International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 18 (2018) pp. 13945-13949

Object Recognition in Dynamic Environments with Convolutional Neural Networks

Robinson Jiménez-Moreno[1], Oscar Avilés Sánchez[2], Diana Marcela Ovalle[3]

[1]*Professor, Department of Mechatronics Engineering, Nueva Granada Military University, Bogotá, Colombia.*

[2]*Professor, Department of Mechatronics Engineering, Nueva Granada Military University, Bogotá, Colombia.*

[3]*Professor, Department of Electronics Engineering, Distrital Frco. Jose de Caldas University, Bogotá, Colombia.*

Abstract

Deep learning techniques have revolutionized pattern recognition systems over the last decade. Within these techniques, convolutional neural networks (CNN) have demonstrated a high performance in the recognition of objects in images. This article presents the evaluation of a CNN with 93% accuracy in the recognition of objects of four categories, trained with images at a fixed distance and the variations of said performance when evaluating said network with images of the trained categories, with variations of distance, evidencing a degradation of said performance by 30%. This makes it possible to establish the need for a structured convolutional neural architecture that, in function of the advantages of training of fixed distance by parallel layers, reinforces learning to accuracy values higher than 90%, whose incidence is demarcated in robotic mobile applications where the camera approaches or distances from the object.

Keywords. Machine vision, convolutional neural network, object recognition, MATLAB, Image RGB-D.

INTRODUCTION

Machine vision algorithms were usually developed in three basic stages, the one of pre-processing of image for adequacy of the same in size, elimination of noise and others, stage of image processing oriented to extraction of characteristics and a stage of recognition of patterns or classification, the latter usually developed through neural networks. With the development of different deep learning techniques [1], and within these, of convolutional neural networks (CNN) [2], pattern recognition systems have become more versatile. A clear example of this, CNNs are now widely used in pattern recognition [3] and image identification applications [4], with very high accuracy ranges above 90% in most cases, and with the ability to integrate the traditional steps described at the beginning.

Although they are still developing techniques [5], CNN offer solutions in different areas [6], such as artificial intelligence, robotics, medicine, intelligent control systems and the development of intelligent city schemes such as traffic control applications, etc. So that, as they are applied in different environments, opportunities for developing these techniques are found, as a function of the capabilities they offer.

In the field of robotics, several developments have been presented that allow the interaction of a robot with its medium, for it is necessary to capture and interpret it, which is achieved by image capture sensors such as cameras and RGB-D sensors, i.e., that apart from the image, they give the depth to which the objects in the captured scene are. CCNs have already been involved in such applications [7] [8] at the level of conventional RGB and depth [9-11]. However, mobile robotics applications are still under development and, as mentioned in [12] [13], the combination of information of RGB-D and CNN presents an important nucleus for the development of applications of this type.

When a robot moves, the distance from the camera to the object varies, which presents a dynamic learning characteristic of the object. It is clear that an object presents different perspectives depending on the distance in which it is observed, as the focus of vision moves away from an object, specific features of the object are lost and the outline is the most relevant information, as it approaches the object specific details appear. This can present a challenge in the performance of a CNN when identifying an object, when it is used from a mobile robot that approaches or distances an object of interest. In the state of the art this analysis is not found, for which it is addressed in the present article.

This article is organized as follows. In section 2 the CNN architecture to be used is presented, the generalities of the training of this type of networks and the results of the training for the recognition of objects at a fixed distance. In section 3 it is presented the analysis of the trained network with distance variations in the test database and a new training including this new database. In section 4 the conclusions of the evaluation of the network in dynamic atmospheres of image capture input for prediction of CNN are presented.

Convolutional neuronal network architecture

Robotic agents handle different path planning schemes to achieve a specific goal. For this task it is fundamental to recognize the environment in which it moves and the objective that seeks, typically this objective is an object, as is the case of

a robotic arm that performs care tasks, where it looks for a tool, a food or approaching the hand of a user to interact. Object recognition for robotic applications can be developed using pattern recognition tools such as CNN. Therefore, in order to validate the performance of these networks in the recognition of objects with distance variations when robotic agent moves, as it would be presented in a care robot, it is established as methodology a case study that allows to determine a network architecture that lets to perform the recognition of objects at a fixed distance, then this architecture is evaluated with distance variations of the trained objects and a solution is proposed facing the degradation in object recognition.

To evaluate the performance of a CNN in the recognition of objects in environments where there is a variation of the distance of the object to the point of capture of the image, a neuronal architecture is initially designed with a database whose capture distance is fixed, later this architecture is subjected to evaluation with images of the same objects already learned, but taken closer and farther than the training ones.

