Potential of embedded vision platforms in development of spatial AI enabled CPS

Branko Brkljač*, Boris Antić, and Zoran Mitrović
{brkljacb*, antic, zoranmit}@uns.ac.rs
Department of Power, Electronic and Telecommunication Engineering,
Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia

***Abstract*—Motivated by the recent trends in the field of embedded vision platforms, we discuss potential of such solutions in providing foundations for the next generation of Cyber-Physical Systems (CPS). Improved capabilities and reduced price of these platforms will have profound effect on their everyday usage and applications. In comparison to speech and natural language processing, which have established speech recognition and machine translation applications as indispensable in many contemporary CPSs, the vision community is still searching for an application that would be so necessary and desirable to make most of the consumers buy specific vision hardware just to run it. That would be the ultimate proof of the core value of the technology in the market. Thus, also vision problems come with a longstanding tradition and history of numerous solutions, it is still hard to point out a single application that would incorporate many specific vision tasks into one device, and which would be ubiquitously useful and affordable to all (e.g. like smartphone has done in the fields of communication and personal computing). However, with development of new miniaturization technologies and spatial AI it is reasonable to expect that there will be more possibilities for designing CPS with capabilities of visual understanding of outdoor, dynamic and uncontrolled environments. One step in such direction are embedded vision platforms that besides powerful computing capabilities also provide multimodal perception, and thus improve the algorithm performance. As an example, we will discuss stereo depth perception in the context of new spatial AI platforms like OAK-D lite, and point out some possibilities for its improvement and integration into future CPS.**

***Index Terms*—CPS, embedded vision, multimodal perception, spatial AI, depth from stereo, OAK-D lite;**

I. Introduction

Visual perception plays important role in understanding the world by humans and the machines [1]. It is a basic step in visual information processing and foundation of vision based inference. Thus, it represents the key enabling technology for mobile robotics, outdoor navigation, autonomous driving and many other contemp. Cyber-Physical Systems (CPS) [2, 3].

Embedded vision platforms are the next step in the design and development of technical devices that provide such capabilities. They offer seamless integration of camera devices and computing platforms that are highly optimized for execution of computer vision algorithms. Instead of being large and expensive, platforms are usually designed to have small size, low cost, low energy consumption and low weight [4]. Although, in some specific applications that are dealing with strict requirements or targeting only a small set of vision tasks (like in some manufacturing processes, measuring devices or machines), all of the above mentioned characteristics are not always the priority and some of them can be discarded. However, in the case of general embedded vision platforms, which are not tailored for some specific need, such features are always welcomed and make the system easier to fit into some larger CPS and the corresponding application scenario.

Small form factor and efficient use of power resources, at low cost, are always contrasting performance and diversity of capabilities that device offers. In that sense, it is hard to find a platform that would be a one-size-fits-all solution, Fig. 1.

This is especially true in vision, where many tasks have different levels of complexity and can vary significantly depending on the environnement. Historically, many of the vision applications that require visual understanding of the scene or precise measurements were oriented towards static and controlled indoor environments. This is not surprising, since it is always easier to search for solution under assumptions that make the problem more constrained and less general. However, challenging problems are always opportunity for wider adoption of vision technologies, so sentiment towards solving of problems 'in the wild' is steadily growing.

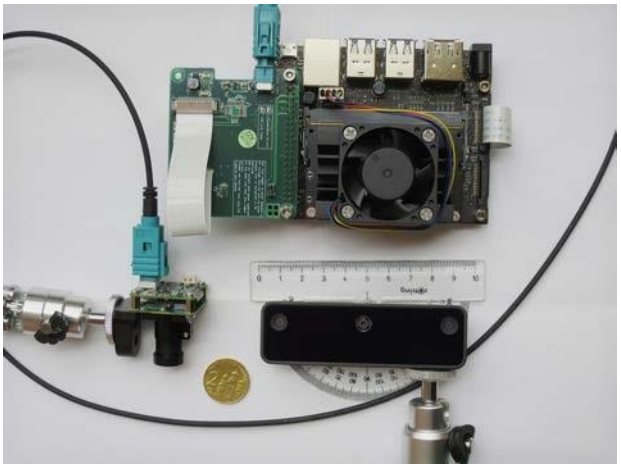

Fig. 1: Jetson Nano [5] with camera extension board and coaxial cable, in comparison to OAK-D lite [6], which integrates three cameras and processing board in a single case.

