International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/2231280453/IJCTT-V67I3P122

Volume 67 Issue 3,113-118, March 2019
© 2019 Seventh Sense Research Group®

Original Article

Object Recognition using Deep Convolutional Neural Network

Ali Raza[1], Qian yurong[2]

[1,2] Department of software college, Xinjiang University (830002) Xinjiang, China.

Abstract *- Recognizing an object in an image is one of the principal difficulties of PC vision frameworks because of the varieties that each item or the particular picture where the object is indicated could have, like enlightenment or perspective. Deep Neural Networks (DNNs) have recently revealed exceptional execution on picture classification tasks [14]. In a recent paper, we go one step further and identify the issue of object detection utilizing DNNs that aren't only classifying but also precisely localizing objects of different periods. We present a simple yet powerful formulation of object identification as a regression issue to object bounding box masks. We characterize a multi-scale deduction system that can create high-resolution object detections at a low cost by a few network applications. Best in class execution of the methodology has appeared on Pascal VOC.*

Keywords **-** *Database, Benchmark, Object recognition, DNN.*

I. INTRODUCTION

Unique mark acknowledgment for security confirmation or legal applications, medical imaging utilize numerous body studies about an individual's mind morphology while the yage, observation to distinguish and screen interlopers, or checking shorelines/pools for drowning are the victims of nowadays to recognize real-world applications of PC vision [1].

As we progressively understand the total image, having increasingly exact and point-by-point object acknowledgment becomes significant. In this specific circumstance, one thinks about ordering pictures, yet in addition to accurately evaluating the class and area of items contained inside the pictures, an issue known as object detection. At the same time, physical design representations concerning shallow discriminatively prepared models have been among the best accomplishment ideal models for the related issue of object classification [2]. However, in recent years, Deep Neural Networks (DNNs) [3] have risen as a powerful machine learning model.

The PASCAL Visual Object Classes (VOC) Challenge comprises two parts: (i) an openly accessible dataset of pictures and explanation, together with standardized assessment programming, and (ii) a yearly challenge and workshop [4]. The VOC2007 dataset comprises commented consumer photographs gathered from the Flickr photo-sharing website [5]. Another dataset with ground truth annotation has been released every year since 2006. There are two main difficulties: classification—"does the image contain any occasions of a specific object class?" (Where the object classes include cars, people, dogs, etc.), and detection—"where are the occurrences of a particular object class in the image (if any)?" Furthermore, there are two backup challenges ("tasters") on pixel-level segmentation—assign each pixel a class mark, and "person layout"—localize the head, hands, and feet of the individual in the picture [5]. The difficulties are issued with due dates each year, and a workshop is held to look at and talk about that year's outcomes and techniques. The datasets and related explanation software are subsequently released and accessible for use.

The goals of the VOC challenge are twofold: first, to give testing pictures and high-quality annotation, together with a standard assessment methodology—a "plug and play" training and testing harness so that the performance of algorithms can be compared (the dataset component); and second to measure state of the art each year (the competition component) [6].

We present a plan which is fit for foreseeing the jumping boxes of various items in a given picture. We define a DNN-based regression that yields a binary mask of the object-bounding box (and portions of the box). Furthermore, we utilize a basic jumping box inference to extract detections from the masks. We apply the DNN mask generation in a multi-scale design on the full image and a few large image crops to expand localization precision, followed by a refinement step [7]. Along these lines, only through a few dozen DNN-regressions can we accomplish state-of-art bounding box localization. In addition, the presented method is quite simple. There is no need to hand-design a model that explicitly captures parts and their relations. This simplicity has the advantage of easy applicability to a wide scope of classes and shows better detection performance across a wider range of rigid and deformable objects. This is presented with state-of-the-art detection results on the Pascal VOC challenge [8] in Sec. 7.

II. RELATED ARCHITECTURES

A. Deep Q-networks

A deep Q-network (DQN) is a deep learning model that combines a deep CNN with Q-learning reinforcement learning. Unlike earlier reinforcement learning agents, DQNs can learn directly from high-dimensional sensory inputs. Preliminary results were presented in 2014, with an accompanying paper in February 2015. The research described an application to Atari 2600 gaming. Other deep reinforcement learning models preceded it.

