www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

FOUNDATIONS FOR CONSTRUCTION 5.0: A REVIEW-BASED TAXONOMY FOR CONSTRUCTION WORKER ACTION UNDERSTANDING

SUBMITTED: November 2024
REVISED: March 2025
PUBLISHED: June 2025
EDITORS: Yang Zou, Mostafa Babaeian Jelodar, Zhenan Feng, Brian H.W. Guo
DOI: 10.36680/j.itcon.2025.038

Sudheer Kumar Nanduri, Ph.D. Student
Department of Civil Engineering, Indian Institute of Technology Bombay
n.sudheer@iitb.ac.in

Venkata Santosh Kumar Delhi, Associate Professor
Department of Civil Engineering, Indian Institute of Technology Bombay
venkatad@iitb.ac.in

SUMMARY: *Construction worker actions, driven by personal and organizational goals, are vital in handling dynamic and unstructured environments. As the industry advances towards Construction 5.0, integrating automation while maintaining a value-oriented approach necessitates understanding the worker's actions in several dimensions. Action understanding goes beyond recognition to interpreting intentions, predicting future actions, assessing behavior, and many others. Yet, in the adoption, several challenges are faced due to complex action hierarchies, domain-specific applications, and a semantic gap between observations and interpretations. This paper addresses the challenges by developing a hierarchical taxonomy. An initial study is conducted on existing literature, narrowing the scope to the two-dimensional RGB camera-based computer vision as the sensing system. Identifying the lack of a structured approach and the existence of a semantic gap between observed features and their assigned meanings, this work embarks on establishing a unified taxonomy. Following the PRISMA protocol, a dual-literature review is conducted across literature across the domains of computer vision and construction automation, and findings are presented in three steps. In the first step, the review papers in the computer vision field were synthesized to develop a taxonomy essential for action understanding. This taxonomy outlines a four-step approach essential for action understanding. In the second step, the construction automation literature is reviewed, and the extant literature is mapped to the taxonomy established. In the third step, a discussion is presented on the current state-of-the-art approaches, the missing elements, the possible future directions specific to different parts of the taxonomy, and the integration with current technologies. Along with the future directions, suggestions also include use cases for the construction industry to improve upon core values in line with Construction 5.0.*

KEYWORDS: *Construction 5.0, Worker Action Understanding, Taxonomy.*

REFERENCE: *Sudheer Kumar Nanduri & Venkata Santosh Kumar Delhi (2025). Foundations for Construction 5.0: A Review-based Taxonomy for Construction Worker Action Understanding. Journal of Information Technology in Construction (ITcon), Special issue: 'Construction 5.0', Vol. 30, pg. 924-962, DOI: 10.36680/j.itcon.2025.038*

1. INTRODUCTION

Construction workers are the driving force behind the construction processes, intentionally operating in dynamic and often unstructured environments to achieve their personal and organizational goals. The worker actions do not merely represent mechanical movement but enable achieving their intentions and goals embedded within the specific process workflows and contexts. In addition to the organizational goals, the worker's personal conditions drive their behavior at the workplace. Given the hazardous and physically demanding nature of the construction work, an automatic understanding of worker actions becomes imperative for the industry to ensure the productivity, safety, and well-being of workers and projects. Such automation has to go beyond just recognition of actions to actually understand them - Interpreting the intentions behind the actions (Blakemore and Decety, 2001), Predicting future actions (Koppula and Saxena, 2016), Assessing behavior (Pentland, 2007; Cristani *et al.*, 2013) are some practical applications. Such advancements from 'action recognition' to 'action understanding' have long been the focus of extant research, especially in the domains of artificial intelligence in robotics and their application to controlled production environments (Demiris, 2007; Zhang *et al.*, 2022). Given the dynamic environment of construction, the development of action understanding models is a significant need for integrating technological developments and improving project metrics like safety and productivity.

With the rapid advancement of technology, the construction industry is increasingly integrating digital tools and automation across the project life cycle, as indicated in Construction 4.0 (Karmakar and Delhi, 2021). The emergence of Industry 5.0 – extending to Construction 5.0 – emphasizes a value-driven approach to technological integration, prioritizing the core values of human-centricity, sustainability, and resilience (European Commission *et al.*, 2021). Considering the essential role of workers in construction, Construction 5.0 aligns with all three values by positioning them at the core of technological advancements. For instance, robots can form teams with workers (Baskaran and Adams, 2023), and automation can help anticipate worker safety risks, ensure workplace resilience, and maintain workflow sustainability. In the evolving landscape of construction automation, applications of action understanding enhance the perception capabilities of different types of technologies (You, Zhou and Ding, 2023), keeping humans at the core, thus making action understanding a key component in adopting Construction 5.0.

Translating the research advancements in action understanding to the construction industry presents significant challenges primarily due to the complex action hierarchies and the organizationally driven goals rather than individual choice alone. Additionally, different applications for metrics like productivity or safety require different approaches. A safety-oriented process will necessitate a specific set of actions and intentions, which differ considerably from productivity-oriented actions. Similarly, various technologies like robots or remote monitoring applications adopt action understanding differently. These applications have commonalities originating from similar works in artificial intelligence research. Thus, building a construction-specific approach for developing tools is necessary while also recognizing the commonalities from the ground up. A systemic approach is needed for integrating structured taxonomies, ensuring relevance across application use cases, and closely aligning with the core values of Construction 5.0 in the industry.

In addition, researchers also identify a 'semantic gap' between low-level action primitives observed on the field and high-level semantic interpretations of the same actions (Zhong *et al.*, 2019; Paneru and Jeelani, 2021). Different supportive components are proposed, including Ontologies and Rule engines (Zhong *et al.*, 2019); Additional technologies and BIM models (Fang, Ding, *et al.*, 2020); Knowledge graphs (Liu and Jebelli, 2022); and Prior knowledge and Posterior inference models (Xu *et al.*, 2021). However, there is no integrated understanding of where and how these components add value to overcoming the semantic gap. This necessitates a hierarchical taxonomy approach, from low-level primitives to high-level interpretations.

In summary, action understanding is a foundational capability that enables a value-oriented approach to adopting different technologies in the construction industry. However, the existing action understanding approaches are limited in their applicability to construction due to the industry-specific action hierarchies, domain-specific use cases, and a lack of a unified framework for integrating various supporting technologies to bridge the semantic gap. This paper proposes to build a hierarchical taxonomy tailored to address these challenges, using an in-depth literature review integrating findings from artificial intelligence and construction automation. The scope and boundaries for the review are established based on the existing studies, detailed in the literature review section. Overall research methodology and PRISMA reporting are presented in the methodology section. The taxonomy is presented in the results section, and studies in construction automation are tagged within the taxonomy in its sub-

section. The discussion section highlights the current state of the art and future directions for research in two layers – underlying algorithms and utilizing technologies.

2. LITERATURE REVIEW

Research on action understanding can focus on equipment and worker actions. Equipment actions are largely deterministic, making action understanding useful for applications such as productivity monitoring (Chen, Zhu and Hammad, 2022). In contrast, workers and operators possess expertise, adaptability, and decision-making capabilities, making their actions uniquely complex. Focusing on understanding worker action is crucial for developing human-centric technologies aligned with Construction 5.0.

In the first step of data collection, earlier studies have explored frameworks (Calvetti *et al.*, 2020) and mapped existing sensors (Gao *et al.*, 2022) for workers. Yet, recent advancements in the sensor domain, such as the 4D mmWave sensor, highlight continuous improvements in the information collection capabilities and subsequent construction applications (Han *et al.*, 2023). Therefore, rather than focusing on sensor technologies, this work centers on using a single technology while sustaining the generalizable approach across different applications.

A prior review of activity recognition (Sherafat *et al.*, 2020) categorizes the methods into audio-based, video-based, and kinematics-based approaches. While this classification provides a useful starting point, it does not account for theoretical approaches to action understanding. Among the methods, video-based approaches excel in capturing worker movements and interactions with surrounding tools and objects, making them particularly effective for comprehensive action understanding. Therefore, this paper focuses on video-based action recognition as it provides the most comprehensive insight into construction activities with minimal preprocessing.

While multiple technologies exist within the computer vision field (Stereo vision, RGB-D cameras), this work limits its scope to only regular cameras with 2D images collected over time. By focusing on regular 2D cameras, this study ensures broader applicability in real-world construction scenarios where such equipment is prevalent while generating insights that can be extended to advanced systems like stereo and RGB-D cameras.

Classic (Arashpour, Ngo and Li, 2021) and modern (Paneru and Jeelani, 2021) computer vision (CV) methods are useful in field applications and also for offsite production (Alsakka *et al.*, 2023). Of the different field applications, progress monitoring (Yang *et al.*, 2015; Pal *et al.*, 2023; Moragane *et al.*, 2024), safety & health monitoring (Seo *et al.*, 2015; Fang, Love, *et al.*, 2020; Liu *et al.*, 2021; Guo *et al.*, 2021) are the most touched upon topics, and newer approaches like quality control (Wang *et al.*, 2021) are also being presented. Computer vision has the potential for scene-based, location-based, and action-based risk identification (Seo *et al.*, 2015), typically utilizing the methods of detection, localization, tracking, and action recognition (Li *et al.*, 2024). However, these applications remain fragmented, focusing primarily on pattern recognition and lacking deeper insights into worker intentions, reasoning, and other cognitive aspects.

Bridging this gap requires drawing from cognitive and neurological research, which has explored how humans perceive, comprehend, and project actions in the real world (Guo *et al.*, 2021). These processes align with computer vision tasks such as detection, assessment, and prediction, indicating better potential for modeling worker actions. Ongoing research has sought to replicate these capabilities in industrial robots (Bonci *et al.*, 2021), yet their application to human-centered construction environments remains limited. Additionally, neurological studies explored the vision-language connection, highlighting how humans connect the visual and language aspects (Willems, Özyürek and Hagoort, 2007). This connection could support more advanced action interpretation while tackling the semantic gap between low-level observations and high-level understanding. Theoretical approaches like the 3Rs - reconstruct, recognize, and reorganize (Wiriyathammabhum *et al.*, 2016) offer additional frameworks for structuring action understanding over time, but their adaptation to dynamic environments like construction remains underexplored.

In summary, various studies have explored action understanding in construction, focusing on machinery and human workers using different types of sensors, and computer vision-based worker action understanding applications stand out to be more helpful. Despite significant advancements in computer vision-based methods, the applications remain limited in their ability to infer worker intentions and contextualize within workflows. Furthermore, while cognitive research offers promising theoretical models, these have not yet been integrated into construction automation research.