This work was partially supported by the European Union's H2020 research and innovation programme under the LEIT project grant agreement No.872614, SMART4ALL - Selfsustained cross border customized cyberphysical system experiments for capacity building among European stakeholders.

General vision tasks 'in the wild' are always considered as harder or at least more sophisticated. E.g. the gait recognition has been the subject of many investigations, but up to recently the quality of cross-view recognition was limited by the characteristics of available datasets that were recorded in controlled environment [7]. Similarly, some complex tasks that are relatively easy for humans require 3D understanding of the environment [8]. Even the optical character recognition (OCR) and text digitization, which have been considered as prime examples of vision tasks performed in controlled environments by the line scan sensors and their variants, have evolved from the methods that are considered as standards for OCR in the software industry [9], into numerous solutions oriented towards uncontrolled environments [10]. As a result, these trends are also reflected into design of novel embedded vision platforms that support development of spatial AI.

Spatial AI is a term that can be attributed to Davison [11]. Accordingly, devices with such capabilities should operate in real-time, in a context and with goals. In that sense, the term goes beyond visual perception of complicated 3D environments and abstract scene understanding. Besides, it also involves learning of optimal signal representations and continuous capturing of right information that will jointly enable real time interpretation and action [12]. CPS incorporating such functionalities can be regarded as spatial AI enabled.

The rest of the paper is organized as follows. In Section II we introduce the multimodal perception paradigm. Next, design of a recently proposed embedded vision platform with passive camera sensors and native depth perception capabilities is discussed in Section III. Finally, in Section IV we briefly reexamine possibilities for improvement of depth perception based on stereo vision and suggest necessary requirements for incorporating such solutions into CPS with depth perception functionalities. At the end, the paper is concluded in Section V, with a reference to future work and possible applications.

II. CPS and multimodal perception

If it is expected to achieve wider adoption of CPS in the future, human-machine interactions will need to be carefully designed functionalities of the next generation CPS. This will improve the overall user experience, but also make the added value brought by CPS more easily recognized by the society. In that sense, vision tasks like facial expression recognition or person identification will require human level performance in order to gain necessary trust among the CPS users. E.g. thanks to advanced attention-based feature fusion, current solutions in the field of facial expression recognition are more robust to occlusions and variant head pose [13]. However, in order to achieve human level performance, or go beyond, such systems will also need to exploit 3D facial information [14] and perform multimodal feature fusion [15] – in addition to novel learning-based feature engineering.

The prevailing opinion in the vision community is that the general trend in the future will be towards providing CPS with multimodal information about its surrounding [16]. Therefore, regardless of that whether the 3D information will be result of: a) advanced 3D reconstruction, like in [14]; b) pretrained neural network models that are tailored for odometry and depth perception based on monocular videos [17, 18]; or c) come from some dedicated hardware like stereo camera rigs or active camera sensors [19, 20], such complementary information will be necessary to ease the problems and accomplish the vision tasks more successfully.

Uncertainty reduction provided by implicit or explicit use of 3D information is expected to bring complex vision functionalities to everyday life and uncontrolled environnements. Therefore, domain of multimodal perception will be of particular interest for product positioning and research field where different solutions will be competing for their place on the market (besides 'standard' characteristics like computational power, efficiency and support for deep neural network (DNN) inference). Embedded platforms capable of providing necessary level of 3D perception with low energy consumption will be a preferred choice for solving vision tasks in future CPS.

III. Native depth from stereo

Implementation of spatial AI goes hand in hand with heterogeneous computing environnements. When it comes to depth perception, embedded vision platforms of new generation, like OAK-D lite [6], Fig. 1 and Fig. 2, have significantly evolved in comparison to their predecessors. E.g. 'Bumblebee' devices [21], once considered as the industry standard for mobile robotics, were designed to have a pair of calibrated stereo cameras in a metal case, but without any processing capabilities. Instead, the computational load of determining depth of the scene was usually transferred from the acquisition device to external computer, which had to detect image correspondences and estimate disparity maps. However, due to lack of an efficient implementation over a specialized hardware, such approach was usually resulting in small frame rates [19], or led to non real-time processing in cases when the perception accuracy was the main design priority.

