Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

A Novel Image Semantic Block Clustering Method based on Artificial Visual Cortical Responding Model

Zhiping. XU, *Student Member, IEEE*, Shiyong ZHANG, *Member, IEEE* Shengxiang MA
Department of Computing and Information Technology
Fudan University, Shanghai, China, 200433
dr.bennix@gmail.com

***Abstract*—This paper proposed a novel visual information process model named Artificial Visual Cortical Responding Model (AVCRM) to obtain the invariable time sequence response feature from the sub-image block in the image. By compressing the time sequence feature and selecting the important points in the sequence, we compared the compressed version of sequences with each other to generate the distance matrix. According to distance matrix, we clustered the sub-images into the initially manually assigned concept categories to attain the semantic distribution map of the image. This mechanism was proved to be effective through the experiments and made a good semantic foundation of the future content based image retrieval research work.**

I. Introduction

ALL the ultimate aims of image classification and indexing are to let the computer automatically label a set of images with a set of semantic categories (e.g. coast, mountain and street). The state-of-art technology for Content Based Image Retrieval (CBIR) system is indexing the images in the database by their low level features or high level features, or both. With the progress achieved in the scope of low level features in the CBIR system, many researchers used low level features such as texture, shape, ECM[1], and color histogram and so on as the metric of image distance. The attained low level feature information proved to be useful and powerful in works of QBIC [2], Photobook [3], Virage [4], VisualSEEK [5], Netra [6], SIMPLIcity [7]. One of the most important high level features used in CBIR system was semantic information. However, the gap between the low level features and high level features is still wide and the system can not directly obtain high level feature set from the low level features. Though many sophisticated algorithms were designed to describe color, shape, and texture features, these algorithms could not adequately model image semantics and have many limitations when dealing with broad content image databases [8]. Extensive experiments on CBIR systems show that low-level contents often fail to describe the high-level semantic concepts in user's mind [9]. Therefore, the performance of CBIR is still far from user's expectations. However, the mammalian visual system is considerably more elaborate than simply processing an input image with one set of inner products or apply another set of filters. Mammalian visual system can identify the object in the image quickly and easily, much more accurate than current algorithms. Most neuron computing operations are done before the step of decisions of the content of image. However, neuroscience is not at all close to understanding all of these operations. First of all, optical information from image input is performed through the eyes. Receptors inside the retina at the back of the eye ball are not evenly distributed nor are they all sensitive to the same optical information. Some receptors are more sensitive to motion, color, or intensity. However, the receptors neurons are interconnected. When one receptor receives optical information, it alters the behavior of other surrounding receptors. A mathematical operation is thus performed on the image before it leaves the eye. The receptors neurons also receive feedback information. For example, the human visual system has the phenomena of vision remnant. Furthermore, feedback information also alters the output of the receptors. After the image information leaves the eye, it is received by the visual cortex, where the information is further analyzed by the brain. The investigations of the visual cortex of the cat [10] and the guinea pig [11] have been the foundation of Artificial Visual Cortical Responding Model (AVCRM) used in this paper. The mammal visual system uses neural pulse to transmit the information and retrieval the information and the time sequences feature generated from the AVCRM also can be used as feature to identify the content of the image.

In this paper, we present a novel model named AVCRM based on the principle of mammalian visual system mechanism. The AVCRM can be used to generate semantic blocks from the image, to make these semantic blocks be the foundation of high level features in CBIR. Our approach is to split the image into several blocks, and use the visual cortical response sequence signal as the invariant feature, and then we use the segment time-wrapping method to identify the

Manuscript received October 31, 2006. This work was supported by National Natural Science Foundation of China under Grand 60473118.

Zhiping XU. Department of Computing and Information Technology, Fudan University, Shanghai, China, 200433 (phone: 86-021-65643194; e-mail: dr.bennix@ gmail.com).

Shiyong. ZHANG. Department of Computing and Information Technology, Fudan University, Shanghai, China, 200433 (e-mail: szhang@fudan.edu.cn).

