**Journal of Computer and Communications, 2015, 3, 66-73**
Published Online March 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.33012

# A New SIFT-Based Image Descriptor Applicable for Content Based Image Retrieval

**Davar Giveki[1], Mohamad Ali Soltanshahi[2], Fatemeh Shiri[3], Hadis Tarrah[4]**

[1]Iranian Research Institute for Information Technology (IranDoc), Department of Information Engineering, Tehran, Iran
[2]Department of Computer Science, University of Tehran, Tehran, Iran
[3]Department of Electrical Engineering, Tarbiat Modares University, Tehran, Iran
[4]Department of Electrical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
Email: Giveki@students.irandoc.ac.ir, Ali.soltanshahi@gmail.com, fatima.shiri@gmail.com, hs.tarrah.88@gmail.com

Received January 2015



## Abstract

**The large amounts of image collections available from a variety of sources have posed increasing technical challenges to computer systems to store/transmit and index/manage the image data to make such collections easily accessible. To search and retrieve the expected images from the data base a Content Based Image Retrieval (CBIR) system is highly demanded. CBIR extracts features of a query image and try to match them with extracted features from images in the data base. This paper introduces two novel methods that can be used as image descriptors. The basis of the proposed methods is built upon Scale Invariant Feature Transform (SIFT) method. After extracting image features using SIFT, k-means clustering is applied on feature matrix extracted by SIFT, and then two new kinds of dimensionality reduction reapplied to make SIFT features more efficient and realistic for image retrieval problem. Using the proposed strategies we can not only take the advantage of SIFT features but we can also highly decrease the memory storage used by SIFT features. Finally, proposed methods are compared with some other state of the art methods and as a result, our proposed retrieval system is faster and more accurate. Experimental results on two popular datasets, Corel (includes 1000 images) and OT data sets (includes 2688 images), show the superiority and efficiency of the proposed methods.**

## Keywords

**Content Based Image Retrieval, Scale Invariant Feature Transform (SIFT), Cardinality Matrix of Cluster-Sets (CMCS), Resultant Vector of Clustered SIFT Features (RVCSF)**

**How to cite this paper:** Giveki, D., Soltanshahi, M.A., Shiri, F. and Tarrah, H. (2015) A New SIFT-Based Image Descriptor Applicable for Content Based Image Retrieval. *Journal of Computer and Communications*, **3**, 66-73.
http://dx.doi.org/10.4236/jcc.2015.33012

## 1. Introduction

In the latest years, the amount of digital images has grown rapidly. Among the main reasons for that, one may mention digital cameras and high-speed Internet connections. Those elements have created a simple way to generate and publish visual content worldwide. That means that a huge amount of visual information becomes available every day to a growing number of users. Much of that visual information is available on the Web, which has become the largest and most heterogeneous image database so far. In that scenario, there is a crucial demand for image retrieval systems [1] [2], which could be satisfied by content-based image retrieval (CBIR) systems. In CBIR systems, the image descriptor is a very important element. It is responsible for assessing the similarities among images. Descriptors can be classified depending on the image property analyzed, like, for example, color, texture or shape descriptors, that analyze color, texture or shape properties, respectively. It is known that many image descriptors are application dependent, that is, their performances vary from one application to another. Therefore, conducting comparative evaluation of image descriptors considering different environments of use is very important. Literature presents several comparative studies for color, texture, and shape descriptors. A recent study [3] compares a large number of image descriptors in five different image collections for tasks of classification and image retrieval. Other studies are specific to certain properties: shape descriptors [4], texture and color descriptors [5]. Although these descriptors are using by the CBIR researchers nowadays, there has been a growing demand for using stronger features and more advanced computer vision and image processing algorithms. Feature detection is the process where we automatically examine an image to extract features that are unique to the objects in the image, in such a manner that we are able to detect an object based on its features in different images. This detection should ideally be possible when the image shows the object with different transformations, mainly scale and rotation, or when parts of the object are occluded. The processes can be divided in to 3 overall steps. The first step is Detection that automatically identifies interesting features, interest points this must be done robustly. The same feature should always be detected irregardless of viewpoint. The second step is Description. Each interest point should have a unique description that does not depend on the features scale and rotation. Finally, in Matching for a given input image, determine which objects it contains, and possibly a transformation of the object, based on predetermined interest points.

