JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

MODIFIED LOCAL DIRECTIONAL PATTERN AND TAMURA FEATURES BASED OBJECT DETECTION AND CLASSIFICATION IN VIDEOS USING DEEP NEURAL NETWORK

Saju A, **Dr. H. N. Suresh**,

Research Scholar (VTU), Bangalore Institute of Technology, V.V. Puram, Bangalore 560004
link2saju@gmail.com
Professor & PG Coordinator, Department of Electronics and Instrumentation Engineering,
Bangalore Institute of Technology, V.V. Puram, Bangalore 560004
hn.suresh@rediffmail.com

Received: 16 March 2020 Revised and Accepted: 18 June 2020

Abstract

In recent scenario, object classification is gaining more attention among the researchers, due to its extensive range of applications like surveillance, image analysis, etc. Recently, many existing methods are developed for object classification, but still it achieves only considerable performance in the conditions; congest situation, complex background, etc. To overcome these concerns, an appropriate feature extraction and classification techniques are proposed in this research article for automatic object classification. At first, the annotated objects are segmented from the video sequences by using Superpixel based Fast Fuzzy C-Means (SFFCM) algorithm. Then, the features from the segmented objects are extracted by applying tamura features and Modified Local Directional Pattern (MLDP). Finally, Deep Neural Network (DNN) classifier is applied to classify the object classes or categories. The Caltech 256 and PASCAL VOC 2007 databases are used to analyze the proposed model performance. The classification performance of the proposed model is evaluated by means of precision, specificity, recall, and accuracy. In the experimental part, the proposed model improvement maximum of 20.19% and 18.29% of precision in PASCAL VOC 2007, and Caltech 256 databases related to existing model; Image Level Hierarchical Structure (ILHS).

KEYWORDS: Deep neural network, modified local directional pattern, superpixel based fast fuzzy c-means, tamura features.

1. Introduction

In recent decades, the deployment of video cameras and the increasing availability results in hundreds of video streams, such videos are sub categorized into number of interested frames [1]. Several information are extracted from the videos such as object behaviour, motion, etc. Recently, the video analytic system performs object detection and classification [2]. The object detection states the detection of object instances, which belongs to the category (for instance; motorcycle, human, etc) [3]. Within a video frame, the objects resides at any location that needs the detection process for studying the dissimilar parts of a video frame to locate the interested objects [4-5]. Some of the existing methods utilized in object classification are Convolutional Neural Network (CNN) [6-8], Deep CNN [9], KNN [10], etc. The weak semantic segmentation annotations are developed in the previous research studies, due to the weak supervised semantic segmentation. Hence, the labels in the weak semantic segmentation annotations are too noisy and also not appropriate for training the deep learning networks. To highlight this concern, a new model is proposed for automatic object detection and classification.

In this research, a new hybrid feature descriptor is proposed with DNN classifier for enhancing the object detection and classification performance. At first, the data is acquired from two online databases; PASCAL VOC 2007, and Caltech 256. Then, the annotated objects are segmented from the video sequences using SFFCM algorithm. The undertaken clustering algorithm delivers better result in overlapped data and comparatively better than other existing algorithms. Then, two feature extraction techniques; MLDP and tamura feature (directionality, contrast, and line-likeness) are utilized for extracting the feature vectors. The major benefit of tamura feature (directionality, contrast, and line-likeness) is extremely robust to noise. Besides, MLDP is invariant to photometric and geometric transformations that helps in extracting the active features. The extracted feature vectors are classified by using DNN classifier, which is very robust in unstructured data

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

conditions. In the resulting part, the proposed model efficacy is analysed by means of precision, specificity, recall, and accuracy.

A few existing research papers on object detection and classification in video sequences is surveyed in the section 2. The mathematical explanation about the undertaken methodologies is indicated in the section 3. Section 4 details about the experimental simulation of the proposed model. The conclusion about the proposed model is given in the section 5.

2. Literature survey

T. Gong, *et al*, [11] implemented a multi-point K-Nearest Neighbor (KNN) classifier for object detection and classification. Initially, the developed approach utilizes the training images for differentiating the spatial correlation between the land cover classes. Then, scan the training images by data template to identify the event matches and record the central class. In addition, estimate the multiple point probabilities by counting the replicates and then incorporated into KNN classifier. In the experimental step, the performance of KNN classifier was related with some existing classifiers like Support Vector Machine (SVM), Bayesian approach and decision tree classifier. From the experimental investigation, the proposed approach showed significant performance in classifying land cover classes; shadow, buildings, road and vegetation. Related to deep learning classifiers, the KNN was a slow learner, so it requires considerable amount of data for classification.

