Kuwait J. Sci. **45** (**1**) pp 54-69, 2018

A novel image retrieval based on rectangular spatial histograms of visual words

Zahid Mehmood[1,2,*], Syed M. Anwar[1], Muhammad Altaf[3]

[1]*Dept. of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*

[2]*Dept. of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*

[3]*Dept. of Mathematics, University of Engineering and Technology, Taxila 47050, Pakistan*

**Corresponding author: zahid.mehmood@uettaxila.edu.pk*

Abstract

Content-based image retrieval (CBIR) provides a solution to search the images that are similar to a query image. From last few years, the bag-of-visual-words (BoVW) model gained significance and improved the performance of CBIR. In a standard BoVW model, an image is represented as an order-less histogram of visual words, by ignoring the spatial layout of the image. The spatial layout carries significant information that can enhance the image retrieval accuracy. In this paper, we present a novel method of image representation, which is based on the construction of histograms over two rectangular regions of an image. Division of the image into two rectangular regions at the time of construction of histograms adds the spatial information to the BoVW model. The proposed image representation uses separate visual words for upper and lower rectangular regions of an image. The experimental analysis carriedout on two image datasets validates that the proposed image representation based on the division of an image into histograms of rectangles increases the performance of image retrieval.

Keywords: Bag-of-visual-words; content-based image retrieval; rectangular spatial histograms; support vector machine.

1. Introduction

The process of retrieving relevant images on the basis of image contentsis referred to as CBIR (Tousch *et al*., 2012). Its applications are in surveillance, data mining,internet image search, video search, geographical information systems, andmedical imaging domain (Datta*et al*., 2008). In CBIR, the query image is used to search for there levant images from an image archive (Datta *et al*., 2008). The low-level features are used to represent images in CBIR (Tousch *et al*., 2012) and the semantic gap between high-level image concepts and low-level image features makes CBIR a challengingresearch problem (Zhang *et al*., 2012). The focus of research in CBIR is to retrieve those images, whose visual contents are similar to the query image (Datta *et al*., 2008; Tousch *et al*., 2012; Zhang *et al*., 2012). From last few years, the BoVW model gained a lot of attention (Lazebnik *et al*., 2006; Philbin *et al*., 2007; Khan *et al*., 2012) in CBIR that significantly increase the efficiency of image search. The spatial information of an image is lost when we represent an image by constructing a single histogram from the whole image based on the traditional BoVW methodology (Sivic &Zisserman 2003, Cao *et al*., 2010). The spatial layout of animagecontains information about the salient objects that enhance the performance of image retrieval (Lazebnik *et al*., 2006; Philbin *et al*., 2007; Khan *et al*., 2012). Large vocabulary size, query expansion, and soft quantization are usedto enhance the performance of content-based image matching. All of these approaches lack spatial information (Philbin *et al*., 2007; Cao *et al*., 2010).

The appearance of similar visual contents in the images of different semantic categories decreases the performance of image retrieval (Mehmood *et al*., 2016). According to the proposed research, spatial information is extracted from an image by dividing it into two rectangular regions. Figure 1 represents the images from four semantic classes (Beach, Africa, Elephants, and Mountains) of Corel-A image dataset. The discriminating information like the sky, people, trees, elephants, and mountains are likely to be in the divided rectangular regions of the image. Division of an image into two rectangles atthe time of construction of histograms of visual words seems to be a solution for the reduction of the semantic gap and it requires aless computational cost. By applying this approach, two separate histograms are constructed from the upper and lower rectangular regions of an image. Each rectangular region contains discriminating information in the form of visual words as shown in Figure 2.

Fig. 1. Corel images of different classes (Beach, Africa, Elephants, and Mountains) with a semantic rectangular relationship

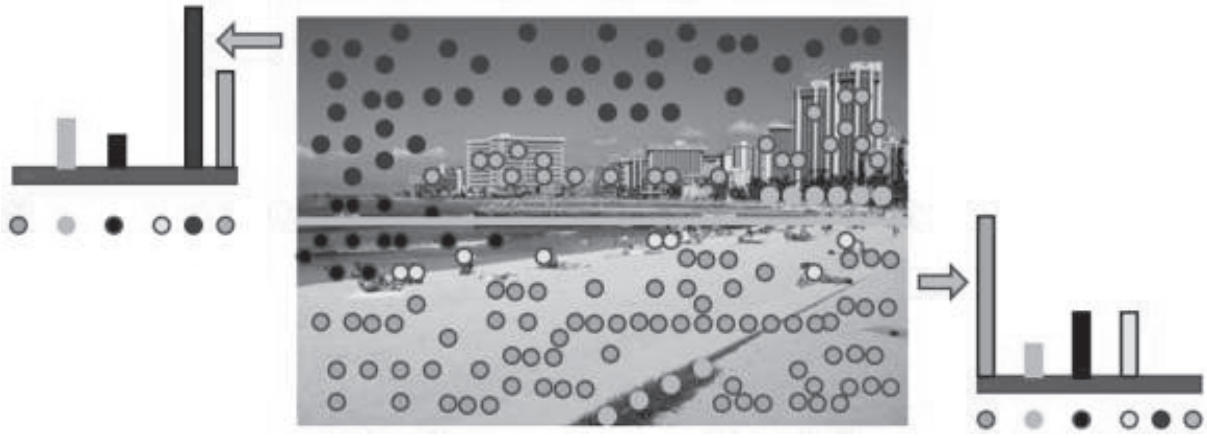

Fig. 2. Proposed technique based on rectangular spatial histograms of visual words

In this paper, we are presenting a novel method of image representation in the form of rectangular spatial histograms of visual words by dividing an image into two equal rectangular regions. Spatial rectangular histograms of visual words are constructed from each rectangular area of an image. The constructed histograms are concatenated and this information is added to the inverted index of BoVW representation. The main contributions of this paper are as follows:

1. Addition of spatial information to the inverted index of the standard BoVW model.
2. An image representation in the form of rectangular spatial histograms ofvisual words.
3. Reduction of semantic gap between high-level concepts and low-level features of an image.
4. Automatic image annotation based on classification scores.

The remaining sections of this article are categorized as follows: Section 2 provides an overview of the relevant research work. Section 3 is about the proposed methodology that is based on BoVW model. Section 4 is about experimental details and the results obtained from Corel-A and Ground truth image datasets and Section 5 conclude proposed technique and discuss future work.

2. Related work

CBIR is tremendously growing research area since last three decades (Rui *et al*., 1999). There are mainly two categories of visual features that are known as domain specific visual features and general visual features (Cao *et al*., 2010). The domain specific visual features are sensitive to their applications and are particularly used for different applications such as object recognition and face recognition. General features are low-level features of an image whichincludes color, shape, and texture (Safar, 2009; Mahmood *et al*. 2017). These features are not dependent on the particular applications and are used for CBI Ras well as for various image processing based applications (Mahmood *et al*., 2015; Mahmood *et al*. 2017). Feature extraction approaches can be categorized into four different classes that are transform-based, structural-based, model-based and statistical-based approaches (Rui *et al*., 1999; Datta *et al*., 2008).