Scale Invariant Feature Transform (SIFT) [6], Speeded Up Robust Features (SURF) [7] and Histograms of Oriented Gradients (HOG) [8] are three algorithms that are being used for describing local features of the images. SIFT is a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images.

The most effective approaches currently used are in CBIR based on a vocabulary of local patches, called visual dictionaries [9]. That method is inspired in text retrieval, where a simple but effective method takes documents simply as “bags” (multi-sets) of words. In the same spirit, visual dictionary representations take images as bags of local appearances. That method has several important advantages, such as compactness (it encodes local properties into a single feature vector) and invariance to image/scene transformations. Creating a visual dictionary takes several steps. First and foremost, local characteristics must be obtained from a set of training images, usually by extracting local patches and describing them. The patches may be taken around Points of Interest (PoI) [10] or by dense sampling [11], and image descriptors, like the SIFT are used to extract feature vectors for each of them. Once the learning set of feature vectors is obtained, they are used to quantize the feature space (using, for example, k-means clustering) to choose a codebook of feature vectors representative of the training set. The clusters tend to contain visually similar patches and each cluster is a visual word of the dictionary. Once the dictionary is available, images are represented by statistical information about how they activate the visual words. The final image feature vector is commonly called bag of (visual) words (BoVW). When creating an image representation, one must be aware of the creation of really discriminating representations. Very small differences between images or objects must be encoded, while still being robust to specific photometric/geometrical transformations related to the domain. Therefore, there presentation must be very precise. However, there is a very important drawback with BoVW model that is the length of feature vector is considerably high. So, in this paper we decided to use SIFT so that we can tackle this important drawback. To this end we propose a new kind of

clustering method that highly decreases a dimension of the features extracted by SIFT while at the same time keeps the discriminative power of the image features. The rest of the paper is organized as follows. In Section 2 proposed approaches are described in details. Section 3 deals with experimental results and implementations. Finally, future work and conclusion are drawn in Section 4.

## 2. Proposed Approaches

In this section we propose two approaches for using SIFT features in CBIR. SIFT features are local features that are highly distinctive and suitable for image matching. However, in case of CBIR, a problem arises from the large number of keypoints extracted by SIFT algorithm. Storing these features and then matching them with images in the large databases is really impractical [12]. Therefore, it seems that new ways for decreasing the size of SIFT descriptor is needed. Data clustering is one way for achieving this goal. So, here we use clustering for decreasing the size of SIFT feature descriptor. By increasing the number of keypoints in images, SIFT feature descriptor gets too big. Suppose that an image A contains a set of *k* keypoints $P=\{p_1,p_2,\cdots,p_k\}$ in which feature vector of each point $p_i$ is $v_i=\{v_{i1},v_{i2},v_{i3},\cdots,v_{i128}\}$.

### 2.1. Cardinality Matrix of Cluster-Sets (CMCS)

In our proposed methods k-means clustering is used to decrease the dimension of feature matrix. To this end, in our first method, feature vector of each keypoint is divided into 16 parts so that each part includes 8 key points. Clustering these parts yields a new feature vector. So, the feature vector $v_j$ is divided into 16 parts as following:

$$\begin{aligned} v_j &= \{v_{j1},v_{j2},v_{j3},\cdots,v_{j128}\} \ \underline{\text{division of Feature vector}} \\ v'_{j1} &= \{v_{j1},v_{j2},v_{j3},\cdots,v_{j8}\},\cdots,v'_{j16}=\{v_{j121},v_{j122},v_{j123},\cdots,v_{j128}\} \end{aligned} \tag{1}$$

