Multi-Camera Multi-Person Localization and Tracking Using Hierarchical Data Association

The localization and tracking of pedestrians, has been always one of the widespread research topics; from the most basic computer vision applications to current hot topics of autonomous robots, self-driving cars, human-computer interaction systems and safety. Trajectories convey valuable information for the scenes analysis and thus generate vital input data for different applications (Klinger et al., 2017). Tracking purpose is the localization and making connections between goals positions over time, goal identifying and tracking each person's path throughout the image sequences (Zhang et al., 2010).

The predictive model in a multi-object tracking environment with challenges like mutual occlusions, ambiguous backgrounds, appearance changes, and complex motions, is on the edge of drifting away from the actual position of objects. The trajectory's errors would accumulate when tracks are followed by consecutive updates based on inaccurate locations (especially in online methods).

Multiple object tracking (MOT) ought to allocate a unique trajectory for each goal across the whole sequences. One sequence of multi object detection and tracking is presented in figure 1 as a block diagram.

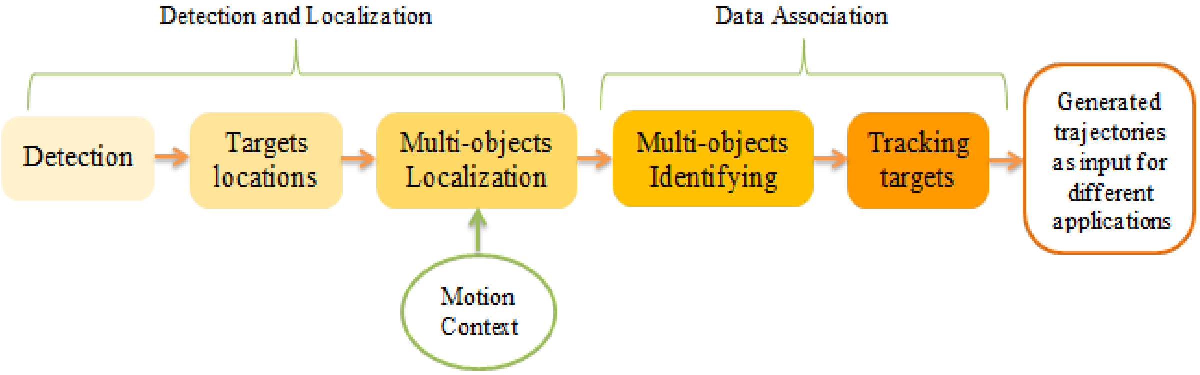
Detection and Localization
Data Association
Detection
Targets locations
Multi-objects Localization
Multi-objects Identifying
Tracking targets
Generated trajectories as input for different applications
Motion Context

Fig. 1. System diagram of each sequence in multiple object detection and tracking.

Detections in individual frames organize the observations in a recursive estimation framework. It is the initiation of many tracking systems. In situations with imprecise observations (resulted by the first block in figure 1), the generated trajectory is deviated easily (Klinger et al., 2017 and Ning et al., 2020). As mentioned before and seen in the diagram, in all tracking-by-detection methods (as the most frequent technique in MOT), tracking performance at the first stage relies on the detection accuracy.

In our previous work, we addressed some of the main difficulties attached to the tracking problem; occlusion, vast number of false positives and false negatives through different methods; while here, as a result of efficiency in applied detection method (YOLO) even in occlusion, the issue of false detections is highly handled that makes the problem much easier to solve. However, complex scenes with long occlusions, unexpected motion changes and approximate detections are still challenging. So, we can focus more seriously on occlusion and identity-switches.

Multi-object identification (next block in figure 1) is one of the most critical challenges in multi-object tracking. In common recursive works, applying classifiers in the target's categorization is usual. However, these strategies sharply diverge from the real object if the training data are stemmed from inappropriate examples. Tracking targets using the Hungarian algorithm and also global optimization methods are widespread. However, these methods not only can't handle occlusions in complex scenes but also have high computational complexity.

