Evaluating modern computer vision techniques for Shape Language classification in meetings

Author: Sorana Stan

Supervising team: Stephanie Tan, Edgar Salas Girones

Automatic understanding of meetings and negotiations

1. Background

1.1 The Shape Language [1]

- Is a system of geometric shapes (e.g., spheres, cubes, pyramids)
- Designed to enhance collaboration and represent abstract ideas.

1.2 Limitations of Computer Vision tools

- Limited exploration in specialized contexts like human-object interactions in negotiations.
- Lack of comparison analysis between different models in these scenarios.

Figure 1: Example of a frame from the dataset containing the Shape Language objects

1.3 The objective

Assess the models' ability to recognize and classify Shape Language objects to improve collaborative tools in organizational contexts.

2. Research question

How well do modern computer vision models perform in recognizing and classifying Shape Language objects during meetings?

3. Related literature

Four models were chosen for this study: **Yolov8, SSD, RCNN and RetinaNet**.

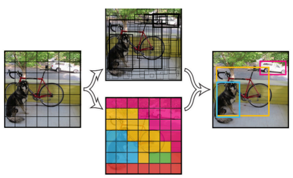

Figure 2: Yolo architecture [2]

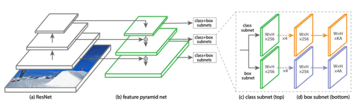

Figure 3: RetinaNet architecture [3]

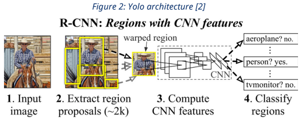

Figure 4: RCNN architecture [4]

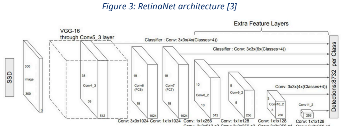

Figure 5: SSD architecture [5]

YOLOv8
- Single-stage, very fast, real-time suitable.
- Balances accuracy and speed.
- Issues with localization in earlier versions.

RetinaNet
- Single-stage, high accuracy, strong for imbalanced datasets (uses FPN).
- Moderate speed, slower than YOLO/SSD.

RCNN
- Two-stage, high accuracy for complex objects.
- Slow and computationally expensive.
- Not real-time capable.

SSD
- Single-stage, fast, real-time capable.
- Moderate accuracy, struggles with small objects.
- Simpler architecture than the other 3 models.

4. Methodology

4.1. Frame annotation: Using Farneback Optical Flow

- Selects frames with substantial motion, minimizing redundant or static frames.

4.2 Fine-tuning the target models: Yolov8, SSD, RCNN and RetinaNet.

- Result: bounding box predictions and class labels, along with confidence scores for each label.

4.3 Evaluating the models

- Result: The predictions of the test set, confusion matrix, F1 score, precision-confidence curve, recall-confidence curve, and precision-recall curve

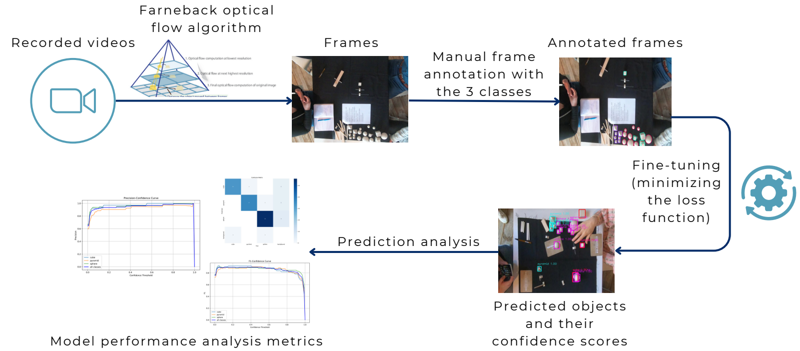

Figure 6 : High-level workflow pipeline [6]

5. Results

Confusion matrices

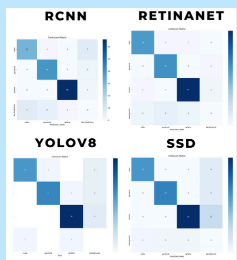

Figure 7: The confusion matrices showing the performance of the 4 models on the test set

F1 confidence

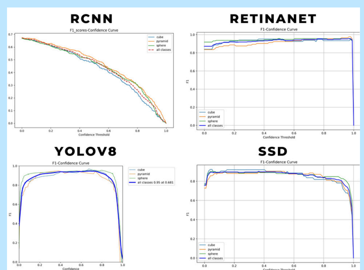

Figure 8 The F1-confidence scores on the 4 models: Higher F1 scores indicate better balance between precision (accuracy of positive predictions) and recall

Precision-confidence

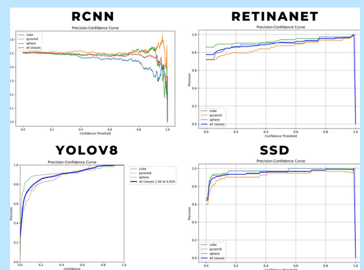

Figure 6: The precision-confidence scores on the 4 models: Higher precision at a broader range of confidence thresholds indicates fewer false positives.

Recall-confidence

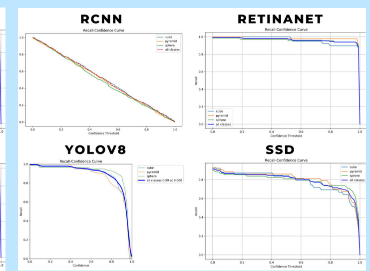

Figure 8: The recall-confidence scores on the 4 models: models with high recall across thresholds are better at detecting positive instances.

- **RetinaNet:** Best overall performance with high F1 scores, precision, and recall across all confidence thresholds and object classes.
- **YOLOv8:** Strong precision-recall metrics, achieving near-perfect precision at moderate thresholds, but recall drops sharply at extreme thresholds.
- **SSD:** Lower precision and recall than the other models, especially for shapes like spheres and pyramids, reducing reliability.
- **RCNN:** Weakest performance, with significant variability in precision and recall, struggling with accurate detection of smaller or less prominent objects.

Overall perfromance conclusion: Overall, RetinaNet emerged as the most balanced and robust model, followed by YOLOv8, while SSD and RCNN had some limitations in handling this custom dataset

6. Limitations and future work

- **Small Dataset**
- **Computational Constraints**
- **High Memory Usage**
- **Dealing with occlusions**
- **Semi-supervised and active learning**

7. Conclusions

- **Objective:** Assess the performance of YOLOv8, SSD, RCNN, and RetinaNet in recognizing and categorizing Shape Language objects in meeting scenarios.
- **Key Findings:**
  1. RetinaNet outperformed all models in precision, recall, and F1 score
  2. Small dataset reduced generalizability of conclusions.
  3. Computational constraints limited hyperparameter tuning and deeper evaluations.

8. References

1. Vormtaal. Vormtaal website. https://www.vormtaal.com/, 2025. Accessed: 2025-01-23.
2. J Redmon. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
3. T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In proceedings of the IEEE conference on computer vision and pattern recognition, pages 2980–2988, 2017
4. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.
5. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016
6. Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, pages 363–370. Springer, 2003