Content Based Image Retrieval Using Adaptive Inverse Pyramid Representation

Mariofanna Milanova[1], Roumen Kountchev[2], Stuart Rubin[3], Vladimir Todorov[4], and Roumiana Kountcheva[4]

[1] Computer Science Departmentt, UALR, USA
mgmilanova@ualr.edu
[2] Department of Radio Communications, Technical University of Sofia, Bulgaria
rkountch@tu-sofia.bg
[3] SSC San Diego, California, USA
stuart.rubin@navy.mil
[4] T&K Engineering, Bulgaria
todorov_vl@yahoo.com, kountcheva_r@yahoo.com

Abstract. This paper presents a new approach for content-based image retrieval using cognitive representation with pyramidal decomposition. This approach corresponds to the hypothesis of the human way for object recognition based on consecutive approximations with increased resolution for the selected regions of interest. The method is based on object model creation with Inverse Difference Pyramid controlled by neural network. The method's basic advantages are the high flexibility and the ability to create general models for various views and scaling with relatively low computational complexity. The method is suitable for great number of applications – medicine, digital libraries, electronic galleries, geographic information systems, documents archiving, digital communication systems, etc.

Keywords: content- based image retrieval, multi-layer representation, IDP decomposition.

1 Introduction

Research in the area of Content Based Image Retrieval (CBIR) has come a long way since it was first introduced by T. Kato in 1992. CBIR has been a focus of intensive research with more than 300 scientific publications per year [1]. Most of the widely known methods for image and video retrieval are based on the use of *quantitative* (low –level) features and *qualitative* (high level) features. Feature design problems include finding how many meaningful visual features do exist and on which spatio-temporal regions of media objects should the selected features be applied on. The classical answer to these problems is the Multi-Resolution Analysis (MRA). The basic MRA hypothesis is that using interactively computed 2D wavelet coefficient matrices as features is sufficient for content retrieval.

Generally, the visual retrieval process aims at finding media objects that are similar to given examples. "Similarity" is a weakly defined term, and therefore difficult to implement in computer systems. Two requirements (the similarity matching and the

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 304–314, 2009.

user feedback) have to be satisfied by visual information retrieval (VIR) systems. The similarity matching has to be performed on media objects represented by feature vectors and the user feedback has to be integrated in the retrieval process. The retrieval therefore is a necessary interactive communication process between user and computer. One major advance of VIR in recent years was achieved by using relevant feedback. Unfortunately, even the most sophisticated algorithms are still not able to satisfy the users' need for similarity-based retrieval sufficiently. The question "How domain knowledge is represented?" is still open. Most of the VIR systems derived from text retrieval concepts, but it is not necessary to use the same or similar mining techniques in VIR systems. The human visual system has the ability to correctly interpret most images even using low resolution images. Search and visual information processing, as seen by psychologists, is observed in the following three basic hypotheses:

- The first is that image resolution exponentially decreases from the fovea to the retina periphery. Unlike digital cameras and their uniform sampling acquisition system, humans do not see the world uniformly, because the retina receptors are not equally distributed on its surface, but are concentrated in the fovea [2]. This hypothesis can be represented computationally with different resolutions. The visual attention points may be considered as the most highlighted areas of the Visual Attention model i.e., these points are the most salient regions in the image. When going further from these points of attention, the resolution of the other areas dramatically decreases. There are existing models where perception of visual environment is based on the fact that the observer first fixates the higher attention level areas and only then he looks at the other areas. Different authors work with various filters and kernel size [3].
- Another interesting question is the role of the visual contextual information in the attention model creation and VIR. Most computational attention models ignore the contextual information provided by the correlation between objects and the scene. Schyns and Oliva [4] showed that a coarse representation of the scene initiates semantic recognition before the identification of objects is performed. Many studies support the idea that scene semantics can be available early in the chain of information processing and suggest that scene recognition may not require object recognition as a first step [5], because humans can recognize the scene even using low-special frequency image.
- Covert attention allows us to select visual information at a cued location, without eye movements. It is proved that covert attention not only improves discriminability, but also accelerates the rate of information processing [6]. Attention affects both spatial and temporal aspects of visual processing. By enhancing the signal, attention improves discriminability and enables us to extract relevant information in a noisy environment by accelerating information processing.

