Efficient Bounding Box Filtering in British Sign Language Datasets

Giacomo Bonomi

Monash University, Faculty of Information Technology

1 Introduction

British Sign Language (BSL) is an essential communication tool for the deaf and hard-of-hearing, enabling effective interpersonal interaction and access to information [Nim20]. Advancements in computer vision and deep learning have lead to significant progress in developing automated sign language recognition systems. These systems leverage object detection models that identify and isolate signers in video frames using bounding boxes [MMD18]. A primary metric for evaluating the accuracy of these object detection models is the Intersection over Union (IoU) score [Now14], which measures the overlap between predicted and ground truth bounding boxes. High IoU scores indicate precise localisation, essential to subsequent translation processes in accurate interpretation of sign gestures. However, such methods face many limitations towards their effectiveness such as unsuitability in the absence of a ground truth box, or when background interference can lead to noise and possible inaccurate box isolation.[Lle22].

This study aims to develop strategies to assess the impact of background interference on bounding box accuracy when ground truth data is unavailable. Specifically, we analyse IoU scores between two segmentation models, EPP-Net-Action [Liu24] and MMDetection [MMD18], using statistical methods to understand their reliability in distinguishing relevant signers from background noise.

1.1 Research Question

"What impact does background interference have on the accuracy of bounding boxes in BSL datasets in the absence of ground truth data?"

2 Background

2.1 British Sign Language

The BBC-Oxford British Sign Language (BOBSL) dataset [Alb21] provides a comprehensive resource for evaluating sign language recognition systems. The dataset was collected over 250 annotated videos of professional signers in BBC official broadcasting, covering 1,839 unique words. The dataset includes a comprehensive annotation of each word's occurrence and their timestamps. Despite the controlled nature of the dataset, instances of background interference (additional non-signer humans, mistakenly identified objects, eft.) or noise are still frequent throughout [Figure 1].

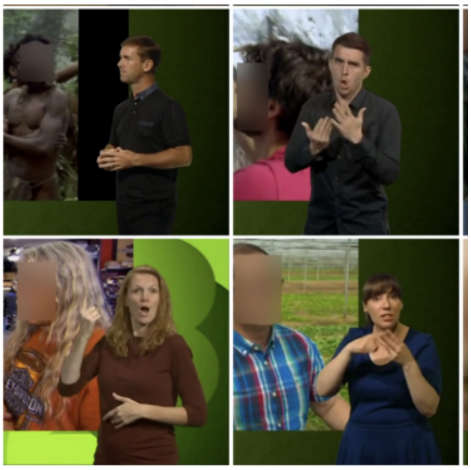

Figure 1: Possibilites for Signer Intereference

2.2 The Bounding Box

Existing research has emphasised the importance of accurate signer isolation for effective sign language recognition. Object detection models like MMDetection [MMD18] or EPP-Net-Action [Liu24] are often leveraged due to their high flexibility and reliable accuracy. These tools rely on the usage of bounding boxes [Suri99], a minimum-area rectangle used to identify

In: M. Meder, A. Rapp, T. Plumbaum, and F. Hopfgartner (eds.): Proceedings of the Data-Driven Gamification Design Workshop, Tampere, Finland, 20-September-2017, published at http://ceur-ws.org

and isolate detected objects [Figure 2]. Only the contents within the box will be ultimately processed by any computation model, meaning it's crucial the box contains and is able to filter for all relevant data.

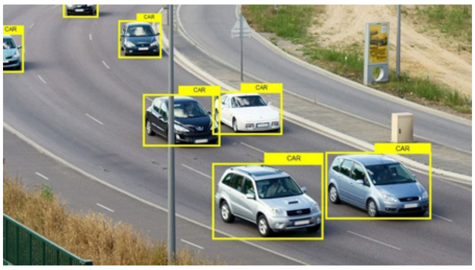

Figure 2: Bounding Boxes drawn for cars [Alfasly19].

