ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

# Large-scale dataset and benchmarking for hand and face detection focused on sign language

Alvaro Leandro Cavalcante Carneiro[1], Denis Henrique Pinheiro Salvadeo[1]
and Lucas de Brito Silva [1] *

São Paulo State University, Institute of Geosciences and Exact Sciences.
Rio Claro, São Paulo - Brazil

**Abstract**. Object detection is an important preprocessing technique for sign language recognition, allowing focus on the most important parts of the image. This paper introduces a new large-scale dataset for hand and face detection in sign language context, mitigating the lack of data for this problem. We evaluated different object detection architectures to find the best trade-off between computational cost and mean Average Precision (mAP). The proposed dataset contains 477,480 annotated images. The most accurate detector (CenterNet) achieved an mAP of 96.7%. Furthermore, the optimizations made to the models reduced the inference time up to 74% in the best scenario.

## 1 Introduction

Disabling hearing loss and deafness are two conditions that affect millions of people around the world, impairing their ability to communicate and making social interactions harder. However, new research is emerging focusing on the creation of sign language recognition systems capable of translating signs into text. One of the primary concerns in developing these systems is to determine the optimal pre-processing technique, as this can significantly improve the accuracy. Pose estimation and object detection techniques, for example, are being widely used to segment the interpreter's hands and face, which contain most part of the necessary information to classify the sign. Object detection, in particular, offers the advantage of being less computationally expensive, making them suitable for execution on edge devices like smartphones.

Despite the benefits of object detectors, many previous works in this area do not document, compare or publish their models. Furthermore, there are only a few publicly available datasets with hand and face annotations, and none of them are focused on the context of sign language recognition. Based on that, this work aims to create a new large-scale dataset for hand and face detection by labeling the images from the AUTSL [1] sign language database, which has a diverse range of interpreters and backgrounds. We subsequently fine-tuned and trained various object detection architectures for the given context, comparing them in a trade-off perspective between the mean Average Precision (mAP) and computational cost. In addition, the trained models, source code, and dataset are publicly available on GitHub[1] to assist future researchers.

*This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil

[1] https://github.com/AlvaroCavalcante/hand-face-detector

## 2 Related Works

Considering previous works, we can find some publicly available datasets containing annotated hands [2, 3], faces [4] and both [5, 6]. The Autonomy [6] dataset stands out among them since it includes new data and combines multiple other datasets with annotated hands and faces that conform to the same annotation standard, totaling 50,365 images. However, all of these datasets are general-purpose, with images captured under diverse conditions that do not mirror those typically encountered in sign language recordings, which usually contain a single person centralized in the image.

Regarding sign language recognition, previous works [7, 8, 9] have successfully utilized object detectors to improve model accuracy by detecting hands and faces. Nevertheless, none of them compare different architectures for the task, and some [7, 9] did not detail the precision obtained by the detector, resulting in simplistic solutions that may not generate the best detection performance.

## 3 Dataset preparation

The large-scale dataset proposed in this research, called HFSL (Hand and Face for Sign Language), is based on the AUTSL dataset, which presents 43 interpreters on 20 different backgrounds, totaling more than 38,639 videos. Besides that, the recordings were made in a resolution of 512x512 pixels and had an average of 61 frames per video, resulting in more than two million images. To facilitate the annotation of this amount of images, we adopted a semi-supervised approach using the predictions of an initial model. This model was trained on a simplified version of the Autonomy collection, which does not have the EgoHands and Viva datasets as they contained images captured in a first-person view, which is not applicable to sign language. The remaining 40,064 images of Autonomy were used to train the EfficientDetD1 [10] initial model.

After that, we developed an algorithm to iterate over the AUTSL videos and extract the desired data according to some predefined rules. Firstly, to prevent the collection of an excessive number of frames or too many similar images, we limited the number of frames per video to 16 and employed a dynamic step parameter as part of our sub-sampling strategy. The step size is calculated by dividing the total number of frames in the video by the maximum number of frames allowed and rounding up to the highest integer.

For each sampled frame, we utilized the initial object detector and converted the model's predictions into PASCAL VOC [5] labels. Additionally, we applied a confidence threshold of 35% to accept the inferences as correct. However, to reduce the amount of noisy data and the effort required for manual label correction, we filtered the predictions to retain only the face and hands with the highest confidence, discarding the frame if the desired objects were not detected.

Finally, we manually reviewed and improve all labels of the selected frames to better fit the objects and guarantee the quality of the data. The new object detection dataset reached a total of 477,480 images, where each one of them has exactly one face and two hands, as illustrated in Figure 1. Moreover, we adhered to the ChaLearn competition [11] guidelines to divide the data into 31 individuals for training, 6 for validation, and 6 for testing, ensuring that the same interpreter did not appear in multiple splits.

