Multimedia Tools and Applications (2019) 78:15751–15777
https://doi.org/10.1007/s11042-018-7031-0

Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features

Muhammad Rashid[1] · **Muhammad Attique Khan**[2] · **Muhammad Sharif**[1] · **Mudassar Raza**[1] · **Muhammad Masood Sarfraz**[3] · **Farhat Afza**[1]

Received: 3 May 2018 / Revised: 22 November 2018 / Accepted: 3 December 2018 /
Published online: 8 December 2018

Abstract

In the area of machine learning and pattern recognition, object classification is getting an attraction due to its range of applications such as visual surveillance. In recent times, numerous deep learning-based methods are presented for object classification but still, set of problems/ concerns exists which reduce the overall classification accuracy. Complex background, congest situtaions, and similarity among different objects are few challenging issues. To tackle such problems, we propose a technique by using deep convolutional neural network (DCNN) and scale invariant features transform (SIFT). First, an improved saliency method is implemented, and the point features are extracted. Then, DCNN features are extracted from two deep CNN models like VGG and AlexNet. Thereafter, Reyni entropy-controlled method is implemented on DCNN pooling and the SIFT point matrix to select the robust features. Finally, the selected robust features are fused in a matrix by a serial approach, which is later fed to ensemble classifier for recognition. The proposed method is evaluated on three publically available datasets including Caltech101, Barkley 3D, and Pascal 3D and obtained classification accuracy of 93.8%, 99%, and 88.6% - clearly showing the exceptional performance compared to existing methods.

Keywords Object detection · Hand crafted features · Deep CNN · Features fusion · Features reduction · Classification

✉ Muhammad Attique Khan
attique.khan440@gmail.com

[1] Department of Computer Science, COMSATS University Islamabad, Wah Campus, Islamabad, Pakistan

[2] Department of Computer Science and Engineering, HITEC University, Museum Road, Taxila, Pakistan

[3] Department of Electrical Engineering, COMSATS University Islamabad, Wah Campus, Islamabad, Pakistan

1 Introduction

Object detection and classification are challenging tasks in the domain of computer vision (CV), which are used to classify the objects according to their class labels. This domain gets much attention due to its enormous applications including video surveillance, target recognition, face detection, optical character recognition, video stabilization, image watermarking, plants diseases recognition, and automated pedestrian detection [24, 48, 51, 52]. Recently, promising results are achieved in this area when dealing with simple images with transparent background. But in few cases, where objects contain a complex background, multiple shapes, and under congest situations [46], it require few enhancements.

Researchers work in this domain from last two decades and try to categorize the challenging problem of object detection and classification including complex background, features extraction, best features selection, execution time and accurate classification. They introduce several features extractions-based methods to detect and classify complex objects using classification methods. The famous features which are used for object classification include color [39, 46], shape (Histogram Oriented Gradients (HOG) [49], texture (Local Binary Pattern (LBP)) [59], local & global (SIFT, PCA), Bag of Features (BoF) [40] and also cover deep features (convolutional neural network (CNN)) [3]. Lazebnik et al. [28] introduce a new technique to deal with this limitation by BoF and made a Spatial Pyramid Matching (SPM) technique, which can divide the image into spatial sub-regions, and computes the histogram of each sub-region, which is later used for the creation of a spatial location sensitive vector. In addition to feature extraction, a fusion strategy is also adopted by several researchers to take advantage of distinct patterns of various descriptors which increases the classification accuracy [31, 34, 35]. The noticeable fusion methods are serial and parallel that is used in several domains like medical imaging, video surveillance, biometrics, and few more. Stochastic discriminate analysis (SDA) [22], fusion of low level and mid-level features, and transfer based features fusion techniques are the eminent algorithms in this domain that enhance the recognition accuracy [37].

To overcome the limits of conventional algorithms of features extraction like handcrafted approaches, deep learning is a suitable candidate for this domain due to its legitimate ability. CNN is a subtype of deep architecture [5]. It has shown improved performance for classification and recognition with successful applications such as machine learning and pattern recognition [41]. Several pre-trained deep CNN (DCNN) models are introduced by several researchers such as VGG [55], AlexNet [27], ResNet [19], YOLO model, and GoogleNet [58]. These model are implemented in several directions such as image classification [47], action recognition [36], medical imaging [10], agricultural plants, and a few more [25]. Jun et al. [38] presented an image classification method based on Group Sparse Deep Stack Network (GS-DSN). The method consists of two modules. In the first module, the interdependencies amongst hidden units are acquired by splitting them into amalgamate groups whereas the second module splits image description into sub-groups to design for clustering of each sample and later gradient descent is used for estimation of weights. The pre-trained VGG CNN models are used for features extraction and classified by group sparse network module (GSNM). Wei et al. [60] presented an image classification method through intra-class CNN feature pyramid. The advantage of lower level layers is to get the structural information and the higher-level layers to acquire the semantic information. Later, AlexNet and VGG16 pre-trained CNN models are utilized for features extraction and give an improved performance on the Caltech-101 dataset.

Recently, feature reduction and selection techniques have been gained much attention in the domain ML because a high number of features reduces the classification accuracy and also increases the computation time [26]. The major aim of feature reduction step is to solve the issue of the curse of dimensionality. The most existing feature reduction methods are Entropy-based feature reduction [16], Genetic Algorithm (GA) [8], Particle Swarm Optimization (PSO) [2, 44], and Canonical Correlation Analysis (CCA) [42]. Jinjoo et al. [57] introduce a new technique for the curse of dimensionality reduction using Structured Sparse Principle Component Analysis (SSPCA). The features are extracted through SIFT points, and optimal features are obtained through the SSPCA approach. Yongsheng et al. [45] describe a decomposition technique, which is used for encryption of frequency and spatial information. Through this approach, break down of the input image is performed into sub-regions using Spatial Pyramid Matching (SPM). The SIFT features are extracted from smaller regions in the initial stage and then by using codebook, the global features are extracted. Later, irrelevant features are reduced using K-means clustering and obtained maximum classification accuracy of 85.78% on the Caltech-101 dataset. Shuangshuang et al. [6] suggested a new sampling-based method for object classification. Three steps are performed in this method such as random, saliency-based, and dense sampling. The objects are categorized into semantic groups using these sampling methods. Thereafter, a supervised dimensionality reduction approach is provided, which removes the irrelevant features and only selects the best features for classification. The introduced method is validated on the STL-10 dataset and achieved a classification accuracy of 67%. In [29], a dynamic weighted discrimination power analysis method is introduced for selects the best discriminant coefficients for achieving the best recognition accuracy. The coefficients are selected according to their distinction strength. Lu et al. [30] introduce a dynamic weighted discriminant power analysis (DWDPA) approach for selection of best DCT features. The pre-masking transom is not required in DWDPA because this approach selects features through high power discrimination. The experiments are performed on three datasets which show significant recognition accuracy. Moreover, several other techniques are also introduced in literature for features reduction such as random projection (RP) [32], 2-dimensional RP [33], and few more [4, 53, 54]. The major advantage of features reduction and optimal features selection is to achieve maximum recongition rate in minimum computational time. The reduced features are finally classified by supervised and unsupervised learning methods such as Linear Support Vector Machine (L-SVM) [15, 56], Cubic-SVM (C-SVM), Quadratic-SVM (Q-SVM), Fine K-Nearest Neighbor (F-KNN), Cubic-KNN (C-KNN), deep learning, Ensemble Subspace-KNN (EKNN) [23, 48], Bayesian model, and Random forest and Naive based classifiers [17].

The above-discussed methods do not work well when dealing with larger datasets which include hundreds of classes. The preprocessing step is very important which is never performed in existing object classification studies to improves the classification accuracy. The preprocessing is an important step due to background factors such as illumination. Moreover, we notice that in [45, 57] SIFT features are extracted from input images but they achieved maximum accuracy of 85.78% on the Caltech101 dataset. As before, no one fuses two pre-trained DCNN models because each model has a different number of inputs, which makes the problem for fusion. Moreover, patterns fusion of two DCNN models provides better classification performance as compared to the individual model. To inspire this approach, in this article, we propose a new DCNN based method for object classification from static images. The proposed method is implemented in two parallel steps. An improved saliency-based method is proposed in the first step and SIFT point features are extracted. Then, VGG and

AlexNet based pre-trained DCNN models are used to extract deep CNN features by employing activation on a fully connected (FC) layer. Thereafter, Reyni entropy-controlled method is proposed, which is employed on DCNN and SIFT Point feature matrices to selects the best features. But the size of inputs for each model is different, which becomes a problem in the fusion process. To resolve this problem, we perform the augmentation and make the both matrices equal in size. Then both feature matrices are fused by using a serial-based method and features are stored in a new matrix, which is later fed to ensemble classifier for classification.

2 Materials and methods

In this research, we used three famous datasets such as Caltech101, PASCAL 3D, and 3D dataset to deal with complex object detection and classification. These datasets contain hundreds of object classes and thousands of images. To overcome the challenges of these datasets such as illumination, color, and similarity among various object classes, we propose a new method for object classification based on DCNN features extraction along with SIFT points. The proposed method consists of two major steps, which are executed in parallel. In the first step, SIFT point features are extracted from mapped RGB segmented objects. In second step, DCNN features are extracted through pre-trained CNN models such as AlexNet and VGG. The both SIFT point and DCNN features are combined into one matrix by a parallel fusion method and the best features are selected for final classification. The detailed description of each step is given below in section 3.2 to 3.4. The comprehensive flow diagram is presented in Fig. 1.

