Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 173 – 180

ISBN 978-83-60810-14-9
ISSN 1896-7094

Image Similarity Detection in Large Visual Data Bases

Juliusz L. Kulikowski
Institute of Biocybernetics and
Biomedical Engineering, Polish
Academy of Sciences
4, Ks. Trojdena Str.
02-109 Warsaw, Poland
E-mail: jkulikowski@ibib.waw.pl

***Abstract*—A method of similarity clusters detection in large visual databases is described in this work. Similarity clusters have been defined on the basis of a general concept of similarity measure. The method is based also on the properties of morphological spectra as a tool for image presentation. In the proposed method similarity of selected spectral components in selected basic windows are used to similarity of images evaluation. Similarity clusters are detected in an iterative process in which non-perspective subsets of images are step-by-step removed from considerations. In the method similarity graphs and hyper-graphs also play an auxiliary role. The method is illustrated by an example of a collection of medical images in which similarity clusters have been detected.**

I. Introduction

VISUAL data bases (*VDB*) are widely used in various experimental investigation areas [2, 3, 5, 6, 21]. Documents stored in *VDB*s consist of a digital representation of an image (e.g. given in the form of a bit-map), of a sequence of formal data identifying the document and, possibly, of a series of qualitative and/or quantitative attributes characterizing image content. One of the main problems in *VDB* exploration is retrieval of visual documents satisfying the formal and content requirements given by the users. A typical query coming from a *VDB* user concerns all available visual documents satisfying formal (type, source, emission data, etc.) as well as content requirements [5, 20]. However, in certain cases another type of queries concerning visual documents is possible: among a class of visual documents satisfying some general formal requirements, find all subsets of documents forming, in a below-defined sense, *similarity clusters.* In this case it is assumed that a concept of *similarity* has been defined by the user (instead of defining the image content attributes). Moreover, the number of similarity clusters is beforehand neither defined nor limited, single documents as similarity clusters being out of interest. On the other hand, some documents, as it will be shown below, can be included into more than one similarity cluster. We call the above-formulated problem a *similarity clusters detection* (*SCD*) *problem*. It should be emphasized that there is a substantial difference between the well known problem of strong classification of objects [1,16] and this of *SCD*. The difference is caused, in general, by non-transitivity of similarity relation, as it can be illustrated by an example of a dactyloscopic database. The classification problem consists in this case in assigning fingerprints to similarity classes strongly defined on the basis of dactyloscopic patterns and minutes (bridges, meshes, forks, line ends, etc.). Similarity classes are pair-wise disjoint and each object belongs to exactly one similarity class. On the other hand, a *SCD* problem may consists in finding all subsets of fingerprints containing, at least, one subset of minutes of the corresponding types forming the same geometrical configuration. A visual document, due to several types of minutes satisfying the above-formulated similarity criterion, can be included into more than one similarity clusters. Solution of the *SCD* problem, as leading to a class of *NP*-complete numerical tasks, in the case of large visual databases leads to high calculation costs.

The aim of this paper is presentation of a method of reducing the calculation costs of *SCD* due to a multi-step similarity detection strategy. The strategy is based on a concept of a step-wise strengthening of similarity criteria connected with elimination of not satisfying them pairs of documents. This concept is realized due to an additional concept of using *morphological spectra* to image description; the concept was formerly used to partial similarity of documents detection [11].

The paper is organized as follows: in Sec. II basic notions of *similarity measure ε-similarity clusters* (Sec. II *A*), and of *representation of images* (Sec. II *B*) are shortly reminded. In Sec. III the concept of multi-step similarity detection is presented generally (Sec. III *A*) and an algorithm of ε-similarity clusters detection in (Sec. III *B*) is described. An example illustrating using the proposed method to a collection of medical images is presented in Sec. IV. Conclusions are formulated in Sec. V.

II. Basic Notions

The notion of *similarity* plays a basic role in pattern recognition. In the strongest sense it can be considered as a synonym of *equivalence*, i.e. a binary relation satisfying the reciprocity, symmetry and transitivity conditions [9]. However, such similarity concept does not suit well to a description of similarity of images where the transitivity condition

978-83-60810-14-9/08/$25.00 © 2008 IEEE

is often not satisfied. In a wider sense, similarity is a sort of *neighborhood* relation (reciprocal and symmetrical) rather than this of equivalence.

I. Similarity measures and similarity clusters

A numerical characterization of similarity in wider sense is possible due to a *similarity measure c*oncept. This concept can be defined in several ways [9,14]. In pattern recognition similarity measure is usually defined on the basis of a *distance measure* or of a *cosine* (angular) measure [11]. That is why the following definition of similarity measure below is given (see also [9,11]):

Definition 1. Let C be a set of elements and let $a, b, c \in C$ be any of its members; then a function:

$$\sigma: C\times C \rightarrow [0,\ldots,1] \qquad (\mathbf{1})$$

satisfying the conditions:

I. $\sigma(a,a) \equiv 1$,

II. $\sigma(a,b) \equiv \sigma(b,a)$,

III. $\sigma(a,b) \cdot \sigma(b,c) \leq \sigma(a,c)$

will be called a *similarity measure* described on C ●

In particular, if C is a metric space [8] and $d(a,b)$ is a distance measure of the given pair of its elements then a function:

$$\sigma(a,b) = exp\,[-\alpha\cdot d(a,b)], \qquad (\mathbf{2})$$

α being a positive scaling coefficient, satisfies the conditions of Definition 1.