Shengxiang. MA. Department of Computing and Information Technology, Fudan University, Shanghai, China, 200433

1-4244-0705-2/07/$20.00 ©2007 IEEE

differences between the sequences. This paper consists of 6 sections. Section 2 introduces the artificial visual cortical responding model and the time sequence signal the model generated. Section 3 shows the region division for local semantics. Section 4 presents the segment time-wrapping alignment algorithm in compare the sequence signal obtained from the AVCRM. Section 5 gives the image semantic block clustering idea. Section 6 shows the experiments and results. Finally, we conclude the paper in Section 7.

II. Artificial Visual Cortical Responding Model

Eckhorn [12] introduced a model of the cat visual cortex. The neuron inside the cortex contains two input components: the feeding and the linking part. The feeding part receives an external stimulus as well as local stimulus. The external stimulus usually are the optical information from the image; the local stimulus are the inter connection neuron activity information. The linking receives local stimulus. The feeding and the linking are combined in a second-order fashion to create the membrane voltage U_m ,which is then compared to a local threshold Θ . The Eckhorn model is expressed by the following equations,

$$F_k(t)=\sum_{i=1}^{N}\left[w_{ki}^{f}Y_i(t)+S_k(t)+N_k(t)\right]\otimes I(V^{\alpha},\tau^{\alpha},t) \tag{1}$$

$$L_k(t)=\sum_{i=1}^{N}\left[w_{ki}^{l}Y_i(t)+N_k(t)\right]\otimes I(V^{l},\tau^{l},t) \tag{2}$$

$$Y_k(t)=\begin{cases}1 & U_{mk}(t)>\Theta_k(t)\\ 0 & U_{mk}(t)\le\Theta_k(t)\end{cases} \tag{3}$$

where, in general

$$X(t)=Z(t)\otimes I(v,\tau,t) \tag{4}$$

is

$$X[n]=X[n-1]e^{-t/\tau}+VZ[n] \tag{5}$$

Here N is the number of neurons, $w_{\square}$ are the synaptic weights, Y are the binary outputs, and S are the external stimuli. The Eckhorn model shows some information about the principle of the visual cortical system. However, Eckhorn model was a little complicated and its computational cost was great. We simplify Eckhorn model into Artificial Visual Cortical Responding Model (AVCRM). The AVCRM consists of neurons that communicate through dynamic connections. The architecture of the one single neuron in this model is illustrated in Fig.1.

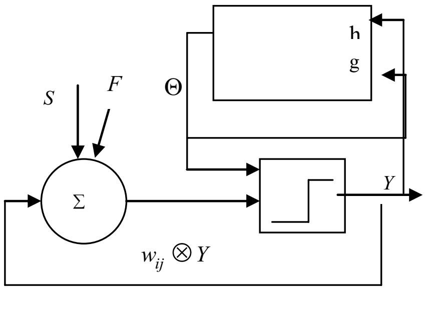


Fig. 1. The Architecture of one single neuron in AVCRM

In the AVCRM, every single neuron has the capability to remember last state $F_{ij}[n-1]$, the state will be attenuation with the time elapsed; the weight of attenuation speed can be affected by the attenuation weight matrix m_{ij} . The AVCRM can be given by,

$$F_{ij}[n]=m_{ij}F_{ij}[n-1]+S_{ij}+w_{ij}\otimes Y \tag{6}$$

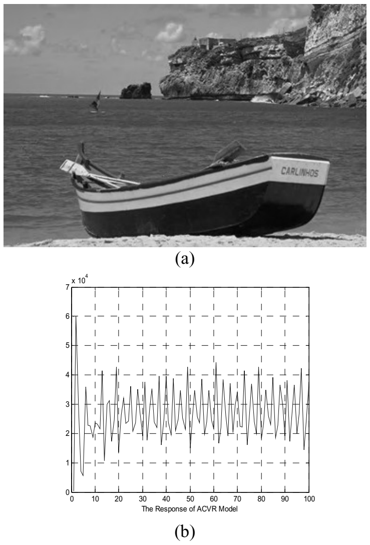
$$Y_{ij}[n]=\begin{cases}1 & F_{ij}[n]>\Theta_{ij}[n]\\ 0 & F_{ij}[n]\le\Theta_{ij}[n]\end{cases} \tag{7}$$

