ViTPoseActivity: A Multifaceted Computer Vision Approach to On-Site Activity Monitoring

Fabian Pfitzner (corresponding author), (fabian.pfitzner@tum.de)
Chair of Computational Modeling and Simulation,
Technical University of Munich, Germany

Alex Braun, (alex.braun@tum.de)
Chair of Computational Modeling and Simulation,
Technical University of Munich, Germany

Frédéric Bosché, (f.bosche@ed.ac.uk)
School of Engineering, University of Edinburgh, Scotland, UK

André Borrmann, (andre.borrmann@tum.de)
Chair of Computational Modeling and Simulation,
Technical University of Munich, Germany

Keywords: activity reasoning, vision transformer, construction monitoring, labor productivity

Abstract

In response to the activity-based productivity concerns in construction environments, we developed a multifaceted computer vision approach merged with BIM models. High-level process information is derived from continuously acquired site images by the following computational processing chain: (1) Worker activity is classified using the proposed vision transformer network ViTPoseActivity, leveraging human pose features to detect worker activities. (2) On-site labor activities are analyzed according to their on-site impact and fused with the corresponding BIM geometry. Our model, ViTPoseActivity, achieved 92.31% accuracy while surpassing previous prediction speeds, demonstrating an effective trade-off between computational cost and precision in activity analysis. Unlike previous studies, our approach was deployed on a large real-world dataset, carefully investigating subtasks and affording productivity insights on reinforcement activities. Integrating as-performed and geometry information supports construction management by facilitating better decision-making regarding worker group definition and task allocation. Our research fills a crucial gap by providing a robust and efficient method to assess on-site labor productivity.

1. Introduction

Computer vision has been shown to be a suitable approach to progress monitoring in construction. While significant advancements have been demonstrated in documenting the construction phase based on building progress, less focus has been given to worker activity itself (Reja et al., 2022). For example, construction progress tracking has revolved around discrete events, such as the completion of new building storeys between two distinct states (Braun et al., 2020). Nonetheless, it remains unclear why specific on-site activities are carried out more efficiently than others. There is significant potential in increasing the temporal resolution of on-site activity monitoring in order to enable work performance to be understood more precisely. Quantified measurement of work progress is the key requirement for productivity assessment of construction resources and supports project managers in controlling productivity during construction activities (Pal & Hsieh, 2021). The question of how many resources, in terms of the number of workers and equipment, are required for individual construction activities must be addressed to allow a detailed performance evaluation (Pal & Hsieh, 2021). This paper investigates to which extent these unknown parameters can be determined in an automated manner using activity monitoring using remote sensing. Activity Monitoring is the task of computing process-related information, like elapsed time and required resources, acquired through data acquired over time on site.

2. Related work

2.1. Vision transformers for pose detection

While the Transformer architecture is widely adopted in natural language processing, its use in computer vision is still limited (Dosovitskiy et al., 2020). Specifically in large-scale image recognition, classic Convolutional Neural Network (CNN) architectures represent the state-of-the-art, with most algorithms pre-trained on a large dataset and fine-tuned on a smaller, task-specific one (Mahajan et al., 2018). Like CNNs, Transformer-based models are often pre-trained on large datasets and then fine-tuned for the task at hand (Dosovitskiy et al., 2020). Although CNNs have been the de-facto standard in computer vision, (Vision-) Transformers offer key advantages in capturing global dependencies and contextual understanding, going beyond CNNs' local feature extraction limitations. A scenario where global context is significant is pose detection: Human pose detection is the task of identifying and localizing the keypoints of individuals based on the body's anatomy. Detecting people is an enormous challenge due to variations in appearance (Forsyth & Ponce, 2012) and complex interactions (Cao et al., 2016).
Compared to CNNs, Vision-Transformers process image information patch-wise. The 2D images $x \in \mathbb{R}^{H \times W \times C}$ are converted into a sequence of flattened 2D patches, also known as patch embedding, $x_p \in \mathbb{R}^{N \times (P^2 C)}$, where (H, W) is the image resolution, C is the channel number, (P, P) is the patch resolution, and $N = \frac{H W}{P^2}$ the sequence length for the Transformer (Dosovitskiy et al., 2020). The recently introduced ViTPose (Xu et al., 2022) outperforms existing methods on the MS COCO Keypoint Detection benchmark, setting a new state-of-the-art by reaching 80.9% average precision (AP). Specifically, thanks to the model structure, a very large flexibility and transferability of knowledge between models is enabled.