In Fig. 1, part of the database is observed, where 100 images of each of the four categories of the objects to be recognized in the training stage and 50 in the validation stage are used. The images are taken at a distance of 60 cm for the case.

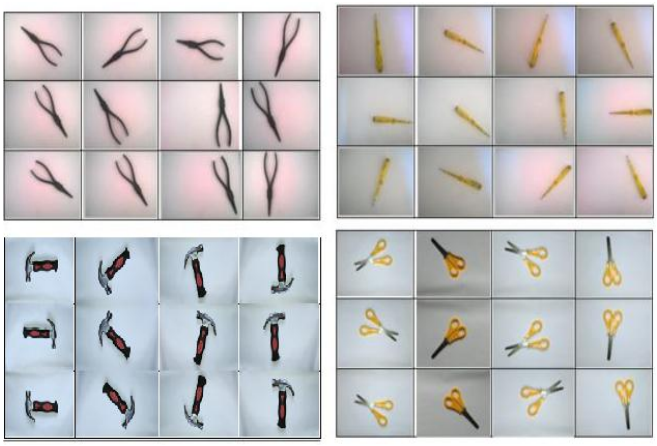

Figure 1. Image samples of the Database.

The architecture of a convolutional neural network is framed in two basic stages, one of learning features (FL) and another of features classification (FC), as illustrated in Fig. 2. The depth of the network is determined by convolution (conv), linear rectification (ReLU) and pooling (pool) combinations, in step FL. The CNN design must determine how many of these layers to use and how to relate them, so that the architecture implemented obeys the structure shown in Table I.

The convolution operation is characterized by equation 1, while the pooling method uses the maximum method, calculated by equation 2 and the ReLU layer simply eliminates negative values.

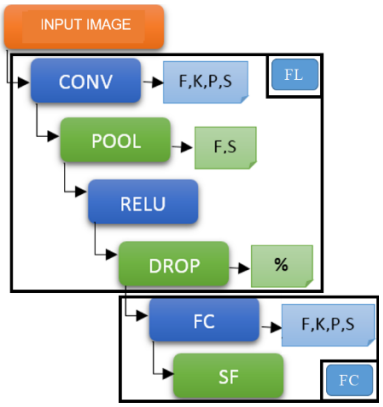

Figure 2. Basic structure of a convolutional neural network.

Table I. CNN Architecture used.

Type	Kernel		Filters
Convolution/ReLU	6x6	S=1	30
MaxPooling	3x3	S=1	
Convolution/ReLU	4x4	S=1	20
MaxPooling	2x2	S=1	
Convolution/ReLU	3x3	S=1	15
MaxPooling	2x2	S=2	
Full-Connected	4		
Softmax	4		

$$h_j^n = \max(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n) \quad (1)$$

$$h_j^n(x,y) = max_{\bar{x}\in N(x),\bar{y}\in N(y)} h_j^{n-1}(\bar{x},\bar{y}) \quad (2)$$

When the volume of the input image, determined by a width (W), height (H) and a depth (D), is operated with a convolutional layer, change the output volume of these layers. Depending on the dimensions of the input volume to the mentioned layers and the existence or not of padding (P), which is a lateral filling of the volume considered in the input, the dimensions of the output volume are calculated by equations 3 and 4. The parameter S corresponds to the displacement of the convolution filter with size F, in the input volume.

$$W_{n+1} = \frac{W_n - F + 2P}{S} + 1 \quad (3)$$

$$H_{n+1} = \frac{H_n - F + 2P}{S} + 1 \quad (4)$$

Fig. 3 illustrates the performance of the network in the characteristic learning stage, where it can be seen that at least 130 iterations are required to reach 100% accuracy, with a

general network performance of 93.8%, which clearly identifies each of the four learning objects, for the case nippers, screwdriver, hammer and scissors. Where the performance is slightly compromised by similarity between the nippers and the scissors, these two classes being the only ones that present confusion between them.

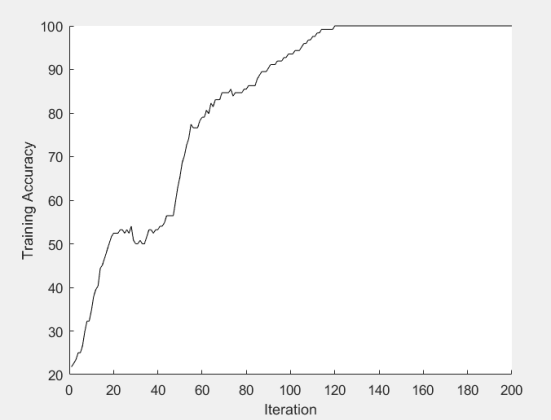

Figure 3. CNN Training accuracy.

