Multimedia Tools and Applications (2020) 79:18033–18058
https://doi.org/10.1007/s11042-019-08264-y

A comprehensive system for image scene classification

Ali Ghanbari Sorkhi[1] · Hamid Hassanpour[1] · Mansoor Fateh[1]

Received: 21 November 2018 / Revised: 10 July 2019 / Accepted: 16 September 2019 /
Published online: 26 February 2020

Abstract

In recent years, image scene classification based on low/high-level features has been considered as one of the most important and challenging problems faced in image processing research. The high-level features based on semantic concepts present a more accurate and closer model to the human perception of the image scene content. This paper presents a new multi-stage approach for image scene classification based on high-level semantic features extracted from image content. In the first stage, the object boundaries and their labels that represent the content are extracted. For this purpose, a combined method of a fully convolutional deep network and a combined network of a two-class SVM-fuzzy and SVR are used. Topic modeling is used to represent the latent relationships between the objects. Hence in the second stage, a new combination of methods consisting of the bag of visual words, and supervised document neural autoregressive distribution estimator is used to extract the latent topics (topic modeling) in the image. Finally, classification based on Bayesian method is performed according to the extracted features of the deep network, objects labels and the latent topics in the image. The proposed method has been evaluated on three datasets: Scene15, UIUC Sports, and MIT-67 Indoor. The experimental results show that the proposed approach achieves average performance improvement of 12%, 11% and 14% in the accuracy of object detection, and 0.5%, 0.6% and 1.8% in the mean average precision criteria of the image scene classification, compared to the previous state-of-the-art methods on these three datasets.

Keywords Scene classification · Semantic feature · Latent topic · Fully convolutional · Two-class SVM-fuzzy · Latent topics

✉ Ali Ghanbari Sorkhi
ali.ghanbari@shahroodut.ac.ir

Hamid Hassanpour
h.hassanpour@shahroodut.ac.ir

Mansoor Fateh
mansoor_fateh@shahroodut.ac.ir

[1] Faculty of Computer Engineering, Shahrood University of Technology, Shahrud, Iran

1 Introduction

In recent years, many researches have been focusing on image scene classification in machine learning and machine vision. In image scene classification, a set of images from scenes that contain different object categories are considered. The goal is to discover these objects and use object occurrences for scene classification. In fact, each image is based on a global descriptor (for example, "coast", "outdoor", "inside city") and may include different objects ("sky", "car", "Tree", etc.). For example, in an image with a street global label, the likelihood of describing it with "car," "man", or "building" is more than "beach" or "sea water".

In the research done in scene classification, mainly either low-level features, such as color, texture, and histogram, or high-level features based on semantic modeling are used. Employing low-level features has simplicity and low computational cost, but with poor performance in image scene classification applications. In contrast, employing the high-level features with the ability of semantic information acquisition has considerably a better performance in many real-world applications. Several methods have been introduced for image scene classification in recent years [61]. In older methods, each image is considered as a separate object and is executed with low-level features for the classification. These techniques are typically used only for the classification of a small number of scenes and cannot be used to categorize real-world images. In recent years, the high-level features of multi-regions have been used which have suitable performance in this application. In modern methods, latent variables are used to construct the topic model [20, 26, 62, 63]. In these systems, image categorization is performed based on the semantic features of the image. These methods are used for cases where the number of classes in the scene is high [12, 78, 81].

To utilize the human understandable ability, a semantic representation can be used. In the same way, in [7], external image archives and apply the concept detectors have been used to semantic representation in event detection in the video. They proposed a bi-level semantic representation analysis method. Due to the source-level, their method learned weights of semantic representation achieved from different multimedia archives. Meanwhile, it limited the negative effect of noisy or irrelevant concepts in the overall concept-level. In reference [8] a new semantic pooling approach for challenging event analysis tasks in long untrimmed Internet videos has been proposed, especially when only a few shots/segments are relevant to the event of interest while many other shots are irrelevant or even misleading. In this reference, a novel notion of semantic saliency has been defined which assesses the relevance of each shot with the event of interest. Furthermore, a new isotonic regularizer that is able to exploit the constructed semantic ordering information has been proposed by them. The resulting nearly isotonic support vector machine classifier exhibits higher discriminative power in event analysis tasks. As shown in these works, gaining meaning can have a great effect on detecting events in images. In order to obtain existing semantic concepts, it is necessary to detect the objects in the image.

The detection and recognition of objects in an image involve some typical challenges, including complex backgrounds, unknown locations, different view angles, different ambient lights, the large number of objects, partial overlaps of objects, etc. An efficient and effective method should overcome these challenges [36, 45, 52, 75].

Recently, object detection systems have been mainly focusing on region proposal algorithms to estimate the location of objects. Because of the time limitations and the various modes of objects, the whole image regions cannot be examined by these methods. For this reason, first, the region proposals of the object's existence must be extracted. Then these

regions are applied to object detectors. Although finding region proposals offers good performance, these techniques require extensive computations [18]. But these calculations are reduced in a region-based convolutional neural network by the share of candidates obtained from convolution [17, 21]. Actually, in [17, 21], a Region Proposal Network (RPN) has been introduced. In this network, features are shared with different convolutions extracted across the image. As a result, the time associated with the region proposals is negligible. RPN simultaneously predicts the boundaries of objects and provides a quantity called objectness score in any position of the image to facilitate the object recoginition scheme.

In this paper, we introduce a comprehensive system for image scene classification based on high-level semantic region and topic modeling. The proposed method involves three stages. In the first stage, objects inside the images are identified based on a combination of a deep network, a two-class fuzzy SVM and SVR. In the second stage, topic modeling is specified due to the extracted areas and labels. Then, based on these areas, latent topics are extracted from the image. In the final stage, classification based on Bayesian classifier is carried out by a hybrid network. The hybrid network considers the latent topics in the images, the features extracted in different layers of the deep network and the proposed labeling. In the following, we describe and express each step individually. Figure 1 illustrates an overview of the proposed method.

The rest of the paper is organized as follows: an overview of object extraction and labeling is provided in Section 2. In Section 3, the R-FCN method is introduced. then the method is improved using SVM-Fuzzy in Section 4. Extracting the bag of visual words is performed in Section 5. In Section 6, a hybrid method is proposed for topic models extraction. In Section 7, the image scene classification method based on the extracted features is defined.

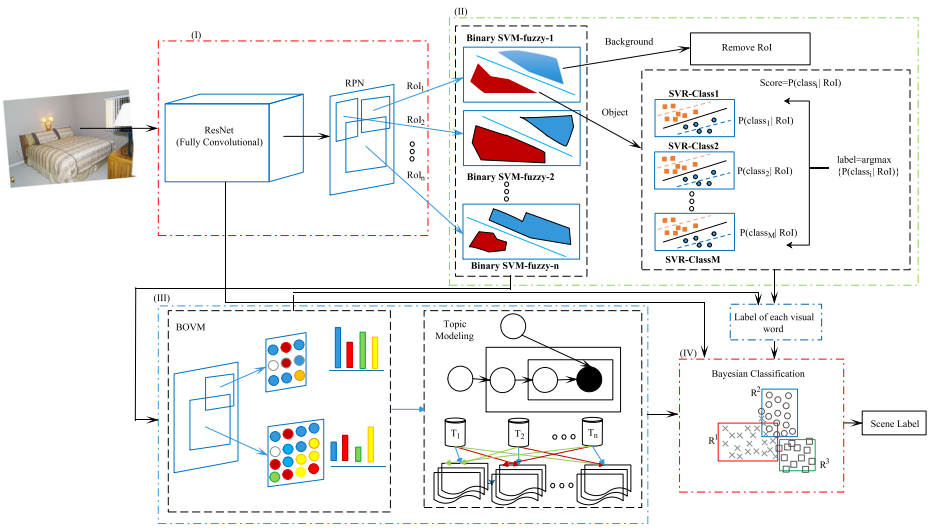

Fig. 1 Different stages of the proposed method for image scene classification. (I), the candidate regions are identified by the deep network. (II), label of regions are detected by fuzzy SVM and SVR. At this stage, in two-class fuzzy SVM method, the background of the objects are separated and the output of this stage (label of object) is as the input of the next stage. (III), the BOVW method is used to extract visual words, and then the topic modeling is used to extract the latent models in the images. (IV), the output of different stages is used as the input of the Bayesian classifier for scene classification

Experimental results on three datasets are provided in Section 8. Finally, conclusion and discussion are presented in Section 10.