A number of interest points based detectors are proposed like Harris-Affine (Tuytelaars and Van Gool 1999), Hessian-Affine (Tuytelaars & Van Gool 1999), Edge-Based Region detector (Tuytelaars & Van Gool 2000), Intensity Extrema (Kadir *et al*., 2004) and salient regions (Liu *et al*., 2010). Lowe (2004) proposed interest points based feature descriptor to match the different scenes and objects in the images named as scale-invariant feature transform (SIFT) that is rotation and scale-invariant. It performs well in extreme conditions like the addition of noise and 3D viewpoint change in the images. J hanwar *et al*. (2004) proposed CBIR model known as motif co-occurrence matrix (MCM), in which each image is divided into square grid and reported that MCM increases the performance of image retrieval. Heikkilä *et al*. (2009) proposed an improvement in local binary patterns (LBP) by using the interest points to characterize the distribution of local patches. They combined the strengths of LBP and SIFT features andis known as center-symmetric LBP. Liu *et al*. (2011) proposed micro-structure based descriptor for CBIR, named as microstructure descriptor (MSD). In this descriptor, they have utilized the edgeorientation similarity with the color in micro structures. MSD has benefits of structural texture description and statistical methods. Bosch *et al*. (2007) proposed a method to recognize the scene categories based on global geometric correspondence, in which an image is partitioned into increasingly finesub-regions and thehistogram is computed on that local features found insideeach sub-region. This is a simple and computationally efficient extensionof BoVW model. In LBP, image micro patterns are gathered in a single histogram of the image. However, it loses the spatial information between LBPs. To overcome the aforementioned problem, Nosaka *et al*. (2011) used the properties of spatial co-occurrence to encode the LBP's of each micro patterns. By using the spatial co-occurrence, more information can be preserved and it also performs well for light variation. Binary descriptors have the advantage that the computational cost is less and it needs less storage space to store, due to its dimensions. However, it decreases the classification rate.

Xie *et al*. (2015) proposed a new framework, in which they have combined the shape and color information of images. They have used SIFT-based BoVW approach and hierarchical MAX (HMAX) model to combine the shape information with color information. Group-independent spanning tree (GIST) algorithm is used for scene classification and Berkeley algorithm is used for segmentation. The main intent of presented CBIR technique based on color and shape features by Jiji & DuraiRaj (2015) was the analysis of dermatology images. Three phases were employed for effective retrieval of skin lesions namely shape and color feature vectors, particle swarm optimization (PSO) technique, and receiver operating characteristics (ROC) curve. The extracted features were stabilized using min-max normalization, and PSO was embedded for multi-class classification, to analyze the search space more efficiently. The ROC curve proved that presented structural design was well promoted to computer-aided diagnosis of skin lesions. Tomašev & Mladenić (2015) proposed a graphical tool for visualization of images. The main intent of the tool was to represent the features and their evaluation in the context of CBIR as well as to give recommendations for efficient image retrieval. The tool allowed its users to select multiple feature representations and matrices in order to obtain the desired performance. This tool was designed by utilizing multiple instances learning, classification, and re-ranking procedures in order to sustain the reliability of the CBIR results. The proposed tool provides representation of global features of images only. Zeng *et al*. (2016) has proposed Gaussian mixture model (GMM) for color quantization to constructs patiogram histogram. In spatiogram histogram based image representation, each color is assigned different weights according to the location of pixels contributing for each color bin. Classification of materials is a difficult task, due to the variation in pose, illumination, shadows, structure, and different texture of objects (Ullah & Baharudin, 2016).

3. The BoVW methodology

In BoV Wmethodology, local features are computed from the image. The local features contain information about salient objects of the image. For CBIR techniques based on the BoVW methodology, the first step involves the extraction of keypoints that are utilized to calculate the features of an image.

A clustering algorithm such as k-mean (Sivic & Zisserman, 2003) is applied to the set of training images to construct the dictionary consisting of t-visual words. To represent an image in the form of visual words, the distance between the feature descriptor is calculated using visual words from the dictionary. The nearest visual words are assigned to the descriptor by calculating the Euclidean distance between visual words of the dictionary and the descriptor. An image is represented in the form of a histogram and the feature vector of each histogram represents the occurrence of the respective visual words. The image representation in the form of order-less histogram provides flexibility to the change in position and view point. A classification algorithm is applied to the training images for learning, and test images are given as an input to the trained classifier to determine the output labels. The classifier output label determines the semantic category of the test image, while the distance between the feature vector of query image to the images placed in an image collection with the same class label determines the output of retrieved images.

3.1. Methodology of the proposed technique

For an image representation based on the BoVW methodology, spatial information about the salient objects of the image is lost, due to the formation of the single histogram from the whole image (Philbin *et al*., 2007). The spatial information provides discriminating details in recognition problems (Philbin *et al*., 2007). The framework of the proposed technique based on rectangular histograms is shown in Figure 3.

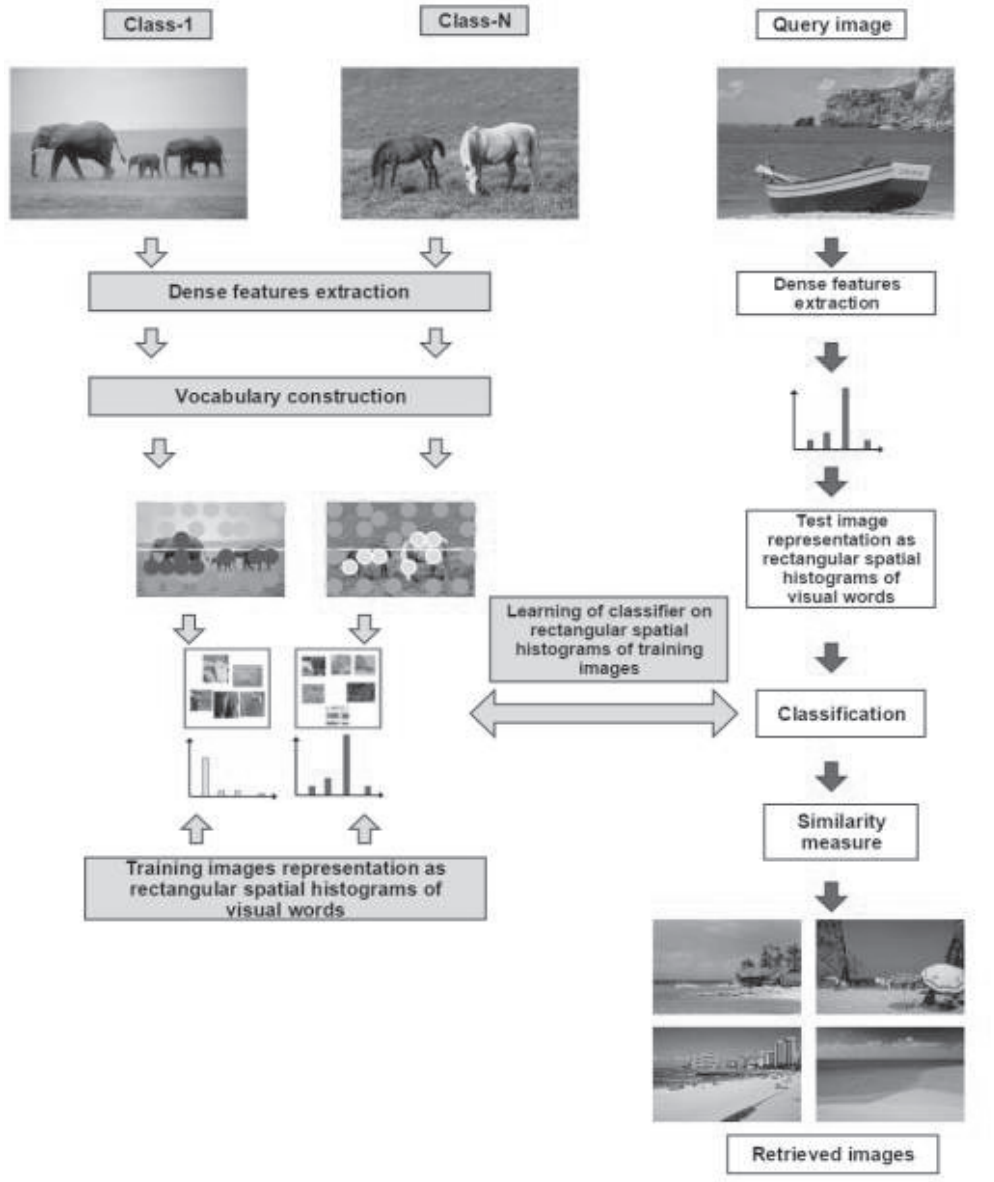

Fig. 3. Framework of the proposed technique based on the image representation in the form of rectangular spatial histograms of visual words

The detailed methodology of the proposed technique is as follows:

1. Consider an image which is represented by the following mathematical equation:

$$F = \left(z_{x,y}\right) \tag{1}$$

Where $\left(z_{x,y}\right)$ is the pixel at position (x, y).