Analysis and results

To evaluate the performance of the convolutional neural network with the proposed objective, a database of the selected categories is set at different distances, according to the resolution of the depth capture camera. Depth information for mobile robotic agents can be taken by sensors such as the RGBD Blaster Senz3D Creative camera, whose 3D vision range is from 0.2 to 1.5 meters [14]. Derived from this, the distances 20, 40, 60, 80 and 100 cm are taken to form the database.

In Fig. 4 it can be seen a sample of the database, which for the case corresponds to 25 images of each category. This allows to validate the network, being presented and labeled one by one the images after the prediction of the network. Table II presents the results of this evaluation.

Table II. Depth identification results.

Depth	Nippers	Screwdriver	Hammer	Scissors
20 cm	40.4%	51,7%	48,4%	33,1%
40 cm	61.6%	74,7%	65,8%	63,1%
60 cm	91.6%	94,7%	95,8%	93,1%
80 cm	51.5%	84,7%	85,1%	61,6%
100 cm	46.3%	54,7%	65,1%	31,6%

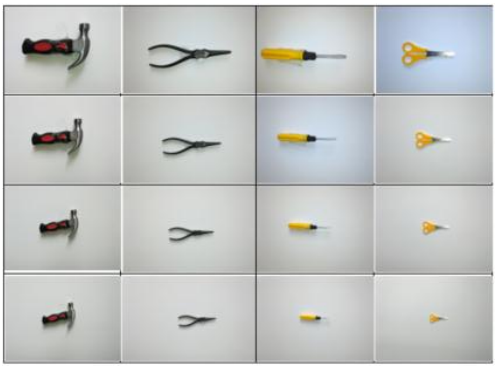

Figure 4. Database for depth.

It can be seen that as the camera approaches or moves away from the object, the accuracy of the prediction decreases, being slightly more harmful to approach the object than to move it away. This is due to the convergence of characteristics, so that when the object is moved away, the general shape of the object tends to be maintained, whereas the approach exhibits specific characteristics of the same, which have not been learned by the network.

An obvious solution would seem to be to expand the database of the initially trained network, including the images of the objects at the different distances evaluated. Under this scheme, Fig. 5 illustrates the performance of the network based on the architecture of Table I, with the joint database.

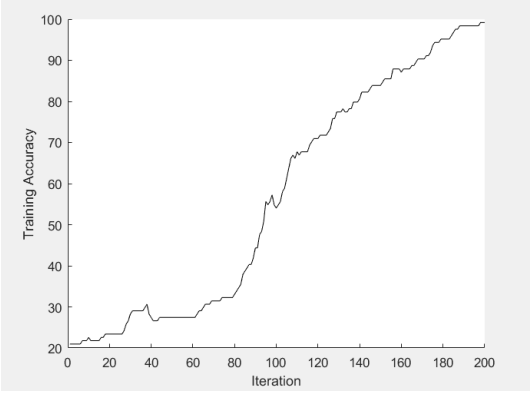

Figure 5. CNN2 Training accuracy.

It is observed that the network needs much more iterations to learn the features, while the performance drops to 63.71%. This leads to a general degradation in the identification of objects under this aspect, which was to be expected, because the learning characteristics now diverge with the new images. The Fig. 6 illustrates part of the identification results, where the confusion of classes is evidenced.

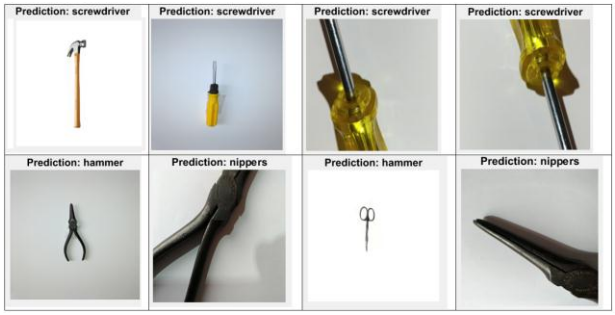

Figure 6. Misclassification in depth.

This allows to conclude that the use of convolutional neural networks to identify objects in dynamic environments, such as mobile robotic navigation, the selection of tools by robotic arms with local cameras, in the gripper for example, requires the development of more complex neuro-convolutional architectures, as could be proposed in Fig. 7.