$$\Theta_{ij}[n]=g\Theta_{ij}[n-1]+hY_{ij}[n] \tag{8}$$

in Eq (6)， s_{ij} is the 2D image feed into the neural network, i,j is the coordinate of every pixel, s_{ij} is the intensity value of the coordinate (i , j). m_{ij} is the attenuation weight matrix. During the AVCRM processing image, which each neuron correspond the pixel in the same position in the image, and each neuron has the state F_{ij} , and w_{ij} is the connection weight matrix among the neurons. In Eq.(7) , Θ_{ij} is the dynamic threshold, Y_{ij} is the output of each neuron, the output value is 1 or 0. In Eq.(8), g ， h are both scale coefficient and $g<1$ ， which guarantee the dynamic threshold will finally lower than the state value and fire value, and h is a huge coefficient to raise the threshold rapidly after the neuron fired and not let this neuron fire again in the next iteration. In other words, the local connections are modified according to previous pulsation patterns. When we integrate the neurons response that pulsed during each iteration, it will produce the response of Artificial Visual Cortical Responding Model, namely Artificial Visual Cortical Response (AVCR). AVCR is invariant to the change in scale and orientation. The AVCR can be obtained by,

$$R_p[t]=\sum_{ij}Y_{ij}[t] \tag{9}$$

After the response is generated from the AVCRM, the


(a)

(b)

Fig. 2. The Response Time Sequence of AVCRM (a) The Beach image from the COREL image database (b) The Response Time Sequence of AVCR Model attained from Fig.2(a)

response can be regarded as a time- series signal:

III. Regional Division for Local Semantics

In order to identify local semantics of image, an image is often separated into several sub-regions. The state-of-art methods for image segmentation can be divided into object-based segmentation and block-based segmentation. An object can be a fundamental unit to represent image semantics. The union of several objects usually represents high-level semantics. Although elaborated object segmentation is the best representation of object semantics, it is time-consuming and costly in computational work to achieve reliable segmentation performance. However, the block-based region segmentation is much simpler as well as faster than the object-based one. In the block-based method, a whole image region is tessellated with a fixed or variable block size. So, there is little need for any complex algorithm. But there is often a trade-off in the block size decision in order to capture potential object semantics well, i.e., the smaller block size could lead time-consuming segmentation work as generating many blocks while the larger block size be difficult to detect small object semantics. Initially, we first manually grouped sub-images into several meaningful categories: skin, grass, sea water, tree, sky, soil, concrete building, flower, metal object, food and so on. Each category contains 10 images manually selected; these images are averaged to gain the average image of their represented category as illustrated in Fig 3. Then, the image itself is divided into a number of sub-blocks according to the block size determined. When each block is attained, we apply AVCRM to each image block to generate the time sequence. Using SWT method in Section 4 and Agglomerative Scheme in Section 5, we cluster each time sequence, which represented sub-image block, into different classes.

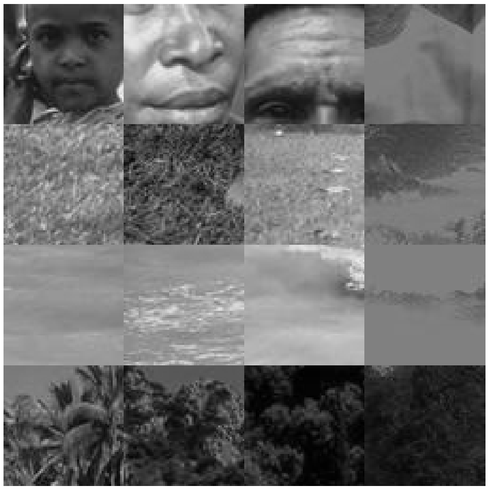


Fig. 3. The manually selected image and the average image in the last column, the image block size is 64X64, first row is skin category, second row is grass category, third row is water category, fourth row is tree category

