INCREMENTAL CLASSIFICATION OF IMAGES AND LARGE DATA SETS BY INFORMATION COMPRESSION AND FUZZY SIMILARITY ANALYSIS

G. Vachkov
Kagawa University, Department of Reliability-based Information Systems Engineering
Hayashi-cho 2217-20, Takamatsu City, Kagawa 761-0396, Japan; vachkov@eng.kagawa-u.ac.jp

Abstract. This paper proposes a computational scheme for classification of images and large data sets based on their similarity analysis. The procedure starts with a small number of known *core images* (or data sets) which form the initial size of the Image (Data) Base. During classification, the similarity degree of every new (unknown) image is computed against each of the core images in the Image Base. As a result, depending on the preliminary given threshold for classification, the new image could join the class of one of the core images in the Image Base or could be classified as “quite different” thus forming a *new class* in the Image Base. We use in the paper an unsupervised learning algorithm (a modification of the *Neural-Gas* leaning algorithm) for initial data compression. This algorithm replaces the original “raw data” (the RGB pixels) from the image with a smaller number of neurons thus creating the so called Compressed Information Model (CIM). Then the similarity analysis is performed as a two-input fuzzy inference procedure that uses the *Center-of-Gravity Distance* and the *Weighted Average Size Difference* between each pair of CIMs. The Fuzzy Rule Base and the Parameter Base should be tuned properly beforehand. The output of the fuzzy inference is the similarity degree for the given pair of images, defined in the range of [*0,1*]. The whole computational scheme is illustrated and discussed in the paper on a test example of *16* images, with several Image Bases that contain different number of core images.

1 Introduction

Similarity analysis, performed over a large amount of images or large data sets is very important step in the procedure for classification of different types of pictorial or process information. This is a very specific area of activity, where in many cases the experienced human performs better and produces more plausible solutions than the currently available computerized systems. One reason for this is the complexity and the vagueness in the definition of the problem. Obtaining a “better” and “more plausible” solution to the problem of similarity is a key factor for success in many applications such as quick search through a large amount of image or process data information, and its proper sorting and classification. The results of this similarity analysis and classification are often used for a proper fault or medical diagnosis and for discovering different abnormalities in the observed systems.

In this paper we present a special two-stage computation scheme to solving the complex problem of unsupervised classification, based on unsupervised learning for information compression, followed by fuzzy similarity analysis. The first stage is dealing with the information compression of the “raw data” (pixels or process operation data) into a respective compressed information model (CIM), which consists of a small number of neurons. Here one version of the off-line Neural-Gas unsupervised learning algorithm [5] is used in the paper as a tool for information compression. In the second stage a special fuzzy inference procedure for similarity analysis is proposed that uses two distinct parameters, extracted from the CIM. The first one is the C*enter-of-Gravity* of the compressed information model, while the other one is the *Weighted Average Size* of the CIM. The differences between these parameters for each pair of compared images are used as important features for the fuzzy procedure of similarity analysis.

Important element of the proposed classification scheme is its ability to make *incremental classification*, which could be very useful in the cases of growing number of images or data sets during time with a little initial information about the number of classes for classification. Then the iterative subsequent performance of the classification scheme would gradually find new classes which could be added to the initial small set of classes (called “core images”). The computational details about the proposed classification scheme are given in the sequel of the paper.

Finally, the flexibility and applicability of this scheme is shown in the paper on an illustrative example, consisting of *16* images of flowers. Several types of initial classes, consisting of different number of “core images” in the Image Base are used for the simulations and the analysis. The problems of tuning the parameters of the fuzzy similarity analysis are also discussed in the paper.

2 The proposed Unsupervised Classification Scheme

The Block-Diagram of the proposed scheme is shown in Fig. 1. This scheme is a further development of our previous ideas for similarity analysis, discussed in [6,8].

The procedure starts with a small number of known *core images* (or *core data sets*) which form the initial size of the Image (Data) Base. During classification, the similarity degree of every new (unknown) image is computed against each of the core images in the Image Base. As a result, depending on the preliminary given *threshold Th* for classification, the new image could join a certain class (*core image*) in the current Image Base or could be classified as "quite different image" thus forming a *new class* in this Image Base. In such way, the proposed general concept of classification is incremental one, allowing the Image Base to gradually grow, when a new CIM (with low similarity degree to any of the other CIMs in the Image Base) is discovered.