Another possibility arises if C is assumed to be a linear unitary space [19]. In such case $\boldsymbol{a}$, $\boldsymbol{b}$, etc. are interpreted as *vectors*, $(\boldsymbol{a},\boldsymbol{b})$ denotes their *scalar product*, $\|\boldsymbol{a}\| = (\boldsymbol{a},\boldsymbol{a})^{½}$ is a *norm* of $\boldsymbol{a}$ and the cosine of the angle $\angle(\boldsymbol{a},\boldsymbol{b})$ between the vectors is given by the well-known formula:

$$\cos(a,b)=\frac{(a,b)}{\|a\|.\|b\|} \qquad (3)$$

However, *cos(**a**,**b**)* cannot be used as a similarity measure satisfying the condition III of Definition 1. For this purpose it will be used the following *angular similarity measure*:

$$\sigma(a,b)=1-\sqrt{1\text{-}\cos^2(a,b)} \qquad (4)$$

On the basis of similarity measure a concept analogous to this of *similarity classes* used in equivalence (i.e. strong similarity) relation can be introduced.

Definition 2. Let C be a set of elements, $\sigma(a,b)$ be a similarity measure defined on C and let ε such that $0 \leq \varepsilon \leq 1$ be an arbitrary constant. Then a subset $\Phi_\varepsilon \subseteq C$ such that:

1 st for any pair of its elements $a, b \in \Phi_\varepsilon$ it is $\sigma(a,b) \geq \varepsilon$, and

2 nd for each element c belonging to $C \setminus \Phi_\varepsilon$ there is at least one element a in Φ_ε such that the inequality $\sigma(a,c) < \varepsilon$ is satisfied,

will be called an *ε-similarity cluster* ●

Let us remark that, excepting the case of ε = 1, there is a substantial difference between the strong *similarity class* concept and this of the *ε-similarity cluster*. It is illustrated in Fig. 1 where a set C consisting of 4 elements on an Euclidean plane is shown. The elements are located in the vertices of a square of unit edge-lengths. A similarity of vertices has been defined on the basis of their Euclidean distance and the ε - similarity clusters consist of subsets of vertices whose distance is not greater than 1. In such case it is clear that two ε - similarity clusters can be established in two alternative ways; such situation in strong similarity classes could not arise.

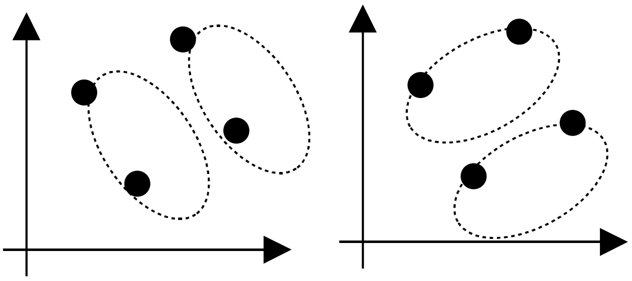

Fig. 1 . Two alternative ways of choosing ε -similarity clusters.

In similarity of objects evaluation a multi-aspect similarity can be taken into consideration by using several similarity measures. For this purpose, if $\sigma_1(a,b)$, $\sigma_2(a,b)$, $\ldots,\sigma_f(a,b)$ are similarity measures satisfying the conditions of Definition 1 and representing different similarity aspects then it can be shown [10] that

$$\sigma(a,b)=\prod_{\phi=1}^{f}\sigma_{\phi}(a,b) \qquad (\mathbf{5})$$

also satisfies the conditions of Definition 1 and as such it can be used as a mutli-aspect similarity measure of the given objects. This property will be used in ε - *SCD* of images.

II. Spectral representation of images

It will be considered representation of monochromatic images in visual databases in a basic form of *bitmaps*, i.e. numerical $I \times J$ rectangular matrices, where I and J denote, respectively, the number of rows and columns. A bitmap representing an image $\boldsymbol{u}$ will be denoted by $\boldsymbol{U}$ while its elements u_{ij}, called pixel values, will be assumed to be integers from a finite interval $[0,\ldots,2^k-1]$, k being a fixed natural number. Below, expansion of a bitmap $\boldsymbol{U}$ into a linear $I\cdot J$-component column vector $\boldsymbol{V}$, as more convenient for calculations, will also be used.

Images can also be represented in several alternative, spectral forms [10,17]. Below, image representation by systems of morphological spectral components will be considered [12,13].

Morphological spectra are basically defined for monochromatic images given in the form of $2^m \times 2^m$ - size bitmaps, where m is a fixed natural number such that $2^m \leq min\,(I,J)$. Morphological spectra form a hierarchical structure, m being the highest level of the hierarchy and the 0^{th} level being identified with the original image. The h-th level spectral components, where $0 \leq h \leq m$, are calculated on *basic windows* of $2^h \times 2^h$ size. Therefore, it is assumed that for calculation of any *h*- th level morphological spectrum the original bitmap is partitioned into a number of adjacent basic windows covering the image area.

#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
#	#	#	#	#	#	#	#	#	0
0	0	0	0	0	0	0	0	0	0

Fig. 2. Partition of an image area into basic windows.

If necessary, the image area can be covered with certain margins as shown in Fig. 2, where an image of a 7×9 (*I*= 7, *J*= 9) size has been covered by basic windows of $2^1 \times 2^1$ (i.e. *h* =1) size, the lacking pixel values in the extreme right column and in the lowest row being filled with 0-s.