2.2. Human-centered activity monitoring in construction

Human-centered activity monitoring is the process of interpreting and understanding human actions and behaviors within a given context to support construction management decision-making. Previous research has developed diverse methods to enhance the reasoning during the construction phase and enable automatic resource monitoring. Khosrowpour et al. (2014) demonstrated first how body postures could be acquired and processed to determine activity rates of construction workers using a Microsoft Kinect Sensor, codebooks, and Support Vector Machines (SVM) classifier. Yang et al. (2016) improved the approach by using data from widespread cameras instead of relatively constrained Kinect Sensors. They applied various image descriptors (HoG, HoF, MNH) to extract features from the image, which were then mapped to the codebook allowing substantial performance and accuracy improvements.
The introduction of CNNs enabled a vast performance and accuracy improvement in de-

tecting worker activities and facilitating real-time monitoring. H. Luo et al. (2018) used a VGG-16 model on three different input streams: RGB, Optical Flow, and Gray Stream to classify the on-site activities. To fuse the results, they applied a one-step reinforcement learning model. X. Luo et al. (2018) showed a similar approach using a CNN (FlowNet 2.0) based on two streams (spatial and temporal) to reason about construction activities. Torabi et al. (2022) further improved previous approaches using a 3D CNN, overcoming the computational and accuracy limitations of previously presented approaches by creating an end-to-end trainable method.

The most recent work by L. Xiao et al. (2024) demonstrates CV-based activity monitoring for construction environments using Spatial-Temporal Graph Convolutional Networks (ST-GCNs). They used OpenPose, an open-source library for real-time multi-person keypoint detection and pose estimation, to generate a sequence of skeletons from RGB images. The ST-GCN was then applied to reason about the activity. Sun et al. (2024) show with their work how 3D body pose information can be used to avoid long-term work-related illnesses. With their novel feature processing method, they are amongst a few that address computational challenges of real-time posture recognition. However, their method relies on high-quality sensor data derived from IMU-based motion capture systems, e.g., the in-lab produced 3D keypoint dataset (Tian et al., 2022), which may not represent real-world conditions.

While the above works show that there have been vast improvements over time, two critical aspects have so far received insufficient attention, impacting real-world applicability: Computational efficiency and actual productivity assessment. The computational effort required for activity monitoring has significantly decreased over the last few years. However, even with the fastest processing algorithms, processing daily activity data based on high-frequency video streams with multiple targets to track in real-time is highly computationally expensive and not feasible with today's GPU hardware. Therefore, the question deserves attention whether a lower frame rate and image number can be applied to monitor parts of on-site processes, achieving reasonable computation times. On the other hand, previously introduced methods often come too short of real-world application. Approaches to utilize the output of published methods for actual productivity monitoring and subsequently supporting construction management for resource planning are surprisingly rare. Some researchers (Torabi et al., 2022; Yang et al., 2015) suggest enhancing process knowledge with resource and geometry data from BIM models. However, research demonstrating a comprehensive implementation of this concept does not exist yet. Our approach, presented in the following, targets both research gaps, introducing a novel deep-learning-based approach to detect workers' activities and fusing it with BIM information to analyze on-site productivity.

3. Methodology

3.1. Scope

The proposed method is illustrated in Fig. 1. Initially, object detection networks are used to process image sequences of the construction site as work progresses. The detected items are mapped onto the BIM model using a geometric approach and a knowledge graph. This process, developed in prior work by Pfitzner et al. (2024), sets the basis for analyzing on-site productivity. Here, we propose a vision transformer designed to monitor human-centered activities, enabling insight into on-site work productivity.

This study uses a low-frequency frame rate, excluding video-based features like optical flow, to reduce computational effort. Construction processes are recognized through body postures in still images. Our approach is designed for construction processes where activities are identifiable by distinct body positions.

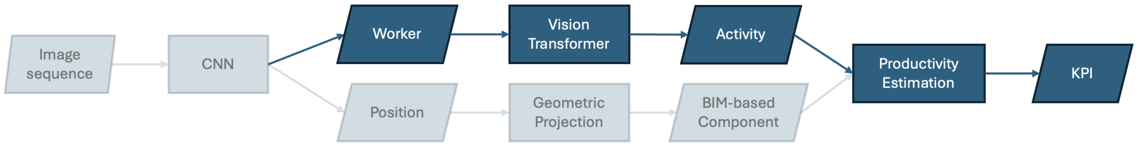


Figure 1: Overview of the method (light-blue: Steps conducted in earlier work (Pfitzner et al., 2024), dark-blue: steps developed in this research).