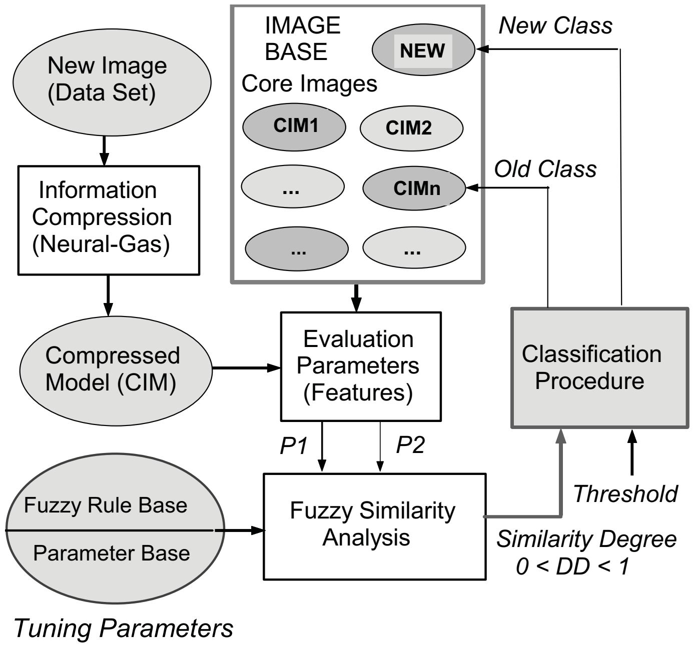


Figure 1. Block diagram of the proposed incremental unsupervised classification scheme based on fuzzy similarity analysis.

3 Unsupervised learning algorithm for Information Compression

The first step before the actual similarity analysis and classification of the images is to find a way to decrease the large amount of the "raw pixel information" contained in the original images. Further on we call this computation step an *Information Compression*. From a computational viewpoint the information compression could be considered as a *transformation* of the original large data set: $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{iK}], i = 1, 2, ..., M$, consisting of M data in the K-dimensional input space into a respective *Neural Model* consisting of N *neurons* in the same space. Here $N << M$. and $CR = M/N$ is the so called *Compression Ratio.*

The *information compression* of the original large data set (pixels or process data) can be perform by using different unsupervised competitive learning algorithms, such as clustering algorithms [2,3], the Self-Organizing (*Kohonen*) Maps [1,4], the Neural-Gas [5,6] and other versions of competitive algorithms [7,9,10] etc. The common point here is that all these algorithms try to find the most appropriate positions of the preliminary fixed number of N neurons (clusters) in the K-dimensional data space so that to resemble as much as possible the density distribution of the original data in the same space.

The essential part of any unsupervised learning algorithm is the so called *updating rule* for the neuron centers $\mathbf{c}_i, i = 1, 2, ..., N$ in the K-dimensional space. The algorithm is performed for a preliminary fixed number of T iterations ($t = 0,1,2,...,T$) as follows:

$$\mathbf{c}_i(t) = \mathbf{c}_i(t-1) + \Delta\mathbf{c}_i(t),\ i = 1,2,...,N. \tag{1}$$

Here the computation of the update $\Delta\mathbf{c}_i(t)$ varies depending on the type of the unsupervised algorithm.

The *Neural-Gas* learning algorithm [5,7], used in this paper, is a *special* version of the basic competitive unsupervised learning, where the amount of the update is computed as:

$$\Delta\mathbf{c}_i(t) = R(t) H_s(t, r_i) \left[\mathbf{x}_s - \mathbf{c}_i(t-1) \right],\ i = 1, 2, ..., N;\ s = 1, 2, ..., M \tag{2}$$

Here $R(t),\ 0 \leq R(t) \leq 1,\ t = 0,1,2,...,T$ is a monotonically decreasing L*earning Rate,* which guarantees the convergence and stability of the learning process:

$$R(t)=R_0\exp\left(-t/T_C\right),\ t=0,1,...,T \tag{3}$$

The so called *Neighborhood Function* in (2) $0 \le H_s(t,r_i) \le 1$ also decreases exponentially with the iterations. It computes the dynamically changing (decreasing) *activity area* for each neuron during the iterations, as follows:

$$H_s(t,r_i)=\exp\left[-(r_i-1)/B(t)\right],\ t=0,1,...,T;\ s=1,2,...,M;\ i=1,2,...,N \tag{4}$$

$$\text{where}\quad B(t)=\exp\left(-t/T_W\right),\ t=0,1,...,T \tag{5}$$

Here $r_i \in [1,2,...,N]$ is an integer number for the so called *ranking position* of the *i*-th neuron ($i = 1,2,...,N$) to the *s*-th data point ($s = 1,2,...M$). This ranking position is defined according to the distance between the *i*-th neuron and the *s*-th data point. The closest neuron (in a sense of a minimal *Euclidean* distance) is called "winning neuron" and gets ranking $r = 1$. The second closest neuron gets $r = 2$ and so on.

The *Initial Learning Rate* R_0 and the *Steepness* parameters: T_C and T_W have to be set prior to the learning. In the further simulation we use the following settings: $T = 500; R_0 = 0.16$ and $T_C = T_W = T/5$. A simple example for information compression of a two-dimensional "raw data" set with $M = 800$ data by using $N = 60$ neurons is given in Figure 2, while the next Figure 3 illustrates the compression of the original *3*-dimensional RGB pixel data of a test image by a fixed number of $N = 62$ neurons.

