NAO-Read: Empowering the Humanoid Robot NAO to Recognize Texts in Objects in Natural Scenes

Diego Alves da Silva, Aline Geovanna Soares, Antonio Lundgren, Estanislau Lima and Byron Leite Dantas Bezerra
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Email: $\{das, ags4, aval, ebl2, byronleite\}$ @$ecomp.poli.br$

***Abstract*—Robotics is a field of research that has undergone several changes in recent years. Currently, robot applications are commonly used for many applications, such as pump deactivation, mobile robotic manipulation, etc. However, most robots today are programmed to follow a predefined path. This is sufficient when the robot is working in a settled environment. Nonetheless, for many tasks, autonomous robots are needed. In this way, NAO humanoid robots constitute the new active research platform within the robotics community. In this article, we present a vision system that connects to the NAO robot, allowing robots to detect and recognize the visible text present in objects in images of natural scenes and use that knowledge to interpret the content of a given scene. The proposed vision system is based on deep learning methods and was designed to be used by NAO robots and consists of five stages: 1) capturing the image; 2) after capturing the image, the YOLOv3 algorithm is used for object detection and classification; 3) selection of the objects of interest; 4) text detection and recognition stage, based on the OctShuffleMLT approach; and 5) synthesis of the text. The choice of these models was due to the better results obtained in the COCO databases, in the list of objects, and in the ICDAR 2015, in the text list, these bases are very similar to those found with the NAO robot. Experimental results show that the rate of detecting and recognizing text from the images obtained through the NAO robot camera in the wild are similar to those presented in models pre-trained with natural scenes databases.**

I. Introduction

Undoubtedly, Deep Learning (DL) methods, especially convolutional neural networks (CNNs) [1], in the last decade has led to major advances in several research fields such as artificial intelligence, pattern recognition, computer vision, and natural language processing. The application of these new approaches has been particularly successful by computer vision community has also embraced these DL methods for various tasks such as scene text recognition, object recognition, character recognition, layout analysis, object detection, and so on. This new paradigm has already attracted the attention of the robot vision community [2] [3].

Despite the great success of these DL methods to solve many robot vision applications, applying DL methods for many robotic vision applications is a challenging task. Therefore, DL methods are usually developed within the computer vision community and then transferred to the robot vision community. However, they cannot always take into account the robotic vision applications, with their specific requirements. For instance, DL methods usually need very large datasets, and architectures with millions of parameters, and neurons. Furthermore, DL often require a lot of computational power, which makes them very challenging for devices that have hardware limitations, such as robots.

Although, robotics is a research field that has trialed several changes in the last years. Currently, robot applications are commonly used for many applications, such as bombs deactivation, mobile robotic manipulation, and so on. In other words, for many tasks, autonomous robots are needed. In this way, the NAO humanoid robots constitute a new active research platform within the robotics community. NAO robots is a social humanoid robot with approximately 60 cm in height and weight of 5 kg. The device is capable of standing, walking, dancing, and speaking. In addition, it has cameras, microphones, speakers, and various sensors such as tactile sensors, pressure, and sonar, that enabling communication with people and interaction with the environment.

In this context, we propose a vision system integrated with a NAO robot to detect and recognize visible text in natural scene images and to use this knowledge to interpret the content of a given scene. Concretely, the proposed vision system is based on DL methods and is designed to be used by NAO robots. Our proposed system is composed of five stages: the first stage is the image capture that is done by the NAO's onboard camera. In the second stage, the objects are detected and classified by YOLOv3 model [4]. YOLOv3 was used instead of others deep models, e.g., (R-CNN [5]; Fast R-CNN [6]; Faster R-CNN [7]; CRNN [8]), due to the low computational cost and accuracy of this model in comparison with other DL approaches. Beyond that, YOLOv3 has been extensively used in many robotic vision applications to object detection and classification.

In the third stage, the user select the objects of interest based on the options detected by YOLOv3. Then, the fourth stage takes as input the objects of interest selection and their coordinates. Then, this information is used to text detection and recognition, based on the OctShuffleNet model [9]. This model was used since it has fewer layers and parameters in comparison with others, which makes this model suitable for devices that have hardware limitations such as robots. Also, this model has outperformed state-of-the-art models in the scene text recognition task.

The final stage is the text synthesis, allowing the NAO robot

to speak the text recognized in the previous stage to the end-user. Additionally, NAO robot was used instead of other robot platforms, because it is increasingly been used in research exploring human-humanoid interaction and its applications, such as investigating non-verbal cues [10], interaction therapy [11], and assisted-living [12].

