From On-Site Video to Crew Balance Charts: A Computer Vision Approach for Worker Productivity Analysis

Sai Kumar Titti[1], Aritra Pal[1], Varun Kumar Reja[2] and Koshy Varghese[1]

[1] Department of Civil Engineering, Indian Institute of Technology Madras, India
[2] Department of Civil Engineering, Indian Institute of Technology Bombay, India
saikumartitti@gmail.com, aritrapal@civil.iitm.ac.in, varunreja@iitb.ac.in, koshy@iitm.ac.in

Abstract –

Construction productivity has remained stagnant for decades, underscoring the need for productivity improvement methods. Crew Balance Charts (CBCs) are an effective tool for analysing crew dynamics and identifying opportunities for productivity improvement. However, manual preparation of CBCs is time-consuming and limits scalability. This paper proposes a Computer Vision-based Crew Balance Chart (CV-CBC) framework for automated CBC generation using construction site videos. The framework comprises: Activity Definition, Data Preparation, Activity recognition model development, Inference pipeline, and CBC generation and Validation. The activity-level outputs are further aggregated into Crew-Based Work Sampling (CBWS) metrics for high-level labour monitoring. The framework is implemented and evaluated on masonry activity, with validation performed through per-second comparison between the automatically generated CBC and the manually prepared CBC. Results show that the activity recognition model achieves a macro-F1 score of 0.54 on the validation set across four activity classes, while the CBC achieves an accuracy of 60%. Activity predictions are aggregated into Value-Adding (VA), Non-Value-Adding (NVA), and Non-Value-Adding, but Necessary (NVAN) categories, enabling the calculation of CBWS metrics. The study introduces automated CBC generation and highlights its potential for scalable crew-level analysis in construction, while performance improvements remain an area for future work.

Keywords –

Computer Vision; Crew Balance Charts; Construction Productivity; Activity Recognition

1 Introduction

The construction industry is one of the largest sectors of the global economy, accounting for nearly 7% of global gross output, and is projected to reach USD 22 trillion by 2040 [1]. Despite this scale and anticipated growth, construction productivity has remained largely stagnant for decades. Between 2000 and 2022, global construction productivity increased by approximately 10%, whereas manufacturing productivity rose by roughly 90%, and the overall economy by 50% [1]. This highlights the need to improve productivity to meet future demands amid persistent labour shortages.

Labour accounts for approximately 30–50% of project costs, yet is often utilised at only 40–60% of its potential due to coordination inefficiencies [3]. Improving labour productivity is challenging due to the nature of construction work. Construction projects are temporary and executed in dynamic environments. Consequently, performance depends on interdependent and sequential activities, rather than isolated tasks [2][3]. At the operational level, work is predominantly executed by crews whose productivity is largely dependent on coordination, task sequencing and interaction among the crew members [3]. These underscore the need for crew-level analysis to quantify productivity losses and identify opportunities for improvement.

Lean Construction emphasises maximising value by eliminating waste, which requires systematic waste identification [4][5]. Crew Balance Charts (CBCs) and Crew-Based Work Sampling (CBWS) are established lean tools for analysing crew-level productivity. CBCs capture the temporal distribution of work among the crew across work cycles, highlighting crew interdependencies, exposing forced idleness, and work imbalance. By making such inefficiencies visible, CBCs facilitate the identification of opportunities for productivity improvement [6][7]. In contrast, CBWS quantifies waste by aggregating each worker's activities into Value-Adding (VA), Non-Value-Adding (NVA), and Non-Value-Adding but Necessary (NVAN) categories and can be applied beyond cyclic activities. While CBCs enable productivity improvements, CBWS enables effective labour monitoring.

Traditionally, CBCs are prepared through manual observation of crews across work cycles, requiring continuous observation of crew members across multiple cycles. Task durations are recorded manually using stopwatches or time-lapse video and then translated into a graphical representation. The process is illustrated in Figure 1. While effective for identifying inefficiencies and supporting improvement decisions, this manual process is labour-intensive and limits the number of cycles that can be analysed [8]. To enable scalability and repeatability, this manual process needs to be automated. Similarly, CBWS requires repeated manual observations, making it time-consuming and challenging to scale.