B. Deep belief networks

Convolutional deep belief networks (CDBN) are very similar to convolutional neural networks and are trained similarly to deep belief networks. Therefore, they exploit the 2D structure of images, as CNN does, and use pre-training like deep belief networks. They provide a generic structure that can be used in many image and signal processing tasks. Benchmark results on standard image datasets like CIFAR have been obtained using CDBNs.

C. Literature

A standout amongst the most heavily studied paradigms for object identification is the deformable part-based model, with is the most prominent example [9]. This technique consolidates a set of discriminatively trained parts in a star model called pictorial structure. It can be considered a 2-layer model – parts being the primary layer and the star display being the second layer. As opposed to DNNs, whose layers are conventional, the work by [10] exploits domain knowledge – the parts depend on physically plane Histogram of Gradients (HOG) descriptors [11], and the structure of the parts is kinematically motivated.

It very well may be characterized as the task of discovering a certain object in an image or even in a video sequence. It is a fundamental vision issue since, unlike humans that can detect and identify with almost no effort, a huge range of objects in images or videos might diverge from the viewpoint, color, size, or even when the object is partially obstructed. This task is a genuine test for item acknowledgment motors [12].

Deep architectures for object detection and parsing have been motivated by part-based models. They traditionally are called compositional models, where the object is expressed as a layered composition of image primitives. A notable example is the And=Or graph [13], where a tree models an object with And-nodes representing different parts and Or-nodes representing different modes of the same part. Similar to DNNs, the And=Or graph consists of multiple layers, where lower layers represent small generic image primitives while higher layers represent object parts. Such compositional models are easier to interpret than DNNs.

On the other hand, they require inference, while the DNN models considered in this paper are purely feed-forward with no latent variables to be inferred. The compositional models for detection are based on segments as primitives [14], focus on shape [15], and use Gabor filters [16] or larger HOG filters [17]. These approaches are traditionally challenged by the difficulty of training and specially designed learning procedures. Moreover, at inference time, they combine bottom-up and top-down processes.

D. Experimental

Performance of the proposed methodology on the test set of the Pascal Visual Object Challenge (VOC) 2007 [7]. The dataset contains approx. 5000 test images for more than 20 classes. Since our approach has many parameters, we train on the VOC2012 preparing and approval set, which has approx. 11K images. At test time, a calculation produces for an image, a set of identification, analyze bounding boxes and their class labels. We use accuracy review curves and average precision (AP) per class to quantify the algorithm's performance. Fig.1 indicates the acknowledgment of altered constituents.

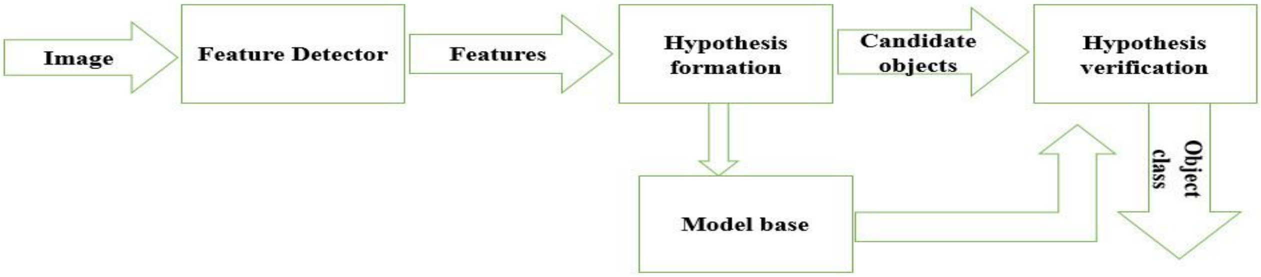

Fig. 1 The object recognition system of different components

III. IMAGE COLLECTION PROCEDURE

For the 2015 test, all images were collected from the Flickr photo-sharing website. The utilization of individual photos that were not taken by, or chosen by, vision/machine learning researchers results in a much "unbiased" dataset. The photos are not taken with a particular purpose in mind, for example, object acknowledgment. Qualitatively the images contain an extremely wide scope of review conditions (pose, lighting, and so on), and images where there is little bias toward images being "of" a particular object, e.g., there are images of motorcycles in a street scene, rather than solely images where a motorcycle is the focus of the picture. The annotation guidelines guided annotators on which images to annotate—essentially everything which could be annotated with confidence. Utilizing a single source of "consumer" images addressed issues experienced in past difficulties [18].