In this paper is presented one new approach for content-based image retrieval based on these main hypotheses. The proposed solution is based on image representation with adaptive inverse difference pyramid (IDP) decomposition controlled by neural network. Such image representation corresponds to the human way of objects perception and is suitable for the creation of flexible objects' models, which to be used for query procedures in image databases in accordance with predefined decision rules. Significant element of the new representation is the use of a feedback, which provides

iterative change of the cognitive models' parameters in accordance with the data mining results obtained.

2 Basic Principles of the IDP Decomposition

The algorithm for recursive IDP coding of halftone digital images comprises the following steps:

Step 1. The matrix [X] of the original image is divided into sub-images of size $2^n \times 2^n$ and each is then processed with a two-dimensional (2D) orthogonal transform (OT) using only a limited number of spectrum coefficients (usually, the low-frequency ones). The values of these transform coefficients constitute the first pyramid layer.

Step 2. Using the values of the transform coefficients, every sub-image is restored by inverse orthogonal transform and then subtracted pixel by pixel from the original one. The difference sub-image with elements $e_p(i,k)$ in the IDP layer p is defined as:

$$e_p(i,k)=\begin{cases} x(i,k)-\tilde{x}_0(i,k) & \text{for} \quad p=0; \\ e_{p-1}(i,k)-\tilde{e}_{p-1}(i,k) & \text{for} \ p=1,2,..,P, \end{cases} \tag{1}$$

where x(i,k) is the pixel (i,k) in a sub-image of size $2^n \times 2^n$ of the input image [X] (Fig.1a); $\tilde{x}_0(i,k)$ and $\tilde{e}_{p-1}(i,k)$ are correspondingly the pixels of the recovered input and the difference sub-images in the IDP layer p.

Step 3. The difference sub-image is divided into 4 sub-images of size $2^{n-1} \times 2^{n-1}$. Each sub-image is then processed with 2D OT again and the values of the used transform coefficients build the second pyramid layer. The image is then restored and the second difference image is calculated. The process continues in a similar way with the next pyramid layers. The block diagram for pyramid of 3-layers is shown in Fig. 1.

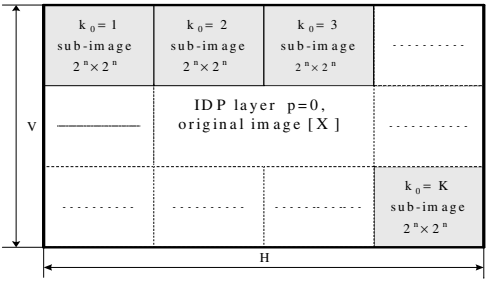
The applications usually do not require all the pyramid layers to be calculated, because the needed image quality is usually obtained in the lower layers. Such pyramid is called "truncated".

The approximation models of the input or difference image in the layer p are represented by the relations:

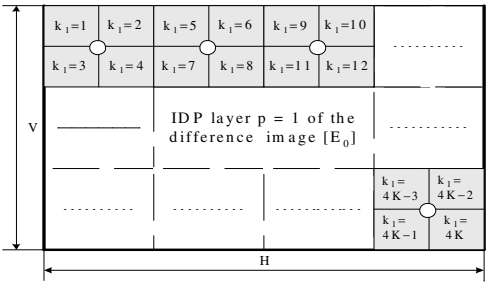
$$y_p(u,v)=T[x(i,k)/e_{p-1}(i,k)] \ \text{and} \ \tilde{x}(i,k)/\tilde{e}_{p-1}(i,k)=IT[\tilde{y}_p(u,v)] \tag{2}$$

where T[•] is the operator for the truncated direct two-dimensional orthogonal transform applied on the input block of size $2^n \times 2^n$, or on the difference sub-image of size $2^{n-p} \times 2^{n-p}$ from pyramid layers p=1,2,..,P (Fig. 2b); IT[•] is the operator for the inverse OT of the spectrum coefficients $\tilde{y}_p(u,v)$ from the layer p of the truncated transform $2^{n-p} \times 2^{n-p}$, obtained in result of the transformation of each ¼ part of the difference sub-image, $e_{p-1}(i,k)$.