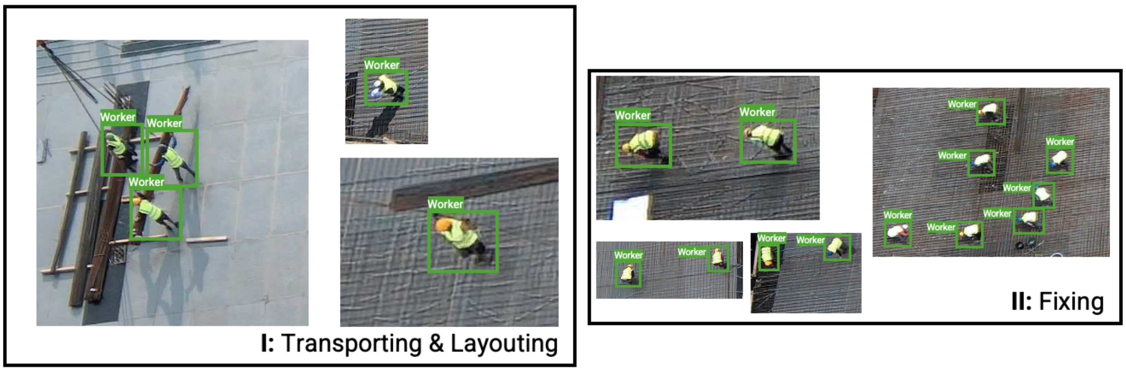


Figure 2: Phase I and phase II of the reinforcement process.

Reinforcement work involves activities that notably correlate with body posture. Therefore, worker poses are utilized to identify worker activities. The reinforcement process contains three phases: I. Layouting and Transporting, II. Fixing, III. Quality Control. In step I, reinforcement bars and meshes are transported and layouted on the ground using cranes or similar equipment. During this step, workers continuously transport materials and are, as such, primarily in an upright position, as depicted in Fig. 2 (left). To facilitate the network's ability to distinguish between transporting and layouting, workers transporting materials are identified by a walking posture with arms extended from the body, while those engaged in layouting are identified by a standing position with the arms close to the body.

During step II, workers fix and secure the reinforcement bars. Fixing is identified by a bent-over position, illustrated in Fig. 2.

Step III, Quality control, is done by construction management and requires comparatively less capacity. Therefore, the focus will be on the first two phases. Gangs of workers are generally split up and assigned to phases I and II. Selecting a crew size is crucial to ensure a stable construction flow. If the first phase takes too long, the iron workers must wait before they can start fixing. On the other hand, when the iron crew is chosen too small, the fixing phase can significantly slow down subsequent processes.

3.2. Human-centered activity reasoning

The chosen DL architecture for human pose detection is a vision transformer, as it surpasses CNNs in global feature detection, as discussed in section 2.1. The model's architecture, designed for human-centered construction activity classification, is shown in Fig. 3. In addition to the vision-based transformer backbone (Xu et al., 2022), including a classic decoder (B. Xiao et al., 2018) to extract and localize keypoints, a feature extraction block, and several fully connected layers to reason about the construction activity are embedded. ReLU and Softmax activation functions are employed to enable non-linear learning. During training, a dropout layer is used to prevent overfitting.

The model supports generating additional features based on the body pose while re-

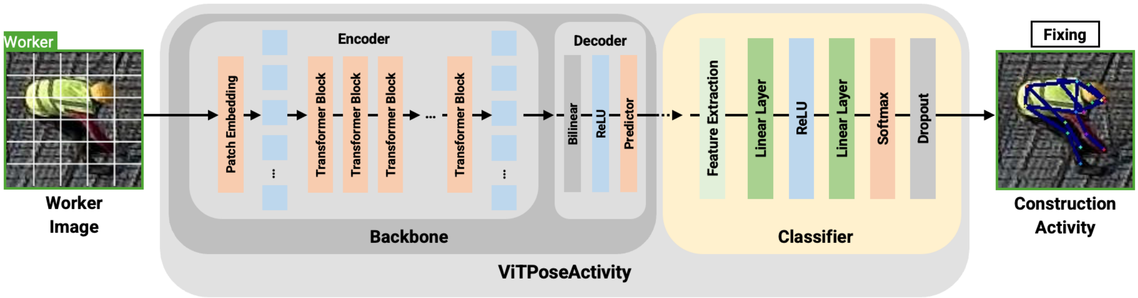


Figure 3: Proposed network architecture ViTPoseActivity for human-centered construction activity reasoning: Encoder based on Xu et al. (2022), Decoder based on B. Xiao et al. (2018), and feature-engineered Classifier.