2.3 Ground Truth Data

Intersection over Union (IoU)[Now14] is a method utilised for comparing total overlap between two bounding boxes, providing a final accuracy result as a percentage from 0-1 signifying no to total overlap respectively.. Normally, IoU computation relies on comparison between generated bounding boxes and ground truth bounding boxes to verify an IoU score as "accurate". However, in the context of emerging datasets such as for BOBSL, there is a lack of such ground truth data, rendering the verifiability of predicted bounding boxes hard to verify using traditional methods. This can pose significant issues to researchers and developers as it makes it difficult to evaluate any generated box, posing serious limitations to determining overall accuracy.

3 Methodology

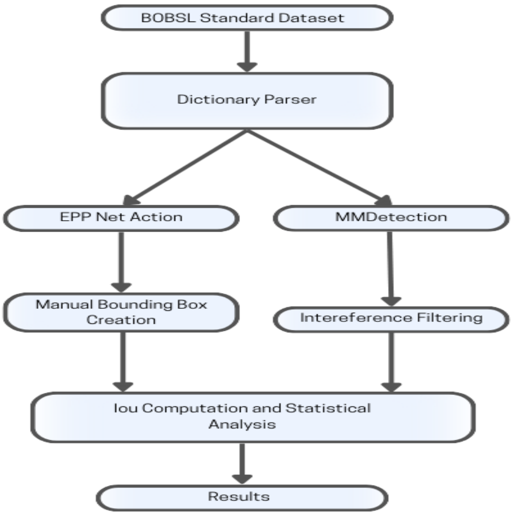

Figure 3: Abstract of Design Process

3.1 General Design Overview

An experimental process to evaluate bounding box accuracy without ground truth data across models was developed [Figure 3]. This model seeks to compare the commonly used model's output against known instances of signage in order to determine their accuracy for specific words, videos, and across a dataset. To achieve this, analysis methods such as IoU computation and box filtering and normalisation were utilised to obtain data to be used for statistical analysis.

3.2 Model Selections

Two object detection models were integrated into this study:

- **EPP-Net-Action**: An open-source human-specific object detection model with a custom pipeline provided by Dr. Kunyuan Xie. [Figure 4] [Now14]
- **MMDetection**: An open-source object detection toolbox that has a more varied detection capability and homemade extraction pipeline. [Figure 4] [MMD18].

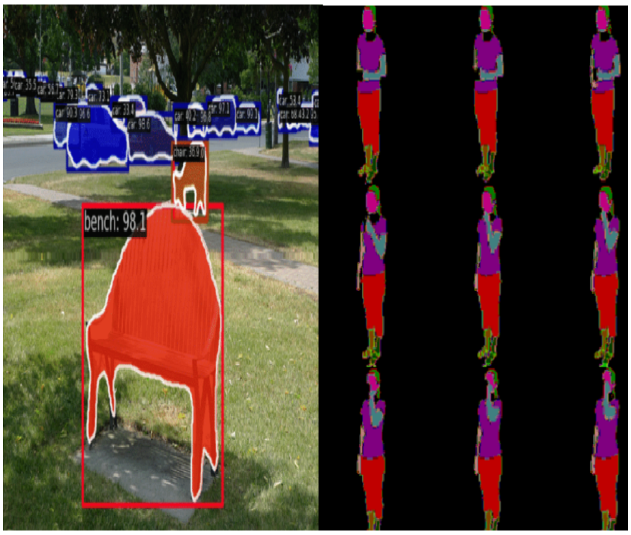

Figure 4: Comparison of bounding box outputs from MMDetection (left) and EPP-Net-Action (right) for a sample frame.

These models were chosen due to their flexibility and widespread usage, specifically in the field of signage translation [Feng21], as well as for key functionalities discussed later.

3.3 Dataset Preparation

The BOBSL dataset contains information for each word detected across all videos. This information contains a list of each video in which the word was signed, as well as the specific frame instances in which this occurred. In order to effectively utilise this data, we implemented a parser to read and store it for future use. Following standard collection practices outlined in the BOBSL paper, the 14 frames before, the 1 target frame, and the 3 frames after [Alb21] were noted and stored for word each occurrence. These frame indexes and video names were stored as numpy [Wal11] files for future reference.

