Similarity measures based on fuzzy sets

Denise Guliato
Universidade Federal de Uberlândia (UFU)

Similarity measures based on fuzzy sets

Alan C. Genari, Denise Guliato*
Faculdade de Computação
Universidade Federal de Uberlândia
Uberlândia - MG - Brasil
Email: alanufu@gmail.com, guliato@ufu.br
Instituto Nacional de Ciência e Tecnologia - Medicina Assistida por Computação Científica
INCT - MACC
Petrópolis - RJ - Brasil

***Abstract*—The images are considered complex data and one of the fundamental operations in images databases is the similarity search. One way to represent an image is through feature vector, which is a numerical representation of the image. In this paper we proposed the representation of each element of the feature vector by a set of fuzzy numbers and how to calculate the measure of similarity using this new representation.**

***Keywords*-similarity measures; fuzzy set; content-based retrieval;**

I. Introduction

The interest in the development of content-based image retrieval (CBIR) system is increasing because of the growth in the number of image databases in many domains such as multimedia libraries, medical images and geographical information systems. In CBIR systems, the comparison of two images is a fundamental operation and is rarely made based on exact match. An image can be represented by a feature vector, where each element is associated to an attribute (or feature) of a image. These attributes are represented, in general, by single numerical values obtained by feature extractors. The similarity of two images is obtained by computing the similarity (or dissimilarity) between their feature vectors [1].

In this paper, we propose a novel approach to derive the similarity between two images, by representing each numerical value of their feature vectors as a fuzzy set, instead of a single value. This representation takes into account the uncertainty presents in the extraction process of features and consequently, increases the precision rate in the image retrieval process. In order to test our new approach, we used two databases and two (di)similarity measures: Euclidian distance and an equality index proposed by Bustince [2]. The results obtained by the proposed approach present higher performance than the traditional ones.

II. Methodology

The traditional approach referred in this work as *Approach One*, uses a single value to represent each feature in the feature vector. The proposed approach, referred in this work as *Approach 2*, represents each feature by a fuzzy set obtained by a predefined fuzzy partition. In both approaches, the objective is to obtain a value that represents the similarity between two images via similarity operators.

Let be X and Y the feature vectors of a query image and a target image, x_i and y_i, numerical values that represent the i^{th} features in the vectors $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$, respectively, and n the number of features. Let v be the value that express the similarity between the query image and the target image.

A. Approach One

In this approach each numerical value x_i in X and y_i in Y are normalized into the interval [0,1]. The normalized feature vectors are now represented as $X_p = [xp_1, xp_2, ..., xp_n]$ and $Y_p = [yp_1, yp_2, ..., yp_n]$. The similarity operator is applied to X_p and Y_p, as shown in Figure 1.

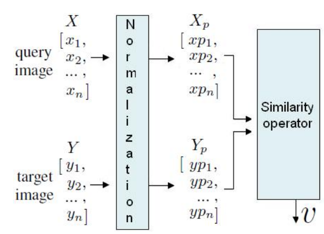

Figure 1. Overview of tradicional approach.

B. Approach Two

In this approach each numerical value x_i in X and y_i in Y are transformed in a fuzzy set of linguistic terms using fuzzy membership functions. To define the fuzzy membership functions there are three possible paths: consult a specialist, use predefined membership functions or obtain the membership functions automatically. In this work, the domain of each attribute is homogeneously divided into three linguistic terms L (low, medium, and high) interpreted as fuzzy membership functions with trapezoidal shape, as shown in Figure 2. The fuzzy function is defined as $\mu_L(x_i) \rightarrow [0,1]$ [3]. Note that different attributes can be fuzzified using different fuzzy partitions, depending on the nature of the feature. After the fuzzification process, the feature x_i is represented by the fuzzy set $xf_i = \{\mu_{low}(x_i), \mu_{medium}(x_i), \mu_{high}(x_i)\}$. Therefore, the vectors X and Y will be represented as $X_f = [xf_1, xf_2, ..., xf_n]$ and $Y_f = [yf_1, yf_2, ..., yf_n]$, respectively.

The (di)similarity between X_f and Y_f is derived by computing the average of individual (di)similarities of each pair of attributes (xf_1 and yf_1; xf_2 and yf_2;...;xf_n and yf_n) obtained by any similarity operator. A general scheme of the proposed approach is shown in Figure 3.

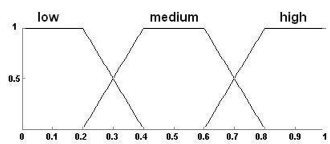

Figure 2. Three membership functions: low, medium and high.