2. The SIFT features (Lowe, 2004) are computed from an image over a dense grid (Vedaldi &Fulkerson 2010), represented as:

$$h(t,i,j) = (k_i k_j * \bar{J}_t)\left(T + m\sigma \begin{bmatrix} x_i \\ y_i \end{bmatrix}\right) \tag{2}$$

$$\bar{J}_t(x) = \omega_{ang}(< J(x) - \theta_t)|J(x)| \tag{3}$$

Where is the scale, θ is the orientation, m is the descriptor magnification factor, T is the affine transformation, J is the gradient, and h is the histogram of descriptors.

The kernels k_i and k_j are defined by the following mathematical equations:

$$k_i(x) = \frac{1}{\sqrt{2\pi}\sigma_{win}} exp\left(\frac{-1(x-x_i)^2}{2\sigma_{win}^2}\right) \omega\left(\frac{x}{m\sigma}\right) \tag{4}$$

$$k_j(y) = \frac{1}{\sqrt{2\pi}\sigma_{win}} exp\left(\frac{-1(y-y_i)^2}{2\sigma_{win}^2}\right) \omega\left(\frac{y}{m\sigma}\right) \tag{5}$$

3. k-means clustering is applied to construct a codebook consistingof visual words, represented as W_v :

$$W_v = \{v_1, v_2, \dots, v_t\} \tag{6}$$

Where v_1 to $v_t$are the visual words.

The dense SIFT features are computed from an image and feature space is quantized. For the construction of a histogram from the upper rectangular region of an image, mapping of each visual word is carried on by using the upper rectangular region of an image. Similarly, for the construction of a histogram from the lower rectangular region of an image, the mapping of each visual word is carried on by using the lower rectangular region of an image. The nearest words are assignedto the quantized descriptors according to the following equation:

$$v(d_m) = arg\ min_{v \epsilon W_v}\ Dist(v, d_m) \tag{7}$$

Where (d_m) is representing the visual word assigned to the m^{th} descriptor, while Dist (v, d_m) is the distance between the descriptor d_m and the visual word v. Each image is represented as a collection of two rectangular regions and each rectangular region is represented by the visual words.

4. Two histograms of M visual words are computed from a single image. The visual words for the upper and lower rectangular regions are mapped to the upper and lower rectangular regions of the image that are selected by applying the Equations (8) and (9) respectively. The four points from each of the upper and lower rectangular regions are selected from each image by applying following equations, respectively:

$$P_{u1} = F(1,1), P_{u2} = F(1,w), P_{u3} = F(h/2,1), P_{u4} = (h/2,w) \tag{8}$$

$$P_{l1} = F\left(\frac{h}{2} + 1,1\right), P_{l2} = F\left(\frac{h}{2} + 1,w\right), P_{l3} = F(h,1), P_{l4} = F(h,w) \tag{9}$$

Where P_{ui} and P_{li} represent the points of upper and lower rectangular regions of the image respectively and i varies from 1 to 4. w and h and represent the width and height of the image, respectively.

5. Both of the computed histograms are concatenated and this information is added to the inverted index of BoVW representation. Consider M as the number of visual words of the codebook. Let N_i be the set of the descriptors that are mapped to the visual word v_j then the j^{th} bin of the histogram of visual words h_i is the cardinality of the set N_i can be represented as:

$$h_i = Card(N_i)\ \ where\ N_i = \left\{d_m, m\epsilon(1,2,\dots,M)|v(d_m) = v_j\right\} \tag{10}$$

3.2 Image classification

support Vector Machine (SVM) is a state-of-the-art supervised learning classification method. The linear SVM separate the two classes by using a hyperplane. The dataset with two classes can be represented as:

$$\left\{\left(P_j, Q_j\right)\right\}_{j=1}^{M}, Q_j = \{+1, -1\} \tag{11}$$

Where P_j and Q_j are input classes and +1 and -1 are the correspondence labels of the classes, respectively. The hyper planes are generated by finding the values of coefficients:

$$w^T.P + b = 0 \tag{12}$$

Where w is weight vector and b is bias. The maximum margin is determined by $2/||w||$ hyper planes and the two classes are separable from each other according to the following equations:

$$w^T.P_j + b = 1 \tag{13}$$

$$w^T.P_j + b = -1 \quad (14)$$

This can be expressed equivalently as:

$$Q_j\left(w^T.P_j + b\right) \geq 1 \quad (15)$$

The kernel method mentioned in Shawe-Taylor & Cristianini (2004) is applied for image classification of the proposed technique and details of the SVM Hellinger kernel are mentioned in Mehmood *et al*.(2016). The histograms constructed over upper and lower rectangular regions of each imageare normalized by applying normalization. The SVM Hellinger kernel (also known as the Bhattacharyya coefficient) (Vedaldi &Zisserman, 2012) is applied on the normalized histograms by applying the following equation:

$$H(n,n') = \sum_j \sqrt{n(j)n'(j)} \quad (16)$$

Where n and n' are the normalized histograms.

One versus one rule is applied and for m number of classes $m.(m-1)/2$ classifiersare constructed and each one trains the data using two classes. The class label of the image is selected by using the maximum score value obtained from $m.(m-1)/2$ classifiers.

In order to find the best efficiency of the proposed technique, the performance is also evaluated by using radial basis function-artificial neural network (RBF-ANN) (Haykin & Network, 2004) classifier and performance evaluation of both classifiers are presented in Section 4.

4. Experimental details and discussion

This section is about the performance measurement parameters, experimental details, and result discussion of the proposed technique based on the image representation in the form of rectangular spatial histograms of visual words. The performance of the proposed technique is measured on the Corel-A and the Ground truth image archives or datasets. The images of each image archive arecategorized into training (70%) and test (30%) sets. The codebook is computed from the training set and mean average precision (MAP) of the proposed technique is reported on the basis of images retrieved from the testset. Each experiment is performed 10 times to report the average value of the MAP performance due to unsupervised behavior of *k*-means clustering technique. The performance measurement parameters (i.e. precision and recall) are calculated on the basis of the following formulas:

$$Precision = \frac{Total\ relevant\ retrieved\ images}{Total\ retrieved\ images\ at\ output} * 100 \quad (17)$$

$$Recall = \frac{Total\ relevant\ retrieved\ images}{Total\ relevent\ images\ in\ a\ class} * 100 \quad (18)$$

The details about the experimental parameters are given below:

1. Size of the codebook: According to Sivic & Zisserman (2003), the CBIR performance is affected by varying the size of the codebook. Increase in the size of codebook at certain level increases retrieval precision and larger size codebook result to over-fit. In order to find the best performance of the proposed technique, codebooks of various sizes are formulated from the training set for the Corel-A and the Groundtruth image archives.
2. Pixel step size: For a precise content-based image retrieval (Hassner *et al*., 2012), we computed dense SIFT feature descriptors using two different scales (i.e. 4 and 6). The step size is one of the major parameters that controls the spatial resolution of the dense grid. In our experiments, we computed dense SIFT feature descriptors by using the pixel step sizes of 5, 10, and 15. For a pixel step sizes of 5, 10, and 15, we computed dense SIFT feature descriptors after every 5th, 10th, and 15th pixel, respectively.
3. Feature percentage for the codebook construction: In our experiments, we constructed the codebook from training images by using different percentages of dense SIFT feature descriptors from each image (10%, 25%, 50%, 75%, and 100%).

