THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 23, Number 3/2022, pp. 289–298

CONTENT-BASED IMAGE RETRIEVAL FRAMEWORK USING MULTI-SCALE DEEP-LEARNING-BASED FEATURE EXTRACTION AND GRAPH-BASED CLUSTERING

Tudor BARBU

Institute of Computer Science of the Romanian Academy – Iași Branch, Iași, Romania
E-mail: tudor.barbu@iit.academiaromana-is.ro

Abstract. A novel content-based color image retrieval framework is proposed in this article. It uses a querying by example technique and a cluster-based image database indexing approach. A multi-scale deep learning-based image analysis is performed for this retrieval task. The scale-space representation is constructed by applying a nonlinear second-order reaction-diffusion based model at various scales, then a high-level CNN-based feature extraction is performed at each scale. The feature vectors obtained for the considered database are then classified unsupervisedly by applying an automatic graph theory based clustering technique. A relevance feedback-based CBIR process that is facilitated considerably by the obtained image clusters is then described here. Retrieval tests are also discussed.

Key words: content-based image retrieval (CBIR), multi-scale image analysis, anisotropic diffusion model, convolutional neural network (CNN), query by example, graph clustering, database index.

1. INTRODUCTION

Image retrieval systems are designed for browsing, searching and retrieving images from voluminous databases. Depending on the formats of their querying techniques, the image retrieval algorithms are divided into two major categories: metadata-based approaches and content-based methods. The systems in the first category add keywords, captioning, titles or various descriptions to the images so that retrieval could be performed over these annotations. The content-based image retrieval (CBIR) techniques aim to avoid the use of these textual descriptions and instead retrieve the images on the basis of their content features [1]. Their querying process is performed by using *query by example*, which means providing a user-supplied query image, or using some user-specified image content features. The image features involved in the retrieval process could be *global* features, such as those based on color, texture, shape and spatial information, or *local* features. The global feature-based CBIR systems developed in the last years include those using invariant color moments [2], color histograms [1], color correlograms [1], color co-occurrence matrices [3], Wavelet transforms [4], gray-level co-occurrence matrices (GLCM) [5], Gabor filters [6], Fourier descriptors [7], affine moment invariants [8], local structure descriptors [1] and color, texture and shape feature fusion [9]. The local feature-based retrieval systems include those using the scale-invariant feature transform (SIFT) [10], binary robust invariant scalable key points (BRISK) [1], speeded-up robust features (SURF) [11], local binary patterns (LBP) [12], co-occurrence histogram of oriented gradients (CoHOG) [13] and the Harris corner detector [1]. Also, some famous general-purpose content-based image retrieval systems developed in the last decades include QBIC, Photobook, Blobworld, Virage, VisualSEEK, WebSEEK and PicHunter.

Although many effective content-based retrieval systems have been created, the *semantic gap* represents still a challenge for the CBIR systems. It reflects the difference between the relatively limited descriptive power of the low-level content features and the richness of the user semantics. Various retrieval techniques have been developed to narrow this semantic gap. They can be categorized into two main classes depending on the degree of the user interactivity: relevance feedback and image database pre-processing using unsupervised classification. A relevance feedback-based system allows a user to provide information which the user considers to be relevant to the query and based on the user feedbacks, the model of similarity measure is dynamically updated [1]. The retrieval methods from the second category classify the images of the database into semantically meaningful groups using low level features. They cluster the image features based on similarity, using various unsupervised machine learning algorithms, such as *K*-means, Gaussian Mixture Models (GMM), Mean Shift, Normalized Laplacian Spectral algorithms and Bayesian fuzzy clustering [14]. Some clustering-based image retrieval systems perform the unsupervised learning process on the images that are retrieved in response to a query [15]. Other CBIR systems apply clustering techniques in order to generate cluster-based database indexing structures that would facilitate the retrieval task [9, 12, 16].

The recent trends for content-based image retrieval are focused on the deep learning networks, which represent more effective solutions for this task, since they are able to extract powerful high-level image content features. Thus, many CBIR systems based on convolutional neural networks (CNN) have been developed in the last decade [1, 17]. Here we consider a query by example based content-based retrieval system for RGB color images that also applies two deep neural networks in the feature extraction stage.

The motivation of the proposed system is to continue at a higher level and to improve substantially our research results in the content-based image indexing and retrieval domain. The most CBIR techniques developed by us in the last 15 years were based on relevance-feedback schemes [18–23] and some of them used Spatial Access Method (SAM) – based indexing structures [19, 20]. They used low-level content features, such as various color - based features [18, 20, 22, 23], or medium-level image characteristics based on histograms of oriented gradients (HOG) [19] or 2D Gabor filters [21]. The novel CBIR framework introduced here performs a much improved high-level content feature extraction based on an effective combination of nonlinear partial differential equations, multi-scale image analysis and deep learning, which is described in the next chapter. The scale-space representation is created by applying the nonlinear second-order anisotropic diffusion model introduced in the first sub-chapter at several scales. A convolutional neural network - based image feature extraction is then performed at each scale and the feature vectors obtained at multiple scales are next concatenated into a final content descriptor. The proposed CBIR system uses an improved feature clustering solution, which outperforms our past clustering (unsupervised classification) methods [24, 25]. The cluster-based indexing structure that is created for its image database by applying the automatic graph-based clustering technique introduced in the third chapter is more performant than the tree-based indexing schemes of our past retrieval models. This framework uses also an efficient relevance-feedback mechanism that is also described in the third chapter. The experiments illustrating the effectiveness of the described framework are discussed in the fourth chapter and the conclusions are drawn in the last one.

2. HIGH-LEVEL MULTI-SCALE IMAGE FEATURE EXTRACTION TECHNIQUE

The feature extraction component of the proposed content-based image retrieval system uses a multi-scale analysis based on nonlinear anisotropic diffusion. The multi-scale image analysis provides much better results than mono-scale analysis and the anisotropic diffusion produces more performant scale spaces than the 2D Gaussian filtering [26]. The second-order partial differential equation (PDE) - based model that is applied to construct the scale-space representation is introduced in the next sub-chapter. Then, a deep neural network-based content feature extraction is performed at multiple scales, as described in the sub-chapter 2.2.