We construct a k-means clustering named K with 8 clusters. So, $K(v'_{jl})=n$ if and only if $v'_{jl}$ belongsto the $n^{th}$ cluster. Therefore, each $v'_{jl}$ belongs to one cluster and feature vector $v_j$ can be represented in a more compact way using 8 clusters. But, for each image, A, the number of keypoints vary and this causes a big problem to define a fixed size feature vectors and for comparing images. Therefore, we have to aggregate extracted features from points. In order to do that, we define the following Cluster-Set:

$$M'_{ij}=\{v'_{lj}\mid K(v'_{ij})=i,v_l\in P\} \tag{2}$$

$M'_{ij}$ is a set of parts belongs to the ith cluster. This set helps to construct a global matrix, *G* as following:

$$G=\begin{bmatrix} |M'_{11}| & \cdots & |M'_{1,16}| \\ \vdots & \ddots & \vdots \\ |M'_{81}| & \cdots & |M'_{8,16}| \end{bmatrix}_{8\times 16} \tag{3}$$

where $|M'_{ij}|$ is cardinality of the set $M'_{ij}$. **Figure 1** shows the experimental results of the first proposed method.

### 2.2. Resultant Vector of Clustered SIFT Features (RVCSF)

In the second method, we first compress local feature vectors based on the concept of SIFT features. To do so, resultant of histogram bins in the same directions is computed. So a new compact feature vector $\bar{v}_j$ is built as following:

$$v_j=\{v_{j1},v_{j2},v_{j3},\cdots,v_{j128}\}\rightarrow \bar{v}_j=\left(\sum_{l=0}^{15}v_{j(l*8+1)},\sum_{l=0}^{15}v_{j(l*8+2)},\cdots,\sum_{l=0}^{15}v_{j(l*8+8)}\right) \tag{4}$$

As a result of this compression the local features tends to be more global while they are highly distinctive yet. This improvement results in better performance in the scope of CBIR. To make improvement we decided to cluster new compact feature vector using k-means, K, with 16 clusters. Hence, $K(\bar{v}_j)=n$ if and only if $\bar{v}_j$

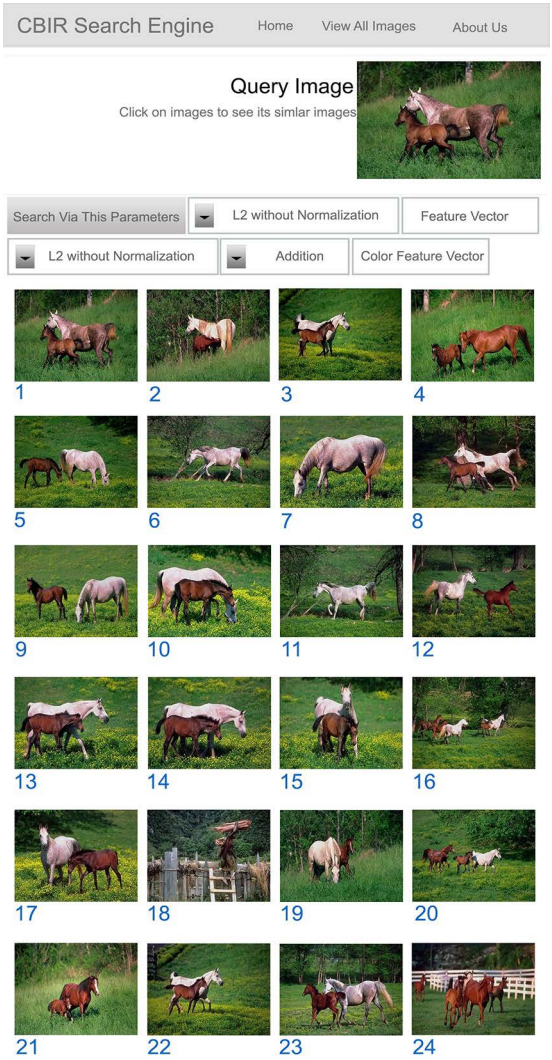


**Figure 1.** The retrieval results of the horse images from corel database [13].

belongs to the nth cluster. By using clustering K, each feature descriptor can be referred as a cluster. Therefore, with aggregating clustering results, we build the global feature vector $G_2$ as following:

$$G_5 = \left[\left|\bar{M}_1\right|, \left|\bar{M}_2\right|, \ldots, \left|\bar{M}_{16}\right|\right] \tag{5}$$

where $\bar{M}_k = \left\{P|K\left(\bar{v}_j\right) = k, p \in P\right\}$, $k = 1, 2, \cdots, 16$.

