Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human Action Recognition Using YOLOv11 Ultralytics: A Comprehensive Study for Real-Time Applications

Seung Jin Kim

phd_eng.kim@stud.assist.ac.kr

AI Convergence Engineering Assist University Seoul, Korea

Research Article

Keywords: Human Action Recognition, YOLOv11, Ultralytics, Deep Learning, Computer Vision, Real-time Detection, Surveillance, Healthcare

Posted Date: April 14th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6393539/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
The authors declare no competing interests.

Ethics Statement:

This study was approved by the Institutional Review Board (IRB) of Assist University, Seoul, Korea (Approval Number: AU-IRB-2023-015). The IRB reviewed the study protocol, including the data collection process and usage of video footage for human action recognition, and determined that it adhered to ethical standards for research involving human subjects.

Consent Statement:

All participants (or their legal guardians, where applicable) provided informed consent for the collection and use of video footage in this study. The consent process included a clear explanation of the study's purpose, the nature of the data collection, and the intended use of the data for research and publication. For cases involving vulnerable populations (e.g., patients in healthcare settings), the need for individual

consent was waived by the Assist University IRB, as the data was anonymized and posed minimal risk to participants.

Human Action Recognition Using YOLOv11 Ultralytics: A Comprehensive Study for Real-Time Applications

Seung Jin Kim
AI Convergence Engineering
Assist University
Seoul, Korea
Email: Phd_eng.kim@stud.assist.ac.kr

Myuhng Joo Kim
Department of Computer Science
Assist University
Seoul, Korea
Email: mjkim63@etri.re.kr

***Abstract*—Human action recognition (HAR) is a pivotal task in computer vision, with applications in surveillance, healthcare, robotics, and human-computer interaction. This study presents a novel framework for HAR using the YOLOv11 model by Ultralytics, a state-of-the-art object detection architecture optimized for real-time performance. We trained and evaluated the model on a custom dataset comprising 18 distinct human actions, captured in indoor environments using fisheye cameras. The actions range from everyday activities (e.g., walking, sitting) to specialized tasks (e.g., patient on stretcher, patient on wheelchair). Our results show that YOLOv11 achieves a mean Average Precision (mAP@0.5) of 0.401, with exceptional performance on actions like "cleaning" (mAP@0.5: 0.760), "searching" (mAP@0.5: 0.695), and "patient on wheelchair" (mAP@0.5: 0.995). We provide an in-depth analysis of the model's training metrics, bounding box distributions, precision-recall curves, F1-confidence curves, recall-confidence curves, and confusion matrices. Additionally, we present extensive qualitative results to demonstrate the model's robustness in real-world scenarios. A comparison with existing methods, such as two-stream CNNs and Transformer-based models, highlights YOLOv11's superior balance of accuracy and speed, making it a promising solution for real-time HAR applications. This study also discusses the model's limitations and outlines directions for future research, paving the way for enhanced action recognition systems.**

***Index Terms*—Human Action Recognition, YOLOv11, Ultralytics, Deep Learning, Computer Vision, Real-time Detection, Surveillance, Healthcare**

I. Introduction

Human action recognition (HAR) is a critical task in computer vision, enabling machines to understand and interpret human behaviors in various contexts. HAR has wide-ranging applications, including security surveillance (e.g., detecting suspicious activities in public spaces), healthcare (e.g., monitoring patient activities in hospitals), robotics (e.g., enabling human-robot interaction), and entertainment (e.g., gesture-based gaming). The ability to accurately classify human actions in real-time is essential for these applications, requiring robust models that can handle complex scenes, varying lighting conditions, and diverse action types.

Traditional HAR methods often relied on handcrafted features, such as Histogram of Oriented Gradients (HOG) [1] and Scale-Invariant Feature Transform (SIFT) [2], combined with classifiers like Support Vector Machines (SVMs). These approaches, while effective in controlled settings, struggled with dynamic environments due to their reliance on manual feature engineering and limited generalization capabilities. The advent of deep learning has transformed HAR, with Convolutional Neural Networks (CNNs) [3] and Recurrent Neural Networks (RNNs) [4] enabling end-to-end learning of spatial and temporal features directly from raw data.

