IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 33, No. 5, October, 2019, pp. 367-372
Journal homepage: http://iieta.org/journals/ria

Object Detection in Convolution Neural Networks Using Iterative Refinements

Vijay Vasanth Aroulanandam[1], Thamarai Pugazhendhi Latchoumi[2*], Battula Bhavya[2], Shaik Sajida Sultana[2]

[1] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai 600062, India
[2] Department of Computer Science and Engineering, VFSTR (Deemed to be University), AP 522213, India

Corresponding Author Email: tpl_cse@vignan.ac.in

https://doi.org/10.18280/ria.330506

Received: 11 June 2019
Accepted: 16 August 2019

Keywords:
convolutional neural networks, object detection, localization refinement, region-based CNN, stochastic gradient descent

ABSTRACT

One of the major problems in computer vision is the object detection through image classification. In recent years, the convolutional neural network (CNN) has been extensively applied for image classification. This paper combines iterative refinement and joint score function into a strategy to accurately localize the detected objects. First, a unified model with fast approximation was proposed to correct the position and range of region proposals, which are often incorrect in conventional linear methods. Focusing on data, the model can acquire knowledge without any cost, and suit different CNN architectures for various datasets. Next, a joint score function was introduced to process the number of candidate regions in the images. The joint score function deals with the relative position of the occluded object, and depends on the image data and output loss. Experimental results show that the proposed strategy achieved a 3.6% higher mean precision than the contrastive method. The research greatly promotes the object detection accuracy in computer vision.

1. INTRODUCTION

In an early period, the effectiveness of CNN in deep learning is regularly joined with an article of the object which creates a small set of indefinite class regions to speed up the determination of object [1-3]. With respect to scale-invariant, the object idea method produces inaccurate Region of Interest (RoI). When RoI is given to CNN and if it is not accurate then it may be incorrectly detected in later stages and it affects the performance. Several procedures were proposed to handle this problem [4, 5]. Multiple object plan model is pooled to acquire accurate RoI. Anyway, it is ineffective to deal with the vast amount of RoI of high extent because localization is not accurately solved a regression method. Based upon a linear regression model, a new bounding box is proposed for every RoI [6, 7].

A sliding window-based region is used to make small adjustments to achieve the desired performance. Fine-tuning of the sliding window-based region proposal method using this regression model which maps the features of the last convolution layer to bound a box location [8-10]. Replacing the last layer of CNN with a half mean squared error loss for regression problems on several parts of CNN to describe an area of map refinement. These approaches require re-training and offline learning for the training when applied to the separate datasets [11-13].

Resolving the previously mentioned obstacle [14], this paper presents a novel method to improve the accuracy of numerical solutions that need to alter inaccurate novel methods of linear equations that are marked by exact position and sizes. As shown in Figure 1, extending the method of a linear equation to contextual reasoning, where the occlusion of the other parts of the object in the image or video is put into relative position with respect to the region proposal. To know precise position and relative extent, a unified model is developed with fast approximate which aims to alter defective field suggestions. When applied to different datasets i.e., concentrated data and information free is best suited with CNN architecture [15-17].

