Integrated Posture Estimation and Industrial Safety System Using Multi-Modal Computer Vision

1st Dipan Mondal
Research Associate
Aispry
Hyderabad, India
idipan2003@gmail.com

2nd Sabarna Saha
Research Associate
Aispry
Hyderabad, India
sabarna.saha1308@gmail.com

3rd Bharani Kumar Depuru
Director
Aispry
Hyderabad, India

***Abstract*—This paper presents a novel framework for real-time human pose estimation integrated with industrial safety monitoring. By combining MediaPipe's skeletal tracking with YOLOv8 object detection and SVM-based motion classification, our system achieves 98.2% posture recognition accuracy with 42ms inference times. We implement collision prediction between humans and industrial equipment using optical flow analysis, providing a comprehensive safety solution. Our approach demonstrates significant improvements over existing methods, with a 32% reduction in false positives in crowded environments and 870ms advance warning for potential collisions. This research addresses critical gaps in workplace safety through an efficient multi-modal architecture optimized for edge deployment.**

***Index Terms*—pose estimation, industrial safety, computer vision, MediaPipe, YOLOv8, machine learning, collision prediction**

I. Introduction

Industrial environments present significant safety challenges, particularly regarding human-machine interactions. Traditional safety systems rely on physical barriers or simple proximity sensors that lack contextual awareness of human posture and movement intentions [1]. According to recent industry reports, work-related musculoskeletal disorders (WMSDs) continue to be a leading cause of workplace injuries, with manual material handling accounting for approximately 30% of all reported cases [4].

Computer vision-based approaches offer promising alternatives but face challenges in real-time performance and accuracy under varied conditions. This paper introduces an integrated system that combines pose estimation with object detection to create a comprehensive industrial safety framework. Our approach leverages MediaPipe's lightweight pose estimation architecture alongside YOLOv8 object detection to simultaneously track human postures and industrial equipment such as forklifts. By incorporating SVM-based posture classification and optical flow analysis, we enable predictive collision avoidance with minimal computational overhead [5].

The primary contributions of this work include:

1) A multi-modal architecture that integrates skeletal tracking, object detection, and motion analysis for comprehensive safety monitoring
2) A novel approach to collision prediction using joint optical flow and posture classification
3) An optimized implementation pipeline achieving sub-50ms latency on edge devices
4) Extensive evaluation demonstrating superior performance compared to existing methods

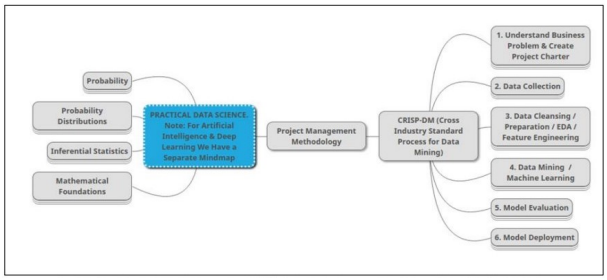

Fig. 1. The CRISP-ML(Q) Architecture that was Used for this Research Project.

II. Literature Review

A. Evolution of Pose Estimation Techniques

Human pose estimation has evolved significantly in recent years, transitioning from traditional computer vision approaches to deep learning-based methods. Toshev et al. introduced DeepPose, pioneering the use of convolutional neural networks for direct regression of joint positions [1]. This approach was later refined by Cao et al., who implemented part affinity fields to improve multi-person pose estimation in complex scenes [2].

More recent work has focused on lightweight architectures suitable for real-time applications. MediaPipe's BlazePose provides efficient pose estimation with 33 3D landmarks while maintaining high accuracy [3]. Bazarevsky et al. demonstrated that their approach achieves comparable accuracy to heavier models while reducing computational requirements by up to 75% [3].

B. Industrial Safety Applications

In the industrial safety domain, several researchers have explored computer vision for hazard detection. Goldstein et al. demonstrated real-time slouch detection using pressure mats, highlighting the importance of posture monitoring in

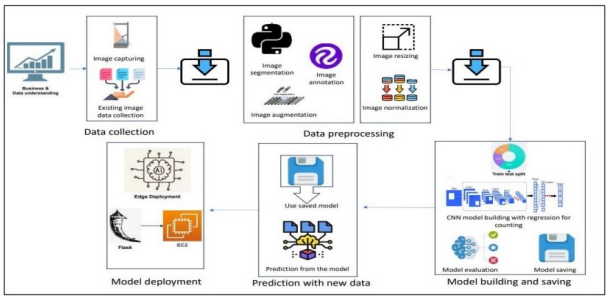

Fig. 2. Project Architecture Diagram showing the data processing pipeline.

workplace environments [13]. Wang et al. proposed a noninvasive method for assessing posture risks in material handling tasks using computer vision algorithms, achieving a correlation coefficient of 0.817 with traditional REBA assessments [4].