Experimental results provided in this paper show that our proposed vision system is able to detect and recognize text in natural scenes with similar detection and recognition rates of the models pre-trained with natural scenes databases such as ICDAR 2015 and COCO-Text. Additionally, the system is capable of handle images with complex backgrounds, with different light environments and partial occlusion.

II. Related Works

DL methods, in particular, deep CNNs currently dominate the field of computer vision in many tasks such as image recognition [1], object detection [6], and in many other computer vision task for instance scene text recognition [9], with high performances to a new level comparing to the traditional methods.

In the field of robotics, CNNs have achieved the highest accuracy on detection and recognition problems. The NAO robot is one of the most used in the field of robotics, largely due to its design and its friendly way of interaction. It is used in various tasks, such as in a study by Tapus et al. (2012), an experiment with four children was carried out, in which the interaction with NAO and humans were compared [13]. Two of these children did not show any effect on the presence of the robot. However, the other two paid more attention to NAO than humans, and one of them showed a great interaction with the robot. Other recent and interesting applications of NAO are: robot football [2], emotion recognition [14] and recognition of human posture and imitation of gestures [15].

Despite being widely used, the NAO robot has a low computational capacity, making the task of implementing computer vision models more challenging. Recognizing and classifying texts in objects by observing images in real-time is one of the hardest tasks to be performed in the field of computer vision due, the backgrounds of these images are clustered, and the text has different styles, such as fonts, languages, sizes, contrasts, lighting conditions, among others.

To the best of our knowledge, this paper presents the first work to use the deep CNNs method with a NAO robot to detect and recognize text in natural scene images.

III. Proposed System

In this part, we describe our proposed system, which follows the steps as depicted in Figure 1. The proposed pipeline involves the following five stages:

A. Image capture by the NAO's onboard camera

The first stage consists of capturing the image. The task was performed using the NAO's camera. NAO operates on his own specialized Linux distribution, called NAOqi accompanied by a software package, including a graphical programming tool called Choregraphe, and an SDK (Software Development Kit). This SDK offers high-level programming functions to control the NAO robot.

Through NAOqi, the program that governs our proposal was made. It consists of a library called ALProxy that guarantees access to the robot's functionalities. Through the ALProxy lib, we can register for an event, to have access to some functions of the robot. For instance, passing as parameter 'ALVideoDevice', we can access the robot's camera, from there we capture and save the image on the computer, with dimensions of 640x480. The saved image will be used as input for the next stage.

B. Objects detection and classification

In the second stage is performed the object detection and classification. For this stage, the YOLOv3 algorithm [4] was used with the pre-trained weights available. YOLO is a new approach to object detection, which frames object detection as a regression problem to spatially separate bounding boxes and associate a class probability for every object detected in scene. A single neural network predicts bounding boxes and class probabilities directly from a full input image in one forward propagation over this neural network.

YOLO was used instead of other any available detection methods because, compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on the background, with the best inference time. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including SSD [16] and R-CNN [5], when generalizing from natural images to other domains like artwork. Since YOLO achieves faster inference times, it is a good suit for real-time object detection, and maybe, for this reason, it has been intensively used in robotics applications.

C. Objects of interest selection

The main goal of this stage is provide to the user a interface to ask his/her object of interest. After the survey of the objects in the previous step, which returns the identified objects and their coordinates, the system presents the interface of all detected objects listed. Them, the user must enter the index of the object that he wants to identify the textual content. Thus, the algorithm cuts the object of interest and saves the image that will be used as input for the next step. In a new version of this interface, the user will interact with NAO by voice commands.

D. Text detection and recognition

The fourth stage is the text detection and recognition. For this stage, the OctShuffleMLT [9] was used to perform the text recognition. The OctShuffleMLT model is a compact network with very fewer parameters and layers. The model is characterized by presenting a completely convolutional architecture, which approaches the text on the scene in a multilingual way as a base, and adapts the extremely light architecture of ShuffleNet, designed for mobile devices. Also,

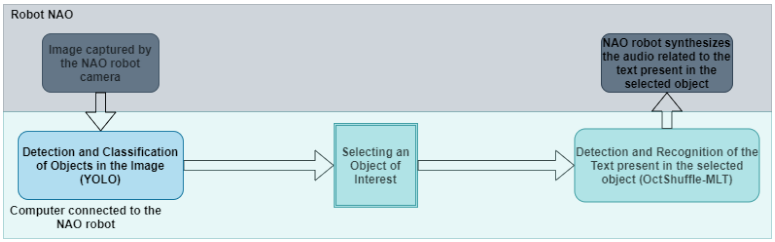

Fig. 1: Overview of the process proposed by this work

Octave Convolutions are used to decrease the computational cost and improve the accuracy of the models.