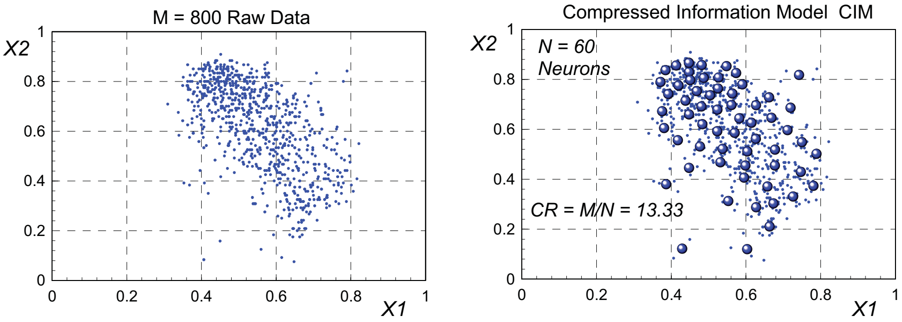


Figure 2. Example for creating Compressed Information Model (CIM) with $N = 60$ neurons from a process data set containing $M = 800$ "raw" data by the explained above Neural-Gas unsupervised learning algorithm.

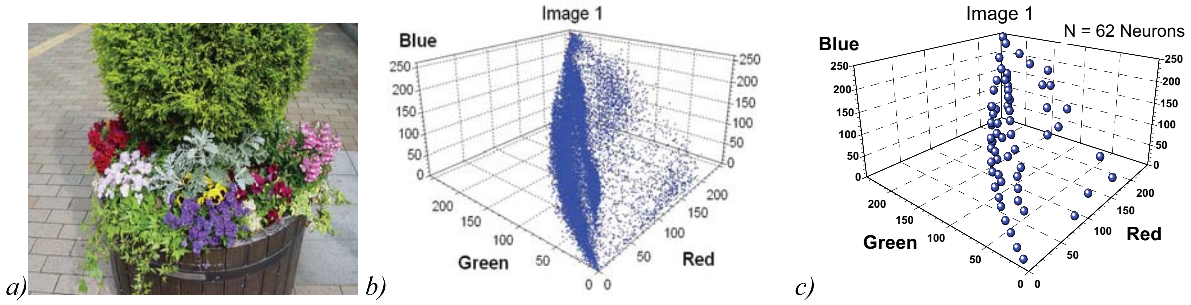


Figure 3. Example of *a)* Image; *b)* Raw Data (RGB pixels) and *c)* Compressed Information Model (CIM).

4 Computing the features for the similarity analysis

As seen from Figure 1 in Section 2, in order to evaluate the similarity between a given pair of mages or data sets, we have to evaluate two parameters (two features) for each pair. There could be several important parameters that characterize each image (based on its respective CIM) and its relation to any other image. In this paper we propose to use the following two parameters: 1) *Center-of-Gravity* and 2) *Weighted Average Size* of the CIM.

1) The *Center-of-Gravity* $\mathbf{CG}=[CG_1, CG_2, ..., CG_K]$ of a *K*-dimensional data set or image ($K = 3$) is a vector can be computed directly from the respective CIM as follows:

$$CG_j = \sum_{i=1}^{N} c_{ij}g_i \bigg/ \sum_{i=1}^{N} g_i,\ j=1,2,...,K \tag{6}$$

Here $c_{ij}, j=1,2,...,K$ denotes the *center* (coordinates) of the *i*-th neuron in the *K*-dimensional parameter space and $0 < g_i \le 1,\ i=1,2,...,N$ are the *normalized weights* of the neurons:

$$g_i = m_i / M;\ i = 1,2,...,N \tag{7}$$

$m_i \leq M$, $i = 1,2,...,N$ is the number of the data points: $\mathbf{x}_s$, $s=1,2,...,m_i$ for which the *i*-th neuron is a *winning neuron* (i.e. the neuron with the shortest *Euclidean* distance to all of these data points, as defined in [5]). Obviously, the following equation holds: $\sum_{i=1}^{N} m_i = M$ and therefore $\sum_{i=1}^{N} g_i = 1$.

2) The *Weighted Average Size* **WAS** that we propose here is a scalar value, which takes into account the normalized weights of all neurons and the Euclidean distance ED_{pq} between all pairs of neurons, $\{p,q\}$, $p=1,2,...,N$; $q=1,2,...,N$ as shown in the next two equations (8) and (9):

$$WAS = \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} ED_{pq} w_{pq} \Bigg/ \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} w_{pq} \ . \tag{8}$$

$$w_{pq} = g_p \times g_q,\quad p=1,2,...,N;\ q=1,2,...,N \tag{9}$$

Then, as input parameters *P1* and *P2* for the similarity analysis scheme in Figure 1, we use the *distance* CGD between the centers-of-gravities as well as the *difference WSD* between the weighted average sizes for each pair {A,B} of images, as follows:

$$P1 = CGD_{AB} = \sqrt{\sum_{j=1}^{K} [CG_j^A - CG_j^B]^2} \tag{10}$$

$$P2 = WSD_{AB} = \left| WAS_A - WAS_B \right| \tag{11}$$

5 Fuzzy rule based similarity analysis

We assume here to use a two-input *Fuzzy Rule Based Inference Procedure* for similarity analysis of a given pair (*A,B*) of images, in which the parameters *P1* and *P2* from (10) and (11) are used as inputs. Since the fuzzy inference procedure has many parameters that can be tuned (manually or algorithmically), it is important to finally optimize these parameters, according to a predefined human preference for similarity.