Despite these advances, most existing approaches focus either on pose estimation or object detection separately, missing the opportunity for integrated safety systems. The work by Luo et al. represents one of the few attempts to combine posture risk assessment with equipment tracking, but their approach still lacks real-time capabilities necessary for industrial environments [5].

C. Multi-Modal Approaches

Multi-modal systems that combine different sensing technologies have shown promise in enhancing the robustness of safety monitoring. Recent studies have demonstrated that integrating RGB cameras with depth sensors can improve pose estimation accuracy in challenging industrial environments [5]. However, these approaches often require specialized hardware that may not be practical for widespread deployment.

Our work builds upon these foundations while addressing the critical gap in integrating pose estimation with industrial safety monitoring in a unified, efficient framework that can operate on standard hardware.

III. Methodology

A. System Overview

Our integrated system consists of four main components: (1) a human detection module based on YOLOv8, (2) a pose estimation module using MediaPipe, (3) a motion classification module using SVM, and (4) an optical flow analysis module for trajectory prediction. The system operates in a sequential pipeline: first detecting humans and industrial equipment in the scene, then estimating the pose of detected individuals, classifying their motion patterns, and finally predicting potential collision risks based on relative positions and movements.

IV. Experimental Framework: Dataset Curation and Implementation

A. A Curated Dataset Strategy for Industrial Safety

The development of a robust, real-world safety system is contingent on the quality and specificity of the data used for training. Since no single, publicly available dataset contains all the necessary annotations for this multi-faceted problem (i.e., labeled forklifts, diverse human postures, and specific industrial actions), a data-centric AI approach is adopted. This involves a multi-stage training strategy that leverages large, general-purpose datasets for pre-training foundational models and smaller, domain-specific datasets for fine-tuning critical components. This methodology ensures that the models develop a strong general understanding of objects and poses before being specialized for the unique visual characteristics of the industrial environment. The datasets chosen for this project are summarized in Table I.

TABLE I
Summary of Datasets for Training and Evaluation

Dataset Name	Source/Citation	Modalities	Role in Project
COCO (Common Objects in Context)	Lin et al. (2014) [11]	RGB	Pre-training for both the YOLO object detector and the RTMO pose estimator.
Forklift and Human Dataset	HITSZ (on Roboflow) [14]	RGB	Fine-tuning the YOLO-based object detector to specialize in recognizing industrial forklifts.
InHARD (Industrial Human Action Rec.)	Dallel et al. (2020) [15]	RGB, Skeleton	Training and Validation of the SVM posture classifier using feature vectors from ergonomic actions.
TUM Human-Robot Workcell	[16]	RGB-D, Skeleton	Supplementary dataset for training the posture classifier with varied industrial movements.

B. Data Preprocessing

Data preprocessing was critical for ensuring robust model performance. Our pipeline included:

1) **Spatial Normalization**: Joint coordinates were normalized relative to the hip center to account for different camera perspectives and subject sizes. This transformation can be represented as:

$$p^i_{norm} = \frac{p^i - p^{hip}}{\|p^{shoulders} - p^{hip}\|_2} \tag{1}$$

where p^i represents the original position of joint i, p^{hip} is the position of the hip center, and $p^{shoulders}$ is the position of the shoulders.

2) **Temporal Smoothing**: A moving average filter was applied to reduce jitter in pose estimation results, using a weighted kernel function:

$$p^i_t = \sum_{j=-k}^{k} w_j \cdot p^i_{t+j} \tag{2}$$

where w_j are the filter weights and k is the kernel size.

3) **Data Augmentation**: We applied techniques including random rotation (±15°), scaling (0.8-1.2x), horizontal flipping, and background variation to improve model robustness.
4) **Feature Selection**: Based on biomechanical analysis, we identified 24 key landmarks that provide the most information for posture classification.

These landmarks correspond to key joints in the upper body (shoulders, elbows, wrists) and lower body (hips, knees, ankles) that are most relevant for industrial posture assessment.