P. Tang, *et al*, [12] developed a new image descriptor on the basis of Artificial Visual Cortex (AVC) model for identifying the objects location in the natural images. The developed descriptor utilizes sparse points instead of image region for selecting the objects. Additionally, a feedback operation was included in AVC model to build the descriptor, since it encompasses more information about the objects. In this literature study, the developed model performance was validated on GRAZ 01 and 02 datasets. From the simulation outcome, the developed model achieved superior performance in object detection and classification by means of accuracy. In large databases, the developed AVC model recognizes more wrong objects location, due to the absence of feature scaling.

C. Zhang, *et al*, [13] developed ILHS classifier for object detection by using semantic and visual similarities. At first, a new image representation was generated by exploring semantic and visual similarities. Then, hierarchically cluster the images for exploring the correlations. In every cluster, the diversity of image classes were utilized to reweight the visual similarities for generating a new image representation. In this literature, the developed model performance was verified on PASCAL VOC 2007, 2012 and Caltech 256 datasets. The simulation outcome shows the efficacy of the developed model in light of precision. In this research, the visually similar images were difficult to correlate, when the image have varied appearance.

M. Rashid, *et al*, [14] developed a new strategy for object detection and classification. Initially, Scale Invariant Feature Transform (SIFT) feature descriptor was used for extracting the feature vectors from collected images. Then, Deep Convolutional Neural Network (DCNN) classifier was applied to classify the objects. In DCNN, Renyi entropy controller approach was used for controlling the extracted feature vectors and to choose the best features. The developed strategy automatically labels the objects with limited human intervention. In this literature, the developed strategy performance was evaluated on Barkley 3D, Caltech101, and PASCAL 3D Plus datasets. From the experimental consequence, the developed strategy achieved significant performance in object classification by means of time, false negative rate, and accuracy. Some of the major drawbacks in DCNN; need lot of data for training and requires GPU based system for implementation, which was a complex task.

T. Mahalingam, and M. Subramoniam, [15] developed a new model for automatic object detection and classification. Initially, modified kernel fuzzy c means algorithm was developed with ant colony optimizer for foreground and background separation. The developed clustering method was effective in segmenting the non-static objects. In this work, the developed model performance was related with other optimizers in light of recall, precision, accuracy, etc. From the experimental consequence, the developed model attained significant performance in object detection on Hall monitor and PETS video sequences. During object detection, the developed model includes a few concerns such as stuck in the local minima and poor conjunction rate.

In order to highlight the above stated issues, a new hybrid feature descriptor is proposed with DNN classifier for enhancing the object detection and classification performance.

3. Proposed model

In computer vision, object detection and classification gained more attention among the researchers, because it plays a dominant role in image analysis and video surveillance. In this research study, the proposed model comprises of four phases for automatic object detection and classification such as **data collection**: Caltech 256 and PASCAL VOC 2007 databases, **object segmentation:** superpixel based fast fuzzy c means, **Extraction of feature vectors:** MLDP and tamura features, and **classification:** DNN classifier. The flow diagram of proposed model is specified in figure 1. The brief explanation about the undertaken approaches is given below.

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

3.1 Dataset description

In first phase, the data is collected from two online datasets such as Caltech-256 dataset [16] and PASCAL VOC 2007 dataset [17]. The Caltech 256 dataset contains 256 object classes or categories and a total number of 30607 images. In the undertaken dataset, each class includes almost 30 to 830 images. Additionally, the PASCAL VOC 2007 database includes 9963 images containing almost 24640 annotated objects. The undertaken database includes twenty classes, where every class includes almost 10 to 20 images.

- **Person:** cat, bird, dog, horse, cow, person and sheep,
- **Vehicle:** train, boat, car, bus, bicycle, aeroplane and motor bike
- **Indoor:** monitor, chair, bottle, potted plant, dining table and sofa.