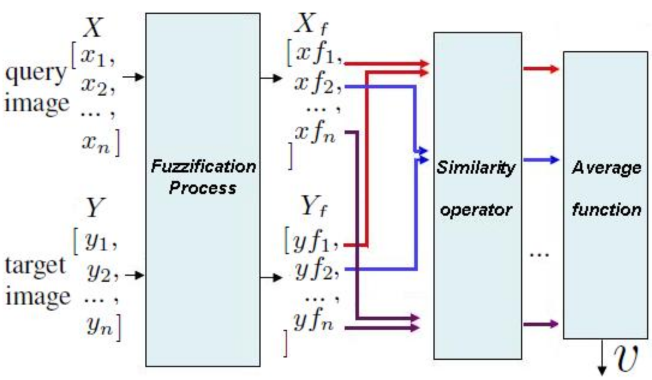

Figure 3. Overview of proposed approach.

III. Results

We implemented two operators of similarities; the first one based on Euclidean distance and the second one based on equality index proposed by Bustince [2], defined as:

$$EQ_{DI}(A,B) = \wedge \{\sigma_{DI}(A,B), \sigma_{DI}(B,A)\}\text{ , where}$$

$$\sigma_{DI}(A,B) = \frac{1}{n}\sum_{i}^{n} I(x_i, y_i) \; and \; I(x,y) = \wedge(1, 1-x+y),$$

$$A = (x_1, x_2, ..., x_n),\; B = (y_1, y_2, ..., y_n), x_i, y_i \in [0,1].$$

The IRIS and Corel 1000 databases were used for the tests. The IRIS database (http://archive.ics.uci.edu/ml/datasets/Iris) contains information on plants, is composed of 150 samples classified into 3 categories with 50 items for each category. Although the IRIS database does not relate to images, it fits to demonstrate the purpose of this work. The Corel 1000 database from Corel Corporation consists of 1000 images classified into 10 distinct visual groups, with 100 images for each group and the feature vectors was formed only by color moments.

The tests was run using leave-one-out cross validation. The precision/recall curve, shown in Figure 4, demonstrates that in both cases the proposed approach has reached better results.

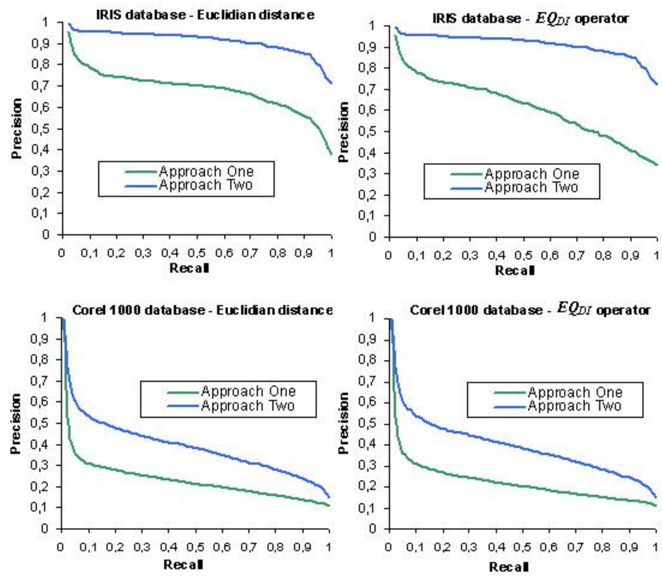

Figure 4. Precison X Recall obtained in experiments.

IV. Conclusion and Future Work

In tests made up to this point, the proposed approach has showed be efficient. The experimental results showed that the proposed approach allowed an increase in the precision rate. Further works are in process to evaluate the proposed approach with other similarity operators and other databases. We are also working in the development of techniques for automatic generation of fuzzy partitions.

Acknowledgment

This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Coordenação de Aperfeiamento de Pessoal de Nível Superior (CAPES), Brazil.

References

[1] *Hjaltason Gisli R., Hanan Samet, Index-driven similarity search in metric spaces (Survey Article).* ACM Trans. Database Syst., 2003.

[2] *Bustince H., M. Pagola, E. Barrenechea, Construction of fuzzy indices from fuzzy DI-subsethood measures: Application to the global comparison of images.* Information Sciences, 2007.

[3] *Klir G.J., B. Yuan, Fuzzy Sets and Fuzzy Logic, Theory and Applications.* Prentice Hall, New Jersey, 1995.