As well known [2,3,8], the most common fuzzy decision procedure consists of the following three main computation steps, as follows:

1) *Fuzzyfication* (with triangular Membership Functions);

2) *Fuzzy Inference* (with Product Operation) and

3) *Defiuzzification* (Weighted Mean Average).

From a theoretical viewpoint, the Fuzzy Rule Based Procedure is a two-input / one output fuzzy system, as follows: $D = \mathbf{F}(P1, P2)$. Here $0.0 \leq D \leq 1.0$ is the *Difference Degree* (or *Dissimilarity Degree*). $D = 0$ denotes that the images A and B are *identical* (equal), while $D = 1$ means that the A and B are *completely different* images.

For the next simulation in the paper, we assume *five triangular membership functions* that characterize linguistically the two inputs (parameters) *P1* and *P2*, as shown in Figure 4. They are used for the *fuzzification step* and have the following linguistic meaning: *VS = Very Small*; *SM = Small*; *MD = Medium*; *BG = Big* and *VB = Very Big*. The positions of these membership functions have been roughly tuned, based on the data from the test example of *16* images, described in the next Section.

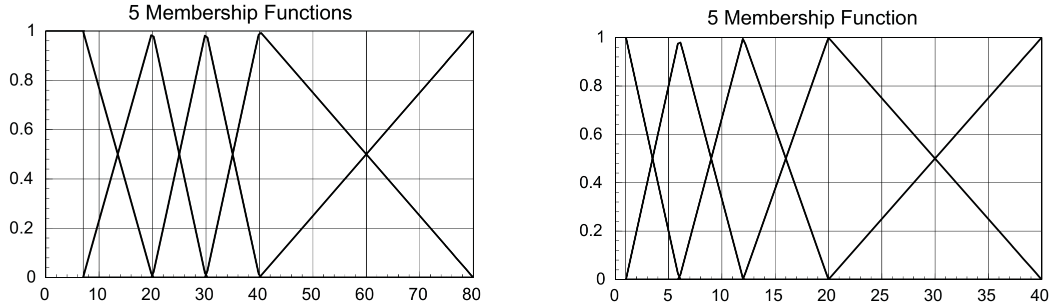
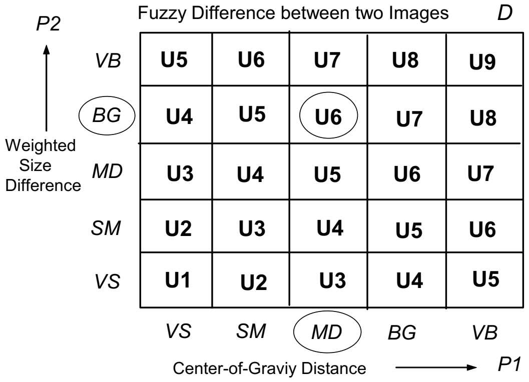


Figure 4. Five triangular membership functions assumed for the Features *P1* and *P2*.

The Fuzzy Rule Base for the Fuzzy Inference procedure is shown in the next Figure 5. It consists of *25* fuzzy rules that have individual outputs in the form of crisp values (*Singletons*): $U_1, U_2, ..., U_9$, as shown in the figure. This fuzzy rule base has been generated by using general *human logic and experience* in evaluating the similarity between the images.



Fuzzy Difference between two Images D

P2 (Weighted Size Difference)					
VB	U5	U6	U7	U8	U9
BG	U4	U5	U6	U7	U8
MD	U3	U4	U5	U6	U7
SM	U2	U3	U4	U5	U6
VS	U1	U2	U3	U4	U5
	VS	SM	MD	BG	VB

Center-of-Graviy Distance → P1

Figure 5. The fuzzy rule base with *25* fuzzy rules, used for the similarity analysis.

The following numerical values for all nine singletons: $U_1, U_2, ..., U_9$ have been assumed in this simulation:

$$\begin{gathered} U_1 = 0.0; \quad U_2 = 0.125; \; U_3 = 0.250; \\ U_4 = 0.375; \; U_5 = 0.500; \; U_6 = 0.625; \\ U_7 = 0.750; U_8 = 0.875; \; U_9 = 1.0 \end{gathered} \tag{12}$$

It is obvious that any changes (tuning) of the singletons would produce different final results from the fuzzy rule based similarity analysis. Therefore their values should be optimized according to a human specified performance criterion.

The *response surface* of the fuzzy inference procedure that uses the Fuzzy Rule Base from Figure 5 and the Membership Functions from Figure 4 is shown in the next Figure 6.