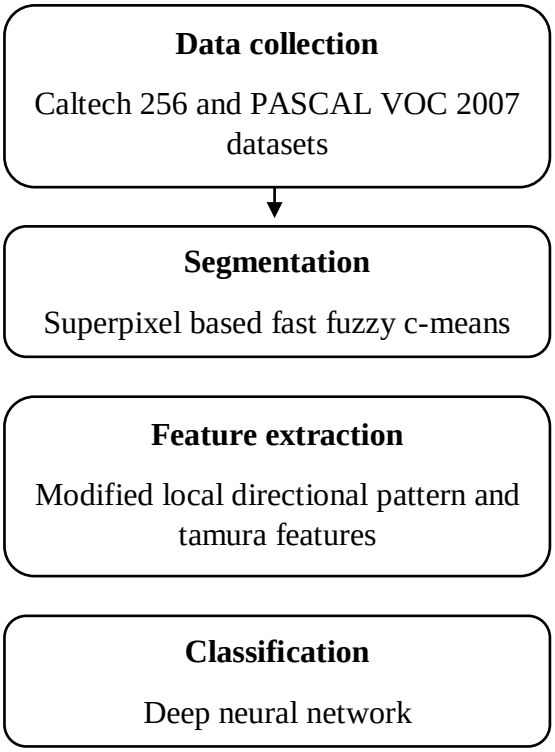

Figure 1. Block diagram of proposed model

3.2 Object segmentation

After the acquisition of video sequences, object segmentation is carried out by utilizing SFFCM algorithm [18]. For color image, the objective function of SFFCM is obtained based on Morphological Gradient Reconstruction-Watershed Transform (MGR-WT). The objective function of SFFCM algorithm is mathematically denoted in equation (1).

$$J_m = \sum_{l=1}^{q}\sum_{k=1}^{c} S_l u_{kl}^m \left\| \left(\frac{1}{S_l}\sum_{p\in R_l} x_p\right) - v_k \right\|^2 \quad (1)$$

Where, q is indicated as number of regions in the superpixel image, l is represented as color level $l \le l \le q$, x_p is stated as color pixel in the l^{th} region of superpixel image, S_l is denoted number of image pixels in the l^{th} region of R_l. The new objective function significantly lessen the computational complexity $l \ll N$, which is defined in equation (2).

$$\tilde{J}_m = \sum_{l=1}^{q}\sum_{k=1}^{c} S_l u_{kl}^m \left\| \left(\frac{1}{S_l}\sum_{p\in R_l} x_p\right) - v_k \right\|^2 - \lambda\left(\sum_{k=1}^{c} u_{kl} - 1\right) \quad (2)$$

Where, λ is stated as lagrange multiplier. The partial dirrential equation of $\tilde{J}_m$ is compuated on the basis of u_{kl} and v_k, as indicated in the equations (3) and (4).

$$\frac{\partial \tilde{J}_m}{\partial u_{kl}} = \sum_{l=1}^{q}\sum_{k=1}^{c} \frac{\partial S_l u_{kl}^m \left\| \left(\frac{1}{S_l}\sum_{p\in R_l} x_p\right) - v_k \right\|^2 - \lambda}{\partial u_{kl}} = 0 \quad (3)$$

$$\frac{\partial J_m}{\partial v_k} = \sum_{l=1}^{q}\sum_{k=1}^{c} \frac{\partial S_l u_{kl}^m \left\| \left(\frac{1}{S_l}\sum_{p\in R_l} x_p\right) - v_k \right\|^2 - \lambda}{\partial v_k} = 0 \quad (4)$$

By combing the equations (3) and (4), the final objective function is updated as shown in the equations (5) and (6). The sample collected and segmented video frames are graphically represented in the figures 2 and 3.

$$v_k - \frac{\sum_{l=1}^{q} u_{kl}^m \sum_{p\epsilon R_l} x_p}{\sum_{l=1}^{q} S_l u_{kl}^m} \quad (5)$$

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

$$u_{kl} = \frac{\left\| (\frac{1}{S_l}\sum_{p\in R_l} x_p) - v_k \right\|^{-2/(m-1)}}{\sum_{j=1}^{c} \left\| (\frac{1}{S_l}\sum_{p\in R_l} x_p) - v_j \right\|^{-2/(m-1)}} \quad (6)$$

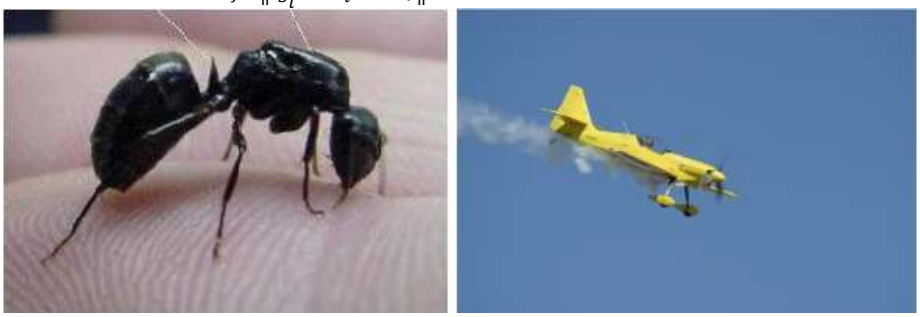

Figure 2. Sample collected images (a single video frame)