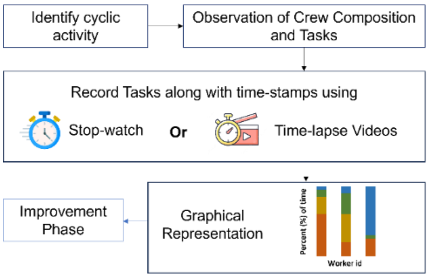

Figure 1 Traditional Crew Balance Chart Process

Prior studies have explored automating activity recognition using wearable sensors such as accelerometers [9]. While these capture detailed motion information, the hardware required for large-scale deployment is intensive and can be uncomfortable for workers. In contrast, recent advances in Computer Vision (CV) enable automated interpretation of construction site videos in a non-intrusive way by detecting, tracking and recognising worker activities [10]. This makes computer vision suitable for capturing crew interactions in labour-intensive construction activities.

Despite the potential of CV for automating construction site observation, existing applications remain fragmented and largely isolated. These approaches rarely translate the outputs into structured analytical tools such as CBC. Consequently, there is no integrated framework that uses computer vision to support crew-level analysis.

Accordingly, this study focuses primarily on CBC as the analytical tool, with CBWS concepts used only to aggregate recognised activities for high-level labour monitoring. Therefore, the objectives of this paper are to:

1. Investigate the feasibility of using CV to automate the CBC generation from construction site videos.
2. Develop an integrated framework, referred to as CV-CBC, that integrates worker detection, worker tracking, and activity recognition to convert video-based observations into CBC.
3. Evaluate the framework on real-world data and identify key technical limitations.

The remainder of this paper is organised as follows. Section 2 reviews related work on computer vision applications in construction productivity. Section 3 presents the proposed CV-CBC framework and its components. Section 4 describes the implementation and validation of CV-CBC framework using masonry activity. Section 5 presents the key results, Section 6 discusses ethical considerations, and Section 7 concludes by outlining future research directions.

2 Related Work

Computer vision applications in construction have largely focused on safety monitoring, while studies on productivity analysis remain limited [11]. The papers that address productivity are also focused on equipment rather than workers. Workers pose various challenges, as their movements are less well-defined and structured than those of equipment operations. Nonetheless, researchers have attempted to model the workers' actions, and a summary of these studies is presented below.

2.1 Equipment-Centric Activity Analysis

Several studies have focused on recognising equipment activities due to their structured, predictable motion patterns. Roberts and Golparvar-Fard proposed an end-to-end pipeline for detecting, tracking, and recognising excavator activities. However, the authors noted that extending such approaches to workers is challenging due to the complexity of human motion [12].

The multi-stage process can lead to error accumulation across modules, which affects the overall performance of the system. Ghelmani et al. developed a single-stage spatio-temporal activity recognition framework for excavator actions. This improved the performance compared to a multi-stage pipeline, highlighting the advantages of single-stage architectures [13]. Nevertheless, the study is limited to excavators and does not consider workers' activity recognition.

2.2 Worker-Centric Activity Analysis

Worker-centric analysis is slightly more challenging than equipment-centric analysis. Toledo et al. used YOLOv5 to detect worker postures and classify them into productive and contributory works, enabling the generation of CBC [6]. While promising, the approach was limited to single-worker scenarios and did not incorporate tracking or temporal modelling. Roberts et al. developed a 2D pose-estimation-based approach for CBC automation that achieved good accuracy but was limited to a single-worker scenario. As a result, the approach does not capture interaction among multiple workers, limiting its applicability for crew-level analysis [14].

Kim et al. proposed a generalised vision-based framework for construction productivity analysis, aiming to improve scalability. The framework addresses both equipment and labour productivity analysis. However, it relies heavily on the synthetic and web-crawled images, rather than video data. In addition, limited emphasis is placed on fine-grained worker activity recognition, which is essential to developing CBC [15].

In summary, existing studies have explored vision-based productivity analysis. Despite these advances, current methods have largely focused on equipment, are limited to single-worker analysis and mostly rely on static images rather than continuous video data. To address these limitations, an integrated CV-CBC framework is developed to enable CBC generation.

