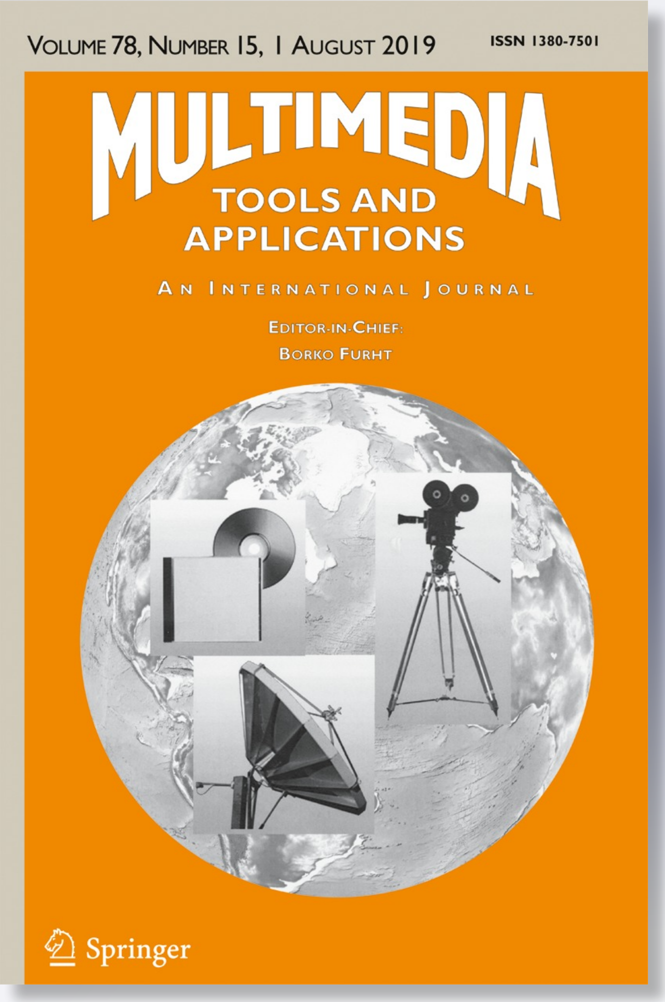
Content-based image retrieval by combining convolutional neural networks and sparse representation

Amir Sezavar, Hassan Farsi & Sajad Mohamadzadeh

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501
Volume 78
Number 15

Multimed Tools Appl (2019)
78:20895-20912
DOI 10.1007/s11042-019-7321-1



Multimedia Tools and Applications (2019) 78:20895–20912
https://doi.org/10.1007/s11042-019-7321-1

Content-based image retrieval by combining convolutional neural networks and sparse representation

Amir Sezavar[1] • Hassan Farsi[1] • Sajad Mohamadzadeh[1]

Received: 21 February 2018 / Revised: 28 January 2019 / Accepted: 31 January 2019 /
Published online: 7 March 2019

Abstract
As stored data and images on memory disks increase, image retrieval has a necessary task on image processing. Although lots of researches have been reported for this task so far, semantic gap between low level features of images and human concept is still an important challenge on content-based image retrieval. For this task, a robust method is proposed by a combination of convolutional neural network and sparse representation, in which deep features are extracted by using CNN and sparse representation to increase retrieval speed and accuracy. The proposed method has been tested on three common databases on image retrieval, named Corel, ALOI and MPEG7. By computing metrics such as P(0.5), P(1) and ANMRR, experimental results show that the proposed method has achieved higher accuracy and better speed compared to state-of-the-art methods.

Keywords Content-based image retrieval · Deep learning · Convolutional neural networks · Sparse representation

1 Introduction

Nowadays, Content-Based Image Retrieval (CBIR) is an important task on image processing and artificial intelligence. As a reason of huge progress in digital imaging and internet usage, amount of digital images saved on hard disks has highly increased. Therefore, to manage and search the images and databases, a robust system is needed. Image retrieval, which means searching in a set of images in order to represent the nearest ones to query, was introduced in 1970, named Text-Based Image Retrieval (TBIR) [27]. After that, content based image retrieval was indicated in 1990. TBIR searches the images using the text that the user applies on the system. Since text

✉ Hassan Farsi
hfarsi@birjand.ac.ir

[1] Department of Electrical and Computer Engineering, University of Birjand, Shahid Avini Highway, Birjand, Iran

searching has many problems, such as misspelling or words low ability of explaining emotions accurately, researches were involved to introduce methods which use contents of image and features to retrieve images. CBIR has been widely used in different fields such as medicine, digital imaging and libraries, crime excluding and other areas which use digital images [34].

Feature extraction is the main part of content-based image retrieval systems. In mostly related methods, content of an image is summarized to a feature vector that its size is smaller than the related image. Since images are compared together according to their features, such as color, texture, shape and etc., the accuracy of a CBIR system mainly depends on the feature vectors which are extracted from different images. This low-level hand crafted features are unable to explain any scene completely and it causes low accuracy in retrieval. On the other hand, if more features are extracted, the system will become more complex and time consuming. Thus, by extracting low level features, there is a tradeoff between retrieval time and the accuracy of performance. In this paper, we use convolutional neural network for extracting deep features from images to reduce the semantic gap between low level features and human concept and then by using sparse representation, the nearest images to query image are represented quickly. Image retrieval system is illustrated in Fig. 1.