OctShuffleMLT first extracts the low-level features of the input image, which are relevant for recognition and detection. In the detection, CNN blocks are used to extract these features and output the boundary box predictions of text regions. In the OCR stage, a completely convolutional neural network was used to output the multilingual label sequence predictions. Once the text has been detected and recognized by the template, the content is passed to the next stage of our proposed pipeline. A pre-trained OctShuffleMLT available is used, avoiding a hard training task.

E. Text synthesis

Finally, the text returned from the previous stage is processed. The goal of this final stage is to transform the recognized text in the corresponding speaks as the output of the designed system. To do this, our system calls the event "ALTextToSpeech" through ALProxy contained in the NAO API. This event gives access to the robot's speech functionality and has the 'say' method capable of receiving a text as a parameter and transforms this text speaks to the robot.

IV. Results and Discussion

To evaluate the efficiency of each model chosen for our system, several experiments were carried out. For example, to analyze the efficiency of YOLOv3, we compared the result of YOLOv3 with other baselines, trained in the COCO database. Figure 2 shows a precision x speed graph compared to other latest generation models.

Regarding text detection and recognition, the above mentioned OctShuffleMLT [9] was selected instead of other deep models, because it is a compact model, in addition to its efficiency in recognizing scene text, in three different data sets, which are : ICDAR 2015, ICDAR 2017 MLT and 2019 MLT. The performance of OctShuffleMLT, in terms of precision and computational cost, was compared with different state of the art models, the results obtained by [9] can be seen in Tables I and II, where table I represents a comparison of the model in terms of precision, table II shows results in terms of compaction in relation to other models, in [9] it is explained how all these data were obtained.

The purpose of this study was to incorporate DL algorithms into NAO to perform scene text recognition. The application of DL methods on the robotics platform is a complicated task

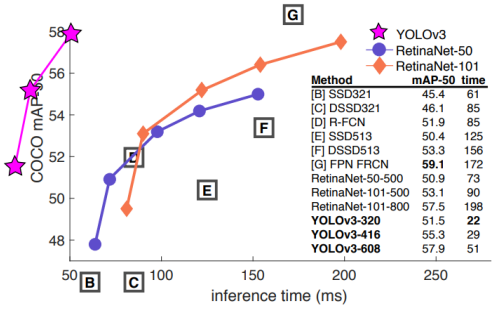

Fig. 2: Comparison between YOLOv3 model and other models. Source: Redmon and Farhadi (2018) [17].

TABLE I: Results obtained, about the accuracy of the model, in the database of ICDAR 2015 and ICDAR 2017 MLT. Source: Lundgren (2019) [9].

	ICDAR 2015		
Method	**Detection**	**Recognition**	**End-to-End**
E2E-MLT	-	-	55,1%
OctShuffleMLT	65,6%	61,7%	60,3%
	ICDAR 2017 MLT		
Method	**Detection**	**Recognition**	**End-to-End**
E2E-MLT	50,1%	65,1%	48,0%
OctShuffleMLT	64,2%	60,7%	58,6%

since DL methods require many computer resources, and the robotics platform is limited by computer resources. Also, this work encountered the technical limitations of the platform. Surprisingly, the results of the study were better than initially expected.

Figure 3 shows each step of the program's operation. Firstly, the user must enter the robot's IP address, allowing our system to connect with NAO. The user will need to press the sensor in front of the robot's head or the 'Q' key on the keyboard to capture an image of the scene from the camera.

TABLE II: Comparison of the OctShuffleMLT model with state-of-the-art models in relation to computational cost, taking into account the use of active memory, number of operations and number of parameters. Source: Lundgren (2019) [9].