Specific for the IDP is that the OT coefficients, used for every pyramid layer, can be different. The coefficients from all pyramid layers are sorted in accordance with their frequency, and scanned sequentially. The obtained one-dimensional massif for



a. The original image of size H×V, divided into K sub-images of size $2^n \times 2^n$ (layer p=0).



a. Each sub-image of the difference image $[Y_0]$ for layer p=0 is divided into 4 sub-images of size $(2^{n-1} \times 2^{n-1})$ in the pyramid layer p = 1

Fig. 1. The IDP layers p=0,1 for an image of H×V pixels

the s-th frequency band of the two-dimensional OT of the input or of the difference image for the IDP layer p is represented by the relation:

$$\tilde{y}_p(s) = \tilde{y}_p[u=\varphi(s), v=\psi(s)] \tag{3}$$

where $u=\varphi(s)$ and $v=\psi(s)$ are functions, which define the transformation for the two-dimensional massif of coefficients in the s-th frequency band for the layer p.

The block diagram of the IDP decomposition is shown in Fig. 2.

The image decoding is performed in reverse order. The processing of color images depends on the color component representation – individual pyramid is build for each component [7].

The object representation based on the IDP decomposition offers the solution for problems, concerning image rotation and translation: for RST-invariant transforms (Fourier-Mellin) [8] the values of the decomposition coefficients are invariant as well. The object representation is done using a single original image, or more than one images (different view, lighting, color, scaling, etc). In the second case the initial object representation (in the lower decomposition layers) is fuzzy and the more exact representation is defined for the higher decomposition layers. The coefficients for every sub-image in the consecutive pyramid layers build the vectors of the object's features, which are then used for the model evaluation. For this the selected coefficients are processed with inverse transform, and the quality of the restored image (i.e. the model error), is estimated. In case that this error is too big, the neural network,

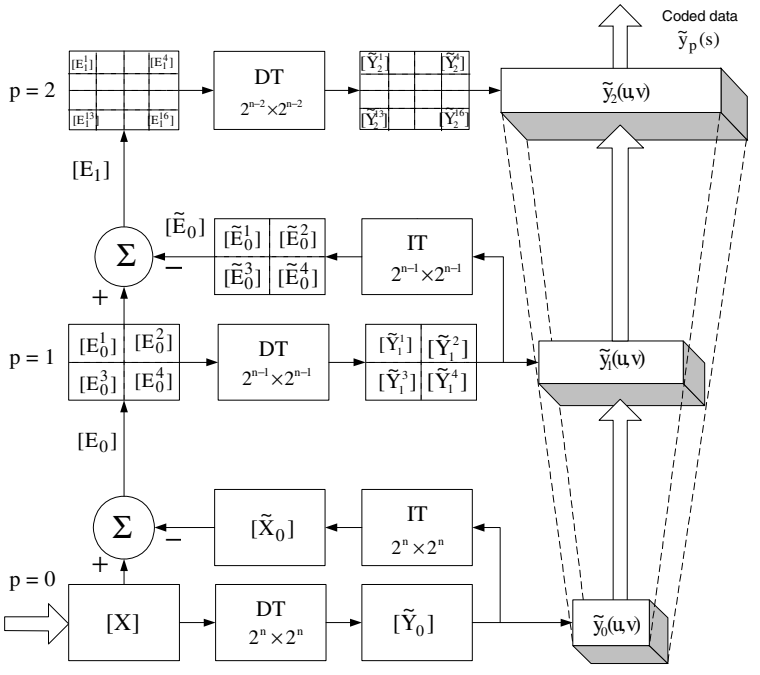


Fig. 2. Block diagram of 3-layer IDP decomposition

which controls the features' selection for the next decomposition layer performs the corresponding model tuning. At the last decomposition layer is obtained the final description of the object model.