To address these gaps, this study:

1. Adopts a dual-literature review approach, systematically integrating insights from both computer vision and construction automation research.
2. Develops a hierarchical taxonomy tailored for construction worker action understanding.
3. Evaluates the current practices in construction and identifies challenges and opportunities for future research.

The following methodology section outlines the approach used in detail.

3. METHODOLOGY

A three-step process is set out for this review, presented in Figure 1. In the first step, a hierarchical categorization is established from past works in computer vision. The literature on computer vision is reviewed to identify the broad categories that are necessary to build a taxonomical structure. In the second step, the construction domain works are tagged following the hierarchy. The discussion section in the third step presents the current state-of-the-art findings, how well they fit into the taxonomy, what is missing, and future directions. For both reviews, the PRISMA protocol is followed for documentation.

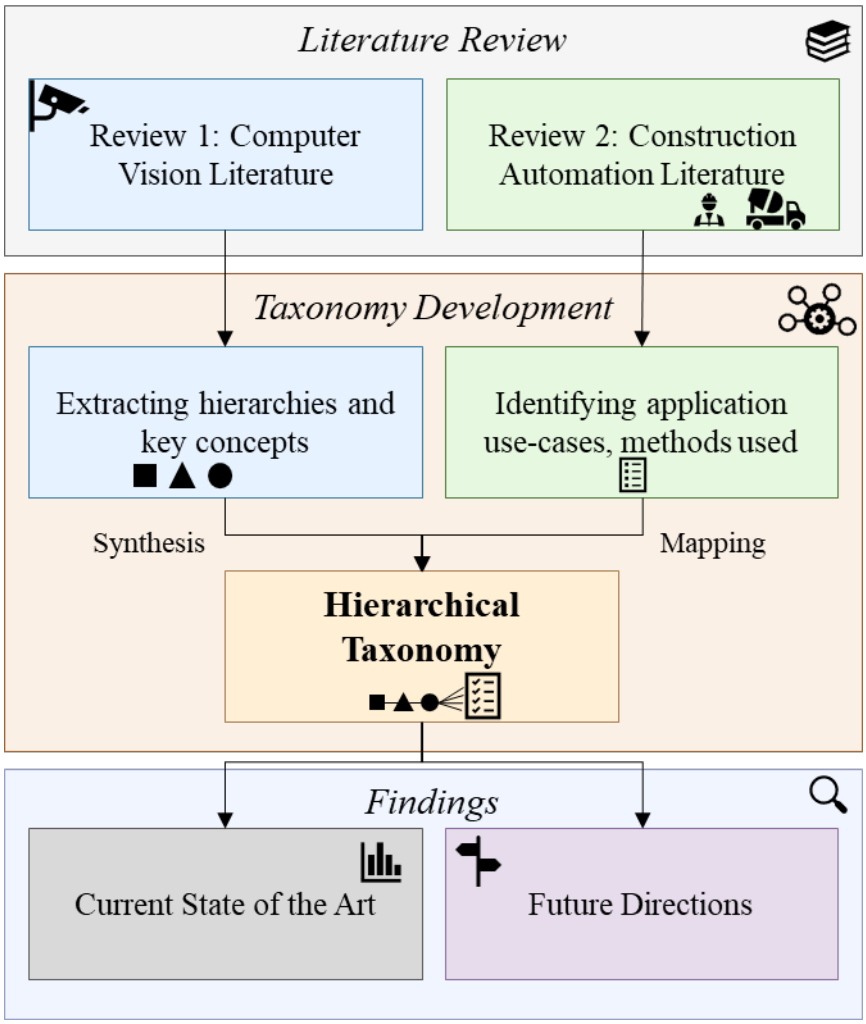

Figure 1: Overall Methodology.

Since the objective in the first step is to derive a categorization, a search was conducted in the IEEE Xplore, ACM Digital Library, Web of Science, and Scopus databases. These databases were chosen due to their extensive coverage of computer vision and construction literature, ensuring a comprehensive review. As the initial search for

'action understanding' yielded no results in IEEE, the search terms were adjusted to 'activity recognition' or 'action recognition. The search term (ALL = (“activity recognition” OR “action recognition”) AND ALL = (review) AND (From 2000 to 2023) is used. The timeframe from 2000 to 2023 was selected to capture the evolution and advancements in action recognition research both using classical and modern computer vision methods. Within each database, the search string varied as they have different service providers. However, only these terms are utilized. Only review papers that align with action understanding are identified by manual verification of title and abstract. This boiled down a database of 4071 papers to 58 documents for use. These papers identify the essential components of an intelligent system, the use cases, and the problems. The PRISMA flowchart is provided in the figure below.

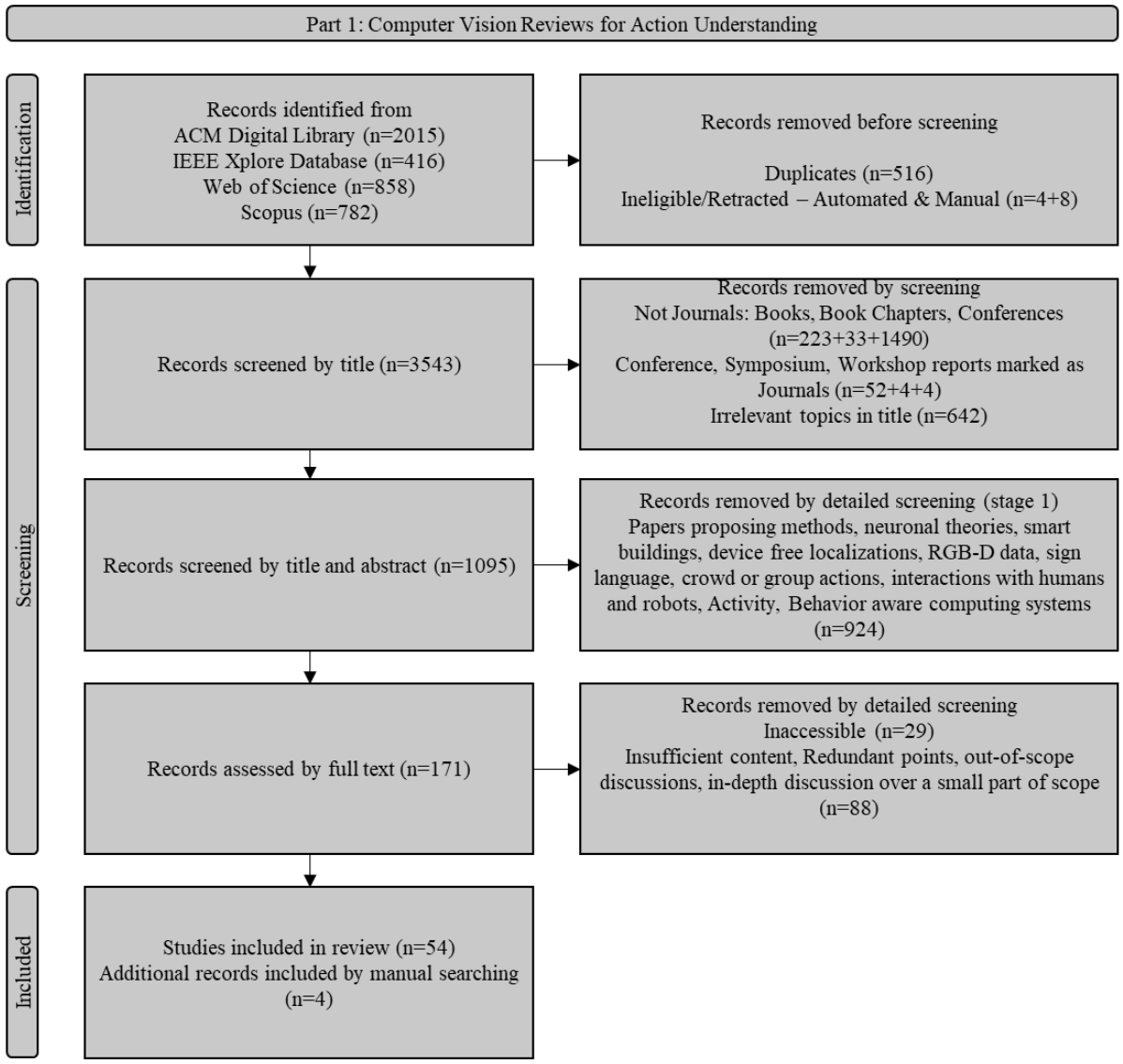

Figure 2: PRISMA Methodology for Review of Computer Vision Literature.

For the second step, focusing on the construction field, again a search was conducted in all the same databases with the search term (ALL = (“activity recognition” OR “action recognition”) AND ALL = (construction)) AND (Up to February 2024). From the results, duplicates and completely irrelevant papers were removed. The remaining papers are separated into groups according to the different themes they cover in their work. All reviews are removed from the documents. Papers that propose algorithms and network architectures have also been removed. Papers that deal with action recognition - in machines, for overall processes, and inside the built environment – are removed. These exclusions ensured that the review focused specifically on studies addressing human action

recognition in construction. Papers that deal with audio, vibration, wearables, and other sensors have also been removed. Finally, we are left with documents on human action recognition within the construction processes using computer vision. The PRISMA flowchart for this part of the work is presented as a figure below.