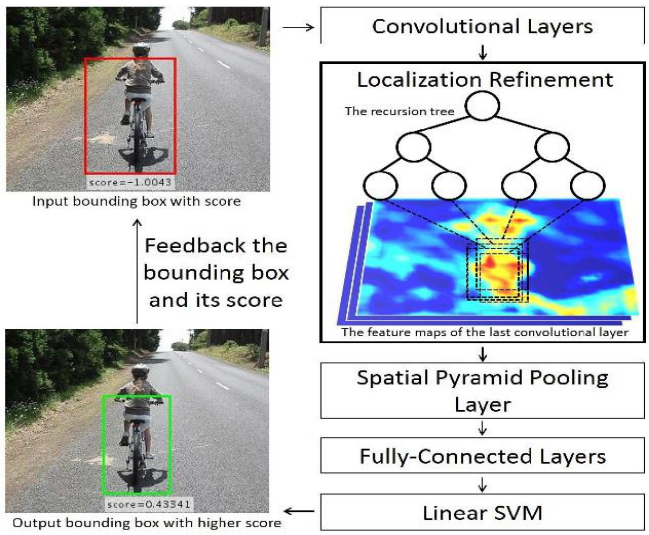

Figure 1. Iterative localization refinement in a CNN

This function verifies the proportion to some object occluded. The probability of all criteria depends on image data and uses the most effective to construct an output loss function. The final unified model merges iterative refinement and the joins score function. Finally, the structure shows the classification of the paper. Related work will be explained in section 2 and section 3 describes techniques. Similarly, the

outcomes will be presented in section 4. Section 5 describes the conclusion.

2. RELATED WORK

Object detection points to discover the instance of an object of specific classes in digital images or videos. This model detects what objects are present and where they situated from the given image or video. Here autonomous automobiles, video supervision, and several applications play an imperative role [18]. Recently CNN has improved object detection problems over the existing system. These algorithms are spilled into two divisions i.e., Region-based algorithm (R-CNN, Fast R-CNN, Faster R-CNN) and Regression-based algorithms (You Only Look Once (YOLO) and Single-Shot multibox Detector (SSD)) [19].

2.1 Region proposal based algorithms

R-CNN uses a selective search method to remove regions from the image is called a region proposal. These region proposals are generated by three steps. First, generates the region proposal by a selective search. Second feature extraction is done by a greedy algorithm while merging similar regions. Third, the classification of the image is completed by Support Vector Machine (SVM). The position with the selective search is that no learning is done, and it takes a long time in training to identify the regions [20, 21].

Fast R-CNN frames with an effort of R-CNN to increase the speed of training and testing time, and to improve the accuracy [22]. An image is shifted to CNN to create an activation map of convolution and then take out an RoI pooling layer to represent numeric or symbolic characteristics for each proposal. Then these characteristics are applied to the fully connected layer and finally, the model causes a SoftMax classifier to identify the bounding box for the detected objects. A problem consisting of this model takes a long time for preprocessing.

Solving the dispute of region, Fast R-CNN and Faster R-CNN were proposed, which creates region ideas by a neural network as an alternative which consists of two modules. Region Proposal Network (RPN), is happening which makes the second part as forwarding regions, the detector of Fast R-CNN to inspect. A small network moves smoothly around a surface RPN, slides over the convolution activation sketch with several anchors at a sliding window location. RPN results in bounding boxes and estimates class as Fast R-CNN. Four-step exchange training is given to share features for multiple parts [23].

Algorithm Sequences in Selective Search:

1. Generate initial sub-segmentation, we generate many candidate regions
2. Use a greedy algorithm to recursively combine similar regions into larger ones
3. Use the generated regions to produce the final candidate region proposals

2.2 Regression-based algorithms - YOLO and SSD

YOLO changes the regression detection problem. SSD based on the grid having existed in advance and takes a CNN to one after another without interrupting the estimation of confidence in classes and bounding boxes makes very fast compared with other approaches on region-based plans [24, 25].

The unified model finds the accurate locations and scales of an object by using an iterative refinement. This unified model constructs a bounding box consists of a single component and involving a pair of potentials. These potentials represent the confidence of the object of RoI based on the image data, whereas this model has property possessed by an array of things that have space in the image of RoI. The joint score function model works jointly to fine-tune the structured surrogate loss [26].

3. PROPOSED METHOD

First, we develop a unified model with fast approximate which aims to solve incorrect region proposals into their position and range. As our proposed method mainly focuses on data and we can the acquisition of knowledge of costless and it is well-suited with different CNN architecture when applied to distinctive datasets. Second, to the number of candidates in the images, we are using a joint score function. The joint score function deals with the relative position of the object occluded. The joint score function depends upon optimizing a structured output loss function and the proposition of image data. The concluding unified approach combines the joint score function and iterative refinement.