2.1 Improved saliency method

An improved saliency method is employed by utilizing existing saliency approach name HDCT, for single object detection. In this step, we extract a single object from an image by an existing saliency method namely HDCT saliency estimation. The idea behind the

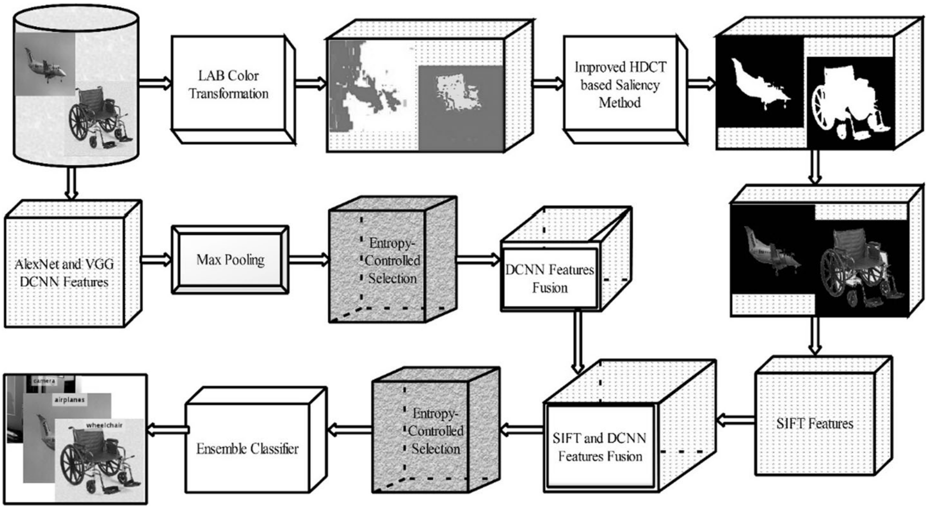

LAB Color Transformation
Improved HDCT based Saliency Method
AlexNet and VGG DCNN Features
Max Pooling
Entropy-Controlled Selection
DCNN Features Fusion
camera
airplanes
wheelchair
Ensemble Classifier
Entropy-Controlled Selection
SIFT and DCNN Features Fusion
SIFT Features

Fig. 1 Flow diagram of the proposed object classification method

improvement of saliency method is to implement the color spaces before it gives the input image to the saliency method. The LAB color transformation is utilized for this purpose, which identifies color in 3 dimensions consisting of L* for lightness, a*, and b* are utilized for color components green-red and blue-yellow respectively. The components L* is brighter white at 100 and darker black at 0, whereas 'a' and 'b' channels show the natural values for the RGB image. This transformation is defined as follows:

Let $U(i,j)$ denotes an input RGB image having length $N \times M$, then for RGB to LAB conversion, first RGB to XYZ conversion is performed through Eqs. 1–10:

$$\begin{bmatrix} \varphi(X) \\ \varphi(Y) \\ \varphi(Z) \end{bmatrix} = [M \times N] \begin{bmatrix} \varphi^r \\ \varphi^g \\ \varphi^b \end{bmatrix} \tag{1}$$

where $\varphi(X)$, $\varphi(Y)$, and $\varphi(Z)$ denote the X, Y, and Z channels, which are extracted from red (φ^r), green (φ^g), and blue channel (φ^b). The φ^r, φ^g, and φ^b channels are defined as:

$$\varphi^r = \sum_{k=1} \frac{\varphi k}{\vartriangle_k}, k = \mathrm{Re}d \tag{2}$$

$$\varphi^g = \sum_{k=2} \frac{\varphi k}{\vartriangle_k}, k = Green \tag{3}$$

$$\varphi^b = \sum_{k=3} \frac{\varphi k}{\vartriangle_k}, k = Blue \tag{4}$$

Then LAB conversion is defined as:

$$\left(\varphi^L = \beta_1 \times \left(f_y - 16\right)\right), \beta_1 = 116 \tag{5}$$

$$\left(\varphi^{*A} = \beta_2\left(f_x - f_y\right)\right), \beta_2 = 500 \tag{6}$$

$$\left(\varphi^{*B} = \beta_3\left(f_y - f_z\right)\right), \beta_3 = 200 \tag{7}$$

where, f_x, f_y, and f_z are linear functions which are computed as:

$$f_x = \left\{ \sqrt[3]{x_r} | \frac{kx_r + 16}{116}, \rightarrow x_r > \in | \ otherwise \right\}, x_r = \frac{X}{Xr} \tag{8}$$

$$f_y = \left\{ \sqrt[3]{y_r} | \frac{ky_r + 16}{116}, \rightarrow y_r > \in | \ otherwise \right\}, y_r = \frac{Y}{Yr} \tag{9}$$

$$f_z = \left\{ \sqrt[3]{z_r} \middle| \frac{kz_r + 16}{116}, \rightarrow z_r > \in \middle| \, otherwise \right\}, z_r = \frac{Z}{Zr} \tag{10}$$

Thereafter, we employ a saliency approach for salient object detection. Salient region detection technique detects the salient region from an image by utilizing a high dimensional color transform. In this work, the superpixel saliency features are used to detect the initial salient regions of the dermoscopic images. The superpixels of the LAB image are formulated by Eq. 11:

$$Y = \{p_1, \ldots .p_N\} \tag{11}$$

For low computational cost and exceptional performance, we utilize the SLIC superpixel [1] with a total number of $N = 400$ superpixels. The color features are computed from LAB color space. The parameters which are used for color features extraction from LAB color space are mean, variance, standard deviation, and skewness. These color features are concatenated with the histogram features because the histogram features are effective for saliency approach. The euclidean distance is calculated between extracted color features by Eq. 12:

$$\overrightarrow{D} = \overrightarrow{D}(A) = \left\| l_i - l_j \right\|_2^2 \tag{12}$$

where l_i and l_j denote the ith and jth features in the given matrix A. In this work, the global contrast/color statistics of objects are used to define the saliency values of the pixels by using a histogram-based method. The saliency values of pixels are defined by Eq. 13:

$$S(\varphi_k) = \sum\nolimits_{\forall \varphi_i \in I} \overrightarrow{D}(A) \tag{13}$$

where $\overrightarrow{D}(A)$ is the color distance between the features l_{ii} and the l_j in the LAB color space. By rearranging the above equation, we get the saliency value for each color by Eq. 14:

$$S(\varphi_k) = \sum_{l=1}^{n} f_l D(c_j, c_l) \tag{14}$$

where n, c_j, f_ldenote the total number of the different pixel color, the color value of the pixe l φ_k, and the frequency of the color pixel respectively. The HOG and the SFTA texture features, are utilized for shape and texture features. After the calculation of the feature vector for each superpixel, the random forest regression is used to estimate the salient degree of each region. Further to identify the very salient pixels calculated from initial saliency map, the Trimap is constructed by using adaptive thresholding. First, the input images divided into 2 × 2, 3 × 3, and 4 × 4 patches and then the Otsu thresholding is apply on each patch individually. Finally, the Trimap is obtained by using global thresholding which is formulated by Eq. 15:

$$T(i) = \begin{cases} 1 \rightarrow T(i) \geq \tau \\ 0 \rightarrow T(i) \leq \tau \\ unknown \ldots else \end{cases} \tag{15}$$

Where τ denotes the global threshold value. After getting the optimal coefficient α (estimate for the saliency map) it constructs the saliency map as follow:

$$S_{LS}(X_i) = \sum_{j=1}^{l} K_{ij}\alpha_j, i = 1, 2, ..., N \tag{16}$$

Where K denotes the high dimensional vector to present the color of the input image. The final map is obtained by adding the spatial and color-based saliency map through Eq. 17:

$$S_{final}(X_i) = S_{LS}(X_i) + S_S(X_i), i = 1, 2, ..., N \tag{17}$$

The final spatial saliency map is defined by Eq. 18:

$$S_S(X_i) = \exp\left(-K \frac{\min_{j\in f}\left(d\left(P_i, P_j\right)\right)}{\min_{j\in\beta}\left(d\left(P_i, P_j\right)\right)}\right) \tag{18}$$

where the K = 0.5, and $min_j \in \beta(d(P_i, P_j))$ and $min_j \in f(d(P_i, P_j))$ are the Euclidian distance from the ith pixel to definite background pixel and to definite foreground pixel respectively. The improved saliency method effects are shown in Fig. 2. In Fig. 2, the 1st row shows input images, second rows present LAB transformation, third row defines improved saliency image in a binary form, and the last row depicts the mapped RGB image.

