Optik 131 (2017) 242–254

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A new image feature descriptor for content based image retrieval using scale invariant feature transform and local derivative pattern

Davar Giveki [a,*], Mohammad Ali Soltanshahi [b], Gholam Ali Montazer [c]

[a] Iranian Research Institute for Information Science and Technology (IranDoc) Tehran, Iran
[b] School of Mathematics, Statistics and Computer Science, Department of Computer Science, University of Tehran, Tehran, Iran
[c] Information Technology Engineering Department, School of Engineering Tarbiat Modares University, Iran

ARTICLE INFO

Article history:
Received 20 June 2015
Received in revised form 3 October 2016
Accepted 7 November 2016

Keywords:
Content based image retrieval
SIFT
HOG
LBP
LTP
LDT

ABSTRACT

This paper presents a new methodology to retrieve images of different scenes by introducing a novel image descriptor. The proposed descriptor works with Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Local Derivative Pattern (LDP), Local Ternary Pattern (LTP) and any other feature descriptor that can be applied on the image pixels. As the proposed descriptor considers a group of pixels together, higher level of semantic is achieved. In this work, a new image descriptor using SIFT and LDP is introduced that is able to find similarities and matches between images. The proposed descriptor produces highly discriminative features for describing image content. Four image datasets are used for evaluating our proposed descriptor. Comprehensive experiments have been conducted using various classifiers and different image features to show the superiority of the proposed method.

1. Introduction

In many computer vision tasks such as image retrieval, image classification, object recognition and moving object detection, the most important stage is image feature extraction which is done by applying strong image descriptors on images [5,29,11,12]. An ideal image descriptor should be robust against some image transformations such as scaling, rotation, partly occlusion, illumination variations, size and pose as much as possible. Semantically recognizing, localizing and describing objects in a complex scene are very challenging problems. Although these problems have been efficiently and successfully solved by the human visual and cognitive system, no method has been yet proposed to offer a human-like performance. Therefore, a question arises here: How this semantic gap can be reduced in terms of image feature description?

Early research works consider the perceptual gap between the low-level image features and what we understand from an image (the high-level conceptual identification) [27]. Research in neuroscience relates the significance role of the feature extraction step for a more accurate visual understanding [1] because the human cognition process is composed of a combination of complex features [23].

* Corresponding author.
E-mail addresses: Giveki@students.irandoc.ac.ir (D. Giveki), ali.soltanshahi@gmail.com (M.A. Soltanshahi), montazer@modares.ac.ir (G.A. Montazer).

http://dx.doi.org/10.1016/j.ijleo.2016.11.046
0030-4026/

Image processing and computer vision society are doing research on biologically inspired feature descriptions to address this issue. These approaches aim at exploring improvements in the feature extraction and description steps of the image understanding framework [46,43,56,31].

This paper explores the feature description process from low-level semantic (pixel-level features) to mid-level semantic (objects or patch −level features). The proposed descriptor can be applied by standard descriptors including SIFT [26], LBP [37], LTP [48], LDP [60] and HOG [6].

1.1. Related works and motivation

Object recognition, image classification and image retrieval have been widely reviewed in the literature [22,10,24,47,38,7,20].

A common classification system consists of four major parts: the first part is local image feature extractor, the second part is local image descriptor, the third part pools encoded descriptors into a global image descriptor, and the final part is a classification unit for identifying the class of the images.

Most of the recent studies, propose feature fusion algorithms to better describe images. Wang and Han Wang and Han, 2009 proposed an integrated HOG-LBP feature on the cell-level as a human detector. It has been reported that the method achieves good performance in the INRIA dataset. Heikkilä [17] derived a new descriptor that combined the strengths of the SIFT and LBP. The proposed descriptor worked by center-symmetric local binary patterns. Zheng [57] combined the SIFT and rotation invariant LBP for an image matching problem, where the LBP descriptor was used to describe the local region centered at the keypoints which were found by the SIFT algorithm. Schwartz [45] introduced a human detection method by integrating edge-based features with texture and color information to form a relative large descriptor set. Yu proposed a feature integration template for image retrieval based on the bag of features model and weighted K-means clustering [55]. In their experiments, authors integrated SIFT, LBP and HOG features. Based on comprehensive experimental studies on benchmark image retrieval datasets, they achieved the best performance by integrating SIFT and LBP. In [18], Hu and Collomosse introduced an image retrieval system for the interactive search of photo collections. They proposed Gradient Field HOG (GF-HOG) which is an adapted form of the HOG descriptor suitable for Sketch Based Image Retrieval (SBIR). In that paper, GF-HOG features were used in Bag of Visual Words (BoVW) framework. Experimental results showed that retrieval based on GF-HOG outperformed retrieval based on SIFT, multi-resolution HOG, Self Similarity, Shape Context and Structure Tensor.

In a common object recognition or classification system, after feature extraction, feature coding plays a vital rule. Huang et al. [20] discussed various methods of feature coding, including motivations and mathematical representations of feature coding. Furthermore, they analyzed relations of different coding methods in theory, and empirically evaluated their performance. In our work, we used bag of visual words for feature coding.

The second step of the object recognition, image classification and image retrieval systems has also been widely reviewed. For encoding a set of local descriptors into a single high dimensional feature vector, the Fisher Vector method in [42] achieves the state-of-the-art performance.