IV. Segment Time-Wrapping Alignment

In this section, we will use the SWT method to measure the similarity among the time series generated by the Section 3 mentioned AVCR model. Due to the length of the time sequence could be very long, we use [13] mentioned important point selection algorithm to compress time sequence data. In the scope of time series matching, there are various techniques such as the Euclidean distance and Dynamic Time Wrapping (DTW) method to solve distance measurement problem. However, in many cases, using the traditional Euclidean distance as the metric of sequence difference is very brittle, because the Euclidean space is too sensitive to time warping and shifting. DTW method is relative robust distance measure for time series. DTW can allow similar shape to match even if they are out of phase in the time axis. Nevertheless, there are some special problems the DTW method can't do well in some sequence comparison, for example the triangular-wave sequence. In the shape form view of the time series generated by the AVCR model, most part of the sequence looks like triangular wave, so we use the STW method instead of DTW in comparing these sequences. STW use a non-uniform time scaling transformation to handle the matching problem, that is, all the elements of the sequence don't have to be stretched by the same factor and the time warping method should be based on segments instead of points. So we can linear interpolate some points into the segment to stretching the segment; that is, given a segment P_iP_{i+1} , if it is stretched n times, the stretched segment $P_iP_{i+1}(s_0, s_1, s_2, ..., s_n)$ should be given by:

$$S_i = P_i + \frac{i}{n}(P_{i+1} - P_i), 0 \le i \le n \qquad (10)$$

where P_i is the starting point of the sequence segment, and P_{i+1} is the end point of the sequence. Eq.(10) remains the shape of the original sequence as much as possible. The shape of the original sequence is preserved by the Eq.(10), the two compressed version of sequence may have the time shifting and time wrapping situation. To solve this kind of situation, we using STW to find the optimal alignment between the segments of two sequence and calculate the STW distance. First, we define the distance for two segments $P(a_i, a_{i+1})$ and $Q(b_j, b_{j+1})$ given by:

$$d(P,Q) = (a_i - b_j)^2 + (a_{i+1} - b_{j+1})^2 \qquad (11)$$

Also, the distance between a stretched long segment and a series of un-stretched segment is defined ,as follow :given one segment $P(a_i, a_{i+1})$,that will be stretched N times to match a series of segments $Q(b_j, b_{j+1})$, $Q(b_{j+1}, b_{j+2})$,…., $Q(b_{j+N-1}, b_{j+N})$,the distance is given by:

$$d(P, Q(j \sim j+N)) = \sum_{k=0}^{N}(a_i + \frac{k}{N}(a_{i+1} - a_i) - b_{j+k})^2 \quad (12)$$

The STW algorithm will build a $n \times m$ matrix to facilitate finding the two sequences $S(s_{1,}s_2....,s_n)$ and $Q(q_0,q_1....,q_m)$ distance.

TABLE I

STW ALGORITHM FOR TWO SEQUENCES S AND Q

Algorithm: SWT
Input :The sequences S and Q;
Output: The optimal matrix D
The three-dimensional matrix PATH which record the optimal warping path.

```
1.   Initial the second row and column of the matrix D ;
2.   Initial the PATH corresponding to the matrix D;
3.   min:=0;value:=0;N:=|S|;M:=|Q|
4.   for i:=3 to N do
5.   begin
6.     for j:=3 to M do
7.     begin
8.       min: =D_{i-1,j-1}+d( $s_i$ , $q_i$ );
9.       if i >3 then
10.      begin
11.        for k:=3 to j-1 do
12.        begin
13.          value:= D_{i-1,k-1}+d( $s_i$ ,q(k~ j));
14.          if value < min then
15.          begin
16.            min:=value;
17.            Record the warping path;
18.          end
19.        end
20.      end
21.      if j>3 then
22.        do similar thing in from step 11 to 19;
23.      end
24.    end
25.  end
```