ceiving transparency and control. Building a robust deep-learning architecture requires exploring different algorithms for classifying and engineering diverse features. In order to achieve robust performance while keeping the computational effort low, we use classical Machine Learning (ML) algorithms, like Support Vector Machines (SVM), for feature engineering before incorporating these features into the Deep Learning model.

Index	Key point
0	Nose
1	Left eye
2	Right eye
3	Left ear
4	Right ear
5	Left shoulder
6	Right shoulder
7	Left elbow
8	Right elbow
9	Left wrist
10	Right wrist
11	Left hip
12	Right hip
13	Left knee
14	Right knee
15	Left ankle
16	Right ankle

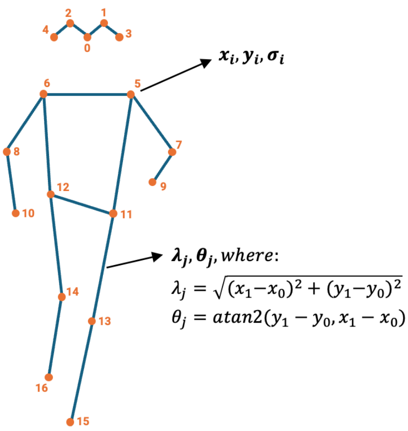


Figure 4: Skeleton-based keypoints and body composition for feature engineering.

Feature engineering is conducted on the vision-transformer outputs, which are detected human keypoints represented as skeleton maps, as illustrated in Fig. 4. The maps include the keypoints' position, the confidence score, and the skeleton composition. In addition to the keypoints, limb length (λ_j) and direction (θ_j) are computed and included as features. Providing the DL architecture with tailored posture features allows the model to learn and adjust according to the human-centered activity classification task. The effectiveness of ViTPoseActivity is validated by comparing its performance with leading classification networks like ResNet and VGG (see Section 4.2.1).

3.3. Productivity estimation

Labor productivity represents the relationship between inputs, such as labor hours, and outputs, like quantity of building components (Hofstadler, 2014). For a detailed productivity understanding, subtasks must be considered so that the bottlenecks of processes causing reduced productivity can be identified. In addition, the value work activity adds to the final product varies. Previous work has approached this by differentiating between direct, indirect, and waste work (Jacobsen et al., 2023; Park et al., 2005): Direct

work contributes directly to the output, indirect work is necessary to conduct direct work, and waste does not add value. In our approach, fixing activities are considered direct work, while layouting and transporting are viewed as indirect work.
We suggest measuring productivity by investigating individual subtasks of construction processes, which differentiate in their contribution to the final product. This fine-granular analysis allows for a more profound understanding of the process. Besides, the geometric projection method developed in prior work (Pfitzner et al., 2024) is used to merge the geometry with detected activities. This enables our approach to provide enriched process insights. Detailed exploration of the impact of specific building components' details is beyond this study's scope but will be addressed in the authors' future work.

4. Experiments

4.1. Data and Setup

To develop and evaluate our deep learning model, we utilized a comprehensive image dataset collected from various construction sites, with images acquired by fixed crane-mounted cameras continuously every 30 seconds over several months. The object detection network and a knowledge graph were employed to locate the workers and cut the images into patches, as depicted in Fig. 5. For training and validation, a dataset comprising 329 samples of the work activities *layouting*, *fixing*, and *transporting* was divided into an 80/20 split.

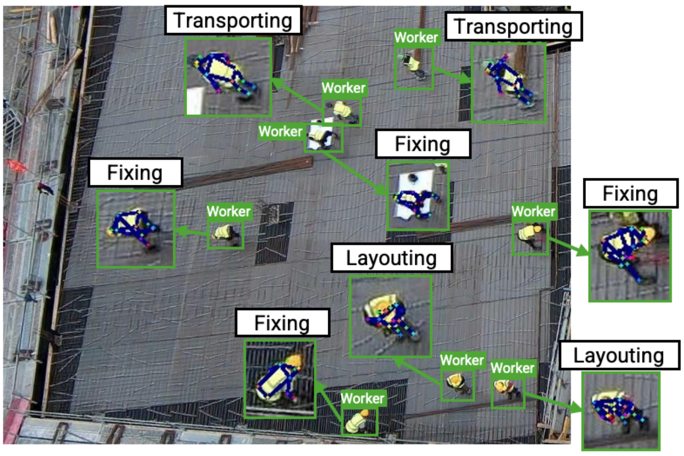


Figure 5: Reinforcement work with diverse construction activities.