In each basic window the values of a fixed *h* -th level spectral component are calculated independently. Therefore, if *p, q* denote, respectively, the number of rows and columns of basic windows covering the image (*p* =4, *q* =5 in the above-shown case) then the values of each *h* -th level spectral component of a total image can be collected in a $p \times q$ real matrix called a *spectral component matrix* . Any *h* -th level morphological spectrum consists of 2^{2h} types of components. For *h* >0, spectral components are labeled by *h*- element strings consisting of symbols *Σ*, *V, H, X* ; the 1st level morphological spectrum contains four components labeled by single symbols *Σ*, *V, H* and *X* only. The 2nd level morphological spectrum contains 2^4=16 components labeled and lexicographically ordered as follows: ***ΣΣ***, *Σ V, Σ H, Σ X, V Σ, VV, VH, VX, H Σ, HV, HH, HX, X Σ, XV, XH* and *XX.* The 3rd level spectral components are denoted by ***ΣΣΣ***, ***ΣΣ*** *V*, ***ΣΣ*** *H*, … etc . The components of morphological spectra can be thus represented by a regular tree whose nodes on a given level are assigned to the given spectral-level components [17]. The spectral component labels are used to a denotation of spectral component matrices. For example, $\boldsymbol{M}_V$, $\boldsymbol{M}_{VX}$, $\boldsymbol{M}_{\Sigma\Sigma H}$, etc. denote, respectively, the spectral component matrices of the spectra *V, VX* and *ΣΣ H.* For a given image size the size of the corresponding spectral component matrices depends on the spectrum level *h* and is decreasing with it. Hence, the number of elements representing a given image on each spectrum level corresponds to the number of pixels in the original image (it is exactly equal to this number if the image area can be covered without margins by the highest-level basic windows).

Morphological spectra can be calculated by using *spectral matrices* [13] . For this purpose for each (*h*-th) spectrum level it is constructed a matrix $\boldsymbol{M}^{(h)}$ of $4^h \times 4^h$ size whose rows correspond to lexicographically ordered spectral components and columns are assigned to the lexicographically ordered pixels in the basic window. For example, the 1st level morphological spectrum can be represented by a matrix:

$$M^{(1)}=\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \end{bmatrix} \qquad (6)$$

Each row of the spectral matrix consisting of the elements +1 and –1 only represents the weights assigned to pixel values whose sum within a basic window should be calculated. Spectral matrices of any level satisfy the following orthogonality condition:

$$\boldsymbol{M}^{(h)} \cdot (\boldsymbol{M}^{(h)})^{tr} = (\boldsymbol{M}^{(h)})^{tr} \cdot \boldsymbol{M}^{(h)} = 4^h \cdot \boldsymbol{I} \qquad (7)$$

where ^{tr} denotes matrix transposition and $\boldsymbol{I}$ is an unity matrix of $2^{2h} \times 2^{2h}$ size. If pixel values of a basic window are presented in the form of a vector:

$$\boldsymbol{V}^{(h)} = [\xi_{1,1},\xi_{1,2},\ldots,\xi_{1,K},\xi_{2,1},\ldots,\zeta_{2,K},\ldots,\xi_{K,1},\ldots,\zeta_{K,K}] \qquad \textbf{(8)}$$

where $K = 4^h$, then the morphological spectrum of the basic window can be calculated as:

$$(\boldsymbol{W}^{(h)})^{tr} = \boldsymbol{M}^{(h)} \cdot (\boldsymbol{V}^{(h)})^{tr}. \qquad \textbf{(9)}$$

A formula for reverse bitmap components calculation from the morphological spectral components then takes the form:

$$\boldsymbol{V}^{(h)} = 4^{-h} \cdot \boldsymbol{W}^{(h)} \cdot \boldsymbol{M}^{(h)}. \qquad \textbf{(10)}$$

An important property of morphological spectra as tools for image representation consists in their multi-scalar structure and possibility to focus image description in given regions of interest covered by selected elements of the spectral component matrices. These properties will be used below in a concept of step-wise *ε-SCD* of images.

III. Multi-step similarity clusters detection

It will be assumed that there is given a finite set *C* of images available in the form of bitmaps and a similarity measure *σ* defined on the Cartesian product *C*×*C*. The problem consists in finding in *C*, for a certain *ε*, $0 \le \varepsilon \le 1$, (usually kept close to 1) all *ε*-clusters. If *N* = |*C*| denotes the number of elements in *C* then the simplest (but very time-consuming) solution method can be based on a direct review of all ½*N*(*N*-1) pairs of analyzed objects (images). The algorithm will then consist of the following (roughly defined) steps:

Algorithm 1 (naïve):

Step 1: take into consideration all unordered pairs of objects $\{ a ,b \} \in C \times C$;

Step 2: for each pair calculate its similarity measure $\sigma(a,b)$;

Step 3: accept all pairs for which the condition $\varepsilon \le \sigma(a,b) \le 1$ is satisfied and reject the other ones;

Step 4: for the accepted pairs construct a set *C** , $C^* \subseteq C$, of all occurring in them objects;

Step 5: construct a *similarity graph* *G* = [*C** , *S, φ*] where *A* is a set of its *nodes, S* is a set of undirected *edges* , and *φ* is a function assigning, in an unique way, to each accepted pair of nodes an edge from the set *S* ;

Step 6: using a standard algorithm of finding *cliques* in graphs find all cliques in *G* ;

Step 7: end ●

The cliques (maximal complete subgraphs) of the similarity graph *G* found by the algorithm are, by the same, the ε-similarity clusters. On the other hand, it should be remarked that if *F* is a family of cliques of the similarity graph *G* then a triple:

$$H = [C^*, F, \psi] \tag{11}$$

defines a hyper-graph in the Berge sense [7], where C^* is a set of nodes, ***F*** is a set of hyper-edges and ψ is a partial function described on the family 2^{C^*} of the subsets of C^*, assigning in an unique way the hyper-edges of *F* to selected subsets of nodes. We shall call *H* a *similarity clusters hyper-graph*; its role in *SCD* algorithms will be shown below.