2 Related works

As a common knowledge, although color-based features are unable to describe images effectively, for years, color histogram was widely used in image retrieval [17]. In this area, image histogram chromaticity was used and many color spaces and descriptors such as Color Layout Descriptor (CLD), Dominant color Descriptor (DCD) and Scale Color Descriptor (SCD) were indicated in [11, 19, 35], respectively. Texture feature extraction is another impressive method to describe an image. It has been revealed that using a mixture of color and texture feature can improve the retrieval accuracy [16]. In [4], Hue-Maximum-Minimum-Difference (HMMD) color space and Hadamard Discrete Wavelet Transform (HDWT) were

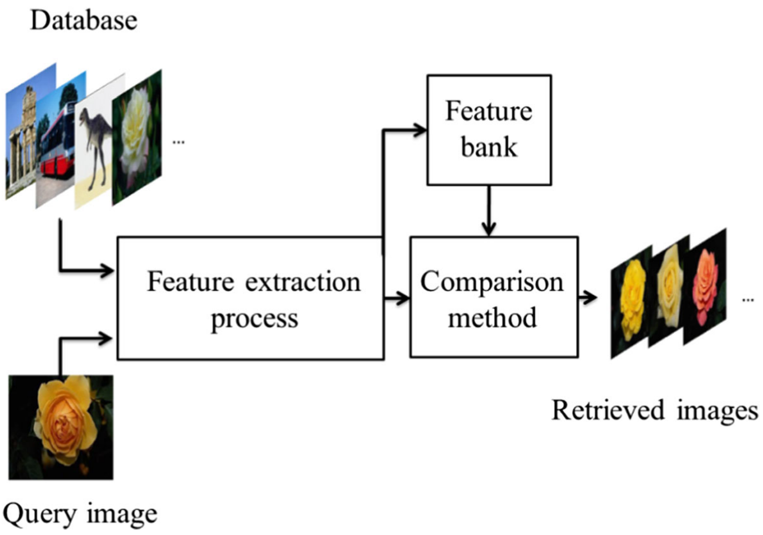


Fig. 1 Mechanism of image retrieval system

combined to represent and retrieve images. Also a combination of HDWT, discrete cosine transform and principal component analysis has been proposed in [5] which could reduce feature vector size to increase the retrieval efficiency. Using filter has shown an appropriate method to extract texture features of image, like Gabor filter used in [21].

Although using color and texture features play a successful role in image retrieval, it is unable to satisfy user's demand; accordingly, researches try to use new artificial intelligence based methods to describe images. Domonkos and Sziranyi [36] introduced a new image retrieval system using convolutional neural networks (CNNs) and hash function, in which CNN and hash function are learnt concurrent and the images are converted to a binary small-size vector. Although reducing the feature vector size can increase the speed of retrieval, they achieve a maximum of 72% precision of retrieval on Oxford database [25]. Another mixture of CNN and hash coding for image retrieval was presented in [24] in which, Peng and Li used a deeper CNN, called VGGNet [28]. Their model achieved 85.5% precision on CIFAR-10 database [12] while making more layers in the CNN causes more computation and lowers the speed. In [26], Qayyum et al. used the CNN for medical image retrieval. Their obtained results show that the CNN based image retrieval can be used in medical tasks. A combination of two CNNs was presented in [18] by H liu et al. in which features are extracted from two different CNNs for retrieval. Their system advantage is obtaining 94.8% precision on Corel (http://caffe.berkeleyvision.org/) database, though using two CNNs requires high computation time.

Tao et al. introduced a tensor based facial recognition system [33] which considered input images as two- order tensors. Their system could achieve effective performance compared to other facial recognition systems. In addition, a novel approach for person re-identification task was proposed by Tao et al [32] which achieved superior performance on relevant challenge databases. These excellent approaches demonstrate that we can improve the retrieval performance by improving feature extraction methods.

3 Theoretical background

In this section, an image retrieval system is proposed which uses the CNN and sparse representation to extract appropriate features and provide fast retrieval. In the following subsections, firstly, a brief of the CNN and sparse representation is provided and then the proposed method is described.

3.1 Convolutional neural networks

Deep learning is a new approach in machine learning, in which high-level features can be extracted from input data through hierarchical layers [8]. It has been shown that deep learning provides an excellent performance in data processing, due to achieving higher accuracy in image, video [1], natural language and audio processing [15]. Among some of the deep learning algorithms, convolutional neural networks are appropriate for image processing. As mentioned in [8], CNNs are constructed specially for two-dimensional (2D) data, like image and video, and they also have a superior accuracy on image processing.

CNNs are important and considered as a useful field of deep learning method which is widely used in image processing. They are special kind of neural networks in which some learnable filter coefficients are convolved with input images during a hierarchical algorithm; that was firstly introduced by LeCun et al. for handwritten recognition [14].

What makes deep learning different from previous processing methods is that the data are fed to the system directly in order to extract features, while in traditional processing hand-crafted features were fed to algorithm for the purpose of processing or classifying, like artificial neural networks and other classifiers as shown in Fig. 2.

A CNN consists of mainly three kinds of layers including convolutional, pooling and fully connected, while additional layers may be used in it. In the convolutional layer, filters are convolved by image pixels in order to create feature maps. If more filters are considered, more feature maps will be created. After the convolution, pooling operator, either max or mean pooling, is used to reduce feature map's size. After many continuous layers, a fully connected layer is considered to convert 2D features to a vector for classification. After lenet-5 was demonstrated by LeCun, Alex Krizhevski et al. constructed a novel CNN named AlexNet [13], which causes a huge revolution in image classification. After that, although many supplementary and deeper networks were implemented, such as VGGnet [29], GoogLeNet [31] and etc.; Alexnet is still known as the basic convolutional neural network. Thus, we decided to use AlexNet for feature extraction in this paper. Recently, most done researches have demonstrated the efficiency of using CNNs on image processing tasks. For example, Gou et al. presented a novel facial expression recognition system by introducing a new deep learning network [7]. Their system achieved superb performance compared to state-of-the-art systems.