We used a pre-trained ViTPose backbone to generate the human keypoints from the images. Support Vector Machines (SVM), random forest, logistic regression, and decision trees were used for feature engineering. Applying GridSearchCV and cross-validation facilitated comprehensive experimenting with different training and validation splits and diverse feature sets. Subsequently, the best-performing feature combination was integrated into ViTPoseActivity. During model training of ViTPoseActivity, the backbone weights of the pose estimator were frozen. The training parameters were defined as follows: Epochs: 150; Learning rate: 0.001; Weight decay: 0.0001; Dropout-rate: 0.5; Loss-function: CrossEntropy. The ResNet-152 and VGG-19 classification models were trained and tested on the same dataset using the same training parameters and loss-function. Image augmentation techniques like scaling, rotating, and resizing were applied for all approaches. Finally, the ViTPoseActivity model was tested on a larger dataset containing 10,020 images of workers. As shown in Fig. 6, this dataset represents a section of reinforcement work on the first-floor slab of a real-world construction project, which took a total of three days (21 labor hours).

The workers' positions were projected to the BIM geometry using the pre-established camera internal and external calibration matrices. The reinforcement area was divided into a grid with a 1.8x1.8 meter cell size for a more granular location-based analysis.

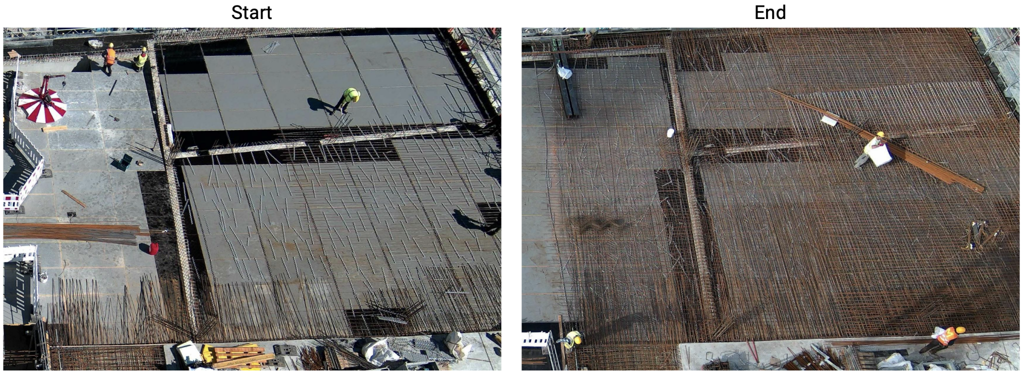


Figure 6: Crane camera view of the start and end of the slab reinforcement work.

4.2. Results

4.2.1. Activity reasoning

Table 1: Results of ML-based feature engineering.

	kp position	kp position + confidence	kp position + confidence + body composition
best model	SVM	SVM	SVM
precision	0.82	0.83	0.85
recall	0.80	0.82	0.85
f1-score	0.81	0.82	0.85

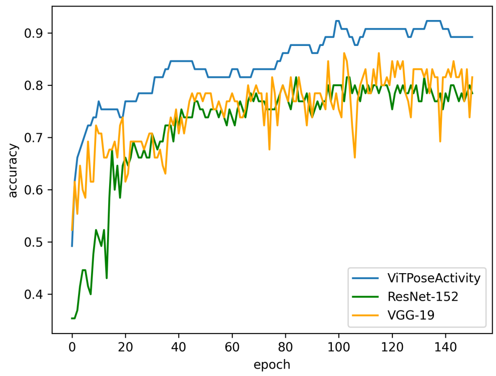
Table 1 displays the feature set outcomes obtained from the classical ML algorithms and obtained by the GridSearchCV approach. SVM emerged as the top performer, which was consistent with prior studies on body pose estimation (Khosrowpour et al., 2014; Yang et al., 2016).