The third step has also reached notable improvements especially when bag of visual words framework is used [22]. Major drawback of the third step is loss of spatial information. To alleviate this problem, Lazebnik used spatial pyramid matching (SPM). In spatial pyramid matching, each image is divided into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction. The number of points in each grid cell is then recorded. This is a pyramid representation because the number of points in a cell at one level is simply the sum over those contained in the four cells it is divided into at the next level. The cell counts at each level of resolution are the bin counts for the histogram representing that level. The soft correspondence between the two point sets can then be computed as a weighted sum over the histogram intersections at each level. Similarly, the lack of correspondence between the point sets can be measured as a weighted sum over histogram differences at each level. Considering SPM for images, image pyramid matching is applied to the two-dimensional image space, and a bag of visual word vector is computed for each grid cell at each pyramid resolution level. So, the 2D points in the example above are replaced by visual words obtaining a pyramid histogram of visual words descriptor for the image.

Typically, matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. The pyramid histogram of visual words is normalized to sum the unity. For a pyramid level of 2, for example, 21 bags are generated. The first bag comes from the image without splitting. In the next level, the image is split into 4 tiles of the same size. The next level splits each of the 4 tiles into another set of 4 tiles. Therefore, there is 1 bag for level 0, 4 bags for level 1 and 16 bags for level 2 [22].

Exploring the final step of the classification system, discriminative classifiers such as SVM [3] and ANN [31] have reached acceptable performances.

By studying the related work, we felt that more work should be done to achieve more reliable image descriptors. However, these descriptors should carry more semantic to better describe image contents.

There are some similar works to ours. The work in [2] proposed mid-level features for object recognition. The article presented a detailed analysis on different levels of pooling strategies. Authors defined macro-feature vectors as jointly encoded small neighborhoods of SIFT descriptors. The neighborhoods were defined by a fixed size of squares that encode multiple SIFT descriptors into one as the macro-feature vector. The main drawback of the work is that it only uses fixed

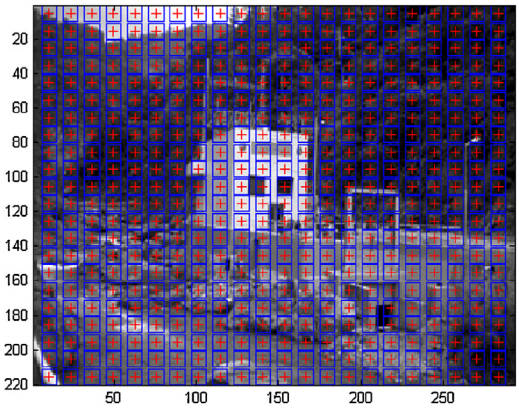

Fig. 1. Diving image into 484 patches of size 9×9 for computing SIFT features.

sized (multiple) blocks without considering the region properties while our proposed method aims at combining spatial information of the patch and encoding it into a descriptor that has flexible and adaptive coverage depending on the spatial region properties. In [35], Parikh considered the role of local and global information in image classification. In addition, he focused on exploring the performance limitations of current techniques. Another study that aimed at labeling image regions depending on the similarity of the extracted features in the training set was presented in [49]. In that study, scene-level matching with global image descriptors was accompanied by spatial pyramid matching of mid-level features. In another paper, Zheng used the low, mid and high-level cues [58]. Individual classifiers were trained on different levels of descriptors and classification outputs were combined for the final decision. Descriptor level grouping has also been studied in [15] where local histograms that from larger neighboring regions, have shown to improve classification performance. The method used a fixed neighborhood definition to aggregate the local histograms; whereas, our method proposes a flexible and more natural region description.

1.2. Main contributions

Proposing a high performance content based image retrieval (CBIR) model using mid-level features is our main motivation of this work. To better explain the proposed descriptor, please consider the SIFT algorithm. By using the SIFT, one pixel in an image can be matched with another pixels in other images with high degree of certainty. This fact shows that features extracted by the SIFT are strong enough in pixel level. We want to take this advantage and extend the ability of the SIFT features from low-level (pixel-level) to higher level of semantic namely objects, segments or patches using a new template. Hence, our main contribution is proposing a new descriptor with the nature of the SIFT which is able to find similarities and matches between images with high accuracy. The proposed descriptor provides highly discriminative features for describing the content of an image. Comprehensive experiments are conducted using feature descriptors such as LBP, LTP, LDP and HOG to show the efficiency of the proposed method. Furthermore, different classifiers have been employed to reach higher performance. Finally, we conclude that integrating SIFT and LTP features through the proposed descriptor achieves the highest performance. The organization of the rest of the paper is as follows: In Section 2 the new image descriptor is proposed. The proposed image retrieval model is shortly described in Section 3. Experimental results are presented in Section 4 and conclusion is drawn in Section 5.

2. Proposed image descriptor

In the proposed method, images are divided into some patches. Then by considering patch positions, the changes of the image descriptor (SIFT, LBP, LTP, LDP and HOG) are described in terms of angle. The proposed descriptor is obtained by quantizing these features (the position of patches and the angle). Figs. 1 and 2 show the image divisions for computing SIFT, LBP and HOG features, respectively.

The proposed image descriptor is explained based on the SIFT features. Applying it with other descriptors is straightforward. In our implementations, patches are obtained by dividing images into regular grids (Figs. 1 and 2).

Given an image I for computing its feature vector, after image division, SIFT features are $\vec{V}_{x,y}$extracted densely. For each patch centered at(x, y), its SIFT feature vector is denoted by

Then angle of the gradient vector of the SIFT features at(x, y), $\theta_{x,y}$, is computed as follows:

$$\theta_{x,y} = \tan^{-1} \frac{\|\vec{V}_{x,y+1} - \vec{V}_{x,y-1}\|}{\|\vec{V}_{x+1,y} - \vec{V}_{x-1,y}\|}$$

where $\|\cdot\|$ is a vector norm. Here Euclidean norm is used.

Fig. 2. Dividing image into 16 patches for computing LBP and HOG features.

