1st Open Day and Workshop
14th July 2022

INSTITUTO DE SISTEMAS E ROBÓTICA
UNIVERSIDADE DE COIMBRA

Deep Learning-based Perception Modules applied in a Dynamic Environment-based Visual Interface System for Brain-actuated Wheelchairs

Ricardo Pereira, PhD Student
Supervisor: Prof. Urbano Nunes Co-Supervisor: Prof. Ana Lopes

Introduction

Around the world, there are a significant number of people who are unable to perform their daily tasks due to severe motor impairments [1]. As a way to increase their autonomy and mobility, brain-actuated wheelchairs have been considered promising assistive devices [1], where visual user interface paradigms are one of the key modules. Most user interfaces used in P300-based BCI provide the possibility of selecting intermediate navigation goals or static low-level commands and do not incorporate surrounding environment information. However, to develop a dynamic environment-based user interface, perception modules able to recognize the surrounding environment, are required. A particular focus on improving the recognition of indoor places has been given, since the same place category may have various configurations and points of view, which become difficult to obtain an appropriate feature representation that is invariant to such conditions [4]. Moreover, as scenarios become more complex and the number of categories increases, the following problems need to be considered: inter-class ambiguity and intra-class variation.

DEVIS

The Dynamic Environment-based Visual Interface System [1] (DEVIS), as shown in Fig. 1, has in view a dynamic selection of navigation targets based on the surrounding context for brain-actuated wheelchairs. It comprises a Dynamic Visual Interface (DVI), an RGB image-based perception module, and a P300-based Brain-Computer Interface (BCI).