3.2 Sparse representation

Sparse representation has been one of the most efficient methods which is able to represent a large signal by few non-zero coefficients. A signal with length of $'n'$ can be shown by $'k'$ nonzero coefficients while $k \ll n$. This means that the signals can be compressed and the number of measurements can be reduced. Therefore, it is an efficient representation that has been used in many fields of researches specially image processing and speech recognition. The core of representing a signal by a linear combination of some basic coefficients is to solve:

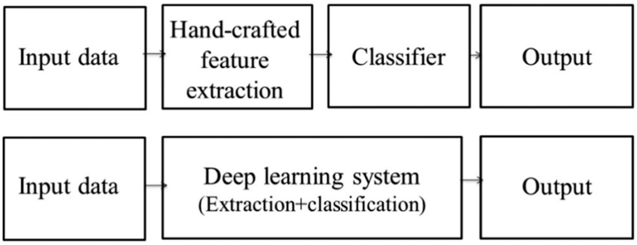

Input data
Hand-crafted feature extraction
Classifier
Output
Input data
Deep learning system (Extraction+classification)
Output

Fig. 2 Advantage of using deep learning in processing systems

$$b = Ax = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \tag{1}$$

Where $b \in R^m$ is the main signal and A is a dictionary in which, $\alpha_i \in R^m$ $(1 \le i \le n)$ are called atoms. It can be simply shown that the solution of the Eq. (1) depends on relation between $'n'$ and $'m'$ and has one unique solution when $'n'$ is equal to $'m'$. When $'n'$ is greater than $'m'$ $(n > m)$, the equation does not have a unique solution. Thus, a condition is considered for the equation to provide the specific solution as following:

$$P_j = \min J(x) \quad subject.to \quad b = Ax \tag{2}$$

Depending on J(x) function, the problem is classified into many forms. To find sparsest solution, J(x) should be the l_0-norm of x, which is defined as [39]:

$$\|x\|_0 = \lim_{p \to 0} \sum_{i=1}^{n} |x_i|^p \tag{3}$$

$$P_0 : \min \|x\|_0 \quad subject.to \quad b = Ax \tag{4}$$

This represents sum of nonzero elements of the vector x. Although solving this problem with l_0-norm function can provide sparsest result, this is non-convex problem and difficult to be solved. It is known that the nearest solution to l_0 is using l_1-norm instead. Thus, the problem is converted to convex optimization [38]:

$$P_1 : min \|x\|_1 \quad subject.to \quad b = Ax \tag{5}$$

As mentioned in [38], signals are usually found with noise in practice, so the problem is changed to Eq. (6) and a well-known problem is given by Eq. (7):

$$p_{1,2} : min \|x\|_1 \quad subject.to \quad \|b - Ax\|_2 \ll \varepsilon \tag{6}$$

$$QP_\lambda : \min \frac{1}{2} \|b - Ax\|_2^2 + \lambda \|x\|_1 \tag{7}$$

Although many methods either classical or proximal were introduced to solve convex Eq. (7) such as Smoothed L0 (SL0), Homotopy method, Fast Iterative Soft-Thresholding Algorithm (FISTA) and Approximate Message Passing (AMP) [23, 31, 39], it has been shown that Dual Augmented Lagrangian Method (DALM) and Primal Augmented Lagrangian Method (PALM) are more efficient than other methods [39]. Thus, these methods are selected to solve optimization problem in this paper. The pseudo code for PALM method is illustrated in Algorithm 1 which is used in [38] for face recognition.

[31]

Algorithm 1. The pseudo codes of an instance of PALM method to solve $L_\lambda(x,e,z) = \|x\|_1 + \|e\|_1 + \frac{\lambda}{2}\|b - Ax - e\|_2^2 + \theta^T(b - Ax - e)$. [31]

```
1. **input**: $b \epsilon R^m$ , $A \epsilon R^{m*n}$ .
Initialize: $x_1 = 0$ , $e_1 = b$ , $z_1 = 0$.
2. **While** not converged, **do**: (for k=1, 2, …)
  3.$e_{k+1} = \text{shrink}\left(b - Ax_k + \frac{1}{\lambda}z_k, \frac{1}{\lambda}\right)$;
  4.$t_1 = 1, z_1 = x_k, w_1 = x_k$;
  5. **While** not converged, **do**: (l=1, 2, …)
    6. $w_{l+1} = \text{shrink}(z_l + \frac{1}{L}A^t\left(b - Az_l - e_{k+1} + \frac{1}{\lambda}\theta_k, \frac{1}{\lambda L}\right)$;
    7. $t_{l+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_l^2})$;
    8.$z_{l+1} = w_{l+1} + \frac{t_l - 1}{t_l + 1}(w_{l+1} - w_l)$;
  9. **End while**;
  10. $x_{k+1} = w_l, \theta_{k+1} = \theta_k + \lambda(b - Ax_{k+1} - e_{k+1})$;
11. **End while**;
12. $x^* = x_k, e^* = e_k$.
```

4 The proposed method

The proposed model as same as any retrieval algorithm has two main steps: feature extraction and online retrieval which will be explained in two subsections. Offline training consists of network fine-tuning for each database and computing feature matrix, while online step is getting query image from the user and representing the nearest images using a combination of sparse representation and CNN. The maim duty of the proposed method is converting the image retrieval problem to a dual convex problem which is solved by sparse representation, while the dictionary of sparse representation is provided by CNN which will be trained in the offline step.

4.1 Offline step

For extracting features, a pre-trained AlexNet is considered to be trained on the datasets which will be mentioned in Section 5. Training steps are listed as following:

a) In pre-processing step, every image should be resized to 224 × 224 pixels in order to be prepared for CNN.
b) Fine-tune AlexNet for each database.
c) Trained network in an iterative process.

There are two reasons for fine-tuning a pre-trained network, instead of training it completely. The first one is that deep convolutional neural networks have a huge amount of parameters which should be trained and to train these amounts of parameters, large scale databases are needed. The second one is concerning about over fitting when the training process diverges. Hence, fine tuning is a suitable option when our databases do not contain large amount of samples. Therefore, a pre-trained, AlexNet which has been trained on 1.2 million images, is used in this research. To fine-tune AlexNet on each database, we consider earlier layers fixedly, while last two fully connected layers will be trained during the fine-tune process.

After training, filter parameters are trained to classify images to their classes. We need a feature vector to be extracted from network layer. Although the features could be extracted from any layer of networks, the content of last layer is considered as feature vector in this paper because of two reasons:

- Last layer has lower size compared to other layers (i.e. 4096 in third fully-connected) and its size depends on the number of classes in each database which is usually less than 100. Thus the computational time is reduced.
- Features in last layer will be fed to a classifier (i.e. softmax in AlexNet) and they consist of information about dependency of images to each class. Therefore closer image can be retrieved.