The abbreviations used in the block diagram are: [X] – the matrix of the processed image (sub-image); DT/IT – direct/inverse orthogonal transform; $[E_p]$ - the difference (error) matrix in layer p; $\tilde{y}_p(u,v)$ – the retained set of coefficients in layer p.

3 Multi-layer Image Retrieval

The image retrieval is performed by comparing the object model of 2 or more layers with the content of the images in the database. The multi-layer search is based on the evaluation of the multi-layer distance, which is defined as a sequence of differences between approximations, obtained in result of the layered IDP decomposition for any couple of compared objects: objects with maximum similar content have smallest multi-layer distance. The initial presumption is that the queried object image is smaller than the database image. In general, the number of search layers corresponds to the number of decomposition layers used for the object model creation. The queried object model is used for the creation of the corresponding pyramid decomposition for a sub-image (window) of same size in every image from the database. The initial position of the search window in one of the database image corners (for example, the lower left corner). For this position is evaluated the distance between the object model vector for the layer 0 and the corresponding vector obtained

for the search window content. After translation by one step in the selected direction (horizontal or vertical), the distance between the compared vectors is evaluated again, etc. When the scanning in the database image is finished for the decomposition layer 0, the search continues in similar way with the next database image for the same decomposition layer until all images are processed. When the analysis for the layer 0 is completed, the database images, containing an object which is close enough to the queried object model for this layer, are separated in a special group. In case that there are no images, which answer the requirements, this group is empty. The described operations are performed in similar way for the decomposition layer 1 of the separated images only. In the consecutive layers the number of images, which answer the requirement to be close enough to the queried one, becomes smaller. For the defined empty groups additional search should be performed, for which through feedback is introduced the next model (different view angle, lighting, etc., if there is such) and the described operations are preformed again.

The search for the closest object in the image database $\{X^t\}$ for t = 1, 2,.., N is represented by the relations below:

For the IDP layer p = 0 the distance between the object model request $[X]$ and the object representation $[X^t]$, from the database, is:

$$D_0\{[\tilde{X}_0],[\tilde{X}_0^t]\}=\sum_{k_0=1}^{K}\sum_{s=1}^{S_0}\left|[\tilde{y}_{0,k_o}(s)-\tilde{y}_{0,k_0}^t(s)]\right| \tag{4}$$

where S_0 is the number of the retained spectrum coefficients in the layer p=0.

For IDP layers p = 1, 2,.., P the distance between the object model for the corresponding layer and the sub-image from an image in the database, is:

$$D_p\{[\tilde{E}_{p-1}],[\tilde{E}_{p-1}^t]\}=\sum_{k_p=1}^{4^pK}\sum_{s=1}^{S_p}\left|[\tilde{y}_{p,k_p}(s)-\tilde{y}_{p,k_p}^t(s)]\right| \tag{5}$$

where S_0 is the number of the retained spectrum coefficients in the layer p.

The distance between object models $[X]$ and $[X^t]$ is calculated for p = 0, 1,…,P:

$$D\{[X],[X^t]\}=D_0\{[\tilde{X}_0][\tilde{X}_0^t]\}+\sum_{p=1}^{P}D_p\{[\tilde{E}_{p-1}],[\tilde{E}_{p-1}^t]\} \tag{6}$$

The *multi-layer search* in the image database comprises the following operations:

- All distances for layer p = 0 of the IDP decompositions between the object request and the images in the database are calculated and is found the smallest one. The image from the database, which has a part, containing an object with smallest distance, is named t_r and is represented by the relation:

$$t=t_r \text{ if } D_0\{[\tilde{X}_0],[\tilde{X}_0^{t_r}]=\min< d_{\min}(0) \text{ for } p=0 \tag{7}$$

where $d_{\min}(0)$ is a threshold for the IDP layer p=0. In case that there is only one such image, the search is successfully finished.