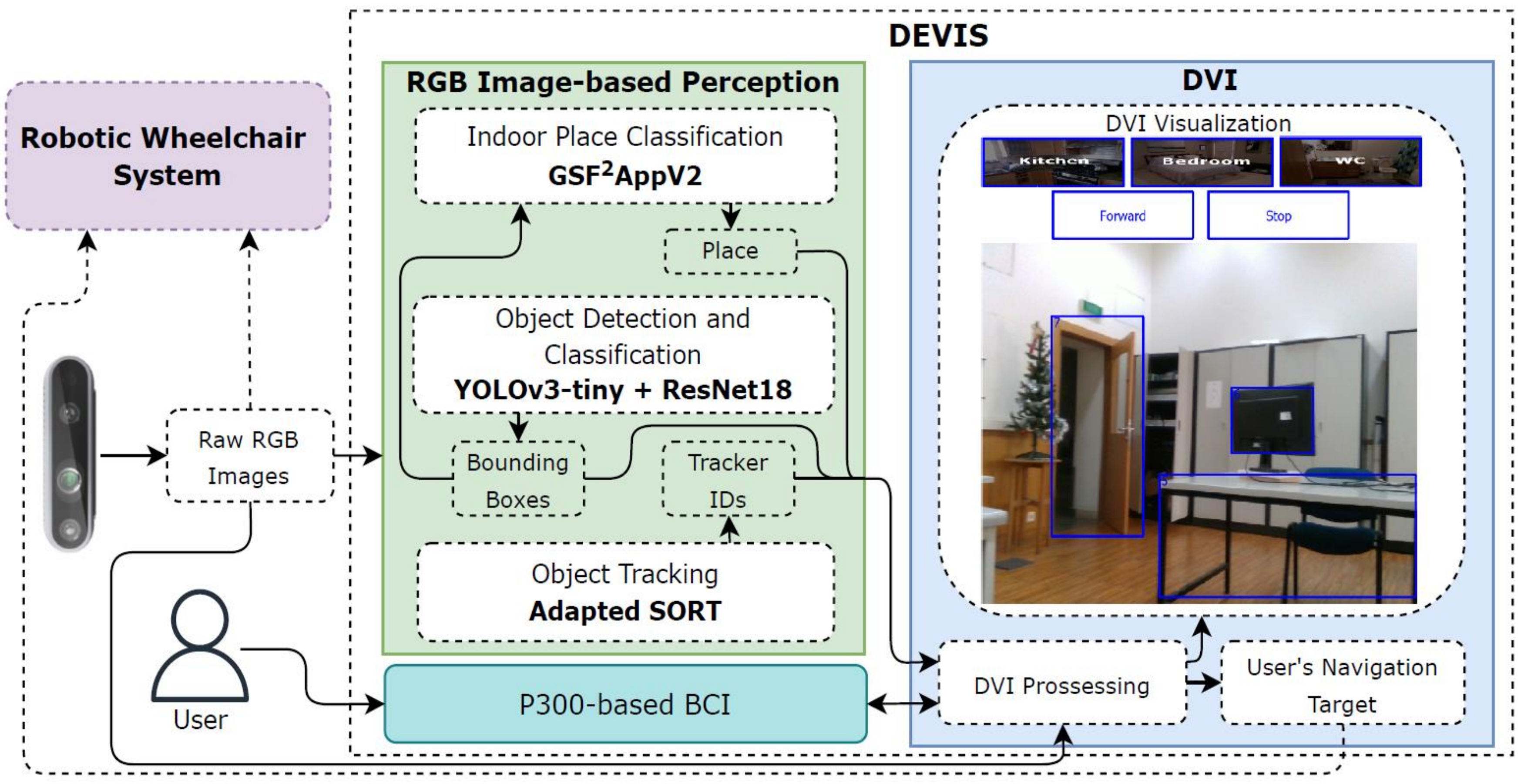

Fig. 1: Overview of DEVIS for brain-actuated wheelchairs.

- **DVI:** It displays potential environment-based navigation goals in two forms of visual cues: an RGB image streaming with bounding boxes overlaid on objects detected and tracked in real-time, and three rectangular boxes for indoor global points of interest (e.g. kitchen, bedroom, and WC). Additionally, two visual cues for static commands, FORWARD and STOP, are also displayed.
- **Perception Module:** It is used to obtain the environment-based information, which is composed of an object detection and classification method (YOLOv3-tiny + ResNet18 [2]), an object tracking method (adapted SORT [3]), and an indoor place classification method (GSF²AppV2 [4]).
- **P300-based BCI:** It allows the user to select, by flashing the DVI visual cues associated with each potential navigation target, his/her navigation target intent.

Validation and Experimental Results

DEVIS was validated with two different P300-based BCI modalities, non-self-paced and self-paced, and the experiments were carried out with 5 participants, who had to sequentially select targets from possible choices that could change according to the environment context. The ISR-RGB-D Dataset [2] was used to provide a dynamic setting for the evaluation.

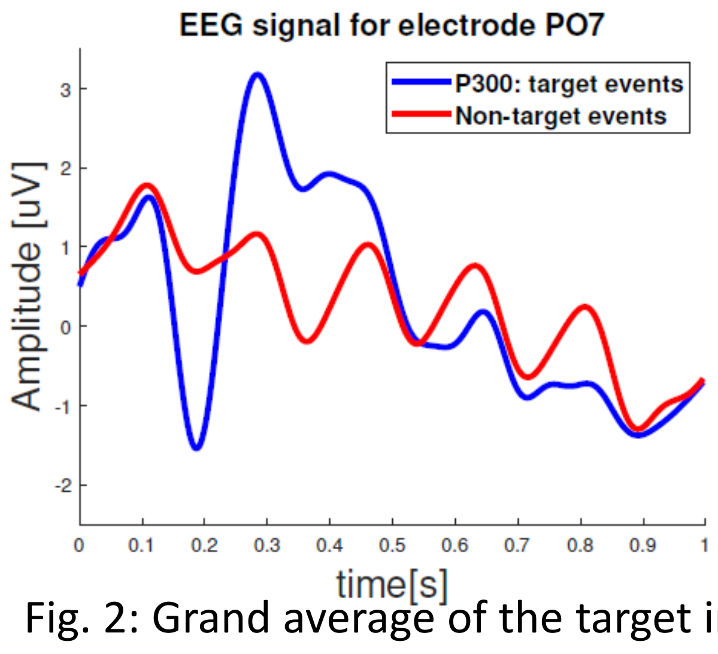

Fig. 2: Grand average of the target in a non-self-paced approach.

Subjects	Non-Self-Paced		Self-Paced	
	Acc	eSPM	Acc	eSPM
S1	93.3	5.0	100.0	5.4
S2	92.6	5.0	88.9	4.8
S3	86.7	4.7	73.3	4.0
S4	86.7	4.7	76.7	4.1
S5	90.0	4.9	93.3	5.0
Average	**89.9**	**4.8**	**86.4**	**4.7**

Table I: Classification Accuracy.