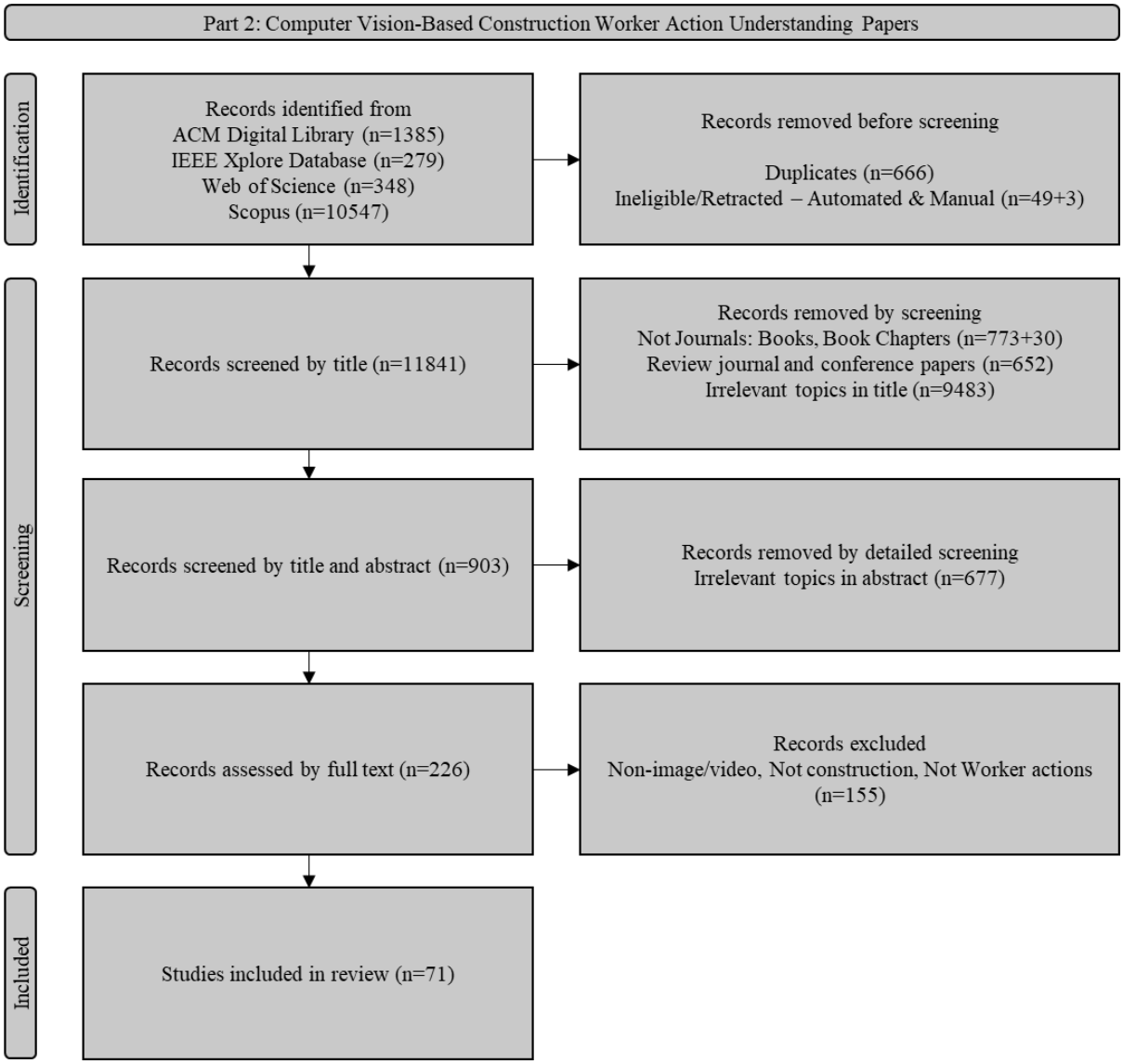

Figure 3: PRISMA Methodology for Review of Construction Automation Literature.

In the third step, the construction literature is discussed within the context of the developed hierarchical structure, drawing connections between the identified themes, highlighting trends and potential use cases, and identifying limitations in the current body of work.

4. RESULTS

4.1 Action Understanding in Computer Vision Research

4.1.1 Taxonomy

Following step 1 of the methodology, many review papers are found on action recognition (Minh Dang *et al.*, 2020) and action segmentation (Gammulle et al., 2023). These papers presented summarizations, frameworks, and taxonomies. Following the aim of the study, documents that focus on taxonomies are studied in detail and presented in Table 1. The table is not an exhaustive listing of taxonomies but a select few that present different elements

necessary for the taxonomy. One major limitation of all the literature is a lack of focus on (any) industry-relevant hierarchies and application use cases.

Table 1: Comparison with Existing Taxonomies and Literature.

Reference	Proposed Taxonomies	Unique Contributions	Gaps for Action Understanding in Construction
(Turaga *et al.*, 2008)	• Low-level – Optical flow, Point trajectories, Background subtracted blobs and shapes, filter responses • Actions: Mid-level (Simple) – Non-parametric, Volumetric, Parametric approaches • Activities: High-level (Complex) – Graphical, Syntactic, Knowledge-based models	• Introduces a hierarchical taxonomy • Identifies the need for different approaches in high-level tasks	• Does not cover low-level features from newer approaches like convolutional neural networks, which are found helpful for the construction domain • Does not discuss contextual reasoning approaches needed for handling the issue of semantic gap in complex activities, observed in construction
(Aggarwal and Ryoo, 2011)	• Single layered – Space-time, Sequential approaches • Hierarchical – Statistical, Syntactic, Description based approaches	• Classifies approaches into two types and identifies appropriate application areas	• No discussion on low-level features and differentiation between action-relevant knowledge and contextual factors for different applications like safety
(Ke *et al.*, 2013)	• Low-level core technology – object segmentation, Feature extraction and representation, activity detection and classification • Mid-level human activity recognition systems – single person, multiple people interaction and crowd behavior, abnormal activity • High-level Applications – Surveillance, Entertainment, Healthcare	• Hierarchical layering leading to applications in the field • Categorization of individual steps further based on existing works • Identifies model-free, indirect, and direct human model approaches	• No discussion on the motion detailing needed for different types of technologies • No discussion on models capturing workplace, tools, and other contextual elements. • No discussion on deep learning approaches
(Guo, Ishwar and Konrad, 2013)	• High-level cues – Human body, Body parts, Objects, Human-Object interaction, Context, or scene • Low-level features – Scale-Invariant Feature Transforms, histogram of Oriented Gradients, Shape Context, Spatial envelop, Others • Action learning – Generative models, Discriminative models, Learning mid-level features, Multiple feature fusion, Spatial saliency, Conditional Random Fields, Pose matching	• Utilizes a bounding box approach for covering the high-level cues, breaking down the body into parts, and also interactions and context • Adds more methods beyond previously identified generative and discriminative approaches	• Limited to image-based action recognition approaches and lacks a discussion on temporal aspects, which is essential to understanding the dynamics of actions
(Rodríguez *et al.*, 2014)	• Learning procedure – Data-driven, Knowledge-driven, Hybrid • Modeling technique – Graphical, Non-graphical, Hybrid • Social interaction • Sensor Infrastructure • Scalability	• Identifies context as a key component for behavior recognition • Extensive discussion on ontologies for human behavior recognition • Proposes the Ambient Intelligence as an application area	• Limits to behavior recognition and does not discuss activity recognition and other applications • No discussion on a single modality-based structured approach, limiting the possibility of developing specific applications using a single modality. • Lack of a hierarchical approach to the taxonomy
(Vrigkas, Nikou and Kakadiaris, 2015)	• Unimodal – Space-time, Stochastic, Rule-based, Shape-based methods • Multimodal – Affective, Behavioral, and Social networking methods	• Decomposes activities into gestures, atomic actions, interactions, group actions, behaviors, events • Incorporates multi-modal information in action understanding	• Mixes all levels of human actions and activities into a single layer, reducing the range of applications for different use cases in construction
(Onofri *et al.*, 2016)	• Statistical approaches – Bayesian belief networks, Probabilistic Petri nets, Hidden Markov Models • Syntactic approaches – Ontologies, Logic rules, Approximate reasoning, Grammars • Description-based approaches	• Separates exploitable knowledge for action recognition into apriori knowledge and context information • Focuses on knowledge-based hierarchical approaches	• Lack of discussion on all low-level features like pose, motion, interest points, and others necessary for computer vision applications in general

(Herath, Harandi and Porikli, 2017)	• Deep architectures – Spatiotemporal, Multistream, Deep generative, Temporal coherency	• Presents local feature representations and feature aggregations for them • Presents the different network structures as an essential parameter	• Does not utilize taxonomical approaches for developing applications • Lacks discussion on any schema or breakdown of actions into components
(Beddiar *et al.*, 2020)	• Activity hierarchy – Elementary human actions, gestures, behaviors, interactions, group actions, events	• Discusses in detail the different features • Presents stages such as detection, tracking, and classification of actions.	• Merges multiple approaches from feature extraction processes, recognition stages, sources of inputs, and learning supervision level – but does not consolidate into a unified taxonomy • No connection between the proposed hierarchy and the presented approaches
(Pareek and Thakkar, 2021)	• Action representation – Interest points, depth, pose, motion, shape, and others • Dimensionality reduction – Principal component analysis, Autoencoders, Reduced basis decomposition, Linear and Kernel Discriminant analysis • Action classification – Traditional machine learning, deep learning	• Presents action recognition system with feature extraction and encoding, dimensionality reduction, action classification steps • Identifies the dimensionality reduction and specific action representations	• Lack of discussion on aspects such as learning type – supervised, semi-supervised, and others – which is relevant to construction, as data availability is limited in many cases • Lack of discussion on hierarchical approaches and knowledge integration for developing applications for specific use cases like safety.
(Kong and Fu, 2022)	• Shallow action representations – Holistic, Local • Shallow action classifiers – Direct, Sequential, Space-time, part-based, manifold learning, Feature fusion • Deep learning – Space-time, multi-stream, hybrid	• Adds action localization, action prediction, and motion trajectory prediction to the tasks • Differentiates action classifiers and relevant action representations • Also discusses different learning methods, useful as a parameter	• Lack of discussion on hierarchical approaches for developing applications for specific use cases • No discussion on feature reduction approaches and knowledge integration for higher-level action understanding-based applications
(Morshed *et al.*, 2023)	• Feature extraction-based methods - Hand-crafted, Deep learning, Attention-based • Activity type-based methods – Atomic action, Behavior, Interaction, Group activities	• Presents a schema as part of the activity-type based methods in the taxonomy • Separates attention-based methods from other deep-learning methods	• Lack of discussion on model-based approaches, feature engineering approaches • Lacks discussion on construction task knowledge and contextual factor integration to develop construction-specific applications for use cases like safety

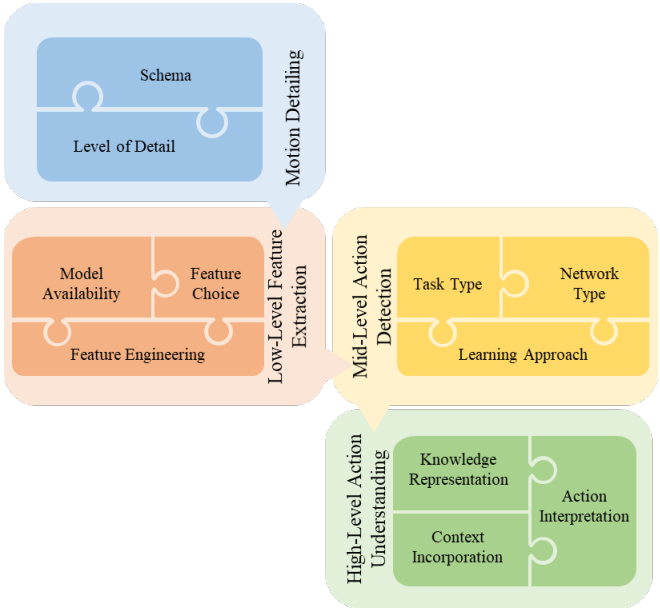

Figure 4: Proposed Taxonomy.