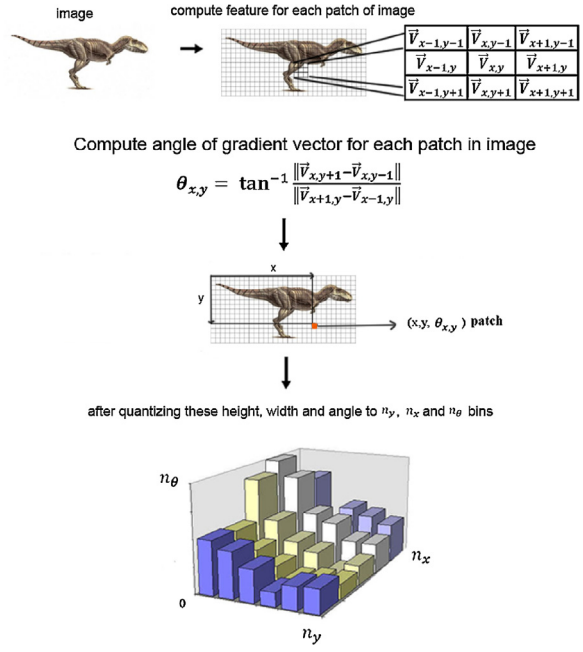

Fig. 3. The framework of the proposed image descriptor.

Therefore, three features for each patch, namely, height, width and angle of the gradient vector of the patch are available. By quantizing these features (height, width and angle) ton_x, n_y and n_θ bins, respectively, the proposed image descriptor is constructed. Uniform quantization has been adopted in this paper. The proposed technique counts occurrences of gradient orientations in the localized portions of an image.

The important idea behind the descriptor is that appearance and shape of the local object/scene within an image are captured by computing the distribution of intensity gradients, edge directions or texture features (in the case that the LBP-based features are used). Fig. 3 shows the framework of the proposed descriptor. The complexity of the proposed method is the same as the complexity of the methods used for extracting image features. In fact, there is no overhead in time complexity. The time complexity of executing the proposed descriptor is $\left(\|\vec{V}\|+1\right) * numberofpatches$.

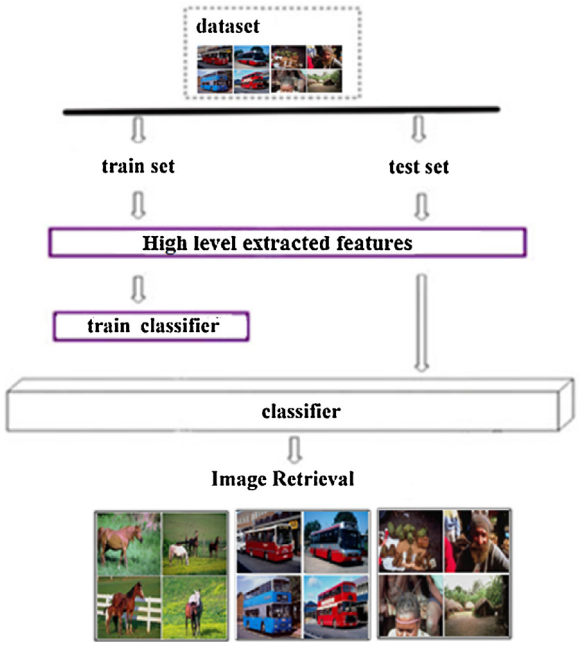

Fig. 4. The proposed scheme for image retrieval.

3. Proposed CBIR system

In this section, each stage of the proposed methodology for image retrieval is explained. As shown in Fig. 4, the proposed CBIR model has three basic stages as follows:

At first, images in datasets are randomly divided to train set and test set. Afterwards, image features are extracted using the proposed image descriptor for every image in the test set and train set. In the next step, features extracted from training images are used to train a classifier. After training and testing the classifier, user selects a query image and then feature vector of the query image is extracted and the trained classifier determines the category of the query image.

SVM, RBF and MLP neural network classifiers are used to evaluate the proposed method. For the SVM classifier, radial basis function kernel is used [4]. In addition, Multi layered feed forward neural network with one hidden layer is trained and tested. The network uses $(2d+1)$ neurons in the hidden layers stipulated by Kolmogorov's theorem [21] where d is dimension of feature vectors. The neurons used in the output layer of the network have sigmoidal activation function. The scaled conjugate gradient algorithm [30] is selected for fast and robust training with the back propagation method.

Radial basis function neural networks introduced in [31] is other classifier that is employed in this paper. The proposed RBF uses an improved version of particle swarm optimization technique for adjusting centers of the Gaussian functions in the hidden layer of the network. Moreover, optimum steepest descent method is implemented for calculating the synaptic weights of the RBFNN.

3.1. Extracted features

As it was stated before, different image features were used to validate the proposed feature descriptor.

The SIFT approach, for image feature generation, takes an image and transforms it into a "large collection of local feature vectors" [26]. Each of these feature vectors is invariant to any scaling, rotation or translation of the image. This approach shares many features with neuron responses in primate vision. To aid the extraction of these features the SIFT algorithm applies a 4- stage filtering approach [26]:

1. Scale- Space Extrema Detection

This stage of the filtering attempts to identify those locations and scales that are identifiable from different views of the same object. This can be efficiently achieved using a "scale space" function. Further, it has been shown under reasonable assumptions it must be based on the Gaussian function. The scale space is defined by the function:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2}$$

Where * is the convolution operator, $G(x, y, \sigma)$ is a variable-scale Gaussian and $I(x,y)$ is the input image.

Various techniques can then be used to detect stable keypoint locations in the scale-space. Difference of Gaussians is one such technique, locating scale-space extrema, $D(x, y, \sigma)$ by computing the difference between two images, one with scale k times the other. $D(x, y, \sigma)$ is then given by:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3}$$

To detect the local maxima and minima of $D(x, y, \sigma)$ each point is compared with its 8 neighbors at the same scale, and its 9 neighbors up and down one scale. If this value is the minimum or maximum of all these points then this point is an extrema.