Selection bias is presented by a researcher physically performing image determination. The “person” category provides a vivid example of how the adopted collection methodology leads to high variability; in past datasets, "person" was essentially synonymous with "pedestrian," whereas, in the VOC dataset, we have images of people engaged in a wide range of activities such as walking, riding horses, sitting on buses, etc., was given inFig.2.

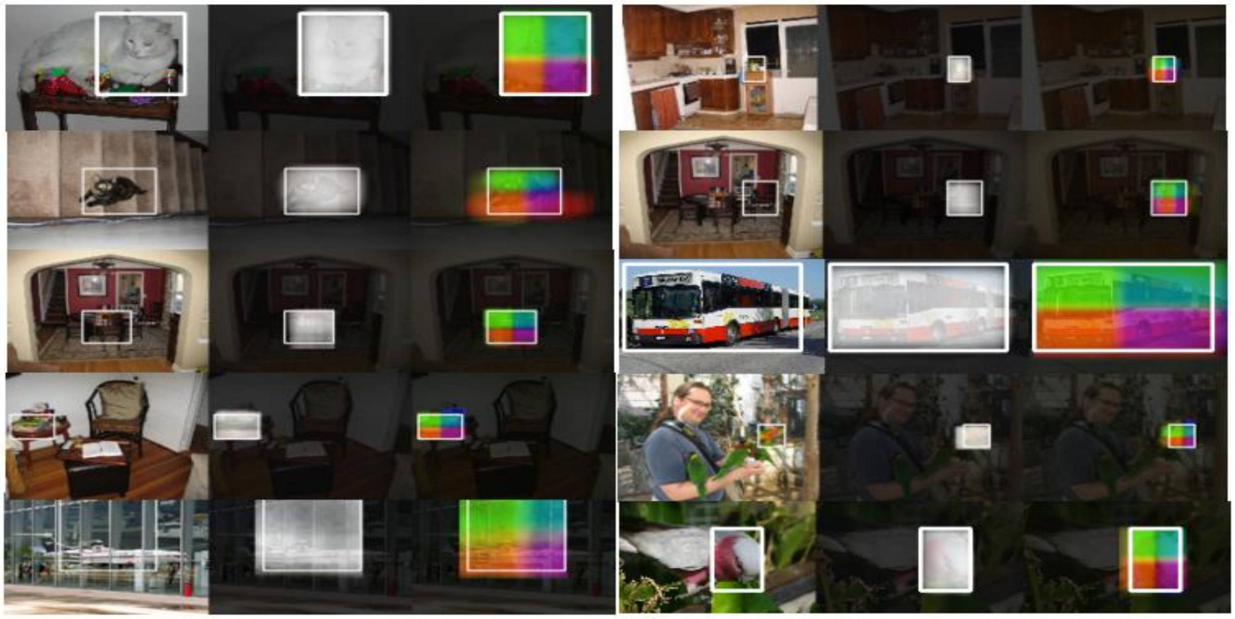

Fig. 2 For each image, we show two heat maps on the right: the first one corresponds to the output of DNN full, while the second one encodes the four partial masks in terms of the strength of the colors red, green, blue, and yellow. In addition, we visualize the estimated object bounding box. All examples are correct detections, except the examples in the last row.

IV. RESULT AND DISCUSSION

A. Convolutional Neural Networks Basic Structure

As indicated by one of the pioneers of CNN's design, LetNet-5, recommended that the CNN basic architecture must have the convolutional layer, pooling layer, and fully associated layers [19]. In general, this design was the basis of another and more recent CNN structure, keeping the core concepts intact regardless of the improvements and modifications.