3 Computer Vision-based Crew Balance Chart (CV-CBC) Framework

Based on previous studies, automated generation of CBCs requires three major CV techniques: object detection, tracking and activity recognition. Object detection enables the localisation of workers, object tracking maintains worker identity over time, and activity recognition classifies the activities performed by the worker [12] [10]. These techniques form the CV pipeline for generating CBC from site videos. The proposed CV-CBC framework comprises five components: Activity Definition, Data Preparation, Activity Recognition model development, Inference Pipeline, and CBC Generation and Validation. The detailed framework is illustrated in Figure 2 and each component is discussed in the following subsections.

3.1 Activity Definition

The first step is to identify the main activity for CBC analysis. CBC analysis is effective when applied to cyclic, labour-intensive, repetitive, and coordination-intensive work. Small improvements in cyclic activities can yield significant productivity gains over time. The selected activity should also be critical to the project in terms of time and cost.

In this study, the term *"main activity"* refers to the overall construction operation selected for CBC analysis. In contrast, *"activity"* refers to the specific actions carried out by workers within the main activity. Once the main activity is selected, it is further broken down into a set of activities that represent the work performed by the individual workers. This breakdown is carried out by observing on-site operations and understanding the crew composition. The objective is to identify all activities that may occur during one complete work cycle. Activities are classified into Value-Adding (VA), Non-Value-Adding (NVA), and Non-Value-Adding but Necessary (NVAN) categories to enable CBWS. To reduce ambiguity, activities are defined to be distinguishable based on visual cues, with visually similar activities grouped into broader categories.

3.2 Data Preparation

This stage converts the site videos into a structured dataset that enables model training in the subsequent stage. It includes Data Collection and Annotation, Clips Generation, Data Augmentation and Dataset Split.

Data Collection: Video data are collected from construction sites using standard camera-based methods. The primary requirement is that workers are clearly visible for a sufficient time to capture their activities. To improve robustness, videos are collected across different work cycles, camera viewpoints and site conditions.

Data Annotation: The videos are manually annotated using standard tools. For each visible worker, a bounding box (bbox) is drawn, and an activity label is assigned based on a pre-defined activity breakdown. A new bbox will be drawn each time the worker changes the activity. The bounding boxes are updated over time in response to the worker's movement. During annotation, the worker is preferably kept near the centre of the bbox, with a small margin, to capture more spatial context. The annotations are exported in a structured format, such as CVAT 1.1, which enables clip generation in the subsequent step.

Clips Generation: Using the annotations, videos are processed to extract worker-level tracks, which are then trimmed into shorter clips for training and validation.

Data Augmentation: It is used to improve the diversity of the dataset and the activity recognition model robustness. Both spatial (horizontal flipping) and photometric (brightness, contrast) augmentation techniques are applied. Data augmentation also helps eliminate the imbalance among activity classes.

Dataset Split: The dataset is split into training, validation, and test sets using a video-level splitting strategy to prevent data leakage and ensure fair evaluation. All clips that belong to the same video are assigned to a single subset.

3.3 Activity Recognition Model Development

This stage involves developing an activity recognition model to classify video clips into pre-defined activity classes, which requires learning both spatial and temporal information. Model development includes model selection, training and evaluation. Model hyperparameters are tuned using validation data to improve performance. Model selection is based on macro F1-score, where F1-score is defined as the harmonic mean of precision and recall, and the macro F1-score is computed by averaging the F1-scores across all classes giving equal importance to each class. This metric provides a balanced assessment of model performance. The output of this stage is a trained activity recognition model that is used in the subsequent inference pipeline.