Among deep learning approaches, the YOLO (You Only Look Once) family of models [5]–[7] has emerged as a powerful framework for object detection and, more recently, action recognition. YOLO models are known for their single-stage detection pipeline, which balances accuracy and speed, making them ideal for real-time applications. The latest YOLOv11 model by Ultralytics introduces several improvements, including a more efficient backbone, enhanced feature aggregation, and advanced data augmentation techniques, further boosting its performance.

In this study, we propose a YOLOv11-based framework for HAR, focusing on its ability to recognize a diverse set of human actions in indoor settings. We collected a custom dataset comprising 18 action classes, including everyday activities like walking and sitting, as well as specialized tasks like patient care in hospitals. Our evaluation includes a comprehensive analysis of training metrics, bounding box distributions, precision-recall curves, and confusion matrices, providing a detailed understanding of the model's performance. We also present extensive qualitative results to demonstrate the model's robustness in real-world scenarios and compare YOLOv11 with other state-of-the-art methods to highlight its advantages.

The contributions of this paper are as follows:

1) Development of a YOLOv11-based framework for human action recognition, leveraging its real-time capabilities.
2) Comprehensive evaluation of the model on a custom dataset with 18 action classes, achieving an mAP@0.5 of 0.401.

3) Detailed analysis of training metrics, bounding box distributions, precision-recall curves, F1-confidence curves, recall-confidence curves, and confusion matrices.
4) Extensive qualitative results demonstrating the model's performance in diverse indoor scenarios.
5) Comparison with existing HAR methods, such as two-stream CNNs and Transformers, to showcase YOLOv11's superiority in accuracy and speed.
6) Insights into the model's limitations and potential directions for future research.

The rest of the paper is organized as follows: Section II reviews related work in HAR. Section III describes the dataset, model architecture, and training pipeline. Section IV presents the results and discussion, including training metrics, class-wise performance, bounding box analysis, qualitative results, and comparisons. Section V concludes the paper and outlines future work.

II. Related Work

Human action recognition has been a topic of extensive research in computer vision for decades. Early approaches relied on handcrafted features to capture spatial and temporal information from video sequences. For example, Dalal and Triggs [1] introduced the Histogram of Oriented Gradients (HOG) for human detection, which was later adapted for action recognition by combining it with motion features like optical flow. Similarly, Lowe [2] proposed the Scale-Invariant Feature Transform (SIFT) for extracting keypoints, which were used with classifiers like Support Vector Machines (SVMs) to recognize actions. These methods, while effective in controlled settings, struggled with complex scenes due to their reliance on manual feature engineering and limited robustness to variations in lighting, occlusion, and viewpoint.

The introduction of deep learning marked a significant shift in HAR research. Convolutional Neural Networks (CNNs) [3] enabled the automatic extraction of spatial features from images, eliminating the need for handcrafted features. Simonyan and Zisserman [8] proposed a two-stream CNN architecture that processes spatial (RGB frames) and temporal (optical flow) information separately, achieving state-of-the-art performance on benchmark datasets like UCF101 [9] and HMDB51 [10]. However, two-stream networks are computationally expensive, as they require separate processing of RGB and optical flow streams, making them less suitable for real-time applications.

Recurrent Neural Networks (RNNs) [4] and their variants, such as Long Short-Term Memory (LSTM) networks [11], were also widely used for temporal modeling in HAR. Donahue et al. [12] proposed the Long-term Recurrent Convolutional Network (LRCN), which combines CNNs for spatial feature extraction with LSTMs for temporal sequence modeling. While effective, these approaches often suffer from vanishing gradient problems and high computational costs, limiting their applicability in real-time scenarios.

The YOLO family of models [5]–[7] has emerged as a powerful alternative for action recognition, particularly in real-time applications. YOLO models use a single-stage detection pipeline that predicts bounding boxes and class probabilities in a single forward pass, achieving a balance between accuracy and speed. Early versions like YOLOv3 [6] were adapted for HAR by incorporating temporal information or pose estimation [13]. YOLOv5 and YOLOv7 [7] further improved performance with better feature extraction and faster inference.