3.1 Localization refinement

Figure 2 given the model of Faster R-CNN. The aim of localization refinement is to define new bounding box Ri within the neighborhood of the initial box (Ri, f (Ri)). It is placed within two layers i.e. convolution and RoI pooling layer. By applying divide and conquer paradigm Faster R-CNN can find the bounding box and solve iterative refinement. The first step is over an RoI to choose the specific bounding box and it is continually divided into searches over smaller subregions.

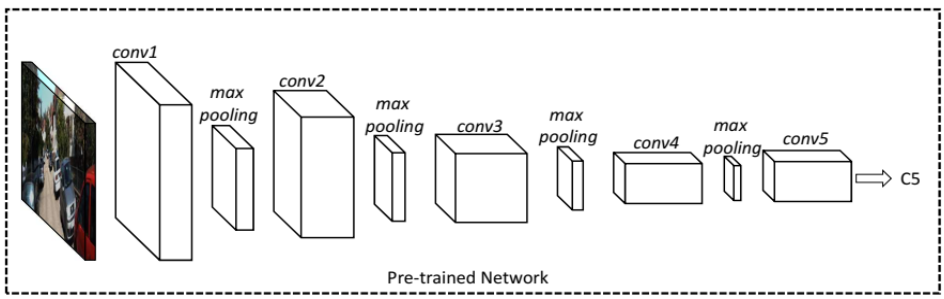

Figure 2. Localization refinement in a Faster R-CNN

Algorithm 1 Localization Refinement Method

Step 1: Define the search point Ri around ri
Step 2: Evaluate P(Ri).
Step 3: Evaluate f (Ri)
Step 4: Assign the new Sub Coordinates to the Ri, if P(Ri) >0
Step 5: Redo steps 1 to 4 until P(Ri) = 0 or if the iteration passes a threshold T
Step 6: Put the new coordinates Ri

For every iteration, Ri is the initial bounding box with score f (Ri) indicating how feasible the ith RoI consists in the region Ri. The region of the top, bottom, left and right correlate with the set of integers can give the bounding box Ri using the four-dimensional vector [Ti, Bi, Li, Ri]. Search region R of sub-regions split into 2 sub-regions X1 (Left Subtree) and X2 (Right subtree) with the smaller subregions we invoked the same search process for every level of the recurrence tree. When the region R (Leaf node) contains several bounding boxes the results of subproblem are added to attain the resulting probability P(Ri) in the bottom out recursion tree. Algorithm 1 showed the outline of this process.

The initial bounding box score is criticized through the RoI pooling and classifier. With the selective search method, it extracts the region proposals. Non-Maximum Suppression (NMS) is used to find those region's proposals with the highest rating as the initial bounding box. After that NMS new search regions are attained by crop overlapping bounding boxes as illustrated in Figure 3.