4.1. Performance measurements on the Corel- Aimage archive

The Corel-A image archive1 is a publicly available image archive that is used for the performance evaluation of proposed technique and comparison of results are performed with the state-of-the-art CBIR techniques (Lin *et al*., 2011; Wang *et al*., 2013; Tian *et al*., 2014; Mehmood *et al*., 2016; Zeng *et al*., 2016). The total number of images in the Corel-A image archive are 1000 that are organized into ten different semantic classes namely: “Food”, “Mountains”, “Flowers”, “Horses”, “Elephants”, “Buses”, “Dinosaurs”, “Buildings”, “Beach”, and “Africa”. Figure 4 is presenting the sample images of all semantic classes of the Corel-A image archive.

Fig. 4. Sample of semantic classes fromthe Corel-A image archive

Each semantic class contains 100 images with a resolution of 256 x 384 pixels or 384 x 256 pixels of each image. Different sizes (20, 50, 100, 200, 300, and 400) of codebook are computed to evaluate the best performance of the proposed method. The mean average precision (MAP) as a function of codebook size and percentage offeatures computed for each image used in the codebook construction is presented in Table 1, Table 2, and Table 3 (for the dense pixel step sizes of 5, 10, and15), respectively.

Table 1. MAP performance of the proposed technique on the Corel-A image archive using pixel step size=5

Codebook size & Features % used	**20**	**50**	**100**	**200**	**300**	**400**
10%	79.99	81.23	83.04	85.33	84.19	83.99
25%	79.53	81.34	83.72	85.37	84.66	84.23
50%	79.36	81.51	83.16	85.51	85.23	84.49
75%	79.04	81.61	84.53	85.72	85.35	84.59
100%	79.29	81.85	84.64	85.98	85.77	84.86
MAP	79.44	81.50	83.81	**85.58**	85.04	84.43
Std. Dev.	0.3533	0.2412	0.7467	0.2697	0.6188	0.3345
Confidence Interval	79.00-79.88	81.20-81.80	82.89-84.74	85.24-85.91	84.27-85.80	84.01-84.84
Standard Error	0.1580	0.1079	0.3339	0.1206	0.2767	0.1496

As shown in Table 1, the 95% confidence interval for the codebook size of 200 visual words is 85.24 -85.91 at 5% level of significance, indicating more better precision result, with moderate standard error as compared to others confidence intervals.

Table 2. MAP performance of the proposed technique on the Corel-A image archive using pixel step size=10

Codebook size & Features % used	**20**	**50**	**100**	**200**	**300**	**400**
10%	78.03	79.32	81.45	83.69	82.76	82.04
25%	78.64	79.49	81.68	83.39	83.03	82.24
50%	78.22	79.63	81.84	83.88	83.54	82.6
75%	78.4	79.99	81.99	83.75	83.79	82.72
100%	78.13	80.41	82.14	84.15	83.93	82.87
MAP	78.28	79.76	81.82	**83.77**	83.41	82.49
Std. Dev.	0.2411	0.4354	0.2684	0.2773	0.4996	0.3443
Confidence Interval	77.98-78.58	79.22-80.30	81.48-82.15	83.42-84.11	82.78-84.03	82.06-82.92
Standard Error	0.1078	0.1947	0.1200	0.1240	0.2234	0.1540

Table 3. MAP performance of the proposed technique on the Corel-A image archive using pixel step size=15

Codebook size & Features % used	**20**	**50**	**100**	**200**	**300**	**400**
10%	77.36	78.11	80.38	81.81	81.84	81.09
25%	77.49	78.23	80.56	81.97	81.99	81.27
50%	77.61	78.28	80.73	82.15	82.07	81.43
75%	77.72	78.71	80.27	82.36	82.14	81.65
100%	77.95	78.82	80.49	82.63	82.29	81.77
MAP	77.62	78.43	80.52	**82.18**	82.06	81.44
Std. Dev.	0.2254	0.3144	0.1752	0.3227	0.1677	0.2758
Confidence Interval	77.34-77.90	78.03-78.82	80.26-80.70	81.78-82.58	81.85-82.27	81.09-81.78
Standard Error	0.1008	0.1406	0.0783	0.1443	0.0750	0.1233

For pixel step sizes of 10 and 15, Table 2 and Table 3 also shows the 95% confidence interval for the codebook size of 200 visual words at 5% level of significance, indicating more better precision results, with fewer standard errors also strengthen the same as compared to other confidence intervals. The MAP performance of the proposed method using different step sizes and codebook sizes is presented in Figure 5.

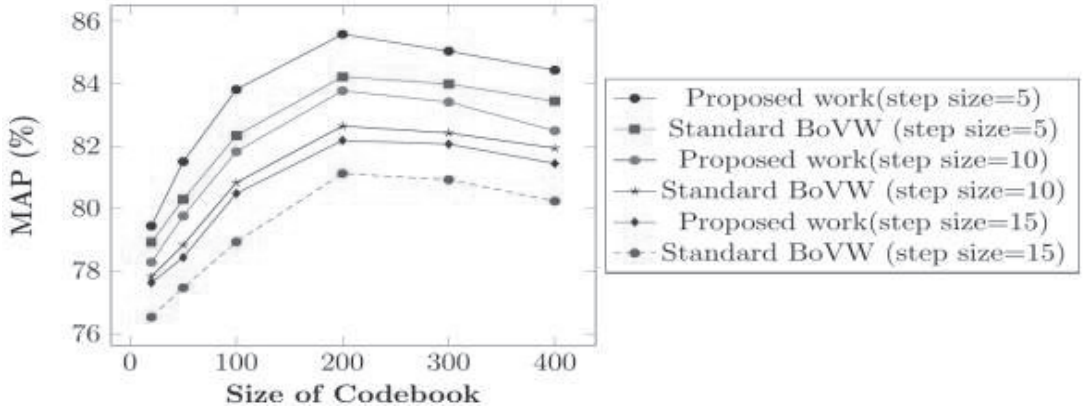

Fig. 5. MAP as a function of codebook size on the Corel-A image dataset.

The experimental results show that the best performance is obtained using the pixel step size of 5 with a codebook size of 200 words and increasing the pixel step size decreases the image retrieval performance and vice versa. In order to present a sustainable performance of the proposed method, the mean precision for top-20 image retrievals is calculated and compared with the state-of-the-art methods for CBIR (Lin *et al*., 2011; Wang *et al*., 2013; Tian *et al*., 2014; Mehmood *et al*., 2016; Zeng *et al*., 2016). Table 4 and Table 5 present the class-wise comparisons of the mean precision and mean recall of the proposed method using SVM and RBF-ANN classifiers (on a codebook size of 200 words) with existing state-of-the-art techniques of CBIR. The first and second top values against each class are mentioned in bold. The MAP performance obtained by using the proposed method on the Corel-A image dataset is presented in Figure 6.