2. Keypoint Localization

This stage attempts to eliminate more points from the list of keypoints by finding those that have low contrast or are poorly localized on an edge. This is achieved by calculating the Laplacian value for each keypoint found in stage 1. The location of extremum, z, is given by:

$$z = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x} \tag{4}$$

If the function value at z is below a threshold value then this point is excluded. This removes extrema with low contrast. To eliminate extrema based on poor localization it is noted that in these cases there is a large principle curvature across the edge but a small curvature in the perpendicular direction in the difference of Gaussian function. If this difference is below the ratio of largest to smallest eigenvector, from the 2×2 Hessian matrix at the location and scale of the keypoint, the keypoint is rejected

3. Orientation assignment

This step aims to assign a consistent orientation to the keypoints based on local image properties. The keypoint descriptor, described below, can then be represented relative to this orientation, achieving invariance to rotation. The approach taken to find an orientation is:

- Use the keypoints scale to select the Gaussian smoothed image L, from above
- Compute gradient magnitude, m

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{5}$$

- Compute orientation, θ

$$\theta(x, y) = \tan^{-1}(L(x, y+1) - L(x, y-1), L(x+1, y) - L(x-1, y)) \tag{6}$$

- Form an orientation histogram from gradient orientations of sample points
- Locate the highest peak in the histogram. Use this peak and any other local peak within 80% of the height of this peak to create a keypoint with that orientation
- Some points will be assigned multiple orientations
- Fit a parabola to the 3 histogram values closest to each peak to interpolate the peaks position

4. Keypoint descriptor

The local gradient data, used above, is also used to create keypoint descriptors. The gradient information is rotated to line up with the orientation of the keypoint and then weighted by a Gaussian with variance of 1.5 * keypoint scale. This data is then used to create a set of histograms over a window centered on the keypoint.

Keypoint descriptors typically uses a set of 16 histograms, aligned in a 4×4 grid, each with 8 orientation bins, one for each of the main compass directions and one for each of the mid-points of these directions. This results in a feature vector containing 128 elements.

In this paper a dense version of the SIFT descriptor is applied.

The original LBP operator extracts information invariant to local grayscale variations in the image [37]. It is computed at each pixel location, considering the values of a small circular neighborhood (with radius of R pixels) around the value of a central pixelq_c.

Formally, the LBP operator is defined as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\left(q_p - q_c\right) 2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{7}$$

where P is the number of pixels in the neighborhood, R is the radius. The histogram of these binary numbers is then used to describe the texture of the image.

For the construction of the patch-based LBP, pattern histograms in cells are built. Then the histograms of the LBP patterns from different cells are concatenated to describe the texture of the current moving window. The notation $LBP^{u}_{P,R}$ is used to

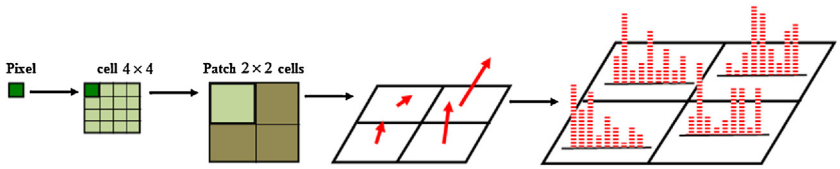

Fig. 5. Visualization of extracting HOG features.

denote LBP feature that takes P sample points with radius R, and the number of 0–1 transitions is no more than u. A pattern that satisfies this constraint is called uniform pattern. For instance, the pattern 0010010 is a non-uniform pattern for LBP^2 and is a uniform pattern for LBP^4 because LBP^4 allows four 0–1 transitions. In this case, instead of original 256 bins, a feature vector of 58 bins is extracted. Moreover, the remaining non-uniform patterns are aggregated to a single bin, the uniform LBP histogram actually contains 59 bins. In this work, using a 8×8 window around each pixel, uniform pattern $\left(LBP_{8,1}^{u_i}\right)$ of each pixel is extracted. These descriptors are of 64-dimensional: $LBP_i = \left[LBP_{8,1}^{u_1}, LBP_{8,1}^{u_2}, \ldots, LBP_{8,1}^{u_{64}}\right]$.

LTP is a recent variant of LBP [48]. The LBP operator is rotation invariant and evaluates the binary difference between the gray value of a pixel x and the gray values of P neighboring pixels on a circle of radius R around x. A problem with conventional LBP is its sensitivity to noise in the near-uniform image regions. The three value encoding scheme of LTP overcomes this problem. In our experiments, the uniform bins with parameters $P = 8$ and $R = 1$ have been used.

As detailed in [60], LDP is a general framework that encodes directional pattern features based on local derivative variations. The LDP templates extract high-order local information by encoding various distinctive spatial relationships contained in a given local region.

As a dense version of the dominating SIFT feature, HOG [6] has shown impressive performance in object detection and recognition. To extract HOG features, each image is partitioned into 4×4 cells. Then for these cells a 1–9 histogram of gradient directions is stacked. Thus, a patch will be described by a $2 \times 2 \times 9 = 36$ dimensional feature vector (Fig. 5).

4. Experimental results and discussion

Many researchers retrieve images by using image classification [53,14,36,41,33]. To this end, first of all, image features are extracted from the images in different categories. Then the features are learned using a classifier. In the next step, user enters a query image and the trained classifier predicts the class of the query image and the most similar images are retrieved from the predicted category. Our approach towards image retrieval is the same.