Feature vector is calculated for each image in database, and we have a bank of feature vectors named feature matrix. Therefore, two necessary things are available after training step: trained parameters of the networks and feature matrix which will be used in online retrieval. It is noticeable that if we need to use the proposed method for a new database, it is just needed to train or fine-tune the CNN on new database in the offline step and the structure of the model would be the same for any database.

4.2 Online retrieval using sparse representation

Feature vector of query image is extracted from the last layer of CNN as described in previous section. Although a simple way to find nearest images to query image sounds computing the distance (i.e. Euclidian distance) between the query feature vector and all feature vectors and sort images that have less distance compared to others, whoever, it is not official and it is time consuming specially for large databases. Therefore, instead of computing all distances, a sparse representation is used to retrieve similar images, as following:

First, feature matrix which has been created in offline step, is considered as dictionary $'\boldsymbol{F}'$: $F_i = [a_{i,1}; a_{i,2}; \dots; a_{i,m}] \in R^{m \times n}$ in which, $a_{i,j}$represents the j-th feature of the i-th image, extracted using CNN. On the other hand, we create the dictionary using the CNN and use it as a sparse dictionary. So, deep features extracted by the CNN can create meaningful atoms. Thus:

$$\boldsymbol{F} = [F_1; F_2; \dots; F_n] \in R^{m \times n} \tag{8}$$

After that, feature vector of the query image named $'q'$ is extracted by using CNN trained accurately in offline step. Now, a feature matrix $'F'$ and a query feature vector $'q'$ are available. Consider query feature vector can be represented by a linear combination of some feature vectors from Dictionary $'F'$. To find minimum number of coefficients, the following problem is defined as:

$$min\|x\|_1 \quad subject.to \ \ \|q - Fx\|_2 \ll \varepsilon \tag{9}$$

To solve this optimization problem, dual lagrangian function is defined:

$$L_\lambda(x, z) = \|x\|_1 + \frac{\lambda}{2}\|q - Fx\|_2^2 + Z^T(q - Fx) = 0 \tag{10}$$

In which, Z is called dual variable. In [38], it has been shown that solving dual problem with ALM principles provides appropriate results to solve optimization problems. Thus, in order to solve the problem using DALM, it can be written as dual minimization. The dual problem is given by:

$$min_{\lambda, z} - q^T \lambda - x^T\left(z - F^T \lambda\right) + \frac{\beta}{2}\left\|z - F^T \lambda\right\|_2^2 \ \ subject.to \ z \epsilon \left\{\|x\|_\infty < 1\right\} \tag{11}$$

To solve this minimization problem, $'x'$, $'z'$ $and'\lambda'$ should be obtained, alternatively. In each step, minimization is performed with respect to one variable while two other variables are considered constant. Based on [38], three equations for updating variables in the l-th step are used as following:

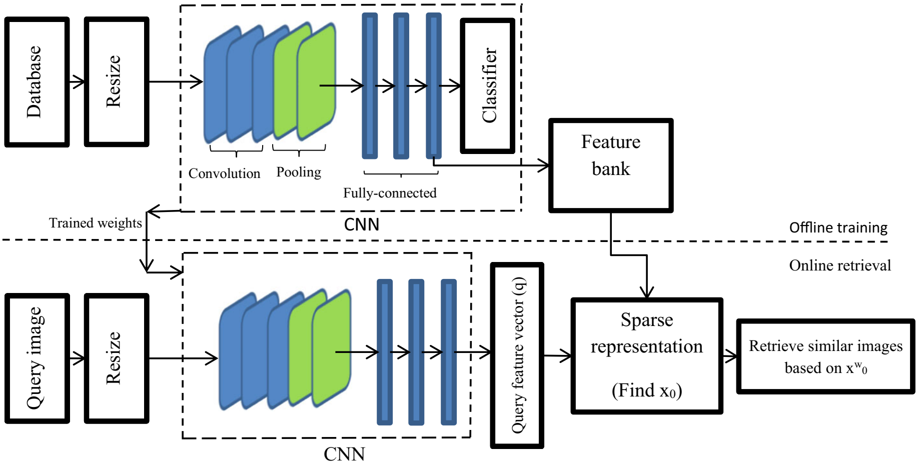


Fig. 3 Flowchart of the proposed system on offline training and online retrieval

$$z_{l+1} = sgn\left(F^T\lambda + \frac{x}{\beta}\right) \times \min\left\{1, \left|F^T\lambda + \frac{x}{\beta}\right|\right\} \tag{12}$$

$$\lambda_{l+1} = \frac{Fz_{l+1} - \frac{Fx_l - q}{\beta}}{\left(FF^T\right)} \tag{13}$$

$$x_{l+1} = x_l - \beta \times \left(z_{l+1} - F^T\lambda_{l+1}\right) \tag{14}$$

If stop condition is satisfied, the obtained x is called x_0. Then we separate elements of $'A'$ and $'x_0'$ into $'k'$ classes for each database which contains $'k'$ clusters:

$$x_0 = \left[\underbrace{a_1;\ a_2; \dots;}_{1} \underbrace{a_j; \dots;}_{2} \dots; \underbrace{a_n}_{k}\right] = \left[x_{0,1}; x_{0,2}; \dots; x_{0,k}\right]$$

$$\boldsymbol{F} = \left[\underbrace{F_1, F_2, \dots,}_{1} \underbrace{F_j, \dots,}_{2} \dots, \underbrace{F_n}_{k}\right] = \left[M_1, M_2, \dots, M_k\right]$$

After that $C_i = M_i x_{0,i} \in R^m$ is defined by using specified elements in the previous step and error vector $'E'$ is calculated by computing Euclidian distance between the query feature vectors $'q'$ and the $'C_i'$ by:

$$e_i = \sqrt{\sum_{l=1}^{m} \left(q_l - C_{il}\right)^2} \tag{17}$$

$$E = [e_1; e_2; \dots; e_k] \tag{18}$$

To determine the value of the elements of x_0 with respect to the query image, each element is weighted by the inverse of its error as following:

$$x_0^w = \left[\frac{x_{0,1}}{e_1}; \frac{x_{0,2}}{e_2}; \dots; \frac{x_{0,k}}{e_k}\right] \tag{19}$$