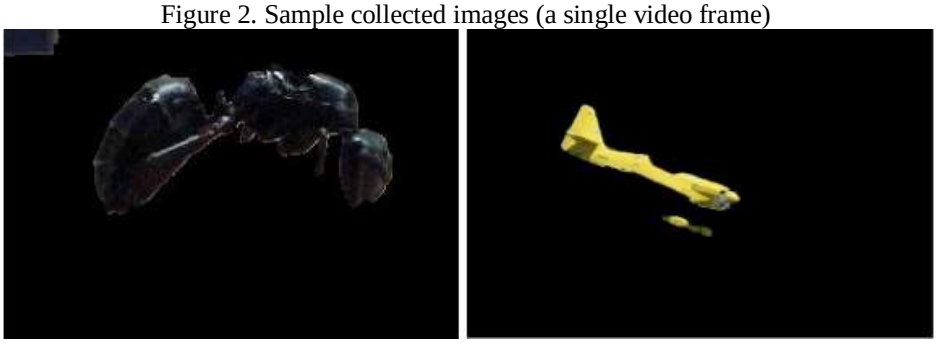

Figure 3. Sample segmented images

3.3 Feature extraction

After segmenting the objects from collected images, feature extraction is accomplished by using MLDP and tamura features. The undertaken methodologies extracts the active feature vectors which are utilized to achieve better classification. The description about the undertaken methodologies is given below.

3.3.1 Modified local directional pattern

The MLDP is an incorporation of LDP and Local Binary Pattern (LBP) techniques for preserving the arbitrary sequences of binarization weights that helps in improving the image orientation dependency. Let, the central pixel intensity is denoted as i_c at image pixel (x_c, y_c), and $i_n(n = 0,1, \dots 7)$ is indicated as pixel intensity in the 3×3 neighborhood of (x_c, y_c) that excludes the central pixel intensity i_c. In the conventional LDP method, eight kirsch masks are utilized for direction orientation that leads to intensity variation. So, the binarization weight is assigned to every neighbourhood pixel on the basis of kirsch output. Since, the LBP algorithm specifies the intensity variation over the neighbourhood pixels in the similar directions, and the kirsch mask output value is used for assigning the decimal to binary weight.

In MLDP, an exponential w_n $(n = 0,1, \dots 7)$ on the basis magnitude rank is assigned to the kirsch masks $m_n$corresponding to pixel intensity $i_n(n = 0,1, \dots 7)$. Then, the pixel (x_c, y_c) is updated in the traditional LDP as mentioned in the equations (7-9).

$$LDP_{k(x_c,y_c)} = \sum_{n=0}^{7} s(m_n - m_k).2^n \quad (7)$$

$$MLDP_{(x_c,y_c)} = \sum_{n=0}^{7} s(i_n - i_c).2^{w_n} \quad (8)$$

$$s(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & otherwise \end{cases} \quad (9)$$

Hence, the MLDP descriptor encodes the rotation invariance into main formulation. In addition, the MLDP descriptor significantly eliminates the empirical assignment to the parametric k value in the conventional LDP method.

3.3.2 Tamura features

Tamura is a texture feature descriptor that resembles human visual perception. Usually, it contains six features such as line-likeness, coarseness, directionality, regularity, contrast and roughness in that directionality, contrast, and line-likeness are used in this study for extracting the features.

Directionality: It considers both directional angle and edge strength of the image I, which is computed by using pixel wise derivative on the basis of prewitt's edge detector. The directional angle and edge strength is mathematically denoted in the equations (10) and (11).

$$Edge\ strength = 0.5(|\Delta_x(x,y)| + |\Delta_y(x,y)|) \quad (10)$$

$$Directional\ angle = arctan\frac{\Delta_x}{\Delta_y} + \frac{\pi}{2} \quad (11)$$

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

Where, Δ_y and Δ_x are indicated as pixel differences in y and x directions.

Contrast: In an image I, contrast measures the gray levels q to identify upto what extend the distribution is biased to white or black. The mathematical equation to calculate the contrast of the image I is denoted in equation (12).

$$Contrast = \frac{\sigma}{\sqrt[4]{\gamma_4}}, where\ \gamma_4 = \frac{\mu_4}{\sigma^2} \quad (12)$$

Where, σ^2 is indicated as variance, and μ_4 is represented as fourth mean value.

Line-likeness: It is defined as the average concurrence of edge directions that occur at pixels, which are separated by a distance d and the direction α. The extracted feature values are given as the input to DNN classifier for classifying the segmented objects.