To deal with occlusion, we propose an Affinity matrix based on the existed trajectories and current detections. We also apply person re-identification methods to control the number of identity-switches.

We introduced this matrix in our previous work on datasets including "PETS 2009 campus", "AVG-TownCentre" and "ETHZ Bahnhof" that resulted in multi-person tracking accuracy 87.74-94.85 and precision 88.57-88.57 percent, outperforming the existing approaches. This happened in conjunction with a new detection method, combining background subtraction, HOG and ACF, with huge number of FP and FN. Applying the proposed matrix improved drastically the proficiency of the detection upon all evaluation parameters.

While here, we apply YOLO, as a state-of-the-art, real-time object detection algorithm with accuracy and precision of 97.6% and 95.7%, respectively for human, on its own (YOLO V8).

Moreover, we used to focus only on goals located in interested areas to control FNs; here, thanks to a proficient detection method, this considerations are useless and impose computational expenses. On the other hand, through our proposed method, YOLO's small number of probable false-positives will cause no obstacle in the tracking process (figure 2).

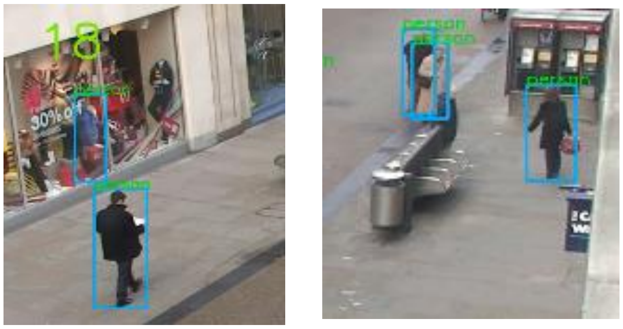

Fig. 2. Left: False Negative of YOLO, **Right:** Efficiency of YOLO's results in occlusion.

In this study, tracking is performed separately for each individual in an additional step. Tracking-by-detection method in combination with the probabilistic approach is used in data association; we also apply the previous information of targets, such as their locations, velocities, and appearances. The appearance dissimilarity is calculated by matching extracted features using convolutional neural network.

In addition, the image sequences or video frames are in two-dimensional environments; while, the real-world coordinate is in three. As a result of the perspective features and image depth, the scale in front of the scene doesn't equal it at the end.

Process of data association, jointly with our defined affinity matrix, are the main contributions of our introduced algorithm.

Here, we are working on multi-camera multi-person tracking, using the "WILDTRACK Seven-Camera HD" dataset (Figure 3). After handling challenges of multi-person localization and tracking, we follow exactly the same trend and lows for more cameras. We even keep to the same order in data association step, considering whole affinity matrices as a unit.

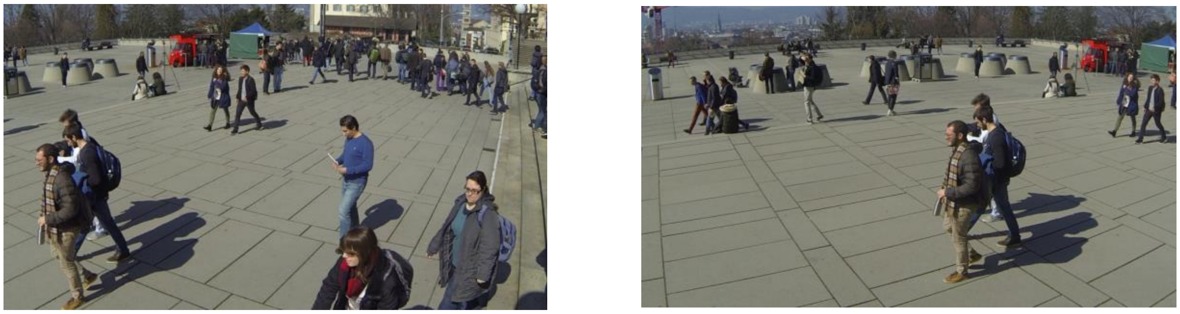

Fig. 3. WILDTRACK Seven-Camera HD.