In this method we used 16 clusters. Experimental results show that by increasing the number of clusters the efficiency of the proposed method is highly increased. **Figure 2** shows the result of the second method with 16 clusters. We should mention that the best results were achieved when we used the second method with 16 clusters and dividing images into 16 parts.

We also noticed that the location of the points in the images play an important role in the SIFT algorithm so to take this point into account, we decided to divide images into some parts and compute their G2 feature. In this case the distance between 2 images is sum of the distances between corresponding parts.

## 3. Experimental Results and Comparisons

The experiments in the retrieval scenario are based on traditional image datasets which comprise the semantic-

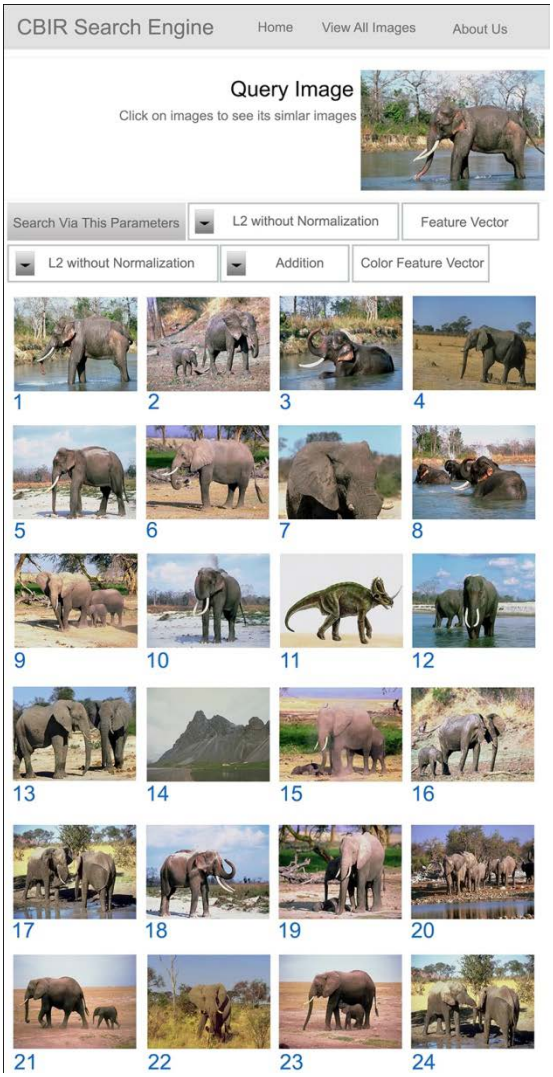


**Figure 2.** Retrieval results of the elephant image from corel database [13].

search application. The description of the datasets used in this paper is as following.

## 3.1. Dataset Description

### 3.1.1. Wang, Li and Wiederhold (Corel) Dataset.

Wang *et al*. [13] dataset contains 1000 images in 10 categories; Africa, Beaches, Building, Bus, Dinosaur, Elephant, Flower, Horses, Mountain and Food. There are 100 images in each category. The images are with the size of 256 × 384 or 384 × 256 pixels.

### 3.1.2. Oliva and Torralba (OT) Dataset

Oliva and Torralba [14] dataset includes 2688 images classified as 8 categories: 360 coasts, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets and 356 tall buildings. Note that river and forest scenes are all considered as forest, moreover there is not a specific sky scene since almost all of the images contain the sky object. These annotations make a higher inter-class variability. Most of the scenes present a large intra-class variability. The average size of each image is 250 × 250 pixels. The class numbers in our experiments are: ‘coast = 1’, ‘forest = 2’, ‘highway = 3’, ‘inside city = 4’, ‘mountain = 5’, ‘open country = 6’, ‘street = 7’ and ‘tall building = 8’.

We analyze the performance of our proposed method by considering the second proposed approach (it should be noted that the performance of the second approach is in average 4% - 6% higher than the first one). Finally, we compare our proposed model with other state of the art methods on these datasets.

**Table 1** shows the results of retrieval on OT dataset using different methods.

**Table 2** shows the comparison results of our proposed method on Corel dataset.

In addition in order to give more insight into the behavior of the proposed descriptor we report the confusion matrices of our method for each category on Corel and OT datasets. **Figure 3** and **Figure 4** show the confusion matrices of proposed method for Corel and OT datasets respectively.