2.1. Nonlinear second-order reaction diffusion-based scale-space representation

We have performed a vast amount of research in the PDE-based image processing and analysis field, a lot of PDE and variational models being introduced by us in the last years [27, 28]. Here we consider a nonlinear second-order anisotropic diffusion-based filter, given by the next parabolic PDE with boundary conditions:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \lambda\eta\left(\left\|\nabla u_{\sigma}\right\|\right)\nabla\cdot\left(\psi\left(\left\|\nabla u\right\|\right)\nabla u\right) + \beta\left(u - u_0\right) = 0, \ \ \forall\left(x,y\right)\in\Omega \\ u(0,x,y) = u_0(x,y), \quad \forall\left(x,y\right)\in\Omega \\ u(t,x,y) = 0, \quad \forall\left(x,y\right)\in\partial\Omega; \end{cases} \tag{1}$$

where $\alpha \in [1,2), \beta \in (0,0.4)$, the image domain $\Omega \subseteq R^2$, the observed image $u_0 \in L^2(\Omega)$ and $u_\sigma = u * G_\sigma$, where the 2D Gaussian filter kernel $G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$. We introduce the next positive and monotonic decreasing diffusivity function that is properly chosen for an efficient denoising [28]:

$$\psi : [0,\infty) \to [0,\infty) : \ \psi(s) = \gamma \left(\frac{\alpha}{\left| \xi s^2 + \delta \right|} \right)^{1/3} \tag{2}$$

where $\xi \in (0,1], \gamma \geq 1, \alpha, \delta \geq 4$. The other positive function used by the filtering model has the form:

$$\eta : [0,\infty) \to [0,\infty) : \ \eta(s) = \sqrt[\kappa+1]{\frac{\nu s^\tau + \rho}{\zeta}}, \tag{3}$$

where $\nu, \rho \in [1,4)$, $\zeta \in (0,0.6)$ and $\kappa \in (0,1)$. This reaction-diffusion based filter removes the additive Gaussian noise, while preserving the essential image details and avoiding unintended effects like blurring or staircase. Thus, it represents a better tool for constructing an effective scale-space than the 2D Gaussian operator that generates the blurring effect and may deteriorate the boundaries. Also, the second-order nonlinear PDE model (1) is non-variational, since it cannot be obtained from the minimization of an energy cost functional, and well-posed, since it exists a unique weak (variational) solution for it. That solution is determined using a numerical approximation scheme that is consistent to (1) and converges to it. The approximation algorithm is constructed applying the finite difference method [29]. First, the spatial and time coordinates are quantized as $x = ih,\ y = jh,\ i \in \{1,...,I\},\ j \in \{1,...,J\}$ and $t = n\Delta t,\ n \in \{0,...,N\}$, where h represents the space size, Δt is the time step and $[Ih \times Jh]$ is the support image's size. Since anisotropic diffusion equation (1) leads to

$$\frac{\partial u}{\partial t} + \beta(u - u_0) = \lambda \eta\left(\|\nabla u_\sigma\|\right)\left(\frac{\partial}{\partial x}\left(\psi(|\nabla u|)u_x\right) + \frac{\partial}{\partial y}\left(\psi(|\nabla u|)u_y\right) \right) \tag{4}$$

we approximate its left component using the central finite differences [29], as:

$$\frac{u_{i,j}^{n+\Delta t} - u_{i,j}^n}{\Delta t} + \beta\left(u_{i,j}^n - u_{i,j}^0\right) = u_{i,j}^{n+\Delta t} \frac{1}{\Delta t} + u_{i,j}^n \left(\beta - \frac{1}{\Delta t}\right) - u_{i,j}^0 \beta. \tag{5}$$

The right term in equation (4) is then approximated. The component $\lambda\eta\left(\|\nabla u_\sigma\|\right)$ is discretized as $\lambda\eta_{i,j} = \lambda\eta\left(\nabla \left\| (G_\sigma * u)_{i,j} \right\|\right)$, where $\|\nabla u_{i,j}\| = \sqrt{\left(\frac{u_{i+h,j} - u_{i-h,j}}{2h}\right)^2 + \left(\frac{u_{i,j+h} - u_{i,j-h}}{2h}\right)^2}$. Then, $\frac{\partial}{\partial x}\left(\psi(|\nabla u|)u_x\right)$ is discretized spatially as $\psi_{i+\frac{h}{2},j}\left(u_{i+h,j} - u_{i,j}\right) - \psi_{i-\frac{h}{2},j}\left(u_{i,j} - u_{i-h,j}\right)$ and $\frac{\partial}{\partial y}\left(\psi(|\nabla u|)u_y\right)$ is discretized spatially as $\psi_{i,j+\frac{h}{2}}\left(u_{i,j+h} - u_{i,j}\right) - \psi_{i,j-\frac{h}{2}}\left(u_{i,j} - u_{i,j-h}\right)$, where $\psi_{i\pm\frac{h}{2},j} = \frac{\psi_{i\pm h,j} + \psi_{i,j}}{2}, \psi_{i,j\pm\frac{h}{2}} = \frac{\psi_{i,j\pm h} + \psi_{i,j}}{2}$.