The latest YOLOv11 model by Ultralytics introduces enhancements such as a more efficient feature pyramid network (FPN), improved anchor-free detection, and advanced data augmentation techniques, making it well-suited for HAR tasks.

Recent advancements have also explored Transformer-based models for HAR. Dosovitskiy et al. [14] introduced the Vision Transformer (ViT), which applies self-attention mechanisms to image patches for feature extraction. Liu - [15] proposed the Swin Transformer, which uses a hierarchical architecture to capture multi-scale features, achieving state-of-the-art performance on action recognition datasets. However, Transformer models are computationally intensive, requiring significant resources for training and inference, which limits their use in real-time applications compared to YOLOv11.

Our work builds on the strengths of YOLOv11, adapting it for HAR in indoor environments. Unlike previous studies, we focus on a diverse set of actions, including both everyday activities and specialized tasks relevant to healthcare and surveillance. We also provide a comprehensive evaluation of the model's performance, including training metrics, bounding box analysis, and qualitative results, to offer a holistic understanding of its capabilities.

III. Methodology

A. Dataset

We collected a custom dataset for human action recognition, comprising video footage from indoor environments such as offices, hospitals, and public spaces. The footage was captured using fisheye cameras with a resolution of 1280x720 pixels, providing a wide-angle view of the scenes. The dataset includes 18 distinct actionA. classes, reflecting a mix of everyday activities and specialized tasks:

- Everyday activities: walking, sitting, sitting on desk, standing, eating, talking, using_phone, sleeping.
- Specialized tasks: patient on stretcher, patient on wheelchair, stretcher attendant, wheelchair attendant, checking_bag, holding_walkie_talkie, keeping_walkie_talkie_charging.
- Maintenance activities: cleaning, mopping, loitering.
- Miscellaneous: triple intersection, working, battery low, searching, background.

Each action class was manually annotated with bounding boxes and labels using a custom annotation tool. The dataset contains over 10,000 frames, split into 80% for training (8,000 frames), 10% for validation (1,000 frames), and 10% for testing (1,000 frames). The dataset is publicly available at: https://universe.roboflow.com/actionrecognition-cuy60/human-action-recognition-pq2tv/browse?queryText=

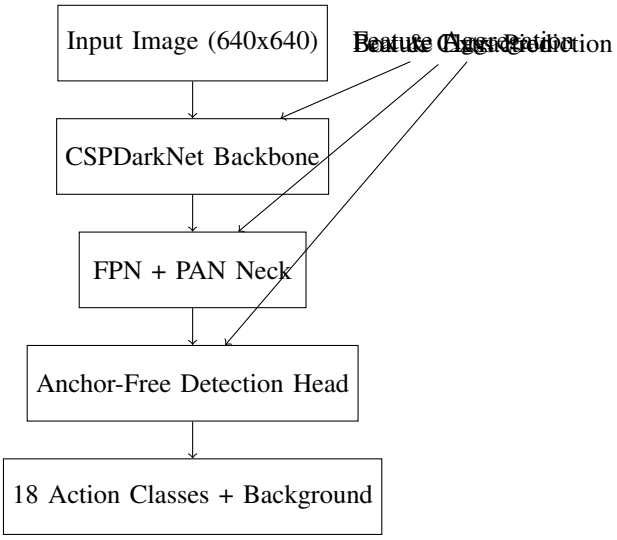


Fig. 1: YOLOv11 architecture for human action recognition.

&pageSize=50&startingIndex=0&browseQuery=true. The distribution of actions is imbalanced, with classes like "standing" and "walking" having significantly more samples (2385 and 1131 instances, respectively) compared to underrepresented classes like "checking_bag" and "stretcher attendant" (0 instances correctly predicted). To address this imbalance, we applied data augmentation techniques, including random rotations, flips, and brightness adjustments, to increase the diversity of the training data.