Stereo vision is often perceived as more robust to ambient illumination than ToF cameras. For example, Kinect [22], is

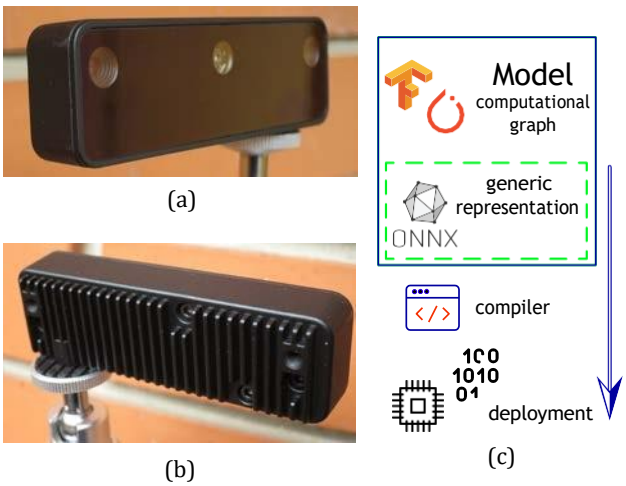

Fig. 2: (a) Global shutter stereo pair, and a 4K video camera; (b) back cover of OAK-D lite; (c) hardware acceleration steps.

not suitable for outdoor applications due to the interference of the direct sunlight with the camera emitter in the near-infrared range. If presence of textureless objects in the scene makes dense image correspondences harder to learn, high resolution depth maps can also be hard to produce by stereo vision, requiring a more complex image descriptors or structured light projection [23]. In general, active illumination makes any approach less suitable for outdoor applications and spatial AI.

Thanks to novel computing architectures, contemporary embedded vision platforms are closer to the concept of edge AI and processing of the information at the place of their acquisition. For the purpose of this paper we have performed some outdoor tests and generated results shown in Fig. 3. Platforms like [6] can produce VGA resolution depth maps with high frame rates [24], but are also capable of other processing tasks, like real time inference using pretrained DNNs, or HEVC 4K hardware video encoding [25]. E.g. Intel's Movidius chip [26], which is positioned on the backside of the processing board, theoretically produces up to 4 TOPS, out of which 1.4 TOPS for 'neural compute engine'. Besides 16 CPU cores for image signal processing, it also provides dedicated hardware accelerator for custom DNN models made in standard deep learning libraries, and several predefined hardware accelerators for standard vision tasks. Although on-device development through firmware change is not allowed, there is a possibility to exploit available neural network interface to implement custom computational graphs corresponding to computationally heavy vision tasks.

As depicted in Fig. 2c, the first step in such procedure is to define a dummy neural network model, which is defined in usual way using standard programming interface in some of mainstream DNN libraries. In the next step, created computational graph description in '.onnx' format [27] is further optimized by 'onnx-simplifier' [28]. Finally, after the generated model description is compiled to necessary '.blob' format, which is standard for similar hardware accelerators based on Movidius MyriadX VPU architecture [29], the model is ready to perform 'neural inference' on input data.

IV. Applications in CPS

Described platform [6] is only one of several available in the market [5, 29, 30]. It was developed as the result of an open funding campaign initiated by the vision community gathered around OpenCV project [31]. In order to make the platform more affordable and widely applicable, some characteristics have been chosen to be less top notch in comparison to capabilities of the chip and the processing board. E.g. spatial resolution of stereo cameras was chosen to be smaller, while an inertial measurement unit was left out, although the allocated space and connectors still exist on the designed circuit board. On the other hand, platforms like [30] highlight the advantages of reconfigurable hardware based on the FPGA technology, which can also be foundation for similar embedded vision platforms. Key advantage of [6, 30] over [5] is a more compact design and no need for additional camera hardware, while [5] has richer I/O interface, GPU with CUDA support, and does not require an additional microcontroller or host computer.