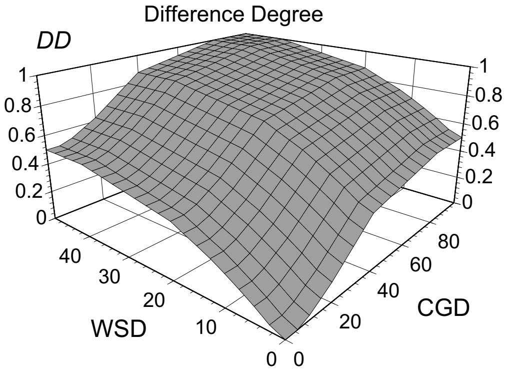


Figure 6. Non-linear response surface of the fuzzy inference procedure by using the fuzzy rule base from Figure 5 and membership functions from Figure 4.

This surface is computed by using the following standard *Defuzzification* procedure:

$$D = \sum_{i=1}^{L} u_i v_i \Big/ \sum_{i=1}^{L} v_i \tag{13}$$

Here $0 \leq v_i \leq 1,\ i = 1,2,...,L$ is the *Firing* (*Activation*) *Degree* of the *i*-th fuzzy rule and $L = 25$ is the total number of the fuzzy rules. All rules have their individual crisp values (Singletons): $u_i \in [U_1, U_2, ..., U_9], i = 1,2,...,L$, according to the notations of the Fuzzy Rule Base from Figure 5 and equations (12). For example the crisp output of the Fuzzy Rule *No. 14* (*i = 14*) from Figure 5 is as follows:

$$\text{IF}(P1 \textit{ is } \text{MD AND } P2 \textit{ is } \text{BG})\text{THEN } u_{14} = U6 = 0.625 \tag{14}$$

6 Test example for similarity analysis and classification of images

In order to illustrate the whole proposed process for similarity analysis and unsupervised classification, we present the following example of *16* images of different flowers, as shown in Figure 7.

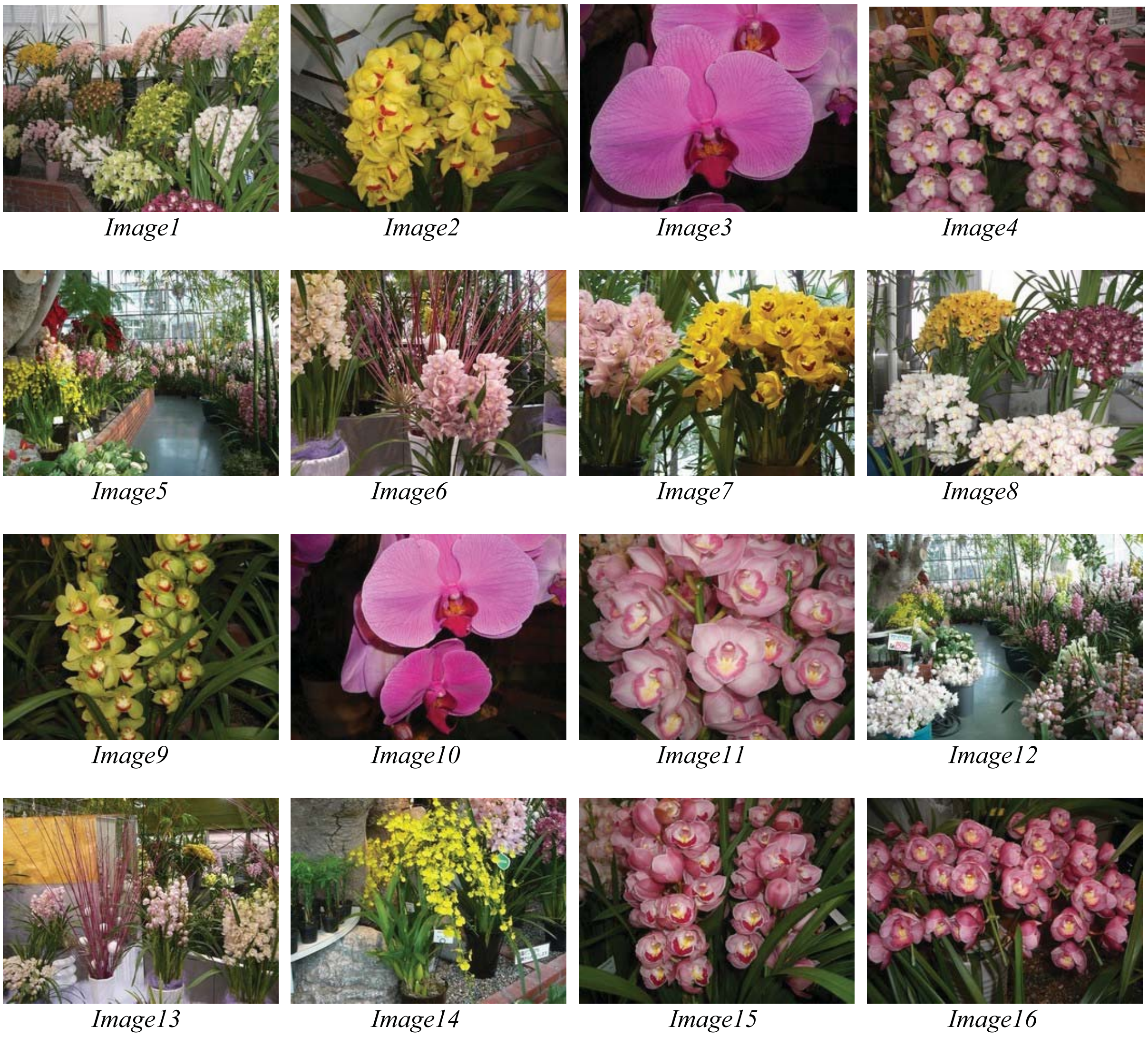


Figure 7. Images of 16 different flowers, used as example for fuzzy similarity analysis and unsupervised classification.