We then consider $h = \Delta t = 1$. Thus, by using all these approximations, the following explicit iterative numerical approximation algorithm is obtained:

$$u_{i,j}^{n+1} = u_{i,j}^n(1-\beta) + u_{i,j}^0\beta + \lambda\eta_{i,j}\left(\psi_{i+\frac{1}{2},j}\left(u_{i+1,j}^n - u_{i,j}^n\right) - \psi_{i-\frac{h}{2},j}\left(u_{i,j}^n - u_{i-1,j}^n\right) + \psi_{i,j+\frac{1}{2}}\left(u_{i,j+1}^n - u_{i,j}^n\right) - \psi_{i,j-\frac{1}{2}}\left(u_{i,j}^n - u_{i,j-1}^n\right) \right) \tag{6}$$

The iterative approximation scheme (6) is stable and consistent to the anisotropic diffusion model (1) that is numerically solved by it and converges fast to its weak solution representing the filtering result. It is succesfully used to create a proper scale-space representation for the multi-scale analysis. Thus, the nonlinear diffusion-based scale-space is created by applying the PDE-based filter on a digital image until various moments of time. The diffusion model (1) works for gray-level images only, while we aim to construct a scale-space representation for RGB color images. An obvious solution to this problem would be to apply (6) on each of the 3 channels of the analyzed image, but since its R, G and B channels may have high levels of correlation, we have considered a better approach. Thus, the RGB image *Im* is converted to a decorrelated color space like CIELAB, then its luminance channel *L* (Im) is processed using (6), $(L(\text{Im}))^n$ representing its filtering after *n* iterations. The filtered *L*a*b* image, denoted as $\left[(L(\text{Im}))^n,\ a(\text{Im}),\ b(\text{Im})\right]$, is then converted back to RGB form. So, one obtains the next multi-scale representation with *K* scales for *Im*:

$$S(\text{Im}) = \left\{\text{Im},\ RGB\left(\left[(L(\text{Im}))^{\tau}, a(\text{Im}), b(\text{Im})\right]\right), \ldots, RGB\left(\left[\left(L(\text{Im})\right)^{\tau(K-1)}, a(\text{Im}), b(\text{Im})\right]\right)\right\}, \tag{7}$$

where $K \geq 5$, the time step $\tau \in [3,10]$ and $RGB(\ )$ returns the RGB conversion of its argument.

2.2. Deep learning-based content feature extraction

The scale space *S* given by (7) is then used in the multi-scale image content feature extraction process. A high-level feature extraction is performed on the analyzed image at each scale $k \in \{0,\ldots,K-1\}$, which has the form $S(\text{Im})\{k\} = RGB\left(\left[(L(\text{Im}))^{\tau k}, a(\text{Im}), b(\text{Im})\right]\right)$. A deep learning-based method is used for this purpose.

The Convolutional Neural Networks (CNNs, ConvNets) represent deep learning networks that provide very effective image content feature extraction results that help reduce the CBIR semantic gap [17]. A CNN could learn successfully high-level feature representations for a large variety of images if it has been trained on voluminous databases storing various types of images. The obtained DL-based image features outperform the content characteristics produced by other descriptors. Here we apply two pre-trained convolutional neural networks to compute the high-level characteristics of each analyzed image. The ConvNets that have been chosen for this purpose are ResNet (Residual Network) 50 and Inception-V3. We consider them much better high-level feature extraction solutions than other pre-trained CNN-based models, such as VGGNet-16, VGGNet-19, GoogleNet and AlexNet, since those networks have much fewer convolutional layers [30].

So, ResNet-50 is a deep residual network characterized by a 50-layer architecture that has been trained on 1.2 million training images of the voluminous ImageNet database which contains 1000 object categories [31]. Since it has learned rich features for a wide range of images, this deep neural network can be used to characterize and classify successfully new images [32]. Inception-V3 is a 48 layers deep convolutional neural network that started as a GoogleNet module [33]. It has been trained on more than a million images from ImageNet and represents a very useful tool in image analysis and object detection. The layers of the 2 CNNs produce activations to the input images, but not all layers have the same feature extraction power. So, the first layers of these deep networks can detect only the low-level characteristics that are then processed by the deeper convolutional layers that combine them to achieve higher level image features [32, 33]. Thus, our technique extracts the required content features from the deep layers of ResNet-50 and Inception-V3. We have considered as optimal content-based feature extraction solutions the Fully Connected Layer with 1000 neurons (FC 1000), which is located right before final classification layer in the ResNet-50 architecture, and the Predictions Layer of Inception-V3 that is also a fully connected layer located before classification layer.

Each $S(\text{Im})\{k\}$ image is pre-processed according to the specifications of the ResNet-50 input layer, first. It is resized at the $[224 \times 224 \times 3]$ required format and a zero center normalization is then performed to it as

$$S(\text{Im})\{k\} := \left(S(\text{Im})\{k\} - \mu\left(S(\text{Im})\{k\}\right)\right) / \sigma\left(S(\text{Im})\{k\}\right),\ \ \forall k \in \{0,\ldots,K-1\},\ \mu - \text{average value}. \tag{8}$$

Also, the size of the mini-batch parameter of this ResNet model is set at 32. Next, these pre-processed RGB images are fed into the ResNet-50 network. So, for each *k* from 0 to *K* – 1, the FC 1000 layer of this CNN produces an activation on $S(\text{Im})\{k\}$ given by (8) that computes high-level characteristics of the image in the form of a content feature vector with 1000 coefficients, $V_1\left(S(\text{Im})\{k\}\right)$. The Inception network is then applied similarly on the initial $S(\text{Im})\{k\}$. The image is resized at the $[299 \times 299 \times 3]$ format required by its input layer, then the normalization (8) is performed to it. It is fed into Inception-V3 model whose fully connected *prediction* layer produces an activation generating 1000 coefficient high-level feature vector, $V_2\left(S(\mathrm{Im})\{k\}\right)$. All 2D feature vectors $V\left(S(\mathrm{Im})\{k\}\right) := \left[V_1\left(S(\mathrm{Im})\{k\}\right);\ V_2\left(S(\mathrm{Im})\{k\}\right)\right]$ of the scale space are concatenated into

$$V(\mathrm{Im}) := \left[V\left(S(\mathrm{Im})\{0\}\right) \ldots V\left(S(\mathrm{Im})\{K-1\}\right)\right] \tag{9}$$

This final $\left[2 \times 1000K\right]$ feature vector $V(Im)$ represents a powerful content descriptor of the color image Im. It outperforms clearly the descriptors based on local features like SIFT, SURF, HOG or LBP, which are used by other retrieval systems [1]. The optimal high-level characterizations provided by the CNN – based feature vectors given by (9) lead to proper discriminations between their images. So, images with similar contents have very close feature vectors, while images with different contents have very distant feature vectors. Thus, this multi-scale DL-based feature extraction leads to the optimal image clustering process described in the next chapter.