The results shown in the different columns of Tab. 1 demonstrate that the additional features do not confuse the network and enable additional learning capabilities. Incorporating keypoint confidence marginally enhanced overall precision, whereas integrating body composition features led to a notable improvement in accuracy.

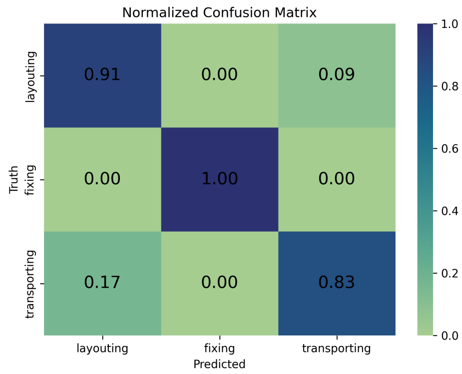
Table 2: Results comparison of DL models.

	precision	recall	f1-score
ViTPoseActivity	**0.92**	**0.92**	**0.92**
ResNet-152	0.82	0.82	0.82
VGG-19	0.86	0.85	0.84

The ViTPoseActivity accuracy during training compared to ResNet-152 and VGG-19 is shown in Tab. 2 and Fig. 7a. Due to the finetuned human pose features, our ViTPoseActivity model has significantly better learning effectiveness. In addition, in the two CNN classification networks VGG-19 and ResNet-152, more noise during training could be



(a) Test accuracy comparison.



(b) Normalized confusion matrix of ViTPoseActivity.

Figure 7: Accuracy and confusion matrix of ViTPoseActivity.

detected, highlighting the networks' uncertainty. Fig. 7b shows the confusion matrix of ViTPoseActivity. The similarity between the classes *transporting* and *layouting* caused slight confusion: In the test set, 3 out of 18 samples were misclassified as *layouting*, and 2 out of 23 samples were misclassified as *transporting*. Nonetheless, the total number of false positives remained relatively small, with 5 out of 66.

4.2.2. Productivity estimation

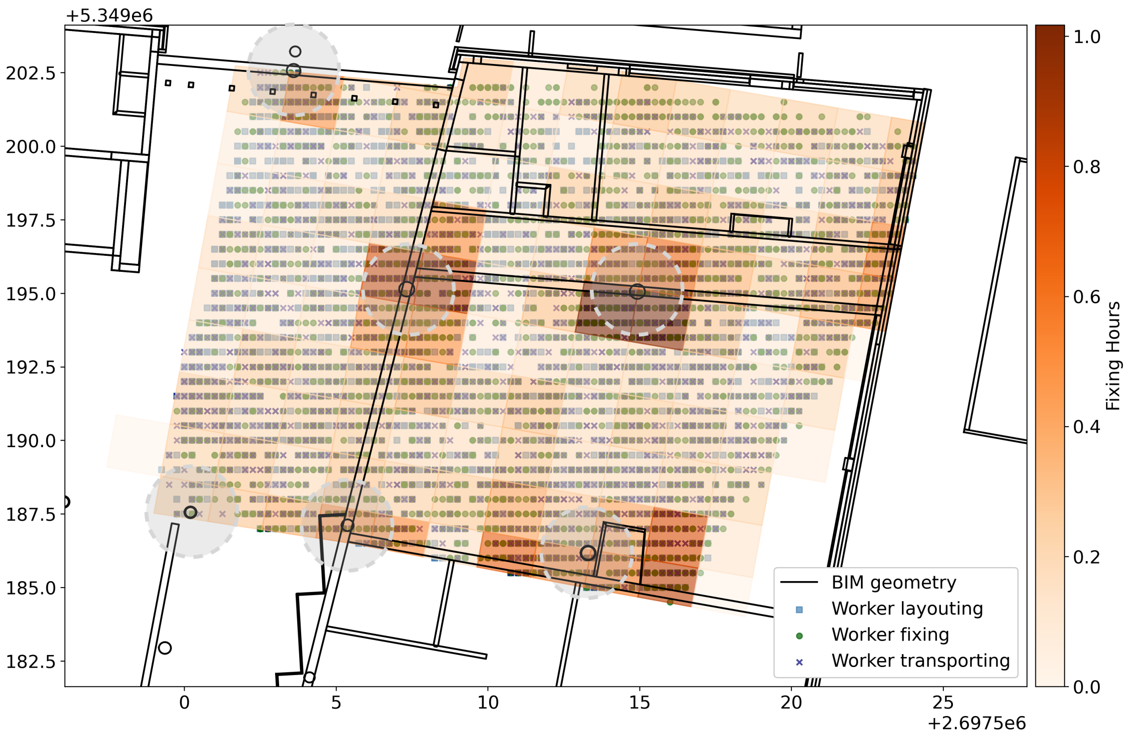


Figure 8: *Fixing* hours spent reinforcing the first-floor slab; connecting areas of load-bearing columns are marked in light-gray.