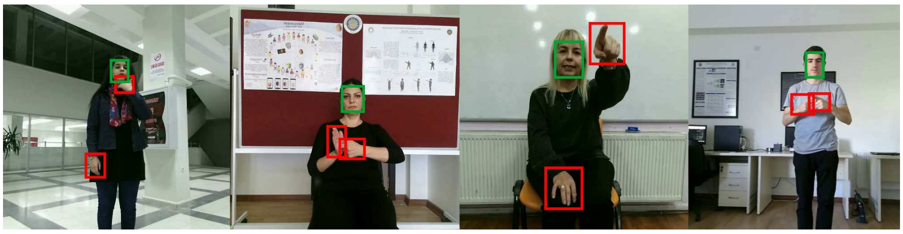

Fig. 1: Example images from the object detection dataset.

## 4 Fine-tuning of object detectors

This research compared several widely-used architectures for object detection: Single Shot multibox Detector (SSD) [12], YOLOV7 [13], Faster R-CNN [14], CenterNet [15] and EfficientDetD0 [10]. Furthermore, to adapt the aforementioned models to the context of hand and face detection for sign language, some architectural modifications were made, since the default configurations suggested by the authors were intended for general-purpose detectors, dealing with objects of various shapes, sizes, and proportions. The primary goal of these changes was to reduce the computational cost while maintaining a high mAP.

The first simplification applied to all models was to decrease the maximum number of predictions per class and generated bounding boxes from 100 to 2 and 4, respectively. This is important since there are only three objects and two classes appearing in the image simultaneously, namely, the face and both hands. By reducing the number of predictions, fewer iterations are required to remove overlapping bounding boxes. Besides that, we employed the Kmeans model for unsupervised adjustment of anchor box proportions, fitting the width and height of the figures to maximize the ratio of Intersection over Union (IoU) in the training set. By doing this, we could better adapt the anchors to the given problem domain, facilitating the model convergence. Further modifications were made individually, considering the architectural details of each object detector.

EfficientDetD0, for example, was modified by reducing the depth of the Bidirectional Feature Pyramid Network (Bi-FPN) from 5 to 4, removing the most superficial layer. In addition, we reduced the iterations of the Bi-FPN from 3 to 2, the feature map kernels from 64 to 48, and the layers of the network responsible for classification and regression from 3 to 2.

Regarding SSD, we also decreased the FPN layers from 5 to 4 and the number of layers of the bounding box predictor from 4 to 3. In contrast to the original paper, we utilized MobileNetV2 as the base network for feature extraction, reducing the model complexity without adversely affecting the results. Finally, two versions of SSD were used in this work, SSD320 and SSD640, which have input image resolutions of 320x320 and 640x640 pixels, respectively.

Concerning Faster R-CNN, we reduced the number of proposals generated by the Region Proposal Network (RPN) from 300 to 50, since there are only a few objects appearing in the image simultaneously. Also, we removed the largest anchor box scale, once the hands and face have a small and fixed size due to the minimum distance the interpreter must keep from the camera. Lastly, we used ResNet-50 as the feature extractor to further reduce the computational cost.

YoloV7, on the other hand, underwent minimal changes to avoid affecting the architecture's functionality. We decreased the number of kernels in each convolutional layer, reducing the number of parameters from 36.4 MM to only 14.1 MM. In addition, we used an input resolution of 512x512 pixels instead of the suggested 640x640 pixels in the original paper.

Finally, the only modification made to CenterNet was the replacement of the default feature extractor with MobileNetV2, as the model's architecture has limited options for problem-specific improvements.

## 5 Experimental setup

To train the object detectors, a uniform procedure was followed. Initially, we used a starting learning rate of 0.001 and a batch size of 16 for 80,000 iterations, where each iteration represents a batch of data. However, due to memory limitations, Faster R-CNN was trained with a batch size of 8 for 160,000 iterations, allowing all detectors to be subject to the same amount of images. Additionally, transfer learning was not employed as a consequence of the architectural changes, ensuring a fair comparison between all models.

Regarding data augmentation, the techniques were carefully selected to suit the specific domain and were standardized across all detectors. The geometric operations included horizontal rotation, changes in image scale, and random crop, and were optimized to avoid cropping the interpreter's hands from the image. For intensity operations, we used grayscale conversion, distortion in RGB color channels, random black patch creation, and adjustments in brightness, contrast, saturation, and hue.