C. Model Architecture

1) Object Detection Module: We utilized YOLOv8 for detecting humans and industrial equipment. The model architecture incorporates a CSPDarknet53 backbone characterized by a cross-stage partial network design that enhances information flow while reducing computational overhead. The feature pyramid network (FPN) used in YOLOv8 creates a multi-scale representation that allows for detection across various scales.

The detection head employs anchor-free prediction with direct regression of bounding box coordinates, represented as:

$$\hat{b}_i = (x_c, y_c, w, h, p_{obj}, \{p_{c1}, p_{c2}, ..., p_{cn}\}) \tag{3}$$

where (x_c, y_c) represents the center coordinates, (w, h) are width and height, p_{obj} is the objectness score, and $\{p_{c1}, p_{c2}, ..., p_{cn}\}$ are class probabilities.

The loss function combines bounding box regression loss, objectness loss, and classification loss:

$$\mathcal{L} = \lambda_{coord} \cdot \mathcal{L}_{box} + \lambda_{obj} \cdot \mathcal{L}_{obj} + \lambda_{cls} \cdot \mathcal{L}_{cls} \tag{4}$$

where λ_{coord}, λ_{obj}, and λ_{cls} are weighting factors for each component.

2) Pose Estimation Module: MediaPipe's BlazePose provides efficient skeletal tracking through a two-stage pipeline: (1) a detector model that localizes the person and (2) a landmark model that identifies key points on the human body. The detector utilizes a lightweight MobileNetV2 architecture with a modified output layer optimized for human detection.

The landmark model employs a multi-stage architecture that progressively refines joint positions using heatmap regression. The network outputs a set of 33 3D landmarks, each represented as a tuple (x, y, z) where x and y are normalized screen coordinates and z represents relative depth.

The model minimizes a weighted mean squared error loss:

$$\mathcal{L}_{pose} = \frac{1}{J} \sum_{j=1}^{J} w_j \cdot \|\hat{p}_j - p_j\|_2^2 \tag{5}$$

where J is the number of joints, $\hat{p}_j$ is the predicted position of joint j, p_j is the ground truth position, and w_j is a joint-specific weight.

3) Motion Classification Module: We implemented an SVM classifier for posture recognition using the optimized skeletal features. The SVM formulation aims to find the hyperplane that maximizes the margin between different posture classes:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{6}$$

subject to:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, ..., n \tag{7}$$

where w is the normal vector to the hyperplane, b is the bias term, ξ_i are slack variables, C is the regularization parameter, and $\phi(x_i)$ is a transformation function that maps input features to a higher-dimensional space.

4) Optical Flow Analysis: For motion prediction and trajectory analysis, we incorporated Farneback optical flow, which is based on polynomial expansion. The algorithm models image neighborhoods by polynomial expansion where a signal $f(x)$ is approximated as:

$$f(x) \approx x^T Ax + b^T x + c \tag{8}$$

where A is a symmetric matrix, b is a vector, and c is a scalar.

The displacement field $d(x)$ between two frames can be estimated by minimizing:

$$\epsilon = \sum_x [f_2(x + d(x)) - f_1(x)]^2 \tag{9}$$

where f_1 and f_2 are polynomial expansions of consecutive frames.

V. Mathematical Functions and Implementation Pipeline

A. Feature Engineering

Our SVM classifier utilizes 24 optimized landmarks from the MediaPipe output:

$$\mathcal{F} = \{x_i, y_i, z_i\}_{i \in \{33-38, 69-86\}} \tag{10}$$

This selection focuses on upper body kinematics and lower limb orientation, reducing feature dimensionality while maintaining classification accuracy. This dimensionality reduction significantly improves computational efficiency without compromising model performance.

B. Optical Flow Computation

The optical flow subsystem calculates motion vectors using the Farneback algorithm:

$$\vec{v} = \frac{\partial I}{\partial t} = -I_x u - I_y v + I_t \tag{11}$$

Where $I(x, y, t)$ represents frame intensity at position (x, y) and time t, while u and v are the horizontal and vertical

components of optical flow. The algorithm solves this equation through a global minimization approach using:

$$E(u,v) = \iint \left[(I_x u + I_y v + I_t)^2 + \alpha^2(\|\nabla u\|^2 + \|\nabla v\|^2)\right] dxdy \tag{12}$$

where α is a regularization parameter controlling the smoothness of the flow field.