However, reducing a *SCD* problem to this of a classical *cliques* finding problem is practically not a satisfying task solution because of its non-polynomial (*NP*) complexity. Despite the fact that there are several approximate algorithms for it described in the literature [4,15], an approach to the calculation cost reduction based on a step-wise restricting of ε-similar cliques is proposed below.

I. General concept description

For the purpose of comparison of images they will be considered as ordered sets of sub-bitmaps according to their partition into basic windows, as shown in Fig. 2. A bitmap ***U*** of a given image ***u*** will be thus represented by a composed matrix of sub-bitmaps:

$$\mathrm{U} = [\mathrm{U}_{\kappa\lambda}] \tag{12}$$

where $\boldsymbol{U}_{\kappa\lambda}$ is a sub-bitmap, $1 \leq \kappa \leq p$, $1 \leq \lambda \leq q$. Similarity of any given pair ***u', u"*** of images means that the pairs of the corresponding sub-bitmaps $\boldsymbol{U'}_{\kappa\lambda}$, $\boldsymbol{U''}_{\kappa\lambda}$ satisfy some similarity criteria. Similarity of any given pair of corresponding sub-bitmaps can be thus considered as an aspect of similarity of the images ***u', u"*** in the whole. According to (5), their similarity measure can be expressed as

$$\sigma(U', U'') = \prod_{\kappa=1}^{p} \prod_{\lambda=1}^{q} \sigma(U'_{\kappa\lambda}, U''_{\kappa\lambda}) \tag{13}$$

Let us denote by *L* , $L = \{(\kappa,\lambda)\}$, the set of all considered pairs (κ,λ). It is clear that due to the general properties of similarity measures the inequalities

$$0 \leq \sigma(\boldsymbol{U'},\boldsymbol{U''}) \leq min_L[\sigma(\boldsymbol{U'}_{\kappa\lambda},\boldsymbol{U''}_{\kappa\lambda})] \tag{14}$$

are held. Therefore, if for a given pair of images ***u', u"*** it is required that $\sigma(\boldsymbol{U'},\boldsymbol{U''}) \geq \varepsilon$ and for a certain pair $(\kappa,\lambda) \in L$ it has been found that $\sigma(\boldsymbol{U'}_{\kappa\lambda},\boldsymbol{U''}_{\kappa\lambda}) < \varepsilon$ then there is no reason to prove the similarity condition in other pairs of sub-bitmaps and the given pair (***U', U"***) can be considered as "non-perspective" from its ε -similarity point of view. Moreover, if we denote by *L'* , $L' \subseteq L$ a certain subset of pairs (κ,λ) of indices for which it has been found that

$$\sigma(U', U'') = \prod_{(L')} \sigma(U'_{\kappa\lambda}, U''_{\kappa\lambda}) = \varepsilon' \tag{15}$$

where ε' is a number $\geq \varepsilon$ then for satisfying the condition $\sigma(\boldsymbol{U'},\boldsymbol{U''}) \geq \varepsilon$ it is necessary that an inequality

$$\sigma_{L \setminus L'}(U', U'') = \prod_{(L \setminus L')} \sigma(U'_{\kappa\lambda}, U''_{\kappa\lambda}) \geq \frac{\varepsilon'}{\varepsilon} \tag{16}$$

is satisfied.

For similar reasons, for a fixed pair of sub-bitmaps ($\mathrm{U}'_{\kappa\lambda}$, $\boldsymbol{U''}_{\kappa\lambda}$) their similarity measure can be considered as a multi-aspect similarity measure of the corresponding pairs of morphological spectra

The below-proposed procedure of ε-*SCD* of images is based on a concept of step-by-step strengthening of similarity criteria connected with removing non-perspective pairs of images from considerations and improving accuracy of the *SCD*.

II. Choosing the ε-similarity thresholds

At an initial state of the procedure it is posed the following

Initial working hypothesis: the similarity measures of the pairs of all corresponding basic windows and spectral components of all pairs (***u', u"***) of images in the analyzed set *C* are equal 1.

By this assumption, the condition $\sigma(\boldsymbol{u'},\boldsymbol{u''}) \geq \varepsilon$ according to (5) and (14) is satisfied and *C* constitutes, as a whole, the first assumed, hypothetical ε -similarity cluster.

In the consecutive iterations of the procedure the working hypothesis by evaluation of similarity measure of selected pairs of basic windows and spectral components (selected similarity aspects) is verified. The condition $\sigma(\boldsymbol{u'},\boldsymbol{u''}) \geq \varepsilon$ is satisfied if it is satisfied by all pairs of corresponding basic windows and spectral components. As a consequence:

i. the objects whose similarity to all other objects does not satisfy the condition $\sigma(\boldsymbol{u'},\boldsymbol{u''}) \geq \varepsilon$ can be removed from considerations as non-perspective ones;

ii. the pairs of perspective (non-removed) objects whose similarity does not satisfy the above-given condition are taken into consideration; however, they bring about a necessity of correction of the currently assumed, hypothetical similarity clusters;

iii. at each iteration of the procedure the similarity threshold levels should be corrected according to the general principle described in Sec. *A* ;

iv. at each iteration the similarity measure of basic windows and of spectral components which yet have not been evaluated remain equal 1;

v. each iteration (excepting the last one) leads to a revised subset of hypothetical similarity clusters which in the next iteration in similar way should be processed.

