Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine

Article *in* Applied Intelligence · January 2018

DOI: 10.1007/s10489-017-0957-5

CITATIONS
24

READS
616

3 authors:

Zahid Mehmood
University of Engineering and Technology, Taxila
58 PUBLICATIONS **474** CITATIONS
SEE PROFILE

Toqeer Mahmood
University of Engineering and Technology, Taxila
35 PUBLICATIONS **350** CITATIONS
SEE PROFILE

Dr Arshad Javid
University of Engineering and Technology, Taxila
33 PUBLICATIONS **86** CITATIONS
SEE PROFILE

Appl Intell
DOI 10.1007/s10489-017-0957-5

Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine

Zahid Mehmood[1] · Toqeer Mahmood[2] · Muhammad Arshad Javid[3]

Abstract In recent years, the rapid growth of multimedia content makes content-based image retrieval (CBIR) a challenging research problem. The content-based attributes of the image are associated with the position of objects and regions within the image. The addition of image content-based attributes to image retrieval enhances its performance. In the last few years, the bag-of-visual-words (BoVW) based image representation model gained attention and significantly improved the efficiency and effectiveness of CBIR. In BoVW-based image representation model, an image is represented as an order-less histogram of visual words by ignoring the spatial attributes. In this paper, we present a novel image representation based on the weighted average of triangular histograms (WATH) of visual words. The proposed approach adds the image spatial contents to the inverted index of the BoVW model, reduces overfitting problem on larger sizes of the dictionary and semantic gap issues between high-level image semantic and low-level image features. The qualitative and quantitative analysis conducted on three image benchmarks demonstrates the effectiveness of the proposed approach based on WATH.

Keywords Content-based image retrieval · Bag-of-visual-words · Support vector machine · Dense SIFT · Image classification

1 Introduction

CBIR imparts an effective way to retrieve similar images on the basis of visual contents of a query image. The exponential growth in the number of image databases makes image retrieval an active research area [1]. In CBIR, we want to obtain those images that are in a semantic association according to the response of a given query image [1]. In CBIR, images are represented in the form of feature vectors that consist of low-level visual features that are based on shape, color, and texture [1, 2]. The comparison between low-level visual features of the images stored in an image database and a given query image evaluate the outcomes of retrieved images [1, 3]. The similarity in visual appearance of the images associated with different semantic categories outcome in the similarity of feature vectors that reduces the effectiveness of CBIR [1]. The real world applications of CBIR include photograph and video search databases (e.g. ImageNet, Flicker, Google image search, and YouTube), retail catalogs, geographical information and remote sensing systems, art collections (e.g. fine arts museum of San Francisco), medical image databases (e.g. the visible human, ultrasound, MRI, and CT), scientific image databases (e.g. earth sciences), general image collections for licensing (e.g. Getty images, and Corbis), nudity detection filters, and the world wide web [4, 5].

✉ Zahid Mehmood
zahid.mehmood@uettaxila.edu.pk

Toqeer Mahmood
toqeer.mahmood@yahoo.com

Muhammad Arshad Javid
arshad.javid@uettaxila.edu.pk

[1] Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan

[2] Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan

[3] Department of Basic Sciences, University of Engineering and Technology, Taxila 47050, Pakistan

Published online: 21 June 2017

In the BoVW-based image representation [6], feature descriptors are used for the extraction of image features. The high-dimensional feature space is quantized by applying a quantization algorithm such as *k*-means to construct the visual vocabulary. An image is represented as an order-less histogram of visual words by ignoring the spatial context of 2D image space. In BoVW-based image representation, histograms of visual words are the feature vectors of images. The spatial information provides discriminating details in classification-based problems and performance of the BoVW model suffers due to the order-less representation of the image [7].

The spatial contents from different semantic regions can be extracted by dividing an image into triangular regions (level-1 and level-2 triangles) [8]. Figure 1 represents the spatial triangular relationship between different images from the Corel image database. The objects or regions of interest are likely to be located within different sub-blocks of triangles. This division is a possible solution for the addition of spatial information and reduction of the semantic gap between high-level image semantic, and low-level feature of the image.

The commonly used approaches applied for the extraction of spatial information are geometric coding [10], random sample consensus (RANSAC) [7] and visual word co-occurrence [11]. These approaches are reported to be computationally expensive and introduce the problem of overfitting on the dictionary of larger sizes [12]. The approach of [13] is to extract the spatial information from different sub-regions of the image by computing histograms of visual words from each of the divided sub-region. This approach divides an image into several rectangular grids and constructs histograms of visual words from each region of the grid. The division of an image into sub-block for extraction of spatial information is reported effective for content-based image retrieval [8, 14]. Keeping in view the effective performance of [13] and spatial context available in different sub-regions, the proposed image representation is extracting the histograms of visual words over the four triangular regions on the basis of weighted average. In our previous work [8], we extracted the histograms of visual words from two and four triangular regions (level-1 and level-2 triangles). In-case of level-1 triangles, for a vocabulary of size N, an image is represented as $2\times N$ visual words. In-case of level-2 triangles, for a vocabulary of size N, an image is represented as $4\times N$ visual words. The proposed image representation in this research article is different from the previous work as we are applying the weighted average on histograms of triangular regions and representing an image in a size that is equal to the size of the dictionary. In-addition to this, dense features are extracted at multi-scales and different step sizes or pixel strides to determine the best performance of the proposed approach, and the problem of overfitting on the dictionary of larger sizes is also addressed in this approach. The proposed approach also outperform as compared to our existing proposed approach [8]. These are the main contributions of this research article:

1. The addition of spatial information to the inverted index of BoVW-based image representation.
2. The image representation on basis of the weighted average of histograms of triangular regions that resolve the problem of overfitting on a dictionary of larger sizes.
3. Reduction of the semantic gap between high-level image concepts and low-level features of the image.
4. Automatic image annotation based on the classification scores.

The rest of the paper is structured as follows: Section 2 reviews the related work. Section 3 presents proposed methodology. Section 4 presents evaluation measures, experimental parameters, and results conducted on three image databases (i.e. Corel-A/1000, Corel-B/1500, and Scene-15), and computational complexity of the proposed approach based on WATH, while Section 5 analyzes the

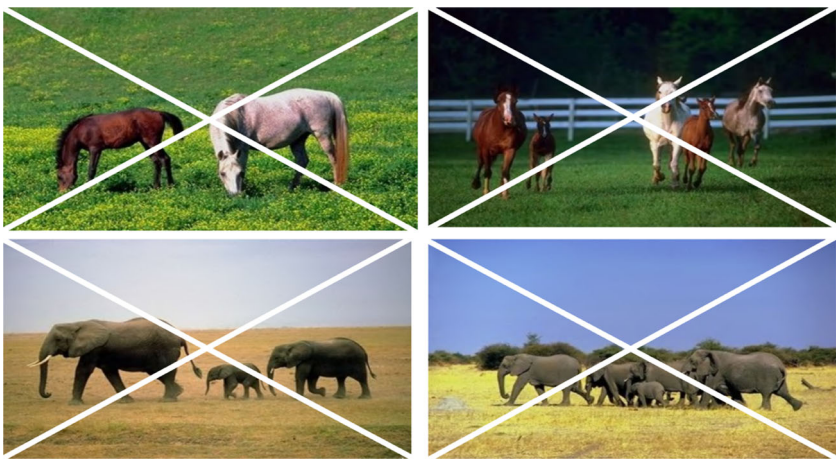

Fig. 1 The content-based spatial triangular relationship between the images of the Corel image database [8, 9]

proposed approach based on WATH as well as present points towards the future directions.

2 Related work

IBM launched the first system for CBIR [3] and later, different image retrieval techniques were introduced based on spatial layout, texture, color, and shape [3, 15–18]. In CBIR, the semantic gap between low-level image representation, and high-level image visual is a common research problem [3]. The interest point detectors like scale-invariant feature transform (SIFT) [19], speeded-up robust features (SURF) [20], histograms of oriented gradients (HOG) [21], maximally stable extremal regions (MSER) [22], and binary robust invariant scalable keypoints (BRISK) [23] were applied in various content-based image matching applications [24–26].