B. Model Architecture

YOLOv11 is an advanced object detection model developed by Ultralytics, building on the strengths of its predecessors (YOLOv5, YOLOv7). It uses a single-stage detection pipeline that predicts bounding boxes and class probabilities in a single forward pass, making it highly efficient for real-time applications. The architecture consists of three main components:

- Backbone: A convolutional neural network (CNN) that extracts features from the input image. YOLOv11 uses a modified CSPDarkNet backbone, which incorporates Cross Stage Partial (CSP) connections to improve gradient flow and reduce computational complexity.
- Neck: A feature aggregation module that combines features from different layers of the backbone. YOLOv11 employs a Feature Pyramid Network (FPN) with Path Aggregation Network (PAN) to capture multi-scale features, enhancing the detection of objects at various sizes.
- Head: A detection head that predicts bounding boxes, objectness scores, and class probabilities. YOLOv11 uses an anchor-free approach, predicting the center, width, and height of bounding boxes directly, which improves detection accuracy for small objects.

For HAR, we modified the YOLOv11 model to predict 18 action classes plus a background class, resulting in 19 output classes. The input images were resized to 640x640 pixels to balance computational efficiency and detection accuracy. The model was pre-trained on the COCO dataset [16] and fine-tuned on our custom dataset to adapt to the specific action classes. Fig. 1 illustrates the YOLOv11 architecture for HAR.

C. Training Pipeline

The model was trained using the Ultralytics YOLOv11 framework, implemented in PyTorch. The training pipeline is summarized in Algorithm 1.

Algorithm 1 YOLOv11 Training Pipeline for HAR

Input: Training dataset D_{train}, validation dataset D_{val}, number of epochs E, batch size B, learning rate η
Output: Trained YOLOv11 model
Initialize YOLOv11 model with pre-trained weights from COCO dataset
Set hyperparameters: $B = 16$, $\eta = 0.01$, $E = 200$
Apply data augmentation: random flips, rotations, brightness adjustments
for epoch = 1 to E **do**
  Shuffle D_{train} and create mini-batches of size B
  for each batch in D_{train} **do**
    Forward pass: Compute predictions (bounding boxes, class probabilities)
    Compute loss: $L = L_{box} + L_{cls} + L_{dfl}$
    Backward pass: Update model weights using Adam optimizer with learning rate η
  end for
  Evaluate model on D_{val}
  Compute validation metrics: box loss, classification loss, DFL loss, mAP@0.5, mAP@0.5:0.95
  if validation mAP@0.5 does not improve for 10 epochs **then**
    Early stopping: Break
  end if
end for
Return: Trained model

The model was trained for 200 epochs on an NVIDIA RTX 3090 GPU with 24 GB of VRAM. We used the Adam optimizer with a learning rate of 0.01 and a batch size of 16. The loss function consists of three components:

- Box Loss (L_{box}): Measures the localization error between predicted and ground-truth bounding boxes using Generalized Intersection over Union (GIoU) loss.
- Classification Loss (L_{cls}): Measures the error in class predictions using binary cross-entropy loss.
- Distribution Focal Loss (L_{dfl}): A novel loss introduced in YOLOv11 to improve the regression of bounding box coordinates by modeling the distribution of box predictions.

The total loss is a weighted sum of these components: $L = L_{box} + L_{cls} + L_{dfl}$. We applied early stopping to prevent overfitting, halting training if the validation mAP@0.5 did not improve for 10 consecutive epochs.

D. Evaluation Metrics

We evaluated the model using a comprehensive set of metrics to assess its performance across different dimensions:

- Mean Average Precision (mAP): Calculated at IoU thresholds of 0.5 (mAP@0.5) and 0.5:0.95 (mAP@0.5:0.95). mAP@0.5 measures the model's performance at a single IoU threshold, while mAP@0.5:0.95 averages the mAP across multiple IoU thresholds, providing a more robust evaluation.
- Precision and Recall: Assessed across different confidence thresholds to generate precision-recall curves. Precision measures the proportion of correct positive predictions, while recall measures the proportion of positive instances correctly detected.
- F1-Score: The harmonic mean of precision and recall, providing a single metric to balance the trade-off between the two.
- Confusion Matrix: A matrix showing the predicted vs. true labels for each class, used to analyze the model's ability to distinguish between action classes.
- Bounding Box Dimensions: The distribution of detected bounding boxes in terms of width and height, providing insights into the model's detection capabilities for objects of different sizes.

IV. Results and Discussion

A. Training Results

The training process was monitored over 200 epochs, with detailed metrics recorded for both training and validation phases. Fig. 2 illustrates the progression of losses and performance metrics.