Apart from the limitations mentioned in Table 1, construction sites struggle with the placement of the sensors as the workplace is dynamically changing. For application-specific use cases, understanding the processes in the industry becomes more relevant, and none of the works attempt to focus on hierarchical breakdown. This gap requires us to define two factors – the breakdown of processes to elemental motions and the categorization of sensor capabilities in collecting the data needed for feature extraction. These factors are grouped under the motion detailing preliminaries that must be defined. From the contribution of different papers, multiple elements of taxonomy are observed – learning methods, network approaches, and low-level features, among others. To unify these different elements, we introduce a four-level taxonomy for action understanding, consisting of:

1. **Motion detailing:** Setting up preliminaries for action understanding for different applications
2. **Low-level feature extraction:** Capturing essential visual elements for action understanding
3. **Mid-level action detection:** Recognizing actions within frames or sequences with temporal and spatial localization
4. **High-level action understanding**: Applying contextual knowledge to interpret actions with a focus on applications

4.1.2 Motion detailing

The task of understanding actions from vision is informed mainly by pattern recognition, particularly the motion patterns that vary according to the context. Two preliminary specifics need to be set up to differentiate between different actions – Schema and Level of detail.

Setting up a schema is the first step after determining the application or use case. **Schema** is necessary to differentiate the patterns observed under different abstractions. Typically, schemas are set up to categorize the motions (or the lack of such) and their combinations in a single person (and sometimes extended to multiple people) into a semantically valuable hierarchy. A simple schema can consist of actions, focusing on individual simple motions and activities, covering complex motions, and having multiple people (Turaga *et al.*, 2008). One schema often quoted by construction literature classifies motion patterns as – Gestures, Actions, Interactions, and Group activities (Aggarwal and Ryoo, 2011). Other proposed schema include Gestures, Atomic actions, Human-object or Human-human interactions, Group actions, Behaviors, and Events (Vrigkas, Nikou and Kakadiaris, 2015) or simply Gestures, Actions, Human-Object interaction, Human-Human interaction, Group activity (Sargano, Angelov and Habib, 2017). SPHERE hierarchy (Woznowski, Kaleshi, *et al.*, 2016) focuses on activities of daily living and proposes physiology, pose, motion, action, activity, and behavior as part of the schema. Based on the application, as new semantic approaches are developed, schemas can be modified with new abstractions (Rodríguez *et al.*, 2014).

These schemas work in conjunction with the **Level of Detail** by which humans are represented in the processing steps. The level of detail for human representation is decided based on the data capture parameters like the type of camera and the camera's intrinsic parameters. At the scene level, humans are represented as part of the scene or by bounding boxes or ellipses. At the full-body level, distinct body parts (head, torso, arms, legs) are represented as constituents of the body and represented by lines, cylinders, and boxes. At a finer body-part level, the constituents of individual body parts (like fingers and eyebrows) are also considered and represented by lines or points. With more detailed representations, more classifications can be made to the schema. This detailing can be regarded as evolving from the three-part classification - scene interpretation (of the whole picture), holistic recognition (of the entire body and parts being used), action primitives and grammars (action hierarchy used for scene description), considered together for action recognition (Afsar, Cortez and Santos, 2015) in a more straightforward way.

These two preliminaries set the boundaries for the applications that can be served based on the sensor capabilities. For example, when we utilize far-field cameras, the level of detail is limited to bounding boxes, and schema is limited to events. In such cases, the low-level task can only detect and track when people are in large numbers together. Mid-level action classes will be events that can be identified based on a large number of people working together. High-level action understanding applications can only determine the causes and effects of such events. Thus, the sensor capabilities need to be carefully matched with the application at hand.

4.1.3 Low-level Feature Extraction

Unlike the classic image processing methods, the low-level tasks we consider are segmentation, detection, and tracking. Although only three low-level tasks are mentioned, the target of low-level tasks is to derive different

feature representations for the actions (Turaga *et al.*, 2008; Ziaefard and Bergevin, 2015; Herath, Harandi and Porikli, 2017; Pareek and Thakkar, 2021) –

1. Motion (Optical Flow, Motion History Image, Motion Energy Image),
2. Trajectories (Points, Parts, Objects, Bodies),
3. Interest points (Space-Time, Color Space-Time, Corners and Edges),
4. Pose (2D, 3D, Skeletal) and Poselets (poses of individual body parts or a subset of the whole body),
5. Shape (Silhouette, HOG, Image moments), and
6. Depth,
7. Others (Texture, Gait)

From a neurological perspective, action recognition occurs in the mind in the ventral and dorsal pathways. Ventral pathways capture the form of the body, and Dorsal pathways capture the body's motion (Yousefi and Loo, 2019). Together, these features enable the brain to perceive and differentiate between actions. The first three feature representations mentioned above relate to the motion, and the remaining relate to the form of the body. Motion History Images, Motion Energy Images, and Spatio-Temporal Interest points can be considered hybrid feature representations containing both the motion and form of the body.

Except for the interest points and poselets, all other features are considered global descriptors; the two are deemed local descriptors. As the word suggests, global descriptors provide features of the overall human body and motion. Local descriptors are more robust against background clutter, illumination changes, and occlusions (Abu-Bakar, 2019).

All the above features are traditionally identified in the literature to help in action recognition. With the introduction of deep learning approaches, specifying these features is shunned, and instead, bounding boxes are utilized to let the models determine useful features themselves. Since the precise features utilized by the deep learning models are unknown, few generalized feature representations can be added to the above list based on the scope of bounding boxes.

8. Body part features
9. Local features
10. Frame features

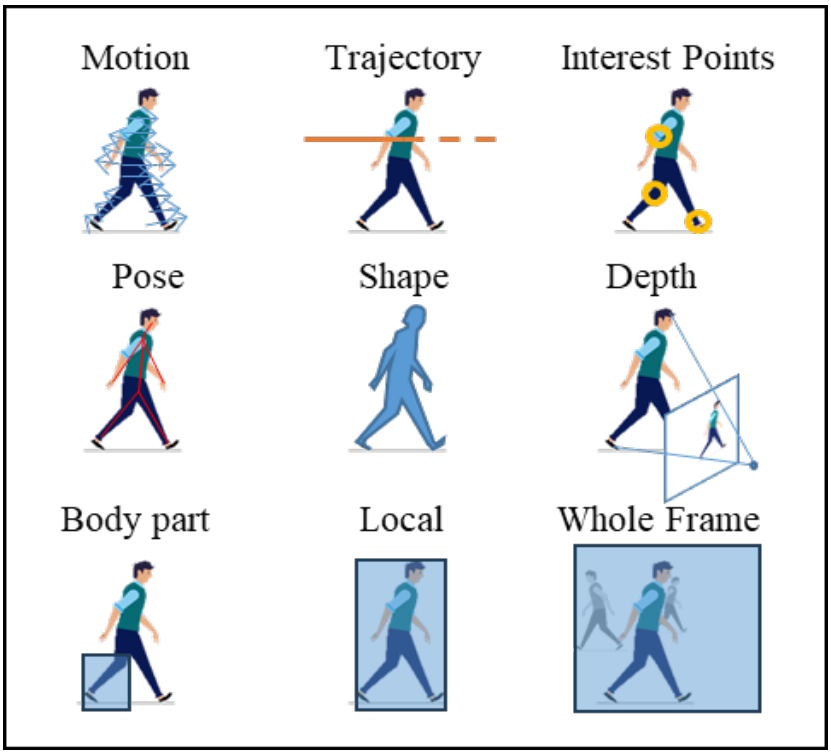

Figure 5: Low-Level Feature Representations.

Model availability is the next consideration. The mirror mechanism theory from neuroscience suggests that action understanding is achieved by transforming the sensory representations of others` behavior into one's motor behavior. Since computers cannot replicate the same approach, modeling and simulation approaches seem apt. A model of the actor, object, and environment can be developed from prior knowledge, with features including appearance, physical capability, physical affordances, and other features that might be useful for understanding the actions. However, modeling new environments and actions is a challenge in itself. We can generally see model-free, indirect, and direct approaches for using apriori models (Ke *et al.*, 2013). The model-free approach uses no prior information regarding the human body. Indirect (or partial) models use information like aspect ratio between limbs, orientations, and relative placements. The direct model uses an explicit geometric representation of the shape, structure, appearance, and movement of the body and its parts (Afsar, Cortez and Santos, 2015). The availability of a model can reduce the dependence on feature extraction. For example, with a full-fledged human body model, simple key-point data is used to simulate a motion and derive all motion features relevant to specific use cases.

Feature engineering is sometimes applied to low-level features to connect them to mid-level tasks. In older works, considering the limited compute capability, the features and models with sizeable dimensional vector sizes are applied with dimensionality reduction methods like Principal Component Analysis, Auto encoders, Reduced Basis Decomposition, and Linear and Kernel Discriminant Analysis. In the current scenario, with increased computing capabilities, feature enrichment methods are adopted for error reduction of data captured through Generative Adversarial Networks and other methods. An alternative approach is the feature conversion from one type to another for subsequent steps.

4.1.4 Mid-level Action Detection

Mid-level **task types** include action recognition, temporal action localization, and spatiotemporal action localization. Action recognition attempts to recognize the action within a frame or a set of frames by identifying the feature patterns. Temporal and spatial localizations provide additional information that is helpful for real-world use cases. Due to the richness of features from low-level tasks, researchers embraced machine learning methods from the early days of these tasks.

Overall, the output of this step is the action recognized and optionally localized spatially and/or temporally. Temporal localization is also considered action segmentation. It is helpful in continuous action recognition, close to real-life scenarios where actions transition from one to another rather than stopping at discrete time points (Gammulle *et al.*, 2023). The temporal action localization will need additional metrics related to action labels and the start-end time proposals (Vahdani and Tian, 2023). Different types of segmentations can be made, like Fixed-size, Overlapping, Hierarchical, and Semantic types of action segments, which break the activity into chunks of actions (Sedmidubsky *et al.*, 2021). Activities can be classified into action sequences – Composite, Concurrent, Sequential, and Interleaved (Kulsoom *et al.*, 2022). In general, for any sensor, the typical processing flow will be segmenting the input into chunks, extracting features, and classifying the segment into some actions (Meng *et al.*, 2020) and combining them back. Spatiotemporal localization adds a spatial dimension. The actions are mapped to the three-dimensional space, which in this case is the construction field to extract relations between different entities and their motions. The additional metrics that can be applied are similar to the localization of objects in a frame at the low-level feature extraction stage. However, the output of spatiotemporal localization is much more useful in extracting relationships between different entities and their actions, enabling a better understanding of the field. At the same time, this approach is more complex than others.