The comparisons evaluated on the Corel-A image dataset shows an overall increase in retrieval accuracy of the proposed method as compared with the existing CBIR approaches. The best MAP performance of 85.58% and 83.78% is obtained by applying SVM and RBF-ANN classifiers, respectively by using the proposed method for a codebook size of 200 words with pixel step size of 5. The top-20 retrieved images for the semantic classes “Buses", "Elephants", and "Beach" are presented in Figure 7, Figure 8, and Figure 9, respectively. At the top of each figure is the query image along with its score, while rest of the images are the

Table 4. Semantic category-wise performance analysis in terms of MAP measure with state-of-the-art CBIR techniques on the Corel-A image archive

Class	Proposed (Step size=5)		Zeng *et al.* (2016)	Mehmood *et al.* (2016)	Tian *et al.* (2014)	Wang *et al.* (2013)	Lin *et al.* (2011)
	SVM	RBF-ANN					
Africa	**73.57**	72.98	72.50	73.03	**74.60**	64	57
Beach	**74.28**	**75.57**	65.20	74.58	37.80	54	58
Buildings	**77.37**	75.13	70.60	**80.24**	70.60	53	43
Buses	94.83	94.81	89.20	**95.84**	**96.70**	94	93
Dinosaurs	97.43	93.61	**100**	97.95	**99**	98	98
Elephants	**90.48**	**88.84**	70.50	87.64	65.90	78	58
Flowers	91.40	85.21	**94.80**	85.13	91.20	71	83
Horses	**93.79**	91.03	91.80	86.29	86.90	**93**	68
Mountains	**81.42**	80.21	72.25	**82.43**	58.50	42	46
Food	**81.23**	**80.42**	78.80	78.96	62.20	50	53

Table 5. Semantic category-wise performance analysis in terms of recall measure with state-of-the-art CBIR techniques on the Corel-A image archive

Class	Proposed (Step size=5) SVM	Proposed (Step size=5) RBF-ANN	Zeng *et al.* (2016)	Mehmood *et al.* (2016)	Tian *et al.* (2014)	Wang *et al.* (2013)	Lin *et al.* (2011)
Africa	**14.71**	14.59	14.5	14.61	**14.92**	12.80	11.40
Beach	**14.85**	**15.11**	13.04	14.92	7.56	10.80	11.60
Buildings	**15.47**	15.02	12.14	**16.05**	14.12	10.60	08.60
Buses	18.96	18.96	17.84	**19.17**	**19.34**	18.80	18.6
Dinosaurs	19.48	18.72	**20**	19.59	**19.80**	19.60	19.6
Elephants	**18.09**	**17.76**	14.10	17.53	13.18	15.60	11.6
Flowers	18.28	17.04	**18.96**	17.03	18.24	14.20	16.6
Horses	**18.75**	18.20	18.36	17.26	17.38	**18.60**	13.60
Mountains	**16.28**	16.04	14.45	**16.49**	11.7	8.4	09.20
Food	**16.24**	**16.08**	15.76	15.79	12.44	10	10.6
Mean Recall	**17.11**	**16.75**	16.11	16.84	14.55	13.94	13.14

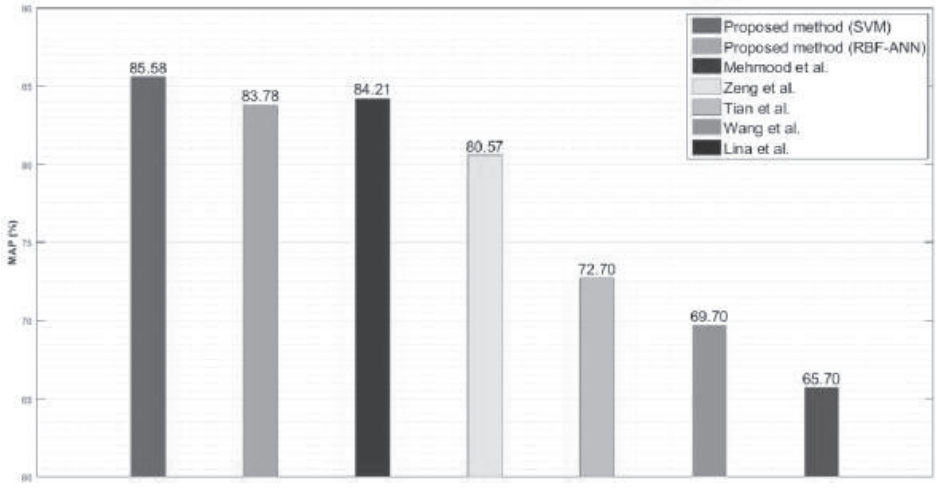

Fig. 6. Comparison of MAP obtained by using proposed method with state-of-the-art CBIR research methods

retrieved images along with scores according to the semantic class of the query image. Similar images are retrieved by applying the Euclidean distance between the score of query image and images placed in the image archive.

Fig. 7. Semantic class "Buses" of the Corel-A image archive shows top-20 image retrievals

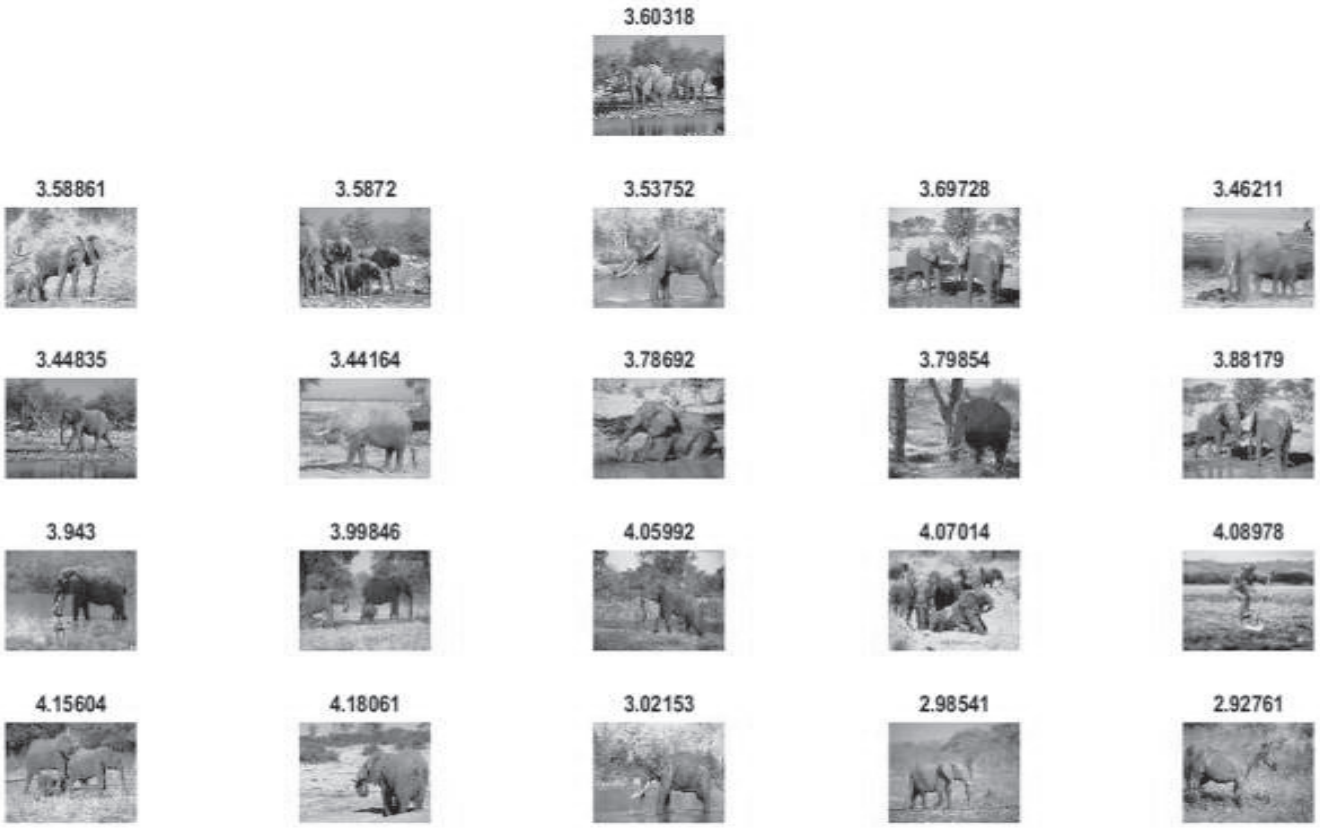

Fig. 8. Semantic class "Elephants" of the Corel-A image archive shows top-20 image retrievals