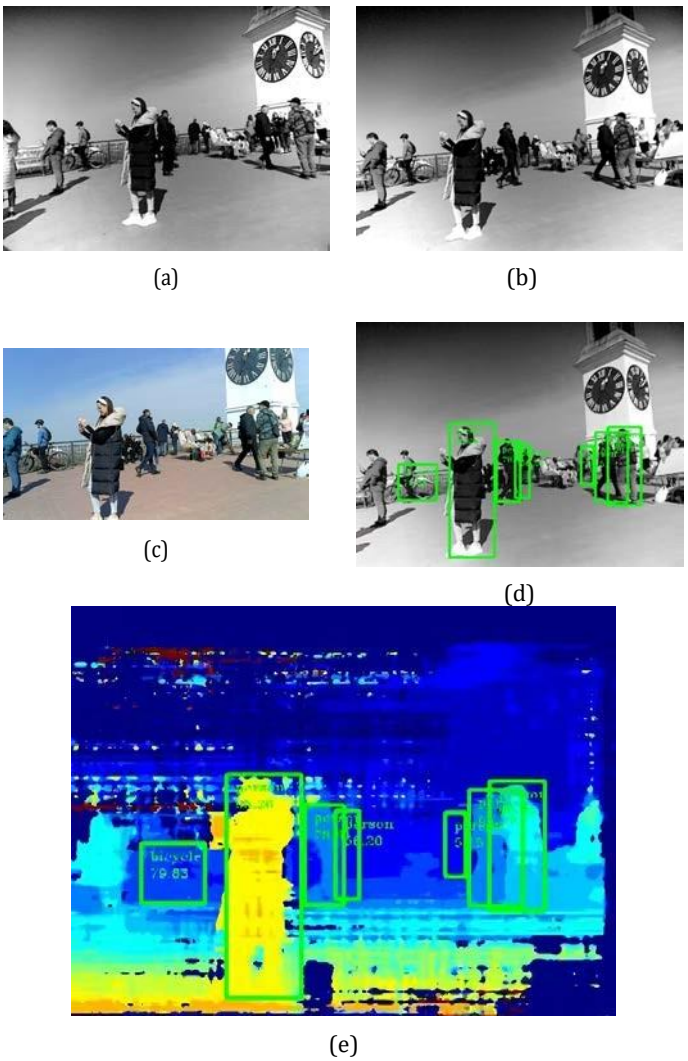

Fig. 3: Real time depth perception, video encoding and neural inference performed simultaneously on device in Fig. 2. (a)-(b) left and right camera view; (c) central camera; (d) detections based on image in (b); (e) computed disparity map.

In this study we have investigated capabilities of platform [6] in the context of potential applications in CPS. As shown in Fig. 3e, it has successfully demonstrated effective depth perception, with high level of hardware integration, Fig. 1. However, our tests have also confirmed that there is space for significant improvement of depth perception quality provided out of the box by the same platform. As the result of additional experiments, illustrated in Fig. 4, which were conducted based on the implementation provided in [32], it was concluded that strategies for possible improvements of depth perception quality should rely on DNNs and their high learning capacity.

In Fig. 4, next to each other are the results of an 'on device' depth perception performed out of the box by [6], and the perception experiment based on the same stereo images by using the method proposed in [33] and the implementation from [32]. The host computer was equipped with high-end GPU and support for tensorRT inference engine [34], which enabled real time performance. However, the hardware requirements

(a) (b)

Fig. 4: Visual comparison of performed stereo depth perception experiments: (a) 'on device' solution based on [6]; (b) 'on host' solution based on method from [33] and the same image pair.

and power consumption were much higher in comparison to [6]. Therefore, a solution for higher quality depth perception in CPS applications would be to combine different embedded platforms, and leverage the best from each of them. For example, Jetson Nano in Fig. 1 also supports tensorRT engine and could perform 'on the host' tasks, like the one from [33].

Some CPS applications that would benefit from such improved perception are e.g. [35] or [36], where embedded vision platform was mounted on small UAV and provided vision capabilities necessary for autonomous drone navigation.

V. Conclusion

Embedded vision platforms are expected to bring novel CPS functionalities and provide multi modal perception in outdoor environments. Different research groups are working towards lowering the cost of such solutions and enabling the technological base for spatial AI. Depth perception solutions that were experimentally compared in this work are only one example of such efforts. Performed tests have confirmed promising characteristics of the recently proposed platform [6] and identified possible advantages of its combination with the existing solutions like [5]. We hope that presented information and experimental comparisons will provide more insights and make integration of similar devices in CPS easier. Our future work will be oriented towards the improvement of the existing algorithmic techniques and their efficient implementation.

References

[1] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. 2ed., MIT press, 2010.

[2] A. Darwish and A. E. Hassanien, "Cyber physical systems design, methodology, and integration: the current status and future outlook," *J of Amb. Intell. and Hum. Comp.*, vol. 9, no. 5, pp. 1541–1556, 2018.

[3] C. V. Lozano and K. K. Vijayan, "Literature review on cyber physical systems design," *Procedia manufacturing*, vol. 45, pp. 295–300, 2020.

[4] B. Kisačanin, S. S. Bhattacharyya, and S. Chai, Eds., *Embedded computer vision*. Springer, 2009.