3. AUTOMATIC CLUSTERING-BASED IMAGE INDEXING AND RETRIEVAL

A typical query by example CBIR technique searches in a large database for images having similar contents with the given query. But assessing the feature similarity between the query image and all the stored images, each time such a query is provided, could become a complex and time-consuming task in case of a voluminous database. Also, the retrieved images could be affected by the already mentioned semantic gap, which means that images with a high feature similarity to the query example could have different semantics than the query. So, the proposed content-based image retrieval framework applies both unsupervised learning and relevance-feedback mechanisms on the considered database in order to solve these issues. An automatic clustering-based image indexing approach that facilitates the retrieval process is proposed first. Then, a relevance-feedback retrieval scheme searching for results only in the most relevant image cluster is provided.

We create a content-based indexing structure for the system database, by applying an automatic graph-based image feature vector clustering technique. If the RGB image database is denoted as $Db = \left\{\mathrm{Im}_1, \ldots, \mathrm{Im}_M\right\}$, then a weighted similarity graph is created for it. Thus, we consider the fully connected undirected graph $G = (V, E)$, whose vertex set V contains M vertices (nodes) corresponding to the images Im_i and $E \subseteq V \times V$ represents the set of linking edges that are weighted by the following weighting function:

$$w : E \to R:\ w_{ij} = w(i, j) = 1000\, e^{-d\left(V(\mathrm{Im}_i), V(\mathrm{Im}_j)\right)},\ \forall i, j \in \{1, \ldots, M\},\ i \neq j, \tag{10}$$

where d is the Euclidian distance, but other metrics may be applied here too. A normalized -cut (*NCut*) based graph clustering algorithm is proposed for G. It represents a much better clustering solution than other graph partitioning methods [14], since it achieves more balanced clusters [34]. The *NCut* measure is described as

$$NCut(A, B) = \frac{\mathrm{Cut}(A, B)}{\mathrm{Cut}(A, V)} + \frac{\mathrm{Cut}(A, B)}{\mathrm{Cut}(B, V)},\quad \mathrm{Cut}(A, B) = \sum_{i \in A, j \in B} w_{ij},\ A \cap B = \varnothing,\ A \cup B = V \tag{11}$$

and a bipartition of G that minimizes it is determined by solving the next generalized eigenvalue system [34]:

$$\left(D - W\right)x = \lambda Dx,\ \ D = \mathrm{diag}\left[\sum\nolimits_i w_{ij}\right], W\left(i, j\right) = w_{ij}, W\left(i, i\right) = 0. \tag{12}$$

The generalized eigenvector with the second lowest eigenvalue is used to bipartition this graph, by considering its median value as a splitting point, and the optimal vertex sets A and B are determined. This *NCut* - based partitioning algorithm is then applied recursively on the two obtained subgraphs. At each step one has to decide if such a subgraph must be further partitioned or not. Various stopping conditions could be applied to this recursive graph clustering process, such as reaching the required number of clusters or *NCut* (A, B) exceeding a certain threshold. Since our graph clustering algorithm is fully automatic, the optimal number of clusters is unknown, so not applicable as a termination condition. So, we consider that if either of partition is too small, the recursion must be stopped. Also, the current vertex set V cannot be partitioned and represents a cluster if all of its weights are close enough to each other. These conditions are formalized as:

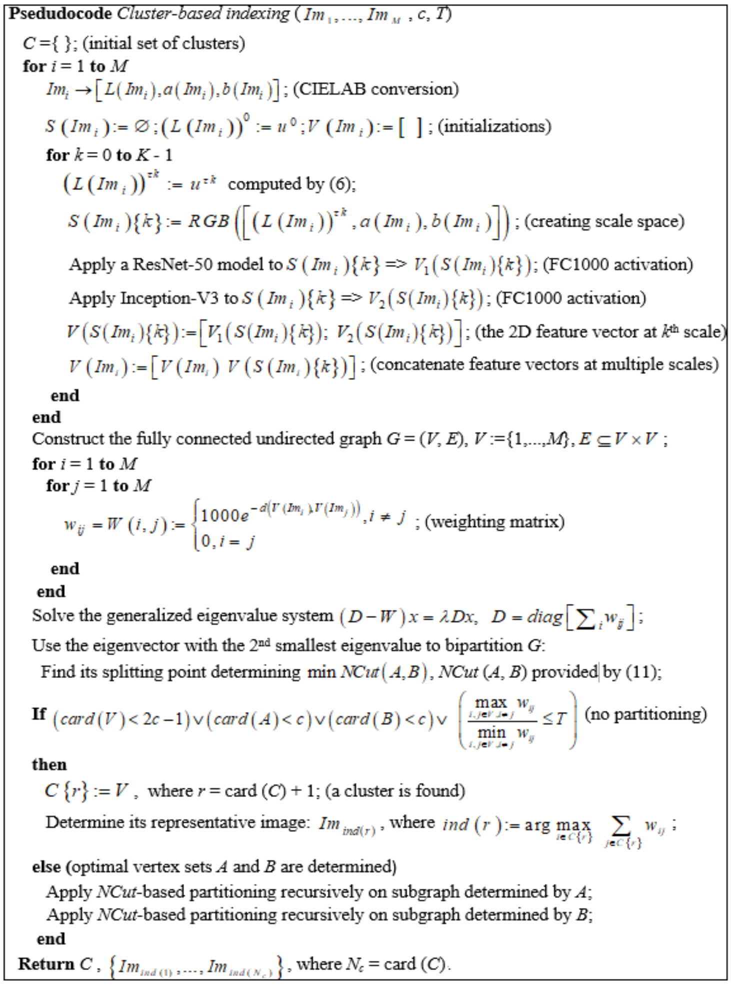
$$\left(\mathrm{card}(V)<2c-1\right)\vee\left(\mathrm{card}(A)<c\right)\vee\left(\mathrm{card}(B)<c\right)\vee\left(\frac{\max\limits_{i,j\in V,i\neq j} w_{ij}}{\min\limits_{i,j\in V,i\neq j} w_{ij}}\leq T\right)\Rightarrow C\{r\}:=V\,, \tag{13}$$

where the node sets A and B are obtained from (10)–(12), c represents the minimum number of images from a cluster, $T\in(1,2]$ is a properly chosen threshold parameter and C is the set of node clusters obtained until that moment (by (12) V becomes the r^{th} detected cluster, $r > 0$). The final form of C produced by this normalized-cut based recursive algorithm represents the graph clustering result. Since this graph partitioning scheme represents a hierarchical divisive clustering procedure creating a tree, C contains the leafs of the tree.