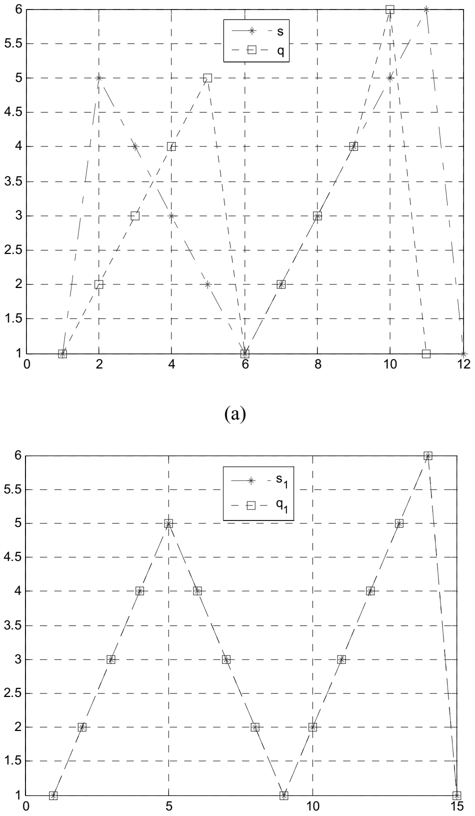
As illustrated in Table 1, the algorithm will produce optimal cumulative-distance matrix D, where the (i^{th}, j^{th}) element of the matrix represent the minimum one among the cumulative distances produced by the all the possible warping path. In order to prove the accuracy, given two sequences, which are of the shape of triangular wave, S=(1,5,4,3,2,1,2,3,4,5,6,1) and Q=(1,2,3,4,5,1,2,3,4,6,1). These two sequence in illustrated in Fig.3(a); we use the DTW to test the similarity and the distance is 11. At the same time, we use the STW method to test the similarity and get a perfect result illustrated in Fig 3(b). The distance between the S and Q is 0.

V. IMAGE SEMANTIC BLOCK CLUSTERING

After each sub-image block's AVCR sequence is compared, we got the distance matrix according to Eq. (11) and (12), given by,

$$DM_{ij} = D(Seq_i, Seq_j) \qquad (13)$$

where Seq_i and Seq_j are the shape preserved important point sequences corresponding to sub-image block i and j. After the distance matrix is obtained, we use Agglomerative Scheme (AS) to cluster these sub-image blocks into different classes.



(a)

(b)

Fig. 4. The verification experiment of two sequence alignment and matching (a) Two sequence S and Q (b) STW alignment and matching result.

TABLE II

AGGLOMERATIVE SCHEME OF CLUSTERING

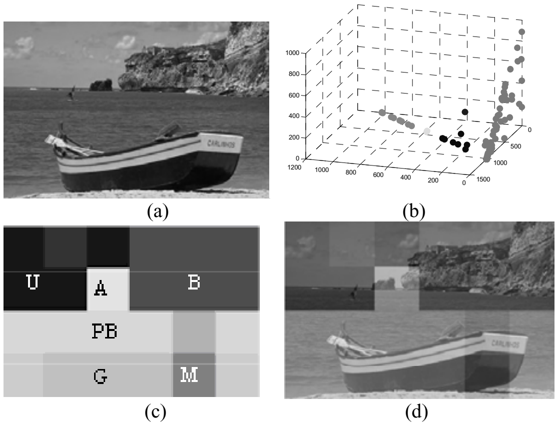
INPUT : N, the Number of clustering number

```
1. Initialization:
   1.1. Choose $\mathbb{R}_0$={C_i={Seq_i},i=1,...,N} as the initial clustering;
   1.2 t:=0;
2. Repeat
   2.1 t:=t+1;
   2.2 Among all possible pairs of clusters (C_r,C_s) in $\mathbb{R}_{t-1}$ , say (C_i,C_j), such that
       $g(C_i, C_j) = DM_{ij}$
   2.3 Define $C_q = C_i \cup C_j$ and produce the new clustering
       $\mathbb{R}_t = (\mathbb{R}_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$
   Until the cluster number reached
```