3.4 Pipeline and Segmentation

A custom pipeline was written to parse the numpy files stored above into corresponding frames for each model. We implemented a pipeline usable for both methods that can provide relevant frames given video names and frame indices as a numpy file.

EPP-Net-Action and MMDetection were modified to accept from this pipeline. Therefore, when we wanted our models to conduct analysis, we simply had to activate the pipeline so that it could parse all the numpy files, and thus the corresponding video frames.

3.5 Segmentation Results Handling

The different models' outputs differed and required specific handling.

- **MMDetection**: Segmentation results were stored as a JSON file a sorted-by-confidence list of bounding boxes with their respective coordinates.

- **EPP-Net-Action**: EPP-Net-Action only ever outputs one bounding box. It outputs an array representing the outline of the body of the signer. This array's cells would correspond to a digit, with a digit representing a human's specific body part.

3.5.1 Manual Box Creation and Filtering

In MMDetection, if the highest confidence frame did not have the label ”0” for humans, this was noted. Later in the analysis, this specific bounding box would be excluded in calculating the final IoU for the word, instead using the highest confidence box with a human label found in the JSON file.

For EPP-Net-Action, we developed a manual bounding box creator. This would parse the segmentation result array and look for the smallest possible bound that could be made. This new bounding box's coordinates were then normalised back to the full size of the frame. Therefore for the sake of IoU computation, EPP-Net-Action's output was used to supplement the ground truth box due to its ability for more strict and singular bounding box creation.

3.6 Analysis Techniques

3.6.1 IoU as a Metric

IoU is a formula used to determine the overlap between a predicted bounding box and the ground truth box, measuring how well the two align. It provides a value between 0 and 1, where 1 indicates a perfect overlap and 0 indicates none at all.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Where A, or our MMDetection result, and B, our EPP-Net-Action one, are the predicted and ground truth boxes respectively.

3.6.2 IoU Computation

Using this formula, an IoU for each frame of each output can be computed across models [Figure 5]. This was done for each frame, with word, video, and total IoU averages being noted and stored.

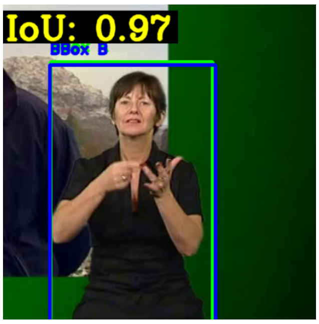

Figure 5: An IoU Score for a specific Frame.

3.6.3 Non-Human or Wrong Boxes

Instances of MMDet results with boxes of the highest confidence being non-human were handled for IoU computations. Percentages of such instances were saved for each word, video, and overall. Additionally, in circumstances where bounding boxes would have little-to-no overlap (IoU score of below 0.3), the instance would be saved for manual analysis.

3.7 Statistical Correlation and Regression

To support our findings, various statistical analyses were conducted. Pearson correlation coefficients were calculated to assess the relationship between Non-Human Frame Percentage and Average IoU scores at both the word and video levels.

Linear regression analysis was used to model the effect of Non-Human Frame Percentage on Average IoU, quantifying how background interference impacts detection accuracy.

ANOVA (Analysis of Variance) tested for significant differences in IoU scores across different words to reveal variability in model performance.

Analyses were executed using Python libraries like SciPy and sci-kit-learn, with visualisations produced via Matplotlib and Seaborn.

4 Conclusion

4.1 Summary of Statistical Results

The evaluation of bounding box accuracy in British Sign Language (BSL) datasets demonstrated that EPP-Net-Action and MMDetection models achieved high overall Intersection over Union (IoU) scores, with an average IoU of approximately 0.8491 and a median IoU of 0.8566. These results indicate that the models effectively isolate signers from background elements. The standard deviation values (0.0987 for overall IoU and 0.0545 for per-video IoU) suggest a relatively consistent performance, although some variability exists, particularly across different videos [Figure 6].