III. The procedure of ε-similarity clusters detection

For a given initial set C of images and a fixed final similarity measure value ε, $0 < \varepsilon \leq 1$, a morphological spectrum level h determining the size of basic windows should be chosen by taking into account that the larger is this size the higher is the probability that differences between the images, if any exist, by the algorithm in each iteration will be detected. On the other hand, the higher is h the larger is the number of spectral components that should be taken into account in the *SCD* algorithm.

The ε-*SCD* procedure consists of a sequence of iterations, each iteration consisting of two phases:

1st reduction, according to the strengthened similarity criteria, of the set of compared pairs of objects;

2nd decomposition and/or reduction of the similarity clusters inherited from the former iteration.

Each (i-th) iteration of the algorithm needs the following data to be entered:

i. a starting threshold level $\varepsilon^{(i)}$, $0 < \varepsilon < \varepsilon^{(i)}$;

ii. a pair$\{(\kappa^{(i)}, \lambda^{(i)})\}$ of addresses of basic windows selected for being used in the current iteration of *SCD* procedure;

iii. a label $\boldsymbol{\Gamma}^{(i)}$ (subset of labels) of the h- th level morphological spectrum component selected for being used in the current iteration of *SCD* procedure.

No strong rules of choosing the above-mentioned data exist. However, the following, heuristic recommendations can be taken into account:

a) according to the formula (16), for any i =1,2,… the threshold level $\varepsilon^{(i)}$ should be chosen as $\varepsilon^{(i)} = \varepsilon^{*(i-1)}/\varepsilon$ where $\varepsilon^{*(i-1)}$ denotes the similarity measure of the given pair of images evaluated in the former, $(i - 1)$ iteration;

b) the basic windows selected for analysis should be, if possible, selected with a preference given to the components having the highest discriminative power.

To meet the recommendations the following, auxiliary objects are defined:

* a symmetric square matrix $\boldsymbol{S} = [\sigma^{(i)}_{\alpha\beta}]$ of $N \times N$ size, N denoting the number of elements of C; elements $\sigma^{(i)}_{\alpha\beta}$ denote the similarity measures of the pairs of bitmaps $(\boldsymbol{U}^{(\alpha)}, \boldsymbol{U}^{(\beta)})$ evaluated at the i-th iteration);

* a binary matrix $\boldsymbol{M} = [m^{(i)}_{\beta\gamma}]$, β denoting a serial number of basic window (equivalent to its (κ,λ) address), γ being assigned to a serial number of spectral component; $m^{(i)}_{\beta\gamma} = 1$ if the β-th basic window and the γ- th spectral component have been already used to the similarity of images assessment, otherwise $m^{(i)}_{\beta\gamma} = 0$.

At the initial state of the procedure all elements of $\boldsymbol{S} = \boldsymbol{S}^{(0)}$ equal 1 and all elements of $\boldsymbol{M} = \boldsymbol{M}^{(0)}$ equal 0.

In the below-presented concept of algorithm a graph representation of the current state of the ε-*SCD* procedure is also useful.

A similarity graph G (see Sec. III) is described by:

a) a subset $C^* \subseteq C$ of nodes representing the images currently considered as "perspective" for ε - *SCD* ;

b) a symmetrical square sub-matrix $\boldsymbol{S}^* \subseteq \boldsymbol{S}$ containing the rows and columns of $\boldsymbol{S}$ corresponding to "perspective" images, the elements of $\boldsymbol{S}^*$ being interpreted as weighed edges of the graph G . $\boldsymbol{S}^*$ plays the role of an adjacency matrix of G .

At a starting point the similarity graph $G = G^{(0)}$ is assumed to be a complete graph identical to its unique clique.

A current state of the ε - *SCD* task solution can also be illustrated by an ε - *similarity cluster hyper-graph* H . Its definition is given by the formula (11) excepting that F denotes a family of hyper-edges representing the currently detected, hypothetical ε - *similarity* cliques Φ_s , s=0,1,2,3,…, in G. At a starting point the hyper-graph $H = H^{(0)}$ contains a single hyper-edge: $F^{(0)} = \{\Phi_0\}$, $\Phi_0 \equiv C^*$.

For similarity of pairs of images assessment the similarity measure of spectral vectors based on the formulae (2) or (4) and (5) can be used. A progress index z also will be used to mark the situations when in the i-th iteration the necessity of similarity clusters hyper-graph H correction has been detected.