The BMP files of the original images with resolution of *256* x *192* = *49152* pixels have been preprocessed in order to extract the respective RGB files with the *3*-dimensional "raw data" that contain all pixel values of the image within the range of *0-255*. Then, the described algorithm for information compression in Section 3 has been used to generate all *16* models (CIMs) of the images by using equal number of neurons, namely *N* = *50*..

After that, the Center-of-Gravities **CG** of all *16* images from Figure 7 have been computed from the obtained CIMs by using equation (6). Their locations in the *3*-Dimensional RGB space are shown in Figure 8. The weighted average sizes: *WAS* for the images were computed by (8) and are shown in Figure 8.

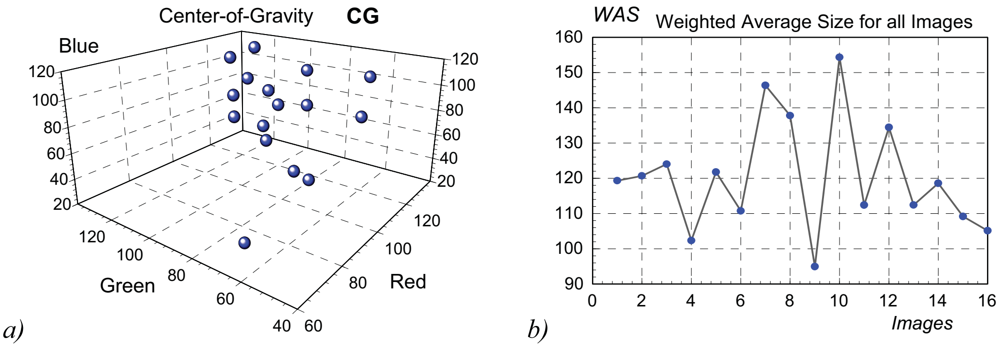


Figure 8. The center-of-gravities *a)* and the weighted average sizes *b)* computed for all *16* test images from their respective Compressed Information Models.

Finally, the parameters *P1* = *CGD* and *P2*= *WSD* used for the fuzzy similarity analysis, according to the flow-chart in Figure 1 were computed by equations (10) and (11) respectively.

7 Results from the similarity analysis and classification

The unsupervised classification scheme, presented in Figure 1 was applied for performing different classifications of the test set of *16* images from Figure 7. For this purpose, *four* different Image Bases with *3, 4, 5* and *6* core images respectively were used for the classification, as follows:

Image Base3 = {Image1, Image2, Image3}
Image Base4 = {Image1, Image2, Image3, Image4}
Image Base5 = {Image1, Image2, Image3, Image4, Image5}
Image Base6 = {Image1, Image2, Image3, Image4, Image5, Image6}

The classification procedure is performed according to the following rule: An external image (with respect to the assumed Image Base) is considered *preliminary classified* as belonging to that class (*core* image from the Image Base), for which it has the best similarity, i.e. the class with the minimal dissimilarity: *D* = D*min*.

Next Figure 9 presents the summarized results from the classification of the *16* test images by using each of the above Image Bases separately, namely: *Image Base3*, *Image Base4, Image Base5* and *Image Base6.* The computed minimal dissimilarity degrees D*min* for all external images (that do not belong to the assumed Image Base) are given in Figure 9*a)*, 9*c)*, 9*e)* and 9*g)*. The right part of Figure 9, namely 9*b)*, 9*d)*, 9f*)* and 9*h)* presents the results from the *preliminary* classification of the external images for all *four* Image Bases.