2 Object extraction and labeling

R-CNN proposed in [18] is one of the first region-based deep convolutional neural networks. This approach employs the AlexNet architecture to detect objects, and uses selective search [67] in its region proposal method. The major drawback of this method is that the use of the fixed size of the input image for the fully-connected layers is necessary. The SPP-net network [21], proposed in 2014, overcame this limitation, and also produced the results, 20 to 60 times faster than the R-CNN network. Followed by R-CNN, networks such as Fast-RCNN [17], Faster R-CNN [57] and R-CNN minus R [33] were introduced to address the R-CNN limitations. In Fast-RCNN, the input image and several regions of interest are fully convolutional inputs. Each region-of-interest (RoI) is converted to a fix-length feature by the pooling layer. The network for each RoI has two output vectors. An output vector determines an estimate of each class label and the other output vector determines regions related to the rectangular window of the object. The R-CNN minus R network is simplified by the RCNN method. In [33], the role of region proposal generation in CNN-based detectors has been studied. In this work, a new detector has been introduced for the region generation. The combination of this detector with the SPP-Net provides a better and faster performance.

In 2016, ProNet [65] introduced a multi-scale fully convolutional network. This network assigns a confidence score to the bounding box in different locations and scales. Cascade or tree methods are used to select the object labels. The Single Shot multi-box Detector (SSD) network [42] uses a single deep neural network to detect an object. The output of bounding boxes is limited to a set of default boxes in different aspect ratios and scales in each location. At prediction time, the network generates scores for each category of objects in each default box. Also, the network creates the box-specific settings for a greater compliance with the shape of the object. Previous methods with the two proposed RPN networks and class estimation were complex due to the slow pipelining, difficulty in optimizing individual pipelines and the need for training pipelines for each section.

In the YOLO method [56], detection is considered as a regression problem. There is also a separate convolutional network for design, which is simple and implemented extremely fast. In this architecture, the image is converted to a fixed number of grid networks. If the center of an object is within the boundaries of a grid network, this grid network can be acceptable for the object detection. The limitation of this method is to detect small size and aspect ratios of objects [56].

The most important works done in 2017 are the SSD + DSSD [15], YOLO9000 [55] and RFCN [11] methods. In the DSSD method, a combination of new class (Residual-101 [22]) with fast detection (SSD [42]) was proposed. The hybrid network has been enhanced by introducing deconvolutional layers to provide good performance in detection of small objects and improving the performance of object detection. YOLO9000 is faster, stronger, and better than YOLO. In YOLO9000, batch normalization [23] with a higher resolution classifier, convolution with anchor boxes, dimension clustering, Fine-Grained features, multi-scale training, hierarchical classification and the classification connection with detection have been used. One of the most important works performed in 2017 is the R-FCN method, whose structure is shown in Fig. 2. This method has the best performance compared to the methods proposed in recent years. The R-FCN method uses fully convolutional with all computations shared on the entire image to perform better than the previous region-based

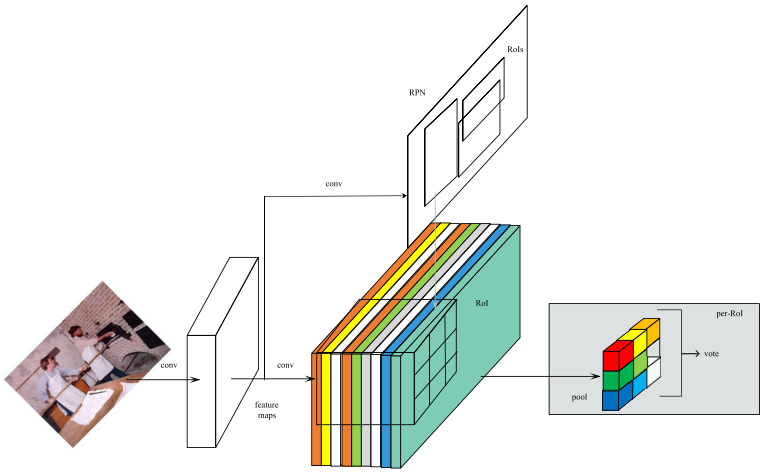

Fig. 2 The structure of R-FCN method [11]

detectors, such as Fast/Faster R-CNN. In this paper, a concept has been used as a position-sensitive score maps. In R-FCN, the channels represent the location of each RoI in the input image. Each channel is designed for a specific location.

The RPN network [57], shown as one of the elements in R-FCN architecture, has been introduced to propose candidate RoIs applied in the score map. The cost of calculating each RoI is negligible. RPN is a fully convolutional network with ResNet [22] architecture. In this network, extracted features are shared between RPN and R-FCN. R-FCN is designed to classify the RoIs into object categories and backgrounds [11].

RetinaNet's deep network has been introduced in [40]. In this method, the backbone network called Feature Pyramid Net is used, which is built on top of ResNet and is responsible for computing convolutional feature maps of an entire image; two subnetworks are responsible for performing object classification and bounding box regression using the backbones output. This also includes a new loss function called "focal loss" in RetinaNet. This focal loss has been introduced by reshaping the standard cross entropy loss. Another method introduced is the RefineDet method, which is a method consisting of two inter-connected modules; the anchor refinement module and the object detection module. The first module has two aims; filtering out negative anchors to decrease search space for the classifier, and regulating the locations and sizes of anchors to provide better initialization for the further regressor. The second module takes the refined anchors as the input from the former to further improve the regression and predict the multi-class label. Meanwhile, a transfer connection block has been designed by them to transfer the features in the anchor refinement module to predict locations, sizes and class labels of the objects in the object detection module [79]. In 2019, TridentNet method claimed introducing a scale invariation method. In this way, a controlled experiment to investigate the effect of receptive fields on the detection of different scale objects is presented.

A novel Trident Network (TridentNet) aiming to generate scale-specific feature maps with a uniform representational power based on the findings from the exploration experiments has been presented. They construct a parallel multi-branch architecture. Also, a scale-aware training scheme has been proposed to specialize in each branch via sampling object instances of proper scales for training. A fast estimation version of TridentNet could

get notable improvements without any additional parameters and computational cost [38]. In this paper, R-FCN architecture has been used to select the region proposal. In the next step, a distinct phase has been introduced for detecting objects and background.

3 R-FCN network

As explained in the previous section, the R-FCN network is one of the newest deep networks for object detection. This network is made up of a combination of RPN network and fully convolutional network. The outputs of these networks are ranked for each RoI according to the location of the objects. In R-FCN, the loss function defined on each RoI is the summation of the cross-entropy loss and the box regression loss. This function is shown below:

$$L(S, t_{x,y,w.h}) = L_{cls}(S_{c^*}) + \lambda[c^* > 0]L_{reg}(t, t^*), \tag{1}$$

where c^* is the ground-truth label of RoI. $c^* = 0$ means the background. In fact, $[c^* > 0]$ specifies 1 if the argument is true, and 0 otherwise. Each overlapping RoI of over 50% with a ground-truth box is considered as a positive example. t and t^* are respectively related to the bounding box and ground-truth box values. $L_{cls} = -\log(S_{c^*})$ is the cross-entropy loss for classification. This function is shown in (2):

$$CEL = -\sum_j y^{*(j)} \log \sigma(o)^{(j)}, \tag{2}$$

where y^* and o are respectively the ground-truth output and the network output. In addition (j) represents the jth dimension of a given vector. In (1), $L_{reg}(t, t^*)$ is box regression loss as expressed in [17].

$$L_{reg}(t, t^*) = \sum_{i \in \{x,y,w,h\}} smooth_{L_1(t_i - t_i^*)}, \tag{3}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } x < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

In the above equations, the loss function $L_1(.)$ is used which is a norm-1 function. The reason for using norm-1 is a less sensitivity to outliers than the norm-2 (L_2). Norm-2 has been used in [18]. Using the proper loss function is very effective in network accuracy. For this reason, this paper discusses the improvement of the R-FCN network using the new loss function. Several known loss functions that can be used in deep learning were introduced in [25]. The results show that the use of the loss function is highly dependent on the application. In this paper, we analyzed and compared the relationship between Cauchy-Schwarz Divergence (CSD) loss and cross-entropy loss. It should be noted that Cauchy-Schwarz Divergence loss function is not used in deep networks to identify objects. Experiments show that the Cauchy-Schwartz difference loss function is more optimal than the loss of entropy in terms of speed and performance [25]. In this paper, the Cauchy-Schwartz Divergence loss function is used. This loss function is defined as (5) [10]:

$$CSD = -\log \frac{\sum_j \sigma(o)^j y^{*(j)}}{\|\sigma(o)\|_2 \|y\|_2}, \tag{5}$$

where y^* , o and σ are the ground-truth output, the network output, and the probability estimate respectively. (j) represents the jth dimensional vector. In R-FCN, the background is considered as a distinct class. In the proposed method, we do not want to increase the number of classes. In this regard, a separate phase is considered for discriminating between

background and objects. First, we design a two-class fuzzy SVM based system (binary fuzzy-svm), which determines whether the region is related to the object or not. Labeling is done for objects only. Usually, for each image, a large number of region candidates are introduced. Only regions that can be considered as objects are evaluated. This practice greatly reduces the labeling time for the whole image. A new approach is used to identify regions related to objects, as outlined in Section 4. The R-FCN method uses voting to select the label of each RoI. However, the proposed method uses SVR. In fact, for each class of objects, an SVM is considered. In the proposed method, each SVR specifies the degree of belonging of each RoI to a class. Eventually, the maximum degree of belonging of each RoI specifies the label for each RoI. The SVRs are implemented in the pipeline in the proposed method to reduce processing time. The general scheme of the proposed method is shown in Fig. 1. In fact, in each SVR, the degree of belonging of each RoI is calculated in a specific class and finally, the class label is determined using (6). In this equation n is the number of classes.