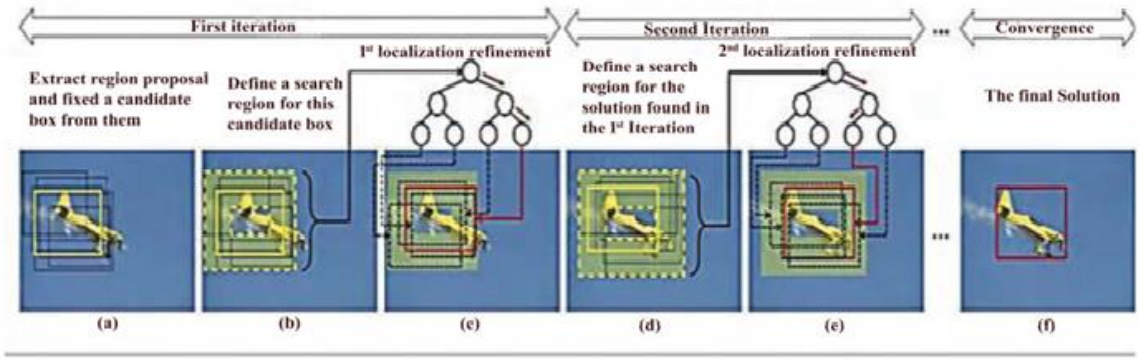

Figure 3. Yellow bounding box before sliding, red bounding box are fine-tuned

The region Ri in the recurrence tree of the leaf node. As inside the correlative region of its ascendants in the recurrence tree region of Ri, it takes the samples of Ri. From Ri, the root node is evaluated from the solution f (Ri) of the bounding box (Ri, f (Ri)).

$$f(Ri) \sim \max \{Ri| R` \in Ri`\} \quad (1)$$

Ri` denotes the set of regions extracted from Ri. Algorithm 1 is slightly modified into an approximate approach in which step 2 and step 3 can be determined by Eq. (1).

3.2 Joint score function

In this section, the joint score function [2] aims to joint multiple RoI. The score functions defined in the previous work [27, 28] the local object detector will define unary potential at a matching position between the candidates using spatial relations and pairwise potential models. These score functions are jointly using a stochastic gradient algorithm for training by reducing the surrogate loss. The cascade CNN was designed to detect the images into three stages. The calibration stage was brought in each of three detection stages and the output of this stage was used to adjust the detection window position for input to the subsequent stage. The main idea was to improve the accuracy by analyzing the existing C-CNN and to apply various optimization techniques such as augmenting data, modeling the optimization and its parameters, and adjusting drop-out. Join score function model was shown in Figure 4.

Consider that each bounding box has a binary variable Ri, there may be a collection of bounding box S extracted from an image, where i $\in$ S assigned to it. The background and RoI are given labels 1 and 0. To all fields of interest in the training images, there are ground-truth tables. Let ε denote pairs of regions in the bounding box, based on scales of the bounding box and relative locations. The edges of the regions are clustered and index denoted by (i,j) $\in$ ε by $k_{ij} \in \{1, \ldots\ldots, k\}$.

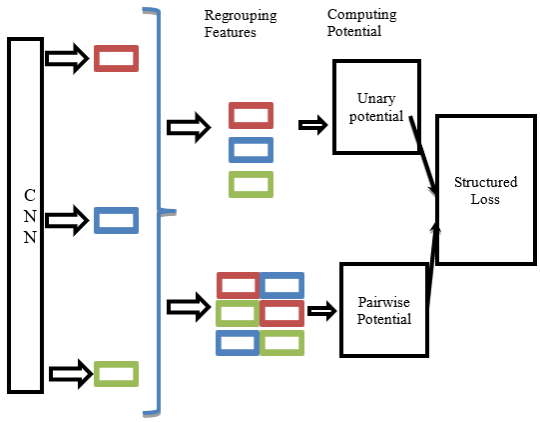

Figure 4. Model of the join score function

The labels of candidates in the same image are placed together by the joint score function S (yy; ωω), where ωω is the trainable parameter, yy is the vector of all binary variables whereas unary $\theta_i^U \theta_i^U$ and pairwise $\theta_{ij}^P \theta_{ij}^P$ potentials depend on it.

$$S(y,w) = \sum_{i \in S} y_i \theta_i^U(\omega) + \sum_{(ij) \in \varepsilon} y_i y_j \theta_{ij,k_{ij}}^P(\omega) \quad (2)$$

The image with potentials is connected using score function Eq. 2 by a feed-forward neural network. The two additional feed-forward networks: unary and pairwise network is used to map the feature vector of RoI. To RoI bounding box and pairwise potential maps are the concatenated pairs of the feature vector of its multiple RoI and it is correlated to unary potential. Using extractor in the joint function, it represents the localization refinement method from the above process leading to 2048 features. Finally, one fully connected layer is taken to output score and bounding box regression. It takes the region proposal as 16 RoI for a single image, but it did not improve the performance in the validation set.