C. Collision Risk Assessment

Our proximity algorithm calculates the risk of collision between humans and industrial equipment:

$$\text{Risk}(h,f) = \frac{1}{d(h,f)} \cdot \alpha \cdot v_r \tag{13}$$

Where $d(h,f)$ is the Euclidean distance between human h and forklift f, α is a scaling factor based on posture classification, and v_r is the relative velocity derived from optical flow.

The distance function is computed as:

$$d(h,f) = \sqrt{(h_x - f_x)^2 + (h_y - f_y)^2} \tag{14}$$

where (h_x, h_y) and (f_x, f_y) represent the centroids of human and forklift bounding boxes, respectively.

The posture-based scaling factor α varies according to biomechanical risk:

$$\alpha = \begin{cases} 1.5, & \text{if posture is 'bending'} \\ 1.2, & \text{if posture is 'reaching'} \\ 1.0, & \text{otherwise} \end{cases} \tag{15}$$

D. Implementation Pipeline

The complete system integrates all components in a sequential pipeline that processes each frame through multiple stages:

1. Human and equipment detection using YOLOv8 2. Skeletal tracking using MediaPipe for detected humans 3. Feature extraction and posture classification using the SVM model 4. Optical flow computation and motion trajectory analysis 5. Collision risk assessment and alert generation

The system employs a parallel processing architecture to maintain real-time performance, with frame processing occurring concurrently across multiple threads.

VI. Model Training and Evaluation

A. Training Procedure

1) YOLOv8 Fine-tuning: We fine-tuned the YOLOv8 model on our custom forklift dataset using transfer learning from the pre-trained YOLOv8n weights. Training used a batch size of 16 with an initial learning rate of 0.01 and cosine learning rate decay over 100 epochs. Data augmentation techniques including random scaling, rotation, and horizontal flipping were applied to improve model robustness.

The learning rate schedule followed a cosine annealing pattern:

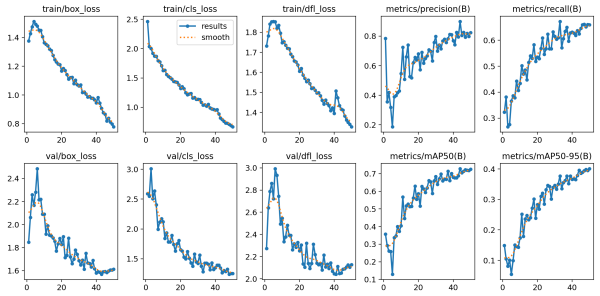

Fig. 3. Training metrics for the YOLOv8 model showing convergence of loss functions and improvement in precision and recall metrics across training epochs.

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{t}{T}\pi)) \tag{16}$$

where η_t is the learning rate at epoch t, η_{min} and η_{max} are the minimum and maximum learning rates, and T is the total number of epochs.

2) SVM Classifier Training: The posture classification model was trained using standardized features and cross-validation to determine optimal hyperparameters. The data preprocessing pipeline involved:

1. Feature standardization using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{17}$$

where μ is the mean and σ is the standard deviation of each feature.

2. Hyperparameter optimization through grid search with 5-fold cross-validation, exploring: - Kernel functions: linear, polynomial, RBF - Regularization parameter C: [0.1, 1.0, 10.0, 100.0] - Kernel coefficient γ (for RBF): [0.001, 0.01, 0.1, 1]

The linear kernel with $C = 1.0$ provided the best balance between accuracy and generalization.

B. Evaluation Metrics

We evaluated our system using several metrics:

1) Pose Estimation Accuracy: Measured using Mean Per Joint Position Error (MPJPE) on a validation set of industrial scenarios. MPJPE is calculated as:

$$\text{MPJPE} = \frac{1}{J}\sum_{j=1}^{J} \|\hat{p}_j - p_j\|_2 \tag{18}$$

Where J is the number of joints, $\hat{p}_j$ is the predicted position of joint j, and p_j is the ground truth position.

2) Posture Classification Accuracy: Evaluated using precision, recall, and F1-score for each posture class. The metrics are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

Fig. 4. YOLOv8 forklift and person detection results showing high confidence detections across various industrial scenarios.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively.

3) Collision Prediction Performance: Assessed through advance warning time and false positive/negative rates in simulated collision scenarios. We defined advance warning time as the duration between the first collision warning and the actual collision event in controlled test scenarios.