$$ClassNumber = \arg\max_{i=1}^{n}\{P(Class_i|RoI)\} \tag{6}$$

4 Fuzzy support vector machine

The support vector machine, as a powerful tool for classification and regression, has been used in many practical applications [6, 24, 76]. Many versions of the SVM have been introduced in recent years. One of the most famous ones is the fuzzy-based version. Most Fuzzy SVMs are used to solve problems where patterns belonging to a class often play a significant role in classification. In the proposed method, after specifying the RoIs existing in each image by the usual SVM, the degree of belonging of each region to the class is determined. In this step, we use a two-class fuzzy SVM to separate the object from the background. In this regard, we use the method presented in [9] for classification. The output of the two-class fuzzy classification defines either an object or a background. The detected background is excluded, but the detected object is labeled in the next step. This action reduces the amount of image labeling. The set of training samples for the binary classification problem is defined as follows:

$$T^* = \{(x_1, m_1), (x_2, m_2), ..., (x_l, m_l)\}. \tag{7}$$

In (7), x_i samples, $m_i \in [0, 1]$ fuzzy membership, evaluate the degree of belonging of the i-th observation x_i to the positive class, and l is the number of training samples. In this method, sample p $\{(\tilde{x}_1, \tilde{m}_1), (\tilde{x}_2, \tilde{m}_2), ..., (\tilde{x}_p, \tilde{m}_p)\}$ of observations is considered as positive samples (object) and sample q $\{(\hat{x}_1, \hat{m}_1), (\hat{x}_2, \hat{m}_2), ..., (\hat{x}_q, \hat{m}_q)\}$ of observations as negative samples (background). Label of samples are calculated from $Y_i = 2m_i - 1$. $Y_1 = diag(\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_p)$ and $Y_2 = diag(\hat{y}_1, \hat{y}_2, ..., \hat{y}_q)$ are defined for positive and negative samples of diagonal matrices which $\tilde{y}_i = 2\tilde{m}_i - 1, (i = 1, 2, ..., p)$ and $\hat{y}_i = 2\hat{m}_i - 1, (i = 1, 2, ..., q)$. When $m_i = 1$, the sample x_i is positive and the associated label is $Y_i = 2m_i - 1 = 1$ and when , $m_i = 0$ the sample x_i is negative and the associated label is $Y_i = 2m_i - 1 = -1$. In this paper, the positive class represents the object and the negative class indicates the background. The optimization is defined as follows:

$$\begin{aligned} &\min_{w_1, b_1, \zeta_2} \frac{1}{2}\|Aw_1 + e_1 b_1\|_2^2 + \frac{1}{2}c_1(w_1^2 + b_1^2) + c_2 e_2^T \zeta_2, \\ &st. Y_2(Bw_1 + e_2 b_1) \geq Y_2^2 e_2 - Y_2^2 \zeta_2, \zeta_2 \geq 0 \end{aligned} \tag{8}$$

$$\min_{w_2,b_2,\zeta_1} \frac{1}{2}\|Bw_2 + e_2b_2\|_2^2 + \frac{1}{2}c_3(w_2^2 + b_2^2) + c_4e_1^T\zeta_1,$$
$$st.Y_1(Aw_2 + e_1b_2) \geq Y_1^2e_2 - Y_1^2\zeta_1, \zeta_1 \geq 0 \tag{9}$$

where c_1, c_2, c_3 and c_4 are penalty parameters, ζ_1 and ζ_2 slack variables, $A \in \Re^{p\times n}$ matrix represents the samples related to the positive class and $B \in \Re^{q\times n}$ are samples related to the negative class. After solving this equation by the Lagrangian multipliers described in details in [9], we find the following equations:

$$v_1 = (H^TH + c_1I)^{-1}G^TY_2^2\alpha \quad where\ v_1 = [w_1^Tb_1]^T, H = [Be_2] \tag{10}$$

$$v_2 = (G^TG + c_3I)^{-1}H^TY_1^2\gamma \quad where\ v_2 = [w_2^Tb_2]^T, G = [HAe_1] \tag{11}$$

After obtaining v_1 and v_2 , for each new sample $x \in \Re^n$ the sample label is obtained by (12):

$$x \in W_k, \quad k = \arg\max_{i=1,2}\{\frac{|w_1^Tx+b_1|}{\|w_1\|}, \frac{|w_2^Tx+b_2|}{\|w_2\|}\}. \tag{12}$$

To calculate the membership value, the fuzzy membership functions introduced in [29] are used. For a positive sample x_i with a positive label (+1), fuzzy membership is expressed as (13):

$$\begin{aligned} m_1(x_i) &= 0.5 + \frac{\exp(C_0(d_{-1}(x_i) - d_1(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - exp(-C_0))}, \\ m_{-1}(x_i) &= 1 - m_1(x_i). \end{aligned} \tag{13}$$

For a negative sample x_i with a negative label (-1), fuzzy membership is expressed as (14):

$$\begin{aligned} m_{-1}(x_i) &= 0.5 + \frac{\exp(C_0(d_1(x_i) - d_{-1}(x_i))/d) - \exp(-C_0)}{2(\exp(C_0) - exp(-C_0))}, \\ m_1(x_i) &= 1 - m_{-1}(x_i). \end{aligned} \tag{14}$$

5 Extracting the bag of visual words

As outlined in the introduction section, obtaining high-level semantic features can greatly enhance the performance of the proposed system for scene classification. As mentioned earlier, the regions and labels of the words in the image are obtained by the deep network.

The results obtained in the previous step indicate that regions of the objects usually exhibit more region than the object itself. For example, Fig. 3 illustrates a sample RoI-calculated by RPN that is considered by the SVM-fuzzy as an object. As it can be seen, many of the region proposals have problems like overlaping with other objects, the loss of a part of the object, and the inclusion of a large part of the background region as object regions. In this phase, high-level visual features are extracted in the identified regions as objects. In the same way, the Bag-of-Visual-Word (BoVW) method is used. In the proposed system, the method of topic modeling is used to extract latent topics. The inputs of these types of systems are words. These words are extracted by the BoVW method.

The BoVW pipeline methods have shown good performance in recent years [70]. Also, these methods have a good effect on the illumination changes and noise [50]. This method involves steps such as feature extraction, feature preprocessing, codebook generation, feature encoding, pooling and normalization.

The feature extraction process uses low-level local features. As outlined in [71], extraction of local features involves detecting a local region and describing the detected region. In earlier methods, local regions are extracted from regions related to the whole image. But in

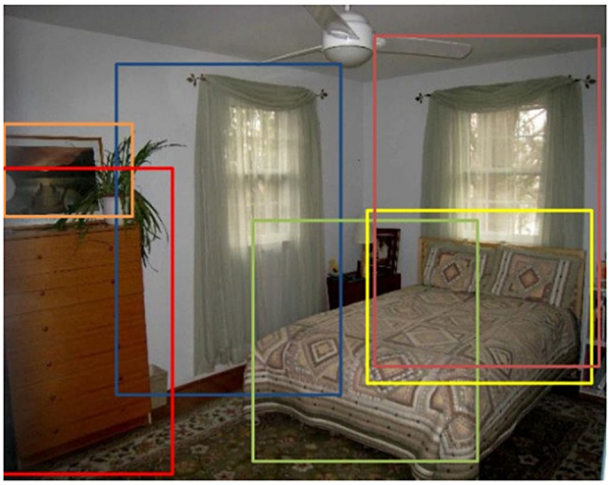

Fig. 3 An example of an RPN network output after applying Fuzzy-SVM

this paper, local regions are extracted in RoI regions. In this paper, the SIFT method is used to extract features. SIFT is calculated for each RoI individually. Low-level local descriptors usually have a high-dimensional and strong correlation [50].

In the same way, in the pre-processing phase, the PCA method [1] is used to reduce dimensionality. After extracting and reducing features, we need to generate codebook. There are two methods for generating codebook. In the first method, feature space is divided into regions. Each region is displayed by its center. Each center is called code-word. In the second method, the generative model is used to represent the probable distribution of the features.