3.4 Classification

The DNN is a feed forward network, which includes multiple hidden layers between the input and output layers. The DNN classifier identifies the mathematical manipulation by using a non-linear or linear relationship. In DNN, the network moves to the next preceding layers by determining the probability of output. The DNN classifier is mathematically represented in the equations (13) and (14).

$$Z^{(l)} = y^{(l-1)}W^{(l)} + b^{(l)} \quad (13)$$

$$y^{(l)} = g(\,Z^{(l)}) \quad (14)$$

Where, $g(.)$ is denoted as non-linear activation layer, $y^{(L)}$ is stated as final layer output, $y^{(l)} \in \mathbb{R}^{n_0}$ is indicated as output layer, $y^{(l-1)}$ is represented as the output of preceding layers $l-1$ and input to the layer l, $y^{(0)}$ is denoted as input to the model, $Z^{(l)}$ is represented as vector of pre-activations layers l, $l \in [1, \dots L]$ is denoted as l^{th} layer, and $W^{(l)} \in \mathbb{R}^{n_i \times n_0}$ is represented as matrix of learnable biases. In this research, ReLU activation function is applied in the hidden layers for better computational efficiency and fast learning. Besides, the output layers utilizes softmax classifier for delivering a probabilistic output that is mathematically stated in equation (15).

$$softmax\big(Z^{(L)}\big) = \frac{expZ_k}{\sum_{k=1}^{K} expZ_k} \quad (15)$$

Where, K is stated as output classes (object categories) and the output layer contains K neurons. In this research, cross entropy loss function is applied to reduce the optimization issue, which is formulated due to fast learning. The cross entropy loss function is mathematically defined in equation (16).

$$C = -\sum_{k=1}^{K} \hat{y}_k \log(y_k^{(L)}) \quad (16)$$

Where, $y^{(L)}$ is represented as DNN model output, and $\hat{y}_k \in \{0,1\}^k$ is denoted as encoded label.

4. Result and discussion

For experimental simulation, MATLAB 2019a software was used in this research study with 16 GB RAM, 4 GB GPU, i5 3.0 GHz processor, windows 10 operating system, and 3 TB memory. In addition, the proposed model performance was compared with ILHS [13] in order to show the benefit of developing the proposed model. **Databases:** In this work, the proposed model performance was analysed on two online datasets; Caltech 256, and PASCAL VOC 2007. The details about the acquired databases were specified in table 1.

Table 1. Details about undertaken databases

Dataset	Released	Categories	Total images	Images per category			
				Minimum	Median	Mean	Maximum
Caltech 256	2006	257	30607	80	100	119	827
PASCAL VOC 2007	2007	20	9963	10	15	14	20

In this study, precision, specificity, recall, and accuracy were utilized for evaluating the proposed model performance. The mathematical expression about the undertaken performance measure is represented in the equations (17-20)

$$Precision = \frac{TP}{FP+TP} \times 100 \quad (17)$$

$$Specificty = \frac{TN}{FP+TN} \times 100 \quad (18)$$

$$Recall = \frac{TP}{FN+TP} \times 100 \quad (19)$$

$$Accuracy = \frac{TP+TN}{FN+TP+FP+TN} \times 100 \quad (20)$$

Where, FN is indicated as false negative, FP is denoted as false positive, TP is stated as true positive, and TN is specified as true negative.

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

4.1 Quantitative investigation on Caltech-256 dataset

In this subdivision, Caltech 256 database is undertaken for analysing the proposed model performance with dissimilar classification techniques like Multi SVM (MSVM), KNN, random forest, and DNN. In this scenario, the proposed model performance is studied with dissimilar classification techniques by means of precision, specificity, recall, and accuracy. Hence, the DNN classifier achieved superior performance in object detection and classification related other classification techniques. In this dataset, the DNN classifier attained 97.58% of accuracy, which showed maximum of 18.9% and minimum of 2.8% improvement in accuracy related to other techniques; MSVM, KNN, and random forest. By inspecting table 2, the DNN classifier also showed significant performance in object classification related to other classifiers. Figure 4 represents the graphical valuation of proposed model with dissimilar classifiers on Caltech 256 database.