Table 1 Specifics of the network used in the paper

Layer name	Spesifics
Conv 1	96 filters of size 11 × 11 and stride 4
Pool 1	a window of size 3 × 3 and stride 2
Conv 2	256 filters of size 5 × 5 and stride 1
Pool 2	a window of size 3 × 3 and stride 2
Conv 3	384 filters of size 3 × 3 and stride 1
Conv 4	384 filters of size 3 × 3 and stride 1
Conv 5	256 filters of size 3 × 3 and stride 1

Finally, the relevant images are determined and presented for user as retrieved images by sorting x_0^w. The flowchart of the proposed system is illustrated in Fig. 3 and the pseudo code of the proposed algorithm is shown in algorithm 2 as following:

Algorithm 2.Pseudo code of online image retrieval system

```
Get query image from user;
resize query to 224× 224;
fed query image to CNN in order to extract last layer's feature, query feature vector;
consider feature matrix as dictionary F:
$\boldsymbol{F} = [F_1; F_2; \dots; F_n] \in R^{m\times n}$;
q = query feature vector;
solve $(p1)$: $min\ \|x\|_1\ \ subject\ to\ b = Ax$ by DALM algorithm and find x0:
  Create augmented Lagrange problem;
  $min\|x\|_1\ \ \ subject.to\ \ \|q - Fx\|_2 \ll \varepsilon$;
  Convert to dual augmented Lagrange minimization;
  $min_{\lambda,z} - q^T\lambda - x^T(z - F^T\lambda) + \frac{\beta}{2}\|z - F^T\lambda\|_2^2\ \ subject.to\ z\epsilon\{\|x\|_\infty < 1\}$;
  While stop condition isn't satisfied, do:
    $z_{l+1} = sgn\left(F^T\lambda + \frac{x}{\beta}\right) \times \min\left\{1, \left|F^T\lambda + \frac{x}{\beta}\right|\right\}$;
    $\lambda_{l+1} = \frac{Fz_{l+1} - \frac{Fx_l - q}{\beta}}{(FF^T)}$;
    $x_{l+1} = x_l - \beta \times (z_{l+1} - F^T\lambda_{l+1})$;
  End
$x_0 = x_{l+1}$
$x_0 = \left[\underbrace{a_1;\ a_2; \dots;}_{1} \underbrace{a_j; \dots;}_{2} \dots; \underbrace{a_n}_{k}\right] = [x_{0,1}; x_{0,2}; \dots; x_{0,k}]$;
$F = \left[\underbrace{F_1, F_2, \dots,}_{1} \underbrace{F_j, \dots,}_{2} \dots, \underbrace{F_n}_{k}\right] = [M_1, M_2, \dots, M_k]$;
$C_i = M_i x_{0,i} \in R^m$;
$e_i = \sqrt{\sum_{l=1}^{m}(q_l - C_{il})^2}$;
$E = [e_1; e_2; \dots; e_k]$;
$x_0^w = \left[\frac{x_{0,1}}{e_1}; \frac{x_{0,2}}{e_2}; \dots; \frac{x_{0,k}}{e_k}\right]$;
$[x_0^w, i] = sort(x_0^w)$;
show nearest images which are pointed in $i$
```

5 Experimental results

In this section, evaluation metrics for testing the proposed system and databases are presented. Then the experimental setup and numerical results are provided. The experimental setup is as following:

As mentioned before, AlexNet is considered to extract deep features. The network specifics are shown in Table 1.

To evaluate the proposed method, the network was fine-tuned on each database in the offline step. For each database, 80% of images were considered as training images (800 images of Corel, 1120 images of Mpeg-7 and 3840 images of ALOI) and 20% are validation images. While the earlier layers were frozen during the training step, two last fully-connected layers were trained on our database. For each database, the feature matrix was extracted. Because the evaluating metrics of retrieval are computed over all images in each database (as explained in Section 5.1), the testing images are the whole images in each database.

5.1 Evaluation metrics

Although precision and recall have been used in many researches for years, three efficient measures with combination of these metrics are used in this paper. Usual precision and recall are defined as Eqs. (20) and (21), respectively:

$$\text{Precision} = \frac{N_r}{N_t} \tag{20}$$

$$\text{Recall} = \frac{N_r}{N_K} \tag{21}$$

Where, N_r is the number of relevant images retrieved, N_t demonstrates total number of images retrieved and N_K is total number of relevant images in database. As defined in [7], the combined metrics are calculated as:

- $P(0.5)$, precision at 50% recall (i.e. precision after retrieving $\frac{1}{2}$ of the relevant images)

Table 2 Feature vector size and evaluated metrics for Corel dataset

Type	Size of feature vector	P(0.5) %	P(1) %	ANMRR
The proposed method	10	96.40	95.59	0.026
IDWT-sparse	64	89.98	89.46	0.082
HMMD-HDWT	3 × 64	41.45	35.44	0.561
MCTF	92	75.43	71.34	0.183
WBCH	512	80.35	76.25	0.120
GGD& KLD	18	65.79	62.35	0.331
WFCTC	384	47.21	40.23	0.548
CLD	12	57.79	41.82	0.539
DCD	32	48.64	36.37	0.582
SCD	11 × 121	43.63	33.65	0.610
PP30	496	38.56	28.65	0.653
HI	1024	40.21	29.92	0.632

Table 3 Feature vector size and evaluated metrics for ALOI dataset

Type	Size of feature vector	P(0.5) %	P(1) %	ANMRR
The proposed method	64	97.46	97.06	0.025
IDWT-sparse	64	89.99	89.89	0.090
HMMD-HDWT	3 × 64	82.29	66.68	0.2551
MCTF	92	81.60	63.33	0.321
WBCH	512	61.59	50.11	0.431
GGD& KLD	18	64.73	55.29	0.402
WFCTC	384	43.03	30.45	0.693
CLD	12	19.49	17.56	0.852
DCD	32	50.21	41.87	0.635
SCD	11 × 121	53.21	43.55	0.521
PP30	496	37.79	7.71	0.902
HI	1024	36.26	6.77	0.918

- $P(1)$, precision at 100% recall (i.e. precision after the retrieving all of the relevant images)

Since using precision and recall separately is not meaningful, a combination of them is used [23]. Another measurement tool is ANMRR, which is an objective measure and summarizes the system performance into a scalar value. This value has been defined from MPEG-7 research group [2, 37]. For calculating this measure, following parameters are considered:

- $NG(q)$: the number of the ground truth images for a query $'q'$.
- $K(q) = \min[4.\ |NG(q)|, 2.\max\{|NG(q)|, \forall q\}]$
- $R(K)$=Rank of an image, k, in retrieval results.