**Table 1.** Comparison of our method to other methods on OT dataset.

| Method | mAP |
|---|---|
| **SIFT-HOG-SVM** | **88.35%** |
| **Proposed Method in [15]** | 72% |
| **Proposed Method in [16]** | 59.2% |
| **Proposed Method in [17]** | 77.4% |
| **Proposed Method in [18]** | 63.5% |
| **Proposed Method in [19]** | 67.4% |

**Table 2.** Comparison of our method to other methods on Corel dataset.

| Method | mAP |
|---|---|
| **SIFT-HOG-SVM** | **84.5%** |
| **Proposed Method in [20]** | 39% |
| **Proposed Method in [21]** | 37.2% |
| **Proposed Method in [22]** | 53.1% |
| **Proposed Method in [23]** | 55.6% |
| **Proposed Method in [24]** | 58.9% |


**Figure 3.** Confusion matrix of our proposed descriptor on Corel dataset.

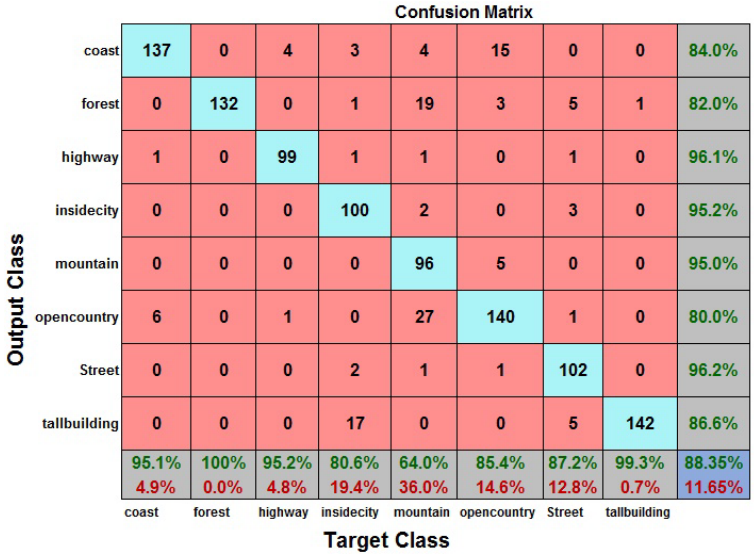


**Figure 4.** Confusion matrix of our proposed descriptor on OT dataset.

## 4. Conclusion and Future Work

This paper proposes 2 new approaches for Content Based Image Retrieval using SIFT method. The aim of these two approaches is tackling 2 important drawbacks of SIFT method namely, its memory usage and its matching time, that prevents it to be used as a reliable method for CBIR problem. In the first approach using k-means clustering, the feature matrix extracted by SIFT method is clustered to 16 clusters so that instead of having a huge matrix with dimension k*128, which k is the number of keypoints, we will have a reduced matrix of size 8*16. In the second approach using the fact that each keypoint is described by 8 dominant directions, the feature matrix is reduced to a vector of 1*18. In both methods there is no loss of discriminative power of SIFT while the size of feature matrix is highly reduced. The proposed approach shows high performance when facing with object images namely the images with an object in them. This fact is because of the concept of SIFT. However, for complex images like nature images the performance of SIFT needs more improvement. To this end we are working on embedding the bag of visual words skim into out proposed methods.

## References

[1] Kherfi, M.L., Ziou, D. and Bernardi, A. (2004) Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Computing Surveys*, **36**, 35-67. http://dx.doi.org/10.1145/1013208.1013210

[2] Datta, R., Joshi, D., Li, J. and Wang, J.Z. (2008) Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, **40**, 1-60. http://dx.doi.org/10.1145/1348246.1348248

[3] Deselaers, T., Keysers, D. and Ney, H. (2008) Features for Image Retrieval: An Experimental Comparison. *Information Retrieval*, **11**, 77-107. http://dx.doi.org/10.1007/s10791-007-9039-3

[4] Mingqiang, Y., Kidiyo, K. and Joseph, R. (2008) A Survey of Shape Feature Extraction Techniques. Pattern Recognition Techniques, Technology and Applications, ISBN: 978-953-7619-24-4, InTech.

