorted list $KNNQuery(o_q,t)=<o_1, o_2, ..., o_t>$ such that $Class(o_i) \neq Class(o_q)$, $\forall o_i \in KNNQuery(o_q,t)$, $1 \leq i \leq t$. We consider that the homogenization process has converged when:

$$\int_0^1((P_t(R_t))\,dR_t - \int_0^1((P_{t+1}(R_{t+1})\,dR_{t+1} \leq 0.01 \tag{3}$$

where *P(R)* is a continuous polynomial function that interpolates the points of Precision-Recall. That is, when there is no longer a significant variation - more than 1% - between two consecutive graphs of Precision-Recall, we consider that the maximum performance has been achieved.

At the moment when there is a convergence to a maximum performance, we state the current value of *t* as corresponding to the complexity of the set of images, which will be considered "H-hard", where *h=t*. We have observed that different datasets converge to a maximum Precision-Recall curve faster than others and, therefore, each set has its characteristic *H* value.

The calculation of the h-metric is described in Algorithms 1 and 2. In Algorithm 1, successive calculations of the area of the Precision-Recall curve are performed, and between each calculation, the pre-processed dataset is homogenized to the *t*-th element. The calculation proceeds iteratively, first by calculating the variation of the Precision-Recall curve, then performing the homogenization for each increment, which characterizes the iteration - please refer to Algorithm 2.

Algorithm 1. Algorithm to calculate the H-Metric

```
Input: PCD: pre-classified dataset
CalculateHMetric(PCD) begin
    t ← 1;
    PRVariation ← 1;
    PreviousPRArea ← 0;
    while (PRVariation ¿ 0.01) do
        PRArea ← CalculatePRArea(PCD);
        Homogenization(t, PCD);
        PRVariation ← PRArea - PreviousPRArea;
        PreviousPRArea ← PRArea;
        t++;
    end
    return t;
end
```

Algorithm 2. Homogenization algorithm

```
Input: PCD: pre-classified dataset, t: current t
Homogenization(PCD, t) begin
    for each class C ∈ PCD do
        if (!C.IsShrinking()) then
            Centroid ← C.GetCentroid();
            ResultSet ← KNN(Centroid, t);
            Result ← ResultSet.GetIthElement(t);
            if (Result.GetClass() != C) then
                (Result.GetClass()).Shrinking(true);
                Result.SetClass(C);
            end
        end
    end
end
```

5 EXPERIMENTS

Experiments were performed with three sets of images called “objects”, “places”, and “landscapes”, illustrated in Figure 1. Each set contains 10 classes, each with 12 items, summing 360 images. In the experiment, we have extracted features of color. More specifically, we have calculated the color histogram of each image, and from the histograms we calculated: mean, standard deviation, smoothness, distortion, uniformity, and entropy. Such characteristics are generally called first-order statistics and, given a histogram, are calculated as follows.

The *n*-th moment of the mean is given by:

$$\mu_n = \sum_{i=0}^{L-1} (Z_i - m)^n \, p(Z_i) \tag{4}$$

where Z_i refers to the *i*-th observed intensity in one color channel (red, blue, green, or gray), $p(Z_i)$ is the relative frequency of intensity Z_i, L is the number of different levels of intensity, and m is the mean of the observed intensities, given by:

$$m = \sum_{i=0}^{L-1} Z_i \, p(Z_i) \tag{5}$$

Thus, a value of $p(Z_i)$ is an estimate of the probability of occurrence of intensity Z_i, so that the entire histogram can be understood as a probability density function. From there, calculate the statistics of first order.

Standard deviation, or average contrast:

$$\sigma = \sqrt{\mu_2(Z)} = \sqrt{\sigma_2} \tag{6}$$

where $\mu_2(Z)$ is the second moment of the mean.

Smoothness, which has value 0 for constant intensities and a value close to 1 for oscillating intensities:

$$R = 1 - \frac{1}{(1+\sigma^2)} \tag{7}$$

Distortion, or third moment around the mean, which approaches 0 for symmetric histograms and tends to positive values for histograms skewed to the right and to negative values for histograms skewed to the left:

$$\mu_3 = \sum_{i=0}^{L=1} (Z_i - m)^3 \, p(Z_i) \tag{8}$$

Uniformity, which tends to its maximum value when all values of $p(Z_i)$ are equal and tends to lower values when the variability is higher:

$$U = \sum_{i=0}^{L-1} p(Z_i)^2 \tag{9}$$

And entropy, which gives an idea of how random the levels $p(Z_i)$ are:

$$e = -\sum_{i=0}^{L-1} p(Z_i) \log_2(p(Z_i)) \tag{10}$$

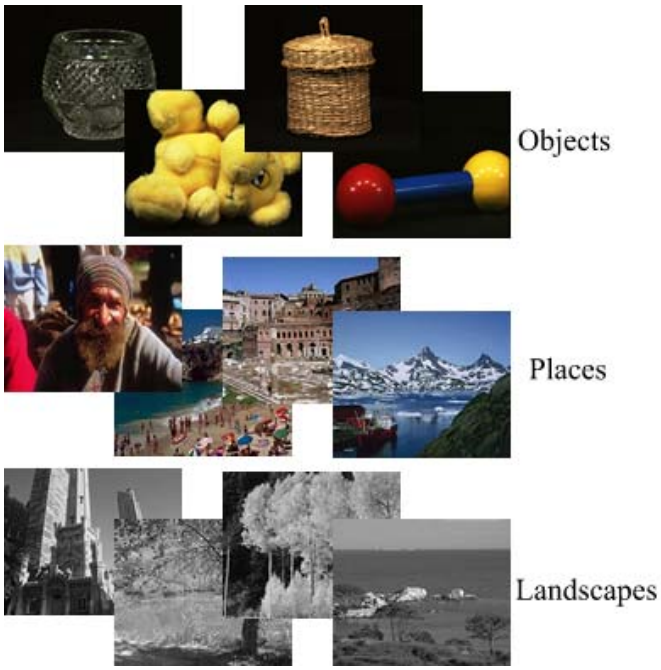

Figure 1. Illustration of the three sets of images used in the experiments

Considering the sets of images, it can be stated that the first set, *objects*, is very homogeneous. In this set, the objects of each class vary only according to the illumination applied to the image. In the second set, *places*, the images are quite complex in terms of shape, but in each class of the set, the images show significant homogeneity of colors. The third set, *landscapes*, uses only shades of gray, and their images are visually confusing.

For each set we have performed 60 queries for the k nearest neighbors, considering k = 120. We used the Euclidean distance. From these queries, we have calculated 12 points of Precision-Recall, each one corresponding to an increase of 1/12 of recall from the total of images. The 60 values calculated for each of the 12 points were aggregated using the arithmetic mean. For each dataset, we performed a sequence of homogenization processes until they reached the condition of convergence.

6 RESULTS

Figure 2 presents Precision-Recall plots for the three groups of images after thirty iterations of homogenization, with values t = 10, 20, 21, and 22. Following this, it is possible to note that the set of images of objects converges with t = 20, so the set is considered 20-hard. Meanwhile, as the set of images of places converges with t = 22, this set is characterized as 22-hard. The set of images of landscapes, in turn, does not converge until the value of t = 22, which demanded additional iterations of homogenization.

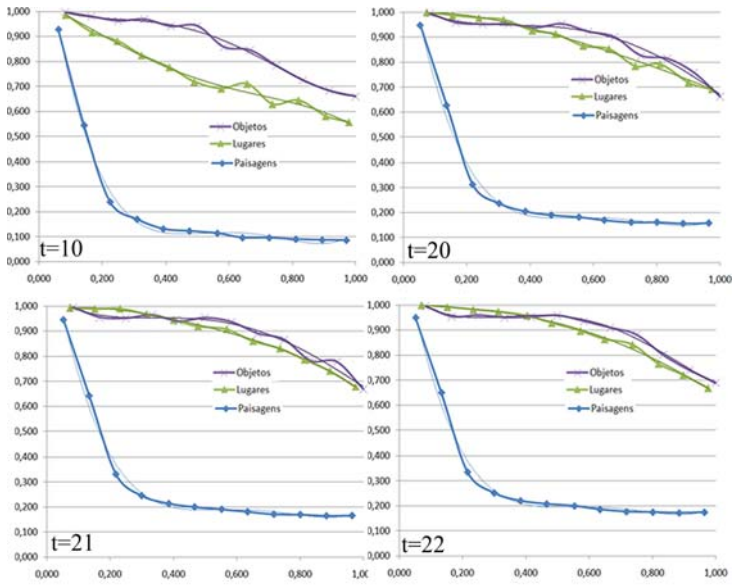

Figure 2. Precision-Recall plots for increasing values of homogenization. With $t=20$, one can see the convergence of the set of images of objects. With t = 22, it IS possible to observe the convergence of the set of images of places.

Figure 3 presents the Precision-Recall plots for values of t = 102, 103, and 104. With the value of t = 102, finally, there is a convergence of the set of landscape images. It is, therefore, 103-hard, which, for a total of 120 images, indicates that this set is totally contrary to the extraction of color characteristics.