Multi-Camera multi-object detection and tracking may seems more challenging; however, through our introduced method and taking the advantages of multiple camera's views, the issue is behaved the same and even less sophisticated than the previous one.

The affinity matrix and necessary kept information of the whole process of detection and tracking steps are defined in a more extended way. Moreover, we make the most of cameras overlaps (specifically in this dataset) and work on coincident points of all cameras simultaneously in an extra step. This provides an additional criteria for comparison and goals re-identification in occlusion.

The main steps of this study include:

- Data preparation
- Thresholds determinations and Initializing the algorithm according to the dataset
- Initializing trajectories
- Detection
- Keeping the goals information
- Affinity Matrix Generation
- Hierarchical Data Association
- Completion and modification of the information
- Applying memory's information/ Filling memory if needed

A simple System diagram of the proposed approach is represented in figure 4.

*** In detail, we define more factors and parameters and use more algorithms such as a short memory for information retrieval, a scale factor according to the scene information for detection modification (depending on the dataset and the detection method), Kalman filtering as an estimation algorithm, a temporal model for next probable area prediction and so on.

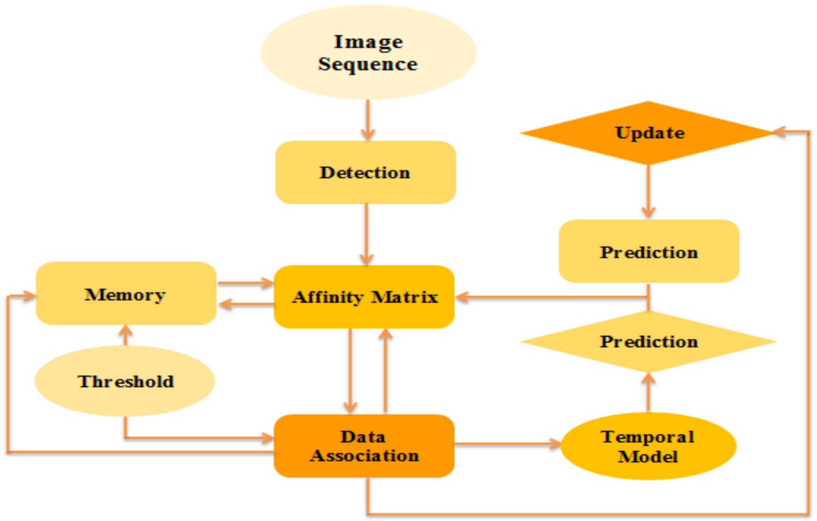

Fig. 4. System diagram. * Colors show the running steps of the algorithm (from light to dark). * Steps related to the Filters (like KF) are displayed by the lozenge, and steps related to the data and prior knowledge represented by oval.

Diagram explanation (for single camera tracking): Image sequences enter the detector. Detection's results, in addition to the average of the filter's prediction (lozenge one) and temporal model's prediction (rectangular one) are applied in generating the affinity matrix. This matrix is completed during next steps (using memory, simultaneously with data association). Memory (defined to decrease the number of FN) is filled at the end of the association, when trajectories lost in the middle of the scene. Finally, positions are updated and relocated to the new positions at the same time (update step of the filter).

Affinity Matrix

After each time step's detection, the proposed affinity matrix is generated between the current detected goals and the existed trajectories. For each goal (pedestrian) in the previous frame, a probable next area is defined in which the person might appear in the next time step (shown in Figure 5). In case any detected goal in the current frame appears in each probable area (each row), the related column would be given a figure.