2.2 SIFT features

Scale Invariant Feature Transform (SIFT) is originally designed in 2004 by [43] and have appeared as a strength descriptors for object detection and recognition. The SIFT features are computed in four steps. In the first step, local key points are determined that are important and stable for given images. Then features are extracted from each key point that explains the local image region samples, which are related to its scale space coordinate image. In the second step, weak features are removed by a specific threshold value. In the third step, orientations are assigned to each key point based on local image gradient directions. Finally, the 1 × 128 dimensional feature vector is extracted, and bi-linear interpolation is performed to improve the robustness of features. The above theory is defined through Eqs. 19–21:

$$\xi(\mu, \nu, \sigma) = \psi_G(\mu, \nu, \sigma) \otimes S_{final}(X_i) \tag{19}$$

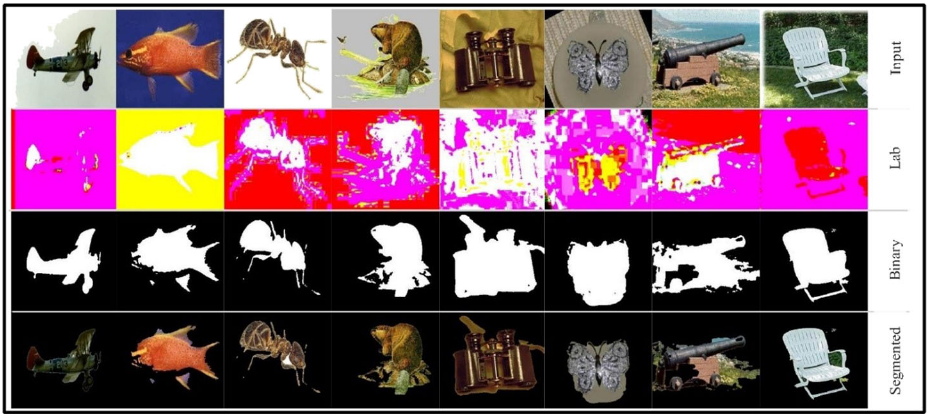


Fig. 2 Proposed improved saliency method results

$$\psi_G(\mu,\nu,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{\mu^2+\nu^2}{2\sigma^2}\right)} \tag{20}$$

$$\begin{aligned} &D(\mu,\nu,\sigma) = (\psi_G(\mu,\nu,k\sigma) - \psi_G(\mu,\nu,\sigma)) \otimes S_{final}(X_i) = \\ &\xi(\mu,\nu,k\sigma) - \xi(\mu,\nu,\sigma) \end{aligned} \tag{21}$$

where $\xi(u, v, \sigma)$ is scale space of an image, $\psi_G(u, v, k\sigma)$ denotes the variable-scale Gaussian, k is a multiplicative factor and $D(u, v\ \sigma)$ denotes the difference of Gaussian convolved with a segmented image.

2.3 Deep CNN features

Recently, in the domain of computer vision, machine, and pattern recognition, deep learning have shows improved performance for image classification on large datasets [20]. The deep learning designs such as deep CNN and recurrent NN have been employed to human action recognition, speech recognition, document classification, agricultural plants, medical imaging, and many other areas and shows superior performance. In object classification, CNN shows much attention due to their ability to automatically determine appropriate contextual features in image categorization problems. A simple CNN model consists of four types of layers. Initially, an input image is passed and computes its neurons by convolution layer, which are connected to local regions of the input. Each neuron is computed by dot product between their small regions and weights, which are connected to in the input volume. Thereafter, activation is performed using ReLu layer. The ReLu layer never changes the size of an input image. Then, pooling layer is performed to reduce the noise effects in the extracted features. Finally, high-level features are calculated by a fully connected (FC) layer.

In this article, we employ two pre-trained deep CNN models such as VGG19 and AlexNet, which are used for features extraction. These models incorporate convolution layer, pooling layer, normalization layer, ReLu layer, and FC layer. As discussed above the convolution layer extracts local features from an image, which is formulated by Eq. 22:

$$g_i^{(L)} = b_i^{(L)} + \sum_{j=1}^{m_1^{(L-1)}} \psi_{i,j}^{(L)} \times h_j^{(L-1)} \tag{22}$$

where $g_i^{(L)}$ denotes the output layer L, $b_i^{(L)}$ is base value, $\psi_{i,j}^{(L)}$ denotes the filter connecting the jth feature map, and h_j denotes the $L-1$ output layer. Then, pooling layer is defined which extract maximum responses from the lower convolutional layer with an objective of reducing irrelevant features. The max pooling also resolves the problem of overfitting and mostly 2 × 2 polling is performed on the extracted matrix. Mathematically, max pooling is described through Eqs. 23–25:

$$m_1^{(L)} = m_1^{(L-1)} \tag{23}$$

$$m_2^{(L)} = \frac{m_2^{(L-1)} - F(L)}{S^L} + 1 \tag{24}$$

$$m_3^{(L)} = \frac{m_3^{(L-1)} - F(L)}{S^L} + 1 \tag{25}$$

where S^L denotes the stride,$m_1^{(L)}$, $m_2^{(L)}$, and $m_3^{(L)}$ are defined filters for feature map such as 2 × 2, 3 × 3. The other layers such as ReLu and fully connected (FC) are defined as:

$$\mathrm{Re}_i^{(l)} = \max\left(h, h_i^{(l-1)}\right) \tag{26}$$

$$Fc_i^{(l)} = f\left(z_i^{(l)}\right) \ with \ z_i^{(l)} = \sum_{j=1}^{m_1^{(l-1)}} \sum_{r=1}^{m_2^{(l-1)}} \sum_{s=1}^{m_3^{(l-1)}} w_{i,j,r,s}^{(l)} \left(Fc_i^{(l-1)}\right)_{r,s} \tag{27}$$

where $Re_i^{(l)}$ denotes the ReLu layer, $Fc_i^{(l)}$ denotes the FC layer. The FC layer follows the convolution and pooling layers. The FC layer is similar to convolution layer and most of the researchers perform activation on the FC layer for deep feature extraction.

2.4 Pre-trained deep CNN networks

In this research, we use two pre-trained deep CNN models such as VGG and AlexNet for deep features extraction. AlexNet deep CNN model is designed by Krizhevsky et al. [27] using ImageNet dataset. This network contains five convolution layers, three pooling layers, and 3 FC layers along with softmax classification function. This network trained on input image size 227x227x3.

VGG-19 CNN network is proposed by Zisserman et al. [20] which contains 16 convolution layers, 19 learnable weights layers, 3 FC layers along with softmax function. This network is trained on ImageNet dataset and shows exceptional performance. This network also uses dropout regularization in the FC layer and apply ReLu activation function on all the convolution layers. The size of the training input images is selected as 224x224x3.

2.5 Features extraction and fusion

In this section, we present our proposed feature extraction and fusion strategy. The features are extracted from pre-trained deep CNN models using the different number of layers. In this work, two pre-trained models are used such as VGG19 and AlexNet for features extraction. The major aim of deep CNN features extraction from two models is to improve the classification accuracy. Because each model has distinct characteristics and gives different features. Therefore, by using this advantage, we extract features by performing activation on the FC7 layer and applying max pooling to remove the noise factors. Thereafter, an entropy-controlled method is implemented for best feature reduction. The proposed feature extraction and reduction architecture are shown in Fig. 3. As shown in Fig. 3, three types of features are extracted such as AlexNet deep CNN, VGG19 CNN, and SIFT. For AlexNet and VGG19, convolution layer is employed as an input layer. Then activation is performed on FC7 layer for both networks to extract deep CNN features. The size of deep CNN features for output layer FC7 is 1×4096 for both networks. The feature size of both output layer is higher. Therefore, we perform max-pooling of filter size of 2×2, which removes the noise effects and selects the maximum value feature of the given filter.

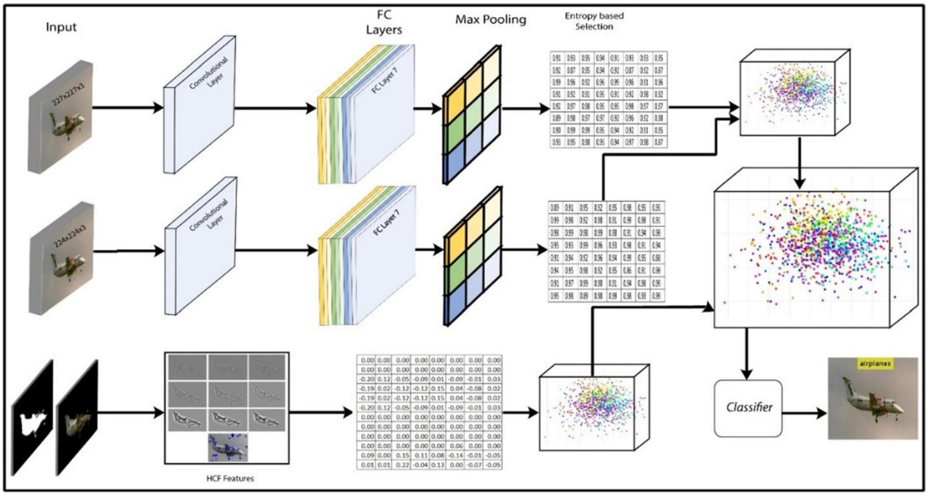


Fig. 3 Proposed deep CNN and SIFT features fusion and reduction method for object classification

After max-pooling, the new feature vectors of size 1 × 2048 are obtained, which are further improved by entropy-controlled feature reduction method. As extracted feature vectors can produce better results, but they increase the execution time. Therefore, our focus is to improve the classification accuracy and decrease the execution time. This problem is resolved by an entropy controlled method. The entropy gives the knowledge about randomness in a signal by showing the system disorder [50]. Due to its capacity to describe system behavior, entropy gives the valuable information which can be employed in features design [7]. Amongst several, we use the Renyi entropy method for feature reduction. In the circumstances of fractal dimension estimation, the Reyni entropy method determines the basis of the theory of generalized dimensions. The fractal dimension estimates the change patterns of the given feature space. The Reyni entropy is defined as follows:

Let $f_1, f_2, \ldots, f_n$ denote the A feature space after max-pooling, $g_1, g_2, g_3, \ldots g_n$ denote the B feature space aftermax-pooling, and $\xi_1, \xi_2, \ldots, \xi_n$ denote the ξ feature space, where $A \in$ AlexNet DCNN features, $B \in$ VGG19 DCNN features, and $\xi \in$ SIFT point feature vector. The dimension of each features space is 1 × 2048, 1 × 2048, and 1 × 128.The entropy is formulated by Eq. 28:

$$E_\alpha(X) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p_i^a\right) \tag{28}$$

where $\alpha \geq 0$ & < 1, $X \in (f_n, g_n, \xi_n)$, and p_i denote the probability value of extracted feature space A, B, and ξ which is defined by $p_i = \Pr(X = i)$ and represents the length of all feature spaces. The entropy function gives a new $N \times M$ feature vector, which controls the randomness of each feature space. Then, each $N \times M$ feature vector is sorted into ascending order and the top 1000 features are selected from A and B vectors and 100 features from the ξ vector. Mathematically, this process is described by Eq. 29:

$$E(A) = \Phi(f_n, \varrho), E(B\ ) = \Phi(g_n, \varrho), E(\xi) = \Phi(\xi_n, \varrho) \tag{29}$$

where $E(A)$ denotes the entropy information of feature space A, $E(B)$ denotes the entropy information of feature space B, $E(\xi)$ denotes the entropy information of feature space ξ, Φ denotes sorting function, and ϱ denotes the ascending order operation. Thereafter, both $E(A)$

and $E(B)$ entropy information features are fused in one matrix by the simple serial based method, which returns a feature vector of size 1 × 2000, which is further fused with SIFT point feature by the serial-based method as shown in the above Fig. 3 and below expression given in Eqs. 30–31:

$$\prod(Fused) = (N \times 1000) + (N \times 1000) + (N \times 100) \tag{30}$$

$$\prod(Fused) = N \times f_i \tag{31}$$

The size of the final feature vector is 1 × 2100, which is fed to ensemble classifier for classification. The ensemble classifier is a supervised learning method, which needs to training data for prediction. Ensemble method combines several classifiers data to produce a better system. The formulation of the ensemble method is given below.

Let we have extracted features and their corresponding labels $((f_1, y_1), (f_2, y_2), \ldots, (f_n, y_n))$, where f_i denotes the extracted features which are typically vectors of form $(f_{i+1}, f_{i+2}, \ldots, f_{i+n})$, then the unknown function is defined as $y = f(\boldsymbol{x})$. An ensemble classifier is a set of classifiers whose individual decisions are combined in on classifier by typical weights and voting. Hence the ensemble classifier is formulated as:

$$\hat{Y} = Sign\left(\sum_{k=1}^{K} \hat{w}_k\, h_k(x)\right) \tag{32}$$

where $h_k(\boldsymbol{x}) = h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_k(\boldsymbol{x})$ and $\hat{w}_k = \hat{w}_1, \hat{w}_2, \ldots \hat{w}_k$. The proposed method is tested on three datasets such as Caltech101, PASCAL 3D+ dataset, and 3D dataset. The sample labeled results are shown in the Figs. 4 and 5.

3 Experimental results

The proposed method is endorsed on three available datasets such as Caltech 101, PASCAL 3D+, and Barkley3D dataset. The Caltech-101 [14] dataset consists of total 102 distinct object classes of 9144 images. Each class consists of approximately 31~800 images. However, this dataset consists of both RGB and gray images, which is a major issue of this dataset. It is because if objects are recognized by their color, then color features are not performed well on

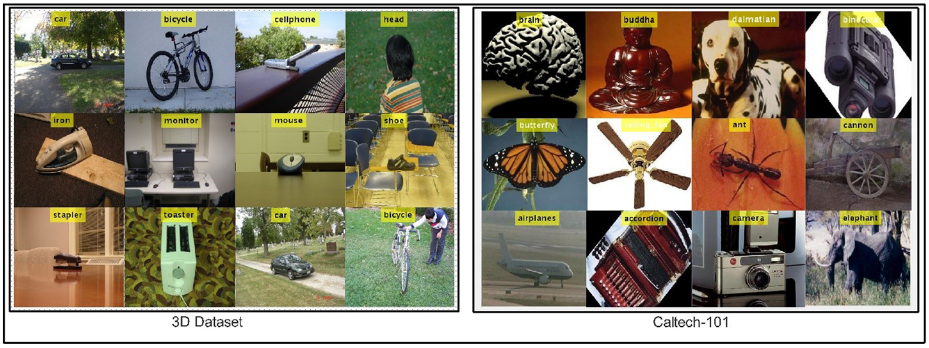


Fig. 4 Proposed labeled classification results for the 3D dataset and Caltech101 dataset

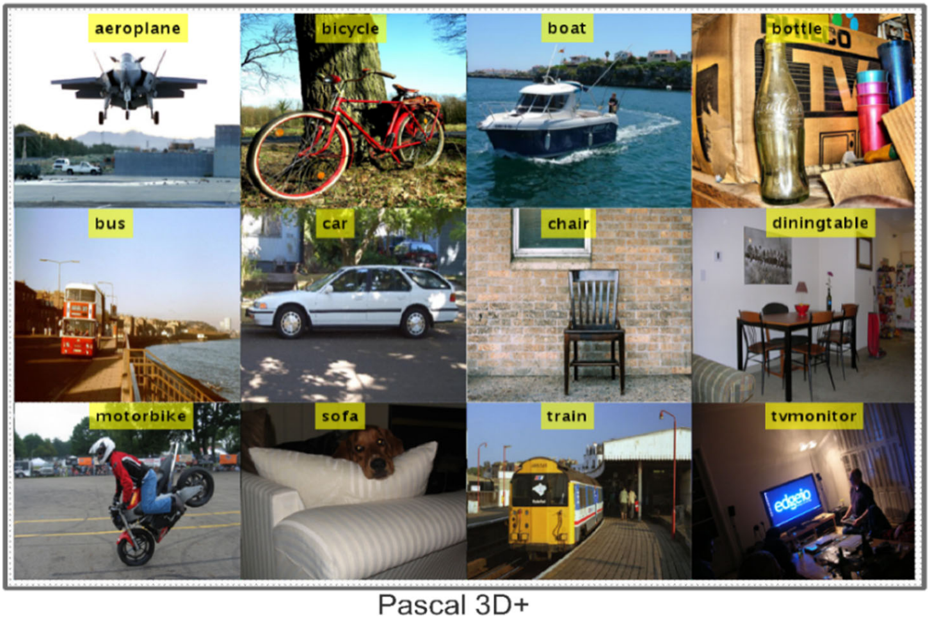


Fig. 5 Proposed labeled classification results for PASCAL 3D+ dataset

grayscale images. Pascal 3D+ dataset [11] is another challenging database which is used for object classification. This dataset is the combination of Pascal VOC 2012 and ImageNet. It contains a total of 22,394 images of 12 unique classes. The classes which are common between PASCAL VOC 2012 and ImageNet are merged into a new database, called Pascal 3D+. Barkley3D object dataset [21] consists of total 6604 images of 10 object classes including bicycle, car, cellphone, head, iron, monitor, mouse, shoe, stapler, and toaster. The number of images in each class range of 474–721. A brief description of each dataset is given in Table 1. For classification, we use Ensemble boosted tree (EBT) classifier and test its performance with Linear SVM (LSVM), Quadratic SVM (QSVM), Cubic SVM (CSVM), Fine KNN (FKNN), Cubic KNN (CKNN), decision tree (DT), and weighted KNN (WKNN). The performance of each classifier is calculated by three measures including accuracy, false negative rate (FNR), and execution time. All results are evaluated on 3.4 Gigahertz Corei7 7th generation desktop computer with a RAM of 16 Gigabytes and a GPU of NVIDIA GeForce 1070 (8GB, 256 bit) having MATLAB 2017b.

3.1 Caltech101 dataset classification results

In this section, we discuss detailed results of our method on selected datasets. For classification results on Caltech 101 dataset, we define three experiments on distinct classes such as 20, 34, 50, and 102. The experiments are a) classification of selected classes using AlexNet DCNN features with entropy-based selection method; b) Classification of selected classes using a VGG-19 DCNN model with entropy-based features selection; c) Fusion of deep CNN and SIFT features along with entropy-controlled selection method. The classification is performed on each class and finally compared the performance of all 102 classes in terms of accuracy and execution time with 20, 32, and 50 numbers of classes. For classification results, 50:50

Table 1 Description of selected datasets

Dataset	Classes	Total Images	Range
Caltech-101	101	9144	31–800
Pascal3D+	12	22,394	536–6704
3D Dataset	10	6604	474–721

approach is opted, and 10-fold cross-validation is employed. The 50:50 approach explains that 50 images from each class are used for training the classifier and remaining 50 for testing. The detailed results are explained in below sections.