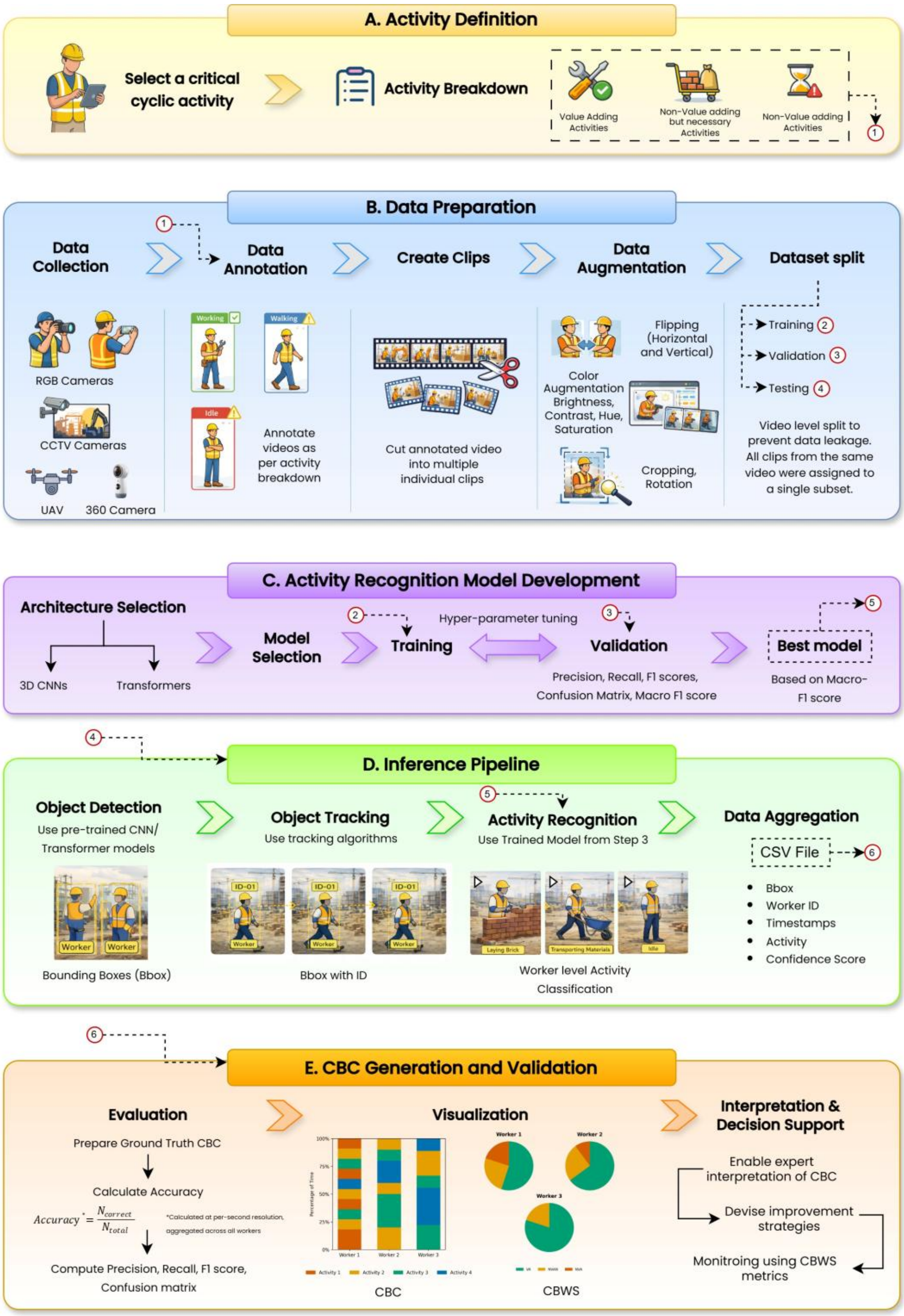

Figure 2 Computer Vision-based Crew Balance Chart Framework

3.4 Inference Pipeline

This stage represents the operational stage of the proposed CV-CBC framework. During inference, workers are detected in each frame, tracked across time with unique IDs, and converted into worker-level clips. The trained activity recognition model is used to predict the activity classes. These predictions are stored in a Comma-Separated Values (CSV) format along with temporal information. This serves as an input for CBC generation in the next stage.

3.5 CBC Generation and Validation

The final stage of the framework converts worker-level activity predictions from the inference pipeline into a CBC and evaluates its accuracy. Predicted activities are temporally organised into continuous timelines for each worker and aligned along a common time axis to generate the CBC. The accuracy is computed by comparing predicted activities with manually annotated ground truth labels at a per-second temporal resolution, aggregated across all workers, as shown in Equation (1)

$$Accuracy = \frac{N_{Correct}}{N_{Total}} \quad (1)$$

where N_{correct} and N_{total} denote the total number of correctly predicted and total activity predictions, respectively, aggregated across all workers.