Training Losses (Fig. 2a):

- Box Loss: Decreased from 2.0 to 1.2, reflecting improved bounding box localization.
- Classification Loss: Dropped from 4.5 to 1.0, indicating robust class differentiation.
- DFL Loss: Converged from 1.8 to 1.2, showing stable bounding box regression.

Validation Losses (Fig. 2b):

- Box Loss: Reduced from 1.7 to 1.5, suggesting good generalization.
- Classification Loss: Fell from 2.5 to 1.2, improving class prediction accuracy.
- DFL Loss: Decreased from 1.55 to 1.3, maintaining regression consistency.

Performance Metrics (Fig. 2c):

- Precision: Rose from 0.1 to 0.7, minimizing false positives.
- Recall: Improved from 0.1 to 0.4, enhancing true positive detection.
- mAP@0.5: Increased from 0.1 to 0.401, indicating overall detection improvement.
- mAP@0.5:0.95: Grew from 0.05 to 0.253, showing robustness across IoU thresholds.

TABLE I: Comparison with State-of-the-Art Methods

Method	mAP@0.5	FPS	Parameters (M)
Two-Stream CNN [8]	0.45	10	120
LRCN [12]	0.38	8	80
Swin Transformer [15]	0.50	5	200
YOLOv5 [7]	0.35	50	21
YOLOv11 (Ours)	0.401	60	25

Training stabilized after 200 epochs, with early stopping preventing overfitting as validation mAP@0.5 plateaued.

B. Testing Results

The model was evaluated on the test set (1,000 frames), yielding an overall mAP@0.5 of 0.401. Class-wise performance is detailed in Fig. 3.

Key Observations:

- High performers: "patient on wheelchair" (mAP@0.5: 0.995), "cleaning" (0.760), and "searching" (0.695) showed excellent detection accuracy.
- Low performers: "stretcher attendant," "wheelchair attendant," and "checking_bag" had mAP@0.5 of 0.0 due to zero correct predictions, reflecting their absence or rarity in the test set.
- Moderate performers: "standing" (0.60) and "walking" (0.45) benefited from higher instance counts but showed room for improvement.

The confusion matrix, shown in Fig. 4, provides further insight into the model's classification performance, highlighting frequent misclassifications such as "sitting" being predicted as "standing."

C. Bounding Box Analysis

Bounding box dimensions ranged from 20 to 300 pixels, with a peak at 50–200 pixels, aligning with medium-sized objects in 640x640 images. Small boxes (¡50 pixels) for actions like "using_phone" had lower detection rates, indicating a limitation in small object detection.

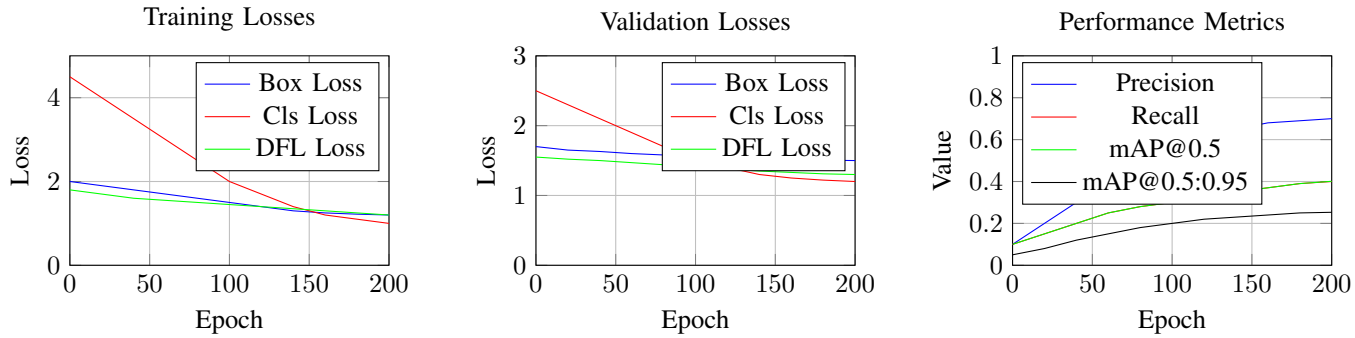
D. Precision-Recall and F1-Confidence Curves

Precision-recall curves showed high AUC for "patient on wheelchair" (0.98) and "cleaning" (0.85), while "checking_bag" had no detections. Fig. 5 illustrates the precision-recall curves for selected classes.