After making the codebook, the encoding method should be done. In general, encoding method can be divided into voting based encoding method [41], reconstruction based encoding method [72] and super vector based encoding method [82]. In fact, in this section, we will determine how close each sample is to the centers of the cluster. Finally, in the pooling and normalization process, the final feature vector is computed by using methods such as Sum-Pooling or Max-Pooling, for each image. The proposed method for extracting BoVW is shown in Fig. 4. As shown in the figure, in the training phase, RoIs that are identified by the fuzzy SVM is used as an object-related region for each image.

In the next step, the SIFT method is used to extract features from these regions. In the next step, PCA is used to reduce dimensionality. Many clustering methods such as k-means clustering, hierarchical clustering and spectral clustering are used to calculate codebook. The k-means method is commonly used in BOVW methods because of the simplicity of implementing codebook build-ups. In this paper, an improved k-means clustering algorithm introduced in [46] is used. The reason for using this method is that it is faster compared to the k-means based method [46]. In fact, at this point, the visual words in the entire training set are extracted. The next steps are carried out for the test set. For training input image, the SIFT features are extracted and these features are mapped to the PCA space. In the encoding phase, the kd-tree method is used to accelerate the calculation of the distance between each sample and the clusters. In fact, the k-dimensional tree is used for the cluster centers. This method reduces the search time complexity for samples from $O(m \times n)$ to $O(m \times \log n)$, where the m and n parameters are the number of clusters and the number of samples,

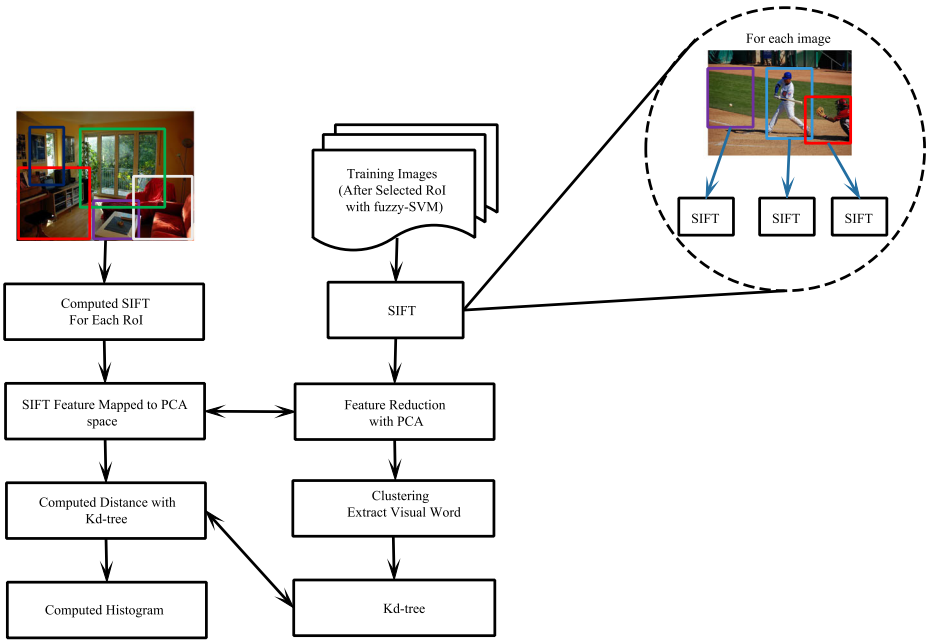

Fig. 4 The general stages of the BOVW extraction process

respectively. Finally, using Max-pooling, the number of repetitions of each sample in the clusters and its histogram are calculated. In fact, for each image, the number of iterations of visual words are specified. At this point, we could make clustering based on the label of each RoI. In fact, at the stage of clustering, we considered the label of each class. Experiments showed that this method reduces performance of the proposed system.

6 Topic model extraction

In recent years, many methods have been introduced for extracting latent topics in images. In these methods, multimodal data is used for modeling [20, 26, 62]. One of the most important methods of modeling is latent Dirichlet allocation (LDA) [3]. This method is essentially introduced for the model production in documents.

The developed LDA methods, such as Corr-LDA [2], Multimodal LDA [51], and MDRF [26], are introduced to compute topic modeling. a method is presented as a Document Neural Autoregressive Distribution Estimator (DocNADE) based on topic modeling in [81]. This method, in comparison with previous methods, demonstrates a better performance in multimodal data. In this paper, this method is used for extracting latent topics. The general schema of the method proposed in [81] is shown in Fig. 5. In this method, the developed system which is supervised by DocNADE as SupDocNADE, has been proposed. This system increases the separation power by teaching latent topics features and shows how to use this method to share different views of the image visual words information, annotations and class labels. Class label is a general label for the image scene. In this way, annotations are the inputs of the system that already exist. However, in this article label of the objects in the image is automatically calculated, using the proposed method.

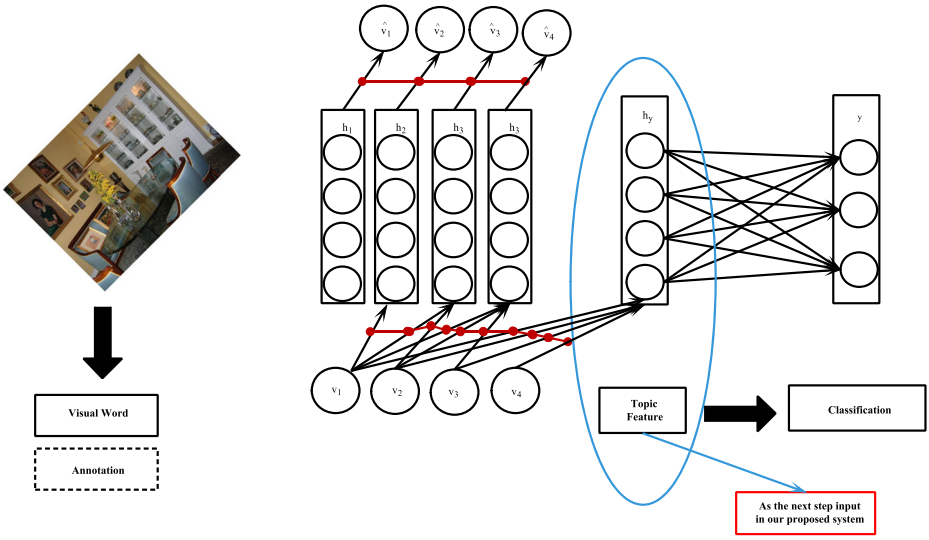

Fig. 5 The method Proposed in [81]. In this paper, the output of the topic feature extraction layer has been used as the next step input in our proposed system

In this method, a deep network is used to calculate the label of each image scene. The last layer represents the label of each scene. The layer before the end calculates latent topics in each image. In this article, the output of the layer before the end is used as a feature of the latent topics for the next step. In fact, the image scene classification is not performed according to the reference method [81]. Only the features related to latent topics are extracted and considered as the input of the next stage of the proposed system. This method is explained in [81]. Also in this paper, the location of the objects is presented as a valuable feature. In fact, each image is divided into equal parts and the location of each visual word is calculated. In this paper, the input image is divided into several regions.

7 Image scene classification

In recent years, many ways have been introduced to classify the image scene based on deep learning. In [43], the pre-trained CNN models have been used for extracting visual features from the middle layer. Also, they used a controlled learning method respectively to train classifiers for each feature. Finally, they used the late fusion strategy to compound the prediction of these classifiers. The authors in [69] show how to use external web text to improve image classification. The keystone of web text-aided image classification is the representation learning for these two modalities of data. A novel semantic CNN (s-CNN) model for high-level text representation learning using task-generic semantic filters based on the pre-trained word vectors has been proposed in [69]. Combining the image CNN models and the s-CNNC models can further enhance image classification named a multi-modal framework (mm-CNN). In this method, online text retrieval has been used to extract similar images and annotations. For this purpose, the search engine of Google has been assisted. As used in recent works, the use of a deep neural network to extract high-level semantic features can greatly affect the classification accuracy. In the same way, in the

preceding sections, the high-level semantic features have been extracted by combining the topic model. In the following, the simple classifier used for classification is described.

In the next step, classification is performed. Simple Bayesian method is used for the classification. Several types of features are considered as inputs in this classification. The features used in this article are shown in Table 1. As shown in the table, the features of the last layer of the ResNet network, latent topics in the image and labels of the extracted regions are the inputs of the final classification. The PCA method is used for dimensional reduction of features. Details of this procedure are expressed in the testing section.

8 Experiments and results

In this section, the experiments and results are analyzed based on the proposed method. Experiments are analyzed for each section separately. In the first step, datasets are introduced. All experiments and extracted parameters are performed separately for each step of the proposed method for each dataset. Extraction of regions related to the objects is an important part in the proposed method. In this regard, the proposed method with famous deep networks introduced in recent years is evaluated. Obtaining visual words is an important part of obtaining latent topics. An analysis is performed to find parameters such as the number of clusters and also the clustering method. The number of latent topics can be very effective in the final accuracy of the proposed method, so the number of topics for each dataset is interpreted. Finally, the proposed method is compared and analyzed for scene classification using famous methods in this area.