The i-th iteration, i = 1,2,3,…, of the *SCD* algorithm applied to a hypothetical ε-similarity cluster detected in a previous iteration has the following structure:

Algorithm 2 (similarity clusters specifying):

```
Step 1: set the initial data values: $\varepsilon^{(i)}$, $z := 0$, $\boldsymbol{S}^{(i)} := \boldsymbol{S}^{(i-1)}$, $\boldsymbol{M}^{(i)} := \boldsymbol{M}^{(i-1)}$ and $\boldsymbol{F}^{(i)} := \boldsymbol{F}^{(i-1)}$;
Step 2: select the next basic window $(\kappa,\lambda) := (\kappa^{(i)}, \lambda^{(i)})$ and the next spectral component $\boldsymbol{\Gamma} := \boldsymbol{\Gamma}^{(i)}$ from the set of non-zero elements of $\boldsymbol{M}^{(i)}$;
Step 3: according to the selection made in Step 2 set $m^{(i)}_{\beta\gamma} := 1$;
Step 4: for all non-zero elements $\sigma^{(i-1)}_{\alpha\beta}$ of adjacency matrix $\boldsymbol{S}^{(i)}$:
    1) select the sub-bitmaps $\boldsymbol{U}^{(\alpha)}_{\kappa\lambda}$ , $\boldsymbol{U}^{(\beta)}_{\kappa\lambda}$ corresponding to the basic windows $(\kappa^{(i)}, \lambda^{(i)})$;
    2) for $\boldsymbol{U}^{(\alpha)}_{\kappa\lambda}$ , $\boldsymbol{U}^{(\beta)}_{\kappa\lambda}$ and the given $\boldsymbol{\Gamma}$ calculate the spectral components $\boldsymbol{W}^{(\alpha)}_{\Gamma;\kappa\lambda}(\alpha)$, $\boldsymbol{W}^{(\beta)}_{\Gamma;\kappa\lambda}$;
    3) calculate the similarity measure values $\sigma^{(i)}_{\alpha\beta} = \sigma[\boldsymbol{W}^{(\alpha)}_{\Gamma;\kappa\lambda}, \boldsymbol{W}^{(\beta)}_{\Gamma;\kappa\lambda}]$ using the formula (2);
Step 4: if $\sigma^{(i)}_{\alpha\beta} = \sigma[\boldsymbol{W}^{(\alpha)}_{\Gamma;\kappa\lambda}, \boldsymbol{W}^{(\beta)}_{\Gamma;\kappa\lambda}] \geq \varepsilon^{(i)}$ then in $\boldsymbol{S}^{(i)}$ set $\sigma^{(i)}_{\alpha\beta} := \sigma[\boldsymbol{W}^{(\alpha)}_{\Gamma;\kappa\lambda}, \boldsymbol{W}^{(\beta)}_{\Gamma;\kappa\lambda}]$, otherwise set $\sigma^{(i)}_{\alpha\beta} := 0$ and $z := 1$;
Step 5: if $z = 0$ then go to Step 8, otherwise:
Step 6: if all elements $\sigma^{(i)}_{\alpha\beta}$ of the $\alpha$ - row (respectively, of the $\beta$-row) of $\boldsymbol{S}^{(i)}$ equal 0 then remove the corresponding row and column and remove $\boldsymbol{u}^{(\alpha)}$ (respectively, $\boldsymbol{u}^{(\beta)}$) from $C^*$ , and from all containing it hyper-edges in $\boldsymbol{F}^{(i)}$ , then go to Step 8, otherwise:
Step 7: in the set $\boldsymbol{F}^{(i)}$ of hyper-edges:
    1) find all hyper-edges $\Phi^{(i)}_m$ containing both nodes $\boldsymbol{u}^{(\alpha)}$ and $\boldsymbol{u}^{(\beta)}$;
```

2) replace each hyper-edge $\Phi^{(i)}_m$ found in 1) by two hyper-edges: $\Phi^{(i)'}_m := \Phi^{(i))}_m \setminus \{\boldsymbol{u}^{(\alpha)}\}$ and $\Phi^{(i)''}_m := \Phi^{(i))}_m \setminus \{\boldsymbol{u}^{(\beta)}\}$;

3) remove $\Phi^{(i)'}_m$ (respectively, $\Phi^{(i)''}_m$) from $\boldsymbol{F}^{(i)}$ if it consists of one node only;

4) assign to the hyper-edges new indexes unifying the enumeration in the set $\boldsymbol{F}^{(i)}$;

Step 8: check whether all elements of $\boldsymbol{M}^{(i)}$ equal 1; if no then go to Step 1, otherwise

Step 9: end of the algorithm ●

Algorithm 2 should be applied to all ε-similarity clusters found in the previous iterations of the ε-*SCD* procedure.

IV. Application to Medical *VDB* Exploration

Large medical *VDB* s may contain thousands of visual documents of various modalities stored in hospitals, specialized medical clinics or in departments and laboratories of medical universities. The *VDB* exploration problems usually concern retrieval of documents based on their formal features or contents. However, in scientific investigations as well as in difficult diagnostic problems *SCD* can also play a significant role.

I. Example

In Fig. 3 a sample of a large collection of typical human brain images obtained by SPECT (Single Photon Emission Tomography) technique is shown. The images are of 124 × 124 pixels size partitioned into basic windows of 16 × 16 pixels size arranged in columns (denoted by numbers from *1* to *8*) and in rows (denoted by symbols from *A* to *H*). The contents of a single basic window can be represented by an 8 × 8 size bitmap or by its 3rd level morphological spectrum consisting of 64 components.

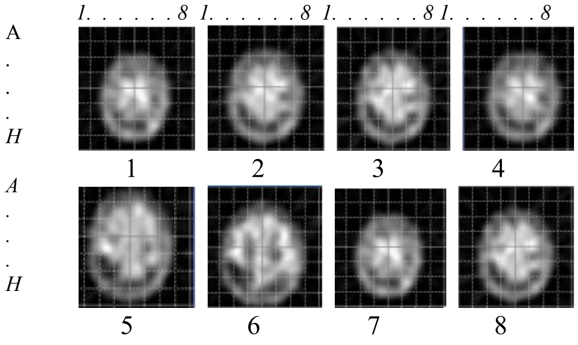

Fig. 3. Cerebral SPECT images stored in a *VDB* .