Processing the reinforcement dataset (10,020 images), covering multiple workers per timestamp, took 276.3 seconds on a Nvidia RTX 8000, resulting in a processing time of 0.028s per image. Fig. 8 illustrates the distribution of *fixing* hours throughout the 21-hour production period. The *layouting, transporting and fixing* labor hours are computed based on the number of hours within the grid cells. In addition, the workers' activities and locations are highlighted.

Our thorough investigation of the daily reinforcement progress, as shown in Fig. 8, has revealed a key observation, taking the *fixing* heatmap into account: Although the distribution of workers appears even, the heatmap reveals differences: the highest density of *fixing* occurred at the connections of load-bearing elements, e.g., columns, highlighted in light-gray, whereas less direct work was conducted elsewhere. The high range of *fixing* hours suggests more labor effort required close to connections. This could be attributed to the relative complexity of reinforcement around connected load-bearing elements, which requires connecting the reinforcement of the new element (here a slab) to the reinforcement of the existing ones (here columns in particular).

5. Discussion and future work

Our experiments have shown promising results in accurately predicting worker activities based on human pose analysis and estimating the productivity of reinforcement work on construction sites. The convincing results suggest ViTPoseActivity could be succesfully applied to other processes like bricklaying and plastering (Roberts et al., 2020). While our method has demonstrated its effectiveness, limitations must be addressed. The geometric projection accuracy is restricted according to the image quality and camera placement. Our approach, designed for computational efficiency, uses less data. But, using less data does not enable detecting complex workflows that would be identifiable with detailed video data and additional features.

However, as shown in the introduced example (Fig. 8), detailed video features are not necessary in every case to identify critical aspects of construction processes. We achieved a computation time per frame of 0.028 seconds. Though this seems slightly better than the fastest known activity prediction network in the domain at 0.04 seconds per frame (X. Luo et al., 2020; Torabi et al., 2022), our method outperforms it significantly. By requiring only one frame instead of 15 FPS, our approach is 20x faster compared to the three-stage method (Torabi et al., 2022) while maintaining a promising detection accuracy of 92.31%. Moreover, using one frame every 30 seconds reduces the processing time by an additional 30 times. This means that, unlike other methods, our approach can be applied in real-time scenarios without major complications.

Lastly, our method helps highlighting process-critical areas that impact overall productivity; in the case of reinforcement, at the connections of load-bearing elements. Given the time-consuming nature of particular tasks within the reinforcement process we have explored, we advocate for future research in this area. Specifically, we suggest leveraging the BIM models to proactively detect these process-critical areas, thereby improving scheduling and resource allocation.

6. Conclusion

This paper introduced a novel, flexible method for identifying worker activities using body postures, delivering insights into construction productivity. Our model, ViTPoseActivity, achieved a 92.31% accuracy rate while surpassing previous prediction speeds, demonstrating an effective trade-off between computational cost and accuracy in activity analysis. We developed our method according to real-world settings, investigating particular tasks according to their contribution to the process, and integrated it with existing BIM data. Moreover, we are among a few other researchers (Jacobsen et al., 2023; H. Luo et al., 2018), who deployed their models on larger datasets (10,020 images) to determine productivity in real-world conditions. This approach allows for detailed analysis of construction processes, identifying on-site productivity bottlenecks more effectively.

Acknowledgements

We thankfully acknowledge Innovation Management Bau GmbH for their financial support and for providing us access to multiple construction sites to collect valuable data. In addition, we would like to thank Siemens Real Estate AG and Max Bögl.