Wang et al. [27] proposed spatial weighting BoF (SWBoF) approach for the extraction of spatial information from different blocks of the image. Adjacent block distance, local entropy, and variance were used for the extraction of spatial information. The spatial weighting is achieved by weighting the corresponding visual words with the help of texture information. Tian et al. [28] introduced a scale and rotation-invariant edge orientation difference histogram (EODH) descriptor. The main orientation of each pixel was extracted by applying a steerable filter, and weighted average scheme was applied for the construction of codebook consisting of EODH and Color-SIFT. Yu et al. [29] proposed feature integration of low-level features for an effective CBIR. For clustering, the *k*-means algorithm was applied, patch-based and image-based integration of mid-level features were proposed. As reported by experimental results of [29], the image-based integration of SIFT-LBP features improves the performance of CBIR as compared to the state-of-the-art techniques.

Zeng et al. [30] introduced an image representation that depends on the generalized histogram of quantized colors. Gaussian mixture models (GMMs) was applied for quantization, and expectation-maximization (EM) algorithm was used for training. Bayesian information criterion (BIC) was applied for the determination of quantized color bins. Images were retrieved on the basis of similarity measure between the respective spatiograms. Walia et al. [31] proposed a fusion approach for color-based image retrieval. The texture and color features were extracted by applying color difference histogram (CDH) and angular radial transform (ART). A modification in CDH algorithm was proposed in order to produce more effective results. Yuan et al. [32] proposed image retrieval approach using features integration SIFT and LBP. The combination of SIFT-LBP was selected to enhance the performance in-case of noisy background and ambiguous objects. Dubey et al. [33] proposed a rotation and scale-invariant hybrid image descriptor for an effective image retrieval. The color features were extracted by quantizing the RGB color space, while the texture was extracted by structuring the patterns that were generated from locally structured elements. Color and texture features were integrated to construct the inherent rotation and scale-invariant hybrid image descriptor (RSHD). Wan et al. [34] utilized deep learning approach based on convolutional neural networks (CNN) for large-scale image retrieval and reported effective performance. According to the experimental results, CNN framework based extracted features produce better results as compared to the traditional feature extraction approaches [34].

Mehmood et al. [35] proposed an effective image representation that was based on local and global histograms (LGH) of visual words, in order to incorporate spatial information into the BoVW model. Guang-Hai et al. [36] emphasize on the edge-based image representation for image retrieval named as multi-texton histogram (MTH). MTH method has the advantages of histogram and co-occurrence matrix. Where the histogram was built using the characteristic of co-occurrence matrix. Liu et al. [37] proposed texture classification method using the attributes of compressed sensing and sparse representation. First of all, from small image patches, random features were computed, and then these features were embedded into the BoVW model for image classification. By using the proposed approach, they have reduced the computational cost and also reduced the storage complexity. They claimed that only using the one-third dimensions of the image patch is enough to represent the salient information of the image.

In order to improve the performance of CBIR, commonly addressed issues in aforementioned CBIR techniques were incorporation of spatial information [7, 8, 10, 11, 13, 27, 35], reduction of the semantic gap between high-level semantic and low-level image features [13, 29, 35], computational complexity [8, 37, 38], and rotation as well as scale-invariant issues [28, 33]. The drawbacks of CBIR techniques [7, 8, 10, 11, 13, 35] were a problem of overfitting on a dictionary of larger sizes as well as these techniques were computationally expensive [12]. The proposed technique improved performance of CBIR, resolve the problem of overfitting on a dictionary of larger sizes, incorporate spatial information, resolve issues of the semantic gap, and automatic image annotation.

3 Proposed methodology

The basic problem of BoVW-based image representation is the lack of spatial information [7]. The visual words are represented in a histogram without considering their locations

in the 2D image space. The spatial information provides discriminating details in recognition problems [7]. The process of computing $128 \times N$ dimensional SIFT feature vector on a regular grid across an image is known as dense SIFT features. After that, a clustering technique like k means is applied, in which nearest features (dense SIFT feature) are grouped or clustered together, and the center of each cluster, called the visual word, represents a patch of the image. The combination of visual words form a dictionary. The procedure of dividing an image into four triangular regions, and constructing a histogram of visual words from each triangular region is known as a triangular histogram. The complete aforementioned process as shown in Fig. 2 for image representation is known as BoVW model. The block diagram of proposed approach is shown in Fig. 3. The detailed working procedure of proposed approach based on WATH is given below:

1. The BoVW representation starts from an image G, represented as

$$G = (s_{j,k}) \tag{1}$$

   Where $(s_{j,k})$ is the pixel at position (j,k).
2. Dense SIFT features are computed from an image G, represented as

$$G = \{i_{d1}, i_{d2}, i_{d3}, i_{d4}, ..., i_{dn}\} \tag{2}$$

   Where i_{d1} to i_{dn} are image descriptors.
3. The hard clustering is applied to the dense features extracted from the whole image through k-means++ algorithm [40] in order to avoid random selection of initial centroid in the clustering to construct a dictionary consisting of m visual words, represented by C_w:

$$C_w = \{w_1, w_2, w_3, w_4, w_5, ..., w_m\} \tag{3}$$

   Where w_1 to w_m are visual words that are used to construct dictionary C_w.
4. For the construction of a histogram from the left triangular region of an image, mapping of each visual word is carried on using the left triangular region of an image. Similar procedure is followed for the top, right, and bottom triangular regions of each image. The nearest words are assigned to the quantized descriptors according to the following equation:

$$w(d_m) = \underset{w \,\varepsilon\, C_w}{\operatorname{argmin}}\ Dist(w, d_m) \tag{4}$$

   Where $w(d_m)$ is representing the visual word assigned to the m^{th} descriptor d_m, while $Dist(w,d_m)$ is the distance between the descriptor d_m and the visual word w. Each image is represented as a collection of four triangular regions and each triangular region is represented by the visual words.
5. Four histograms of m visual words are computed from a single image. The visual words of each triangular region are mapped to the corresponding triangular region of the image that is selected using the equations 5–8. The three points of each triangular region are selected from each image using the following equations, respectively:

$$T_{l1} = G(1,1),\ T_{l2} = G(h,1),\ T_{l3} = G(h/2, w/2) \tag{5}$$
$$T_{r1} = G(1,w),\ T_{r2} = G(h,w),\ T_{r3} = G(h/2, w/2) \tag{6}$$
$$T_{t1} = G(1,1),\ T_{t2} = G(1,w),\ T_{r3} = G(h/2, w/2) \tag{7}$$
$$T_{b1} = G(h,1),\ T_{b2} = G(h,w),\ T_{b3} = G(h/2, w/2) \tag{8}$$

   Where T_{lk}, T_{rk}, T_{tk}, and T_{bk} represents the points of left, right, top and bottom triangular regions of each image, respectively and k varies from 1 to 3 for each triangular region. w and h represent the width and height of the image, respectively.
6. Histograms are computed using the visual words of each triangular region of the image and each histogram is multiplied by weight W. The value of weight W is selected according to the experimental details mentioned in Section 4. The resultant four weighted histograms are concatenated and the resultant information is incorporated into the inverted index of the BoVW-based image representation. Consider m as the number of visual words of the codebook C_w. Let S_n be the set of the descriptors that are mapped to the visual word w_m then the n_{th} bin of the histogram of visual words h_n, is the cardinality of the set S_n can be represented as

$$h_n = Card(S_n)$$

   and

$$S_n = \{i_{dn}, dn \in (1, \ldots, m) \,|\, w(i_{dn}) = w_m\} \tag{9}$$

3.1 Image classification

Support vector machine (SVM) is a state-of-the-art supervised learning classification method [16]. The linear SVM separates the two classes by using a hyperplane. The kernel method [41] is applied using SVM to calculate the dot