[5] Otávio A.B. Penatti, Eduardo Valle, Ricardo da S. Torres. (2012) Comparative study of global color and texture descriptors for web image retrieval. J. Vis. Commun. Image R. 23, 359-380. http://dx.doi.org/10.1016/j.jvcir.2011.11.002

[6] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, 346-359, 2008. http://dx.doi.org/10.1016/j.cviu.2007.09.014

[8] N. Dalal and B. Triggs. Histograms of Oriented Gradients f or Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 886-893.

[9] Sivic, J. and Zisserman, A. (2003) Video Google: A Text Retrieval Approach to Object Matching in Videos. In: *Inter-*

*national Conference on Computer Vision*, Vol. 2, 1470-1477. http://dx.doi.org/10.1109/ICCV.2003.1238663

[10] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and Gool, L. (2005) A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, **65**, 43-72. http://dx.doi.org/10.1007/s11263-005-3848-x

[11] van de Sande, K.E.A., Gevers, T. and Snoek, C.G.M. (2010) Evaluating Color Descriptors for Object and Scene Recognition. *Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1582-1596. http://dx.doi.org/10.1109/TPAMI.2009.154

[12] Beis, J., and Lowe, D.G. (1997) Shape Indexing Using Approximate Nearest-Neighbor Search in High-Dimensional Spaces. *Conference on Computer Vision and Pattern Recognition*, Puerto Rico, 1000-1006.

[13] Wang, J.A., Li, J. and Wiederhold, G. (2001) SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries. *IEEE Trans. Pattern Anal. Mach. Intell*, **23**, 947-963. http://dx.doi.org/10.1109/34.955109

[14] Oliva, A. and Torralba, A. (2001) Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, **42**, 145-175. http://dx.doi.org/10.1023/A:1011139631724

[15] Lazebnik, S., Schmid, C. and Ponce, J. (2006) Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2169-2178.

[16] Pujari, J. and Hiremath, P.S. (2007) Content Based Image Retrieval Using Color, Texture and Shape Features. *Proceedings of the International Conference on Advanced Computing and Communications*, 780-784.

[17] Pietikäinen, M., Ahonen, T. and Takala, V. (2005) Block-Based Methods for Image Retrieval Using Local Binary Patterns. *Proceedings of the* 14*th Scandinavian Conference on Image Analysis*, 2005, 882-891.

[18] Triggs, B. and Jurie, F. (2005) Creating Efficient Codebooks for Visual Recognition. In: *ICCV*, Vol. 1, 604-610.

[19] Wang, H., Teng, P. and Liang, W. (2011) Packed Dense Interest Points for Scene Image Retrieval. *IEEE Sixth International Conference on Image and Graphics* (*ICIG*), 2011, 789-794.

[20] Huang, P.W. and Dai, S.K. (2003) Image Retrieval by Texture Similarity. *Pattern Recognition*, **36**, 665-679. http://dx.doi.org/10.1016/S0031-3203(02)00083-3

[21] Gelzinis, A., Verikas, A. and Bacauskiene, M. (2007) Increasing the Discrimination Power of the Co-Occurrence Matrix-Based Features. *Pattern Recognition*, **40**, 2367-2372. http://dx.doi.org/10.1016/j.patcog.2006.12.004

[22] Lin, C.-H., Chen, R.-T. and Chan, Y.-K. (2009) A Smart Content-Based Image Retrieval System Based on Color and Texture Feature. *Image and Vision Computing*, **27**, 658-665. http://dx.doi.org/10.1016/j.imavis.2008.07.004

[23] ElAlami, M.E. (2011) A Novel Image Retrieval Model Based on the Most Relevant Features. *Knowledge-Based Systems*, **24**, 23-32. http://dx.doi.org/10.1016/j.knosys.2010.06.001

[24] ElAlami, M.E. (2014) A New Matching Strategy for Content Based Image Retrieval System. *Applied Soft Computing*, **14**, 407-418. http://dx.doi.org/10.1016/j.asoc.2013.10.003