Each node cluster $C\{r\}$ corresponds to an image cluster, since it contains the indices of the database stored images. Thus, the database clustering result $C=\{C\{1\}, ..., C\{N_c\}\}$, where N_c is the detected number of clusters, works as a cluster-based indexing structure for *Db*. So, one has to select a representative image for each index. We consider the image coresponding to the maximum sum of within-cluster weights as representative for the $C\{r\}$ cluster, therefore that image is $\mathrm{Im}_{ind(r)}\in Db$, where

$$\mathrm{ind}(r):=\arg\max_{i\in C\{r\}}\sum_{j\in C\{r\}} w_{ij} \tag{14}$$

Therefore, one obtains a set of N_c *key images* that are representative for the entire database *Db*: $\left\{\mathrm{Im}_{ind(1)},...,\mathrm{Im}_{ind(N_c)}\right\}$. These key images are then used successfully in the content-based retrieval process. A pseudocode of this multi-scale feature extraction-based database indexing algorithm is described in Fig. 1.



```
Psedudocode Cluster-based indexing ($Im_1,...,Im_M$ , $c$, $T$)
  $C=\{\ \}$; (initial set of clusters)
  for $i = 1$ to $M$
     $Im_i\rightarrow\left[L(Im_i),a(Im_i),b(Im_i)\right]$; (CIELAB conversion)
     $S(Im_i):=\varnothing;(L(Im_i))^0:=u^0;V(Im_i):=[\ ]$; (initializations)
     for $k = 0$ to $K$ - 1
        $(L(Im_i))^{\tau k}:=u^{\tau k}$ computed by (6);
        $S(Im_i)\{k\}:=RGB\left(\left[(L(Im_i))^{\tau k},a(Im_i),b(Im_i)\right]\right)$; (creating scale space)
        Apply a ResNet-50 model to $S(Im_i)\{k\}\Rightarrow V_1(S(Im_i)\{k\})$; (FC1000 activation)
        Apply Inception-V3 to $S(Im_i)\{k\}\Rightarrow V_2(S(Im_i)\{k\})$; (FC1000 activation)
        $V(S(Im_i)\{k\}):=\left[V_1(S(Im_i)\{k\});\ V_2(S(Im_i)\{k\})\right]$; (the 2D feature vector at $k^{th}$ scale)
        $V(Im_i):=\left[V(Im_i)\ V(S(Im_i)\{k\})\right]$; (concatenate feature vectors at multiple scales)
     end
   end
   Construct the fully connected undirected graph $G=(V,E)$, $V:=\{1,...,M\}$, $E\subseteq V\times V$;
   for $i = 1$ to $M$
     for $j = 1$ to $M$
        $w_{ij}=W(i,j):=\begin{cases}1000e^{-d(V(Im_i),V(Im_j))}, i\neq j\\ 0, i=j\end{cases}$ ; (weighting matrix)
     end
    end
   Solve the generalized eigenvalue system $(D-W)x=\lambda Dx$,  $D=diag\left[\sum_i w_{ij}\right]$;
   Use the eigenvector with the 2nd smallest eigenvalue to bipartition $G$:
    Find its splitting point determining $\min NCut(A,B)$, $NCut(A,B)$ provided by (11);
   If $(card(V)<2c-1)\vee(card(A)<c)\vee(card(B)<c)\vee\left(\frac{\max_{i,j\in V,i\neq j} w_{ij}}{\min_{i,j\in V,i\neq j} w_{ij}}\leq T\right)$ (no partitioning)
   then
     $C\{r\}:=V$ ,  where $r$ = card ($C$) + 1; (a cluster is found)
     Determine its representative image: $Im_{ind(r)}$, where $ind(r):=\arg\max_{i\in C\{r\}}\sum_{j\in C\{r\}} w_{ij}$ ;
   else (optimal vertex sets $A$ and $B$ are determined)
     Apply NCut-based partitioning recursively on subgraph determined by $A$;
     Apply NCut-based partitioning recursively on subgraph determined by $B$;
    end
 Return $C$ , $\{Im_{ind(1)},...,Im_{ind(N_c)}\}$, where $N_c$ = card ($C$).
```

Fig. 1 – The pseudocode of the proposed cluster-based image indexing technique.

So, the query image Q received as input by our query by example-based CBIR system is not compared to all the images stored by *Db*, but only to its representative images. The one that is the most similar to the provided query must have the closest feature vector to the query feature vector, so it is $Im_{ind(j)}$, where

$$j = \arg\min_{i \in \{1,\dots,N_c\}} d\left(V(Q), V\left(\mathrm{Im}_{ind(i)}\right)\right). \tag{15}$$

The retrieval system should provide this most relevant representative, $Im_{ind(j)}$, to the user if it is not too dissimilar to the query, so a threshold-based verification solution is proposed here. One considers that there is no relevant image in *Db* to Q if the feature vector distance in (15) exceeds a threshold representing the average of the distances between the feature vectors of the representative images. Thus, we have:

$$d\left(V(Q), V(\mathrm{Im}_{ind(j)})\right) > 2\frac{\sum\limits_{i,k \in \{1,\dots,N_c\}} d\left(V(\mathrm{Im}_{ind(i)}), V(\mathrm{Im}_{ind(k)})\right)}{N_c\left(N_c - 1\right)} \Rightarrow \text{no images retrieved} \tag{16}$$

The most relevant cluster $C\{j\}$, represented by $\mathrm{Im}_{ind(j)}$, could be sent to the user as the image retrieval result, but it may be difficult to display in case of a large size, and also, if a semantic gap still exists, some images from it may not represent exactly what the user is looking for. So, a relevance-feedback mechanism is then used to search for better retrieval results in the database index given by the cluster $C\{j\}$. The most relevant R images to Q are retrieved from $C\{j\}$ and ranked according to their relevance given by the feature vector distance value $d\left(V(Q), V(\mathrm{Im}_i)\right)$, $i \in C\{j\}$. All or some of them may be extracted as retrieval output and *the best* of them (usually the 1st-ranked one) may be selected as a new query Q and the retrieval process continues the same way. The architecture of the proposed CBIR framework is described in Fig. 2.