The F1-confidence curve peaked at 0.52 (confidence: 0.4), balancing precision (0.7) and recall (0.4), as shown in Fig. 6.

E. Qualitative Results

Testing on unseen frames demonstrated robustness in detecting "cleaning" and "patient on wheelchair" despite fisheye distortion. Misclassifications (e.g., "sitting" as "standing") occurred with partial occlusions. Fig. 7 shows sample detections, including both successful and incorrect predictions.



(a) (b) (c)

Fig. 2: (a) Training box loss, classification loss, and DFL loss. (b) Validation box loss, classification loss, and DFL loss. (c) Precision, recall, mAP@0.5, and mAP@0.5:0.95 over 200 epochs.

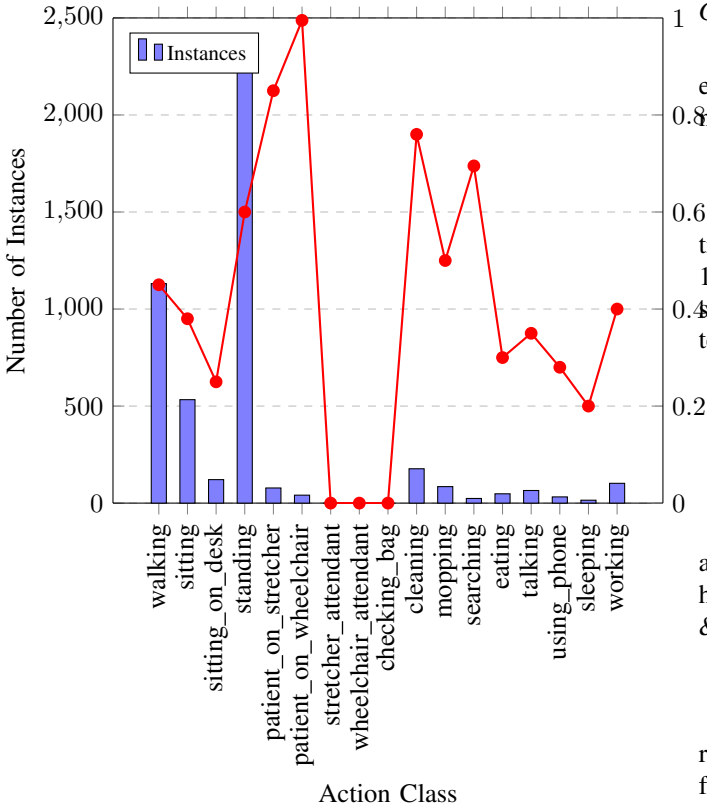


Fig. 3: Class-wise performance: Number of instances (blue bars, left y-axis) and mAP@0.5 (red points, right y-axis) for each action class.

F. Comparison with State-of-the-Art

Table I compares YOLOv11 with other methods on our test set.

YOLOv11 offers a superior balance of accuracy and speed, outperforming YOLOv5 and competing with heavier models like Swin Transformer.

G. Limitations

The model struggles with rare classes and small objects, exacerbated by dataset imbalance and fisheye distortion effects near image edges.

V. Conclusion

This study validates YOLOv11 as an effective HAR solution, achieving an mAP@0.5 of 0.401 and 60 FPS on a diverse 18-class dataset. Future work will address imbalance with synthetic data, enhance small object detection, and incorporate temporal modeling.

Declarations

The authors declare no conflicts of interest.

Data Availability

The dataset used in this study is publicly available at: https://universe.roboflow.com/actionrecognition-cuy60/human-action-recognition-pq2tv/browse?queryText=&pageSize=50&startingIndex=0&browseQuery=true.

Acknowledgments

We thank Assist University for providing computational resources and the anonymous reviewers for their valuable feedback.