3.1.1 AlexNet deep CNN with entropy-controlled selection

In the first step, we extract DCNN features of top 20 object classes by pre-trained AlexNet model and select the best features using an entropy-based method. The selected features are feed to classifiers and achieved the best classification accuracy of 86.5%, which is achieved on ensemble boosted tree (EBT). The classification accuracy of EBT classifier is given in Table 2. The testing time of EBT classifier is 105.00 s which is the best as compared to other classification methods. The second-best execution time is 114.25 s for quadratic SVM, which achieves classification accuracy 83.70% and FN rate is 16.30%. In the second step, classification is performed on 34 classes and obtained maximum classification accuracy of 84.6% on EBT classifier with an FN rate of 15.4%. Also, the classification is performed on some other classification methods and the second highest accuracy of 78.2% is achieved for cubic SVM as given in Table 2. The best execution time on the classification of 34 classes is 172.63 s which shows that the execution time is increased with the addition of more number of classes. In the third step, classification is performed on 50 number of classes and obtained maximum classification accuracy of 83.5% for EBT classifier, which decreases 1% as compared to 20 and 34 number of classes. This problem is caused, when an increase in a number of more object classes.

Moreover, the best execution time for 50 object classes is 193.00 s which is better than other classification methods as shown in Table 2 but it increases as compared to a classification of 20 and 34 classes. Finally, classification is performed on 100 classes and obtained maximum correct classification rate 71.7% on ensemble classifier. However, the FN rate is increased up to 28.3%, which is higher than 20, 34, and 50 classes object classification. Moreover, the execution time of ensemble classifier on 100 classes is 620.42 s, which is better as compared to other classification methods as given in Table 2 but the overall execution time for 20 object classes is better, which shows that the increase in the number of classes' effects on both classification accuracy and execution time.

3.1.2 VGG-19 deep CNN with entropy-controlled selection

In this experiment, the classification is performed on 20, 34, 50, and 100 object classes using VGG-19 deep CNN along entropy-controlled best features selection approach. In the first step, 20 object classes are randomly selected and classification is performed. For 20 object classes, the best classification accuracy of 92.0% with FN rate of 8.0% on EBT classifier. The classification results of EBT classifier are also compared with other supervised learning methods and obtained the second-best accuracy of 91.1% with FN rate is 8.9% on CSVM as presented in Table 3. The execution time of ensemble classifier is also calculated and

Table 2 Classification accuracy for Caltech101 dataset using AlexNet deep CNN along entropy-controlled features selection

Method	No of classes				Performance measures		
	20	34	50	100	Accuracy (%)	FNR (%)	Time (seconds)
Ensemble boosted tree	✓				**86.5**	**13.5**	**105.00**
		✓			**84.6**	**15.4**	**172.63**
			✓		**83.5**	**17.0**	**193.00**
				✓	**71.7**	**28.3**	**620.42**
Linear SVM	✓				82.0	18.0	197.71
		✓			75.6	24.4	266.74
			✓		78.6	21.4	859.90
				✓	67.9	32.1	6270.00
QSVM	✓				83.7	16.3	114.25
		✓			76.3	23.7	332.70
			✓		81.7	28.3	1325.00
				✓	70.3	29.7	20,355
CSVM	✓				83.5	16.5	118.77
		✓			78.2	21.8	339.92
			✓		81.7	18.3	1879.00
				✓	70.4	29.6	12,105.00
FKNN	✓				79.8	20.2	144.54
		✓			68.7	31.3	327.07
			✓		77.2	32.8	270.76
				✓	65.2	34.8	714.21
CKNN	✓				82	18	138.16
		✓			70.8	29.2	251.37
			✓		78.3	21.7	379.01
				✓	65.7	34.3	2038.00
Decision tree	✓				78.5	21.5	136.35
		✓			67.3	32.7	267.00
			✓		74.3	25.7	434.82
				✓	58.9	41.1	949.13
WKNN	✓				79.8	20.2	232.06
		✓			69.9	30.1	264.89
			✓		76.4	23.6	378.13
				✓	65.3	34.7	845.73

The bold values shows the best results

obtained the best testing time of 88.129 s, which is efficiently well as compared to other classification methods as LSVM, QSVM, and few more in Table 3. In the second step, 34 object classes are selected randomly and performed classification. The best classification accuracy is achieved as 84.6% with FN rate 15.4% on EBT classifier as presented in Table 3. The classification performance of EBT classifier is compared with seven other supervised learning methods and achieved the second-best accuracy of 78.2% on CSVM.

Moreover, the best execution time for classification is achieved for 34 object classes is 172.63 s on EBT, which is significantly good as compared to other methods. However, the worst execution time for classification of 34 object classes is 327 s on Fine KNN. Thereafter, 50 object classes are selected randomly for classification. The increase in the number of classes effects on the classification accuracy and execution time. However, using VGG deep CNN features with entropy-controlled method achieve the best classification accuracy of 86.0%, which is increased up to 2.55% as compared to AlexNet deep features. Moreover, the performance on VGG deep features is also improved and achieved the best computation time of 168.66 s which is better as compared to AlexNet deep features and other supervised learning methods as given in Table 3. Finally, classification is performed on all 102 object classes and achieved the best classification accuracy of 73.8% on EBT, which is executed in 454.270 s. The classification accuracy of EBT is increased up to 1.8% as compared to performance on AlexNetmodel but the execution time of EBT classifier is lower than the WKNN, which is 341.83 s as presented in Table 3. The above discussion, it is clear that entropy-controlled selection method performs well along with VGG-19 deep CNN features. Moreover, the execution time for object recognition on VGG features is improved as compared to AlexNet features on 20, 50, and 101 object classes.

3.1.3 VGG-19 deep CNN and AlexNet CNN features fusion and selection

Features fusion is an important step in the domain of machine learning because each feature extraction technique has unique characteristics. Therefore, in this study, we use two pre-trained

Table 3 Classification accuracy for Caltech101 dataset using VGG deep CNN features with the entropy-controlled method

Method	No of classes				Performance measures		
	20	34	50	100	Accuracy (%)	FNR (%)	Time (seconds)
Ensemble boosted tree	✓				**92.0**	**8.0**	**88.129**
		✓			87.5	**12.5**	**198.480**
			✓		**86.0**	**14.0**	**168.660**
				✓	**73.8**	27.2	454.270
Linear SVM	✓				86.7	13.3	180.709
		✓			81.0	19.0	256.700
			✓		78.9	21.1	944.080
				✓	54.8	45.2	1122.200
QSVM	✓				90.6	9.4	147.250
		✓			83.0	17.0	299.100
			✓		80.5	19.5	1205.900
				✓	54.6	45.4	820.010
CSVM	✓				91.1	8.9	191.395
		✓			82.9	17.1	311.200
			✓		81.0	19.0	1624.00
				✓	52.5	47.5	795.360
FKNN	✓				84.3	15.7	114.719
		✓			78.0	22.0	23.620
			✓		75.3	24.7	127.55
				✓	59.4	40.6	1119.490
CKNN	✓				82.6	17.4	315.130
		✓			77.5	22.5	99.800
			✓		77.6	22.4	160.780
				✓	59.2	40.8	81.680
Decision tree	✓				85.5	14.5	351.162
		✓			78.9	21.1	65.790
			✓		77.7	22.3	106.90
				✓	73.3	26.7	431.22
WKNN	✓				83.3	16.7	215.49
		✓			78.7	21.3	72.39
			✓		75.8	24.2	201.6
				✓	65.6	24.4	**341.83**

The bold values shows the best results

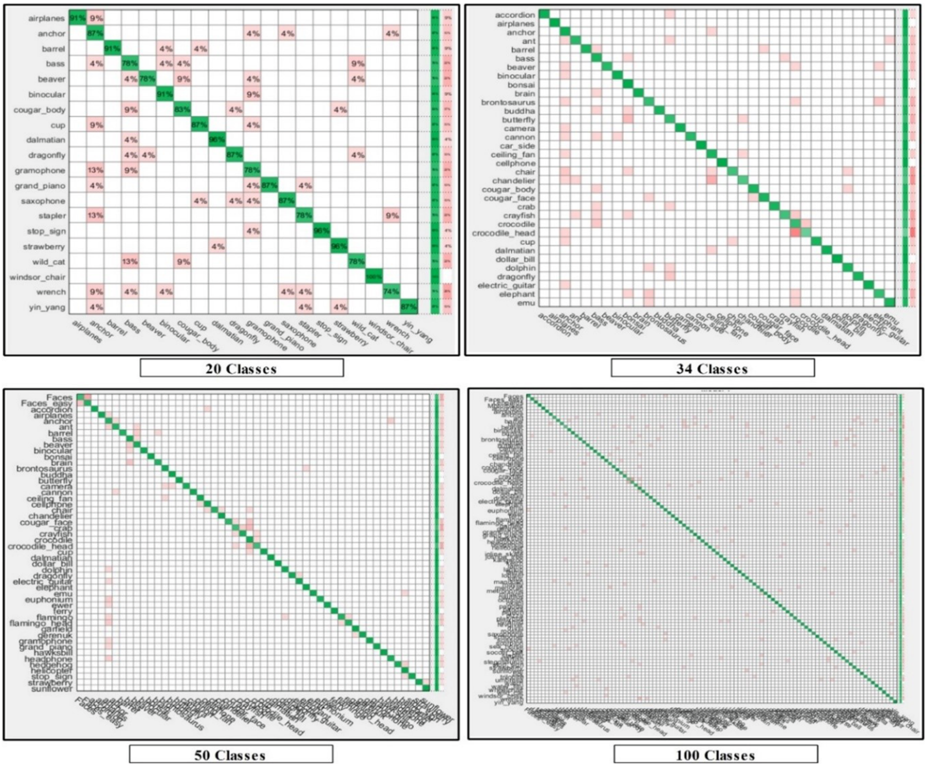
deep CNN models for features extraction and select the best features from each model by an entropy-controlled method. Thereafter, we extract SIFT features from RGB silhouette image and fused along with deep CNN selected features by the parallel approach. Finally, the fused features are feed to classifiers for recognition accuracy. The best-achieved classification accuracy for 20, 34, 50, and 100 classes is 86.5%, 93.8%, 93.5%, and 89.7% on EBT classifier, as presented in the Table 4. The classification accuracy of EBT classifier on 34, 50, and 100 classes is significantly improved as compared to individual AlexNet and VGG-19 deep CNN features with an entropy-based selection approach. However, we notice the execution time of the proposed method on EBT classifier is increased as compared to Tables 2 and 3. The proposed classification performance is proved by their confusion matrices are given in Fig. 6.