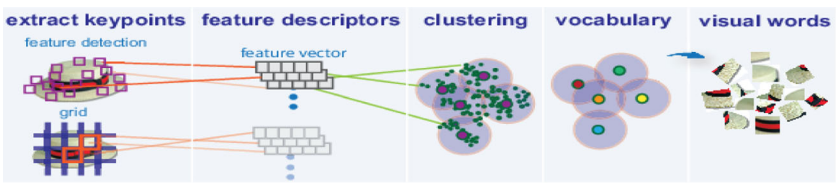

Fig. 2 The methodology of BoVW model for image representation [39]

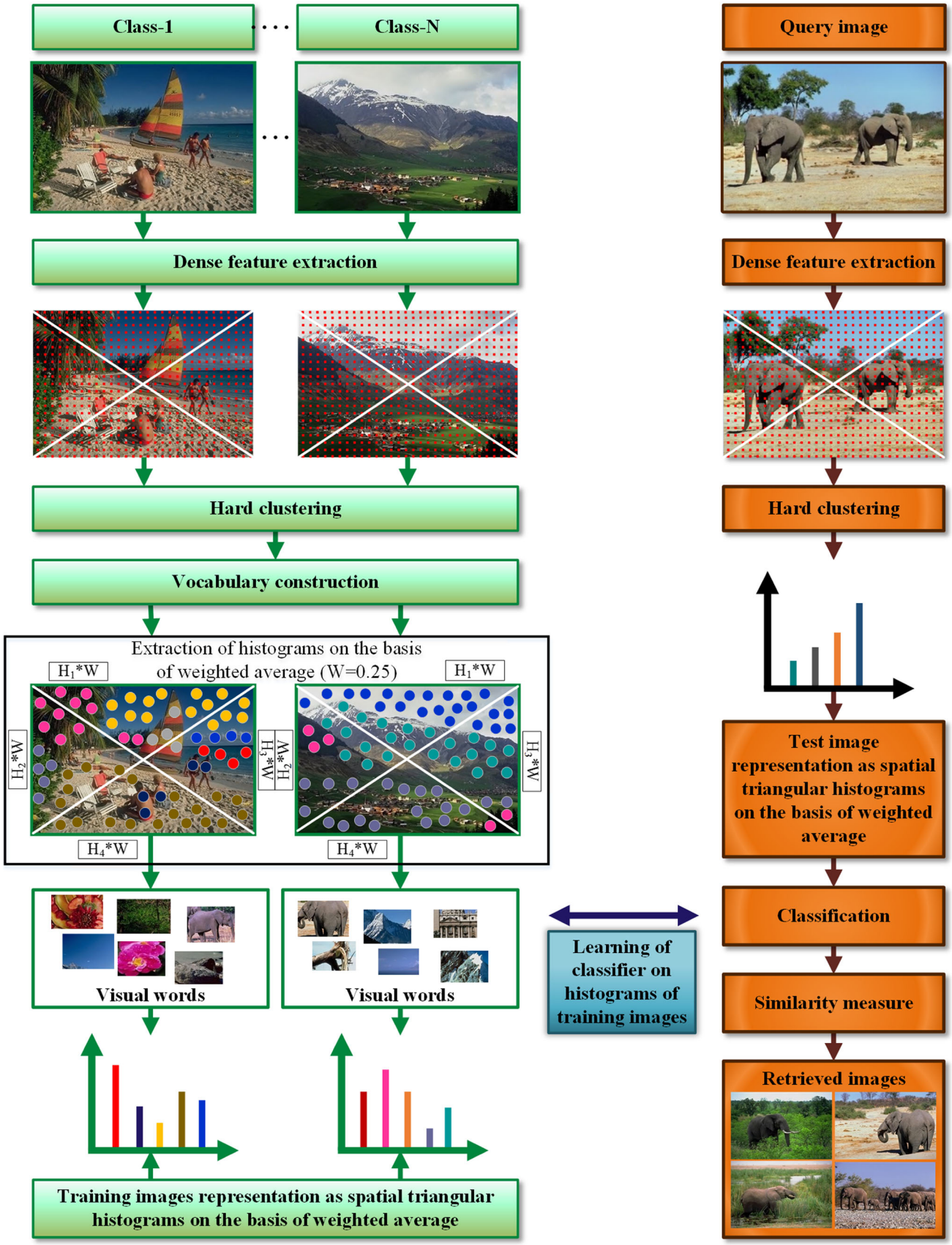

Fig. 3 The proposed approach of CBIR based on the WATH

product in the high dimensional feature map and enables it to produce non-linear decision regions. The histograms of visual words constructed on the basis of the weighted average of triangular regions are concatenated and normalized by applying $l2$ normalization. In order to normalize the weighted average based triangular histograms of visual words, we have applied the Hellinger kernel [42] function of the SVM on them. Due to the low computational complexity of SVM Hellinger kernel, it is chosen for classification. The SVM Hellinger kernel computes the feature space explicitly, instead of calculating the kernel values and classifier still remains linear. For the SVM Hellinger kernel, regularization parameter C is evaluated by applying 10-fold cross validation function on the set of training images.

4 Experimental parameters and results

The proposed technique of WATH is examined by utilizing two subsets (Corel-A and Corel-B) of the Corel image database [9] and Scene-15 image database [13] and the comparison is performed with state-of-the-art techniques of CBIR [8, 27, 30, 31, 35]. For all experiments, we have divided 70% images for training and remaining 30% images for testing. Keeping in view the unsupervised nature of clustering using k-means, every experiment is repeated 10 times and during every run, images are automatically randomly chosen for testing and training processes. The set of chosen images for training are used for the construction of the dictionary (also known as codebook or visual vocabulary) and image retrieval performance is reported using test database. The images are retrieved by calculating the closeness of the classifier score values. The classifier output label determines the class of image, while the Euclidean distance is applied between scores of the images stored in an image database and score of a given query image in order to determine the output of retrieved images. The detail about the experimental parameters, performance evaluation, and results are mentioned in the following sub-sections.

1. **The value of weight W:** The value of weight W is used to calculate the weight of each triangular region to build a histogram of visual words. In-order to assign equal weight distribution, we selected the value of W to 0.25 to compute a histogram of visual words from each triangular region. The value of W is selected to 0.25 because we are constructing four triangular histograms from each image, which result in assigning a weight of 0.25 ($1/4 = 0.25$) to each histogram. According to [35, 39, 43], increasing the size of the dictionary at some certain level increases the performance of image retrieval, which also results in a reduction of the semantic gap, while larger size of dictionary tends to overfit. In our previous proposed approach [8], we get the best performance (i.e. MAP of 86.25%) of image retrieval on a dictionary size of 200 visual words (i.e. 50 visual words per triangular region $\times$ 4 triangular regions $=$ 200 visual words for the dictionary) on the Corel-A image database. Larger sizes of the dictionary in [8, 35] tend to overfit and also result in a decrease of the image retrieval performance as well as an increase in the semantic gap between high-level image semantic and low-level image features. In the proposed WATH based approach, we get best performance (i.e. MAP of 87.85%) of image retrieval on larger size of dictionary (i.e. dictionary size of 600 visual words) due to assigning weight of 0.25 to each triangular histogram of the image on the Corel-A image database, which also result in the reduction of semantic gap due to increase in the performance of image retrieval. The performance of proposed WATH based approach is also analyzed by assigning different weights instead of assigning the same weight to each triangular histogram of visual words, which makes the proposed approach biased and inconsistent, and the performance of image retrieval decreases on the reported CBIR databases in this research article.
2. **Dictionary size and percentage of features for the dictionary construction:** The size of dictionary have an impact on the outcomes of CBIR [39, 43]. An increase in the size of the dictionary, increases the performance of CBIR, while a larger size dictionary tends to over-fit. We constructed different sizes of dictionary from the training database in order to evaluate the best performance of the proposed approach. To reduce the computational cost of the training process, we have randomly selected 50% of features from each image from the training database for the construction of dictionary.
3. **The scale of dense features:** The scale of extracted dense features also affects the performance of content-based image matching [39, 43]. Extraction of dense features at multiple scales increases the performance of image retrieval. We have extracted dense-SIFT descriptors at four different scales that are 4, 6, 8, and 10.
4. **Dense pixel stride or step size:** The dense pixel stride is utilized in order to control the steps of the dense grid. A larger pixel stride makes the grid coarser, while a smaller pixel stride makes the grid finer. We have evaluated proposed approach using four different pixel strides that are 5, 10, 15, and 20. For dense pixel strides of 5, 10, 15, and 20, SIFT descriptor is calculated after every 5^{th}, 10^{th}, 15^{th}, and 20^{th} pixel, respectively as shown in Fig. 4.