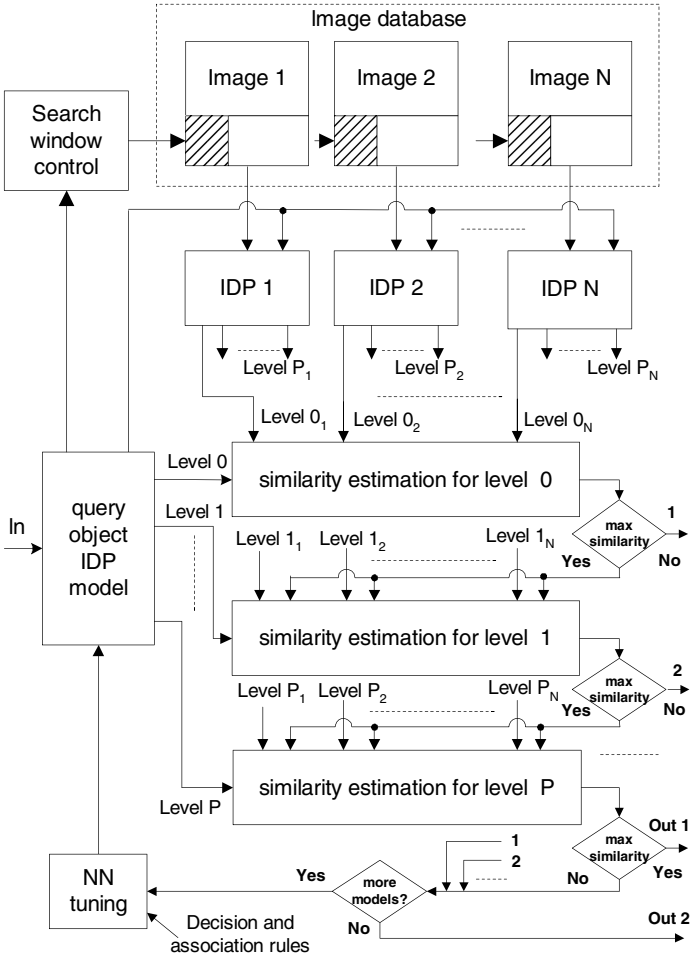


Fig. 3. Block diagram of the method for cognitive multi-layer image retrieval

If the images with smallest distance for p=0 are more than one, they are separated in a group, and the search continues in similar way for the next IDP layers, for the so defined group only:

$$t=t_r \text{ if } D_p\{[\tilde{E}_{p-1}],[\tilde{E}^{t_r}_{p-1}]\}=\min< d_{\min}(p) \text{ for } p=1,2,..,P, \quad (8)$$

where: $d_{\min}(p)$ are the thresholds (the values of the thresholds define the required accuracy of the performed search process for the corresponding IDP layers); t_r is the image from the database, whose distance is the smallest for the decomposition layer p. Maximum similarity is obtained for the case, when the function, which represents the multi-layer distance (6), has a minimum.

The block diagram of the image retrieval method based on the IDP decomposition is shown in Fig. 3. In the block diagram are shown two possible outputs: for detected

closest image from the database (Out 1) and for missing similar image (Out 2). Significant element is the use of a feedback, through the block named NN tuning, which provides iterative change of the cognitive models' parameters in accordance with the data mining results obtained. This block modifies the object model in two main cases: 1. the needed similarity is not achieved; 2. in accordance with predefined decision and association rules.

4 Experimental Results

The experiments were performed with more than 50 test objects. In the experiments was evaluated the efficiency of the models and the ability for their recognition. For the object models creation was used the software implementation of the IDP decomposition of 7 layers. The basic models were created with truncated pyramid of