Performance is evaluated using standard metrics, including accuracy, F1-score, precision, recall, and confusion matrices. Similarly, CBWS metrics are calculated by aggregating recognised activities into the Value-Adding (VA), Non-Value-Adding (NVA), and Non-Value-Adding but Necessary (NVAN) categories, enabling high-level crew monitoring.

4 Implementing CV-CBC Framework

4.1 Activity Definition

This section presents the implementation of the proposed CV-CBC framework using masonry as the main activity. Masonry was selected because it is a labour-intensive, repetitive process that requires close crew coordination, making it prone to work imbalance and enforced idleness. These characteristics make masonry well-suited for CBC analysis. Based on on-site observations of the masonry work, the main activity was broken down into a set of worker-level activities that occur during execution. These activities were defined to represent distinct and meaningful work states observed repeatedly across work cycles. The selected activities and their descriptions are summarised in Table 1.

Table 1 Description of selected activities

Activity	Category	Description
Applying Mortar	VA	Applying mortar between blocks and preparing bed
Idle	NVA	No productive or supporting work
Laying Blocks	VA	Placing and positioning masonry blocks
Shifting Materials	NVAN	Transporting/ Repositioning materials or tools

4.2 Dataset Preparation

Video data were collected from multiple construction sites under normal working conditions using a static camera. Most of the videos were not recorded specifically for computer vision analysis, but rather captured for manual observation. A total of seven videos were collected, with crew sizes ranging between three and four workers. The videos were annotated using the Computer Vision Annotation Tool (CVAT), and the XML files were exported in CVAT 1.1 format, which enabled clip generation. Worker-level clips were generated, each with a fixed duration of 8 seconds. Data Augmentation was applied using horizontal mirroring, brightness, and contrast adjustments. The details of the clips, class-wise, are presented in Table 2.

Table 2 Details of Clips

Class	Clips (No.)	Duration (min.)
Applying Mortar	305	42.62
Idle	357	47.75
Laying Blocks	263	35.15
Shifting Materials	318	40.81

The dataset was split using the following strategy: one video was reserved for validation, one for inference, and the remaining for training.

4.3 Model Training

A transformer-based architecture was selected for activity recognition, as transformer models are well-suited for learning spatio-temporal information from videos. In this experiment, the VideoMAE-Small model pre-trained on the Kinetics-400 dataset was used.

The small variant was chosen to mitigate overfitting given the limited dataset size. Each 8-second clip was uniformly sampled and resized to 224 × 224 pixels to match the pre-training input configuration. The proposed model was then fine-tuned on the annotated videos using supervised learning. A partial fine-tuning strategy was employed by unfreezing 6 of the 12 backbone layers, enabling greater adaptation to construction while preserving common spatial and temporal characteristics learned during pre-training. The trained model achieved a macro F1-score of 0.54, indicating balanced recognition performance across activity classes. During training, class-wise F1-scores were 0.44 for Applying Mortar, 0.56 for Idle, 0.49 for Laying Blocks, and 0.66 for Shifting Materials.

To validate the model choice, a 3D CNN-based I3D model was also evaluated, achieving a macro F1-score of 0.43, with class-wise scores ranging from 0.36 to 0.48. In comparison, VideoMAE achieved a higher macro F1-score of 0.54 and was therefore selected for subsequent experiments.

4.4 Inference Pipeline

During inference, the test set was used to generate the CBC using an end-to-end pipeline. The pipeline comprises object detection, object tracking, activity recognition and post-processing. The test video is 164 seconds long; a segment was trimmed to include only activity classes on which the model is trained, enabling fair evaluation. The crew size is three, and the work cycle is defined as the construction of one masonry course.

4.4.1 Object Detection

The YOLO11x object detection model, pre-trained on the COCO dataset, was used to identify workers in each video frame. In cluttered construction site environments, detection sensitivity and localisation accuracy were balanced using an Intersection-over-Union (IoU) threshold of 0.5 and a confidence threshold of 0.3. Frame-wise bounding boxes generated by the detector were used as inputs for tracking. High detection performance is achieved with a precision of 98.8% and a recall of 98.6%, indicating reliable worker localization.