8.1 Datasets description

The datasets used in this article are Scene15, UIUC Sports, and MIT-67 Indoor. These datasets are introduced below.

Scene15: This dataset contains 15 categories of different scenes for indoor and outdoor environments. The number of images in these scenes is 4485, each category containing 200 to 400 images in gray space. In this article, 100 images from each category are selected for training, 50 images for validation and the remaining images for testing.

UIUC Sports: This dataset contains eight categories of scenes related to sports events. The number of images in these scenes is 1579, each containing 137 to 250 images. In this article, 70 images are selected for training, 20 images for validation and the remaining images for testing.

MIT-67 indoor scenes: This dataset contains 67 categories from different scenes of indoor environments. The number of images in these scenes is 15620. Each category contains around 100 images. In this article, 70 images of each category are selected for training, 10 images for validation and the remaining images for testing.

Table 1 Extracted features for Bayesian classification

	Mthods		
	ResNet	BoVW+ SupDocNADE	SVR
Features	Deep Feature of Convolutional layers	Topic Feature	Label of Objects (BOVM)

8.2 Experimental setup

The mean average precision (MAP) method is used to evaluate the proposed method in the region extraction phase related to the object [14]. In the following, the accuracy criterion has been used to identify the global label of each image.

Implementations were performed in Python programming language on hardware configuration, Nvidia GeForce GTX 1080 graphics card, 32G memory, and Core i7 4790k 4GHz CPU. We used Caffe as a famous framework for creating, training, evaluating and deploying deep neural networks. In this article, deep models of pre-trained models in the ImageNet dataset are used for the initial weight.

8.3 Labeling experiments and objects detection

At this step, the combined method based on the deep network, the SVM-fuzzy and SVR method have been used. The fully-convolutional ResNet network was used in the RPN extraction section. The proposed method was compared with four existing methods, including Fast RCNN, Faster R-CNN, SPPNet [21] and R-FCN. Figure 6 shows examples of

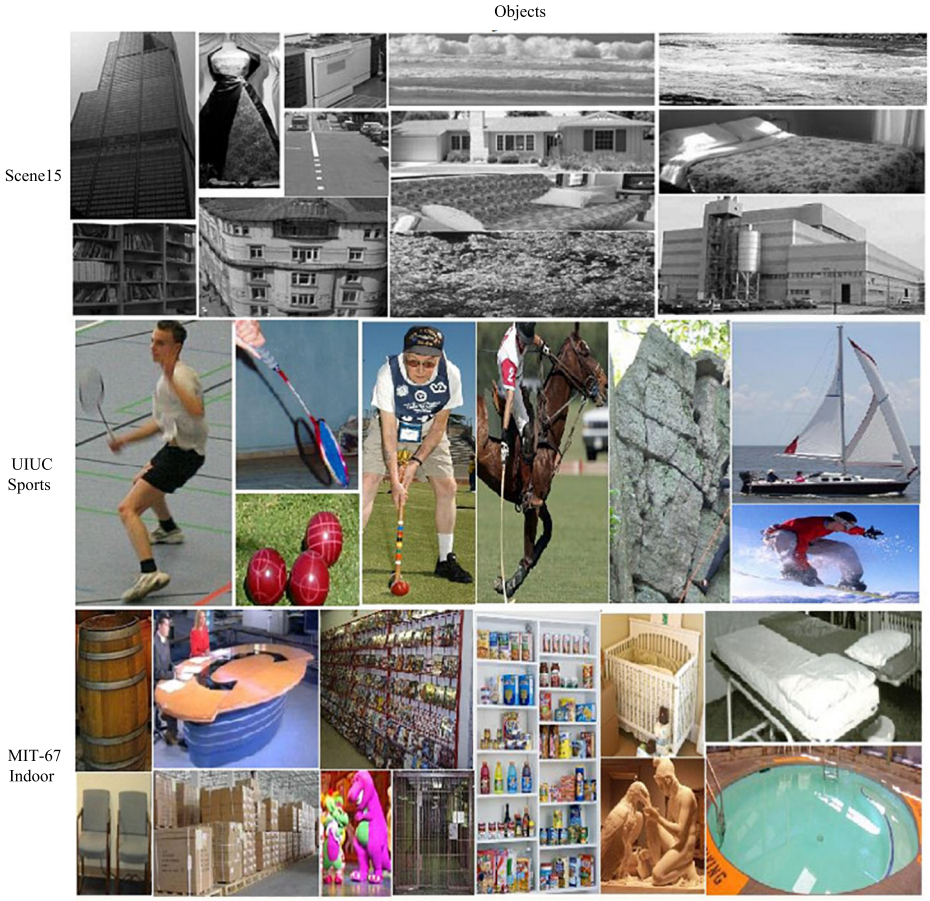

Fig. 6 Examples of objects extracted from the Scene15, UIUC Sports and MIT-67 Indoor datasets

Table 2 Comparison of R-FCN networks with different loss function in terms of mAP

Methods datasets	R-FCN with cross-entropy loss function	R-FCN with Cauchy-Schwartz loss function
Scene15	63.23%	**66.7%**
UIUC sports	77.12%	**81.15%**
MIT-67 indoor	41.70%	**48.11%**

The best result are shown in bold

objects in Scene15, UIUC Sports and MIT-67 indoor scenes. In these experiments, repetitive objects are selected in each category, and non-repetitive objects are not included in the training and testing. In Scene15, UIUC Sports and MIT-67 indoor scenes, 27, 14 and 103 objects are considered, respectively. In the proposed regions, regions with overlapping of over 50% with ground-truth are considered as correct. For a better evaluation of the results and showing performance of the proposed system, several experiments have been implemented.

A new loss function for labeling the objects was introduced to improve the proposed method. In the same way, the R-FCN network [11] with two loss functions is compared in Table 2. As shown in this table, the new loss function in all three sets improves the performance of the R-FCN method. Table 3 shows the results of the proposed method introduced in this phase (object detection) compared to other deep methods. As shown in this table, this method has a better performance than the other methods.

For a better evaluation of deep networks, the various architectures for object detection are shown in Table 4. ZF, VGG and ResNet have 5, 19 and 101 convolutional layers respectively. The results show that in most methods, the architecture of fully-convolution ResNet has a better performance. The latest experiment in this section is the evaluation of the number of ResNet network layers. Table 5 shows the results of the proposed method according to the ResNet architecture with 50, 101 and 152 layers. ResNet with the 101-layer architecture has the best performance. In this paper, the output of this architecture is used with 101 layers.

8.4 Extracting BoVW and topic features

As outlined in the proposed method, after regions extraction and labeling of object regions, the latent topics should be extracted in the image. This phase consists of two steps. In the first step, BOVWs are extracted, and in the next step, the latent topics are extracted based

Table 3 Comparison of different methods for detecting objects using mAP criteria

Datasets	Method				
	ResNet-SVMF-SVR	R-FCN	Fast RCNN	Faster R-CNN	SPPNet
Scene15	**74.9%**	63.23%	59.12%	61.40%	58.9%
UIUC sports	**89.45%**	77.12%	71.87%	72%	78.86%
MIT-67 indoor	**59.32%**	41.70%	39.20%	44.23%	45.4%

The best result are shown in bold

Table 4 Results of object detection in terms of mAP criteria with different architectures

Architectures datasets	ResNet	ZF	VGG
Scene15	**74.9%**	59.1%	61.77%
UIUC sports	**89.45%**	68.1%	77.21%
MIT-67 indoor	**59.32%**	44.5%	51.33%

The best result are shown in bold

on the visual words. In this phase, only two steps of the method presented in [62] are used for separate evaluation. In fact, the analysis and results of classification in this phase are performed by the method presented in [81]. The codebook size in the BoVW method can affect the accuracy of the categorization method. For this purpose, experiments of various sizes from the codebook are presented. The results of the experiments are shown in Fig. 7. It should be noted that the visual words, proportional to the number of clusters, were extracted from the image.

In the next step, the region related to the visual word is extracted and its label is considered as a feature. It should be noted that the number of topics in this phase is 100, and also in the use of PCA, the number of features is computed in (15). In this equation, λ represents the extracted eigenvalue, n represents dimension of features and k specifies the number of features that contain 98% of the total eigenvalue.

$$FeaturePCA = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \geq 0.98 \tag{15}$$

As shown in Fig. 7, the number of codebooks varies in all three datasets. The best number of codebooks is 300 in Scene15 dataset, 300 in UIUC Sports dataset and 450 in MIT-67 Indoor scenes dataset. The experiments are based on this number of codebooks. In clustering of the extracted regions from the previous stage, two views can be considered. In the first view, all regions are clustered regardless of the label of these regions. In the second view, these regions are clustered according to the labels of the regions. In this view, the SVR output label is also considered as input to this section. The purpose of clustering based on the second view is to cluster each region separately with the shared label and combining the clusters. The results of these two views are shown in Fig. 8a. As shown in this figure, the first view performs better than the second view. As mentioned, in this paper, we have used the improved k-means in [46]. In Fig. 8b, performance of the improved k-means and standard k-means are shown.