The experiments in the retrieval scenario are based on traditional image datasets which comprise the semantic-search application. It should be mentioned that in different experiments best results were achieved by $n_x = 10$,$n_y = 10$ and$n_\theta = 15$. The description of the datasets used in this paper is as following.

4.1. Dataset description

4.1.1. Oliva and Torralba (OT) dataset

Oliva and Torralba dataset [34] includes 2688 images classified in 8 categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets and 356 tall buildings images. Note that river and forest scenes are all considered as forest, moreover, there is not a specific sky scene since almost all of the images contain the sky object. These annotations make a higher inter-class variability. Most of the scenes present large intra-class variability. The average size of each image is 250×250 pixels.

4.1.2. Fei-Fei and Perona (FP) dataset

Fei-Fei and Perona dataset [13] contains 13 categories and is only available in grayscale. This dataset consists of the 2688 images (8 categories) of the OT dataset plus: 241 suburb residence, 174 bedroom, 151 kitchen, 289 living room and 216 office images. Note that the presence of man-made images (e.g. bedroom, living room etc.) allows a low inter-class variability since these images have a similar structure. The average size of each image is approximately 250×300 pixels.

4.1.3. Lazebnik, Schmid and Ponce (LSP) dataset

Lazebnik dataset [22] contains 15 categories and, as with FP, is only available in grayscale. This dataset consists of the 13 categories of the FP dataset plus: 315 store and 311 industrial images. The average size of each image is approximately 250×300 pixels.

4.1.4. Wang, Li and Wiederhold (Wang) dataset

Wang dataset [52] contains 1000 images in 10 categories; Africa, Beaches, Building, Bus, Dinosaur, Elephant, Flower, Horses, Mountain and Food. There are 100 images in each category. The images are with the size of 256×384 or 384×256 pixels.

Table 1
Description and values of the parameters used in RBFNN classifier.

Parameters	Description	Considered value
k_1	speed inertial factor	0.1
k_2	the inertia factor variance	0.35
w_0	inertia factor	0.25
C1	Local search coefficient	1.25
C2	Social search coefficient	1.25
Population size	Number of particles	25
T	Number of iterations	500
Number of runs	–	40

Table 2
Parameter Selection of the RBFNN classifier.

Dataset	No. of train	No. of test	No. of hidden neurons	No. of epochs
OT	806	1882	285	60,000
FP	1351	2508	550	150,00
LSP	1570	2250	600	200,000

The performance of the proposed method is analyzed by considering different set of features namely SIFT, HOG, LBP, LDP, LTP and their most effective combinations of them. Finally, the proposed model is compared with other state of the art methods on these datasets.

In order to compare the performance of the proposed CBIR model, mean Average Precision (mAP) is used.

4.2. Image retrieval performance

In order to show the efficiency of the proposed method as well as to reach high rate of accuracy, various experiments were implemented.

Table 1 shows the parameters used for training RBFNN classifier proposed in [31]:

In the case of SVM classifier, as it was stated, standard radial Gaussian kernel was used. The following procedure was investigated to train the SVM classifier:

- Conducting simple scaling on the data
- Considering the RBF kernel $K(x, y) = e^{-\gamma x-y^2}$
- Using cross-validation to find the best parameter C ($C>0$ is the penalty parameter of the error term) and γ
- Using the best parameter C and γ to train the whole training set
- Testing

There are two parameters for an RBF kernel: C and γ . It is not known beforehand which C and γ are best for a given problem; consequently, some kind of model selection (parameter search) must be done. The goal is to identify good (C,γ) so that the classifier can accurately predict unknown data (i.e. testing data). Note that it may not be useful to achieve high training accuracy (i.e. a classifier which accurately predicts training data whose class labels are indeed known). As discussed above, a common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the "unknown" set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation. In K-fold cross-validation, at first, the training set is divided into K subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining K−1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data that are correctly classified.

The cross-validation procedure can prevent the overfitting problem.

A "grid-search" strategy on C and γ is employed so that various pairs of (C,γ) values are tested and the one with the best cross-validation accuracy is picked. It was found that trying exponentially growing sequences of C and γ is a practical method to identify good parameters [28]. Therefore, the following parameters:$C = 2^{-6}, 2^{-5}, \ldots, 2^{16}, \gamma = 2^{-16}, 2^{-15}, \ldots, 2^{-8}$ are tested. In our experiments, it was observed that the best results are achieved by $(C,\gamma)=(25,20)$.

To evaluate the performance of the proposed method, first the training and testing sets were randomly selected by setting 40% of data as the training set and rest as testing set.

To do parameter tuning for SVM (C,γ), 10-fold cross validation was used to find the best pair which obtained best performance of SVM on training data. Within each fold, as it was mentioned, grid search was adopted. Then these parameters were passed to SVM to test the testing fold. Due to using random selection for training and testing data, we did the entire process 10 times and finally reported the average classification accuracy.

Table 2 shows the parameters used by RBFNN on each dataset:

Table 3
Performance of different features on OT dataset.

Our Proposed Method	Precision using MLP Classifier	Precision using SVM Classifier	Precision using RBFFNN Classifier	Length of the feature vector
HOG	76.4	78.3	81.5	1500
LBP	75.1	76.3	88.9	1500
SIFT	**86.4**	**88.4**	**91.2**	1500
LTP	79.6	81.4	83.4	1500
LDP	80.6	82.5	84.6	1500
SIFT-LTP	88.3	90.5	92.2	3000
SIFT-LDP	**90.5**	**92.4**	**94.7**	3000

Table 4
Performance of different features on FP dataset.