VI. EXPERIMENTS

Here we evaluate the performance of our method. We first divide the input image into several sub-image blocks according the block-size 64X64 pixels size. Then apply AVCR model to each sub-image to generate the visual response sequence. To shorten the length of generated sequences and to preserve the outline shape of the sequences, the sequences go through the important point selection filter to gain the compressed version of sequence. After the compressed versions of sequences are obtained, we use segment time wrapping algorithm to the compare the distance between two pairs of sequences to generate the distance matrix. According to the distance matrix, we use agglomerative scheme to cluster these sequences, each sequence represents one sub-image block in the original input image. After sub-image blocks are clustered into different categories, each category contains several images, we average these images and compare these averaged version of images to the manually assigned the category average image to find out the best counter-part and assign the meaning to this block. We choose images from the COREL image CD, which contained several categories of images, like African scene, flowers, beach scene view and so on, to prove the proposed mechanism. As illustrated in Fig 5 and 6, the proposed mechanism works perfectly in obtain the semantic meaning from the image, and human could be easily draw text information from the semantic distribution map, like in Fig 5 and 6. Fig 5 conveys the meaning of sky in upper part, with mountains on the right, the mountain in connected with gorge, the gorge and mountains is above the sea water, the shore with a metal object on it is near the sea water. Fig 6 conveys the meaning of “Trees in upper left and right part, with soils on the top, an object with skin and a small part of skin in middle of the image, and an unknown object in the image”.

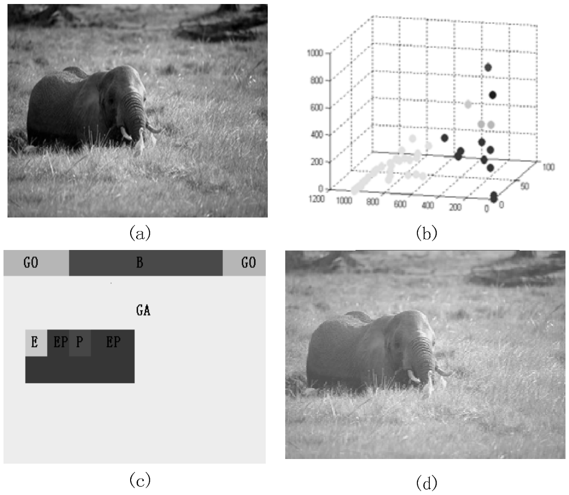
When given a set of query images with certain semantic meanings and retrieve, we evaluated the accuracy of the return result sets in the image database that share similar semantic meaning as illustrated in Table III. The image database contains 40000 images including different categories. From the Table III, we could see that the average accuracy of the proposed framework is 78.56%.The semantics like grass, sea water, fresh water, and sky show good performance due to the sub-image shows good textural properties.



(a) (b)

(c) (d)

Fig. 5. The Result of Artificial Visual Cortical Image Semantic Block Clustering (a) the original input image in COREL image CD; (b) The AVCR model generated time sequence clustering result, the image is divided into five semantic categories; (c) The semantic meaning mapping to the division sub-image blocks, U-labeled part is for the sky meaning, A-labeled part is for the gorge meaning, B-labeled part is for the mountain meaning, PB-labeled part is for the sea water meaning, G-labeled part is for shore meaning, and M-labeled part is for the metal skin meaning; (d) the composite version image blending with semantic meaning distribution map and original image.