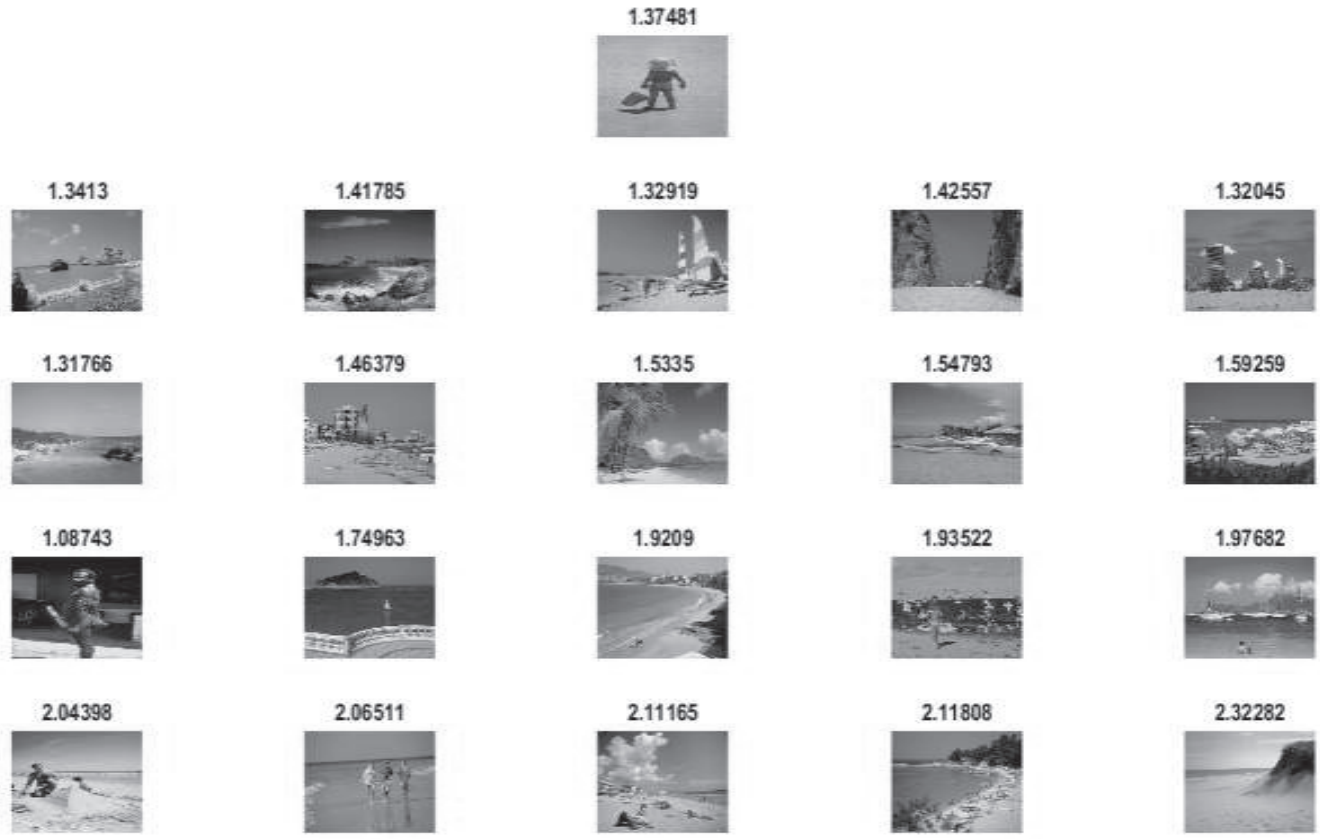

Fig. 9. Semantic class "Beach" of the Corel-A image archive shows top-20 image retrievals

Figure 10 and Figure 11 present the results of automatic image annotation (AIA) based on the top three classification scores. There are 10 semantic classes in the Corel-A image archive. The pre-defined semantic class labels that are used for the 10 classes are: “Restaurants”, “Food”, “Landscape”, “Mountains”, “Grass”, “Horses”, “Garden”, “Flowers”, “Forest”, “Elephants”, “Animals”, “Dinosaurs”, “ Transport”, “Buses”, “Architecture”, “Buildings”, “Sky”, “Beach”, “People”, and “Africa”. Three labels are assigned to every image on the basis of the occurrence of the top three classification scores.

Fig. 10. Semantic class "Horses" of the Corel-A image archive shows result of the AIA based on the occurrence of the top three classification scores

Fig. 11. Semantic class "Mountains" of the Corel-A image archive shows result of the AIA based on the occurrence of the top three classification scores

4.2 Performance measurements on the Ground truth image archive

Ground truth image archive is also publicly available image archive and has been used inthe performanceevaluation of different CBIR techniques (Yildizer *et al*., 2012a; Yildizer *et al*., 2012b; Cardoso *et al*., 2014). There is atotal of 1109 images in Ground truth image archive that are organized into 22 semantic classes. In order to perform a clear comparison with existing state-of-the-art methods of CBIR, we selected 228 images from the 5 different classes (since the same classes of images are used for the performance evaluation

of referred research (Yildizer *et al*., 2012a; Yildizer *et al*., 2012b; Cardoso *et al*., 2014)). The sample images selected for the performance evaluation of the proposed technique are presented in Figure 11, while the names of the semantic classes are presented in Table 6. Different sizes (10, 20, 30, 40, and 50) of the codebook are constructed and MAP performance is reported on these sizes of the codebook. The best performance from the proposed method is obtained by using the codebook size of 40 words and pixel step size of 5 with MAP performance of 87.57% by applying SVM classifier, while 85.21% by applying RBF-ANN classifier. The class-wise mean precision that is obtained using the proposed method is presented in Table 6, while MAP performance is graphically presented in Figure 12.