3.3 Structured SVM (SSVM)

In this section, the surrogate loss is minimized. The SSVM is taken as an objective to handle the surrogate loss for the prediction of the RoI. Since the SSVM does not give positive outcomes in the precision-recall measure and it is less suited for the detection. Here, the surrogate loss can be optimized by the model parameters that are minimized by using structured surrogate loss using a Stochastic Gradient Descent (SGD) algorithm.

The following steps are done in the SGD algorithm.

Step 1: On that score, apply NMS on top of the score.

Step 2: Select a set of fields of interest produced by faster R-CNN.

Step 3: By using perform feed-forward pass calculate potentials of the joint score function.

Step 4: To calculate gradients and structured loss.

Step 5: Perform suggestions throughout the backpropagate model of the gradient.

To optimize the surrogate loss, gradients are computed according to the model guidelines. Through backpropagation, optimize the surrogate loss. By backpropagate the gradients over the model parameters, the scores of the bounding box can be calculated accurately. Compute the loss with respect to unary (consisting of one element) and the Pairwise model using equation 3. Feature extractor is given in equation 4 with the help of the backpropagation procedure.

$$\frac{dl}{d\omega^U} = \sum_{i \in S} \frac{dl}{d\theta_i^U} \frac{d\theta_i^U}{d\omega^{U'}} \quad \frac{dl}{d\omega^P} = \sum_{(i,j) \in k} \sum_{k=1}^{N} \frac{dl}{d\theta_{ij,k}^P} \frac{d\theta_{ij,k}^P}{d\omega^P} \quad (3)$$

$$\frac{dl}{df_i} = \frac{dl}{d\theta_i^U} \frac{d\theta_i^U}{df_i} + \sum_{j:(i,j) \in \varepsilon} \frac{dl}{d\theta_{ij,k_{ij}}^p} \frac{d\theta_{ij,k_{ij}}^P}{df_i} + \sum_{j:(i,j) \in \varepsilon} \frac{dl}{d\theta_{ji,k_{ji}}^p} \frac{d\theta_{ji,k_{ji}}^P}{df_i} \quad (4)$$

4. EXPERIMENTAL RESULTS

Object detection in terms of the mean Average Precision (mAP) to process iterative refinement that considers the PASCAL VOC 2007 dataset. The dataset is split into 5010 training and 4950 validation as endorsed and consist of more than 80 thousand of exploratory objects in the bottleneck. Iterative refinement within Faster R-CNN produces 248 vectors as the result by taking the entire image as the input. With the help of the recurrence tree, filter each output vector. The ground truth bounding box is given as result 0.3 to perform a maximum number of iterations Intersection over Union (IoU) ratio. Reduce the sum of log losses by the SGD algorithm with weight decay 0.0005, momentum 0.9 and learning rate 0.00001. The weight of the unary and pairwise potential was initialized by Gaussians with a standard deviation of 0.01, weight decay 0.00005, momentum 0.9 and learning rate 0.00001. Optimize the objective of the structured surrogate objective and SGD using mAP scope to calculate the detection performance depends on the precision-recall curve. While detection with a high ratio (IoU> 0.3) then Ground truth, value is positive. Assign the same ground truth for multiple detections are considered as faulty, positives.

Table 1. Detection results on PASCAL VOC 2007 dataset

METHOD	INPUT SIZE	FPS	mAP
R-CNN	~ 1000X600	0.84	54.2
Fast R-CNN	~ 1000X600	1.12	55.2
Faster R-CNN	~ 1000X600	2.4	59.1
Faster R-CNN +IR	~ 1000X600	2.8	62.7

Figure 5. Detection result of the proposed methodology (green bounding boxes)

Table 1 The iterative refinement with R-CNN and SPP-net compared by Faster R-CNN. The refinement step is to upgrade

the accuracy of the regions. Observes that IR in Faster R-CNN outperforms R-CNN and SPP-net by 3.6 mAP. The IR in Faster R-CNN achieves greater mAP of 61.2 that R-CNN BB and SPP-net BB with bounding box regression. The bounding box is adjusted to their correct size and position with the help of iterative refinement. If the number of true positive increases, then it improves performance. Based on the IR Faster R-CNN it shows detection results in Figure 5.