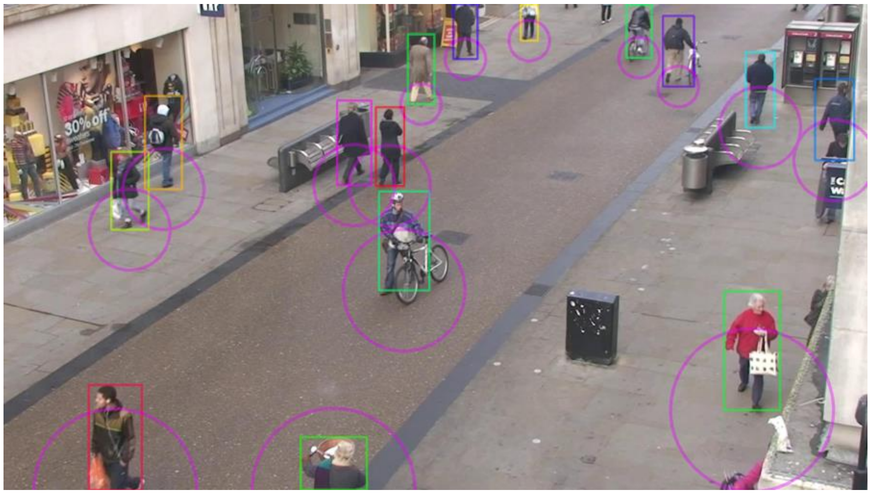

Fig. 5. Scale Factor and Next Probable Area representation according to the dataset's characteristics.

This is a zero and one matrix, in which the rows represent the goals in the previous time step (existed trajectories in the current frame) and the columns show detections in the current frame. In an ideal circumstance, for each column there is one and only one column with a nonzero magnitude in rows. However, in real situation, some special states are possible to happen. Some of them are presented in figure 6 and 7.

Fig. 6. Possible situations of the scene due to occlusion. Missing a pedestrian, entrance or leave the scene, that all result to the detections wrong assignment to trajectories.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 1 \\ 0 & 0 & \mathbf{0} & 0 & 1 & 0 \\ 0 & 1 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Fig. 7. Possible Exceptional states in Affinity Matrix.

In the following, the process of step-by-step association according to predefined rules and the frame’s corresponding matrices, the relation between columns and rows in case of appearance similarity are depicted in Fig. 8 and 9.

Figure 10 shows one more step of non-maximum suppression and then a picture of tracking people in dataset “AVG-Town Centre” in figure 11.