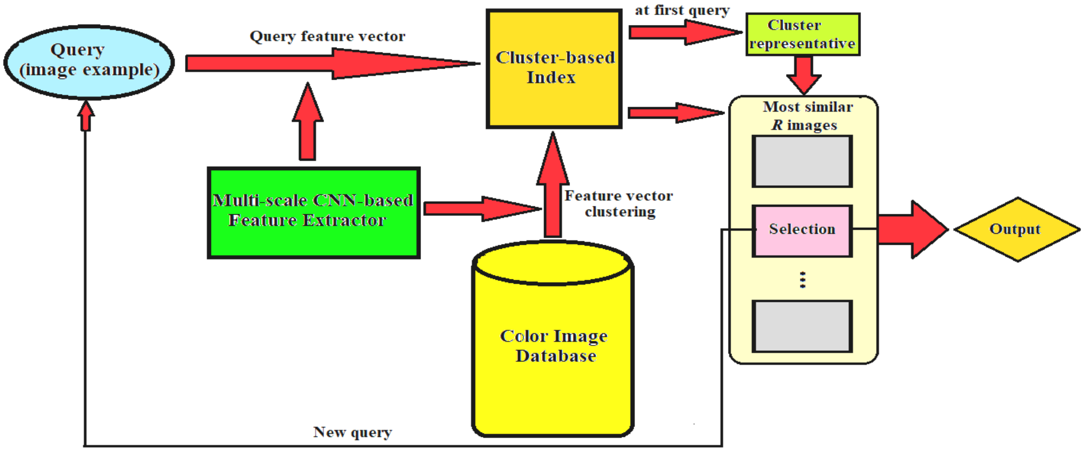


Fig. 2 – The architecture of the proposed content-based image retrieval system.

4. EXPERIMENTS

The multi-scale deep learning-based content-based image retrieval framework proposed in this work has been tested successfully on thousands of RGB color images. These CBIR simulations have been performed using MATLAB R2020a on Intel (R) Core (TM) i7-6700HQ CPU 2.60 GHz processor on 64 bits, operating Windows 10.

We have used the ODIDS (*OutDoor Images Data Set*) database containing over 5000 color images, grouped into 43 categories, which was developed at the Institute of Computer Science of the Romanian Academy [35], for our experiments. Since some ODIDS categories do not represent classes of similar images, we have created a validation data set based on this collection, which contains 1750 images belonging to 34 similarity classes and 80 images of ODIDS that do not belong to any of those classes. We have used 1300 of those class-belonging images as database images to be retrieved ($M = 1300$) and the remaining 450 as first query examples. The 80 no-class images have been also used for queries.

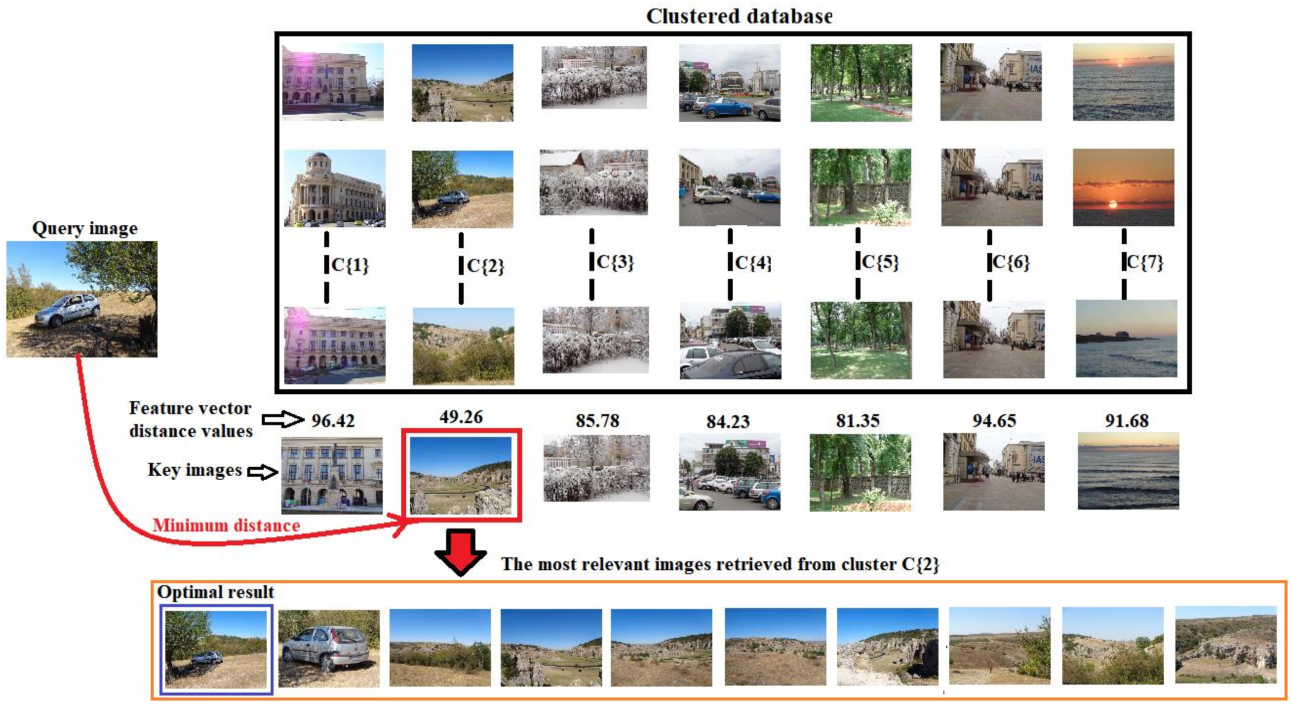


Fig. 3 – Example of a CBIR experiment.