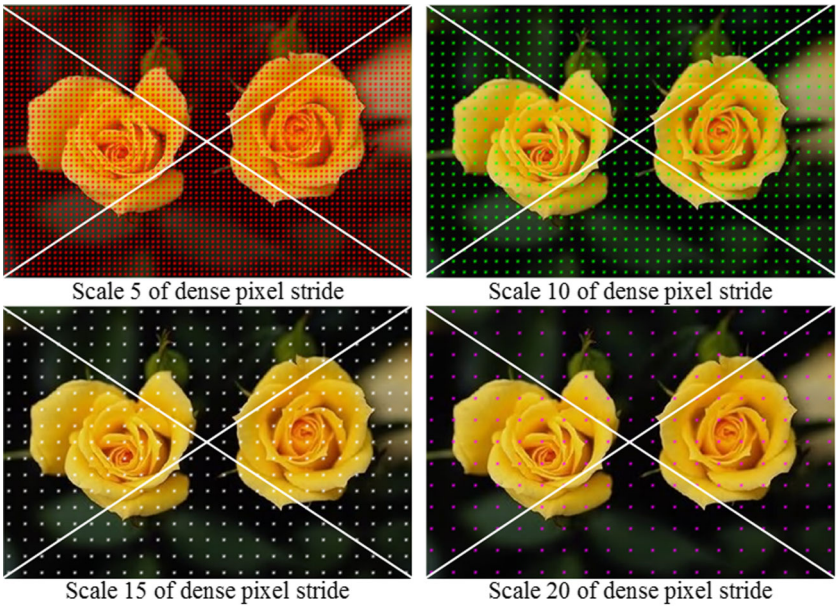

Fig. 4 Different scales of dense SIFT features on sample image of the Corel-A database [44]

4.1 Evaluation measures

The performance of proposed approach is evaluated by measuring the precision and recall parameters. The precision determines the number of correctly retrieved images over the total number of retrieved images from the test image database and it measures the specificity of image retrieval system, represented as:

$$Precision = \frac{R_c}{R_t} \tag{10}$$

Where R_c is the total number of correctly retrieved images and R_t is the total number of retrieved images. The ratio of correctly retrieved images over the total number of images of that semantic class in the image database is known as recall and it measures the sensitivity of the image retrieval system, represented as:

$$Recall = \frac{R_c}{T_s} \tag{11}$$

Where T_s is the total number of images in the semantic class s and R_c is the correctly retrieved images.

4.2 Performance evaluation on the Corel-A image database

The Corel-A image database [44] is a subset of the Corel image database [9] that is commonly utilized for the evaluation of CBIR approaches [8, 27, 30, 31, 35]. There are 10 semantic classes in Corel-A image database and each semantic class contains 100 images. Figure 5 shows a sample of images from 10 semantic classes of the Corel-A image database. Figure 6 represents the mean average precision (MAP) values for top-20 image retrievals obtained from the proposed approach, using the dense pixel strides of 5, 10, 15, and 20, respectively. We have selected the result of top-20 image retrievals because state-of-the-art image retrieval techniques [8, 27, 30, 31, 35] are also reported for top-20 image retrievals.

Fig. 5 Sample images from 10 semantic classes of the Corel-A image database [44]

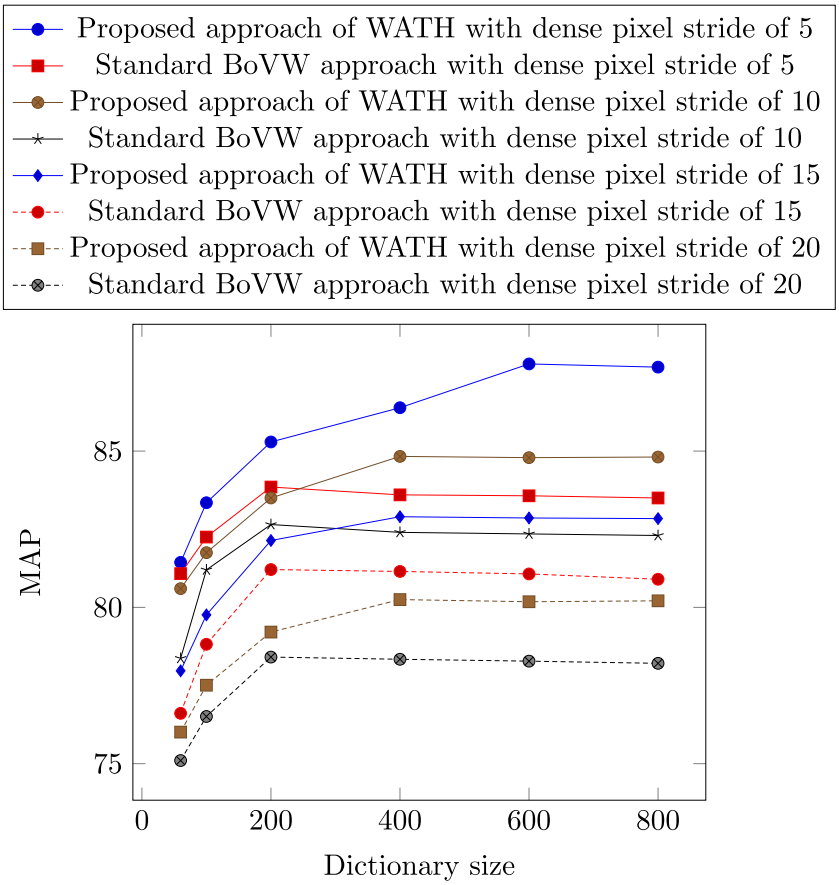

Fig. 6 Comparison of proposed approach based on WATH with baseline competitor on the Corel-A image database

According to the experimental results shown in Table 1, on a dictionary size of 600 visual words, standard 95% value of confidence interval is 87.67-87.89 with 5% significance level result in better MAP and moderate standard error as compared to other reported values of dictionary sizes. Further statistical investigation is also performed to strengthen the findings of proposed approach based on WATH. In Table 1, we have also applied the Wilcoxon matched-pairs signed-rank test to compare results on the dictionary size of 600 visual words with dictionary sizes of 60, 100, 200, 400, and 800 visual words as well as with proposed approach of [8]. According to the statistical analysis results of the Wilcoxon matched-pairs signed-rank test, the dictionary size of 600 visual words shows statistical significance with

Table 1 Statistical analysis of MAP on the Corel-A image database with a dense pixel stride of 5 (bold values indicate the highest MAP and lowest standard error)