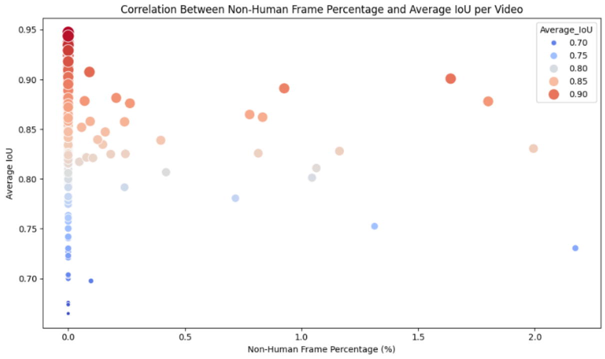

Figure 6: Non-Human Frame Percentage against IoU score

4.2 Clustering and Correlation Insights

Clustering analysis revealed three distinct groups based on IoU scores and non-human frame percentages. Cluster 0 comprised videos with low non-human frame percentages and high IoU scores, indicating optimal detection performance in cleaner environments. In contrast, Cluster 2 included videos with higher non-human frame percentages (up to 2.24%) and slightly reduced IoU scores (average around 0.8274), highlighting the impact of increased background interference [Figure 7]. Correlation analysis showed a minor negative relationship between non-human frame content and IoU scores at the video level (r = -0.1008). However, the low R-squared value (0.0102) suggests that non-human frames account for only a small portion of the variance in IoU scores. Additionally, ANOVA results indicated no statistically significant differences in IoU scores across the identified clusters. This indicates that although present, grouping for non-human interference is unlikely to identify statistically significant score groupings.

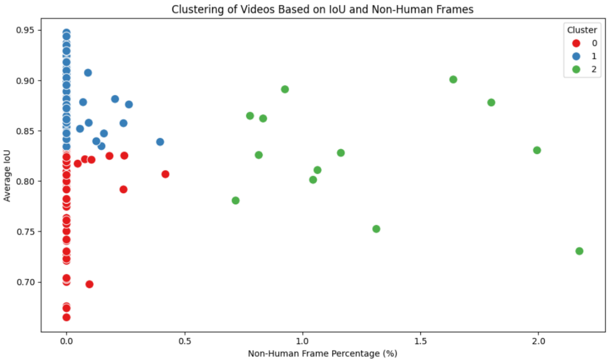

Figure 7: Video Clustering Based on Non-Human Percentage Frames

The correlation analysis supported these findings, showing a minor negative relationship between non-human frame content and IoU scores (r = -0.1008 at the video level). Despite this, the small R^2 value (0.0102) indicated that non-human content alone explained only a minimal part of the variance in IoU scores [Figure 8]

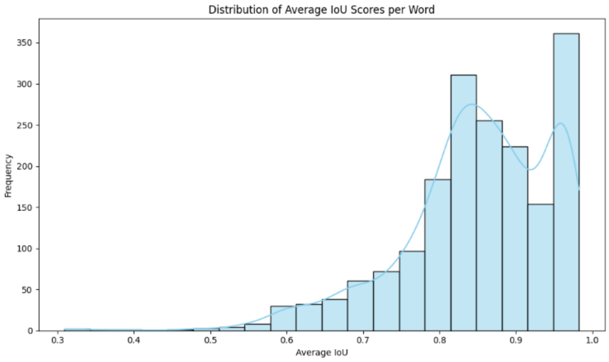

Figure 8: Average IoU score per word distribution

4.3 Implications and Real-World Application

The findings highlight that while background interference can have a negative effect on detection accuracy, the impact appears marginal. This suggests that current state-of-the-art models like EPP-Net-Action and MMDetection can still perform reliably in real-world scenarios with moderate levels of background activity.

4.4 Limitations and Future Directions

4.4.1 Metric and Model Scope

This study focused solely on IoU as the primary metric for evaluating detection performance. While it provides a strong indication of bounding box accuracy, it does not capture other performance aspects that might be critical to accuracy. Expanding the evaluation to include other metrics and models would provide a more comprehensive understanding of detection efficacy.