Finally, we compare our proposed results with existing methods in Table 5. Jun et al. [38] propose a deep stack network (DSN) for object classification and achieved a classification accuracy of 89%. In. [57] sparse structure PCA method is presented for object classification, which is based on SIFT features and SVM classifier. The presented method reports

Table 4 Classification accuracy for Caltech101 dataset using a fusion of CNN features and selection with the entropy-controlled method

Method	No of classes				Performance measures		
	20	34	50	100	Accuracy (%)	FNR (%)	Time (seconds)
Ensemble boosted tree	✓				86.5	**13.5**	**75.70**
		✓			93.8	**6.2**	289.90
			✓		93.5	**6.5**	**178.20**
				✓	**89.7**	**10.3**	**302.50**
Linear SVM	✓				82.0	18.0	97.71
		✓			87.2	12.8	495.22
			✓		87.8	12.2	1457.60
				✓	60.7	39.3	5613.70
QSVM	✓				83.7	16.3	114.25
		✓			88.2	11.8	654.53
			✓		88.9	11.1	1655.10
				✓	60.0	40	9846.00
CSVM	✓				83.5	16.5	118.77
		✓			88.8	11.2	1010.60
			✓		88.9	11.1	1754.10
				✓	55.0	45	10,322.00
FKNN	✓				79.8	20.2	14.540
		✓			82.2	17.8	71.486
			✓		82.6	17.4	71.950
				✓	71.3	28.7	66.749
CKNN	✓				82.0	18	38.160
		✓			82.9	17.1	110.560
			✓		81.0	19	327.660
				✓	70.7	29.3	595.820
Decision tree	✓				78.5	21.5	36.350
		✓			81.3	18.7	99.080
			✓		85.1	14.9	106.200
				✓	88.2	11.8	470.750
WKNN	✓				79.8	20.2	32.060
		✓			86.3	13.7	18.110
			✓		80.8	19.2	68.810
				✓	65.7	34.3	321.97

The bold values shows the best results



20 Classes

34 Classes

50 Classes

100 Classes

Fig. 6 Confusion matrix for 20, 34, 50, and 100 object classes using Caltech101 dataset

classification accuracy 83.9% on a Caltech101 dataset. Qing et al. [38] used a combination of YCbCr transformation and Extreme Learning (EL) for object classification and obtained an accuracy of 78% on the Caltech101 dataset. Yongsheng et al. [45] use the K-means based reduction on the SIFT descriptors and achieved 85.78% accuracy. However, in this research, our proposed method shows improved performance in both accuracy and execution time. The proposed method achieves classification accuracy 86.5%, 93.8%, 93.5%, and 89.7% for 20, 34, 50, and 100 object classes on Caltech101 dataset. The execution time of the proposed method is also plotted in Fig. 7.

3.2 Pascal3D + v1.1 dataset results

In this section, we present the proposed algorithm results on PASCAL 3D dataset. The results are calculated in four different steps: a) AlexNet deep CNN features extraction along with entropy-controlled feature selection, b) VGG features extraction and entropy-controlled selection, c) fusion of VGG and AlexNet deep CNN features along with selection method, and d) fusion of SIFT and deep CNN features along with entropy-controlled method. Three parameters (i.e., accuracy, FNR, and time) are used to measure the performance of each classifier. As discussed above, this dataset consists of total 22,394 images of 12 unique object classes. For validation of the proposed method on this dataset, we opt an approach of 50:50 for training and testing. This approach is followed for each step. The achieved best classification accuracy for AlexNet deep CNN features along with entropy-controlled selection method is 76.8% on ensemble classifier. The FN rate on

Table 5 Comparison with existing methods

Paper	Year	Features	Technique	Accuracy (%)	Time (s)
Jun et al. [38]	2018	2000 after PCA on 4096	Deep Stack Network	89	
Jinjoo et al. [57]	2018	SIFT- > SVM	SSPCA	83.9	
Qing et al. [38]	2018	YCbCr-SIFT	YCbCr-SIFT+LSC + ELM	78	
Yongsheng et al. [45]	2018	SIFT	Reduction using K Means	85.78	
Xiaozhao et al. [13]	2018	Salient features using unsupervised	PCA- > 1500	76	
Ridha et al. [9]	2017	Deep CNN	Fast wavelet	75.6	
Our	2018	Deep CNN and SIFT Features	Fusion of Deep CNN and SIFT Features along with entropy-controlled selection method	**20 Classes: 86.5** **34 Classes: 93.8** **50 Classes: 93.5** **100 Classes: 89.7**	**75.70** 289.90 **178.20** **302.50**

The bold values shows the best results

ensemble classifier is 23.2% and testing execution time is 154.5 s. The recognition results of an ensemble classifier are also compared with other state-of-the-art classification methods as presented in Table 6. In the second step, the classification is performed by using VGG-19 deep CNN features and achieved maximum classification accuracy 81.8%, which is improved as compared to AlexNet features. But the execution time on VGG-19 deep features along with the selection method is increased on ensemble classifier and best-achieved execution time is 240.86 s on decision tree as given in Table 6. In the third step, selected AlexNet DCNN and VGG DCNN features are fused by a serial-based method and perform classification. The best-achieved classification accuracy is 87.4% on ensemble classifier, which is significantly improved after fusion of DCNN features. The execution time of ensemble classifier for step 3 is 230.2 s, which is higher than the FKNN as presented in Table 6.

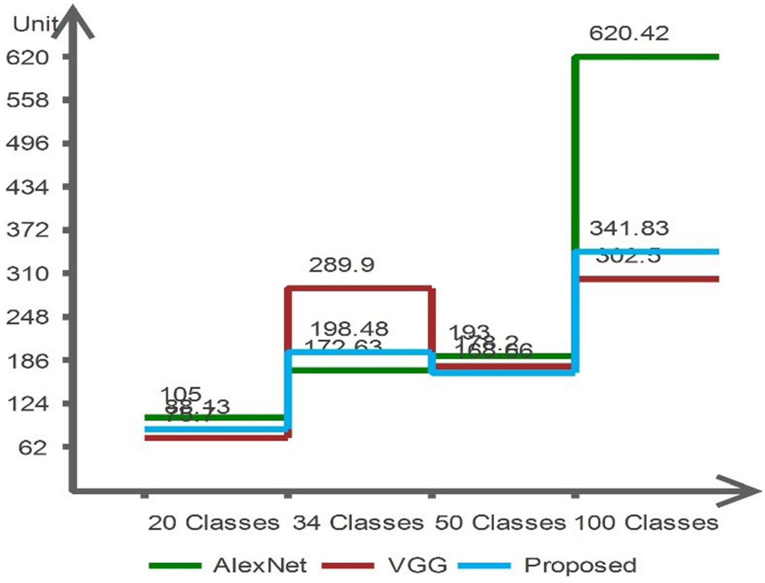


Fig. 7 Comparison of the execution time of all defined experiments on the Caltech101 dataset

Table 6 Classification accuracy on PASCAL 3D dataset

Method	No of classes				Performance measures		
	AlexNet-Entropy	VGG-19-Entropy	Fused DCNN-Entropy	Fused SIFT+DCNN+Entropy	Accuracy (%)	FNR (%)	Time (seconds)
Ensemble boosted tree	✓				**76.8**	**23.2**	**154.5**
		✓			**81.8**	**18.2**	304.8
			✓		87.4	**12.6**	230.2
				✓	**88.6**	**11.4**	**111.99**
Linear SVM	✓				71.0	29.0	437.18
		✓			78.1	21.9	626.3
			✓		56.6	43.4	834.9
				✓	82.8	17.2	175.26
QSVM	✓				75.6	24.4	641.2
		✓			80.6	19.4	698.2
			✓		86.9	13.1	1821.7
				✓	81.3	18.7	210.39
CSVM	✓				73.6	26.4	600.57
		✓			79.1	20.9	765.77
			✓		86.6	13.4	1211.2
				✓	81.4	18.6	217.5
FKNN	✓				64.0	36.0	195.068
		✓			70.8	29.2	222.43
			✓		75.0	25.0	**198.71**
				✓	23.4	76.6	134.934
CKNN	✓				71.5	28.5	2149.1
		✓			78.9	21.1	5277.1
			✓		82.6	17.4	4133.5
				✓	26.5	83.5	659.35
Decision tree	✓				70.6	29.4	164.36
		✓			78.0	22.0	240.86
			✓		82.6	17.4	398.78
				✓	73.85	26.15	185.585
WKNN	✓				64.7	35.3	1087.5
		✓			71	29	2457.3
			✓		74.8	25.2	2020.2
				✓	78.2	21.8	409.78

The bold values shows the best results

Finally, SIFT point and DCNN features are fused by the parallel approach and perform classification. For classification, the EBT method is used and obtained a maximum accuracy of 88.6% and FN rate is 11.4%, which is significantly improved as compared to step 1, 2, and 3. Moreover, the best execution time is 111.99 s for EBT as given in Table 6. From Table 6, the performance of EBT classifier is compared with several other supervised learning methods such as LSVM, WKNN, FKNN, and few more. These supervised learning methods also perform well by using proposed features fusion and selection method, which gives the authenticity of the proposed method. Moreover, the classification performance of ensemble classifier is validated by Table 7.