Rank (K) is obtained by:

$$Rank(K) = \begin{cases} R(K) & \text{if } R(K) \ll K(q) \\ 1.25K & \text{otherwise} \end{cases} \tag{22}$$

Using Eq. (22), average rank, AVR(q), for query, q, is given by:

$$AVR(q) = \langle Rank(K) \rangle \tag{23}$$

However, with ground truth sets with different sizes, the AVR(q) value depends on NG(q). To minimize the influence of variations in NG(q), the modified retrieval rank, MRR(q), is obtained by:

Table 4 Feature vector size and evaluated metrics for MPEG-7 dataset

Type	Size of feature vector	P(0.5) %	P(1) %	ANMRR
The proposed method	70	82.54	77.49	0.177
IDWT-sparse	64	59.41	58.31	0.425
HMMD-HDWT	3 × 64	70.41	53.90	0.428
MCTF	92	66.23	46.68	0.501
WBCH	512	39.69	30.21	0.660
GGD& KLD	18	40.23	35.40	0.713
PP30	496	77.77	37.21	0.682
HI	1024	88.34	57.56	0.443

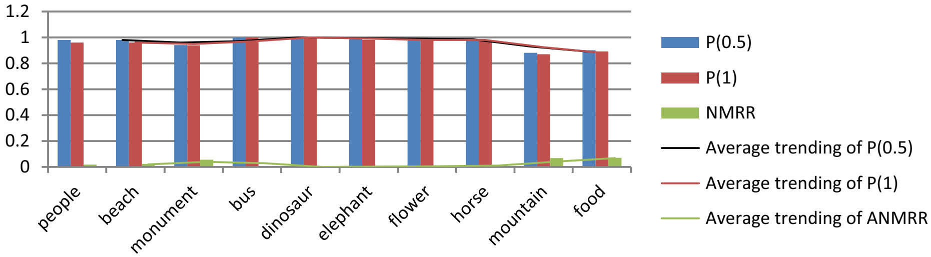


Fig. 4 Results for each class separately in Corel database

$$NMRR(q) = \frac{AVR(q) - 0.5[1 + |NG(q)|]}{1.25K(q) - 0.5[1 + |NG(q)|]} \tag{24}$$

This measure is zero for perfect performance and reaches to one as performance worsens. The ANMRR of the dataset is finally given by averaging the NMRR (q) overall every q as follow:

$$ANMRR = \langle NMRR(q) \rangle \tag{25}$$

5.2 Datasets

Three kinds of common datasets are considered to evaluate the performance of the system and to achieve a comparison between the proposed and state-of-the-art methods. (i) the Corel dataset [3] including 1000 images in variable sizes which are classified into 10 classes of human being, horse, elephant, flower, bus, manmade thing and natural scenery. (ii) The Amsterdam library of object image (ALOI) dataset [6] including 72,000 images in variable sizes which are defined into 1000 classes and it is an appropriate dataset for evaluating robustness of the system against rotation. (iii) the MPEG-7 dataset [10] including 1400 variable size images considered to 70 classes and is a difficult database for color-based image retrieval in which, the images are composed of photos and sequences of video frames, instead of including objects. Note that, we have used only the photos part of each image.

5.3 Indexing results

The obtained results of the proposed method and other algorithms are presented in this section, by computing $P(0.5)$, $P(1)$ and ANMRR metrics. In order to solve Eq. (1), DALM algorithm is used. The results of the proposed method, sparse IDWT [22], HMMD-HDWT [4], multi-

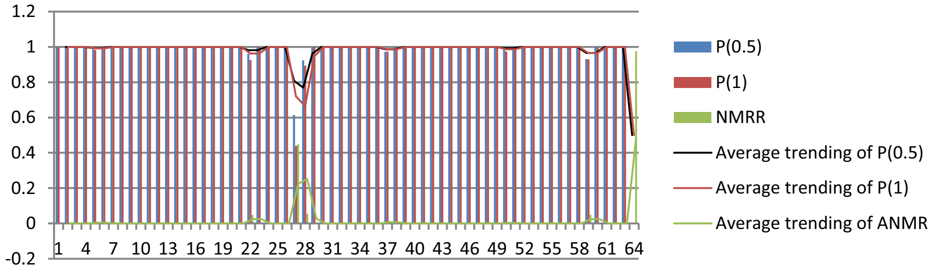


Fig. 5 Results for each class separately in ALOI database

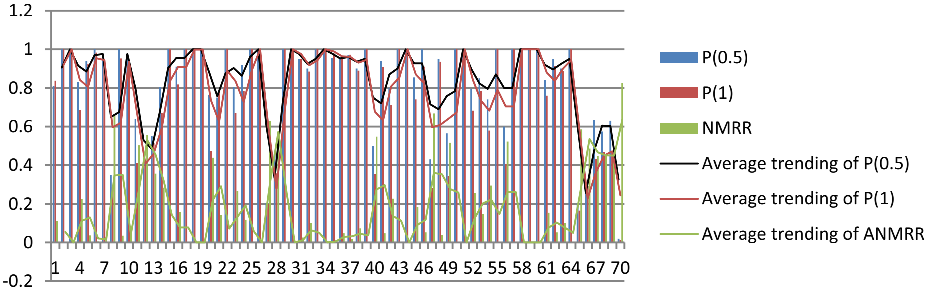


Fig. 6 Results for each class separately in MPEG7 database

resolution color and texture features (MCTF) [2], wavelet-based color histogram (WBCH) [30], generalized Gaussian density and Kullback–Leibler distance (GGD and KLD) [20], wavelet-based features for color texture classification (WFCTC) [9], color layout descriptor (CLD) [35], dominant color descriptor (DCD) [11], scalable color descriptor (SCD) [19], Padua point (PP) and histogram intersection (HI) [23] on three pre-introduced databases are presented.