4.4.2 Object Tracking

The BoT-SORT multi-object tracking algorithm with re-identification (ReID) enabled was used to track detected workers across frames. This was made to assign distinct worker IDs consistently. Only tracked IDs with a minimum continuous duration of 30 frames were retained for further analysis to eliminate workers not relevant to the work cycle. A total of two ID switches were observed across all workers, with Multi-Object Tracking Accuracy of (MOTA) of 97.4%. For better visualisation, tracking IDs were grouped for subsequent analysis.

4.4.3 Activity Recognition

For each tracked worker, bounding box–based crops were used to generate worker-centric clips. Activity recognition was performed using the trained VideoMAE model with an input resolution of 224 × 224 pixels. A 2-second sliding window, consisting of 16 frames, was applied with a stride of 1 second, enabling per-second activity predictions. For overlapping windows, the activity with the highest confidence was selected.

4.4.4 Post-processing

Before CBC generation, worker-level activity timelines were stabilised using post-processing rules. Idle predictions were filtered using a temporal persistence rule, assigning the idle label only if it persisted for more than two seconds and exceeded a confidence threshold of 0.75. This reduces false idle predictions caused by short pauses or low-motion intervals that occur during continuous productive work (such as laying blocks). The worker-level activity timelines were then combined to generate a CBC.

4.5 CBC Generation and Validation

Activity recognition performance was assessed using class-wise F1-score. The model shows limited generalisation compared to training performance. Laying Blocks achieved the highest F1-score (0.62), followed by Shifting Materials (0.38), while Applying Mortar and Idle recorded lower F1-scores of 0.18 and 0.09, respectively. The confusion matrix in Figure 3 shows that Applying Mortar and Laying Blocks are the most frequently misclassified activities, with many Applying Mortar instances incorrectly predicted as Laying Blocks. This confusion arises because both activities exhibit similar motion cues, such as bending and hand actions with a trowel or masonry blocks, making it difficult to distinguish between them.

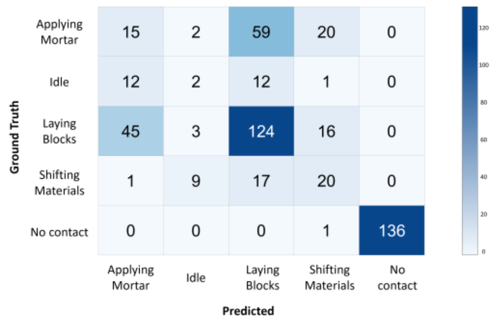

Figure 3 Confusion Matrix

The automated CBC is produced by aggregating worker timelines, whereas the manual CBC was prepared using the traditional process. To ensure ground truth reliability, inter-annotator agreement (IAA) was evaluated on the test set with two annotators and Cohen's κ, achieving 93.30% agreement (κ ≈ 0.90), indicating almost perfect agreement. When the worker is out of the frame, it is termed "No Contact". Figure 4 illustrates the manually created Crew Balance Chart with automated CBCs generated by the CV-CBC framework.

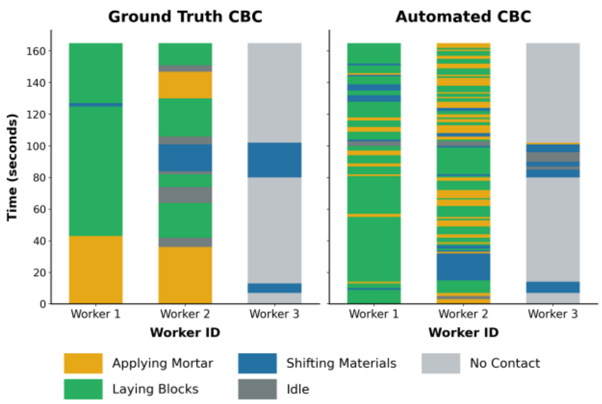

Figure 4 Ground truth CBC and Automated CBC

CBC accuracy was computed by comparing predicted and ground-truth activity labels at a per-second, per-worker resolution (Equation 1). The automated CBC achieved an accuracy of 60.0 % (297 accurate predictions out of 495 total) including No-Contact periods and 45% excluding them. This metric reflects the framework's ability to reconstruct the temporal structure of activities.