Our Proposed Method	Precision using MLP Classifier	Precision using SVM Classifier	Precision using RBFFNN Classifier	Length of the feature vector
HOG	57.2	59.2	61.7	1500
LBP	54.7	55.9	62.4	1500
SIFT	**83.4**	**85**	**86.4**	1500
LTP	71.8	72.2	73.8	1500
LDP	75.4	75.9	77.3	1500
SIFT-LTP	85.6	86.3	87.1	3000
SIFT-LDP	**86.3**	**87.4**	**89.4**	3000

Table 5
Performance of different features on LSP dataset.

Our Proposed Method	Precision using MLP Classifier	Precision using SVM Classifier	Precision using RBFFNN Classifier	Length of the feature vector
HOG	53.4	55.9	58.3	1500
LBP	50.8	53.1	54.6	1500
SIFT	**82.3**	**83.7**	**86.4**	1500
LTP	68.5	69.4	71.6	1500
LDP	67.4	72.1	74.4	1500
SIFT-LTP	84.1	84.9	86.5	3000
SIFT-LDP	**85.3**	**86.4**	**87.6**	3000

Table 6
Performance of different features on Wang dataset.

Our Proposed Method	Precision using MLP Classifier	Precision using SVM Classifier	Precision using RBFFNN Classifier	Length of the feature vector
HOG	77.6	79.3	82.9	1500
LBP	74.9	76.2	80.6	1500
SIFT	**82.4**	**84.6**	**86.4**	1500
LTP	79.7	77.8	79.4	1500
LDP	80.1	82.7	83.9	1500
SIFT-LTP	81.6	84.3	86.7	3000
SIFT-LDP	**83.9**	**85.7**	**88.5**	3000

For each dataset, the effectiveness of different features and classifiers can be seen in Tables 3–6 . The bolded numbers in these Tables show the results of the most effective method.

It should be noted that, SIFT-LDP and SIFT-LTP are obtained by directly concating SIFT with LDP features and SIFT with LTP features, respectively.

Important and interesting conclusions can be drawn from Tables 3–6. In all the experiments, using the proposed descriptor, SIFT features reaches the best performances followed by LDP and LTP features. Moreover, by using different classifiers, different results are obtained. The experiments show that the proposed RBFNN in [31] reaches higher accuracy in comparison with SVM and MLP classifiers. In addition, SVM classifier overcomes MLP classifier. Another interesting observation is that by fusing SIFT and LDP features, better performances are achieved on the four datasets. The reason is that these features are complement to each other. SIFT method fails in the homogeneous areas where gradient of the image is small. In order to overcome this problem higher order texture features extracted by LDP can be a good remedy and a complement to SIFT features.

Fig. 6 reports the confusion matrix of OT data set using SIFT-LDP features and RBFNN classifier. The class numbers are as follows: 1 = 'coast', 2 = 'forest', 3 = 'highway', 4 = 'inside the city', 5 = 'mountain', 6 = 'open country', 7 = 'street', and 8 = 'tall buildings'.

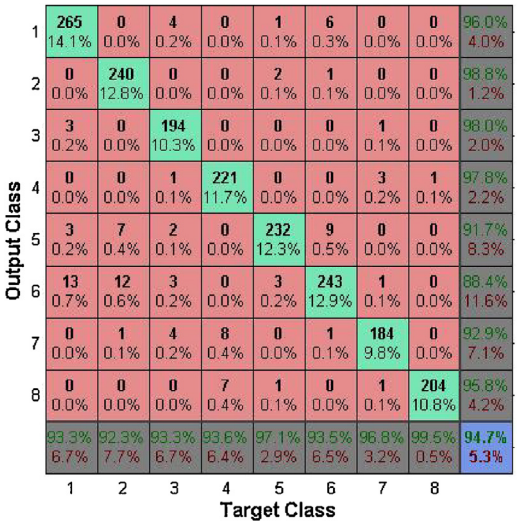

Fig. 6. Confusion matrix of the proposed SIFT-LDP descriptor on OT dataset.

Fig. 7. Confusion matrix of the proposed SIFT-LDP descriptor on Wang dataset.

This confusion matrix shows that RBFNN classifier cannot properly recognize coast and forest categories with open country. Furthermore, open country is confused with coast and mountain categories. There is also a conflict between tall buildings and inside city categories where there are many strong edges in the images.

Figs. 7–9 demonstrate the confusion matrices of FP, Wang and LSP datasets, respectively.

We further compare our proposed method with the state-of-the-art methods in the literature. The bolded numbers in Tables 7–10 show the result of the most effective method.

5. Conclusion

This paper investigates CBIR problem by introducing a new image descriptor. The proposed descriptor can be used by SIFT, LBP, LDP, LTP and HOG features. The aim of the proposed descriptor is capturing higher level of semantic by extending the ability of pixel-based descriptors from pixel level to objects, segments or patches. Hence, the descriptor can meaningfully capture the appearance and shape of the local object within an image by calculating the distribution of intensity gradients, edge directions or texture features. The proposed method was tested on four datasets and promising results were achieved by combining SIFT and LDP features. In future, the proposed descriptor will be used in the framework of bag of features on large datasets.

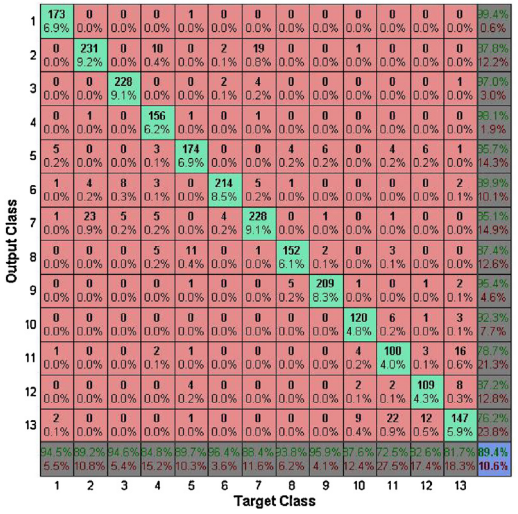

Fig. 8. Confusion matrix of the proposed SIFT-LDP descriptor on FP dataset.