As mentioned earlier, machine learning is the preferred choice of researchers in this field. **Learning approaches** like supervised or transfer learning are typically utilized to train the models with data. However, it does not discount the possibility of using other approaches like unsupervised learning, reinforcement learning, and others. Popular **network types** use convolutional neural networks (CNN) for extracting spatial features, recurrent neural networks (RNN) for extracting temporal features, transformers for extracting features from extended context lengths of any feature type, and graph neural networks (GNN) for encoding and extracting relational features between entities. In some cases, CNNs are also repurposed to collect spatial features from multiple frames, equating to collecting spatiotemporal features. For the tasks mentioned across the taxonomy, the approaches of CNNs, RNNs, GNNs, and Transformers or Attention Networks are found helpful. At the current state-of-the-art, skeletal pose-based action recognition using transformers (Xin *et al.*, 2023) is utilized. However, the network structures can differ based on the application.

4.1.5 High-level Action Understanding

High-level steps bring the reasoning and understanding aspects, move beyond computer vision alone, and are application-oriented. From a neurological perspective, activities can be broken into goals and plans, and each can have different approaches to recognition (Van-Horenbeke and Peer, 2021). Tasks like action prediction, intention recognition, and activity recognition (Kong and Fu, 2022) can be considered part of this level. Some researchers also consider the mid-level and high-level tasks as single and hierarchical approaches (Aggarwal and Ryoo, 2011). The vision-language connection is explored by extending the neurological observations of overlap between visual and language aspects in Broca's area (Willems, Özyürek and Hagoort, 2007) and theoretical lens like the 3R framework (Wiriyathammabhum *et al.*, 2016).

The high-level task of action understanding is application-oriented as we attempt to understand the actions with specific use cases. For this high-level understanding, we require coherence between our interpretations of visual input and our understanding of knowledge about the world. To achieve coherence, past researchers adopted three steps –Represent knowledge of the world and the context of the action (Representation), Capture the context specific to the observations (Incorporation), and match the observations with these representations to derive new knowledge (Interpretation). There is also a proposal to utilize apriori knowledge of actions and contextual information, each valuable for motion categorization into sub-events and event detection from these sub-events, respectively (Onofri *et al.*, 2016). However, the current work presents the categorization of motions as a mid-level task. Providing appropriate context is verified to be more helpful in understanding actions (Wurm and Schubotz, 2017). Though the three steps seem simple, realizing the tasks is nontrivial. The reasons behind such difficulty are – the long temporal interdependencies, complexity, and quantity of possible actions, relevance of associated semantics, and the existence and interaction of several actors in the same environment (Rodríguez *et al.*, 2014). Considerable overlap of the methods within the three steps can be observed in the knowledge engineering domain, which identifies several knowledge-based systems and their required components (Kendal and Creen, 2007).

Knowledge representations can be differentiated based on how objects and relations are represented and the ease of knowledge extraction. Hierarchical modeling approaches break down an activity into its constituents straightforwardly. Key-value models, mark-up scheme models, semantic web technologies, event-based representations, vector representations, and object-oriented approaches are other straightforward encoding methods to represent knowledge. Knowledge graphs capture entities and relationships between entities for efficient querying. Entities can be actions, and relationships can be evaluated from context. Ontology-based systems provide a structured way to represent concepts, relationships, and axioms for reasoning within a domain. Like knowledge graphs, ontologies can encode actions as concepts and context in relationships. The only difference is the presence of axioms, which are pre-defined and fixed for an ontology, reducing the flexibility of querying. Finite automata can be applied to model actions as sequences of states and transitions are used based on the application. Description-based methods involve detailed descriptions of actions, capturing their attributes and contexts. Semantic descriptions of actions can be captured in these methods and used to generate higher-level descriptions (Guo and Lai, 2014).

When studied in different contexts, the same action can provide different understandings. For example, a lifting action is done by a worker to place a material at some height. The quantity of material placed in one lift and the lift speed are necessary for productivity applications. For health applications, the postures adopted, the repetitions made over a given time, and the weight of material lifted are more critical. For safety applications, the location where the lift happened and the conditions of the location before and after the lift are the necessary aspects. Separating the context allows us to reuse the same low and mid-level task outputs for multiple applications more efficiently.

Thus, **context incorporation** is the most informative step for action understanding towards different applications. The word context is defined as "ambiance, attitude, circumstance, dependence, environment, location, occasion, perspective, phase, place, position, posture, situation, status, standing, surroundings, and terms" and "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction …" (Sezer, Dogdu and Ozbayoglu, 2018). Following these definitions, we can classify the contextual factors within the construction domain as Regulatory (for safety and other compliances), Physical environment (the working space and supporting structures, work layout, and work handled), Social (team dynamics, supervision, and management styles), External environment (weather conditions), Project (the time constraints, project specifications, technology used), and Conditions of the worker (health, motivation, emotions,

experience). Though not exhaustive, this list indicates a need for a more nuanced understanding of contextual factors for developing newer applications. Context can be explicitly modeled within the knowledge representation or implicitly modeled while adopting interpretation approaches, or additional inputs can be made for the interpretation step. Due to the intricacies of capturing context, sensors and interpretation methods differ while incorporating different contexts.

Finally, **Interpretation** involves making sense of observations within the context of the represented knowledge. Rule-based systems depend on manually pre-defined rules between conditions, actions, and relationships. Logic-based systems allow formal logic, including first-order and predicate logic, by encoding the logical rules and inference mechanisms. Causal analysis focuses on understanding actions by identifying cause-and-effect relationships in each scenario. Physics constraints involve modeling actions according to the physical laws governing the real world, ensuring the actions are plausible and physically feasible. The syntactic analysis utilizes the grammatical structure of actions to comprehend their meaning (Aggarwal and Park, 2004). By reasoning under uncertainty without considering plausibility, three approaches can be used to interpret the actions. Graphical representation models like Bayesian networks and Markov models utilize the probabilistic relationship between variables. Other representations can utilize Fuzzy logic and Confidence factors for reasoning. Like regular expressions, Pattern searching and matching algorithms identify patterns within the data to match the knowledge representations. Learning approaches like machine learning and deep learning extend these aspects to match the patterns within data of larger dimensions. Exploratory learning schemes like self-supervised learning and reinforcement learning are also covered under this approach, which is similar to mid-level learning approaches.

With these theoretical concepts identified from computer vision literature, we review the works in the construction domain, which cover the aspects of action recognition and understanding. As the field is more oriented towards application, we also consider the application contexts in which they are used.

The Part-1 literature review yielded the taxonomy and the relevant details of each heading. For example, the most popular low-level features are identified and discussed in the relevant section above. Extensive work is done in the computer vision field to understand actions for generalized applications. Comparatively, application-oriented research in the construction field is still in a nascent stage. Hence, there is a need to establish the direction while identifying the current need, particularly the lack of appropriate schema. Thus, the following section reviews past construction literature to fill the gaps and determine the current status and possible future directions.

4.2 Action Understanding in Construction

Following the taxonomy above, the construction literature is classified for each step of the taxonomy. Of the total seventy-one papers found, fifty-seven are journal papers, and fourteen are conference papers. The publication trend is increasing, particularly in recent years. The grid of column charts in Figure 6 shows the trends in individual construction use cases over time.

An increasingly heavy focus is observed in safety-related applications. Consistent efforts have been made to understand, monitor, and measure productivity over the years. Occasionally, applications on quality and health also pop up. Human-robot collaboration (HRC) as an application has been observed in recent years.

The tasks adopted at different levels of the taxonomy and the related use cases are presented in Figure 7. The colors of the nodes are purposefully kept consistent with the taxonomy in Figure 4. Specifically, the nodes represent feature choice at the low level (in orange), task types at the mid-level (in yellow), and action interpretation in the high-level (in green) components of the taxonomy. The other parts of the taxonomy and the results are presented separately in dedicated sections following the current section. The last column of nodes (in grey) represents the use cases in the construction industry, as per the reviewed literature. Application-agnostic works focused on detecting actions and activities without specific use cases, and multi-application studies utilized the detected elements for multiple use cases like safety and productivity.

Few works have utilized more than one task under each level, and this work quotes each separately in Figure 7. For example, one paper might have utilized pose and local features. Then, it is counted in both these works, thus making the numbers more numerous than the total papers.

FIG 6: Sankey Diagram of Number of Works Using Different Low-Level Features, Mid-Level and High-Level Tasks for Construction Use-Cases

Of the seventy-one papers, forty papers utilize only mid-level tasks, twenty-three utilize only high-level tasks, and only eight utilize both mid-level and high-level tasks. In all the works, low-level features are extracted and used.

In works that consider high-level tasks, there are distinct application areas for different high-level tasks. The behavior and condition recognition tasks are exclusively considered in safety applications. Some safety and understanding applications utilize human-human interaction recognition as a high-level task. Applications of human-robot collaboration (HRC) required the intention recognition task. The prediction task is helpful for safety and HRC and is also proposed to be useful for all purposes. For more context, the predictions are the trajectory predictions that the people will follow in their motion on the site.

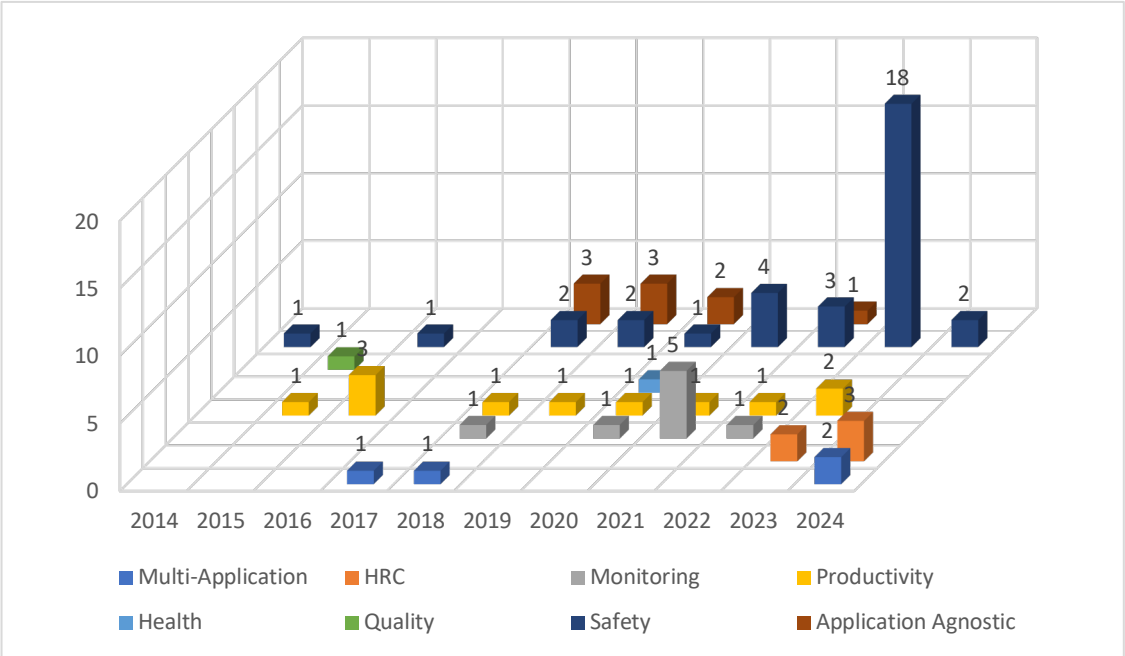

Figure 7: Year-wise Action Understanding Related Works Published on Different Use-Cases in Construction.