Finally, the proposed method results on PASCAL 3D dataset are compared with existing methods as presented in Table 8. In Table 8, Chi et al. [12] extract deep CNN features for object classification and reported classification accuracy of 81.8%. In [18] CNN based features are extracted for object classification and perform experiments on PASCAL 3D dataset and achieved accuracy 83.92%. However, our proposed method

Table 7 Confusion matrix for PASCAL 3D dataset using proposed features fusion and selection method

Class	Classification Class										
	Airplane	Bicycle	Boat	Bottl e	Bus	Car	Chai r	Dining table	Motor bike	Sofa	Train	Tv Monitor
Airplane	**93**		2	<1		1	1	<1	1	<1	2	<1
Bicycle	<1	**80**	1	1	1		1	1	12	1	2	1
Boat	1	<1	**93**	1		1	<1		1		3	
Bottle		1	2	**64**	<1		3	14	3	8	1	5
Bus	1				**91**	1	<1	<1	1		5	
Car	1	1	1	<1		**86**	1	1	4	<1	1	1
Chair			1	2		<1	**47**	19	2	15	2	11
Dining table		1		4			7	**65**	1	13	1	8
Motorbike	<1	7	<1	2	2	2	1	1	**81**	<1	2	<1
Sofa			<1	3			8	14	1	**62**	1	10
Train			1	1	2	1	<1	1	3	<1	**90**	1
Tv monitor	<1			2		1	6	7	<1	15	<1	**69**

shows improved performance on PASCAL 3D dataset and achieved a classification accuracy of 88.60%.

3.3 Barkley 3D dataset

The 3D dataset consists of total 6604 images of 10 object classes including bicycle, car, cellphone, head, iron, monitor, mouse, shoe, stapler, and toaster. The number of images in each class range of 474–721. For validation of the proposed method on a 3D dataset, 50:50 approach is opted for training and testing the classifier. To analyze the performance of a proposed method, we employ four distinct experiments. In the first experiment, Alexnet deep CNN features are extracted and select best features using the entropy method. The best-achieved classification accuracy for the first experiment is 97.90% on EBT classifier with FN rate is 2.1%. The execution time for experiment 1 is 978.00 s on EBT, which is higher than the other classifiers whereas the best execution time for experiment 1 is 245.68 s as given in Table 9. In the second experiment, the VGG deep CNN features are extracted and select the best features by an entropy-controlled method. 10-fold cross validation is performed for testing the classification performance and achieved the best accuracy 97.5% with FN rate is 2.5%. The execution time of ensemble classifier for VGG features is 900.5 s as given in Table 9, which shows that FKNN performs fast and execute in 113.5 s. In the third experiment, to improve the

Table 8 CA Comparison with state of the art techniques on the Pascal3D+ dataset

Author	Year	Features	Technique	Accuracy (%)	Time (s)
Chi Li [12]	2018	CNN	Deep supervision Object reconstruction	81.8	
Alexander [18]		CNN	CNN based multi-view learning	83.92	
Our	2018	SIFT and deep features	Fusion of point and DCNN features along with selection method	**88.6**	**111.9**

The bold values shows the best results

Table 9 Proposed classification results on the 3D dataset

Method	No of classes				Performance measures		
	AlexNet-Entropy	VGG-19-Entropy	Fused DCNN-Entropy	Fused SIFT+DCNN+ Entropy	Accuracy (%)	FNR (%)	Time (seconds)
Ensemble boosted tree	✓				**97.9**	**2.1**	978.00
		✓			**97.5**	**2.5**	900.5
			✓		98.8	1.2	5342.2
				✓	**99.7**	**0.3**	177.49
Linear SVM	✓				96.1	3.9	220.29
		✓			96.2	3.8	132.01
			✓		98.7	1.3	453.66
				✓	99.5	0.5	**94.242**
QSVM	✓				97.5	2.5	231.44
		✓			97.3	2.7	338.24
			✓		**99**	**1.0**	566.5
				✓	99.7	0.3	120.52
CSVM	✓				97.5	2.5	245.68
		✓			*97.3*	2.7	363.6
			✓		99	1	651.6
				✓	99.7	0.3	124.92
FKNN	✓				96.9	3.1	55.29
		✓			97.3	2.7	113.15
			✓		97.9	2.1	141
				✓	62.6	37.4	19.102
CKNN	✓				95.4	4.6	1133.6
		✓			94.8	5.2	1167.8
			✓		97	3	2248.7
				✓	34	66	359.51
Decision tree	✓				92.3	7.7	101.08
		✓			92.8	7.2	128.71
			✓		94.5	5.5	151.44
				✓	85.8	14.2	153.816
WKNN	✓				97	3	596
		✓			97.2	2.8	1862.9
			✓		98	2	1163.1
				✓	98.6	1.4	215.24

The bold values shows the best results

classification accuracy and execution time, we fuse both VGG and AlexNet deep CNN features and achieve classification accuracy 98.8% on ensemble classifier. The fused matrix improves the classification accuracy as compared to an individual selected deep CNN features as presented in Table 9. The execution time of fused approach is increased up to 5342 s on ensemble classifier. To resolve this issue, in experiment 4 we fuse SIFT features along with deep CNN features and achieved classification accuracy 99.7% with FN rate is 0.3%. The execution time of the proposed method is also reduced on ensemble classifier and achieved testing time is 177.49 s. Moreover, the classification accuracy of ensemble classifier for experiment 4 is confirmed by a confusion matrix in Table 10.

3.4 Graphical discussion

In this section, the classification results are presented in the graphical format. In Fig. 8a, the overall comparison of three experiments is plotted for the Caltech-101 dataset.

Table 10 Confusion matrix of proposed method results on the 3D dataset

Class	Classification Class									
	Bicycle	Car	Cellphone	Head	Iron	Monitor	Mouse	Shoe	Stapler	Toaster
Bicycle	100									
Car		100								
Cellphone			99			<1	<1			<1
Head				100						
Iron					100					
Monitor						100				
Mouse							>99		<1	
Shoe								>99	<1	
Stapler							<1		>99	
Toaster							<1			>99

Figure 8a explains that the minimum, average, and maximum accuracy which is achieved through all conducted experiments in Section 3.1. The minimum achieved accuracy for all classification methods is 52.5% on CSVM, average accuracy is above 60%, and maximum accuracy of 89.7% for EBT classifier. The EBT classifier achieves the minimum accuracy of 71.7% through AlexNet along entropy-controlled selection approach, whereas the maximum accuracy of 89.7% through proposed fusion and selection approach. Similar in Fig. 8b classification results for all datasets are computed on four different steps as explains in Section 3.2. The results are presented in the form of minimum, average, and

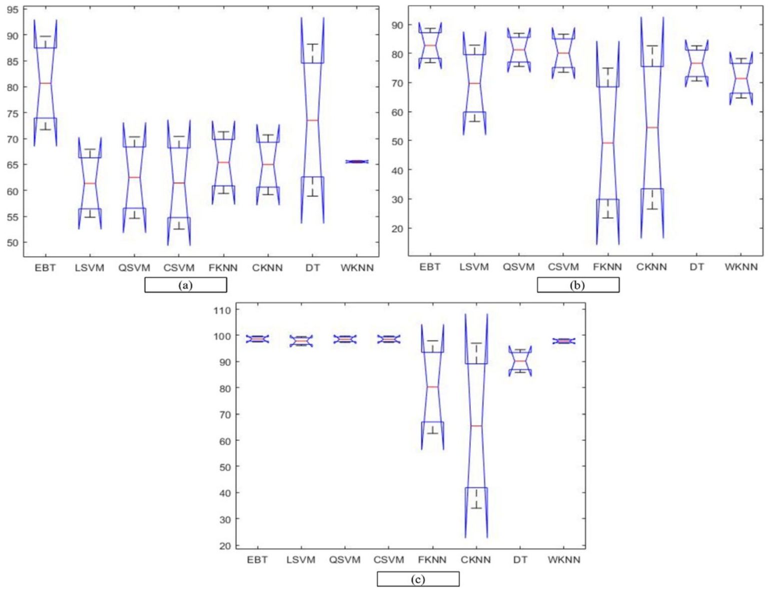


Fig. 8 Overall range of classification accuracy of all datasets. **a** Caltech 101 dataset, **b** PASCAL 3D Plus dataset, and **c** Barkley 3D dataset

maximum. Finally, classification results are calculated using Barkley 3D Dataset through four steps. All steps are explained in Section 3.3. The EBT classifier outperforms on all datasets for proposed fusion and selection method.