References

Braun, A., Tuttas, S., Borrmann, A., & Stilla, U. (2020). Improving progress monitoring by fusing point clouds, semantic data and computer vision. *Automation in Construction*, *116*. https://doi.org/10.1016/j.autcon.2020.103210

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2016). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. http://arxiv.org/abs/1611.08050

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929v2

Forsyth, D. A., & Ponce, J. (2012). *Computer Vision: A Modern Approach* (2nd ed.). Pearson. https://doi.org/https://dl.acm.org/doi/abs/10.5555/580035

Hofstadler, C. (2014). *Produktivität im Baubetrieb*. https://doi.org/10.1007/978-3-642-41633-0

Jacobsen, E. L., Teizer, J., & Wandahl, S. (2023). An On-body Sensor-based Visual Management Tool for Work Task Progress Monitoring. *EG-ICE*. https://orbit.dtu.dk/en/publications/an-on-body-sensor-based-visual-management-tool-for-work-task-prog

Khosrowpour, A., Niebles, J. C., & Golparvar-Fard, M. (2014). Vision-based workface assessment using depth images for activity analysis of interior construction operations. *Automation in Construction*, *48*, 74–87. https://doi.org/10.1016/j.autcon.2014.08.003

Luo, H., Xiong, C., Fang, W., Love, P. E., Zhang, B., & Ouyang, X. (2018). Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Automation in Construction*, *94*, 282–289. https://doi.org/10.1016/j.autcon.2018.06.007

Luo, X., Li, H., Cao, D., Yu, Y., Yang, X., & Huang, T. (2018). Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks. *Automation in Construction*, *94*, 360–370. https://doi.org/10.1016/j.autcon.2018.07.011

Luo, X., Li, H., Yu, Y., Zhou, C., & Cao, D. (2020). Combining deep features and activity context to improve recognition of activities of workers in groups. *Computer-Aided Civil and Infrastructure Engineering*, *35*(9), 965–978. https://doi.org/10.1111/mice.12538

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der, L., & Facebook, M. (2018). Exploring the Limits of Weakly Supervised Pretraining. https://openaccess.thecvf.com/content_ECCV_2018/html/Dhruv_Mahajan_Exploring_the_Limits_ECCV_2018_paper.html

Pal, A., & Hsieh, S. H. (2021, November). Deep-learning-based visual data analytics for smart construction management. https://doi.org/10.1016/j.autcon.2021.103892

Park, H.-S., Thomas, S. R., & Tucker, R. L. (2005). Benchmarking of Construction Productivity. *Journal of Construction Engineering and Management*, *131*(7), 772–778. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(772)

Pfitzner, F., Braun, A., & Borrmann, A. (2024). From data to knowledge: Construction process analysis through continuous image capturing, object detection, and knowledge graph creation. *Automation in Construction*, *164*. https://doi.org/10.1016/j.autcon.2024.105451

Reja, V. K., Varghese, K., & Ha, Q. P. (2022, June). Computer vision-based construction progress monitoring. https://doi.org/10.1016/j.autcon.2022.104245

Roberts, D., Torres Calderon, W., Tang, S., & Golparvar-Fard, M. (2020). Vision-Based Construction Worker Activity Analysis Informed by Body Posture. *Journal of Computing in Civil Engineering*, *34*(4). https://doi.org/10.1061/(asce)cp.1943-5487.0000898

Sun, X., Li, X., Ren, B., & Chen, J. (2024). Construction posture recognition with primitive joints extended planar normal vector quaternions. *Automation in Construction*, *161*. https://doi.org/10.1016/j.autcon.2024.105356

Tian, Y., Li, H., Cui, H., & Chen, J. (2022). Construction motion data library: an integrated motion dataset for on-site activity recognition. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01841-1

Torabi, G., Hammad, A., & Bouguila, N. (2022). Two-Dimensional and Three-Dimensional CNN-Based Simultaneous Detection and Activity Classification of Construction Workers. *Journal of Computing in Civil Engineering*, *36*(4). https://doi.org/10.1061/(ASCE)CP.1943-5487.0001024

Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11210 LNCS*, 472–487. https://doi.org/10.1007/978-3-030-01231-1

Xiao, L., Yang, X., Peng, T., Li, H., & Guo, R. (2024). Skeleton-Based Activity Recognition for Process-Based Quality Control of Concealed Work via Spatial–Temporal Graph Convolutional Networks. *Sensors*, *24*(4), 1220. https://doi.org/10.3390/s24041220

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv*. http://arxiv.org/abs/2204.12484

Yang, J., Park, M. W., Vela, P. A., & Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, *29*(2), 211–224. https://doi.org/10.1016/j.aei.2015.01.011

Yang, J., Shi, Z., & Wu, Z. (2016). Vision-based action recognition of construction workers using dense trajectories. *Advanced Engineering Informatics*, *30*(3), 327–336. https://doi.org/10.1016/j.aei.2016.04.009