Within works that consider mid-level tasks, only a few exclusively identify them separately. Prediction tasks that use trajectories and poses are straightforward motion predictions that do not consider the actions handled. Instead, they are only concerned with the next instant of motion. Activity recognition without mid-level tasks works directly by using local features and poselets. Behavior and condition recognition without mid-level tasks also utilized low-level features like pose, trajectories, body part, local and frame features, and shape features. However, all applications utilize action recognition as the main mid-level task. Action segmentation is utilized only for the productivity aspect in one paper.

For low-level feature extraction tasks, features like pose, local image features, trajectories, motion, frame features, interest points, shape, body part features, and poselets are utilized in the works reviewed.

4.2.1 Motion Detailing for Construction

Motion detailing is the first step in the presented taxonomy, yet very few works in construction literature are found to utilize or define similar steps explicitly. Considering that various technologies differ in their capability to collect motion information of workers in construction sites, these preliminaries are necessary to match the capabilities with applications. The current sub-section defines a schema relevant to construction while also considering the concept of level of detail.

One of the old depictions (Everett and Slocum, 1994) categorizes construction operations at seven levels - Project, Division, Activity, Basic Task, Elemental Motion, Orthopedics, and Cell. Most recent literature uses the term ‘process’ rather than division in the construction. According to recent works (Seo, 2016), using wall masonry as an example process, tasks like setting up, placing mortar, laying blocks, leveling, and rechecking are involved. Each task has operations like lifting, moving, placing, and tapping followed in a sequence. However, if we refer

to ISO 9001:2015, the definition of a process is "A set of interrelated or interacting activities which transforms inputs into outputs". We take the help of the level of detail to set up the schema for ease of reference using the spatial boundaries. The level of detail here considers the scale at which motion is observed as the basis, with an increasingly larger area relating to more complex motion and their resultant schema elements. The schema is presented in Figure 8, and a few examples of the schema elements.

Humans adopt postures relevant to their actions without any motion in the body. At an individual body part level, we can observe elemental motions. Examples of these motions include moving the hand, lifting the head, raising eyebrows, and bending the knee. These motions can give some information by themselves. A combination of these motions can be considered as Actions. Actions can be observed at the whole-body level and classified into expressions, gestures, general body movement, and object manipulation. Though there can be other types of motions and actions in general usage, the elements mentioned here are the most relevant work-related types. Beyond the body level, the actions can be combined as an activity within a locality. Activities include interactions with objects and humans, individual behavior, and events by a group. A series of activities within an environment can be considered a process. The critical aspect is the level of detail, which can differentiate and provide context for observing movement.

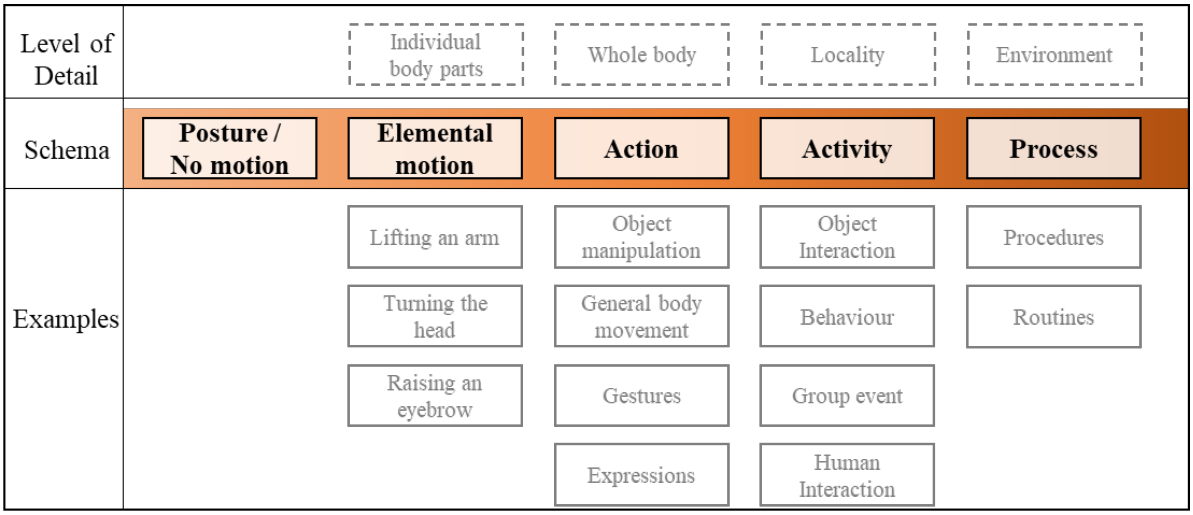

Figure 8: Proposed Schema and Relation with Level-of-Detail, along with Examples.

For example, a process like masonry occurs in the construction site environment, constituting several activities. Activities occur within a workstation or a location, involving object interactions with material and tools, human interactions with other workers to communicate, and following different behaviors underlying these interactions. To elaborate, bricklaying activity within the masonry process can be achieved while interacting cooperatively with other workers or passively neglecting using specific tools for the job. The behaviors arise from the worker's mood, emotions, and other internal factors which can dictate their interactions. Each activity constitutes several actions within the confines of the human body. While laying bricks, the workers must grasp, lift, place, tap the blocks, give hand signs, or express themselves to co-workers for interactions. The motion aspect combines several body parts in conjunction. Each body part has specific movements that need to be achieved to complete the actions. The actions and elemental motions depend on the postures adopted for the actions, but the postures by themselves do not constitute any motion. This is a very simplified example, covering the core hierarchy according to the schema above. In real life, there can be several elements that are not mentioned in the examples above. Considering the existing literature, the schema focuses only on the work-related motions and their hierarchical presentation.

The review found that sixty-eight of the seventy-one can fall within this schema. However, three works propose slightly different schema. The differing schema is - Action, Activity, Step, Subtask, Task (Pan and Yu, 2024a, 2024b) and Interaction, Activity (Fang *et al.*, 2018). The first schema was proposed for intention recognition tasks and human-robot collaboration applications, possibly used since the focus is also on allowing the robot to understand the task context and human action context together. The second schema was proposed to understand applications. Yet, it differs from our current proposal as atomic activities can be considered actions, and

interactions can be regarded as activities within our proposal. This means that the schema is used in reverse in the work. The proposed schema is also close to the SPHERE hierarchy (Woznowski, Burrows, *et al.*, 2016). Still, it varies by considering the behavior as part of the activity and adding processes above the activity level. This is done since the mentioned work is focused on daily living, whereas we are focused on construction activities.

4.2.2 Low-level Feature Extraction

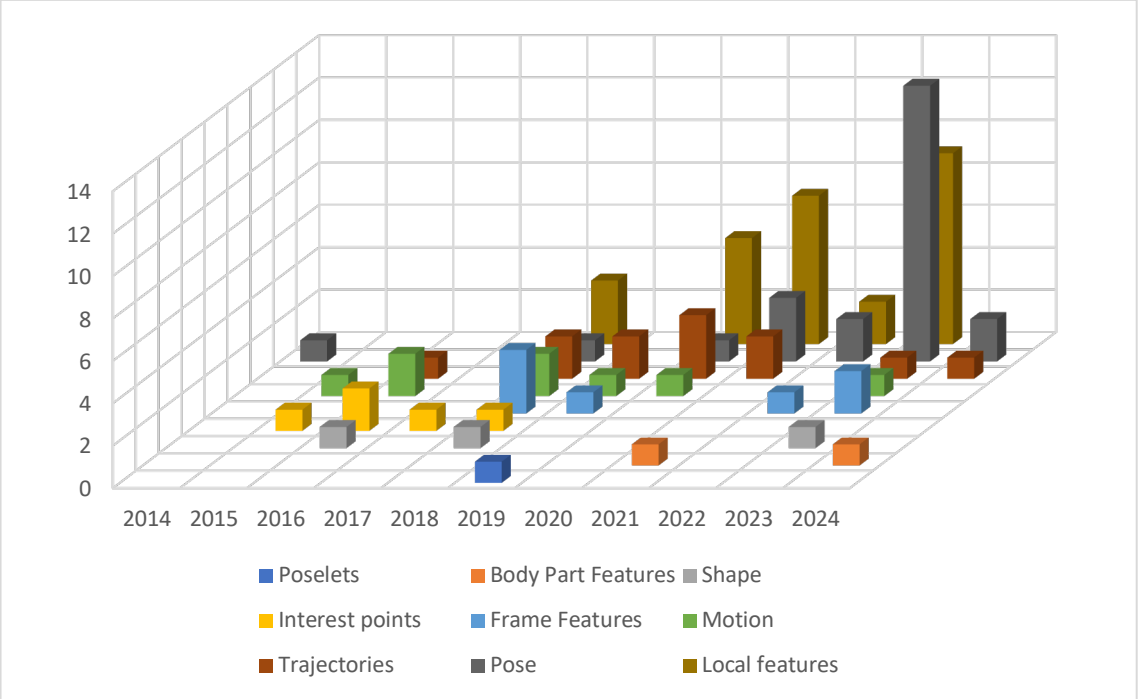

Figure 9: Year-Wise Usage of Different Low-Level Features in Construction Literature.

As presented in Figure 9, there has been a marked increase in the usage of pose (i.e., 2D & 3D body key points) and local features (i.e., bounding boxes around humans) in recent years, owing to their simplicity. Trajectories are utilized sometimes. Interest points are entirely out of favor. A few works have experimented with poselets, body part features, and shape. Few works utilize the whole frame as features and motion features (through optical flow or similar feature extraction).