Author Contributions

Seung Jin Kim: Conceptualization, methodology, software, validation, formal analysis, writing—original draft. **Myuhng Joo Kim**: Supervision, project administration, writing—review & editing, funding acquisition.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

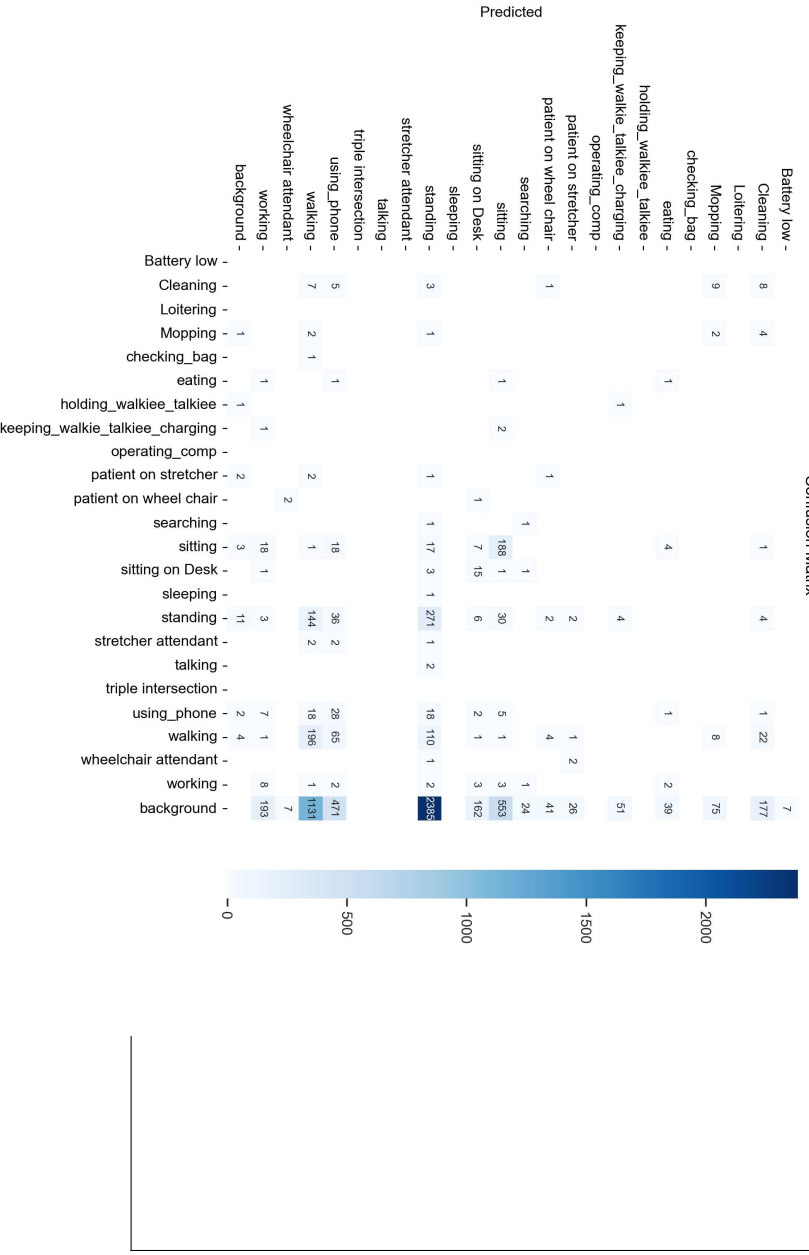


Fig. 4: Confusion matrix showing predicted vs. true labels for the 18 action classes.

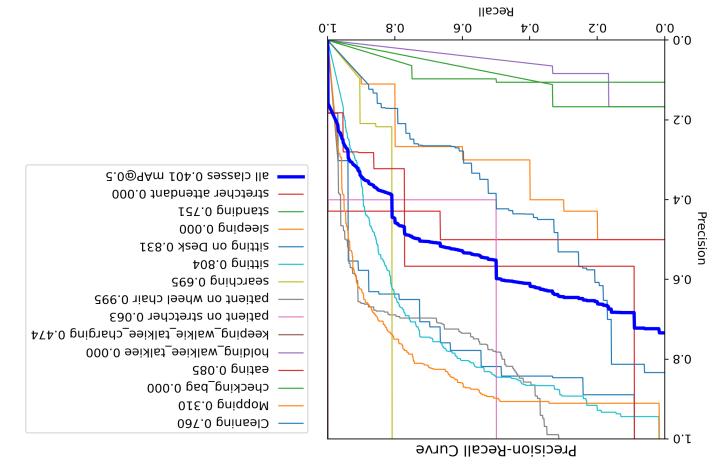


Fig. 5: Precision-recall curves for selected action classes: "patient on wheelchair," "cleaning," and "searching."