GSF²AppV2

To obtain a more meaningful scene representation, the Global and Semantic Feature Fusion Approach V2 [4] (GSF²AppV2), as shown in Fig. 3, simultaneously exploits CNN-based global features and semantic features (object-related features). It is a two-branch CNN architecture that uses a CNN-based state-of-the-art architecture to extract global features, and two different types of semantic features: Semantic Feature Vector [5] (SFV) and Semantic Feature Matrix [4] (SFM).

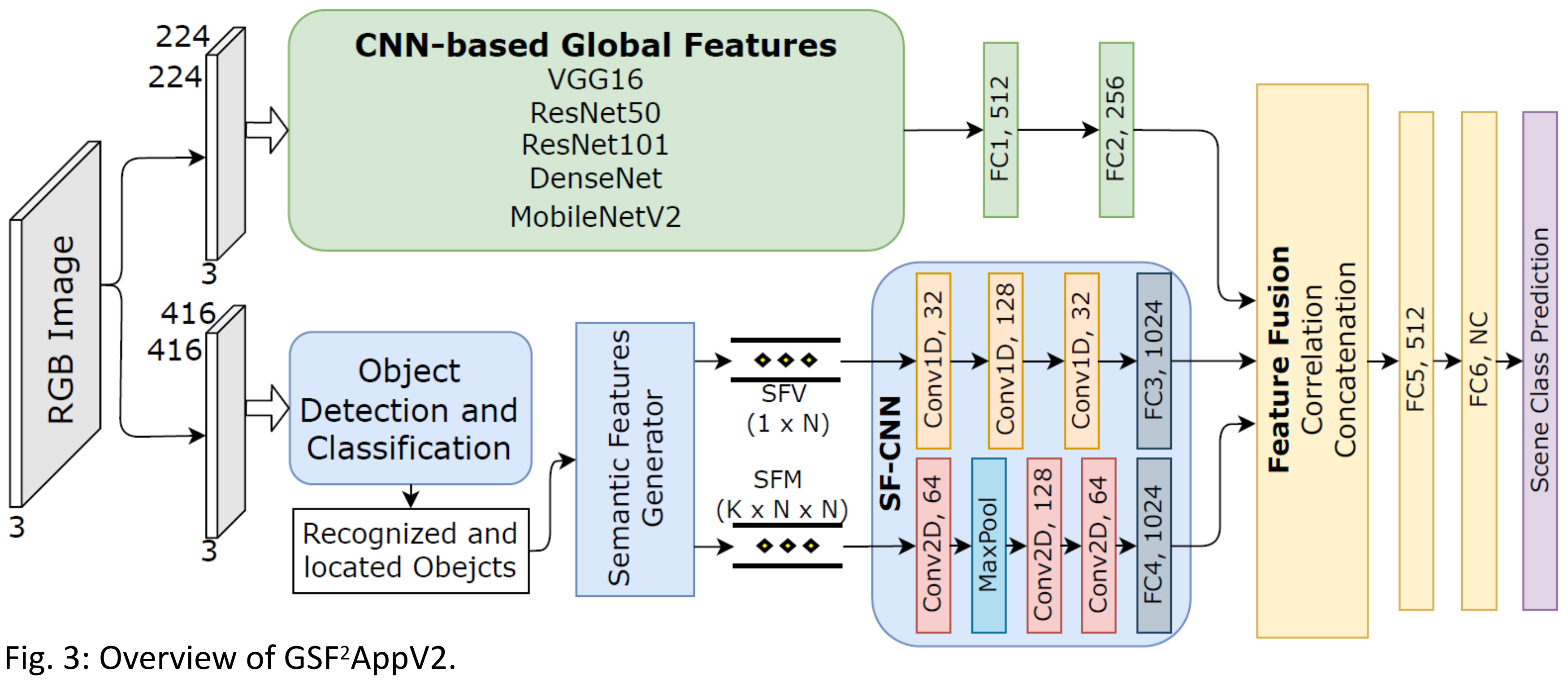

Fig. 3: Overview of GSF²AppV2.