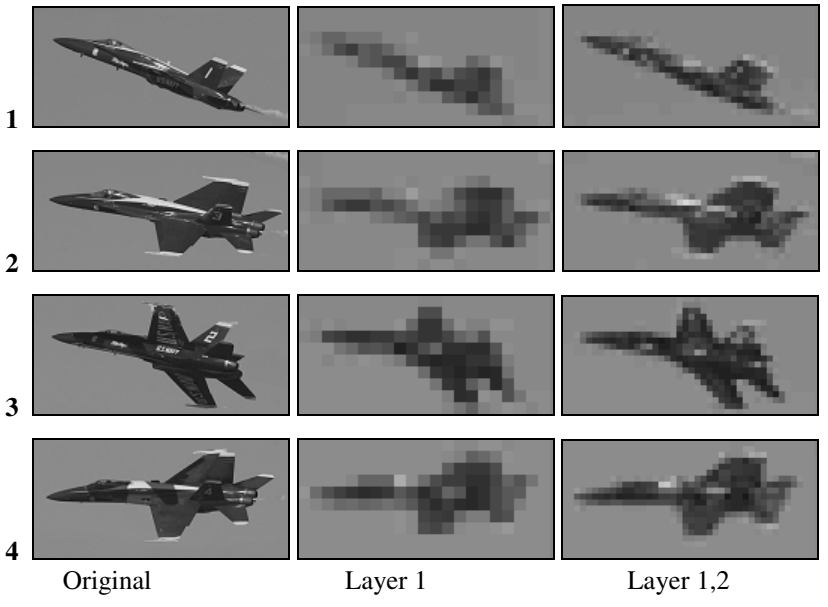


Fig. 4. Object representation from multi-view images of same plane

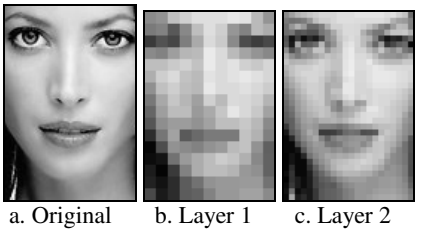


Fig. 5. Test image "Chris"

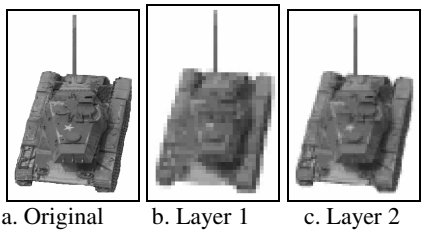
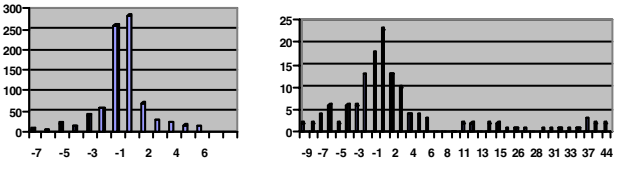


Fig. 6. Test image "Tank"

Table 1.

Image	Layer 1 size [B]	Bit-rate [bpp]	CR	PSNR [dB]	Layer 2 size [B]	Bit-rate [bpp]	CR	PSNR [dB]
Plane1	67	0,039	203	22,10	216	0,127	63,0	24,21
Layer1+2					229	0,134	59,4	24,21
Plane2	84	0,049	162	20,20	261	0,150	52,0	22,37
Layer 1+2					270	0,159	50,4	22,37
Plane3	89	0,053	150	21,29	264	0,155	51,52	23,25
Layer 1+2					262	0,153	51,91	23,25
Plane4	81	0,047	168	20,94	264	0,155	51,52	22,85
Layer 1+2					256	0,150	53,13	22,85
Chris	122	0,086	92	19,80	442	0,310	25,48	22,24
Layer 1+2					393	0,280	28,66	22,24
Tank	700	0,054	147	22,11	2000	0,149	53,43	24,60
Layer1+2					1880	0,140	56,78	24,60



a. “Plane 1” model b. “Chris” model

Fig. 7. Graphic representation of the object models for “Plane 1” and “Chris” - Layer 1