For the method presented in [81], the number of topic features is very influential in the classification accuracy. Similarly, experiments with a number of different topic features are presented, the results of which are shown in Fig. 9. As shown in this figure, the method in [81] with 240, 200 and 360 topics for the three datasets of Scene15, UIUC Sports and MIT-67 Indoor has the best performance. According to the experiments, the structure and various parameters of this phase are selected as in Table 6. It is shown in this table that the extracted features of the proposed architecture include the convolutional features associated with the

Table 5 Results of detection of objects in terms of mAP with ResNet architecture with different layers

Layers datasets	50	101	152
Scene15	70.23%	**74.9%**	73.21%
UIUC sports	85%	**89.45%**	86.7%
MIT-67 indoor	53.7%	**59.32%**	57.12%

The best result are shown in bold

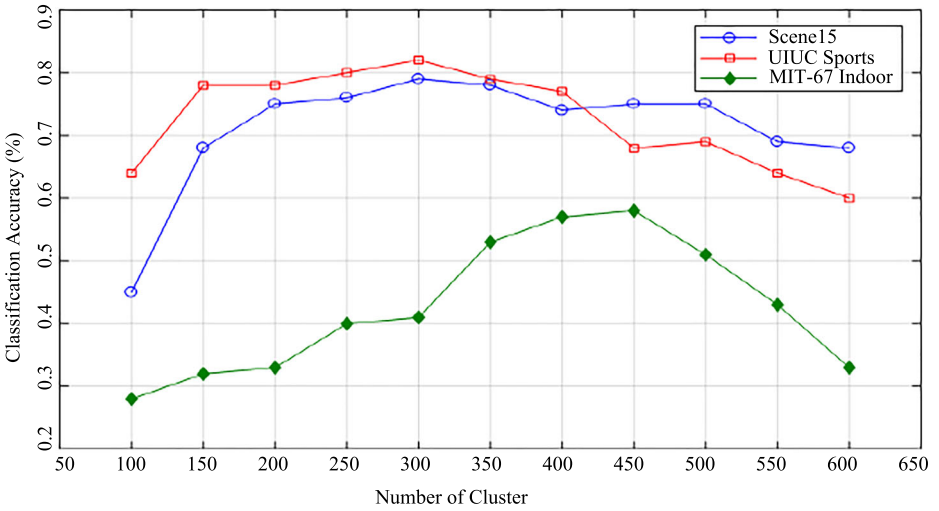

Fig. 7 Comparing the number of clusters

last two layers of the ResNet network after applying the PCA, the topic features, and the label of each visual word.

It should be noted that the number of visual words labels is equal to the size of the codebook. After extracting visual words, it is determined which visual words belong to which region, and then the label of each region (object label) is considered as a feature. In fact, we will have the object label corresponding to the number of visual words. Output of the proposed architecture output is equal to the number of scene classes. The number of extraction features from the last two convolutional layers for each dataset is 4500. In this paper, according to (15), after applying the PCA, these features are reduced to 380, 290, and 500 respectively for the dataset UIUC Sport, Scene15 and MIT-67 Indoor.

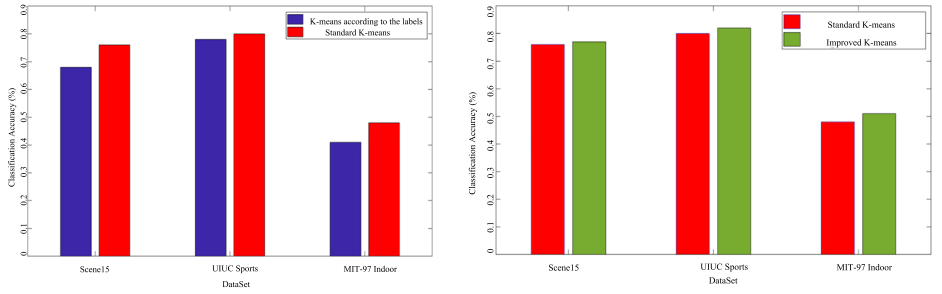

Fig. 8 Comparison of clustering methods, **a** two different perspectives: clustering with regard to label of regions and clustering regardless of labels; **b** comparing performance of the standard k-means method and the improved k-means

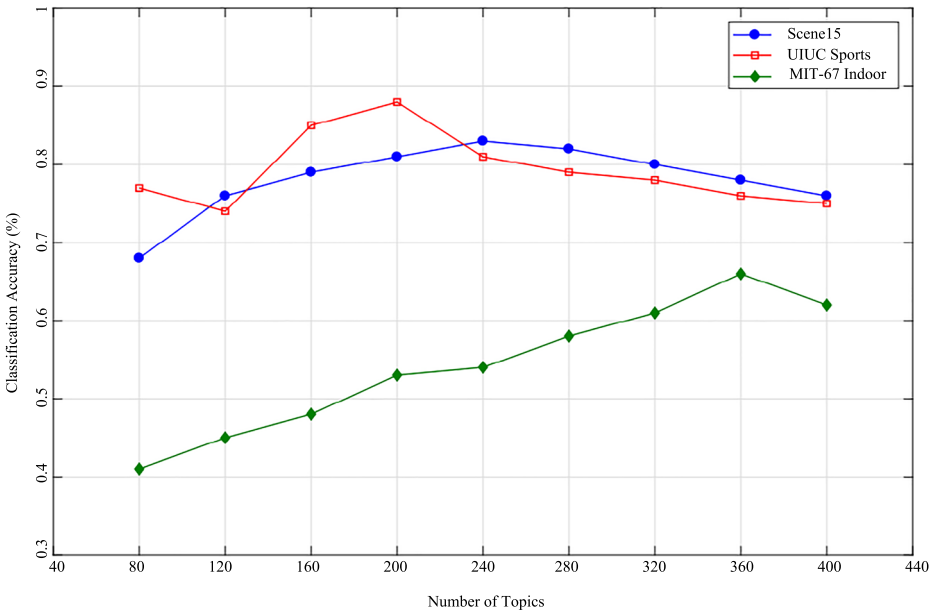

Fig. 9 Comparing the number of topics

8.5 Comparison with other baselines

In this section, the accuracy of image scene classification in the proposed method is compared with the famous methods in recent years. Tables 7, 8, and 9 show the results of different methods in the three datasets UIUC Sport, Scene15 and MIT-67 Indoor. As the results show, accuracy of image scene classification using the proposed method, in comparison to the best method on each datasets, is improved by 0.5, 0.6, and 1.8 percent respectively.

As noted, detection of objects is an important phase in recognizing the class of image scene. Additionally, the implementation time of this part also has a great impact when implementing the entire proposed system. In this regard, in this section, a comparison is made between the implementation time of the existing methods and the proposed method for detecting objects. The average test time for an image for objects detection in three sets

Table 6 Structure of input and output for the proposed network on three datasets

Datasets	Input			Output
	ResNet-Features	Topic Features	Lables of BoVW	
Scene15	$PCA\begin{pmatrix}3\times3, Conv, 512\\3\times3, Conv, 512\end{pmatrix}=380$	240	300	15
UIUC sport	$PCA\begin{pmatrix}3\times3, Conv, 512\\3\times3, Conv, 512\end{pmatrix}=290$	200	300	8
MIT-67 indoor	$PCA\begin{pmatrix}3\times3, Conv, 512\\3\times3, Conv, 512\end{pmatrix}=500$	360	450	67

Table 7 Comparing different methods for categorizing the images from the UIUC Sport dataset

UIUC 8-sport dataset	
Methods	Acc(%)
GIST-color [47]	70.7
MM-Scene [83]	71.7
Graphical [34]	73.4
Object Bank [35]	76.3
Object Attributes [36]	77.9
CENTRIST [74]	78.2
RSP [27]	79.6
SPM [32]	81.8
SPMSM [31]	83.0
Classemes [66]	84.2
HIK [73]	84.2
LScSPM [16]	85.3
LPR-RBF [58]	86.2
Hybrid Parts + GIST + SP [80]	87.2
LCSR [59]	87.2
VC + VQ [37]	88.4
IFV [68]	90.8
ISPR [39]	89:5
DRCF [30]	98.7
Proposed method	99.2

of data is shown in Table 10. As shown in the results of this table, the proposed method shows an appropriate performance in terms of runtime for all the three datasets compared to the other methods.