5. CONCLUSION

Here, we introduce a novel scheme to improve localization accuracy and report a gain of 3.6mAP. The to regulate the inaccurate bounding boxes gained in the existing object proposal approach we developed the Iterative Refinement approach, generally used in Convolution Neural Network-based object detection. The aim of the joint score is to join the multiple regions of interest by applying the score functions. The surrogate loss is minimized for the prediction of the region of interest. The laboratory results show that mean Average Precision increased with respect to the region proposal feature vector. Another possible future action is to consider motion data to perform long term tracking.

REFERENCES

[1] Cheng, K.W., Chen, Y.T., Fang, W.H. (2016). Iterative localization refinement in convolutional neural networks for improved object detection poster. IEEE International Conference on Image Processing (ICIP), pp. 3643-3647. https://doi.org/10.1109/ICIP.2016.7533039

[2] Vu, T.H., Osokin, A., Laptev, I. (2015). Context-aware CNNs for person head detection. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2893-2901. https://doi.org/10.1109/ICCV.2015.331

[3] Salloum, R., Ren, Y., Kuo, C.C.J. (2018). Image splicing localization using a multi-task fully convolutional network (MFCN). Journal of Visual Communication and Image Representation, 51: 201-209. https://doi.org/10.1016/j.jvcir.2018.01.010

[4] Zhang, J., Liu, P., Zhang, F., Song, Q. (2018). CloudNet: ground-based cloud classification with deep convolutional neural network. Geophysical Research Letters, 45(16): 8665-8672. https://doi.org/10.1029/2018GL077787

[5] Li, J., Si, Y., Lang, L., Liu, L., Xu, T. (2018). A spatial pyramid pooling-based deep convolutional neural network for the classification of electrocardiogram beats. Applied Sciences, 8(9): 1590. https://doi.org/10.3390/app8091590

[6] Zhang, J., Malmberg, F., Sclaroff, S. (2019). Unconstrained salient object detection. In Visual Saliency: From Pixel-Level to Object-Level Analysis, pp. 95-11. https://doi.org/10.1007/978-3-030-04831-0_6

[7] Latchoumi, T.P., Ezhilarasi, T.P., Balamurugan, K. (2019). Bio-inspired weighed quantum particle swarm optimization and smooth support vector machine ensembles for identification of abnormalities in medical data. SN Applied Sciences, 1(10): 1137. https://doi.org/10.1007/s42452-019-1179-8

[8] Abbas, S.M., Singh, S.N. (2018). Region-based object detection and classification using faster R-CNN. International Conference on Computational Intelligence & Communication Technology (CICT), IEEE, pp. 1-6. https://doi.org/10.1109/CIACT.2018.8480413

[9] Maisano, R., Tomaselli, V., Capra, A., Longo, F., Puliafito, A. (2018). Reducing complexity of 3D indoor object detection. International Forum on Research and Technology for Society and Industry (RTSI), IEEE, pp. 1-6. https://doi.org/10.1109/RTSI.2018.8548514

[10] Ranjeeth, S., Latchoumi, T.P., Victer Paul, P. (2019). Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. Recent Advances in Computer Science and Communications. https://doi.org/10.2174/2666255813666191116150319

[11] Dou, J., Qin, Q., Tu, Z. (2019). Background subtraction based on deep convolutional neural networks features. Multimedia Tools and Applications, 78(11): 14549-14571. https://doi.org/10.1007/s11042-018-6854-z

[12] Cheng, G., Han, J., Zhou, P., Xu, D. (2018). Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing, 28(1): 265-278. https://doi.org/10.1109/TIP.2018.2867198.