B. The Convolutional Layer

The possibility of a convolution when talking of CNNs is to separate the highlights from the picture, preserving the spatial association between the pixels and the learned features inside the picture with small equally-sized tiles.

The learned features result from scientific activity between every component from the input image and the channel lattice. In other words, the filter, also known as the feature detector, slides through all picture components and is increased by everyone delivering the sum of multiplication yield a single matrix named Feature Map was shown in Fig.3.

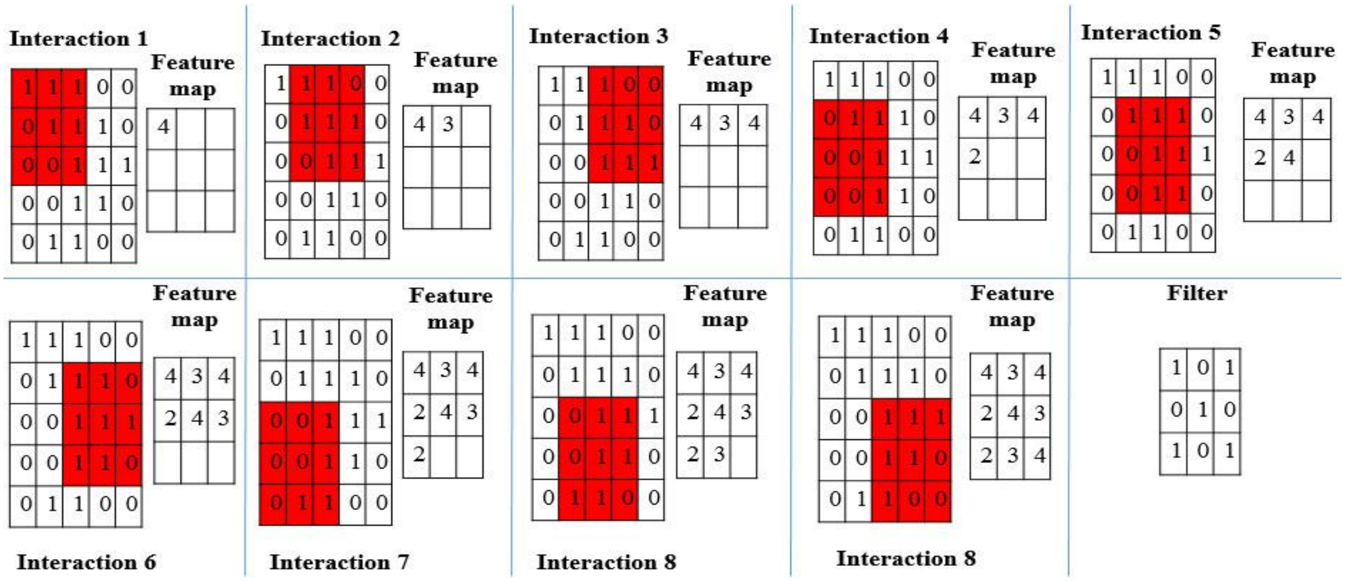

Fig. 3 Example of a convolution

The total assessment of the VOC2007 test is given in Table 1. We think about our methodology, named Detector Net, to three related approaches. The first is a sliding window version of a DNN classifier [20]. After preparing this system as a 21-way classifier (VOC classes and foundation), we produce bounding boxes with 8 distinctive aspect ratios, and 10 unique scales paced 5 pixels apart. The littlest scale is one10th of the image size, while the biggest spreads the entire image. These outcomes in approximately 150; 000 boxes per image. Each case is mapped affinely to the 225 x 225 receptive field. The softmax classifier figures the recognition score. We reduce the quantity of the boxes by non-maximum suppression using Jaccard similarity of at least 0:5 to discard boxes. After the underlying preparation, we performed two rounds of hard negative mining on the preparation set. This added two million guides to our original training set and has chopped down the proportion of false positives. Table 1 demonstrates the mean average precision (mAP), FPS, and the number of bounding boxes by each detection system on the PASCAL VOC 2007 dataset.