Table 2. Performance investigation on Caltech-256 dataset with dissimilar classification techniques

Classifier	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
MSVM	94.20	93.26	95.10	94.78
KNN	90.14	89.23	92.36	92.66
Random forest	76.67	78.25	80.13	78.66
DNN	**97.29**	**97.01**	**96.76**	**97.58**

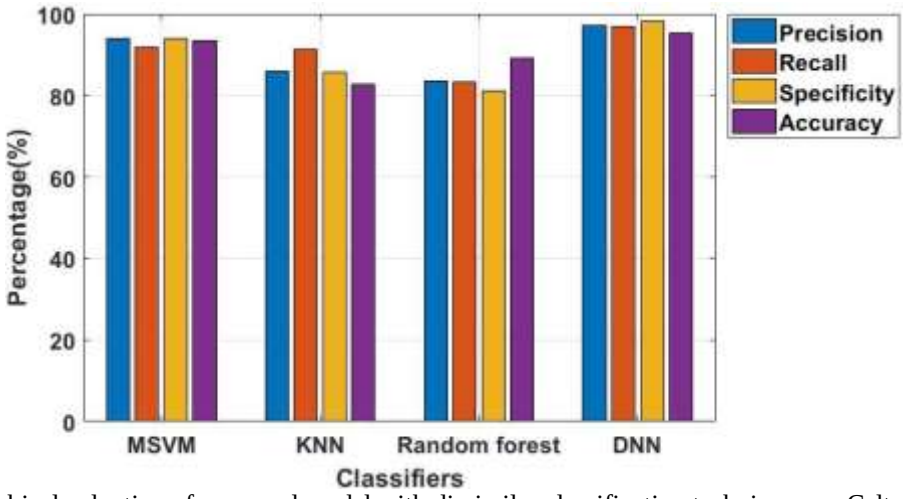

Figure 4. Graphical valuation of proposed model with dissimilar classification techniques on Caltech-256 dataset

In addition, the proposed model performance is investigated with different combination of features on Caltech-256 dataset. By investigating table 3, the proposed model (hybrid feature with DNN) attained 97.58% of classification accuracy, which showed maximum of 14.6% and minimum of 10.3% improvement in accuracy related to the individual features; MLDP-DNN, and tamura-DNN. Additionally, the proposed model showed significant performance in object classification compared to the individual features by means of specificity, recall, and precision. Figure 5 states the graphical valuation of proposed model with dissimilar feature combinations on Caltech-256 dataset.

Table 3. Performance investigation on Caltech-256 dataset with different combination of features

Features	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
MLDP-DNN	89.33	86.448	84.97	87.28
Tamura-DNN	74.51	79.297	81.75	82.92
Hybrid feature-DNN	**98.29**	**98.01**	**98.76**	**97.58**

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

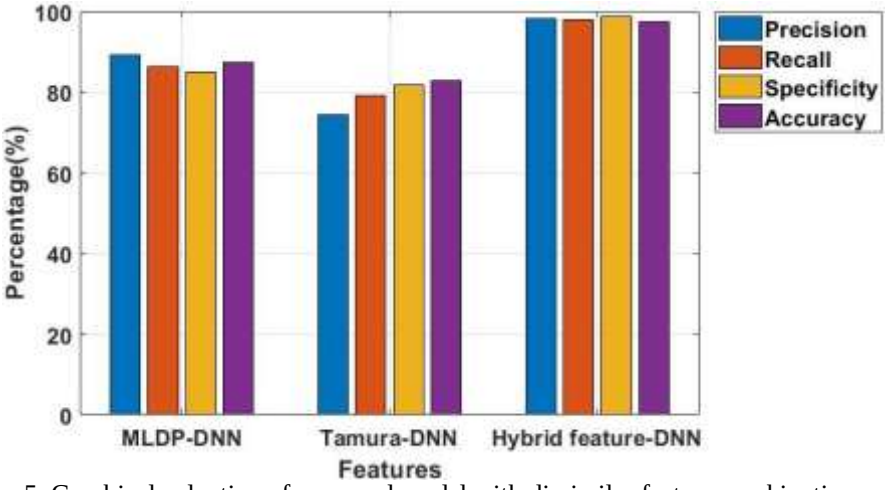

Figure 5. Graphical valuation of proposed model with dissimilar feature combinations on Caltech-256 dataset

4.2 Quantitative investigation on PASCAL VOC 2007 dataset

In this segment, PASCAL VOC 2007 database is undertaken for analysing the proposed model performance with dissimilar classifiers (MSVM, KNN, random forest, and DNN), and feature combinations (MLDP-DNN, Tamura-DNN, and Hybrid feature-DNN). By studying the tables 4 and 5, the proposed model (Hybrid feature-DNN) attained good performance in object classification by means of recall, precision, accuracy, and specificity.

Table 4. Performance investigation on PASCAL VOC 2007 dataset with dissimilar classification techniques

Classifier	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
MSVM	94.07	91.75	94.09	93.42
KNN	85.95	91.44	85.75	82.79
Random forest	83.43	83.33	81.27	89.259
DNN	**97.19**	**97.05**	**98.28**	**95.26**

Table 5. Performance investigation on PASCAL VOC 2007 dataset with dissimilar combination of features

Features	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)
MLDP-DNN	90.72	91.48	92.43	91.13
Tamura-DNN	72.30	75.29	74.20	72.86
Hybrid feature-DNN	**97.19**	**97.05**	**98.28**	**95.26**

Hence, the figures 6 and 7 represents the graphical valuation of proposed model with dissimilar classifiers and feature combinations on PASCAL VOC 2007 dataset.