Dictionary size, Features % used	60	100	200	400	600	800
10 %	80.96	82.78	85.15	85.65	87.63	87.52
25 %	81.11	82.91	85.23	85.80	87.81	87.65
50 %	81.40	83.66	85.28	86.84	**87.85**	87.73
75 %	81.79	83.69	85.31	86.83	87.83	87.74
100 %	81.94	83.70	85.49	86.82	87.82	87.79
MAP	81.440	83.348	85.292	86.388	87.788	87.686
Std. Dev.	0.422	0.461	0.126	0.607	0.089	0.105
Conf. Interval	80.91-81.96	82.77-83.92	85.13-85.44	85.63-87.14	87.67-87.89	87.55-87.81
Std. Error	0.188	0.206	0.056	0.271	**0.040**	0.047
Statistical results of Wilcoxon matched-pairs signed-rank test						
Z-Value	2.023	2.023	2.023	2.032	[8] 2.023	2.023
P-Value	0.0431	0.0431	0.0431	0.0422	[8] 0.0431	0.0431

Table 2 Performance measure of MAP for top-20 image retrievals on the Corel-A image database (bold values indicate class-wise top two MAP)

Class/Method	Proposed WATH approach	Spatial level-2 approach [8]	Spatial LGH approach [35]	GMM spa-tiogram [30]	Color fusion [31]	Spatial BoF [27]
Africa	**77.82**	69.08	**73.03**	72.50	51	64
Beach	**79.56**	72.20	74.58	65.20	**90**	54
Buildings	**80.75**	**84.85**	80.24	70.60	58	53
Buses	95.74	**95.75**	**95.84**	89.20	78	94
Dinosaurs	98.12	**100**	97.95	**100**	**100**	**98**
Elephants	**89.54**	**89.99**	87.64	70.50	84	78
Flowers	86.87	94.01	85.13	**94.80**	**100**	71
Horses	89.41	86.38	86.29	91.80	100	**93**
Mountains	85.78	82.85	82.43	72.25	**84**	42
Food	**84.92**	**85.88**	78.96	78.80	38	50
MAP	**87.85**	**86.27**	84.21	80.57	78.30	69.70

all the other considered dictionary sizes as well as with proposed approach of [8], because the P-value is less than the level of significance $\alpha = 0.5$ in all cases.

According to the experimental results shown in Fig. 6, the best MAP for top-20 image retrievals is obtained using dense pixel strides of 5, 10, 15, and 20 is 87.85%, 84.83%, 82.90%, and 80.25%, respectively using 50% feature percentage, and dictionary size of 600 visual words. For every experiment, the image retrieval precision decreases with the increase in pixel stride and vice versa. In order to show the performance of proposed approach based on WATH, the precision and recall for top-20 image retrievals obtained using the proposed approach on a dictionary of size 600 visual words (the dense pixel stride of 5 and 50% feature percentage) are compared with state-of-the-art image retrieval techniques [8, 27, 30, 31, 35] and presented in Tables 2 and 3, respectively.

The experimental analysis in Table 2 shows that best MAP of the proposed approach using the dense pixel stride of 5 is 87.85%. The best values of average precision and recall are mentioned as bold in Table 2 and Table 3. The MAP obtained from the proposed approach on the basis of WATH is better than the existing state-of-the-art image retrieval techniques [8, 27, 30, 31, 35].

Figures 7 and 8 shows the top-20 image retrieval results obtained on the basis of a similarity measure for the semantic classes "Buses" and "Dinosaurs", respectively. The numeric value shown at the top of every image is the score or classifier decision value that is utilized to retrieve similar images by applying the Euclidean distance between the score of the query image and the images placed in an image database. The image shown at the top of Figs. 7 and 8 is the query image, while rest of the images are the retrieved images in the response to given query image.

Table 3 Performance measure of recall for top-20 image retrievals on the Corel-A image database (bold values indicate class-wise top two recall)

Class/Method	Proposed WATH approach	Spatial level-2 approach [8]	Spatial LGH approach [35]	GMM spa-tiogram [30]	Color fusion [31]	Spatial BoF [27]
Africa	**15.56**	13.82	14.61	14.50	10.20	12.80
Beach	**15.91**	14.44	14.92	13.04	**18.00**	10.80
Buildings	**16.15**	**16.97**	16.05	14.12	11.60	10.60
Buses	19.15	**19.50**	**19.17**	17.84	15.60	18.80
Dinosaurs	**19.62**	**20.00**	19.59	**20.00**	**20.00**	19.60
Elephants	**17.91**	**18.00**	17.53	14.10	16.80	15.60
Flowers	17.37	18.80	17.03	**18.96**	**20.00**	14.20
Horses	17.88	17.28	17.26	18.36	**20.00**	**18.60**
Mountains	17.15	16.57	16.49	14.45	**16.80**	08.40
Food	**16.98**	16.54	15.79	15.76	07.60	10.00
Mean	**17.37**	**17.25**	16.84	16.11	15.66	13.00

Fig. 7 Image retrieval result shows reduction of a semantic gap for the semantic class "Buses" of the Corel-A image database

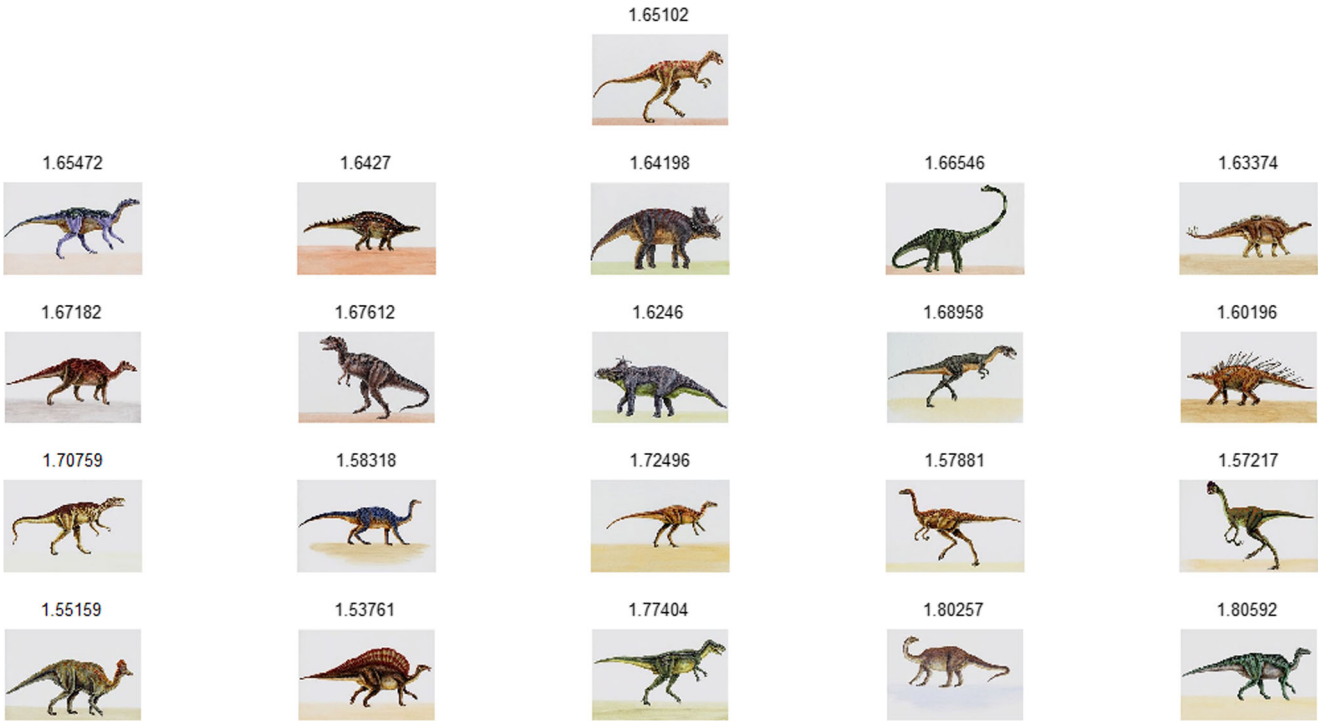

Fig. 8 Image retrieval result shows reduction of a semantic gap for the semantic class "Dinosaurs" of the Corel-A image database