Fig. 4 shows intensity maps of several spectral components of an image are (symbol *S* stands here for Σ). For comparison of images a region of interest (*ROI*) consisting of *K*, *K* ≤ 64, basic windows not belonging to the background should be considered.

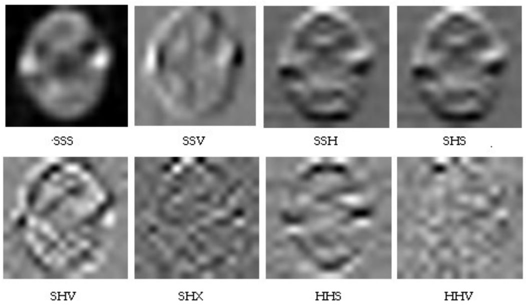

Fig. 4. Selected examples of spectral components' intensities of a given image.

Looking at Fig. 3 it can be remarked that, probably, the subsets of images: {#1, #7} and {#2, #4, #8} form similarity clusters, while images #5 and #6 do not match to any other ones. However, this observation is not based on an objective similarity measure, hence, for a more precise task solution a 0.9-*SCD* problem will be considered.

First, *ROI*s identical in size and form in the set of images should be chosen. Next, a similarity measure based on a distance measure of vectors:

$$d(v', v'') = \sum_{(k)} |v'_k - v''_k| \tag{17}$$

as the simplest one will be chosen.

In the case of using Algorithm 1 each of *K* (*K* ≤ 64) basic windows containing 64 pixels, should be compared with respective basic windows of all other images. Then, all pairs of vectors not satisfying the inequality σ(***v',v''***) ≥ 0.9 should be removed, a similarity graph *G* consisting of 8 nodes and of all edges satisfying the given inequality should be constructed and, according to, the cliques finding problem should be solved.

An alternative, on Algorithm 2 based approach to the problem consists in its iterative solution connected with step-wise elimination of non-perspective pairs of vectors.

At the first iteration the basic windows (*D*,*4*) in the images using the $\Sigma\Sigma\Sigma$ (a sum of 64 pixel values filling a single basic window) spectral component will be compared. As a result, it can be established that the similarity between image #6 and all other ones is below the threshold 0.9 and #6 should be removed from considerations. The problem is thus reduced to this of finding cliques in a complete similarity graph $G^{(1)}$ consisting of 7 nodes: #1, #2, #3, #4, #5, #7 and #8. $G^{(1)}$ is by assumption identical to its unique clique; however, this working hypothesis by testing the similarity measure of selected pairs of basic windows and spectral components should verified.

At the second iteration of the procedure for testing there have been selected (in principle, randomly) the (*F,6*) basic windows and the ΣHV spectral component. As a result two further dissimilarities between the pairs of vectors corresponding to the images (#1,#2) and (#1,#4) have been detected. Therefore, value 0 should be assigned to the elements σ_{12}, σ_{21}, σ_{14} and σ_{21} of the adjacency matrix $\boldsymbol{S}^{(2)}$ of the similarity graph $G^{(2)}$. Moreover, the non-perspective edges

(#1,#2) and (#1,#4) of the graph lead to its replacement by maximal complete sub-graphs non-containing the prohibited pairs of nodes. This means that $\boldsymbol{S}^{(2)}$ should be replaced by two its sub-matrices $\boldsymbol{S}^{(2,1)}$ and $\boldsymbol{S}^{(2,2)}$ based, respectively, on the following rows (columns): {#2, #3, #4, #5, #7, #8} and {#1, #3, #5, #7, #8}. Consequently, the similarity hypergraph $H^{(2)}$ will consist of the set of 7 nodes:

$$C^* = \{\#1, \#2, \#3, \#4, \#5, \#7, \#8\}$$

and of the set of hyper-edges:

$$F = \{(\#2, \#3, \#4, \#5, \#7, \#8),\ (\#1, \#3, \#5, \#7, \#8)\}.$$

Next iterations will be based on increased threshold levels $\varepsilon^{(i)}$; the results will be presented after removing single-node subsets and subsets included by some larger ones. Let us select for being tested the basic windows (F ,3) and spectral components $\Sigma H\Sigma$. Then dissimilarity of the pairs (#2, #7), (#4, #7), (#5, #7) and (#7, #8) can be detected. This preserves the former set C^* of nodes but it leads to the following hypothetical set of hyper-edges:

$$F = \{(\#2, \#3, \#4, \#5, \#8), (\#1, \#3, \#7), (\#1, \#3, \#5, \#8)\}.$$

Next iteration, based on the components (D,6), ΣHX, detects dissimilarity of the pairs (#1, #3), (#1, #4), (#1, #5), (#1, #8), (#2, #5), (#2, #8), (#3, #5) and (#4, #5). This leads to the set of hyper-edges:

$$F = \{(\#3, \#4, \#8), (\#2, \#3, \#4), (\#1, \#7), (\#3, \#7), (\#3, \#5, \#8)\}.$$

II. Comments

In the above-presented example the results have been reached due to comparison of several basic windows and spectral components only instead of comparing full images.

The procedure is stopped when in an iteration all ε-similarity clusters disappear or if the number and form of the clusters does not change for several iterations.