In this architecture the learning of characteristics by distances is proposed, where each network in parallel is trained with a database conformed by the different categories according to the 5 set distances, obtaining an overall performance of 90.23%.

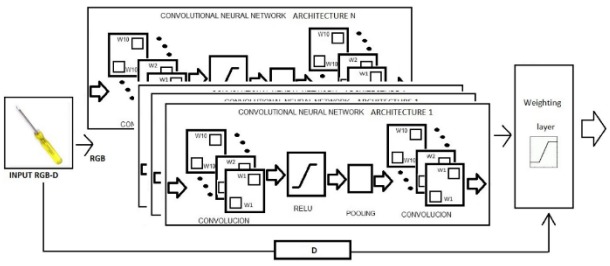

Figure 7. Suggested neuronal structure for learning in depth.

CONCLUSIONS

It was possible to demonstrate the high performance provided by convolutional neural networks for the identification of objects, for the case in the learning of four categories of tools.

It was evidenced that by varying the capture distance of an image and evaluating this by a CNN, there is degradation in the prediction performance of the network. This has its origin in the learned patterns that are demarcated by the training database, which, although it is the same object, to move it away or to bring it closer, exhibits different characteristics of the same.

The inclusion of a more complete database, respecting to the view of different depths from the camera, does not generate a satisfactory solution, since, although the network is able to train with a higher computational cost, the performance of the prediction decreases. However, it can be inferred that parallel CNN architectures, each trained with databases at a given depth may be a viable option, given the high performance achieved by a CNN, for this type of learning.

REFERENCES

[1] Rodríguez, M., Sandobalin, S., Pozo, D., Morales, L., Rosero J., and Rosales, A. 2014, “Mapeo de Laberintos y Búsqueda de Rutas Cortas MedianteTres Mini Robots Cooperativos,” Politécnica, 34(2), pp. 101-106.

[2] J. Schmidhuber, “Deep learning in neural networks: An overview”, Neural Networks, Volume 61, January 2015, pp. 85–117.

[3] Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham.

[4] Walid, R.; Lasfar, A, "Handwritten digit recognition using sparse deep architectures," Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on, vol., no., pp.1, 6, 7-8 May 2014. doi: 10.1109/SITA.2014.6847284.

[5] Krizhevsky A, Sutskever I, Hinton G. (2012) ImageNet classification with deep convolutional neural networks. University of Toronto. Conference on Advances in neural information processing systems, 2012, pp 1097-1105.

[6] Ming Liang, Xiaolin Hu. (2015) Recurrent convolutional neural network for object recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Institute of Electrical and Electronics Engineers (IEEE); 2015 Jun; Available from: http://dx.doi.org/10.1109/cvpr.2015.7298958.

[7] Z. Fadlullah; F. Tang; B. Mao; N. Kato; O. Akashi; T. Inoue; K. Mizutani, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow’s Intelligent Network Traffic Control Systems," in IEEE Communications Surveys & Tutorials , May-2017, no.99, pp.1-1.

[8] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, 2015, pp. 1316-1322. doi: 10.1109/ICRA.2015.7139361.

[9] Wang Z, Li Z, Wang B, Liu H. (2016) Robot grasp detection using multimodal deep convolutional neural networks. Advances in Mechanical Engineering [Internet]. SAGE Publications; 2016 Sep 23;8(9). Available from: http://dx.doi.org/10.1177/1687814016668077.

[10] Maturana D, Scherer S. (2015) VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [Internet]. Institute of Electrical and Electronics Engineers (IEEE); 2015 Sep; Available from: http://dx.doi.org/10.1109/iros.2015.7353481.

[11] S. Yang, D. Maturana, and S. Scherer, “Real-time 3D scene layout from a single image using convolutional neural networks,” in Proc. Int. Conf. Robot. Autom., 2016, pp. 2183–2189.

[12] H. Schulz, N. Hoft, and S. Behnke, “Depth and height aware semantic RGB-D perception with convolutional neural networks,” in Proc. Eur. Symp. Artif. Neural Netw., 2015, pp. 463–468.

[13] L. Porzi, S. R. Buló, A. Penate-Sanchez, E. Ricci and F. Moreno-Noguer, "Learning Depth-Aware Deep Representations for Robotic Perception," in IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 468-475, April 2017.

[14] Creative Technology Ltd, Creative Senz3D. Consulted on September 5, 2018, [Online]. Available in: https://us.creative.com/p/web cameras/creative-senz3d