The second approach is the 3-layer compositional model, considered a deep architecture. As a co-winner of VOC2011, this approach has shown excellent performance. Finally, we compare it against the DPM [21, 22].

Even though our examination is fairly out of line, as we prepared on the larger VOC2012 training set, we show state-of-the-art performance on most models: we outperform on 8 classes and perform on par on other 1. Note that it may be conceivable to tune the sliding window to perform on par with Detector Net. However, the sheer amount of network assessments makes that approach infeasible, while Detector Net requires only (#windows x #mask types) 120 crops per class to be assessed. On a 12-core machine, our execution took around 5-6 secs for every image for each class.

We indicate instances of the detections in Fig. 3, where both the detected box and all five generated masks are visualized. It can be seen that the DetectorNet is capable of precisely finding extensive and little items. The generated veils are limited and have almost no reaction outside the object. Such high-quality detector responses are hard to achieve and, in this case, possible because of the expressive intensity of the DNN and its common method for fusing context.

C. Choice of Classes

Fig.2 shows the 20 classes selected for annotation in the VOC2007 dataset. The classes can be considered in taxonomy with four main branches—vehicles, animals, household objects, and people. The figure also shows the challenge year in which a particular class was included. In the original VOC2005 challenge, which used existing annotated datasets, four classes were annotated (car, motorbike, bicycle, and person). This number was increased to 10 in VOC2006 and 20 in VOC2007, as indicated in Fig.4.

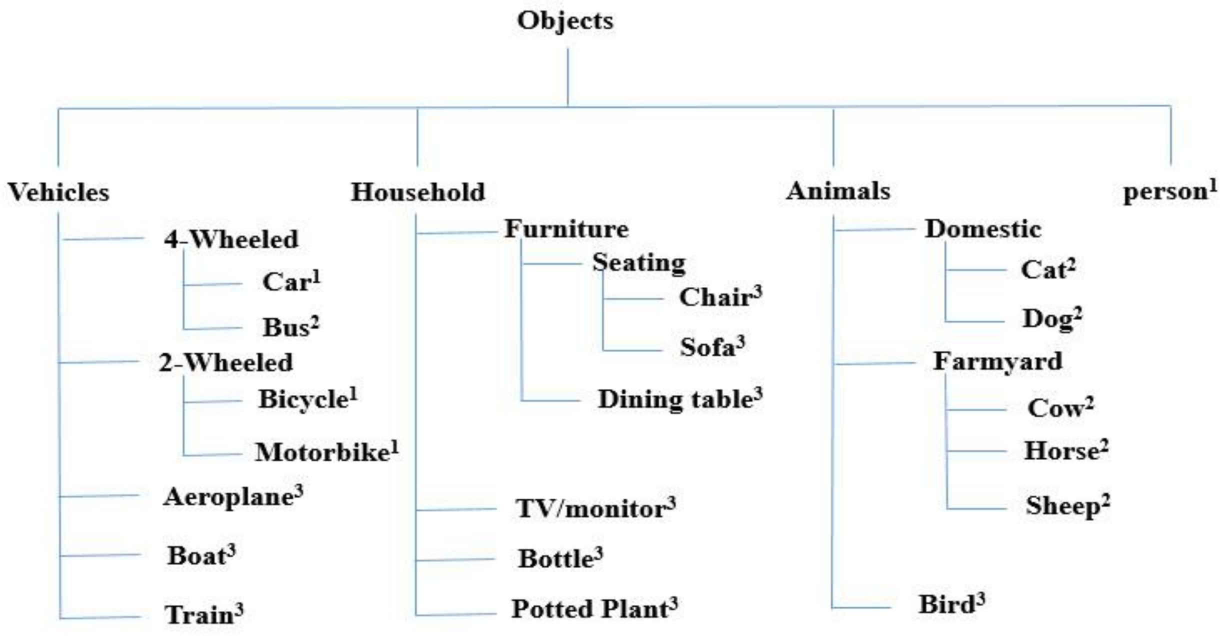

Fig. 4 VOC2007 Classes. Leaf nodes correspond to the 20 classes. The year of inclusion of each class in the challenge is indicated by superscripts: 2005^{1}, 2006^{2}, and 2007^{3}. The classes can be considered a notional taxonomy, with successive challenges adding new branches (increasing the domain) and leaves (increasing detail).