V. Conclusions

The following properties of the above-described method of ε- *SCD* should be remarked:

i. Each iteration of the algorithm consists of a finite number of repetitions of a finite sequence of steps.

ii. The number of iterations is finite.

iii. Each iteration leads to an approximation of the solution satisfying more rigid ε-similarity criteria, hence it preserves or reduces the size of formerly found ε -similarity clusters.

iv. Each next iteration consists in finding cliques in a similarity graph G containing a non-increased (in most cases – reduced) set of nodes and set of edges.

v. In the less favorable case the algorithm leads to an exact ε-*SCD* task solution in a finite number of steps.

Hence, the properties of the Algorithm 2 do not guarantee that the calculation cost of an ε-*SCD* task solution is in each case lower than if Algorithm 1 type one is used. However, it does guarantee that in most cases it is lower due to the reduction of the number of nodes and edges in the similarity graph G leading to an exponential reduction of the cliques finding calculation cost. An exact The above-presented images are of 124×124 pixels size partitioned into basic windows of 16×16 pixels size arranged in columns (denoted by numbers from *1* to *8*) and in rows (denoted by symbols from A to H). The contents of a single basic window can be represented by an 8×8 size bitmap or by its 3rd level morphological spectrum consisting of 64 components. evaluation of the real gain in calculation cost reduction is not possible to be done by analytical methods. It needs a series of experiments to be performed on statistically representative sets of large (i.e. hundreds of nodes containing) similarity graphs.

Acknowledgement

I owe the medical images used in Sec. IV *A* to the courtesy of Prof. MD. L. Krolicki, Head of the Department of Nuclear Medicine of the Warsaw Medical University and of Dr M. Przytulska from Dept. II of the Institute of Biocybernetics and Biomedical Engineering PAS,Warsaw. This work was partially supported by Ministry of Science and Higher Education of Poland, project nr. 4211/BT02/2007/33.

References

[1] Aivazyan S. A., Buchstaber V. M., Yenyukov I. S., Meshalkin L. D., *Applied Statistics. Classification and Reduction of Dimensionality* (in Russian), Moscow: Finansy I Statistika, 1989.

[2] Apers P., Blanken H., Houtsma M., *Multimedia Databases in Perspective.* New York: Springer-Verlag, 1997.

[3] Arisawa H., Catarci T., Eds., *Advances in Visual Information Management. Visual Database Systems* . Kluwer Academic Publishers, Boston Dordrecht London, 2000.

[4] Auguston J. G., Minker J., "An Analysis of Some Graph-Theoretical Cluster Techniques." *J. ACM* 17(4), 1970, pp. 571-588, Errata: *J.ACM* 19(4) , 1972, pp. 244-247.

[5] Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval* . ACM-Press, New York, Addison-Wesley, Harlow Eng. Reading Mass., 1999.

[6] Bakker A. R., "HIS, RIS and PACS." *Comp. Med. Imag. Graph*, No 15, 1991, pp. 157-160.

[7] Berge C.. *Graphs and Hypergraphs.* Amsterdam: North-Holland, 1973.

[8] Duda R.O., Hart P. E., Stork D.G. *Pattern Classification and Scene Analysis* . New York, John Wiley & Sons, 2000.

[9] Kulikowski J. L., "Recognition of Similarities in Image Databases", *17th International CODATA Conference, Book of Abstracts,* Baveno, 2000, pp. 31-32

[10] Kulikowski J. L., "Pattern Recognition Based on Ambiguous Indications of Experts," in *Komputerowe Systemy Rozpoznawania KOSYR'2001,* Kurzyński M., Ed., Wrocław: Wyd. Politechniki Wrocławskiej,, 2001, pp. 15-22.

[11] Kulikowski J. L., Przytulska M., "Partial Similarity Based Retrieval of Images in Distributed Database", in *Advances in Intelligent Web Mastering, AWIC'2007,* Wegrzyn-Wolska K., Szczepaniak P. S., Eds. Berlin, Heidelberg, New York: Springer, 2007, pp. 186-191.

[12] Kulikowski J. L., Przytulska M., Wierzbicka D., "Recognition of Textures Based on Analysis of Multilevel Morphological Spectra", *GESTS Intern. Trans. on Computer Science and Eng,* 38(1), 2007, pp. 99-107.

[13] Kulikowski J. L., Przytulska M., Wierzbicka D., "Morphological Spectra as Tools for Texture Analysis", in *Computer*

Recognition Systems 2 , Kurzynski M., Puchala E., Wozniak M, Zolnierek A., Eds, Berlin, Heidelberg, New York: Springer, 2007, pp. 510-517.

[14] Marek T., *Cluster Analysis in Empirical Investigations. SAHN Methods* (in Polish). PWN, Warsaw, 1989.

[15] Mulligan G. D., “Algorithm for Finding Cliques of a Graph.” *Tech. Report* No 41, Toronto: Univ. of Toronto, 1972.

[16] Noworol C., *Cluster Analysis in Empirical Investigations Fuzzy Hierarchical Models* (in Polish). Warsaw: PWN, 1989.

[17] Pratt W.K., *Digital Image Processing* . New York: John Wiley & Sons, 1978.

[18] Rasiowa H., Sikorski R., *The Mathematics of Metamathematics.* Warsaw: PWN, 1968.

[19] Reinhardt F., Soeder H., „ *dtv-Atlas Mathematik* “. Munich: Deutscher Taschenbuch Verlag, 1977.

[20] 20. Salton G., McGill M. J., *Introduction to Modern Information Retrieval.* New York: McGraw-Hill Book Co., 1983.

[21] Wong S. T. C., Ed. *Medical Image Databases.* Kluwer Academic Publishers, Boston Dordrecht London, 1998.