Method	**Memory (MB)**	**Flops (G)**	**Parameters (M)**
OctShuffleMLT	81.61	0.829	1.6
E2E-MLT	93.98	2.946	4.8
FOTS	113.95	9.997	34.98
EAST	155.19	4.685	24.1

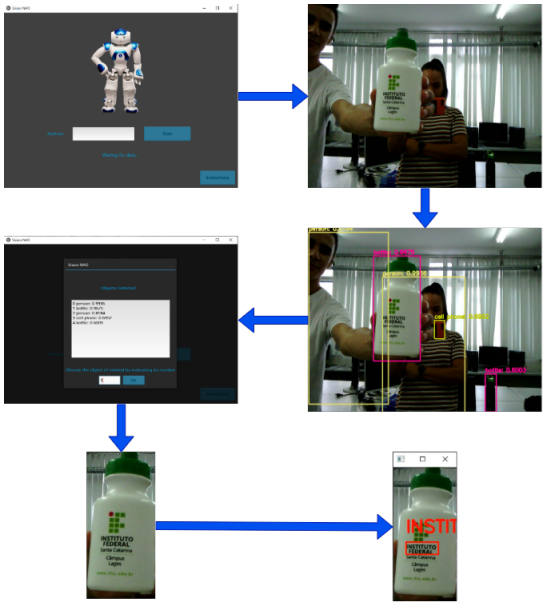

Fig. 3: Stages of the program's operation

The image is saved, and the program executes the YOLOv3 model. In this step, we define thresholds for the accuracy of the objects found. Above 90% we certainly consider that the object identified by the model is correct, between 60% and 80% we consider that perhaps the identified object is the correct one, and below 60% we are not sure if the identified object is the correct one, for each of these cases NAO speaks a different sentence.

Next, the image with the identified objects is displayed on the screen, and the interface will show these objects as a list; then, the user is asked to choose the object of interest so that the text content is recognized. The target region of interest is passed to the OctShuffleMLT text recognition model. Then, it detects and recognizes the text inside the target region. Finally, the NAO robot speaks the recognized text that is also displayed on the computer screen.

V. Conclusion

In this paper, we investigated the challenge of enabling a low-compute humanoid robot like NAO to recognize texts in objects in natural scenes in real-time using DL algorithms. More precisely, in this work, we presented an end-to-end vision system that is based on DL methods, specially designed for NAO robots. However, with this work, we show that using DL in NAO robots is indeed feasible and that it is possible to achieve promising results in the task of object text recognition in daily-life environments. Since the proposed pipeline to achieve real-time abilities are generic and modular, as future work we indeed to extend the stage 3 to improve the user interaction with NAO through voice recognition capabilities. In addition, we plan to evaluate the proposed system in daily-life activities, such as checking if the model can distinguish which drug should be given to a visually impaired based on a set of options seen by the robot.

Acknowledgment

This study was financed in part by the founding public agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, FACEPE and CNPq.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] D. Albani, A. Youssef, V. Suriani, D. Nardi, and D. D. Bloisi, "A deep learning approach for object recognition with nao soccer robots," in *Robot World Cup*. Springer, 2016, pp. 392–403.

[3] S. Chatterjee, F. H. Zunjani, and G. C. Nandi, "Real-time object detection and recognition on low-compute humanoid robots using deep learning," in *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2020, pp. 202–208.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[9] A. Lundgren, D. Castro, E. Lima, and B. Bezerra, "Octshufflemlt: A compact octave based neural network for end-to-end multilingual text detection and recognition," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 4. IEEE, 2019, pp. 37–42.

[10] J. Han, N. Campbell, K. Jokinen, and G. Wilcock, "Investigating the use of non-verbal cues in human-robot interaction with a nao robot," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2012, pp. 679–683.

[11] S. Shamsuddin, H. Yussof, L. Ismail, F. A. Hanapiah, S. Mohamed, H. A. Piah, and N. I. Zahari, "Initial response of autistic children in human-robot interaction therapy with humanoid robot nao," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. IEEE, 2012, pp. 188–193.

[12] J. P. Vital, M. S. Couceiro, N. M. Rodrigues, C. M. Figueiredo, and N. M. Ferreira, "Fostering the nao platform as an elderly care robot," in *2013 IEEE 2nd international conference on serious games and applications for health (SeGAH)*. IEEE, 2013, pp. 1–5.

[13] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A. S. Rusu, and D. O. David, "Children with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments," *Interaction studies*, vol. 13, no. 3, pp. 315–347, 2012.

[14] J. M. Sá, I. V. d. S. T. Pereira, and A. M. A. Maciel, "Integração de um modelo de reconhecimento de emoções ao robô humanoide nao," *Revista de Engenharia e Pesquisa Aplicada*, vol. 5, no. 1, pp. 110–116, 2020.

[15] A. Aly, "Human posture recognition and gesture imitation with a humanoid robot," *arXiv preprint arXiv:2002.01779*, 2020.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[17] A. Rosebrock, "Yolo object detection with opencv," *PyImageSearch, viewed*, vol. 20, 2018.