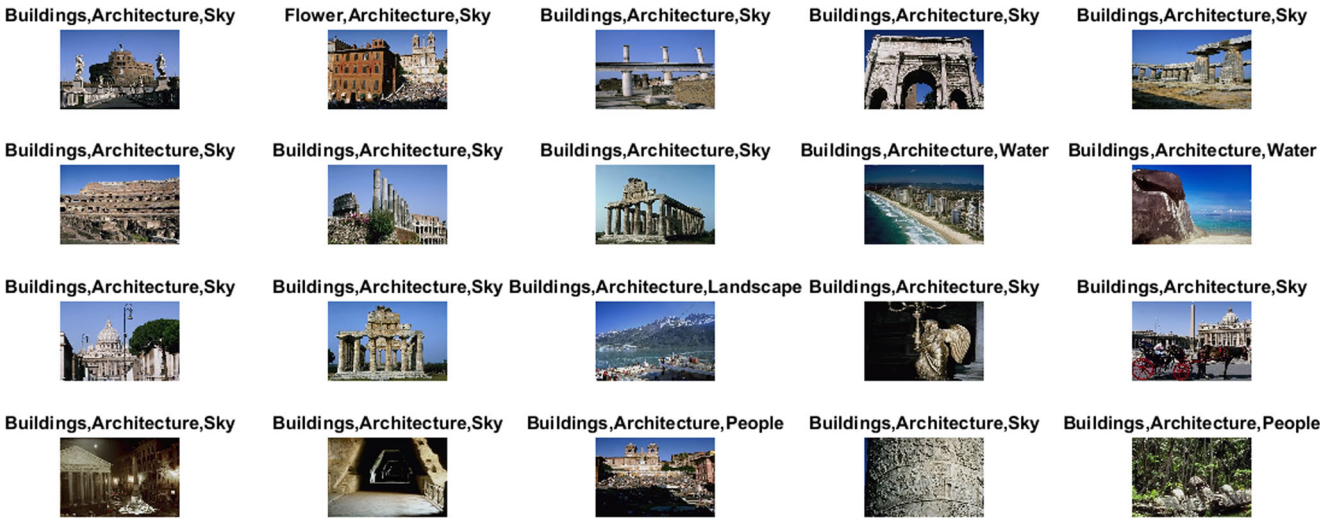

Fig. 9 Top-20 retrievals of semantic auto-image annotation using keyword "Buildings" for the semantic class "Buildings" of the Corel-A image database

In order to retrieve the images by using text-based search, instead of giving query image, we have also retrieved the images using automatic image annotation (AIA) [16], whose results are shown in the Figs. 9 and 10. AIA [16] is a technique that is utilized to represent an image in the form of high-level language keywords (annotations). The predefined semantic class labels that are used for the 10 semantic classes of Corel-A image database are Africa, People, Beach, Sky, Buildings, Architecture, Buses, Transport, Dinosaurs, Animals, Elephants, Forest, Flowers, Garden, Horses, Grass, Mountains, Landscape, Food, and Restaurants. We have evaluated our proposed approach by assigning 3 annotations per image. Top two classification scores are obtained after classification for each image and first two annotations are assigned on the basis of occurrence of first classification score, while the third annotation is assigned on the basis of occurrence of second classification score for each image.

4.3 Performance evaluation on the Corel-B image database

There are 1500 images in the Corel-B image database [9] that are categorized into 15 semantic categories. Figure 11 shows a sample of images from each semantic category of the Corel-B image database. The Corel-B image database

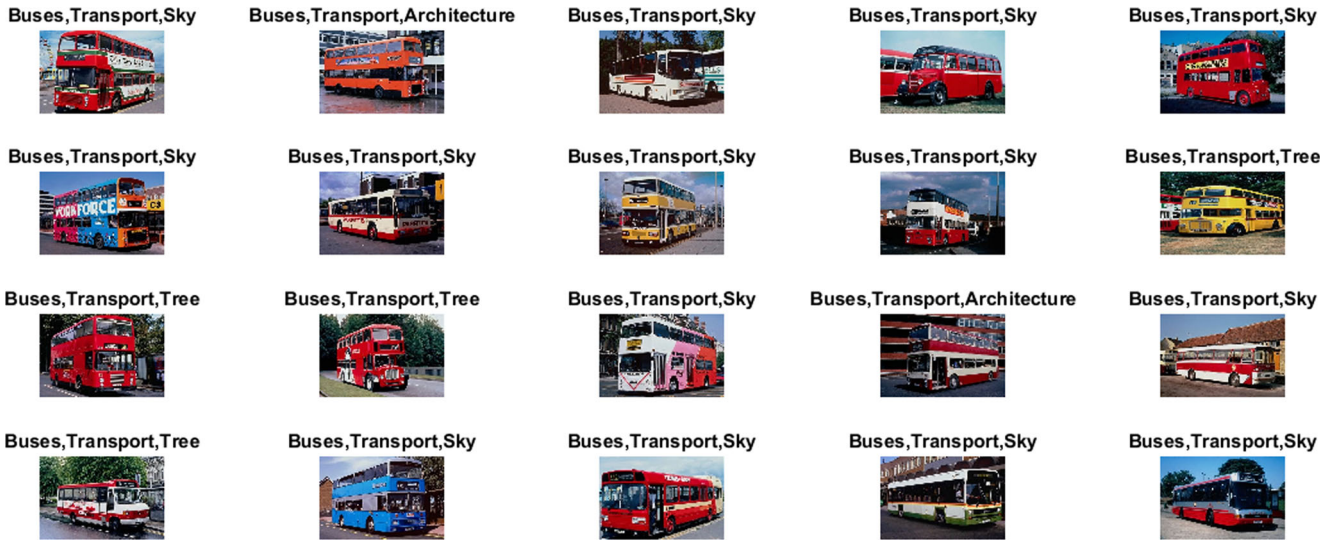

Fig. 10 Top-20 retrievals of semantic auto-image annotation using keyword "Buses" for the semantic class "Buses" of the Corel-A image database

Fig. 11 Sample images from 15 semantic classes of the Corel-B image database [44]

was also utilized for the performance measure of image retrieval approaches [30, 45]. Figure 12 shows the MAP values obtained from the proposed approach using dense pixel strides of 5, 10, 15, and 20, respectively.

According to the experimental results as shown in Fig. 12, the best MAP obtained from the proposed approach on a dictionary size of 800 visual words using dense pixel strides of 5, 10, 15, and 20 with a dictionary size of 800

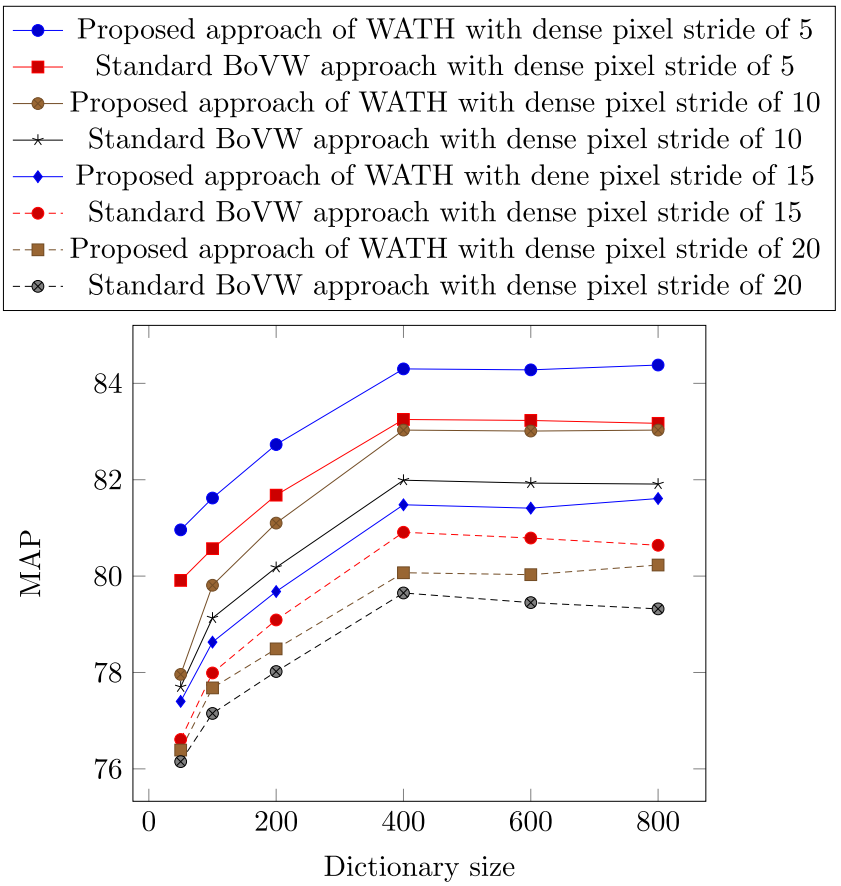

Fig. 12 Comparison of proposed approach based on WATH with baseline competitor on the Corel-B image database