[13] Han, Y., Zhang, P., Zhuo, T., Huang, W., Zhang, Y. (2018). Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recognition Letters, 107: 83-90. https://doi.org/10.1016/j.patrec.2017.08.015

[14] Nguyen, V.D., Tran, D.T., Byun, J.Y., Jeon, J.W. (2018). Real-time vehicle detection using an effective region proposal-based depth and 3-channel pattern. IEEE Transactions on Intelligent Transportation Systems, 20(10): 3634-3646. https://doi.org/10.1109/TITS.2018.2877200

[15] Du, P., Zhang, H., Ma, H. (2019). Classifier refinement for weakly supervised object detection with class-specific activation map. In 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3367-3371. https://doi.org/10.1109/ICIP.2019.8803672

[16] Vijayvasanth, A., Venkatachalapathy, K., Latchoumi, T.P., Latha, P., Ezhilarasi, T.P. (2017). A survey on cache route schemes to improve OoS in ADHOC networks. Pakistan Journal of Biotechnology, 14(11): 265-269.

[17] Liu, W., Liao, S., Hu, W. (2019). Towards accurate tiny vehicle detection in complex scenes. Neurocomputing, 347: 24-33. https://doi.org/10.1016/j.neucom.2019.03.004

[18] Yaseen, M.U., Anjum, A., Rana, O., Hill, R. (2018). Cloud-based scalable object detection and classification in video streams. Future Generation Computer Systems, 80: 286-298. https://doi.org/10.1016/j.future.2017.02.003

[19] Dai, H., Lin, M., Jiang, W. (2018). Object detection based on visual memory: A feature learning and feature imagination process. Enterprise Information Systems, 1-17. https://doi.org/10.1080/17517575.2018.1539775

[20] Liu, Q., Li, Z., Sun, F., Tian, Y., Zeng, W. (2018). Image recognition and classification by deep belief-convolutional neural networks. Journal of Tsinghua University (Science and Technology), 58(9): 781-787. https://doi.org/10.16511/j.cnki.qhdxxb.2018.22.034

[21] Jmour, N., Zayen, S., Abdelkrim, A. (2018). Convolutional neural networks for image classification.

International Conference on Advanced Systems and Electric Technologies (IC_ASET), IEEE, pp. 397-402. https://doi.org/10.1109/ASET.2018.8379889
[22] Mao, M., Zhang, H., Li, S., Zhang, B. (2019). SEMANTIC-RTAB-MAP (SRM): A semantic SLAM system with CNNs on depth images. Mathematical Foundations of Computing, 2(1): 29-41. https://doi.org/10.3934/mfc.2019003
[23] Latchoumi, T.P., Jayakumar, L., Latha, P., Janakiraman, S. (2016). OFS method for selecting active features using clustering techniques. In Proceedings of the International Conference on Informatics and Analytics, ACM. https://doi.org/10.1145/2980258.2982108
[24] Shaoqing, R., Kaiming, H., Ross, G., Jian, S. (2018). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, 91-99.
[25] Shyam, D., Kot, A., Athalye, C. (2018). Abandoned object detection using pixel-based finite state machine and single shot multibox detector. IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. https://doi.org/10.1109/ICME.2018.8486464
[26] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2018). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788.
[27] Zhu, A., Wang, T., Snoussi, H. (2018). Hierarchical graphical-based human pose estimation via a local multi-resolution convolutional neural network. Aip Advances, 8(3): 035215. https://doi.org/10.1063/1.5024463
[28] Everingham, M., Van, G.L., Williams, C. K. I., Winn, J., Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

NOMENCLATURE

FPS	Frames per second
mAP	Activation map

Greek symbols

y	Vector for all binary variables
ω	Trainable parameter
ϕ	solid volume fraction
Θ	dimensionless temperature
μ	dynamic viscosity, kg. $m^{-1}.s^{-1}$