CBWS metrics were derived by aggregating activity predictions into VA, NVAN and NVA categories. For each worker, VA, NVAN and NVA time proportions were computed from ground truth and predicted activities and compared using absolute percentage deviation, as reported in Table 3. Workers 1 and 2 primarily engage in value-adding tasks, whereas Worker 3 is involved only in NVA and NVAN. This distribution is consistent with typical mason–helper role patterns and demonstrates the potential for identifying worker-level contributions to support high-level labour monitoring.

Table 3 CBWS comparison between ground truth and predictions

Worker	Category	Ground Truth (%)	Predicted (%)	Absolute Error (%)
1	VA	98.79	90.91	7.88
	NVAN	1.21	7.27	6.06
	NVA	0.00	1.82	1.82
2	VA	73.33	81.21	7.88
	NVAN	10.30	15.76	5.46
	NVA	16.36	3.03	13.33
3	VA	0.00	3.45	3.45
	NVAN	100.00	68.97	31.03
	NVA	0.00	27.59	27.59

5 Discussion

This study demonstrates that the proposed CV-CBC framework can automatically extract activity-level information to generate CBC. The activity recognition model achieved a validation macro F1 score of 0.54, while the automated CBC achieved an accuracy of 60%, comparable to prior work. Toledo et al. reported a 58% CBC accuracy under site conditions using a fixed camera, with contributory work not detected (0%) due to occlusions and viewpoint limitations [6]. The literature does not define strict deployment thresholds; however, prior studies suggest that accuracy levels around 78–85% are sufficient for practical use [14][15]. Accordingly, the results should be interpreted as a proof of concept.

The reported CBC accuracy of 60% includes no-contact periods, which inflates performance as it is easier to detect the absence of a worker; excluding them reduces accuracy to 45%, which better reflects the activity recognition performance. Given the high performance of detection and tracking stages, the overall pipeline performance is primarily constrained by activity recognition, which limits the CBC accuracy. Most errors arise from activity misclassifications, while detection and tracking contribute comparatively less. Although only two ID switches were observed, identity consistency remains critical for CBC generation, as even minor track fragmentation can affect worker-level timelines, suggesting the potential for post-processing strategies to maintain consistent identities.

These results should be interpreted in the context of the dataset. The dataset is small, and the framework is validated only on masonry activity, limiting generalizability. Extending to other activities or larger datasets may introduce class imbalance (e.g., overrepresented idle class), requiring resampling or augmentation. Due to the data constraints, the study was not extended to other activities.

At the activity level, significant confusion was observed between laying blocks and applying mortar because of their similar motion patterns. Table 3 shows that the mean absolute error for VA activities is within 8%, indicating higher reliability in detecting productive work. In contrast, NVA and NVAN activities exhibit higher error rates. The distinction between these two categories is subjective, as even human observers may classify the same activity differently during manual CBC generation. Higher error rates observed for Worker 3 can be attributed to the worker's limited presence in the video, which amplifies relative errors.

The generated CBCs reveal enforced idleness and work imbalance, supporting productivity improvement strategies during meetings. Extending activity level outputs to CBWS further enables high-level monitoring of worker contributions within a crew. Such monitoring enables supervisors to identify any sustained deviations from the expected value-adding work, allowing earlier intervention compared to post-hoc productivity reporting. Overall, the framework enables automation of Lean tools such as CBC and CBWS, supporting scalable crew-level productivity monitoring and continuous improvement.

6 Ethical Considerations

The video data used in this study were collected for manual observation of construction activities. To preserve worker anonymity, no images or visual data are presented in the manuscript. The analysis is limited to activity-level information, and no individual worker identities are used or reported.

The framework is aligned with Lean Construction Principles, where the activity classifications are intended for waste identification and process improvement. Periods labeled as "Idle" or "Non-Value-Adding" should be interpreted as indicators of workflow constraints (e.g., work front unavailability) and not as worker inefficiency.

A tolerance for short durations of idle activity should be accounted for allowing the natural rhythm of the worker. The framework is not intended for individual-level performance monitoring or punitive use, but for identifying recurring process-level inefficiencies, and outputs should be interpreted at the crew level.