Table 4 Performance measures of precision and recall on the Corel-B image database (bold values indicate highest MAP and recall)

Performance measures	Proposed WATH approach	SQ + Spatiogram [30]	GMM + mSpatiogram [30]	I-CFS-EM approach [45]
Precision	84.38	63.95	74.10	84.00
Recall	16.88	12.79	13.80	16.80

visual words is 84.38%, 83.03%, 81.61%, and 80.23%, respectively. For every experiment, the image retrieval precision obtained using dense pixel stride of 5 is better than that of 10, 15, and 20 pixel strides. Table 4 shows the performance measures of precision and recall obtained using the proposed approach (with dense pixel stride of 5 and dictionary size of 800 visual words) with existing state-of-the-art image retrieval approaches [30, 45]. The comparative analysis in Table 4 shows that the proposed approach based on WATH outperform as compared to the state-of-the-art CBIR approaches [30, 45].

4.4 Performance evaluation on the Scene-15 image database

There is total of 4485 images in the Scene-15 image database [13] that are categorized into 15 semantic classes, and each category consists of 200–400 semantic images. Figure 13 shows a sample of each image from all the categories of the Scene-15 image database. The MAP obtained using proposed approach based on WATH is compared with 2×2 spatial rectangular histograms based approach. The MAP as a function of different dictionary sizes is graphically presented in Fig. 14.

According to the experimental results shown in Fig. 14 and Table 5, the MAP of 81.94% is obtained using WATH approach on a dictionary size of 1000 visual words (with dense pixel stride of 5), while MAP obtained using 2×2 spatial rectangular histograms based approach is 80.08% (with pixel stride of 5 and dictionary size of 800 visual words). The proposed approach on pixel strides of 5, 10, 15, and 20 outperforms 2×2 spatial rectangular histograms based approach on a dictionary of all sizes, as well as state-of-the-art CBIR approaches [8, 13, 46].

4.5 Computational complexity

The computational complexity of proposed approach is calculated on desktop PC with following specifications; Intel Pentium (R) Core i7 2.4 GHz microprocessor, 2 GB GPU, and 8 GB RAM using Windows 7 operating system. The proposed approach is implemented in MATLAB 2015a, and by utilizing the training database, the dictionary is formulated offline, and by utilizing the test database, tested at run time. For only feature computation, average CPU time required using proposed approach of image resolution 256×384 or 384×256 is shown in Table 6. The computational complexity of feature computation to image retrieval for Corel-A image database is shown in Table 7.

5 Conclusion and future directions

This paper presents a novel image representation that adds the spatial information to the inverted index of BoVW representation on the basis of WATH approach, reduces overfitting problem on a dictionary of larger sizes and semantic gap issue. Equal weight ($W = 0.25$) is assigned to the histograms of four divided triangular regions for the computation of histograms of visual words. Dense features are extracted at different scales and pixel strides in order to determine the best performance of the proposed approach based on WATH. Keeping in view the low computational cost, classification is performed using SVM Hellinger

Fig. 13 Sample images from 15 semantic classes of the Scene-15 image database

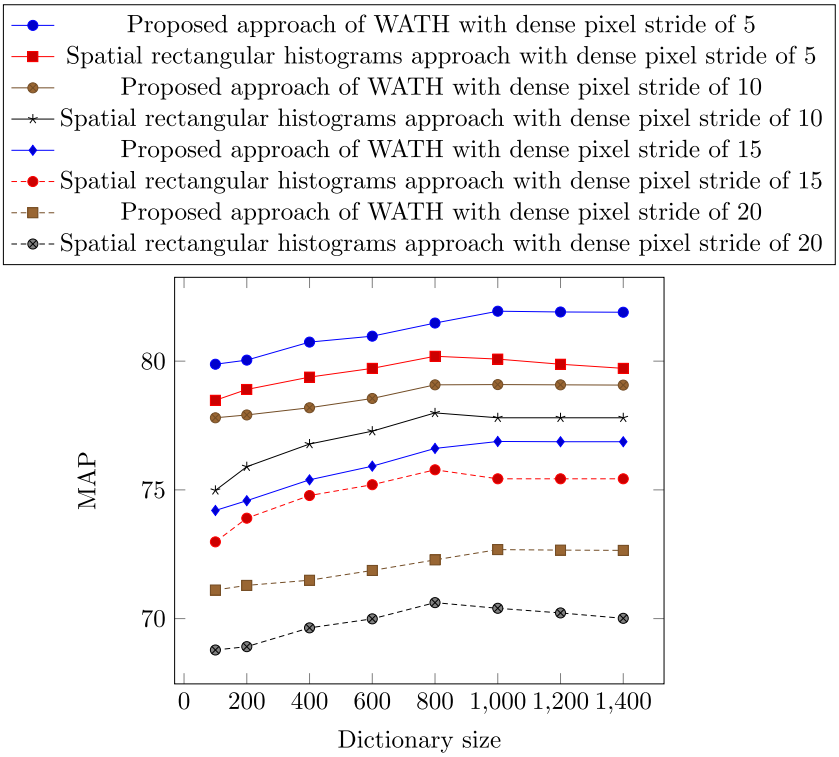

Fig. 14 Comparison of proposed approach based on WATH with 2×2 spatial rectangular histograms based approach on the Scene-15 image database

Table 5 Performance measures of precision and recall on the Scene-15 image database (bold values indicate highest MAP and recall)

Performance measures	Proposed WATH approach	Spatial level-2 approach (DBN) [8]	Spatial level-2 approach (SVM) [8]	SPM approach [13]	HOG 2×2 approach [46]
Precision	**81.92**	79.70	77.00	81.10	81.00
Recall	**16.38**	15.94	15.40	16.22	16.20

Table 6 Comparison of computational cost (in seconds) required for feature extraction only

Proposed WATH approach	Spatial level-2 approach [8]	BoVW approach	RSHD approach [33]			EODH approach [28]
			SEH	CDH	RSHD	
0.0745	0.0821	0.0641	0.186	1.709	0.375	5.6

Table 7 Comparison of computational cost (in seconds) of the proposed approach (complete framework) with state-of-the-art CBIR approaches

Number of images retrieved	Proposed WATH approach	Spatial level-2 approach [8]	BoVW approach
Foremost-05	0.3584	0.3726	0.3415
Foremost-10	0.4811	0.5178	0.4674
Foremost-15	0.6799	0.7050	0.6589
Foremost-20	0.8491	0.8882	0.8213
Foremost-25	1.0121	1.0599	0.9913

kernel. The proposed approach is evaluated on three image benchmarks and results are compared with state-of-the-art CBIR approaches. According to the experimental results, the proposed approach based on the WATH outperforms state-of-the-art CBIR approaches. In future, we will evaluate the performance of proposed approach using deep learning for large-scale image retrieval on ImageCLEF and ImageNET image databases.