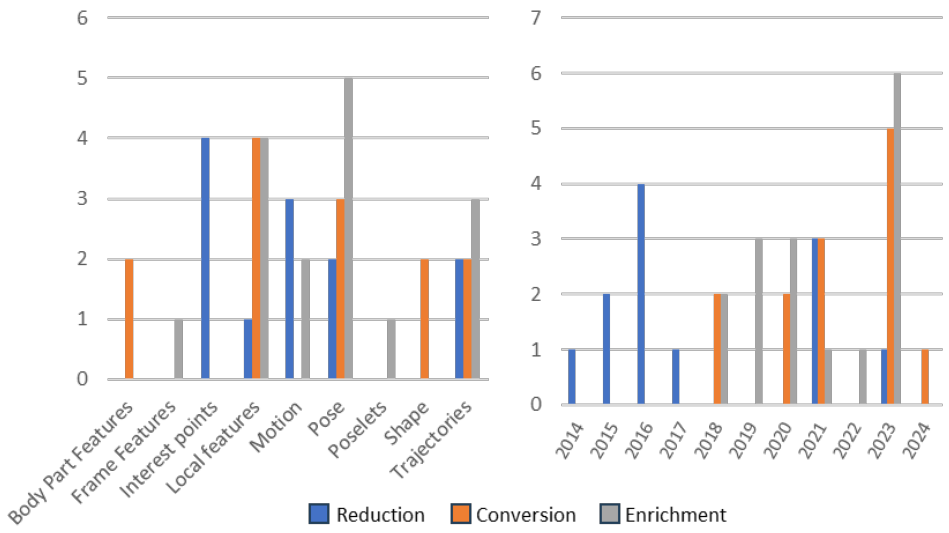

Figure 10: Feature Engineering Adopted for Different Features (Left) and Year-wise Usage of Techniques (Right).

Interest points and motion features are large-size features requiring feature reduction, and these approaches were used in older machine learning tasks. Current methods mostly use pose and local features and convert or enrich the features for further use as can be observed from the adoption trends in Figure 10.

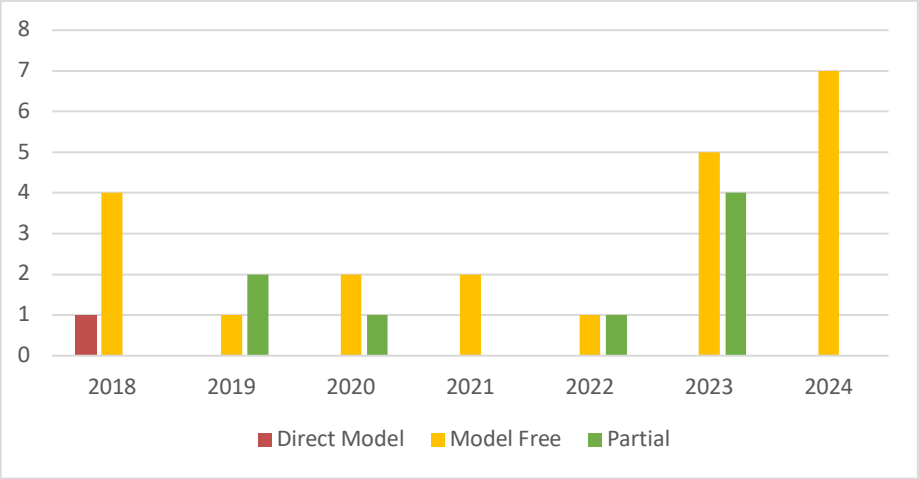

Figure 11: Year-wise Usage of Models in Different Approaches.

Figure 11 presents the year-wise trends of different modeling approaches used. There is also a marked increase in the use of both model-free and partial model approaches, whereas direct model use is completely reduced.

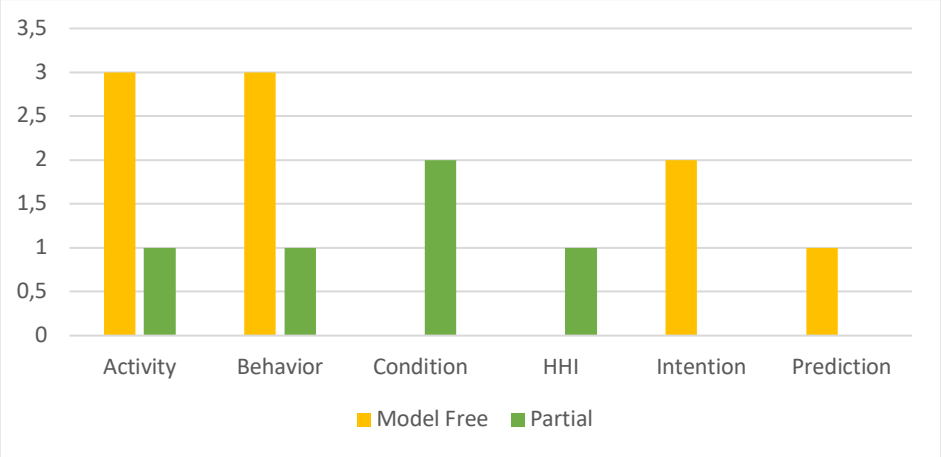

Figure 12: Model Usage for Different High-Level Tasks.

However, only model-free or partial models are used for high-level tasks, as observable in Figure 12. Particularly for condition and interaction detection, some details regarding the human body are necessitated in the literature. Partial models are also helpful in understanding postures and elemental motion in the schema.

4.2.3 Mid-Level Action Detection

In mid-level tasks, as mentioned previously, forty works stop with mid-level tasks, and eight works extend to high-level tasks from mid-level. The older works utilized machine learning approaches like support vector machines and k-nearest neighbors. Newer works mostly prefer deep learning methods. As we deal with images, convolutional networks are the most popular approach. Graph neural networks have been increasingly used in the past few years. Attention networks are another frequently used network type. Only one work in 2016 adopted an action segmentation task using the Bayesian learning approach. Forty-seven works adopted action recognition tasks, one work adopted action segmentation, and twenty-three did not have a mid-level task.

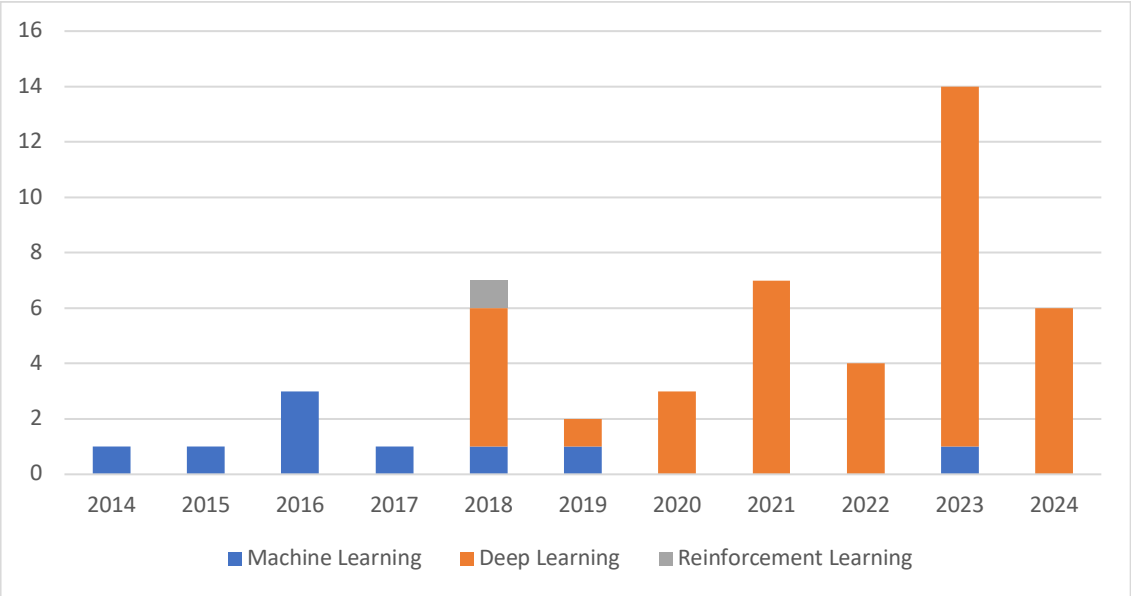

Figure 13: Year-wise Usage of Different Network Types for Mid-Level Tasks.

4.2.4 High-Level Action Understanding

In high-level tasks, the most common tasks are activity, behavior, and condition, which are closely related to safety. There is a reduced interest in activity recognition in general and an increase in behavior and condition recognition. Newer tasks also focus on human-human interaction (HHI), prediction of trajectories, and intention recognition.

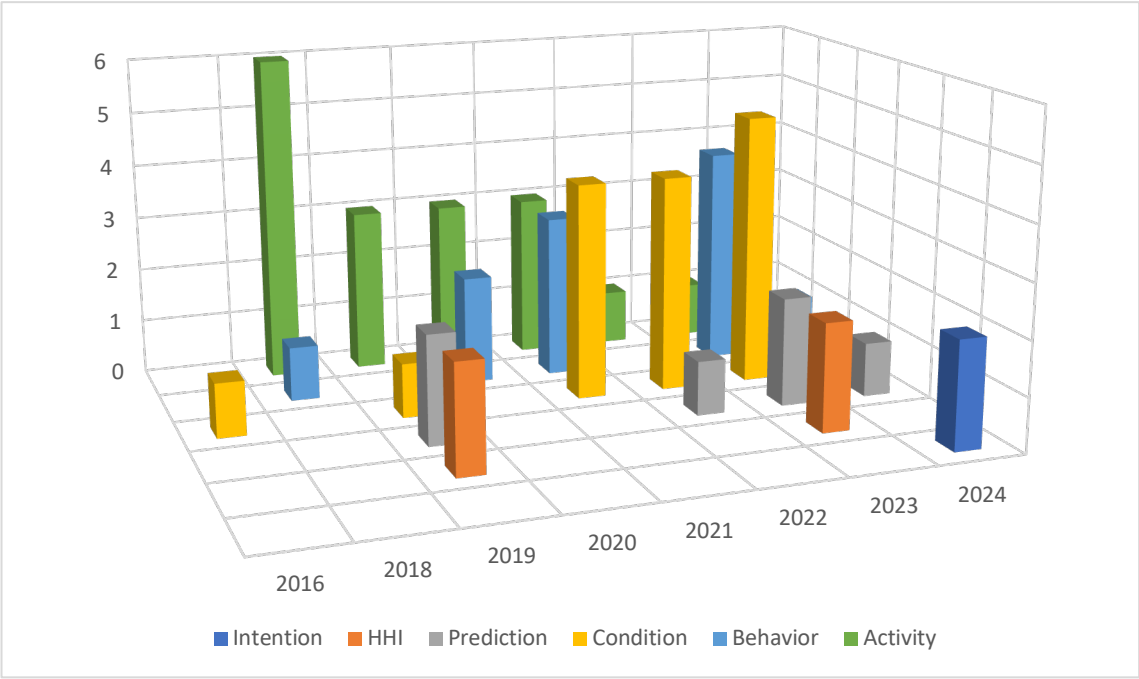

Figure 14: Year-wise Changes in Different High-Level Tasks Adoption.