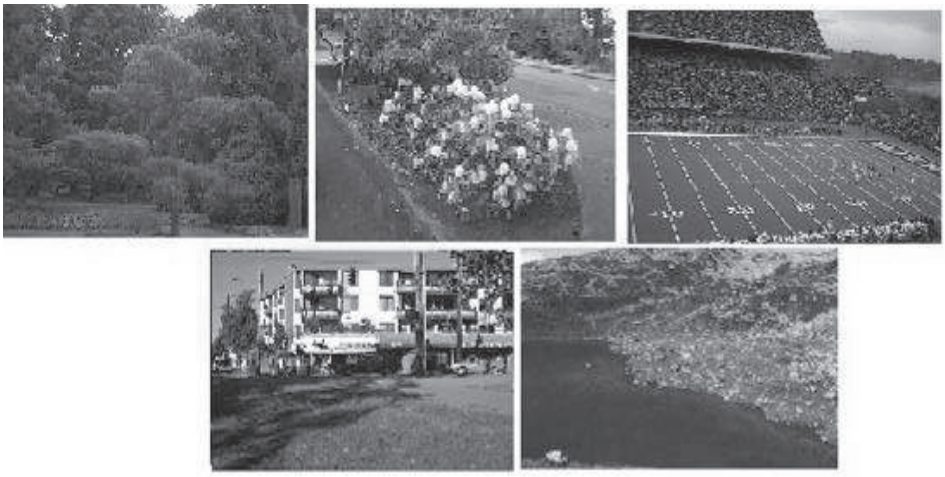

Fig. 11. Sample of semantic classes from the Ground-truth image archive

Table 6. Semantic category-wise performance analysis in terms of MAP measure with state-of-the-art CBIR techniques on the Ground truth image archive

Class	Proposed Method (Step size=5)		Cardoso *et al.* (2014)	Yildizer *et al.* (2012a)
	SVM	RBF- ANN		
Abrogreens	**76.74**	74.14	**80**	66.66
Cherries	**78.18**	75.38	**80**	50
Football	**97.08**	94.26	**100**	75
Greenlake	**90.04**	**89.11**	80	50
Swiss Mountain	**95.94**	**93.18**	66.67	50

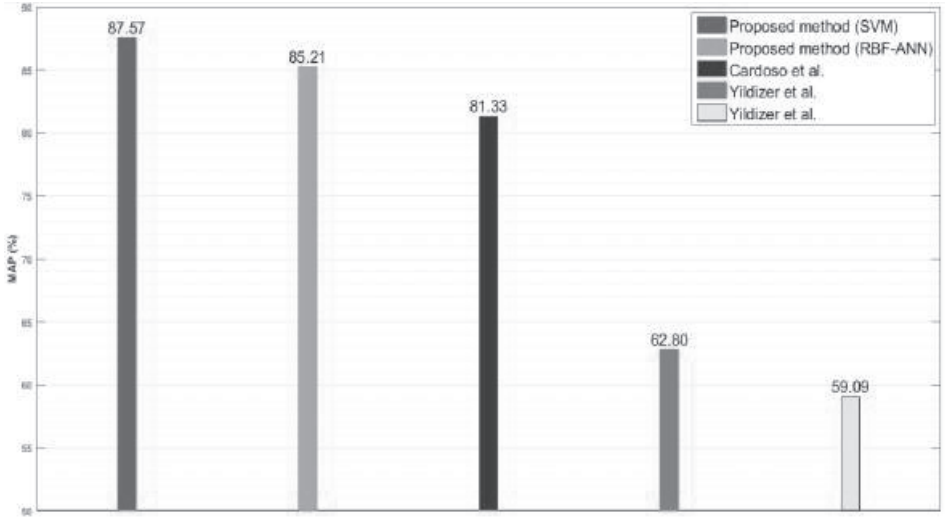

Fig. 12. Comparison of MAP performance obtained by using the proposed technique with state-of-the-art CBIR techniques on the Ground truth image archive

Experimental results and comparisons conducted on the Ground truth image dataset show the robustness of the proposed research work based on the rectangular spatial histograms of visual words approach. The MAP performance obtained from theproposed research method by using SVM outperforms the existing state-of-the-art research methods (Yildizer *et al*., 2012a; Yildizer *et al*., 2012b; Cardoso *et al*., 2014).

4.3. Computational complexity

The computational cost of the proposed algorithm is calculated on desktop PC with following specifications; Intel Pentium (R) 3.0 GHz microprocessor with 2 GB RAM by using Windows 7 operating system. The proposed algorithm is implemented in MATLAB 2013 and codebook is constructed offline and tested at run time. The average CPU time required for feature extraction by using the proposed technique of image resolution 256 x 384 of the Corel-A image archive is shown in Table 7. The required computational complexity from the feature extraction to image retrieval is shown in Table 8 for the Corel-A image archive.

Table 7. Comparison of average CPU time (in seconds) required to extract feature descriptor

Proposed method	BoVW method	Dubey *et al.* (2015) SEH CDH RSHD	Tian *et al.* (2014)
0.1021	0.0991	0.186 1.709 0.375	5.6

Table 8. Comparison of the computational cost of the proposed method (time in seconds) with state-of-the-art CBIR methods

Retrieved images	Proposed method		LGH method-SVM Mehmood *et al.* (2016)	Spatial level 2 method-SVM Ali *et al.* (2016)	WATH method-SVM Mehmood *et al.* (2017)
	SVM	RBF- ANN			
Top-5	0.2614	0.2731	0.2872	0.3584	0.3726
Top-10	0.3751	0.3811	0.3972	0.4811	0.5178
Top-15	0. 6074	0.6123	0.6470	0.6799	0.7050
Top-20	0.7365	0.7599	0.7837	0.8491	0.8882
Top-25	0.8417	0.8673	0.9184	1.0121	1.0599

5. Conclusion and future directions

In this paper, we proposed a novel image representation based on the rectangular spatial histograms of visual words that add the spatial information to the inverted index of BoV Wmodel. The standard BoVW based image representation is not sufficient for the efficient image retrieval, as images of different classes with close visual appearance result in the closeness of visual words in the histogram and it decreases the image retrieval performance. Construction of two separate histograms of visual words for two rectangular regions of each image is a possible solution for the reduction of semantic gap and addition of image spatial attributes to the image retrieval. The proposed research work is evaluated on two image datasets. The rectangular spatial histograms of visual words based approach is found robust and outperform other referred techniques including the standard BoVW based image representation. In future, we will extend proposed research by using adopted rectangular approach (that is extracting dense features over two scales of the color image by division of the color image into two rectangular regions prior to dense features extraction) with deep neural networks for a large scale image retrieval (ImageNet or Flicker). We also plan to replace BoVW model with vector of locally aggregated descriptors (VLAD) or Fisher kernel framework to evaluate the proposed research for large-scale image retrieval.

References

Ali, N., K. B. Bajwa, R. Sablatnig &Z. Mehmood (2016). Image retrieval by addition of spatial information based on histograms of triangular regions. Computers & Electrical Engineering, **54**:539-550.

Bosch, A., A. Zisserman &X. Munoz (2007). Image classification using random forests and ferns. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE.

Cao, Y., C. Wang, Z. Li, L. Zhang &L. Zhang (2010). Spatial-bag-of-features. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE.

Cardoso, D. N. M., D. J. Muller, F. Alexandre, L. A. P. Neves, P. M. G. Trevisani &G. A. Giraldi (2014). Iterative technique for content-based image retrieval using multiple SVM Ensembles." J. Clerk Maxwell, A Treatise on Electricity and Magnetism, **2**:68-73.

Datta, R., D. Joshi, J. Li &J. Z. Wang (2008). Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR),**40**(2):5.

Dubey, S. R., S. K. Singh &R. K. Singh (2015). Rotation and scale invariant hybrid image descriptor and retrieval. Computers & Electrical Engineering,**46**:288-302.

Hassner, T., V. Mayzels &L. Zelnik-Manor (2012). On sifts and their scales. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE.

Haykin, S. &N. Network (2004). A comprehensive foundation. Neural Networks, 2(2004).

Heikkilä, M., M. Pietikäinen &C. Schmid (2009). Description of interest regions with local binary patterns. Pattern recognition,**42**(3):425-436.

Jhanwar, N., S. Chaudhuri, G. Seetharaman &B. Zavidovique (2004). Content based image retrieval using motif cooccurrence matrix. Image and Vision Computing,**22**(14): 1211-1220.