[5] (2019) NVIDIA corporation. Jetson Nano. [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano

[6] (2022) Luxonis corporation. OAK-D lite. [Online]. Available: https://docs.luxonis.com/projects/hardware/en/latest/pages/DM9095.html

[7] Z. Zhu *et al.*, "Gait recognition in the wild: A benchmark," in *Int. Conf. on Computer Vision*, 2021, pp. 14 789–14 799.

[8] Z. Cao, I. Radosavović, A. Kanazawa, and J. Malik, "Reconstructing hand-object interactions in the wild," in *Int. Conf. on Computer Vision*, 2021, pp. 12 417–12 426.

[9] R. Smith, "An overview of the Tesseract OCR engine," in *Int. Conf. on Document Analysis and Recognition*, vol. 2. IEEE, 2007, pp. 629–633.

[10] X. Chen *et al.*, "Text recognition in the wild: A survey," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–35, 2021.

[11] A. J. Davison, "FutureMapping: The computational structure of spatial AI systems," *arXiv preprint arXiv:1803.11288*, 2018.

[12] ——. (2020) From SLAM to spatial AI. [Online]. Available: https://www.youtube.com/watch?v=lGIM2WVp5t0

[13] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[14] T. S. Ly *et al.*, "A novel 2D and 3D multimodal approach for in-the-wild facial expression recognition," *Image and Vision Computing*, vol. 92, p. 103817, 2019.

[15] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, pp. 1–32, 2021.

[16] G. Gkioxari. (2022) Trends in computer vision. [Online]. Available: https://www.youtube.com/watch?v=eJ4IeChWVz0

[17] A. Gordon *et al.*, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Int. Conf. on Computer Vision*, 2019, pp. 8977–8986.

[18] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[19] W. Kazmi *et al.*, "Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison," *ISPRS J of Photogrammetry and Remote Sensing*, vol. 88, pp. 128–146, 2014.

[20] P. T. Boufounos, "Time-of-flight sensor," U.S. patent 0 100 926A1, 2018.

[21] Point Gray Inc. Triclops/Bumblebee reference manual. [Online]. Available: http://csis.pace.edu/robotlab/papers/TriclopsManual.pdf

[22] J. Smisek, M. Jancosek, and T. Pajdla, "3D with Kinect," in *Consumer depth cameras for computer vision*. Springer, 2013, pp. 3–25.

[23] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2003, pp. 1–8.

[24] B. Gilles. (2021) 'OAK-D lite' tear down. [Online]. Available: https://www.youtube.com/watch?v=qFnXyr9iFyo

[25] G. J. Sullivan *et al.*, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[26] (2017) Intel corporation. Movidius Myriad X VPU. [Online]. Available: https://newsroom.intel.com/wp-content/uploads/sites/11/2017/08/movidius-myriad-xvpu-product-brief.pdf

[27] (2022) Open Neural Network Exchange - ONNX. [Online]. Available: https://onnx.ai

[28] (2022) ONNX simplifier. [Online]. Available: https://github.com/daquexian/onnx-simplifier

[29] (2018) Intel corporation. Neural compute stick 2. [Online]. Available: https://www.intel.com/content/dam/develop/public/us/en/documents/ncs2-data-sheet.pdf

[30] (2020) IDS GmbH. AI for all - NXT inference camera. [Online]. Available: https://en.ids-imaging.com/files/downloads/knowledgebase/pdf/technical article/en gb/ids-nxt-ai-for-all.pdf

[31] (2022) Open source Computer Vision library - OpenCV. [Online]. Available: https://opencv.org/about/

[32] (2022) C++ wrapper interface for OAK-D device. [Online]. Available: https://github.com/iwatake2222/play with depthai

[33] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 362–14 372.

[34] (2022) NVIDIA corporation. TensorRT SDK for high-performance DL inference. [Online]. Available: https://developer.nvidia.com/tensorrt

[35] C. A. Luna, C. Losada-Gutiérrez, D. Fuentes-Jimenez, and M. Mazo, "People re-identification using depth and intensity information from an overhead camera," *Expert Systems with Applications*, vol. 182, pp. 115 287:1–9, 2021.

[36] L. O. Rojas-Perez and J. Martinez-Carranza, "Towards autonomous drone racing without GPU using an OAK-D smart camera," *Sensors*, vol. 21, no. 22, pp. 7436:1–19, 2021.