The common misdetections are similar-looking items (left object in the last row of Fig. 3) or imprecise localization (right object in the last row). The last issue is due to the ambiguous definition of object extended by the training data – in certain images, only the head of the bird is visible, while in others, the full body. While most of the time, we might observe a detection of both the body and face if they are both present in a similar image. Whereas table 1 is used to analyze the execution and speed of the object detector system.

Table 1. Comparing the performance and speed of the object detector system on Pascal VOC 2007.

Models	mAP	FPS	Number of boxes
RCNN	57.9	5.8	-
Fast-RCNN	69.7	0.6	-
Faster-RCNN(VGG16)	72.6	7.1	300.5
YOLO	64.1	44.6	98.3
Fast YOLO	53.3	156.1	98.1
SSD300(VGG)	71.7	58.4	7308.4
SSD500(VGG16)	74.8	23.2	20097.7

V. APPLICATIONS

A. Image Recognition and Compression

Neural Networks have already been deployed in Image Recognition Applications[23]. Neural Networks have the property of creating a lower-dimensional internal representation of the input. This has been tapped to create algorithms for image compression. These techniques fall into three main categories - direct development of neural learning algorithms for image compression, neural network implementation of traditional image compression algorithms, and indirect applications of neural networks to assist with those existing image compression techniques [24].

B. Speech Recognition

Most current speech recognition systems use Hidden Markov Models (HMMs) to deal with the temporal variability of speech and Gaussian Mixture Models (GMMs) to determine how well each state of each HMMs fits a frame or a short window of frames of coefficients that represent the acoustic input. Deep neural networks with many hidden layers trained using new methods have been shown to outperform GMMs on various speech recognition benchmarks, sometimes by a large margin [25].

C. Medical Diagnosis

There are vast amounts of medical data in-store today, in medical images, doctors' notes, and structured lab tests. Convoluted Neural Networks have been used to analyze such data. For example, in medical image analysis, designing a group of specific features for a high-level task such as classification and segmentation is common. But detailed annotation of medical images is often an ambiguous and challenging task. It has been shown that deep neural networks have been effectively used to perform these tasks [26].

VI. CONCLUSION

In a recent project, it was built up a deep learning model depended on convolutional neural networks (CNNs) that proposed to be capable of recognizing and detecting fashion items in a static image. Moreover, we leverage the expressivity of DNNs for the object identifier. We demonstrate that the basic formulation of detection as DNN-base object mask regression can yield a solid outcome when utilizing a multi-scale coarse-to-fine procedure. These outcomes originate at some computational cost at training time – one needs to prepare a network per object type and mask type. In future work, we aim to reduce the cost by utilizing a single network to identify objects of various classes and thus expand to a larger number of classes.

ACKNOWLEDGEMENT

The Department of Software Engineering, Xinjiang University China, is highly acknowledged for providing the research facilities.

REFERENCES

[1] Li, P., Wang, Q., Zeng, H., et al.: 'Local log-Euclidean multivariate Gaussian descriptor and its application to image classification, IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39, (4), pp. 803–817.

[2] Park, D.-C.: 'Multiple feature-based classifiers and its application to image classification. IEEE Int. Conf. Data Mining Workshops, 2010, pp. 65–71.

[3] Yu, K., Zhang, T.: 'Improved local coordinate coding using local tangents .'Proc. 27th Int. Conf. Machine Learning (ICML), 2010, pp. 1215–1222.

[4] Zhang, J., Marszalek, M., Lazebink, S., et al.: 'Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vis., 2007, 73, (2), pp. 213–238.

