YEEHaD: YOLO-based Extremely Efficient Hand Detection

Gibran Benitez-Garcia and Hiroki Takahashi

***Abstract*— This paper presents YEEHaD, an Extremely Efficient Hand Detection approach based on YOLO architecture. We introduce modifications to the Cross Stage Partial (CSP) block and include depthwise separable convolutions to reduce the computational overhead of CSP-Darknet and CSP-PAN in YOLOv5. YEEHaD can perform real-time at 640x640 resolution on different embedded systems, achieving less than 3 GFLOPs with fewer than 1.1M parameters. We evaluate our proposal on two public datasets, including cross-dataset validations, to highlight its robustness. Additionally, we contribute with manual annotations of hand locations for the NVGesture dataset. Finally, we explore how detection models can be fine-tuned for hand gesture recognition (HGR).**

I. Introduction

Reliable hand detection in images and videos is critical to a wide range of computer vision tasks, including hand gesture recognition, hand pose estimation, and sign-language recognition. As technology evolves, there's an increasing demand for real-time hand detection solutions that perform on low-powered devices while maintaining high accuracy. In such environments, computational cost is a key factor due to the inherent limitations of embedded systems. This research introduces YEEHaD, a YOLO-based solution for these challenges, offering both efficiency and robust performance. We evaluate our proposal with two public datasets, emphasizing cross-dataset validations. We also provide manual annotations for hand locations of about 40K frames from the NVGesture dataset. Furthermore, we explore the potential of fine-tuning our detection models specifically for HGR. We mainly analyze which layers, initially trained for hand detection, can also benefit HGR. In this way, we prove the adaptability of YEEHaD, validating its value as a versatile tool for diverse computer vision tasks.

II. Proposed Method

We design YEEHaD around the YOLO single-stage architecture, known for its real-time processing and exceptional accuracy in object localization. Our selection is the YOLOv5 framework [1], adhering to the classical structure comprising the backbone, neck, and head. The backbone combines the efficiency of Cross Stage Partial networks (CSP) [2] with the robustness of the Darknet53 architecture [3]. The neck employs a Path Aggregation Network (PAN) built on CSP blocks, ensuring a balanced feature distribution. Finally, the head adopts the multi-scale version of YOLOv3 [3]. In particular, we introduce modifications to the CSP block.

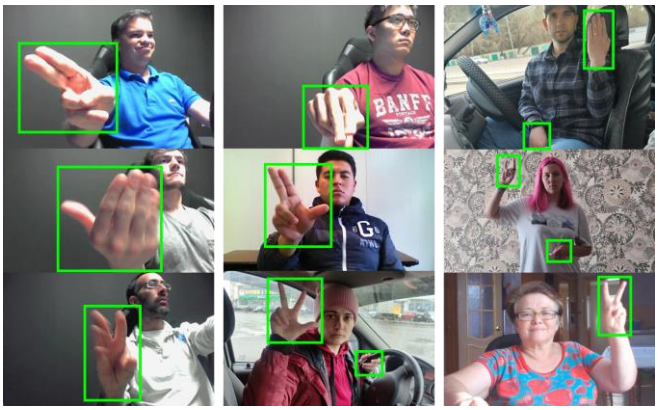

Figure 1. Hand annotations from the two datasets used in this work.

The CSP modifications aim to optimize computational efficiency while retaining detection precision. Additionally, we replace the conventional 3x3 convolutions with 3x3 depthwise separable convolutions to speed up the process in the backbone and neck. YEEHaD's architectural design is constructed to keep its size below 1.1M parameters and not to surpass 3 GFLOPs at 640x640 input resolution. Consequently, we ensure its real-time performance on desktop GPUs and even on embedded systems like Jetson Xavier NX.

III. Experiments

YEEHaD was tested on the NVGesture and Hagrid datasets, depicted in Figure 1. Experiments were conducted on the Jetson Nano and Jetson Xavier NX platforms to validate its application in embedded systems. Leveraging pre-existing knowledge, we fine-tuned YEEHaD for HGR with multiple classes from each dataset. We present a systematic evaluation, identifying the layers that should be trained with HGR-specific data, building upon the foundational hand detection training. This evaluation is extended through cross-dataset validations. Our experiments show that YEEHaD can achieve real-time performance while maintaining high accuracy for both hand detection and HGR tasks. Further details and benchmarks will be included in the full paper.

References

[1] Wang, C. Y.,, et al., “ultralytics/yolov5: v6.2,” https://github.com/ultralytics/yolov5/ (accessed 2023-05-30).
[2] CY. Wang, et al., “CSPNet: A new backbone that can enhance learning capability of CNN,” in CVPRW, pp. 390-391, (2020).
[3] J. Redmon, and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767 (2018).

*Research supported by a Research Grant (S) at Tateisi Science and Technology Foundation.

G. Benitez-Garcia is with the Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan (G. Benitez-Garcia, gibran@ieee.org, corresponding author).

H. Takahashi is with the Graduate School of Informatics and Engineering, Artificial Intelligence eXploration Research Center (AIX), and Meta-Networking Research Center (MEET), The University of Electro-Communications, Tokyo, Japan.