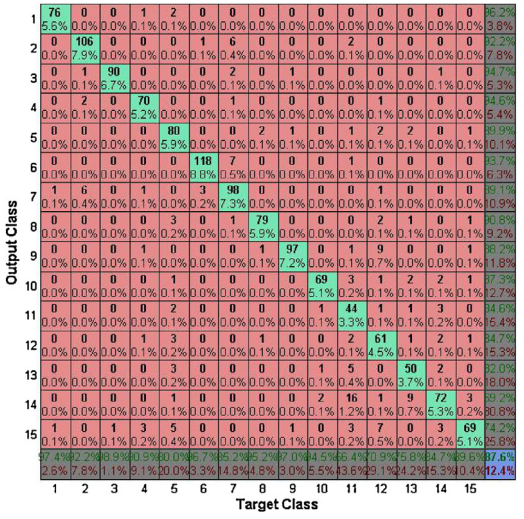

Fig. 9. Confusion matrix of the proposed SIFT-LDP descriptor on LSP dataset.

Table 7
Performance comparison on OT dataset.

Method	mean Average Precision (mAP%)
SIFT-LDP	**94.7**
PHOG	76.4
GIST	80.4
Gabor LBP	79.9
Gradient LBP	73.3
SIFT	86.6
Proposed method in [34]	89
Proposed method in [3]	91.49
Proposed method in [59]	91.1
Proposed method in [50]	90.2
Proposed method in [61]	89.2

Acknowledgments

The authors are grateful to the anonymous reviewers for the insightful comments and constructive suggestions. Part of this research has been funded by the Iranian Research Institute for Information Science and Technology (IranDoc) (No. TMU92-03-44).

Table 8
Performance comparison on FP dataset.

Method	mean Average Precision (mAP%)
SIFT-LDP	**88.6**
PHOG	71.2
GIST	76.1
Gabor LBP	72.3
Gradient LBP	68.9
SIFT	83
Proposed method in [13]	65.2
Proposed method in [22]	74.7
Proposed method in [59]	85.9 ± 1.3
Proposed method in [39]	87.63 ± 2.3
Proposed method in [61]	86.1

Table 9
Performance comparison on LSP dataset.

Method	mean Average Precision (mAP%)
SIFT-LDP	**88.3**
PHOG	63.5
GIST	70.6
Gabor LBP	70.8
Gradient LBP	65.4
SIFT	80.1
Proposed method in [22]	(81.4 ± 0.5)
Proposed method in [3]	85.62%
Proposed method in [59]	(83.7 ± 1.3)
Proposed method in [39]	(85.16 ± 1.62)
Proposed method in [40]	(87.39 ± 2.48)
Proposed method in [61]	84.2%

Table 10
Performance comparison on Wang dataset.

Method	mean Average Precision (mAP%)
SIFT-LDP	**88.5**
Proposed Method in [19]	39
Proposed Method in [16]	37.2
Proposed Method in [25]	53.1
Proposed Method in [8]	55.6
Proposed Method in [9]	58.9

References

[1] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives? IEEE Trans. PAMI 35 (8) (2013) 1798–1828.
[2] Y. Boureau, F. Bach, Y. LeCun, Learning mid-level features for recognition, Proc. IEEE Conf. Comput. Vision Pattern Recogn. (2010) 2559–2566.
[3] A. Bolovinou, I. Pratikakis, S. Perantonis, Bag of spatio-visual words for context inference in scene classification, Pattern Recogn. 46 (3) (2013) 1039–1053.
[4] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011), 27:1-27:27.
[5] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2) (2008) 1–60.
[6] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2005) 886–893.
[7] T. Deselaers, D. Keysers, H. Ney, Features for image retrieval: an experimental comparison, Inf. Retr. 11 (2) (2008) 77–107.
[8] M.E. ElAlami, A novel image retrieval model based on the most relevant features, Knowl.-Based Syst. 24 (2011) 23–32.
[9] M.E. ElAlami, A new matching strategy for content based image retrieval system, Appl. Soft Comput. 14 (2014) 407–418.
[10] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 Results", 2012 http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html.
[11] H. Farhidzadeh, D.B. Goldgof, L.O. Hall, R.A. Gatenby, R.J. Gillies, M. Raghavan, Texture feature analysis to predict metastatic and necrotic soft tissue sarcomas, In Systems, Man, and Cybernetics (SMC), IEEE International Conference on (2015) 2798–2802.
[12] H. Farhidzadeh, J.Y. Kim, J.G. Scott, D.B. Goldgof, L.O. Hall, L.B. Harrison, Classification ofprogression free survival with nasopharyngeal carcinoma tumors, In SPIE Medical Imaging (2016), 97851I.
[13] Fei-Fei L., Perona P., 2005, A Bayesian hierarchical model for learning natural scene categories, In Proc.CVPR.
[14] J. Félix Serrano-Talamantes, C. Avilés-Cruz, J. Villegas-Cortez, J.H. Sossa-Azuela, Self organizing natural scene image retrieval, Expert Syst. Appl. 40 (2013) 2398–2409.
[15] B. Fernando, E. Fromont, T. Tuytelaars, Effective use of frequent itemset mining for image classification, ECCV 7572 (2012) 214–227.
[16] A. Gelzinis, A. Verikas, M. Bacauskiene, Increasing the discrimination power of the co-occurrence matrix-based features, Pattern Recogn. 40 (2007) 2367–2372.
[17] M. Heikkilä, M. Pietikäinen, C. Schmid, Description of interest regions with local binary patterns, Pattern Recogn. 42 (2009) 425–436.
[18] R. Hu, J. Collomosse, A performance evaluation of gradient field HOG descriptor for sketch based image retrieval, Comput. Vision Image Underst. 117 (2013) 790–806.
[19] P.W. Huang, S.K. Dai, Image retrieval by texture similarity, Pattern Recogn. 36 (3) (2003) 665–679.