To evaluate the results, three metrics were employed, the average inference time in milliseconds (ms) per image, and the mAP considering the thresholds of 50% and 75% of IoU, referred to as mAP@50 and mAP@75, respectively. The mAP scores were calculated on the test set, highlighting the performance of the models in interpreters and scenarios not previously observed. Finally, we measured the average inference time of each frame using the same hardware, consisting of an Intel Core i5 10400 CPU and an Nvidia RTX 3060 GPU.

## 6 Results

The results are summarized in Table 1, where the values in parentheses correspond to the inference time before applying any optimizations. As expected, all detectors performed satisfactorily for hand and face detection, considering the simplicity of the task, achieving an mAP above 85%, except for the SSD320 due to its lower input resolution and weaker performance for small objects.

| Architecture | Inf. time CPU | Inf. time GPU | mAP@50 | mAP@75 |
|---|---|---|---|---|
| SSD640 | 53.2 (108.0) | 11.6 (44.1) | 98.5 | 95.0 |
| SSD320 | **15.7** (32.7) | 9.9 (25.7) | 92.1 | 73.1 |
| EfficientDet D0 | 67.8 (124.5) | 16.1 (53.4) | 96.7 | 85.8 |
| YoloV7 | 123.9 (211.1) | **7.4** (7.6) | 98.6 | 95.7 |
| Faster R-CNN | 281.0 (811.5) | 26.3 (79.1) | **99.0** | 96.2 |
| CenterNet | 40.0 | 7.9 | **99.0** | **96.7** |

Table 1: Object detectors results.

EfficientDet, despite employing a base CNN with more layers and a Bi-FPN, delivered inferior results compared to SSD640. These findings suggest that complex models may not be essential for solving simpler problems. YoloV7 outperformed SSD by a small margin, demonstrating the effectiveness of the new versions of this architecture. However, it was unable to surpass Faster R-CNN, since the complexity of this model is attributed to the multiple proposals and anchor boxes generated during inference, which guarantees higher success rates, even for smaller objects. Ultimately, CenterNet proved to be the most successful approach for this task, possibly due to the sparsely populated scenes and the objects' distance from each other, allowing a higher rate of success in estimating the centroids and their boxes from the heat maps generated by the model.

In addition, we can observe that most of the models showed a significant reduction in inference time, starting with the more complex architectures such as Faster R-CNN and EfficientDet, which decreased the time on GPU by 66.75% and 69.8%, respectively. Even architectures built on SSD showed reductions of 73.69% and 61.4% for the 640 and 320 versions, respectively. Considering YoloV7, the difference was only noticeable on the CPU, decreasing the inference time by 41.3%. However, the optimized model has 61% fewer parameters, reducing the memory requirements and making it more suitable for edge devices.

## 7 Conclusion

The object detection technique is fundamental for some sign language recognition systems. That said, the HFSL dataset introduced in this research represents an important advance in this area, allowing future research to train hand and face detectors focused on common scenarios for sign language. Besides that, we demonstrated that it is possible to optimize the architectures to reduce the inference time by up to 73.69% while achieving an mAP of over 95%. This can pave the way for real-time preprocessing and deployment of those solutions on edge devices such as smartphones. In future work, we aim to increase the size of the dataset and train new object detection architectures, providing further details on their performance, such as the distribution of the model mistakes by interpreter and background.

## References

[1] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[2] Arpit Mittal, Andrew Zisserman, and Philip HS Torr. Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference*, pages 75.1–75.11. BMVA Press, 2011.

[3] Vincent Spruyt, Alessandro Ledda, and Wilfried Philips. Robust arm and hand tracking by unsupervised context learning. *Sensors*, 14(7):12023–12058, 2014.

[4] Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based graph matching for robust facial landmark localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1025–1032, 2013.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[6] Sepehr MohaimenianPour and Richard Vaughan. Hands and faces, fast: mono-camera user detection robust enough to directly control a uav in flight. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5224–5231. IEEE, 2018.

[7] Qinkun Xiao, Xin Chang, Xue Zhang, and Xing Liu. Multi-information spatial–temporal lstm fusion continuous sign language neural machine translation. *IEEE Access*, 8:216718–216728, 2020.

[8] Yanqiu Liao, Pengwen Xiong, Weidong Min, Weiqiong Min, and Jiahao Lu. Dynamic sign language recognition based on video sequence with blstm-3d residual networks. *IEEE Access*, 7:38044–38054, 2019.

[9] Shiwei Xiao, Yuchun Fang, and Lan Ni. Multi-modal sign language recognition with enhanced spatiotemporal representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[10] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020.

[11] Ozge Mercanoglu Sincan, Julio Junior, CS Jacques, Sergio Escalera, and Hacer Yalim Keles. Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3472–3481, 2021.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer International Publishing, 2016.

[13] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.

[15] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.