7 Conclusions

This paper proposes a Computer Vision–based Crew Balance Chart (CV-CBC) framework for automatically generating CBCs from construction site videos by integrating worker detection, tracking, and activity recognition. The framework was implemented on the masonry activity and validated through per-second comparison with manually prepared CBCs. The results show that the activity recognition model achieves a macro F1-score of 0.54, while the automated CBC attains an accuracy of 60%. Analysis indicates that the primary source of error arises from confusion between temporally similar activities, Laying Blocks and Applying Mortar, which share similar motion cues. Despite these results, the study has several limitations. The dataset is relatively small and was not originally captured for computer vision–based analysis. Furthermore, the evaluation is restricted to a single construction activity, which limits the extent to which the results can be generalised. Moreover, the inclusion of No-Contact periods can inflate overall CBC accuracy, highlighting the need for more representative CBC-level evaluation metrics.

Future work will focus on implementing the framework on larger, more diverse datasets and on multiple construction activities. Incorporating vision–language models to capture richer spatial context, such as identifying the tools and materials workers handle, could enable zero-shot activity recognition. Given their strong ability to learn temporal information, video transformers can be ensembled with vision-language or keypoint-based models to improve robustness and better distinguish visually similar activities. Overall, the proposed framework highlights how computer vision can support crew-level productivity monitoring and productivity improvement.

References

[1] J. Mischke, K. Stokvis, K. Vermeltfoort, and B. Biemans. Delivering on construction productivity is no longer optional. On-line: https://www.mckinsey.com/capabilities/operations/our-insights, Accessed: 29/03/2026.

[2] P. E. Eriksson and H. Szentes. Managing the tensions between exploration and exploitation in large construction projects. *Construction Innovation*, 17(4):492–510, 2017.

[3] S. Loganathan and P. Forsythe. Unravelling the influence of teamwork on trade crew productivity: a review and a proposed framework. *Construction Management and Economics*, 38(11–12), 2020.

[4] A. S. da Barbosa and D. B. Costa. Productivity monitoring of construction activities using digital technologies: a literature review. In *Proceedings of the International Group for Lean Construction, pages 707–716, 2021.*

[5] W. Shou, J. Wang, P. Wu, and X. Wang. Value adding and non-value adding activities in turnaround maintenance process: classification, validation, and benefits. *Production Planning and Control*, 31(1):60–77, 2020.

[6] M. J. Toledo, M. Lorca, and M. Mora. Use of computer vision and object detection to measure construction productivity using crew balance charts. In *Proceedings of the International Group for Lean Construction,* pages 810–821, 2025.

[7] J. A. Kuprenas and A. S. Fakhouri. A crew balance case study—improving construction productivity. *CM eJournal*, 1(1):10–27, 2001.

[8] C. H. Oglesby, H. W. Parker, and G. A. Howell. *Productivity Improvement in Construction*. 1989.

[9] L. Joshua and K. Varghese. Accelerometer-based activity recognition in construction. *Journal of Computing in Civil Engineering*, 25(5):370–379, 2011.

[10] B. Sherafat et al. Automated methods for activity recognition of construction workers and equipment: state-of-the-art review. *J Constr Eng Manag*, 146(6):03120002, 2020.

[11] J. Li *et al.* A review of computer vision-based monitoring approaches for construction workers' work-related behaviors. *IEEE Access*, 12:7134–7155, 2024.

[12] D. Roberts and M. Golparvar-Fard. End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. *Autom Constr*, 105:102811, 2019.

[13] A. Ghelmani, G. Torabi, A. Hammad, and C. Chen. Single-stage spatiotemporal activity recognition of excavators: a case study. In *Proceedings of the International Symposium on Automation and Robotics in Construction*, pages 800-807, 2024.

[14] D. Roberts, W. Torres Calderon, S. Tang, and M. Golparvar-Fard. Vision-based construction worker activity analysis informed by body posture. *Journal of Computing in Civil Engineering*, 34(4):04020017, 2020.

[15] J. Kim, J. Hwang, I. Jeong, S. Chi, J. O. Seo, and J. Kim. Generalized vision-based framework for construction productivity analysis using a standard classification system. *Autom Constr*, 165:105504, 2024.