[20] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature coding in image classification: a comprehensive study, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 493–506.
[21] Kolmogorov, A.N., 1975. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition.Dokl.Akad.Nauk SSSR 144, 679-681; Amer. Math. Soc. Tansl. 28, 55-59.
[22] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2 (2006) 2169–2178.
[23] Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. et al. (2012). Building high-level features using large scale unsupervised learning. In ICML.
[24] M.S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: state of the art and challenges, ACM Trans. Multimedia Comput. Commun. Appl. 2 (1) (2006) 1–19.
[25] C.-H. Lin, R.-T. Chen, Y.-K. Chan, A smart content-based image retrieval system based on color and texture feature, Image Vision Comput. 27 (2009) 658–665.
[26] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[27] P.L.G. Martens, R. Walle, Bridging the semantic gap using human vision system inspired features, Self-Organ. Maps (2010).
[28] Meyer, D., Wien, F.T., 2015. Support vector machines. The Interface to libsvm in package e1071.
[29] S.L. Michael, S. Nice, D. Chababe, J. Ramsesh, Content-based multimedia information retrieval: state of the art and challenges, ACM Trans. Multimed. Comput. Commun. Appl. 2 (1) (2006) 1–19.
[30] M.F. Moller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Netw. 6 (4) (1993) 525533.
[31] G.A. Montazer, D. Giveki, An improved radial basis function neural network for object image retrieval, Neurocomputing 168 (2015) 221–233.
[33] S. Mukhopadhyay, J.K. Dash, R.D. Gupta, Content based texture image retrieval using fuzzy class membership, Pattern Recogn. Lett. 34 (2013) 646–654.
[34] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.
[35] D. Parikh, Recognizing jumbled images: the role of local and global information in image classification, IEEE International Conference on Computer Vision (ICCV) (2011) 519–526.
[36] S-Sung Park, K-Kwang Seo, D-Sik Jang, Expert system based on artificial neural networks for content-based image retrieval, Expert Syst. Appl. 29 (2005) 589–597.
[37] M. Pietikäinen, T. Ahonen, V. Takala, Block-based methods for image retrieval using local binary patterns, In: Proceedings of the 14th Scandinavian Conference on Image Analysis (2005) 882–891.
[38] A. Pradnya, Vikhar, content-based image retrieval (CBIR): State-of-the-art and future scope of research, IUP J. Inf. Technol. 6 (2) (2010) 64–84.
[39] J. Qin, N.H. Yung, Scene categorization via contextual visual words, Pattern Recogn. 43 (5) (2010) 1874–1888.
[40] J. Qin, N.H. Yung, Feature fusion within local region using localized maximum-margin learning for scene categorization, Pattern Recogn. 45 (4) (2012) 1671–1683.
[41] S. Sadek, A. Al-Hamadi, B. Michaelis, U. Sayed, Cubic-splines neural network- based system for Image Retrieval, 16th IEEE International Conference on Image Processing (ICIP) (2009) 273–276.
[42] Sanchez, J., Perronnin, F., Mensink, T., Verbeek J. 2013. Image Classification with the Fisher Vector: Theory and Practice. [Research Report] RR-8209.
[43] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, Proc. IEEE Conf. Computer Vision and Pattern Recognition (2005).
[45] Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S., 2009. Human detection using partial least squares analysis, in: ICCV, pp. 24–31.
[46] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Trans. Pattern Anal. Mach. Intell. (2007) 1–19.
[47] N.S. Vassilieva, Content-based image retrieval methods, Program. Comput. Softw. 35 (3) (2009) 158–180.
[48] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, Anal. Modell. Faces Gestures LNCS 4778 (2007) 168–182.
[49] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with super pixels, ECCV 6315 (2010) (2010) 352–365.
[50] Wang, Y., Gong, S., 2007. Conditional Random Field for Natural Scene Categorization.In BMVC (pp. 1-10).
[52] J.A. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 947–963.
[53] W. Tak Wong, Sh.-Hsun Hsu, Application of SVM and ANN for image retrieval, Eur. J. Oper. Res. 173 (2005) 938–950.
[55] J. Yu, Z. Qin, T. Wan, X. Zhang, Feature integration analysis of bag-of-features model for image retrieval, Neurocomputing 120 (2013) 355–364.
[56] J. Zhang, Y. Barhomi, T. Serre, A new biologically inspired color image descriptor, ECCV (2012).
[57] Y. Zheng, X. Huang, S. Feng, An image matching algorithm based on combination of SIFT and rotation invariant LBP, J. Comput.-Aided Des. Comput. Graph. 2 (2010) 286–292.
[58] S. Zheng, Z. Tu, A. Yuille, Detecting object boundaries using low-,mid-, and high-level information, IEEE Conference on Computer Vision and Pattern Recognition (2007) 1–8.
[59] S. Zhang, Q. Tian, G. Hua, Q. Huang, W. Gao, ObjectPatchNet: towards scalable and semantic image annotation and retrieval, Comput. Vision Image Underst. 118 (2014) 16–29.
[60] B. Zhang, Y. Gao, S. Zhao, J. Liu, Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor, Image Processing IEEE Trans. 19 (2) (2010) 533–544.
[61] L. Zhou, Z. Zhou, D. Hu, Scene classification using a multi-resolution bag-of-features model, Pattern Recogn. 46 (1) (2013) 424–433.