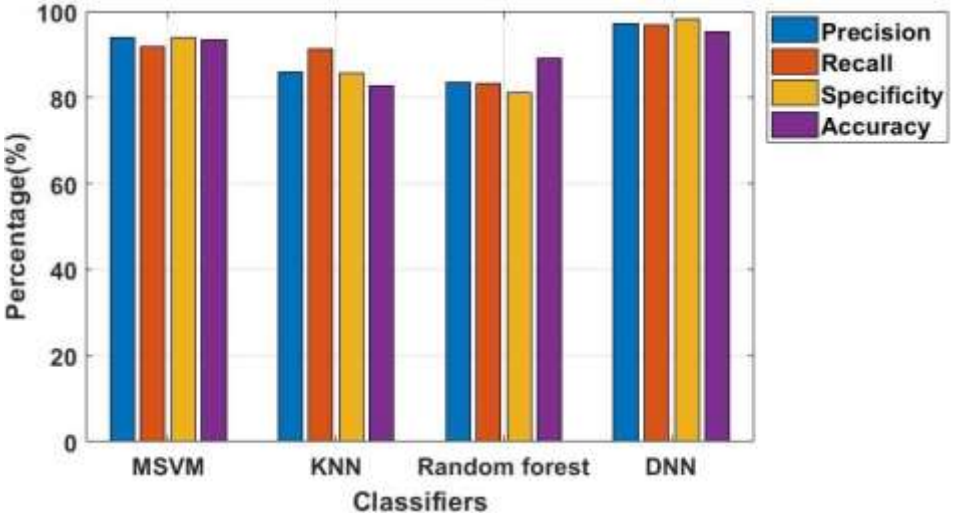

Figure 6. Graphical valuation of proposed model with dissimilar classification techniques on PASCAL VOC 2007 dataset

ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

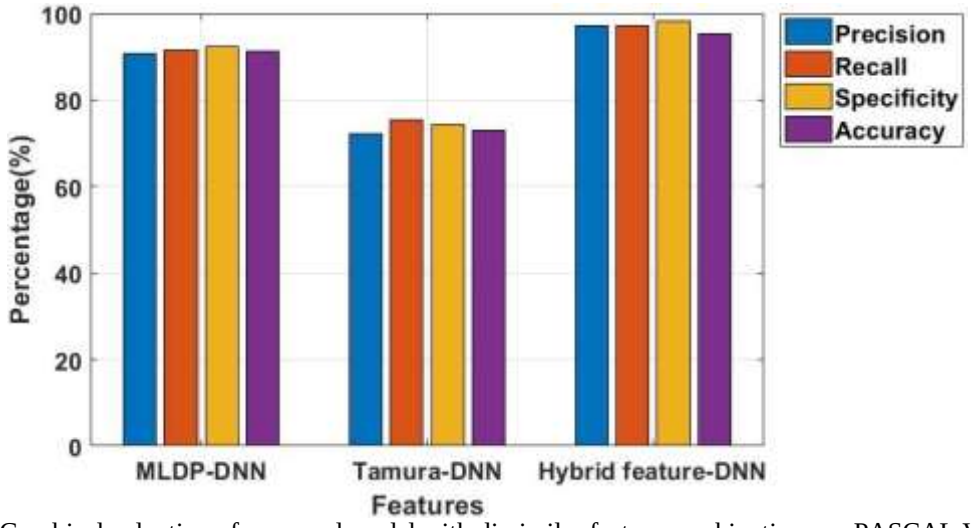

Figure 7. Graphical valuation of proposed model with dissimilar feature combinations on PASCAL VOC 2007 dataset

4.3 Comparative study

In this segment, table 6 indicates the comparative study of proposed and existing model. Zhang, *et al*, [13] presented ILHS classifier for object detection by utilizing semantic and visual similarities. By exploring both the semantic and visual similarities, a new image representation was generated. Then, cluster the images for exploring the correlations and reweight the visual similarities for generating a new image representation. In this literature, the developed model performance was validated on PASCAL VOC 2007 and Caltech 256 databases. Therefore, the developed model averagely attained 77% of precision in PASCAL VOC 2007 dataset and 80% of precision in Caltech 256 dataset. Related to this research paper, the proposed model attained superior performance in object classification using video sequences.