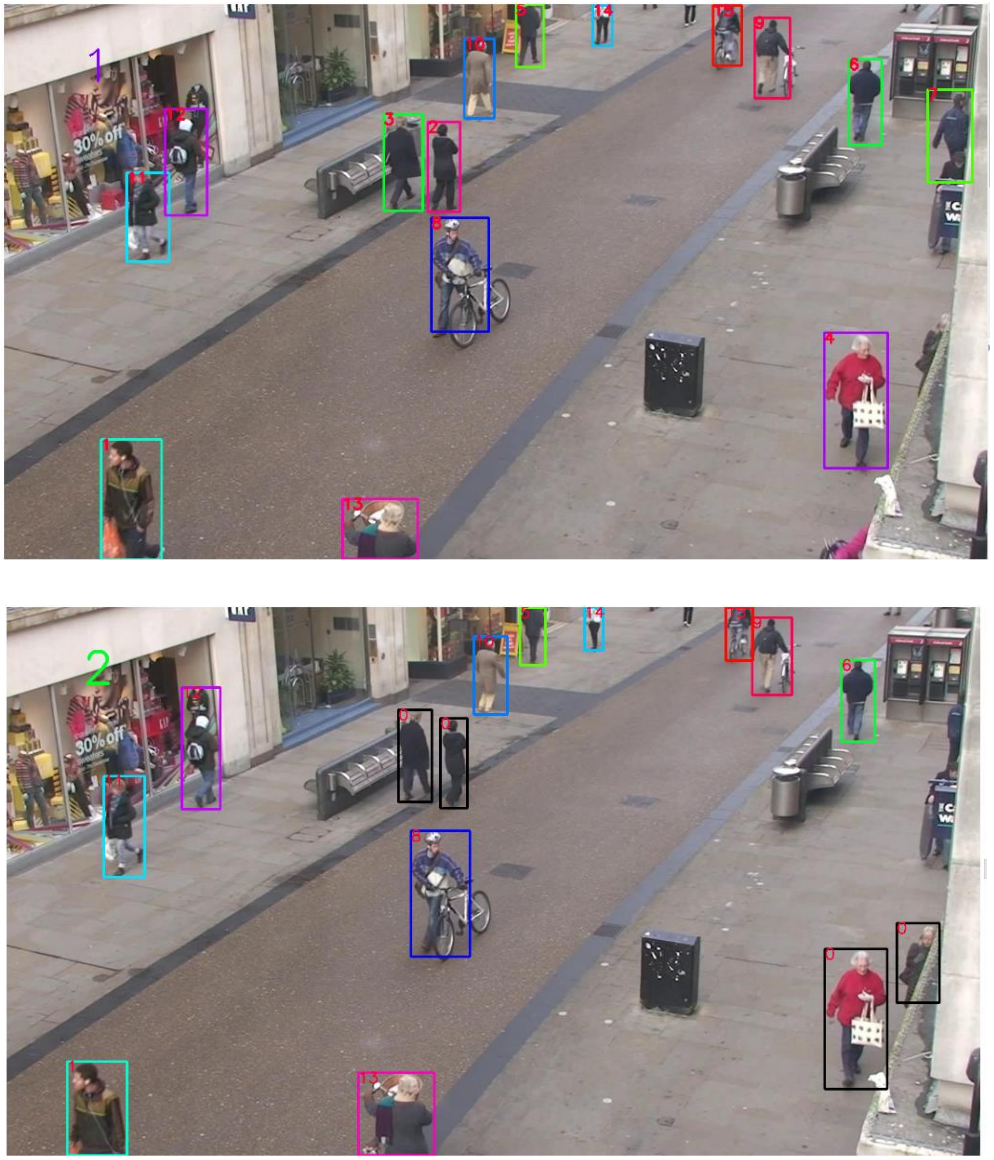

Fig. 8. Data Association is done step-by-step. As it is can be seen, some detected objects remained unassigned (black). For some reasons such as their distances from each other or their similarity in appearances they would be examined in another step, through predefined rules. These exceptions can be seen and predicted in their corresponding matrix.

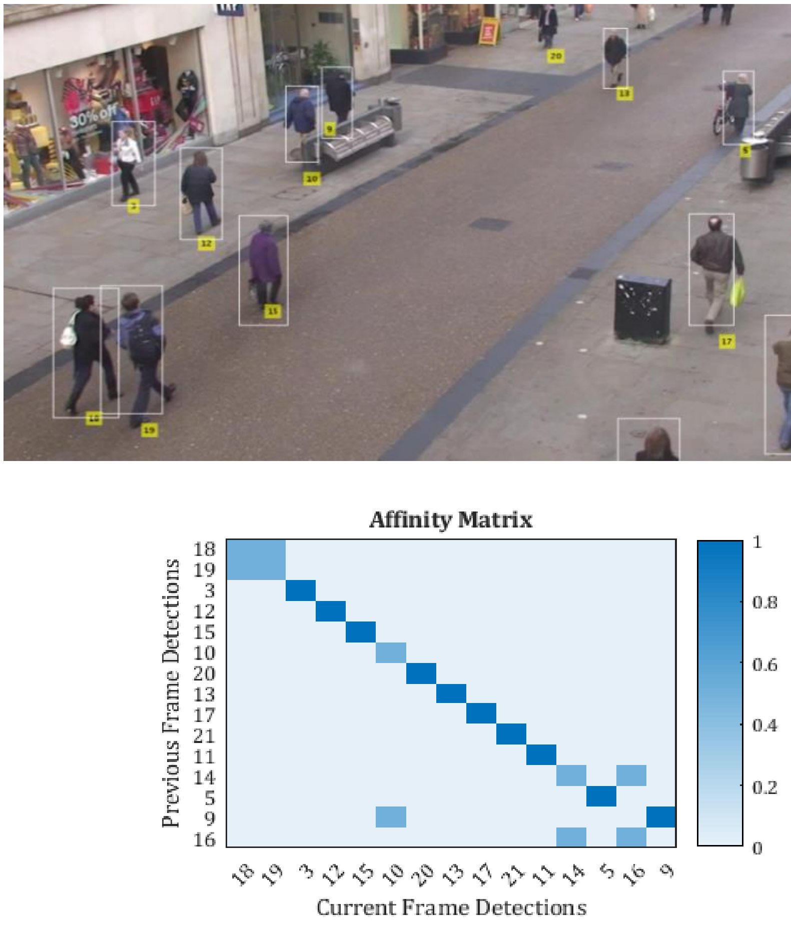

Fig. 9. Affinity Matrix in a time step and its corresponding frame.