In the experiments, the values of the H-Metric give an idea of the challenge presented by each set, which provides the researcher a sense of how each group can be treated and what to expect from each metric configuration. With the metric, the choice of one of these sets for content-based image retrieval becomes straight over a single reference value. Based on this value, we can affirm that the sets of objects and landscapes should show different results but always obey a distance proportional to the value of the H-Metric. One can also say that the set of landscapes should not be considered in settings where the metrics are oriented to color histograms because it has a natural complexity that prevents this approach.

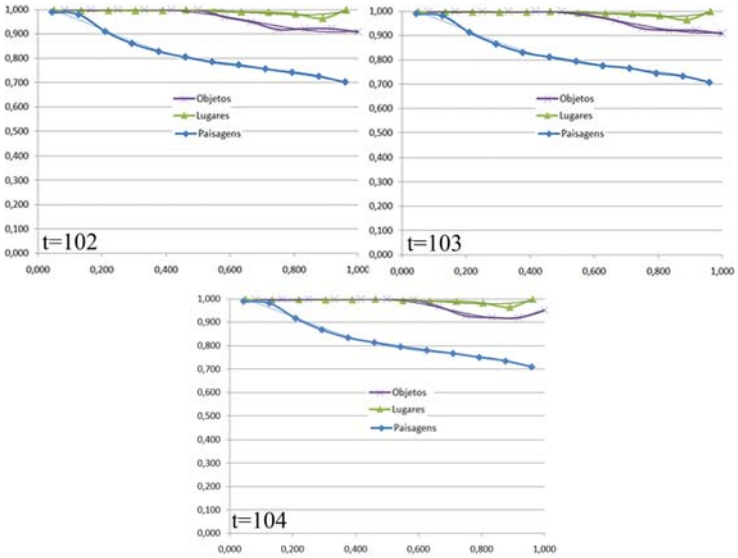

Figure 3. Precision-Recall graphs for even higher values of homogenization. Only after t = 103, does the Precision-Recall plot of the set of landscape images converge to a stable curve. At this degree of

homogenization, the sets of objects and places reach an artificial state, under which most elements are re-classified to the same class.

7 CONCLUSIONS

This paper presents the research and development over a content-based image retrieval system. To this end, we have touched the themes of distance function, features extraction, and evaluation of metric indexing techniques, discussing how these concepts can be put together to form a data recovery system. We considered three sets of images, from which we extracted first order statistics derived from their color histograms. Over these data, we performed experiments concerning the proposed methodology, named H-Metric. The H-Metric is based on series of KNN queries whose results map to Precision-Recall plots; each query in a series is followed by a progressive homogenization of the classes of a given set of images. Initial results reflect the expectations for the use of Precision-Recall along the spectrum of configuration defined by the classes in the pre-classified image sets. As future work, we foresee the systematic use of H-Metric in combination with metric F-Score, aiming at leveraging a ground truth test framework to guide researchers in the task of evaluating data retrieval techniques. The framework shall comprise multiple kinds of images and extracted features as well as the calculi of both metrics in comparative fashion.

8 ACKNOWLEDGMENTS

This work was partly supported by Microsoft Research, FAPESP (São Paulo State Research Foundation), CAPES (Brazilian Committee for Graduate Studies), and CNPq (Brazilian National Research Foundation).

9 REFERENCES

Aggarwal, C.C., Hinneburg, A., & Keim, D.A. (2001) On the surprising behavior of distance metrics in high dimensional spaces. In *International Conference on Database Theory* (ICDT), Springer, 420-434.

Baeza-Yates, R.A. & Ribeiro-Neto, B.A. (1999) *Modern Information Retrieval*. ACM Press/Addison-Wesley.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer.

Felipe, J.C., Traina, A.J.M., & Traina Jr., C. (2005) Global warp metric distance: Boosting content-based image retrieval through histograms. In *IEEE International Symposium on Multimedia-ISM2005*, IEEE Press, 295-302.

Powers, D.M.W. (2007) Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Technical Report SIE-07-001*, School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Reeker, L.H. (2001) Theoretic constructs and measurement of performance and intelligence in intelligent systems. In *Performance Metrics for Intelligent Systems*. National Institute of Standards and Technology.

Sheshadri, H.S. & Kandaswamy, A. (2006) Breast tissue classification using statistical feature extraction of mammograms. *Medical Imaging and Information Sciences*, 23(3):105-107.

Zhang, D.S. & Lu, G.J. (2001) Shape retrieval using fourier descriptors. In *Intl. Conference on Multimedia and Distance Education*, 1-9.

(Article history: Received 17 February 2011, Accepted 3 December 2011, Available online 15 January 2012)