9 Discussion and future works

As the results of the proposed method have shown, the use of a comprehensive system based on deep neural network and topic models can greatly influence the accuracy of image scene classification. In the following, the future works that can improve the proposed method are suggested.

1. A very important phase in objects detection systems is the acquisition of region proposal, which in recent years the deep networks were used to calculate these regions. In this paper, convolutional networks based on architectures such as ResNet, ZF and VGGNet have been used to calculate region proposal. With regard to the results obtained with existing architectures, it is not possible to further improve the extraction of region proposal. Therefore, the presentation of a new architecture with a new structure can be effective in the accuracy of the proposed method.
2. In this paper, we used the visual words method based on the SIFT feature and the extended k-means clustering to calculate the bag of words. This method showed a good performance in extracting the bag of words. According to practical observations, only a small portion of the neurons are involved when images are received in the brain. Therefore, it can be seen that a higher level representation, and possibly sparse, is involved in

Table 8 Comparing different methods for categorizing the images from the Scene15 dataset

15-category scene dataset	
Methods	Acc(%)
GIST-color [47]	69.5
RBoW [49]	78.6
Classemes [66]	80.6
Object Bank [35]	80.9
SPM [32]	81.4
SPMSM [31]	82.3
LCSR [59]	82.7
SP-pLSA [4]	83.7
CENTRIST [74]	83.9
HIK [73]	84.1
OTC [44]	84.4
ISPR [39]	85.1
VC + VQ [37]	85.4
LMLF [5]	85.6
LPR-RBF [58]	85.8
Hybrid Parts + GIST + SP [80]	86.3
CENTRIST+LCC+Boosting [77]	87.8
RSP [27]	88.1
IFV [68]	89.2
LScSPM [16]	89.7
DRCF [30]	94.5
Proposed method	95.1

this process. In addition, experiments in this paper indicate that the extracted matrix of visual words is sparse. In fact, for this type of systems, the sparse coding method can be used by changing the initial coding mechanism. Although its consideration can reduce the reconstruction error, the coding was performed regardless to the sparse parts of the image in this research.

3. Extracting the latent topics in the image can greatly enhance the accuracy of the image classification systems. Due to the methods introduced in recent years and the experiments performed, these methods have shown good accuracy compared to the previous methods. But the time complexity of this type of method is very high. In this regard, a new method based on the topic models that has a better run time can be very important in this application.
4. One of the most important applications of machine vision is the use of content-based systems for auto-image description. As a human after analyzing the content of the image describes it automatically, it can use the system presented in this article for the next step, such as auto-image description. In fact, this method can be used in Semantic-Based Image Retrieval (SBIR). At SBIR, the user inputs his meaning as a text phrase and looks for images with the contents related to that phrase. In fact, in this system, the relationship between the meaning and visual content of the images is examined. The most logical way to search for high semantic meanings in SBIR systems is to assign a meaning-based text label to images in the database and compare those labels with the search term to retrieve the related images. In the following, we can use the output of the

Table 9 Comparing different methods for categorizing the images from the MIT-67 Indoor dataset

MIT-67 indoor scene dataset	
Methods	Acc(%)
ROI + GIST [53]	26.1
MM-Scene [83]	28.3
SPM [32]	34.4
Object Bank [35]	37.6
RBoW [49]	37.9
Weakly Supervised DPM [48]	43.1
SPMSM [31]	44.0
LPR-LIN [58]	44.8
BoP [28]	46.1
Hybrid Parts + GIST + SP [80]	47.2
OTC [44]	47.3
Discriminative Patches [60]	49.4
ISPR [39]	50.1
D-Parts [64]	51.4
VC + VQ [37]	52.3
IFV [68]	60.8
MLRep [13]	64.0
CNN-MOP [19]	68.9
CNNaug-SVM [54]	69.0
DRCF [30]	71.8
Proposed method	73.6

proposed method for labeling in the SBIR system. Content labels from the image are extracted by the proposed method and applied as input to the SBIR system.

5. Datasets with multiple objects are used in this research. Performance of the proposed method highly dependents on regions extraction of these objects. In fact, visual words are derived from the regions associated with these objects. In the existing methods of image content classification, the entire image is used to extract visual words. Indeed, in these methods, visual words are computed based on local regions extracted from the entire image. One can develop a technique to combine the proposed method with a method based on local regions extraction, not dependent to objects label, for image scene classification.

Table 10 Comparison of the average implementation time of the test stage

Datasets	Methods				
	Fast RCNN [17]	Faster RCNN [57]	SPPNet [21]	R-FCN [11]	Proposed method
Scene15	0.49	0.43	0.38	0.17	**0.13**
UIUC sport	0.34	0.32	0.31	0.18	**0.14**
MIT-67 indoor	0.48	0.45	0.36	0.27	**0.2**

The best result are shown in bold

10 Conclusion

Obtaining concept from an image can increase the accuracy of image classification. In fact, getting the semantic high-level concept provides robust information about the image content. In this regard, this paper presents a new approach for categorizing content-based image scenes. This article presented a comprehensive system for image scene classification based on deep networks and latent topics. Image content can be interpreted based on objects in the image. In this regard, a method for extracting objects was introduced. It has been shown that using deep networks and choosing its proper architecture can be very effective in extracting objects in the image. The ResNet architecture in the R-FCN deep network has shown a proper performance in detecting objects. For this reason in this article, a method to improve this network based on the fully-convolutional deep network, fuzzy SVM and SVR was presented to extract regions and objects label. In addition, a new loss function was also used. The results of experiments on Scene15, UIUC Sport and MIT-67 Indoor datasets with 27, 14 and 103 objects represent the proper performance of extracting the object. In the proposed method, a fuzzy SVM network was introduced to separate the object from the background. This has greatly improved the speed of object detection. Fully-convolutional network layers have semantically valuable information on their own which in this paper, the last two-layer convolution information was used after applying the PCA feature reduction method. An important part of this article is the extraction of the latent topics in the images. In this regard, SupDocNADE and BoVW based approaches were presented. Different perspectives for BoVW extraction have been introduced and analyzed. Ultimately, the method based on the SIFT visual features and improved k-means algorithms have shown the best performance. The topic features layer is used to extract the latent topics in the image. It has been shown that the number of topics can be very effective in categorizing accuracy. Eventually, classification of the scene was performed using a Bayesian classification based on extraction properties of deep network layers, object labels and topics. The proposed method has been evaluated on Scene15, UIUC Sport and MIT-67 Indoor dataset, and it has been shown that this method has the best performance compared to other existing methods.

References

1. Bishop C (2006) Pattern Recognition and Machine Learning (Information Science and Statistics) chapter 3:138–147
2. Blei DM, Jordan MI (2003) Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. ACM, pp 127–134
3. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
4. Bosch A, Zisserman A, Muñoz X (2008) Scene classification using a hybrid generative/discriminative approach. IEEE Trans Pattern Anal Mach Intell 30(4):712–727
5. Boureau Y-L, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognition. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 2559–2566
6. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Mining Knowl Discov 2(2):121–167
7. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann AG (2016) Bi-level semantic representation analysis for multimedia event detection. IEEE Trans Cybern 47(5):1180–1197
8. Chang X, Yu Y-L, Yang Y, Xing EP (2016) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell 39(8):1617–1632
9. Chen S-G, Wu X-J (2017) A new fuzzy twin support vector machine for pattern classification. Int J Mach Learn Cybern, 1–12
10. Czarnecki WM, Jozefowicz R, Tabor J (2015) Maximum entropy linear manifold for learning discriminative low-dimensional representation. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 52–67

11. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387
12. Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 2974–2983
13. Doersch C, Gupta A, Efros AA (2013) Mid-level visual element discovery as discriminative mode seeking. In: Advances in neural information processing systems, pp 494–502
14. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338
15. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC Dssd: deconvolutional single shot detector, arXiv:1701.06659
16. Gao S, Tsang IW-H, Chia L-T, Zhao P (2010) Local features are not lonely–Laplacian sparse coding for image classification. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3555–3561
17. Girshick R Fast r-cnn, arXiv:1504.08083
18. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
19. Gong Y, Wang L, Guo R, Lazebnik S (2014) Multi-scale orderless pooling of deep convolutional activation features. In: European conference on computer vision. Springer, pp 392–407
20. Guillaumin M, Verbeek J, Schmid C (2010) Multimodal semi-supervised learning for image classification. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 902–909
21. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
23. Ioffe S, Szegedy C Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167
24. Isa D, Lee LH, Kallimani V, Rajkumar R (2008) Text document preprocessing with the Bayes formula for classification using the support vector machine. IEEE Trans Knowl Data Eng 20(9):1264–1272
25. Janocha K, Czarnecki WM On loss functions for deep neural networks in classification, arXiv:1702.05659
26. Jia Y, Salzmann M, Darrell T (2011) Learning cross-modality similarity for multinomial data. In: 2011 IEEE International conference on computer vision (ICCV). IEEE, pp 2407–2414
27. Jiang Y, Yuan J, Yu G (2012) Randomized spatial partition for scene recognition. In: Computer vision–ECCV 2012. Springer, pp 730–743
28. Juneja M, Vedaldi A, Jawahar C, Zisserman A (2013) Blocks that shout: distinctive parts for scene classification. In: 2013 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 923–930
29. Keller JM, Hunt DJ (1985) Incorporating fuzzy membership functions into the perceptron algorithm. IEEE Trans Pattern Anal Mach Intell 6:693–699
30. Khan SH, Hayat M, Bennamoun M, Togneri R, Sohel FA (2016) A discriminative representation of convolutional features for indoor scene recognition. IEEE Trans Image Process 25(7):3372–3383
31. Kwitt R, Vasconcelos N, Rasiwasia N (2012) Scene recognition on the semantic manifold. In: European conference on computer vision. Springer, pp 359–372
32. 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
33. Lenc K, Vedaldi A R-cnn minus r, arXiv:1506.06981
34. Li L-J, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. In: IEEE 11th International conference on computer vision, 2007. ICCV 2007. IEEE, pp 1–8
35. Li L-J, Su H, Fei-Fei L, Xing EP (2010) Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems, pp 1378–1386
36. Li L-J, Su H, Lim Y, Fei-Fei L (2010) Objects as attributes for scene classification. In: European conference on computer vision. Springer, pp 57–69
37. Li Q, Wu J, Tu Z (2013) Harvesting mid-level visual concepts from large-scale internet images. In: 2013 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 851–858
38. Li Y, Chen Y, Wang N, Zhang Z Scale-aware trident networks for object detection, arXiv:1901.01892

39. Lin D, Lu C, Liao R, Jia J (2014) Learning important spatial pooling regions for scene classification. In: 2014 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3726–3733
40. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
41. Liu L, Wang L, Liu X (2011) In defense of soft-assignment coding. In: 2011 IEEE International conference on computer vision (ICCV). IEEE, pp 2486–2493
42. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
43. Liu X, Zhang R, Meng Z, Hong R, Liu G (2019) On fusing the latent deep CNN feature for image classification. World Wide Web 22(2):423–436
44. Margolin R, Zelnik-Manor L, Tal A (2014) Otc: a novel local descriptor for scene classification. In: European conference on computer vision. Springer, pp 377–391
45. Mesnil G, Rifai S, Bordes A, Glorot X, Bengio Y, Vincent P (2015) Unsupervised learning of semantics of object detections for scene categorization. In: Pattern recognition applications and methods. Springer, pp 209–224
46. Na S, Xumin L, Yong G (2010) Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 Third International symposium on intelligent information technology and security informatics (IITSI), IEEE, pp 63–67
47. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
48. Pandey M, Lazebnik S (2011) Scene recognition and weakly supervised object localization with deformable part-based models. In: 2011 IEEE International conference on computer vision (ICCV). IEEE, pp 1307–1314
49. Parizi SN, Oberlin JG, Felzenszwalb PF (2012) Reconfigurable models for scene recognition. In: 2012 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 2775–2782
50. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
51. Putthividhy D, Attias HT, Nagarajan SS (2010) Topic regression multi-modal latent Dirichlet allocation for image annotation. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3408–3415
52. Qi X, Li C-G, Zhao G, Hong X, Pietikäinen M (2016) Dynamic texture and scene classification by transferring deep image features. Neurocomputing 171:1230–1241
53. Quattoni A, Torralba A (2009) Recognizing indoor scenes. In: 2009 IEEE Conference on computer vision and pattern recognition, 2009. CVPR. IEEE, pp 413–420
54. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: 2014 IEEE Conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 512–519
55. Redmon J, Farhadi A Yolo9000: better, faster, stronger, arXiv preprint
56. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
57. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
58. Sadeghi F, Tappen MF (2012) Latent pyramidal regions for recognizing scenes. In: European Conference on computer vision. Springer, pp 228–241
59. Shabou A, LeBorgne H (2012) Locality-constrained and spatially regularized coding for scene categorization. In: 2012 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3618–3625
60. Singh S, Gupta A, Efros AA (2012) Unsupervised discovery of mid-level discriminative patches. In: Computer vision–ECCV 2012. Springer, pp 73–86
61. Socher R, Fei-Fei L (2010) Connecting modalities: semi-supervised segmentation and annotation of images using unaligned text corpora. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 966–973
62. Srivastava N, Salakhutdinov RR (2012) Multimodal learning with deep Boltzmann machines. In: Advances in neural information processing systems, pp 2222–2230
63. Srivastava N, Salakhutdinov RR (2013) Discriminative transfer learning with tree-based priors. In: Advances in neural information processing systems, pp 2094–2102
64. Sun J, Ponce J (2013) Learning discriminative part detectors for image classification and cosegmentation. In: 2013 IEEE International conference on computer vision (ICCV). IEEE, pp 3400–3407

65. Sun C, Paluri M, Collobert R, Nevatia R, Bourdev L (2016) Pronet: learning to propose object-specific boxes for cascaded neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3485–3493
66. Torresani L, Szummer M, Fitzgibbon A (2010) Efficient object category recognition using classemes. In: European conference on computer vision. Springer, pp 776–789
67. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
68. Vedaldi A, Fulkerson B (2010) Vlfeat: an open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM international conference on multimedia. ACM, pp 1469–1472
69. Wang D, Mao K (2019) Task-generic semantic convolutional neural network for web text-aided image classification. Neurocomputing 329:103–115
70. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: 2013 IEEE International conference on computer vision (ICCV). IEEE, pp 3551–3558
71. Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British machine vision conference. BMVA Press, pp 124–1
72. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: 2010 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 3360–3367
73. Wu J, Rehg JM (2009) Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 630–637
74. Wu J, Rehg JM (2011) Centrist: a visual descriptor for scene categorization. IEEE Trans Pattern Anal Mach Intell 33(8):1489–1501
75. Wu R, Wang B, Wang W, Yu Y (2015) Harvesting discriminative meta objects with deep cnn features for scene classification. In: 2015 IEEE International conference on computer vision (ICCV). IEEE, pp 1287–1295
76. Yen S-J, Wu Y-C, Yang J-C, Lee Y-S, Lee C-J, Liu J-J (2013) A support vector machine-based context-ranking model for question answering. Inform Sci 224:77–87
77. Yuan J, Yang M, Wu Y (2011) Mining discriminative co-occurrence patterns for visual recognition. In: 2011 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE, pp 2777–2784
78. Zhang F, Du B, Zhang L (2015) Saliency-guided unsupervised feature learning for scene classification. IEEE Trans Geosci Remote Sens 53(4):2175–2184
79. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Single-shot refinement neural network for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4203–4212
80. Zheng Y, Jiang Y-G, Xue X (2012) Learning hybrid part filters for scene recognition. In: European conference on computer vision. Springer, pp 172–185
81. Zheng Y, Zhang Y-J, Larochelle H (2016) A deep and autoregressive approach for topic modeling of multimodal data. IEEE Trans Pattern Anal Mach Intell 38(6):1056–1069
82. Zhou X, Yu K, Zhang T, Huang TS (2010) Image classification using super-vector coding of local image descriptors. In: European conference on computer vision. Springer, pp 141–154
83. Zhu J, Li L-J, Fei-Fei L, Xing EP (2010) Large margin learning of upstream scene understanding models. In: Advances in neural information processing systems, pp 2586–2594

Ali Ghanbari Sorkhi received the B.S. degree in computer engineering from Iran University of Science and Technology, Tehran, Iran, in 2010, the M.S. degree computer engineering from Shahrood University of Technology, Shahrood, Iran, in 2012, and since 2014 to now has Ph.D. student of Shahrood University of Technology, Shahrood, Iran. His research interests include image processing, text syntax analyzing and deep learning.

Hamid Hassanpour received the B.S. degree in computer engineering from Iran University of Science and Technology, Tehran, Iran, in 1993, the M.S. degree in computer engineering from Amirkabir University of Technology, Tehran, Iran, in 1996, and the Ph.D. from the Queensland University of Technology, Brisbane, Australia, in 2004. His research interests include biomedical signal processing, time-frequency signal processing and analysis, new architectures in computer design, text syntax analyzing and image processing.

Mansoor Fateh received the B.S. degree in electronic engineering from Shahrood University of Technology, Shahrood, Iran, in 2004, the M.S. degree Biomedical engineering form Tarbiat Modares University, Tehran, Iran, in 2009, and the Ph.D. from Tarbiat Modares University, Tehran, Iran, in 2014. His research interests include reinforcement learning and image processing.