Experimental results for Corel dataset are shown in Table 2. Three pre-defined metrics and the size of feature vector for each method are shown.

As observed in Table 2, the best score is achieved P(0.5) = 96.40%, P(1) = 95.59% and ANMRR = 0.026 by the proposed CNN-sparse method and is superior compared to other

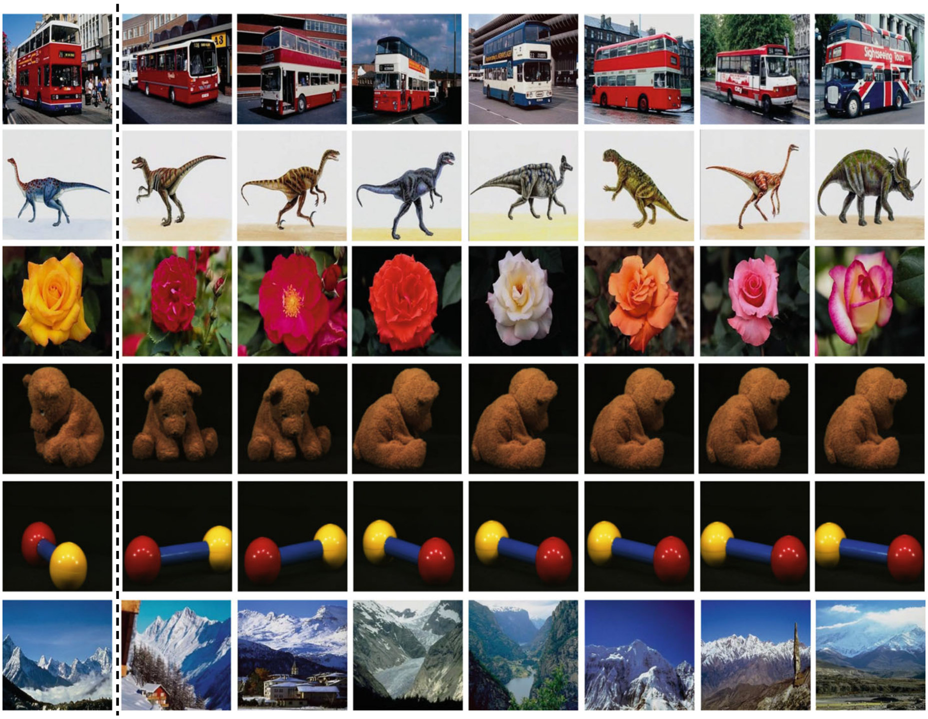


Fig. 7 Some of query images (left) which were fed to the proposed method, and their retrieved images (right)

results. In addition, the size of feature vector in the proposed method, either with sparse representation or without sparse, is less than other methods, which is equal to 10.

Table 3 illustrates the performance on ALOI dataset. The proposed CNN-sparse system achieved P(0.5) = 97.46%, P(1) = 97.06% and ANMRR = 0.025. Although the accuracy in the proposed method without sparse system seems to be a little better, sparse based model has higher speed than computing totally distance.

The performance of system on MPEG-7 dataset, is presented in Table 4 in which, the proposed method has achieved P(0.5) = 82.54%, P(1) = 77.49% and ANMRR = 0.177 in sparse-based CNN. The accuracy of the proposed method without sparse system seems to be the best among other methods; but, as mentioned before, using sparse provides higher speed. It has to be mentioned that the efficiency of any system could not be considered just by its accuracy and retrieval system time processing is important as well. Thus, the efficiency of the system in the proposed CNN-sparse system is better than other methods to the cause that it has improved retrieval accuracy and provides a fast retrieval.

In order to illustrate the robustness of the proposed method among different images, the metrics of all classes in each database are shown separately in Figs. 4, 5 and 6. It can be observed that in Corel and ALOI databases, trending precisions is almost stable except in few classes. Thus, the proposed method seems to be stable in many images classes. Although trending of precisions is unfair in MPEg7 database, the total precision on database is better than state-of-the-art methods shown in Table 4. To show the performance of the system visually, some query images and their relevant images retrieved by the proposed method are shown in Fig. 7.