Compliance with Ethical Standards

Competing Interest All the authors declare no competing interest.

References

1. Alzu'bi A, Amira A, Ramzan N (2015) Semantic content-based image retrieval: A comprehensive study. J Vis Commun Image Represent 32:20–54
2. Castellano G, Fanelli AM, Sforza G, Torsello AM (2016) Shape annotation for intelligent image retrieval. Appl Intell 44(1):179–195
3. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surv 40(2):5
4. Chua T-S, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. ACM, p 48
5. Yousaf RM, Rehman S, Dawood H, Ping G, Mehmood Z, Azam S, Khan AA (2017) Saliency based object detection and enhancements in static images. In: International Conference on Information Science and Applications. Springer, pp 114–123
6. Sivic J, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: Proceedings of the 9th IEEE International Conference on Computer Vision, 2003. IEEE, pp. 1470–1477
7. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07. IEEE, pp 1–8
8. Ali N, Bajwa KB, Sablatnig R, Mehmood Zahid (2016) Image retrieval by addition of spatial information based on histograms of triangular regions. Computers & Electrical Engineering
9. Li J, Wang JZ (2008) Real-time computerized annotation of pictures. IEEE Trans Pattern Anal Mach Intell 30(6):985–1002
10. Zhou W, Li H, Yijuan L, Tian Q (2013) Sift match verification by geometric coding for large-scale partial-duplicate web image search. ACM Trans Multimed Comput Commun Appl 9(1):4
11. Khan R, Barat C, Muselet D, Ducottet C (2012) Spatial orientations of visual word pairs to improve bag-of-visual-words model. In: Proceedings of the British Machine Vision Conference. BMVA Press, pp 89–1
12. Anwar H, Zambanini S, Kampel M, Vondrovec K (2015) Ancient coin classification using reverse motif recognition: Image-based classification of roman republican coins. IEEE Signal Process Mag 32(4):64–74
13. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on Computer vision and pattern recognition, vol 2. IEEE, pp 2169-2178
14. Mehmood Z, Anwar SM, Altaf M, Ali N (2017) A novel image retrieval based on rectangular spatial histograms of visual words. Kuwait Journal of Science, Kuwait
15. Ashraf R, Bashir K, Mahmood T (2016) Content-based image retrieval by exploring bandletized regions through support vector machines. J Inf Sci Eng 32(2):245–269
16. Zhang D, Md MI, Guojun L (2012) A review on automatic image annotation techniques. Pattern Recogn 45(1):346–362
17. Liu Y, Zhang D, Guojun L, Ma W-Y (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282
18. Das R, Thepade S, Ghosh S (2015) Multi technique amalgamation for enhanced information identification with content based image data. SpringerPlus 4(1):1–26
19. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
20. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: Computer visionECCV 2006. Springer, pp 404-417
21. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR, vol 1. IEEE, pp 886–893
22. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 22(10):761–767
23. Leutenegger S, Chli M, Siegwart RY (2011) Brisk: Binary robust invariant scalable keypoints. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 2548–2555
24. Mukherjee D, Wu QMJ, Wang G (2015) A comparative experimental study of image feature detectors and descriptors. Mach Vis Appl 26(4):443–466
25. Krajnik T, Cristóforis P, Nitsche M, Kusumam K, Duckett T (2015) Image features and seasons revisited. In: 2015 European conference on mobile robots (ECMR). IEEE, pp 1–7
26. Mahmood T, Nawaz T, Ashraf R, Shah M, Khan Z, Irtaza A, Mehmood Z A survey on block based copy move image forgery detection techniques. In: 2015 international conference on emerging technologies (ICET). IEEE, pp 1–6
27. Wang C, Zhang B, Qin Z, Xiong J (2013) Spatial weighting for bag-of-features based image retrieval. In: Integrated uncertainty in knowledge modelling and decision making. Springer pp 91–100
28. Tian X, Jiao L, Liu X, Zhang X (2014) Feature integration of eodh and color-sift: Application to image retrieval based on codebook. Signal Process Image Commun 29(4):530–545
29. Jing Y, Qin Z, Wan T, Xi Z (2013) Feature integration analysis of bag-of-features model for image retrieval. Neurocomputing 120:355–364
30. Zeng S, Huang R, Wang H, Kang Z (2016) Image retrieval using spatiograms of colors quantized by gaussian mixture models. Neurocomputing 171:673–684
31. Walia E, Pal A (2014) Fusion framework for effective color image retrieval. J Vis Commun Image Represent 25(6):1335–1348
32. Yuan X, Yu J, Qin Z, Wan T (2011) A sift-lbp image retrieval model based on bag of features. In: IEEE International Conference on Image Processing
33. Dubey SR, Singh SK, Singh RK (2015) Rotation and scale invariant hybrid image descriptor and retrieval. Comput Electr Eng 46:288–302
34. Wan J, Wang D, Hong Hoi SC, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the ACM international conference on multimedia. ACM, 157–166
35. Mehmood Z, Anwar SM, Ali N, Habib HA, Rashid M (2016) A novel image retrieval based on a combination of local and global histograms of visual words. Math Probl Eng 2016

36. Gupta R, Patil H, Mittal A (2010) Robust order-based methods for feature description. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 334–341
37. Liu L, Fieguth WP (2012) Texture classification from random features. IEEE Trans Pattern Anal Mach Intell 34(3):574–586
38. Mahmood T, Nawaz T, Mehmood Z, Khan Z, Shah M, Ashraf R (2016) Forensic analysis of copy-move forgery in digital images using the stationary wavelets. In: 2016 6th international conference on innovative computing technology (INTECH). IEEE, pp 578–583
39. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. vol 1. Prague, pp 1–2
40. Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, pp 1027–1035
41. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge university press, Cambridge
42. Vedaldi A, Zisserman A Sparse kernel approximations for eff ient classification and detection. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2320–2327
43. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: Computer Vision–ECCV 2006. Springer, pp 490–503
44. Wang JZ, Li J, Wiederhold G (2001) Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans Pattern Anal Mach Intell 23(9):947–963
45. Li R, Bhanu B, Krawiec K (2007) Hybrid coevolutionary algorithms vs. svm algorithms. In: Proceedings of the 9th annual conference on genetic and evolutionary computation. ACM, pp 456–463
46. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3485–3492

Zahid Mehmood is serving at Department of Software Engineering, University of Engineering and Technology (UET), Taxila, Pakistan. He has been completed his BS(Hons) Computer Engineering in 2009 from COMSATS University of Sciences and Technology, Pakistan, and MS Electronic Engineering in 2012 with specialization in Signal and Image Processing from International Islamic University (IIU), Islamabad, Pakistan. He has been completed his Ph.D. Computer Engineering in 2016 from UET, Taxila, Pakistan. He is a reviewer for international journals and conferences such as Pattern Recognition, Computer & Electrical Engineering, PAMI, CVPR, Neurocomputing etc. His research interests are content-based image retrieval (CBIR), medical imaging, image forensic, computer vision, and machine learning.

Toqeer Mahmood is serving as a senior researcher at University of Engineering and Technology (UET), Taxila, Pakistan. He has been completed his MS Computer Engineering in 2010 from Center for Advanced Studies in Engineering (CASE), Islamabad, Pakistan. He has been completed his Ph.D. Computer Engineering in 2016 from UET, Taxila, Pakistan. His research interests are Image Forensic, Image Processing, Image Retrieval, Computer Vision, and Computer Networks.

Muhammad Arshad Javid is serving at Department of Basic Sciences, University of Engineering and Technology (UET), Taxila, Pakistan. He has been completed his Ph.D. Physics in 2013 from Quaid-e-Azam University, Islamabad, Pakistan. He has been visited in 2012 as a research scholar Wayne State University, USA. His research interests are Computational Physics, MRI Contrast Agents, CT, and Neuro Imaging.