- **SFV:** Contains the number of object occurrences recognized in the image per object class ($O_{\{i=1:N\}}$), represented as follows: $SFV = \begin{bmatrix} O_1 & O_2 & \cdots & O_N \end{bmatrix}$
- **SFM:** It is a histogram-like approach that represents inter-object distance relationships in terms of how close or apart objects are between two pairs of object classes. Each class-pair inter-object relationship bin ($b_{i,j,k}$) is expressed as follows:

$$b_{(i,j,k)} = \sum_{m=1}^{|C^{[i]}|} \sum_{n=1}^{|C^{[j]}|} f(C_n^{[i]}, C_m^{[j]}, k) \qquad \text{with,} \quad f(C_A, C_B, k)) = \begin{cases} 1 & \text{if } \rho \frac{|C_A - C_B|}{d_{max}} = k, k \in [1, K] \\ 0 & \text{otherwise} \end{cases}$$

where N is the number of classes, K the number of distance bins, $C^{[i]}$ and $C^{[j]}$ represent the detected object's bounding boxes for each class (i and j), ρ is a scaling factor for the number of distance bins ($\rho = K$), and d_{max} is a normalization distance. Inter-object relationship features are represented as follows (SFM):

$$SFM_{(N,N,K)} = \begin{bmatrix} \begin{bmatrix} b_{(1,1,1)} & \cdots & b_{(1,N,1)} \\ b_{(2,1,1)} & \cdots & b_{(2,N,1)} \\ \vdots & \ddots & \vdots \\ b_{(N,1,1)} & \cdots & b_{(N,N,1)} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} b_{(1,1,K)} & \cdots & b_{(1,N,K)} \\ b_{(2,1,K)} & \cdots & b_{(2,N,K)} \\ \vdots & \ddots & \vdots \\ b_{(N,1,K)} & \cdots & b_{(N,N,K)} \end{bmatrix} \end{bmatrix}$$

Results

Backbone	mAC (%)						
	CNN-only	+ SFV		+ SFM		GSF²App v2	
		FCor	FCon	FCor	FCon	FCor	FCon
VGG16	69.1	71.4	71.2	71.1	71.2	71.8	71.7
ResNet50	**70.8**	71.5	72.2	71.5	72.3	72.0	72.7
ResNet101	70.7	**71.6**	**72.4**	**71.7**	**72.5**	**72.2**	**73.1**
DenseNet	67.7	68.3	68.9	68.4	69.4	69.0	69.8
MobileNetV2	67.8	68.2	69.5	68.5	69.7	69.2	70.1

Table II: Achieved Mean accuracy class on the NYU Depth Dataset v2 [6] with K=3.

Future work: Improve the global feature representation with new Deep Learning techniques. Evaluate the influence that each object class may have in each indoor scene category. Improve the semantic scene representation through the development of new types of object-related information.

Acknowledgments

Ricardo Pereira has been supported by the Portuguese Foundation for Science and Technology (FCT) under the PhD Grant SFRH/BD/148779/2019. Works have been also partially supported by the project B-RELIABLE, with reference SAICT/30935/2017, and by ISR-UC through FCT Grant UIDB/00048/2020.

References

[1] R. Pereira, A. Cruz, L. Garrote, G. Pires, A. Lopes, and U. J. Nunes, "Dynamic Environment-based Visual User Interface System for Intuitive Navigation Target Selection for Brain-actuated Wheelchairs", in *IEEE Int. Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[2] R. Pereira, T. Barros, L. Garrote, A. Lopes, and U. J. Nunes, "An Experimental Study of the Accuracy vs Inference Speed of RGB-D Object Recognition in Mobile Robotics," in *IEEE Int. Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2020.

[3] R. Pereira, G. Carvalho, L. Garrote, and U. J. Nunes, "Sort and Deep-SORT Based Multi-Object Tracking for Mobile Robotics: Evaluation with New Data Association Metrics," *Applied Sciences*, vol. 12, no. 3, 2022.

[4] R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "A Deep Learning-based Indoor Scene Classification Approach Enhanced with Inter-Object Distance Semantic Features," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2021.

[5] R. Pereira, N. Gonçalves, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "Deep-Learning based Global and Semantic Feature Fusion for Indoor Scene Classification," in *IEEE Int. Conf. on Autonomous Robot Systems and Competitions (ICARSC)*, 2020.

[6] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGB-D Images," in *ECCV*, 2012.