4.4.2 Case by Case Analysis

Although our current model can detect the performance of specific words or videos, it is unable to identify key causes behind this variation beyond just the interference. Further testing with low-performing words or videos should seek to identify the variables, lighting, angle, etc. outside the scope of our current model that could be impacting these individual cases.

4.4.3 Bounding Box Ground Truth

The absence of ground truth bounding boxes for comparison poses a limitation. Using one model's output as a reference may introduce bias. Further studies could integrate human-annotated ground truth data to validate and refine the findings. Alternatively, the usage of more models to refine our notion of a 'Ground Truth' box can be helpful in mitigating the bias in our current limited model set.

5 Conclusion and Future Prospects

The use of these models as tools to measure the impact of non-human interference on detection accuracy is partially valid. While they can effectively detect general trends and variances, the limitations outlined (e.g., absence of ground truth bounding boxes and other environmental factors) suggest that caution should be exercised when using them for definitive analysis. The findings are promising but highlight that these models alone cannot comprehensively determine the full impact of background noise without supplementary data or more robust comparative benchmarks.

In conclusion, while EPP-Net-Action and MMDetection provide a solid starting point for understanding background interference in BSL detection, more comprehensive analysis frameworks are necessary to draw conclusive insights and guide future enhancements in detection models.

6 Acknowledgements

- **Dr. Kalin Stefanov and Dr. Kunyuan Xie**: For their continued support, guidance, and mentorship throughout the research process.
- **MASSIVE and Monash University**: For providing access to the M3-Monash Bioinformatics Plaftorm Server.
- **Dr.Kunyuan Xie**: For granting access to a pretrained and developed EPP-Net-Action Segmentation Model.
- **Mahsa Salehi, Chief Examiner of FIT2082, and Monash University**: For approving and supporting the undertaking of this research.

References

[Nim20] KP Nimisha and Agnes Jacob. A Brief Review of the Recent Trends in Sign Language Recognition. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 186–190, 2020. DOI: 10.1109/ICCSP48568.2020.9182351.

[Liu24] J. Liu, R. Ding, Y. Wen, N. Dai, F. Meng, S. Zhao, and M. Liu. Explore Human Parsing Modality for Action Recognition. *arXiv preprint arXiv:2401.02138*, 2024. https://arxiv.org/abs/2401.02138.

[MMD18] MMDetection Contributors. OpenMMLab Detection Toolbox and Benchmark. https://github.com/open-mmlab/mmdetection, 2018. Licensed under Apache-2.0.

[Alb21] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BBC-Oxford British Sign Language Dataset. https://arxiv.org/abs/2111.03635, 2021.

[Suri99] S. Suri, P. M. Hubbard, and J. F. Hughes. Analysing bounding boxes for object intersection. *ACM Trans. Graph.*, 18(3):257–277, July 1999. Association for Computing Machinery, New York, NY, USA. DOI: 10.1145/336414.336423.

[Now14] S. Nowozin. Optimal Decisions from Probabilistic Models: The Intersection-over-Union Case. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[Lle22] J. Llerena, L. Kirsten, and C. Jung. Can we trust bounding box annotations for object detection? In *Proceedings of the CVPR Workshop*, June 2022. DOI: 10.1109/CVPRW56347.2022.00528.

[Alfasly19] S. Alfasly, B. Liu, Y. Hu, Y. Wang, and C.-T. Li. Auto-Zooming CNN-Based Framework for Real-Time Pedestrian Detection in Outdoor Surveillance Videos. *IEEE Access*, 7:105816–105826, August 2019. DOI: 10.1109/ACCESS.2019.2931915.

[Wal11] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, 2011. DOI: 10.1109/MCSE.2011.37.

[Feng21] Y. Feng, X. Wu, X. Yin, J. Jia, and Y. Fu. Sign language recognition with multiple large-scale datasets and adaptive multi-modal fusion. *arXiv preprint arXiv:2111.03635*, 2021. https://arxiv.org/abs/2111.03635.