Table 6. Comparative study

Methods	Datasets	Average precision (%)
ILHS [13]	PASCAL VOC 2007	77
	Caltech 256	80
Hybrid feature-DNN	**PASCAL VOC 2007**	**97.19**
	Caltech 256	**98.29**

5. Conclusion

In this research article, a new hybrid feature descriptor is proposed with DNN classifier for enhancing the object detection and classification performance. Initially, SFFCM approach is undertaken for segmenting the annotated objects from the frames and then the active feature values extracted by applying hybrid features. The corresponding extracted feature values are classified as object classes by implementing DNN classifier. In the result section, the proposed model (Hybrid feature-DNN) performance is validated in light of recall, precision, specificity and accuracy. The hybrid feature-DNN showed 20.19% and 18.29% improvement in average precision related to ILHS in PASCAL VOC 2007, and Caltech 256 datasets, respectively. In future work, an optimization technique can be included in the proposed model for decreasing the dimension of the extracted features to further improve the object classification performance in the video sequences.

6. References

[1] A. Sasithradevi, and S.M.M. Roomi, "Video classification and retrieval through spatio-temporal Radon features", *Pattern Recognition*, vol.99, pp.107099, 2020.

[2] M.K. Geetha, S. Palanivel, and V. Ramalingam, "A novel block intensity comparison code for video classification and retrieval", *Expert Systems with Applications*, vol.36, no.3, pp.6415-6420, 2009.

[3] M.U. Yaseen, A. Anjum, O. Rana, and R. Hill, "Cloud-based scalable object detection and classification in video streams", *Future Generation Computer Systems*, vol.80, pp.286-298, 2018.

[4] N. Najva, and K.E. Bijoy, "SIFT and tensor based object detection and classification in videos using deep neural networks", *Procedia Computer Science*, vol.93, pp.351-358, 2016.

[5] Y. Gurwicz, R. Yehezkel, and B. Lachover, "Multiclass object classification for real-time video surveillance systems", *Pattern Recognition Letters*, vol.32, no.6, pp.805-815, 2011.

[6] Y. Tang, L. Jing, H. Li, and P.M. Atkinson, "A multiple-point spatially weighted k-NN method for object-based classification", *International journal of applied earth observation and geoinformation*, vol.52, pp.263-274, 2016.
[7] W. Zhang, P. Zhou, S. Li, and H. Yan, "Arbitrary size ratio canonical object classification based on Convolutional Neural Network", *Procedia computer science*, vol.147, pp.102-108, 2019.
[8] D.E. Hernández, E. Clemente, G. Olague, and J.L. Briseño, "Evolutionary multi-objective visual cortex for object classification in natural images", *Journal of Computational Science*, vol.17, pp.216-233, 2016.
[9] J. Xu, L. Zhao, S. Zhang, C. Gong, and J. Yang, "Multi-task learning for object keypoints detection and classification", *Pattern Recognition Letters*, 2018.
[10] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "A robust approach for object matching and classification using Partial Dominant Orientation Descriptor", *Pattern Recognition*, vol.64, pp.168-186, 2017.
[11] T. Gong, B. Liu, Q. Chu, and N. Yu, "Using multi-label classification to improve object detection", *Neurocomputing*, vol.370, pp.174-185, 2019.
[12] P. Tang, X. Wang, Z. Huang, X. Bai, and W. Liu, "Deep patch learning for weakly supervised object classification and discovery", *Pattern Recognition*, vol.71, pp.446-459, 2017.
[13] C. Zhang, J. Cheng, and Q. Tian, "Image-level classification by hierarchical structure learning with visual and semantic similarities", *Information Sciences*, vol.422, pp.271-281, 2018.
[14] M. Rashid, M.A. Khan, M. Sharif, M. Raza, M.M. Sarfraz, and F. Afza, "Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features", *Multimedia Tools and Applications*, vol.78, no.12, pp.15751-15777, 2019.
[15] T. Mahalingam, and M. Subramoniam, "ACO–MKFCM: An Optimized Object Detection and Tracking Using DNN and Gravitational Search Algorithm", *Wireless Personal Communications*, vol.110, no.3, pp.1567-1604, 2020.
[16] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset", 2007.
[17] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge", *International journal of computer vision*, vol.88, no.2, pp.303-338, 2010.
[18] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A.K. Nandi, "Superpixel-based fast fuzzy C-means clustering for color image segmentation", *IEEE Transactions on Fuzzy Systems*, vol.27, no.9, pp.1753-1766, 2018.

Dataset links:

Caltech-256 dataset: https://www.kaggle.com/jessicali9530/caltech256

PASCAL VOC 2007 dataset: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/