6 Conclusion

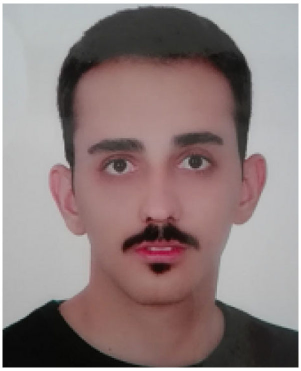
In this paper, a new method for content-based image retrieval was proposed by combination of CNN and sparse representation, in which a CNN as deep learning algorithm was used to extract deep features from images instead of low-level hand-crafted features, and sparse representation was considered to improve the performance of the system. Although some appropriate researches have been reported for image retrieval, extracting low-level features caused achieving low precision. Thus, the main difference of the proposed method was using small-size high-level features which were extracted from convolutional networks. The proposed method used deep features and in order to find relevant images to query image, the retrieval system was modeled by sparse representation which has been a successful method in compressed sensing. This proposed combination of deep learning and sparse representation for image retrieval achieved better results compared to state-of-the-art methods mainly in a result of using deep high-level features and using a fast method to represent compressed sensing, sparse representation. The proposed method was evaluated by computing P(0.5), P(1) and ANMRR on three common datasets which have been used frequently to evaluate retrieval systems. The experimental results on Corel, ALOI and MPEG-7 datasets have shown that, the proposed system has achieved superior performance, either in accuracy or in speed, compared to other methods used for content-based image retrieval.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning - a new frontier in artificial intelligence research [research frontier]. IEEE Comput Intell Mag 5:13–18
2. Chun YD, Kim NC, Jang IH (2008) Content-based image retrieval using multi-resolution color and texture features. IEEE Trans Multimedia 10(6):1073–1084
3. Coral dataset, last referred on June 2009, Available at http://wang.ist.psu.edu/docs/related/
4. Farsi H, Mohamadzadeh S (2013) Colour and texture feature-based image retrieval by using Hadamard matrix in discrete wavelet transform. IET Image Process 7(3):212–218
5. Farsi H, Mohamadzadeh S (2013) Combining Hadamard matrix, discrete wavelet transform and DCT features based on PCA and KNN for image retrieval. Majlesi Journal of Electrical Engineering 7(1):9–15
6. Geusebroek JM, Burghouts GJ, Smeulders AWM (2005) The Amsterdam library of object images. Int J Comput Vis 61:103–112
7. Gou Y, Tao D, Yu j XH, Li Y, Tao D (2016) Deep neural networks with relativity learning for facial expression recognition. In: IEEE international conference on Multimedia & Expo Workshops (ICMEW)
8. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS (2016) Deep learning for visual understanding: a review. Neurocomputing 187:27–48
9. Hiremath PS, Shivashankar S, Pujari J (2006) Wavelet based features for color texture classification with application to CBIR. IJCSNS International Journal of Computer Science and Network Security 6(9):124–133
10. International organization for standardization, MPEG-7 overview 2004. Available at http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm. accessed 15 Nov 2011
11. Ka-Man W, Lai-Man P, Kwok-Wai C (2007) Dominant color structure descriptor for image retrieval. In: IEEE international conference on image processing (ICIP)
12. A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images". 2009
13. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS 2012)
14. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324
15. Lee H, Largman Y, Pham P, Ng A (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems 22 (NIPS'09)
16. Li F, Dai Q, Xu W, Er G (2008) Multi-label neighborhood propagation for region-based image retrieval. IEEE Trans Multimed 10(8):1592–1604
17. Liapis S, Tziritas G (2004) Color and texture image retrieval using chromaticity histograms and wavelet frames. IEEE Trans Multimedia 6(5):676–686
18. Liu H, Li B, Lv X, Huang Y (2017) Image retrieval using fused deep convolutional features. Procedia Comput Sci 107:749–754
19. Manjunath BS, Ohm JR, Vasudvan VV, Andyamada A (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol 11(6):703–715
20. Minh ND, Vetterli M (2002) Wavelet-based texture retrieval using generalized Gaussian density and kullback–leibler distance. IEEE Trans Image Process 11(2):146–158
21. Mohamadzadeh S, Farsi H (2014) Image retrieval using color-texture features extracted from Gabor-Walsh wavelet pyramid. Journal of Information Systems and Telecommunication 2(1):31–40
22. Mohamadzadeh S, Farsi H (2016) Content-based image retrieval system via sparse representation. IET Comput Vis 10:95–102
23. Montagna R, Finlayson GD (2012) Padua point interpolation and Lp-norm minimization in color-based image indexing and retrieval. IET Image Process 6(2):139–147
24. Peng T q, Li F (2017) Image retrieval based on deep convolutional neural networks and binary hashing learning. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1742–1746
25. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–8
26. Qayyum A, Anwar SM, Awais M, Majid M (2017) Medical image retrieval using deep convolutional neural network. Neurocomputing
27. Silva Júnior JA, Marçal RE, Batista MA (2014) Image retrieval importance and applications. In: Workshop de Visao Computacional - WVC 2014
28. Karen Simonyan, and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition". http://arxiv.org/abs/1409.1556, 2014

29. K. Simonyan, and Zisserman, A. "Very deep convolutional networks for large-scale image recognition, " Published as a conference paper at ICLR 2015
30. Singha M, Hemachandran K (2012) Content based image retrieval using color and texture. Signal & Image Processing: An International Journal (SIPIJ) 3(1):39–57
31. Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
32. Tao D, Guo Y, Song M, Li Y, Yu Z, Yan Tang Y (2016) Person re-identification by dual-regularized KISS metric learning. IEEE Trans Image Process 25(6):2726–2738
33. Tao D, Guo Y, Li Y, Gao X (2018) Tensor rank preserving discriminant analysis for facial recognition. IEEE Trans Image Process 27(1):325–334
34. Torres RDS, Falcao AX (2006) Content-based image retrieval theory and applications. RITA 8
35. Troncy R, Huet B, Schenk S (2011) Feature extraction for multimedia analysis: multimedia semantics, desktop edition (XML): metadata, analysis and interaction, 1st edn. Wiley, New York
36. Varga D, Szirányi T (2016) Fast content-based image retrieval using convolutional neural network and hash function. In: IEEE international conference on systems, Man, and cybernetics (SMC), pp 2636–2640
37. Veganzones MA, Graña M (2012) A spectral /spatial CBIR system for hyper spectral images. IEEE J-STARS 5:488–500
38. Yang AY, Zhou Z, Ganesh A et al (2013) Fast l1-minimization algorithms for robust face recognition. IEEE Trans Image Process 22(8):3234–3246
39. Zhang Z, Xu Y, Yang J et al (2015) A survey of sparse representation: algorithms and applications. IEEE Access 3:490–530



Amir Sezavar received the B.Sc. and M.Sc. degrees in telecommunication engineering from university of Birjand, Birjand, Iran in 2015 and 2017, respectively. He is currently Ph.D. student in university of Birjand, Birjand, Iran. His research interests in digital image processing and retrieval, visual signal processing, deep learning and artificial intelligence. His Email is: a.sezavar@birjand.ac.ir.

Hassan Farsi received the B.Sc. and M.Sc. degreed from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D. in the Center of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D. degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as professor in communication engineering in department of Electrical and Computer engineering, university of Birjand, Birjand, Iran. His Email is: hfarsi@birjand.ac.ir.

Sajad Mohamadzadeh received the B.Sc. degree in electrical engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. degree in telecommunication engineering from university of Birjand, Birjand, Iran, in 2012. He is currently an academic staff in Faculty of Technical and Engineering of Ferdows, university of Birjand, Birjand, Iran. His area research includes image processing and retrieval, pattern recognition, digital signal processing and sparse representation. His Email is: s.mohamadzadeh@birjand.ac.ir.