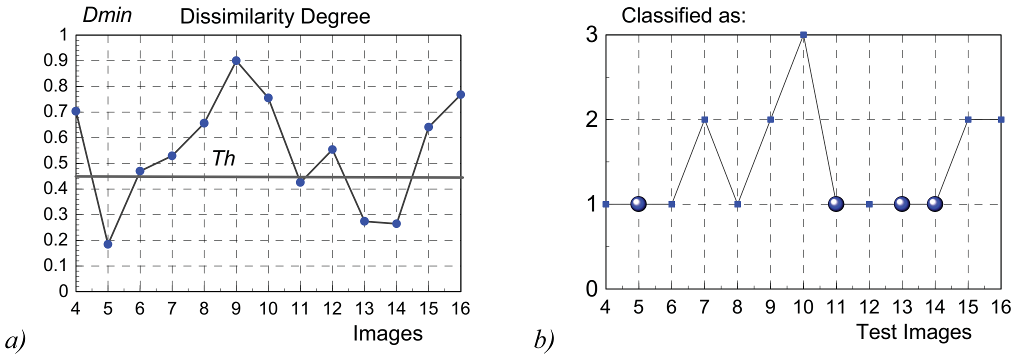


Image Base3 = {Image1, Image2, Image3}

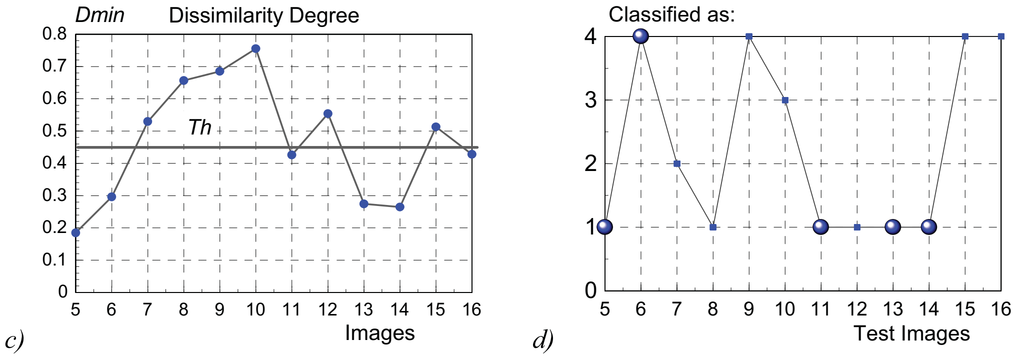


Image Base4 = {Image1, Image2, Image3, Image4}

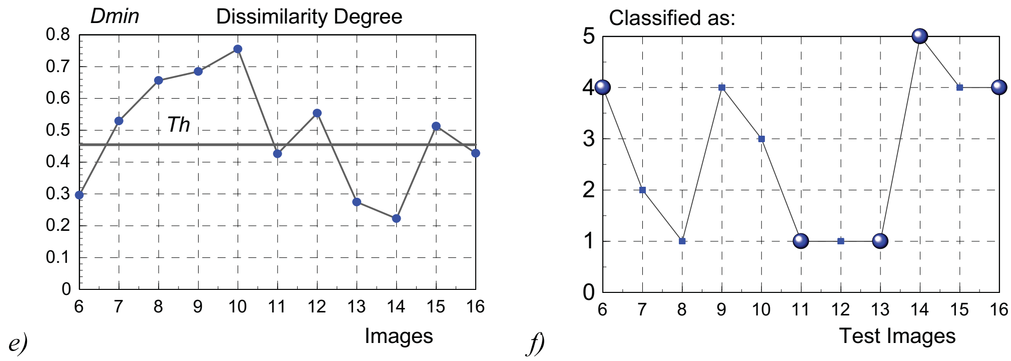


Image Base5 = {Image1, Image2, Image3, Image4, Image5}

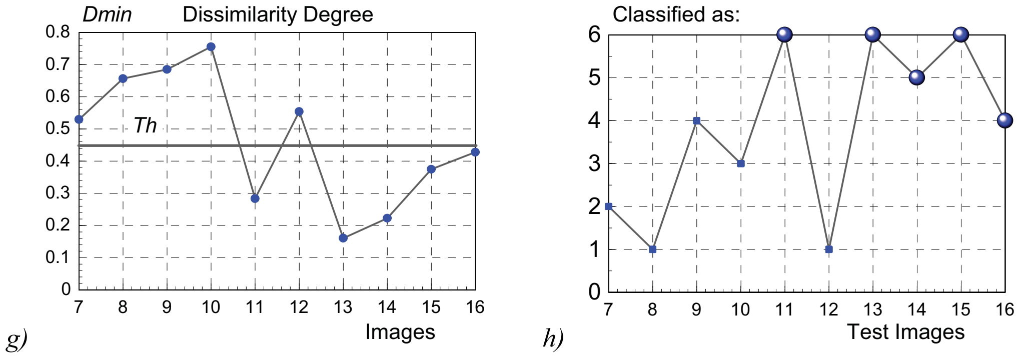


Image Base6 = {Image1, Image2, Image3, Image4, Image5, Image6}

Figure 9. Results from the classification of the test *16* images with four different Image Bases, that contain *3,4,5* and *6* images, respectively. Left part of the Figure depicts the computed minimal Dissimilarity Degrees *Dmin* for each external image. Right part of the Figure shows the classification results. Classifications shown by using large balls symbols are assumed as *final classifications*, according to the pre-specified threshold: *Th = 0.45.*

Now, if a human defined threshold *Th* is established beforehand, then a more *plausible* classification could be obtained, based on the following general rule:

If *Dmin < Th,* **then** the classification of the current image is *confirmed*, i.e. the image is *finally* classified as belonging to that class of the core Image Base with a dissimilarity value of *Dmin*; otherwise the image is *rejected* and considered as a *new* and *different image*. In such case it can be included as a *new member* of the Image Base, thus creating a *new class* in it.

If a predefined threshold *Th = 0.45* is used, as shown in Figure 9, several images are confirmed as *finally classified* with respect to the respective *Image Base.* They are shown as large ball symbols. The other results, shown as small circles, correspond to the *rejected* images that are quite different from the others and could be used to create new classes.