Optimal color image retrieval results have been achieved for the following values of the system's parameters, which have been identified empirically, using the trial and error method on the validation set: the number of scales $K = 4$, the time step of the scale-space $\tau = 5$, the clustering threshold $T = 1.5$, the minimum cluster size $c = 3$ and the number of images retrieved at each query $R = 10$. Our CBIR system has a high retrieval accuracy rate, given the effectiveness of its clustering and feature extraction components. These components have a quite high complexity, due to the computational costs of the multi-scale analysis, CNN-based processes, anisotropic diffusion model approximation and graph partitioning schemes they involve, which may determine a high execution time for the clustering-based database indexing procedure. However, its running time depends also on the database size M and the sizes of the stored images. Unlike the image indexing, the retrieval process executes quite fast, being facilitated by that indexing structure.

One of our CBIR tests is partially described (only 1 query) in Fig. 3. A subset of ODIDS storing $M = 540$ outdoor images is indexed in $N_c = 7$ clusters (only few images are displayed for each of them) and their representatives are then identified. The cluster $C\{2\}$, whose key image is closest to the query (feature vector distance = 49.26), is used for retrieval. The optimal result is the first of its most relevant $R = 10$ images, so no relevance-feedback operation is required in this case.

The main component of this retrieval system, representing the content-based image clustering (indexing) technique, achieves high values of the evaluation measures *Precision* and *Recall*, which means that the obtained clusters contain very few *false positives* and *false negatives*. That implies effective retrieval results at each first query, almost all the images of the retrieved cluster being quite relevant to the query example. The performance of this CBIR framework has been assessed by using the *Mean Average Precision* (MAP) metric that combines Recall and Precision for ranked retrieval results [1]. A high value of this measure have been obtained when MAP is applied on the considered ODIDS–based validation image set and computed on the basis of first queries only.

Our technique has been compared to many other cluster-based CBIR approaches, which use global image features, like color moment invariants [2], GLCM [5] and 2D Gabor filters [6], combined global features [9], or local features, such as LBP [12], SURF [11], SIFT [10, 24] or MKSIFT [16], with clustering algorithms like K-means, hierarchical agglomerative clustering, Bayesian fuzzy clustering or other ML schemes [14]. It outperforms most of them, getting higher MAP values, and is slightly outperformed only by one of them (MKSIFT + BFC), as illustrated by the retrieval method comparison results displayed in Table 1.

Table 1

Method comparison: the MAP values achieved by some cluster-based CBIR models

CBIR technique	**Mean Average Precision (MAP)**
The proposed framework	0.9384
Color Moment Invariant + hierarchical agglomerative clustering	0.6471
GLCM + *K*-means	0.6843
SIFT + *K*-means	0.7302
MKSIFT + Bayesian fuzzy clustering	0.9453
Color, Shape, Texture features + agglomerative clustering algorithm [9]	0.8697
LBP features + *K*-means	0.7815
SURF + hierarchical agglomerative clustering	0.7521
2D Gabor filters + hierarchical agglomerative clustering	0.8162

5. CONCLUSIONS

A novel cluster-based CBIR system has been described in this research work. It groups successfully the images of a large database in an optimal number of similarity classes, without using a training set, then retrieves the images of the database that are relevant to various query images, by using the created cluster-based database indexation.

Its major component, which represents also the main contribution of this paper, is the content-based image indexing technique. It is based on a high-level color image feature extraction, which is performed using a nonlinear diffusion-based multi-scale analysis combined to deep learning networks, and an automatic graph-based image clustering approach. The obtained cluster-based indexing structure partitions successfully the database in clusters containing quite semantically similar images and facilitates considerably the content-based image retrieval process. The CBIR experiments, that have been performed on a voluminous color outdoor image database recently developed at our research institute (ODIDS), show the effectiveness of the proposed framework. It provides good retrieval results, achieves high performance metric values and outperforms other cluster-based CBIR models.

The convolutional neural networks involved in our image feature extraction process play an important role in this technique's performance, but the proposed multi-scale image analysis and *NCut* -based fully automatic clustering methods have essential roles as well. The scale-space representation created here by applying the approximation scheme that solves numerically a well-posed second-order PDE-based filter by converging to its variational solution improves considerably the feature extraction process, leading to some powerful image content descriptors. The size of these descriptors may represent a shortcoming of this CBIR system, since using feature vectors with thousands of coefficients implies a high computational cost of the indexing process. A dimension-reduction solution would be applying a Principal Component Analysis (PCA) to these 2D vectors [14], but without reducing the mean average precision score. This improvement of the CBIR system will represent the focus of our future research.

REFERENCES

1. I.M. HAMEED, S.H. ABDULHUSSAIN, B.M. MAHMMOD, *Content-based image retrieval: A review of recent trends*, Cogent Engineering, **8**, *1*, art. 1927469, 2021.
2. X. DUANMU, *Image retrieval using color moment invariant*, 2010 Seventh International Conference on Information Technology: New Generations, Las Vegas, Nevada, 2010, pp. 200–203.
3. G. QIU, *Color image indexing using BTC*, IEEE Transactions on Image Processing, **12**, *1*, pp. 93–101, 2003.
4. R. ASHRAF, M. AHMEDM, S. JABBAR, S. KHALID, A. AHMAD, S. DIN, G. JEON, *Content based image retrieval by using color descriptor and discrete wavelet transform*, Journal of Medical Systems, **42**, *3*, art. 44, 2018.
5. B. RAMAMURTHY, K. R. CHANDRAN, *Content based medical Image retrieval with texture content using gray level co-occurrence matrix and k-means clustering algorithms*, Journal of Computer Science, **8**, *7*, pp. 1070–1076, 2012.
6. D. ZHANG, A. WONG, M. INDRAWAN, G. LU, *Content-based image retrieval using Gabor texture features*, IEEE Transactions PAMI, Proc. of 1st IEEE Pacific Rim Conference on Multimedia, University of Sydney, Australia, 2000, pp. 392–395.