(a) (b)

(c) (d)

Fig. 6. The Result of Artificial Visual Cortical Image Semantic Block Clustering (a) the original input image in COREL image CD; (b) The AVCR model generated time sequence clustering result, the image is divided into six semantic categories; (c) The semantic meaning mapping to the division sub-image blocks, GO-labeled part is for the tree meaning, B-labeled part is for the soil meaning, GA-labeled part is for the grass meaning, E-labeled part is for the unknown meaning, EP-labeled part is for skin meaning, and P-labeled part is for the darker skin meaning; (d) the composite version image blending with semantic meaning distribution map and original image.

This proposed mechanism make a good foundation of the future work in the content based image retrieval research based on the semantic meaning understanding.

VII. Conclusion

In this paper, we proposed a model named Artificial Visual Cortical Responding Model, which can obtain the semantic content of the image. Also, we illustrated the correctness of AVCRM method in the image semantic meaning feature retrieval by experiments. The proposed ACVRM can obtain the meaning contained in the images in the semantic level. The mechanism would be very helpful in the filling of the gap between the low level feature and high level feature, and make a good foundation of the future CBIR research work.

TABLE III
SEMANTIC MEANING QUERY RESULTS

Semantic	Accuracy (%)	Semantic	Accuracy (%)
skin	75.78	Darker skin	56.39
grass	87.35	soil	79.62
sea water	91.67	concrete building	83.67
fresh water	89.34	flower	79.94
tree/leaf	74.34	metal object	63.21
sky	93.45	food	67.89

ACKNOWLEDGEMENT

The authors would like to acknowledge the comments of the three unknown reviewers whose insightful comments helped to improve this paper.

REFERENCES

[1] Z.P. XU, Y.P ZHONG，S.Y ZHANG, Fast Shape Index Framework based on Principle Component Analysis using Edge Co-occurrence Matrix, KES 2006, Part III, LNAI 4253, pp. 390－397.

[2] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic and W. Equitz, Efficient and effective querying by image content, *J. Intell. Inf. Syst*. 3 (1994) (3–4), pp. 231–262.

[3] A. Pentland, R.W. Picard and S. Scaroff, Photobook: content-based manipulation for image databases, *Int. J. Comput. Vision* 18 (1996) (3), pp. 233–254.

[4] A. Gupta and R. Jain, Visual information retrieval, *Commun. ACM* 40 (1997) (5), pp. 70–79.

[5] J.R. Smith, S.F. Chang, VisualSeek: a fully automatic content-based query system, *Proceedings of the Fourth ACM International Conference on Multimedia*, 1996, pp. 87–98.

[6] W.Y. Ma, B. Manjunath, Netra: a toolbox for navigating large image databases, *Proceedings of the IEEE International Conference on Image Processing*, 1997, pp. 568–571.

[7] J.Z. Wang, J. Li and G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, *IEEE Trans. Pattern Anal. Mach. Intell*. 23 (2001) (9), pp. 947–963..

[8] A. Mojsilovic, B. Rogowitz, Capturing image semantics with low-level descriptors, *Proceedings of the ICIP*, September 2001, pp. 18–21.

[9] X.S. Zhou, T.S. Huang, CBIR: from low-level features to high-level semantics, Proceedings of the SPIE, *Image and Video Communication and Processing*, San Jose, CA, vol. 3974, January 2000, pp. 426–431.

[10] R. Eckhorn, H. J. Reitboeck, M. Arndt, P. Dicke: Feature linking via synchronization among distributed assemblies: Simulations of results from Cat Visual Cortex. *Neural Comp*. 2 (1990), pp. 293–307.

[11] I.A. Rybak, N.A. Shevtsova, V.A. Sandler: The model of a Neural Network visual processor. *Neurocomputing* 4 (1992), pp. 93-102.

[12] R. Eckhorn, H. J. Reitboeck, M. Arndt, P .Dicke. Feature linking via synchronization among distributed assemblies: Simulations of results from Cat Visual Cortex. *Neural Comp*. 2, (1990), pp. 293–307.

[13] E.Fink, K.B.Pratt, Indexing of compressed time series, Data Mining In Times Series Databases, *Machine Perception Artificial Intelligence*, (2004) (57), pp 43-56.