4 Conclusion

A DCNN and SIFT point features fusion, and selection-based approach is proposed in this article. The proposed method works in two parallel steps. In the first step, improved saliency method is implemented, and SIFT point features are extracted from RGB mapped image. Then, in the second step DCNN features are extracted using pre-trained CNN models. The max-pooling is performed on extracted features matrices to remove the noisy information. Thereafter, a Reyni entropy-controlled method is proposed which control the randomness of extracted features and select the best features. The selected features are finally fed to ensemble classifier for object classification. The proposed method automatically detects and labeled object from a large number of sample images with minimum human intervention. The proposed approach performs classification under the supervised method and achieves the maximum classification accuracy 93.8%, 88.6%, and 99% on Caltech101, PASCAL 3D Plus, and Barkley 3D dataset, which shows exceptional performance as compared to existing methods. Moreover, the proposed method efficiently reduces the computation time, which shows the importance of selection methods. In the future, we implement a new generic method for multiple object detection and classification using deep learning. Moreover, we apply method on real-time object classification.

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

References

1. Achanta R et al (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34(11):2274–2282
2. Agrawal S et al (2018) A comparative study of fuzzy PSO and fuzzy SVD-based RBF neural network for multi-label classification. Neural Comput & Applic 29(1):245–256
3. Akcay S, et al (2018) Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery. IEEE Trans Inf Forensics Secur
4. Akram T, et al (2018) Skin lesion segmentation and recognition using multichannel saliency estimation and M-SVM on selected serially fused features. J Ambient Intell Humaniz Comput:1–20
5. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE Comput Intell Mag 5(4):13–18
6. Chen S et al (2018) Local patch vectors encoded by fisher vectors for image classification. Information 9(2):38
7. Cheng G et al (2016) Study on planetary gear fault diagnosis based on entropy feature fusion of ensemble empirical mode decomposition. Measurement 91:140–154
8. Dong H, et al (2018) A novel hybrid genetic algorithm with granular information for feature selection and optimization. Appl Soft Comput
9. Ejbali R, Zaied M (2018) A dyadic multi-resolution deep convolutional neural wavelet network for image classification. Multimed Tools Appl 77(5):6149–6163
10. Esteva A et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115
11. Everingham, M., et al., The pascal visual object classes challenge: a retrospective. Int J Comput Vis, 2015. 111(1): p. 98–136

12. F.a.F. (2018) https://melanoma.canceraustralia.gov.au/statistics
13. Fang X, et al (2018) Approximate low-rank projection learning for feature extraction. IEEE Transactions on Neural Networks and Learning Systems
14. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611
15. Fondón I, et al (2018) Automatic classification of tissue malignancy for breast carcinoma diagnosis. J Comput Biol Med
16. Ghose U, Mehta R (2018) Attribute reduction method using the combination of entropy and fuzzy entropy. In: Networking Communication and Data Knowledge Engineering. Springer, p 169–177
17. Gomathi D, Seetharaman K Object classification techniques using tree based classifiers
18. Grabner A, Roth PM, Lepetit V (2018) 3d pose estimation and 3d model retrieval for objects in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
19. He K, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
20. Hu F et al (2015) Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens 7(11):14680–14707
21. Janoch A, et al (2013) A category-level 3d object dataset: putting the kinect to work. In: Consumer depth cameras for computer vision. Springer, p 141–165
22. Juuti M, Corona F, Karhunen J (2018) Stochastic discriminant analysis for linear supervised dimension reduction. Neurocomputing
23. Khan MA et al (2017) License number plate recognition system using entropy-based features selection approach with SVM. IET Image Process 12(2):200–209
24. Khan MA, et al (2018) An implementation of optimized framework for action classification using multilayers neural network on selected fused features. Pattern Anal Applic:1–21
25. Khan MA et al (2018) CCDF: automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep CNN features. Comput Electron Agric 155:220–236
26. Khan MA et al (2018) An implementation of normal distribution based segmentation and entropy controlled features selection for skin lesion detection and classification. BMC Cancer 18(1):638
27. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems
28. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer vision and pattern recognition, 2006 IEEE computer society conference on. IEEE
29. Leng L, et al (2010) Dynamic weighted discrimination power analysis in DCT domain for face and palmprint recognition. In: Information and Communication Technology Convergence (ICTC), 2010 International Conference on. IEEE
30. Leng L, Zhang J, Khan MK, Chen X, Alghathbar K (2010) Dynamic weighted discrimination power analysis: a novel approach for face and palmprint recognition in DCT domain. International Journal of Physical Sciences 5(17):2543–2554
31. Leng L, et al (2011) Two dimensional PalmPhasor enhanced by multi-orientation score level fusion. In: FTRA International Conference on Secure and Trust Computing, Data Management, and Application. Springer
32. Leng L, et al (2011) Two-directional two-dimensional random projection and its variations for face and palmprint recognition. In: International Conference on Computational Science and Its Applications. Springer
33. Leng L, et al (2012) Two-dimensional cancelable biometric scheme. In: Wavelet Analysis and Pattern Recognition (ICWAPR), 2012 International Conference on. IEEE
34. Leng L, Li M, Teoh ABJ (2013) Conjugate 2dpalmhash code for secure palm-print-vein verification. In: Image and Signal Processing (CISP), 2013 6th International Congress on. IEEE
35. Leng L et al (2017) Dual-source discrimination power analysis for multi-instance contactless palmprint recognition. Multimed Tools Appl 76(1):333–354
36. Li B, et al (2017) Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In: Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on. IEEE
37. Li K et al (2018) Multi-modal feature fusion for geographic image annotation. Pattern Recogn 73:1–14
38. Li Q, et al (2018) Improving image classification accuracy with ELM and CSIFT. Comput Sci Eng
39. Liaqat A, et al (2018) Automated ulcer and bleeding classification from wce images using multiple features fusion and selection. Journal of Mechanics in Medicine and Biology:1850038
40. Liu L, Wang L, Liu X (2011) In defense of soft-assignment coding. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE

41. Liu W et al (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26
42. Liu W, Yang X, Tao D, Cheng J, Tang Y (2018) Multiview dimension reduction via hessian multiset canonical correlations. Information Fusion 41:119–128
43. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
44. Naeini AA et al (2018) Particle swarm optimization for object-based feature selection of VHSR satellite images. IEEE Geosci Remote Sens Lett 15(3):379–383
45. Pan Y, et al Locality constrained encoding of frequency and spatial information for image classification. Multimed Tools Appl:1–17
46. Qin C, Sun M, Chang C-C (2018) Perceptual hashing for color images based on hybrid extraction of structural features. Signal Process 142:194–205
47. Rastegari M, et al (2016) Xnor-net: imagenet classification using binary convolutional neural networks. In: European Conference on Computer Vision. Springer
48. Raza M et al (2018) Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning. Futur Gener Comput Syst 88:28–39
49. Roy PK, Om H (2018) Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos. In: Advances in soft computing and machine learning in image processing. Springer, p 277–294
50. Sankar AS, Nair SS, Dharan VS, Sankaran P (2015) Wavelet sub band entropy based feature extraction method for BCI. Procedia Computer Science 46:1476–1482
51. Sharif M, Khan MA, Akram T, Javed MY, Saba T, Rehman A (2017) A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. EURASIP Journal on Image and Video Processing 2017(1):89
52. Sharif M, et al. (2018) A framework for offline signature verification system: best features selection approach. Pattern Recogn Lett
53. Sharif M et al (2018) Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. Comput Electron Agric 150:220–234
54. Siddiqui S, Khan MA, Bashir K, Sharif M, Azam F, Javed MY (2018) Human action recognition: a construction of codebook by discriminative features selection approach. International Journal of Applied Pattern Recognition 5(3):206–228
55. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
56. Singh C, Walia E, Kaur KP (2018) Enhancing color image retrieval performance with feature fusion and non-linear support vector machine classifier. Optik 158:127–141
57. Song J, et al (2018) Structure preserving dimensionality reduction for visual object recognition. Multimed Tools Appl:1–17
58. Szegedy C, et al (2015) Going deeper with convolutions. Cvpr
59. Wei G et al (2018) Content-based image retrieval for lung nodule classification using texture features and learned distance metric. J Med Syst 42(1):13
60. Yu W, et al (2018) Hierarchical semantic image matching using CNN feature pyramid. Comput Vis Image Underst

Muhammad Rashid : Recently, he is student of Master in Science in Computer Science at COMSATS University Islamabad, Wah Campus, Pakistan. His research interest including Object classification, Medical Imaging, and Image retrieval.

Muhammad Attique Khan : Received his MCS and MS (CS) degree from COMSATS University ISLAMABAD, Wah Campus in 2016 and 2018. Currently, he is Lecturer of Department of Computer Science and Engineering at HITEC University, taxila, Pakistan. Him reserach areas including Video Surveillance, Biometric (Human gait recognition), Object Classifcation, Medical imaging, and Agriculture Plants.

Muhammad Sharif : Ph.D. in Computer Science from COMSATS University Islamabad, Wah campus, Pakistan. Associate Professor, Department of Computer Science, COMSATS Institute of Information Technology Quaid Avenue, Wah Cantt, Pakistan.

Mudassar Raza : Completing PHD from USTC China. His current research area including pattern recognition, computer vision, and machine learning. Moreover, he works recently in several famous area like human action recognition and medical Imaging.

Muhammad Masood : Ph. D in USTC China. His research areas including Medical imaging, pattern recognition, and engineering development.

Farhat Afza : Completing her MS from COMSATS University islamabad, Wah campus, Pakistan. She is currently Ph. D student at COMSATS Wah. Her research interest including pattern recognition, and computer vision.