Jiji, G. W. &P. J. DuraiRaj (2015). Content-based image retrieval techniques for the analysis of dermatological lesions using particle swarm optimization technique. Applied Soft Computing,30:650-662.

Kadir, T., A. Zisserman &M. Brady (2004). An affine invariant salient region detector. Computer Vision-ECCV 2004, Springer:228-241.

Khan, R., C. Barat, D. Muselet &C. Ducottet (2012). Spatial orientations of visual word pairs to improve bag-of-visual-words model. Proceedings of the British Machine Vision Conference, BMVA Press.

Lazebnik, S., C. Schmid &J. Ponce (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE.

Lin, C.-H., D.-C. Huang, Y.-K. Chan, K.-H. Chen &Y.-J. Chang (2011). Fast color-spatial feature based image retrieval methods. Expert Systems with Applications,**38**(9):11412-11420.

Liu, G.-H., Z.-Y. Li, L. Zhang &Y. Xu (2011). Image retrieval based on micro-structure descriptor. Pattern Recognition,**44**(9):2123-2133.

Liu, G.-H., L. Zhang, Y.-K. Hou, Z.-Y. Li &J.-Y. Yang (2010). Image retrieval based on multi-texton histogram. Pattern Recognition,**43**(7):2380-2389.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision,**60**(2):91110-.

Mahmood, T., T. Nawaz, R. Ashraf, M. Shah, Z. Khan, A. Irtaza & Z. Mehmood (2015). A survey on block based copy move image forgery detection techniques. Emerging Technologies (ICET), 2015 International Conference on, IEEE:1-6.

Mahmood, T., A. Irtaza, Z. Mehmood & M. T. Mahmood (2017). Copy–move forgery detection through stationary wavelets and local binary pattern variance for forensic analysis in digital images. Forensic Science International **279**:8-21.

Mahmood, T., Z. Mehmood, M. Shah & Z. Khan (2017). An efficient forensic technique for exposing region duplication forgery in digital images. Applied Intelligence:1-11.

Mehmood, Z., S. M. Anwar, N. Ali, H. A. Habib &M. Rashid (2016). A novel image retrieval based on a combination of local and global histograms of visual words. Mathematical Problems in Engineering, 2016.

Mehmood, Z., T. Mahmood &M. A. Javid (2017). Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Applied Intelligence,**8**:1-16.

Nosaka, R., Y. Ohkawa &K. Fukui (2011). Feature extraction based on co-occurrence of adjacent local binary patterns. Advances in image and video technology, Springer: 82-91.

Philbin, J., O. Chum, M. Isard, J. Sivic &A. Zisserman (2007). Object retrieval with large vocabularies and fast spatial matching. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE.

Rui, Y., T. S. Huang &S.-F. Chang (1999). Image retrieval: Current techniques, promising directions, and open issues. Journal of visual communication and image representation,**10**(1):39-62.

Safar, M. (2009). Image approximation to efficiently support direction queries. Kuwait J. Sci. Eng, **36**(1B):147-166.

Shawe-Taylor, J. &N. Cristianini (2004). Kernel methods for pattern analysis, Cambridge university press.

Sivic, J. &A. Zisserman (2003). Video Google: A text retrieval approach to object matching in videos. Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE.

Tian, X., L. Jiao, X. Liu &X. Zhang (2014). Feature integration of EODH and Color-SIFT: Application to image retrieval based on codebook. Signal Processing: Image Communication, **29**(4):530-545.

Tomašev, N. &D. Mladenić (2015). Image hub explorer: Evaluating representations and metrics for content-based image retrieval and object recognition. Multimedia Tools and Applications,**74**(24): 11653-11682.

Tousch, A.-M., S. Herbin &J.-Y. Audibert (2012). Semantic hierarchies for image annotation: A survey. Pattern Recognition,**45**(1):333-345.

Tuytelaars, T. &L. Van Gool (1999). Content-based image retrieval based on local affinely invariant regions. Visual Information and Information Systems, Springer.

Tuytelaars, T. &L. J. Van Gool (2000). Wide baseline stereo matching based on local, affinely invariant regions. BMVC.

Ullah, A. &B. Baharudin (2016). Pattern and semantic analysis to improve unsupervised techniques for opinion target identification. Kuwait Journal of Science,**43**(1).

Vedaldi, A. &B. Fulkerson (2010). VLFeat: An open and portable library of computer vision algorithms. Proceedings of the 18th ACM international conference on Multimedia, ACM.

Vedaldi, A. &A. Zisserman (2012). Sparse kernel approximations for efficient classification and detection. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE.

Wang, C., B. Zhang, Z. Qin &J. Xiong (2013). Spatial weighting for bag-of-features based image retrieval. Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer:91-100.

Xie, J., L. Zhang, J. You &S. Shiu (2015). Effective texture classification by texton encoding induced statistical features. Pattern Recognition,**48**(2):447-457.

Yildizer, E., A. M. Balci, M. Hassan &R. Alhajj (2012a). Efficient content-based image retrieval using multiple support vector machines ensemble. Expert Systems with Applications,**39**(3):2385-2396.

Yildizer, E., A. M. Balci, T. N. Jarada &R. Alhajj (2012b). Integrating wavelets with clustering and indexing for effective content-based image retrieval. Knowledge-Based Systems,**31**:55-66.

Zeng, S., R. Huang, H. Wang &Z. Kang (2016). Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models. Neurocomputing,**171**:673-684.

Zhang, D., M. M. Islam &G. Lu (2012). A review on automatic image annotation techniques. Pattern Recognition,**45**(1):346-362.

Submitted: 10/01/2016
Revised : 12/06/2016
Accepted : 13/06/2016

طريقة جديدة لاسترجاع الصور
تعتمد على مدرجات مكانية تكرارية مستطيلة للكلمات المرئية

زاهد محمود[1،2،*]، سيد أنور[1]، محمد ألطف[3]

[1] قسم هندسة البرمجيات، جامعة الهندسة والتقنيات، تاكسيلا 47050، باكستان

[2] قسم هندسة الحاسوب، جامعة الهندسة والتقنيات، تاكسيلا 47050، باكستان

[3] قسم الرياضيات، جامعة الهندسة والتقنيات، تاكسيلا 47050، باكستان

*zahid.mehmood@uettaxila.edu.pk

خـلاصـة

تُعطي طريقة استرجاع الصور اعتماداً على المحتوى (CBIR) حلاً للبحث في صور تتشابه مع الصورة محل الاهتمام. في السنوات الأخيرة، اكتسبت طريقة شنطة الكلمات المرئية (BoVW) أهمية وأدت إلى تحسن في أداء CBIR. في نموذج BoVW يتم تمثيل الصورة على شكل مدرج تكراري غير مرتب للكلمات المرئية وذلك بإهمال التفاصيل المكانية للصورة. التفاصيل المكانية تحتفظ بمعلومات هامة قد تؤدي إلى تدعيم درجة الدقة لاسترجاع الصور. في هذا البحث، نقدم طريقة مبتكرة لتمثيل الصور تعتمد على بناء مدرجات تكرارية على منطقتين مستطيلتين للصورة، تقسيم الصورة إلى منطقتين مستطيلتين عند بناء المدرجات التكرارية يؤدي إلى إضافة البيانات المكانية لنموذج BoVW. الطريقة المقترحة تستخدم كلمات مرئية مختلفة للمنطقة المستطيلة العلوية والمنطقة المستطيلة السفلية للصورة. وضح التحليل التجريبي لبيانات صورتين أن الطريقة المقترحة صالحة للاستخدام.