[5] Gehler, P.-V., Nowozin, S.: 'On feature combination for multiclass object classification. Proc. IEEE 12th Int. Conf. Computer Vision (ICCV), 2009.

[6] Khan, F.-S., van de Weijer, J., Vanrell, M.: 'Modulating shape features by color attention for object recognition, Int. J. Comput. Vis., 2012, 98, pp. 49–64.

[7] Xiao, J., Hays, K., Ehinger, A., et al.: 'Sun database: large-scale scene recognition from abbey to zoo .'Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3485–3492.

[8] Zeiler, M.D., Fergus, R.: 'Visualizing and understanding convolutional networks, 2014, pp. 818–833.

[9] Dixit, M., Chen, S., Gao, D., et al.: 'Scene classification with semantic Fisher vectors .'Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2974–2983.

[10] Lin, G., Zhu, H., Kang, X., et al.: 'Feature structure fusion modelling for classification,' IET Image Process., 2015, 9, (10), pp. 883–888.

[11] Zhu, Q.-H., Wang, Z.-Z., Mao, X.-J., et al.: 'Spatial locality-preserving feature coding for image classification, Appl. Intell., 2017, 47, (1), pp. 148–157.

[12] Sun, M., Han, T.-X., Liu, M.-C., et al.: 'Latent model ensemble with autolocalization .'Proc. Int. Conf. Pattern Recognition (ICPR), 2016.

[13] Khan, S.H., Hayat, M., Bennamoun, M., et al.: 'A discriminative representation of convolutional features for indoor scene recognition,' IEEE Trans. Image Process., 2016, 25, (7), pp. 3372–3383.

[14] Hayat, M., Khan, S.H., Bennamoun, M., et al.: 'A spatial layout and scale invariant feature representation for indoor scene classification,' IEEE Trans. Image Process., 2016, 25, (10), pp. 4829–4841.

[15] He, K., Zhang, X., Ren, S., et al.: 'Deep residual learning for image recognition. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.

[16] Dai, J., Li, Y., He, K., et al.: 'R-FCN: object detection via region-based fully convolutional networks, Adv. Neural Inf. Process. Syst., 2016, pp. 379–387.

[17] Oquab, M., Bottou, L., Laptev, I., et al.: 'Learning and transferring mid-level image representations using convolutional neural networks. Proc. IEEE Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1717–1724.

[18] Zhang, Y., Shi, B.: 'Improving pooling method for regularizing convolutional networks based on the failure probability density, Opt. – Int. J. Light Electron Opt., 2017, 145, (Suppl. C), pp. 258–265.

[19] Snoek, J., Rippel, O., Swersky, K., et al.: 'Scalable Bayesian optimization using deep neural networks. Int. Conf. Machine Learning (ICML), 2015, pp. 2171–2180.

[20] Thangarajah, A., Wu, Q.J., Yimin, Y.: 'Fusion-based foreground enhancement for background subtraction using multivariate multi-model Gaussian distribution,' Inf. Sci., 2018, 430, pp. 414–431.

[21] Y. Bassil "Phoenix-The Arabic Object Oriented Programming Language" International Journal of Computer Trends and Technology 67.2 (2019): 7-11.

[22] Bahrampour, S., Nasrabadi, N.M., Ray, A., et al.: 'Multimodal task-driven dictionary learning for image classification, IEEE Trans. Image Process. 2016, 25, pp. 24–38.

[23] Chen, S., Yang, J., Luo, L., et al.: 'Low-rank latent pattern approximation with applications to robust image classification, IEEE Trans. Image Process., 2017, 26, (11), pp. 5519–5530.

[24] Gao, Z., Fatih, P., Hongdong, L.: 'Robust visual tracking with deep convolutional neural network based object proposals on pets .'Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 26–33.

[25] Wenling, S., Kihyuk, S., Diogo, A., et al.: 'The extraordinary link between deep neural networks and the nature of the universe,' MIT Technol. Rev., 2016.

[26] Fei-Fei, L., Fergus, R., Perona, P.: 'Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories', Comput. Vis. Image Underst., 2007, 106, (1), pp. 59–70.