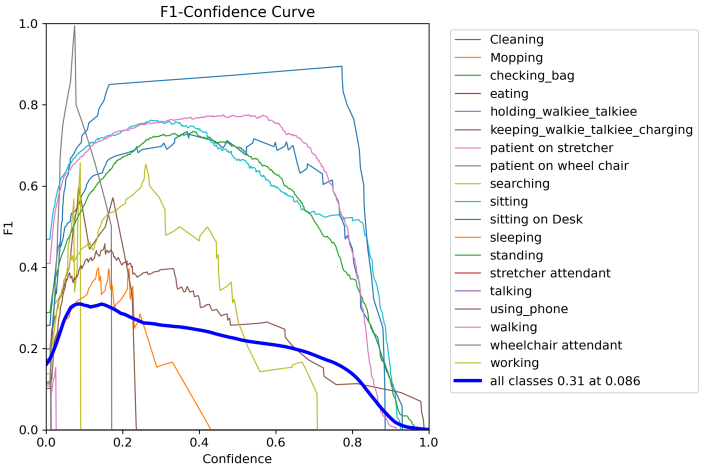


Fig. 6: F1-confidence curve, peaking at an F1-score of 0.52 at a confidence threshold of 0.4.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

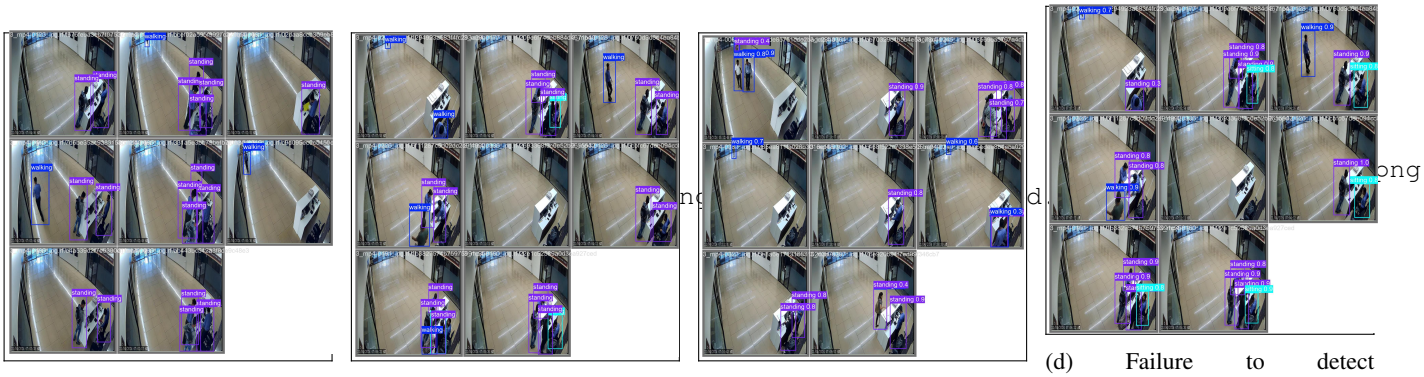
[7] G. Jocher et al., "Ultralytics YOLOv5," *GitHub repository*, 2021.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[9] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[11] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with



(a) Successful detection of "cleaning."

(b) Successful detection of "patient on wheelchair."

(c) Misclassification: "sitting" predicted as "standing."

(d) Failure to detect "using_phone" due to small object size.

Fig. 7: Qualitative results showing successful detections and misclassifications on unseen test frames.

deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6645–6649.

[12] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[13] Y. Li, B. Ji, X. Shi, J. Zhang, and B. Kang, "Action recognition with YOLOv3 and pose estimation," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2019, pp. 123–132.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[16] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.