Let us assume that the *Image Base6* with *6* core images is used for the classification. Then, according to the results shown in Figure 9*g*) and 9*h*), the following *5* external images are *finally classified,* such that:

Image11 is similar to *Image6* (questionable);
Image13 is similar to *Image6* (almost right);
Image14 is similar to *Image5* (right);
Image15 is similar to *Image6* (questionable);
Image16 is similar to *Image4* (right);

It is seen that the *computer classification* decision sometimes differs from the *human classification* (shown in the parenthesis). With appropriate tuning of all the parameters in the fuzzy inference procedure from Section 5, it is possible to increase the accuracy (the plausibility) of the classifications, thus making it closer by results to the expert human decisions.

8 Conclusions

A two stage unsupervised classification scheme for images is proposed in this paper that can be used for incremental classification of images and large process data sets. It is based on unsupervised learning for information compression of the raw data with subsequent fuzzy similarity analysis, based on a two-input fuzzy inference procedure.

As a first step, the “raw” pixel data from the images are used by a special unsupervised learning algorithm for creating respective *compressed information model* (CIM) with small number of neurons in the *3*-dimensional (RGB) space. After that two essential model parameters, namely the *center of gravity* and the *weighted average size* are computed from the obtained CIM. Then the *Center-of-Gravity Distance* and the *Weighted Average Size Difference* for each pair of images are used as inputs in the proposed fuzzy inference procedure for similarity analysis. This similarity analysis is made for the pairs that include each new image and all the core images from the predetermined Image Base. The similarities are expressed numerically as *difference degrees* (*dissimilarities*) between *0* and *1* for all pairs of images and are sorted in increasing order for the final classification result.

An application example for classification of *16* test images of flowers is shown in the paper, by using *4* different Image Bases with *3, 4, 5* and *6* “core” images, respectively.

A special human decided *threshold* is introduced and used in order to properly classify (or reject) the new image as belonging to one of the core images or to reject is, as unknown (or new) image.

It should be noted that the tuning parameters of the fuzzy similarity procedure can affect significantly the final classification results, i.e. the decisions of "accepting" or "rejecting" an image to the certain "core" image.

If a special human preference (or solutions) for some classification cases are available, then the computation scheme would become a kind of *semi-unsupervised* (or *human-assisted*) classification procedure. Here the tuning parameters of the procedure for fuzzy similarity analysis could be appropriately tuned so as to fit as much as possible the human expert decisions and preferences.

The tuning parameters of the fuzzy similarity analysis can be divided into two groups, namely the membership functions parameters and the singletons (outputs of the fuzzy rules). Basically this is a *multivariate optimization* problem with many possible local optima so that genetic algorithms or particle swarm optimization algorithms could be a good choice.

Another point of special attention is the construction of the *optimization criterion* that should make the most use of the information for the human preferences. Finally, the method and the strength of *penalizing* the unsuccessful guesses of the procedure is also another point of consideration.

The further research is aimed in the above mentioned directions for improving the plausibility of the proposed scheme for similarity analysis and classification.

9 References

[1] Alahakoon, D., Halgamuge, S.K. and Srinivasan, B.: *Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery*, IEEE Trans. On Neural Networks, vol. 11, No. 3, (2000), 601 - 614.

[2] Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[3] Bishop Ch.M.: *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 2003.

[4] Kohonen, T.: *Self-Organizing Maps*, Third Edition, Springer Series in Information Sciences, Springer, Berlin, 2001.

[5] Martinetz T., Berkovich S. and Schulten, K.: *Neural-Gas Network for Vector Quantization and Its Application to Time-Series Prediction*, IEEE Trans. Neural Networks, vol. 4, No. 4, (1993), 558 - 569.

[6] Vachkov, G.: *Classification of Machine Operations Based on Growing Neural Models and Fuzzy Decision*, In: *CD-ROM* Proc. 21st European Conference on Modelling and Simulation, ECMS 2007, Prague, Czech Republic, June, 2007, 68 – 73.

[7] Vachkov, G.: *On-Line Unsupervised Learning for Information Compression and Similarity Analysis of Large Data Sets*, In: CD-ROM Proc. 2007 IEEE Int. Conference on Mechatronics and Automation, ICMA 2007, Harbin, China, August, 2007, 105 - 110.

[8] Vachkov, G.: *Classification of Images Based on Information Compression and Fuzzy Rule Based Similarity analysis,* In: Proc. World Congress on Computational Intelligence WCCI 2009, Hong Kong, June, 2008, 2326 – 2332.

[9] Xu, L., Krzyzak A. and Oja, A.: *Rival Penalized Competitive Learning for Clustering Analysis, RBF Net and Curve Detection*, IEEE Trans. Neural Networks, vol. 4, No. 4, (1993), 636 - 649.

[10] Zhang, Ya-Jun and Liu, Zhi-Qiang: *Self-Splitting Competitive Learning: A New On-Line Clustering Paradigm*, IEEE Trans. on Neural Networks*,* vol. 13, No. 2, (2002), 369 - 380.