7. D. ZHANG, G. LU, *A comparative study on shape retrieval using Fourier descriptors with different shape signatures*, Proceedings of the International Conference on Intelligent Multimedia and Distance Education, 2001, pp. 1–9.
8. T. SUK, J. FLUSSER, *Affine moment invariants generated by graph method*, Pattern Recognition, **40**, *2*, pp. 2047–2056, 2011.
9. S. PANDEY, P. KHANNA, *Content-based image retrieval embedded with agglomerative clustering built on information loss*, Computers & Electrical Engineering, **54**, pp. 506–521, 2016.
10. G.A. MONTAZER, D. GIVEKI, *Content based image retrieval system using clustered scale invariant feature transforms*, Optik (Stuttg), **126**, *18*, pp. 1695–1699, 2015.
11. H. BAY, A. ESS, T. TUYTELAARS, L. VAN GOOL, *Speeded-Up Robust Features (SURF)*, Computer Vision and Image Understanding, **110**, *3*, pp. 346–359, 2008.
12. D. KULSHRESHTHA, V. SINGH, A. SHRIVASTAVA, A. CHAUDHARY, R. SRIVASTAVA, *Content-based mammogram retrieval using k-means clustering and local binary pattern*, Proc. ICIVC 2017, Chengdu, China, June 2–4, 2017, pp. 634–638.
13. T. WATANABE, S. ITO, K. YOKOI, *Co-occurrence histograms of oriented gradients for human detection*, IPSJ Transactions on Computer Vision and Applications, **2**, pp. 39–47, 2010.
14. C. M. BISHOP, *Pattern recognition and machine learning*, Springer, 2006.
15. Y. CHEN, J.Z. WANG, R. KROVETZ, *CLUE: Cluster-based retrieval of images by unsupervised learning*, IEEE Transactions on Image Processing, **14**, *8*, pp. 1187–1201, 2005.
16. B.M. KUMAR, R. PUSHPALAKSHMI, *An approach for image search and retrieval by cluster-based indexing of binary MKSIFT codes*, The Computer Journal, **63**, *6*, pp. 857–879, 2020.
17. J. WAN, D. WANG, S.C.H. HOI, P. WU, J. ZHU, Y. ZHANG, J. LI, *Deep learning for content-based image retrieval: A comprehensive study*, Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 157–166.
18. T. BARBU, M. COSTIN, A. CIOBANU, *Color-based image retrieval approaches using a relevance feedback scheme*, In: *New Concepts and Applications in Soft Computing* (eds. V.E. Balas, A.V. Koczy, J. Fodor), Series: Studies in Computational Intelligence, Vol. 417, pp. 47−55, Springer-Verlag, Berlin, 2013.
19. T. BARBU, M. LUCA, *Content-based iris indexing and retrieval model using spatial access methods*, The 12th International Symposium on Signals, Circuits and Systems (ISSCS 2015), IEEE, Iasi, Romania; July 9–11, 2015.
20. T. BARBU, A. CIOBANU, M. LUCA, *SAM-based image indexing and retrieval system using LAB color characteristics*, Latest Trends in Circuits, Systems, Signal Processing and Automatic Control (Proc. of the 5th Intl. Conf. on Circuits, Systems, Control, Signals, CSCS'14), Salerno, Italy, June 3–5, 2014, pp. 266–270.
21. T. BARBU, *Content-based image retrieval system using Gabor filtering*, Proceedings of the 20th International Workshop on Database and Expert Systems, DEXA'09, IEEE, Linz, Austria, August, 31 – September 4, 2009, pp. 236–240.
22. T. BARBU, M. COSTIN, A. CIOBANU, *Histogram intersection based image retrieval technique using relevance feedback*, Proceedings of the Third International Workshop on Soft Computing and Applications, SOFA 2009, Szeged (Hungary), Arad (Romania), July 29 – August 1, 2009, pp. 65−67.
23. T. BARBU, A. CIOBANU, *Color-based image retrieval technique using relevance feedback*, 3rd International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2009, Piteşti, Romania, July 3−5, 2009, Vol. 4, pp. 105−108.
24. T. BARBU, *Unsupervised SIFT-based face recognition using an automatic hierarchical agglomerative clustering solution*, Procedia Computer Science, **22**, pp. 385−394, 2013.
25. T. BARBU, *An automatic unsupervised pattern recognition approach*, Proceedings of the Romanian Academy, Series A: Mathematics, Physics, Technical Sciences, Information Science, **7**, *1*, pp. 73−78, 2006.
26. J. SPORRING, M. NIELSEN, L. FLORACK, P. JOHANSEN, *Gaussian scale-space theory*, vol. 8, Springer Science & Business Media, 2013.
27. T. BARBU, A. MIRANVILLE, C. MOROSANU, *A qualitative analysis and numerical simulations of a nonlinear second-order anisotropic diffusion problem with non-homogeneous Cauchy-Neumann boundary conditions*, Applied Mathematics and Computation, **350**, pp. 170−180, 2019.
28. T. BARBU, *Novel diffusion-based models for image restoration and interpolation*, Book Series: Signals and Communication Technology, Springer International Publishing, 2019.
29. P. JOHNSON, *Finite difference for PDEs*, University of Manchester, School of Mathematics, semester I, 2008.
30. J. MURPHY, *An overview of convolutional neural network architectures for deep learning*, Microway Inc., 2016.
31. J. DENG, W. DONG, R. SOCHER, L. J. LI, K. LI, L. FEI-FEI, *ImageNet: A large-scale hierarchical image database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248−255.
32. K. HE, X. ZHANG, S. REN, J. SUN, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770−778.
33. C. SZEGEDY, V. VANOUCKE, S. IOFFE, J. SHLENS, Z. WOJNA, *Rethinking the inception architecture for computer vision*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818−2826.
34. J. SHI, J. MALIK, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**, *8*, pp. 888−905, 2000.
35. A. IGNAT, M. LUCA, I. PĂVĂLOI, C.L. LAZĂR, *DENOL: efficient descriptors selection for automatic image retrieval*, 9th International Workshop on Soft Computing Applications, SOFA 2020, Arad, Romania, November 27−29, 2020, http://odids.iit.academiaromana-is.ro.

Received December 18, 2021