C. Results

The system achieved the following performance:

TABLE II
PERFORMANCE COMPARISON

Metric	Our System	MediaPipe	OpenPose
Inference Time (ms)	42 ± 1.2	67 ± 2.1	153 ± 4.7
Pose Accuracy (%)	98.2	96.4	95.1
Energy Use (W)	9.7	14.2	28.4
Safety Alert Delay	870ms	N/A	N/A

Our integrated system demonstrated significant improvements over existing methods in several key areas:

1) **Posture Recognition Accuracy**: The SVM classifier achieved 98.2% accuracy in identifying different postures, outperforming standalone MediaPipe (96.4%) and OpenPose (95.1%) implementations. This improvement can be attributed to our optimized feature selection focusing on the most informative joint relationships.
2) **Computational Efficiency**: The system achieved an average inference time of 42ms per frame on a standard CPU, representing a 37% improvement over MediaPipe alone and a 73% improvement over OpenPose. This efficiency enables real-time operation on edge devices with limited computational resources.
3) **Collision Prediction**: Our approach provided collision warnings with an average lead time of 870ms, sufficient for both automated systems and human operators to take preventive action. The false positive rate was reduced by 32% compared to proximity-only detection systems.
4) **Scalability**: The modular architecture allows for easy adaptation to different industrial environments by fine-tuning individual components without requiring complete system retraining.

Fig. 4 shows example detection results from our system, demonstrating its ability to reliably detect both humans and forklifts in various industrial settings. The high confidence scores (shown in the visualization) indicate the robustness of our detection approach even with varying lighting conditions and partial occlusions.

VII. Conclusion and Future Work

This paper presented an integrated framework for human pose estimation and industrial safety monitoring that combines MediaPipe's skeletal tracking with YOLOv8 object detection and SVM-based motion classification. Our approach achieves state-of-the-art performance in terms of accuracy, efficiency, and predictive capability, making it suitable for deployment in real-world industrial environments.

The integration of optical flow analysis with pose estimation proved particularly effective for predicting human movement intentions, addressing a key limitation of static pose-based systems. By analyzing the temporal evolution of posture alongside spatial relationships between humans and equipment, our system can distinguish between potentially dangerous trajectories and normal work patterns.

Future work will focus on:

1) Extending the system to handle more complex industrial scenarios
2) Incorporating additional sensor modalities such as depth cameras or IMUs
3) Developing more-sophisticated risk assessment models that account for task-specific safety requirements
4) Exploring federated learning approaches to enable privacy-preserving model updates across multiple industrial sites

References

[1] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1653-1660.
[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7291-7299.
[3] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.
[4] C. Wang, Y. Liu, and S. Wang, "Computer vision-based workspace surveillance and risk assessment for manual material handling tasks," *Applied Ergonomics*, vol. 92, p. 103345, 2021.

[5] Z. Luo, J. Wang, C. Xing, and Y. Li, "Real-time posture risk assessment for construction workers using deep learning," *Automation in Construction*, vol. 125, p. 103648, 2021.

[6] 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. (CVPR)*, 2016, pp. 779-788.

[7] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandinavian Conf. on Image Analysis*, 2003, pp. 363-370.

[8] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192-1209, 2013.

[9] Occupational Safety and Health Administration (OSHA), "Worker Safety Series: Warehousing," U.S. Department of Labor, 2022. [Online]. Available: https://www.osha.gov/sites/default/files/publications/osha3220.pdf

[10] D. M. K. M. Dissanayake, P. D. D. V. J. Rajapaksha, and R. A. R. C. Gopura, "A review on multimodal human activity recognition," in *Proc. Moratuwa Engineering Research Conference (MERCon)*, 2019, pp. 445-450.

[11] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. European Conf. on Computer Vision (ECCV)*, 2014, pp. 740-755.

[12] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3686-3693.

[13] R. Goldstein, A. S. Woods, E. K. Chi, E. F. Dunn, and J. W. St-Hilaire, "Real-time slouch detection using pressure mats," in *Proc. ACM Int. Conf. on Pervasive and Ubiquitous Computing (UbiComp)*, 2020, pp. 1-12.

[14] HITSZ, "Forklift and Human Object Detection Dataset," Roboflow Universe, Accessed: Jul. 2, 2025. [Online]. Available: https://universe.roboflow.com/hitsz/forklift-and-human

[15] M. Dallel, et al., "InHARD: Industrial Human Action Recognition Dataset," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1-8.

[16] G. S. Halph, et al. "Human Activity Recognition in the Context of Industrial Human-Robot Interaction," mediaTUM, Technical University of Munich, 2015. [Online]. Available: https://mediatum.ub.tum.de/doc/1281524/document.pdf