2 layers (initial layer with sub-block of size 8×8 pixels). The retained coefficients were 4 for the lower layer and 3 - for the next one. The 2D transform was Walsh-Hadamard. All test images “Plane” are grayscale, of size 170×80 pixels. The image “Chris” is of size 88×128 pixels and the image “Tank” – 272×400. Some of the obtained results are shown in Table 1. The column “Layer 1” gives the information about the size of the data for the lower layer of the object model and “Layer 2” - the size of the compressed data for the next layer. For each layer are given the bit-rate and the compression ratio (CR). In rows “Layer 1+2” is given the size of the compressed 2-layer object model data. The experiments prove the efficiency of the presented method for object model creation (the bit-rate of the object models for layers 1 and 2 is very low). The PSNR, i.e. the similarity between the object and the model for these two layers is low, but enough to recognize plane, face or tank. For more difficult tasks (which model is the plane, etc.) we need more complicated representation, using Layer 3 or even higher. Different views are necessary as well.

In Fig. 4 are shown the original test images Plane 1 – Plane 4 (multi-view) with their models for Layers 1 and 2.

In Figs. 5 and 6 are shown the test images “Chris” and “Tank” with their corresponding models for decomposition layers 1 and 2.

The graphics in Fig. 7 represent the histograms of the values of the coefficients, which build the object models (4 coefficients, Layer 1) for a plane (Plane 1) and

human face (Chris). The graphic representations for all planes are quite similar and the graphics for the remaining three test objects are not given here. The difference with the object representation of "Chris" is quite clear, both in range and allocation.

5 Conclusion

The new method for *objects search* was simulated with MATLAB using two or more pyramid decomposition levels. The 2D transform was Fourier-Mellin. The matching obtained for images in several image classes (forest, city, desert, etc.) was more than 80%.

The new method for content based image retrieval ensures faster search in large databases, because the layered processing permits significant part of the images to be excluded from further search at the end of the Layer 1 analysis.

The use of the IDP decomposition permits the creation of efficient multi-view models. Another important advantage is the multi-scale representation, based on the relations between transform coefficients in adjacent layers, which offer significant reduction of the transform coefficients, needed for the object model creation [9]. The introduction of a flexible feedback in the process of object model creation and search makes this approach close to the human way of thinking.

The new method permits to develop flexible models for some basic image kinds, for example: texts/graphics, cartoon images, medical images, natural grayscale or color images, etc., which to be later defined more accurately, in accordance with the object peculiarities. This approach, which is based on preliminary knowledge, will facilitate the object representation and search.

Acknowledgment. This paper is supported by the National Fund for Scientific Research of the Bulgarian Ministry of Education and Science (Contract No VU-I 305), NSF grant 0619069 Development of IAEL and by funding from the U.S. Defense Threat Reduction Agency (DTRA-BA08MSB008).

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40(2), article 5, 60 (2008)
2. Hubel, D.: Eye, Brain and Vision Scientific American Library, vol. 22. W. Freeman, New York (1989)
3. Mancas, M., Gosselin, B., Macq, B.: Perceptual Image Representation. EURaSIP Journal on Image and Visual Processing, article ID 98181 (2007)
4. Schyns, P., Oliva, A.: From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition. Psychol. Sci. 5, 195–200 (1994)
5. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Computer Vision 42, 145–175 (2001)
6. Carrasco, M., McElree, B.: Covert attention accelerates the rate of visual information processing. PNAS 98, 5363–5367 (2001)

7. Kountchev, R., Milanova, M., Ford, C., Kountcheva, R.: Multi-layer Image Transmission with Inverse Pyramidal Decomposition. In: Halgamuge, S., Wang, L. (eds.) Computational Intelligence for Modelling and Predictions, ch.13, vol. 2, pp. 179–196. Springer, Heidelberg (2005)
8. Derrode, S., Ghorbel, F.: Robust and efficient Fourier-Mellin transform approximations for gray-level image reconstruction and complete invariant description. Computer vision and image understanding 83(1), 57–78 (2001)
9. Kountchev, R., Kountcheva, R.: Image Representation with Reduced Spectrum Pyramid. In: Tsihrintzis, G., Virvou, M., Howlett, R., Jain, L. (eds.) New Directions in Intelligent Interactive Multimedia, pp. 275–284. Springer, Heidelberg (2008)