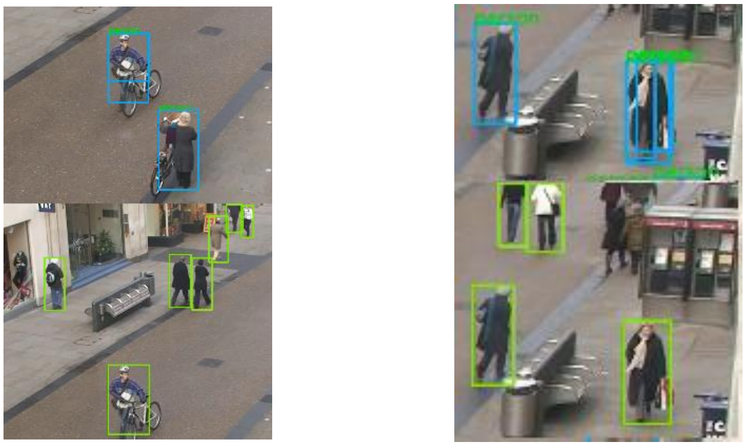

Fig. 10. Data Preparation's Step; Applying More Non-maximum Suppression.

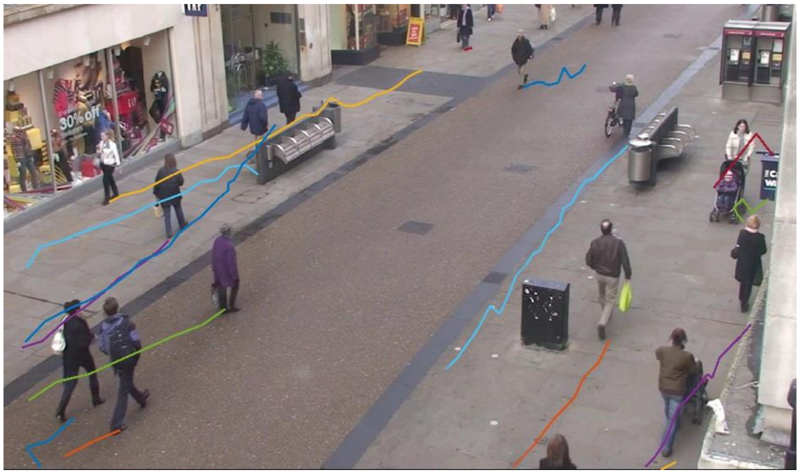

Fig. 11. Tracking people in dataset AVG-Town Centre.

References

Klinger, T.; Rottensteiner, F.; Heipke, C.; "Probabilistic multi-person localisation and tracking in image sequences," ISPRS Journal of Photogrammetry and Remote Sensing, no. 127, pp. 73–88, 2017.

Ning, C.; Menglu, L.; Hao, Y.; Xueping, S.; Yunhong, L.: "Survey of pedestrian detection with occlusion; Complex & Intelligent Systems pp. 1/11, 2020.

The WILDTRACK Seven-Camera HD Dataset: https://www.epfl.ch/labs/cvlab/data/data-wildtrack/

YOLO V8: An improved real-time detection of safety equipment in different lighting scenarios on construction sites - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Performance-metrics-calculated-for-YOLOv8-model_tbl2_379522326

Zhang, D.; Xia, F.; Yang, Z.; Yao, L.; Zhao, W.; "Localization Technologies for Indoor Human Tracking," Future Information Technology (FutureTech), 5th International, 2010.