Besides deep learning, rule-based approaches are also used for different high-level tasks. Deep learning has shown its versatility in application to various tasks. However, the significant applications of activity, behavior, and condition recognition utilized rules similar to deep learning. This also has a relation to contextual factors. The Bayesian approach is used in cases where there is a need for probability exists. Machine learning and feature matching are much less used.

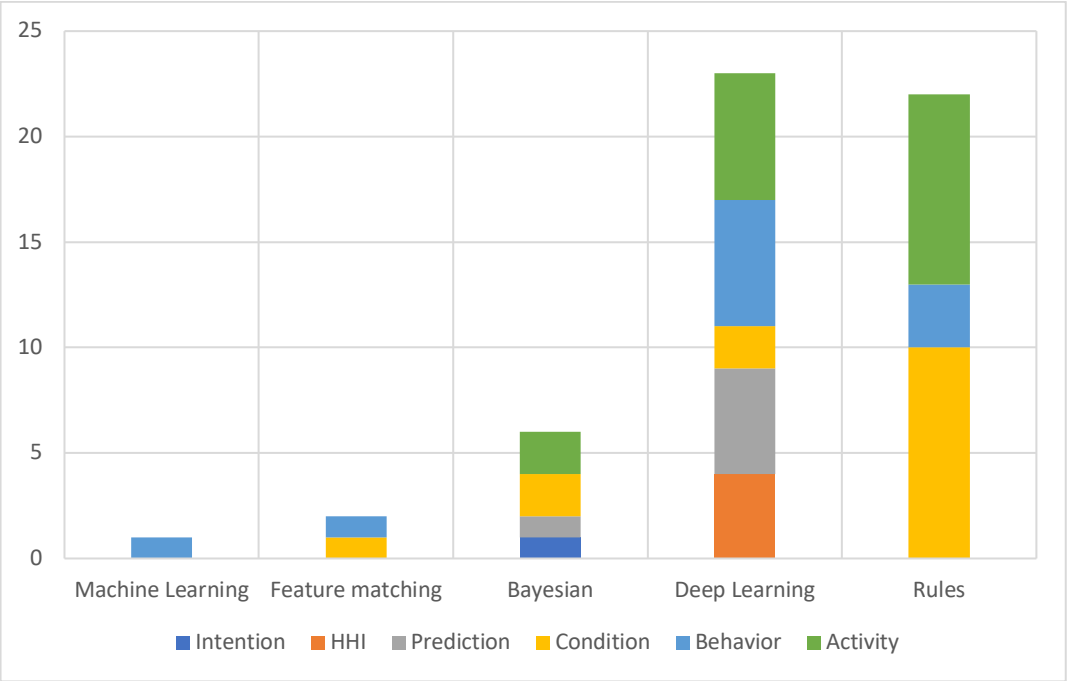

Figure 15: Overall Number of Works Applying Different Approaches Across High-Level Tasks.

Considering that the work focused on computer vision, a large number of works are identified with CNN networks. There has been an increase in GNN and attention network usage in recent years. Traditional methods of machine learning are less in use.

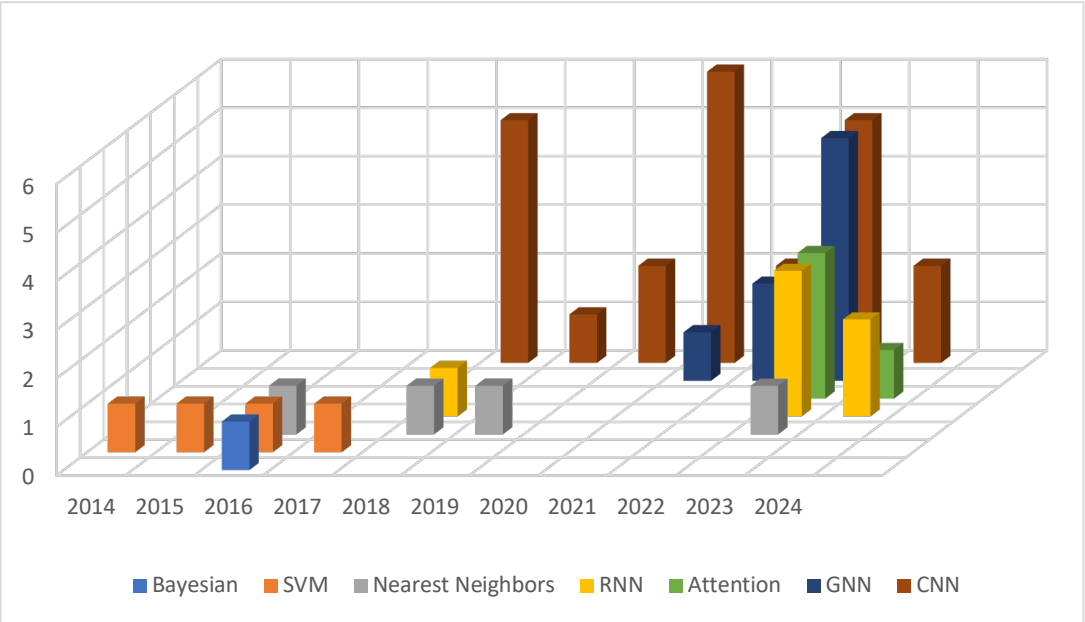

Figure 16: Year-wise Number of Works with Different Networks for High-Level Tasks.

One newer trend is integrating the regulatory and external environment as contextual factors. Regulatory information like safety rules is increasingly considered a context to ground the worker action understanding within the field. The other usual contexts are the project-related task factors, the physical work environment around the worker, and the worker conditions. The external environment is one of the newest additions to the contextual factors.

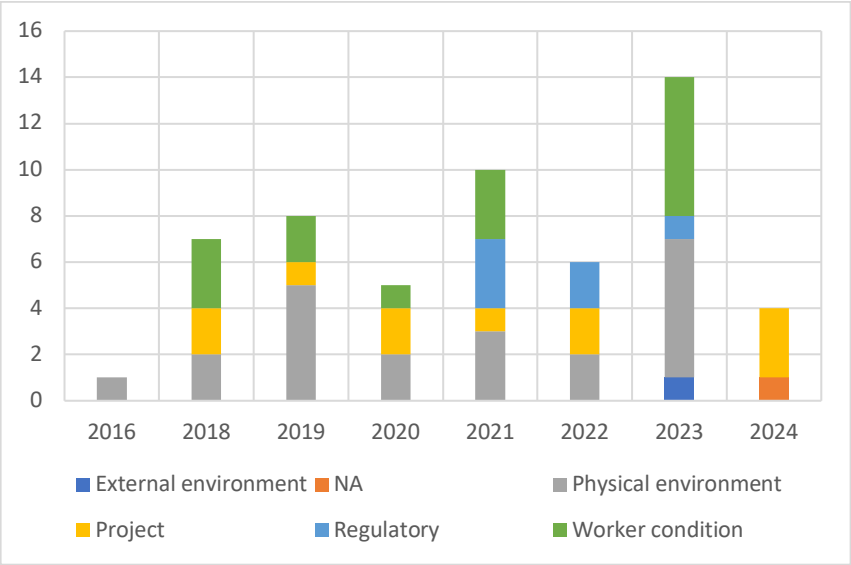

Figure 17: Year-wise Count of Works Using Different Contextual Factors.

Interestingly, the work on site condition recognition has yet to utilize project task factors. The physical work environment is the most utilized context, followed by the worker's condition. However, the work on intention recognition utilized the project task information rather than others, suggesting that it is more advantageous for HRC to have knowledge of the tasks to understand the worker's intention. Additionally, the external environment is also helpful for condition recognition.

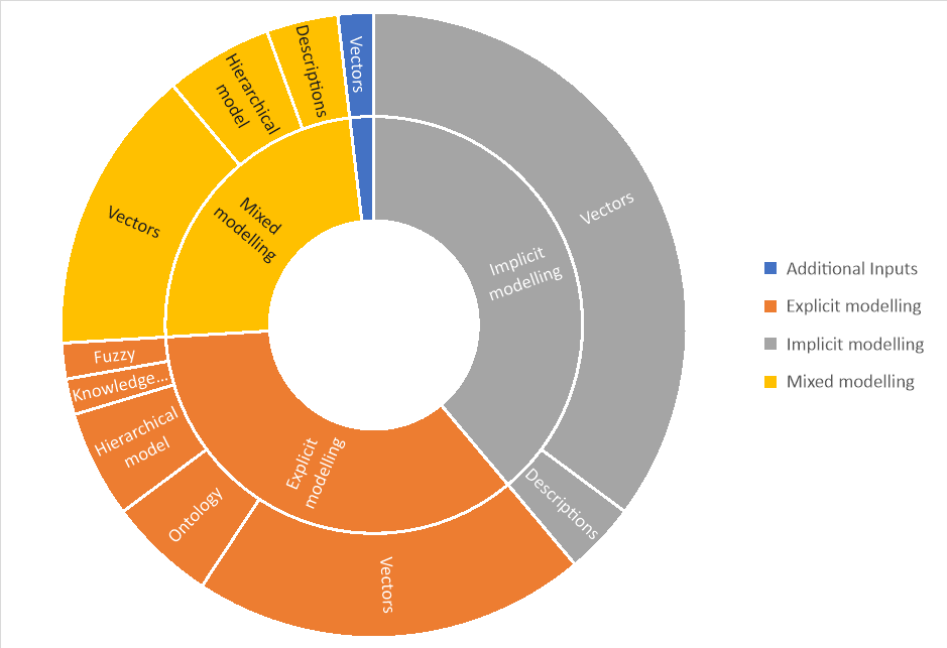

Figure 18: Context Representation Approaches.

Context representation is incorporated in the existing works, as depicted in Figure 18. Implicit modeling approaches include the context as part of the knowledge representation. Explicit modeling presents the context separately at the interpretation step. Additional inputs consider the context to be a separate input from other sensors. Mixed modeling breaks context into parts and includes them both explicitly and implicitly. Across all types of modeling approaches, vector representation is the most in number. This also reflects the large number of learning approaches depicted in Figure 15. The second most used approach from Figure 15 is the rules-based approach. For

this type of approach, text descriptions are utilized for semantic rule matching in both implicit and mixed modeling approaches. Explicit modeling provided more flexibility in adding context through different methods like fuzzy models, knowledge graphs, ontologies, and hierarchical models. The additional inputs are expected from various modalities, and sensors are typically fed vectors to let the learning models decide the importance of the context from them.

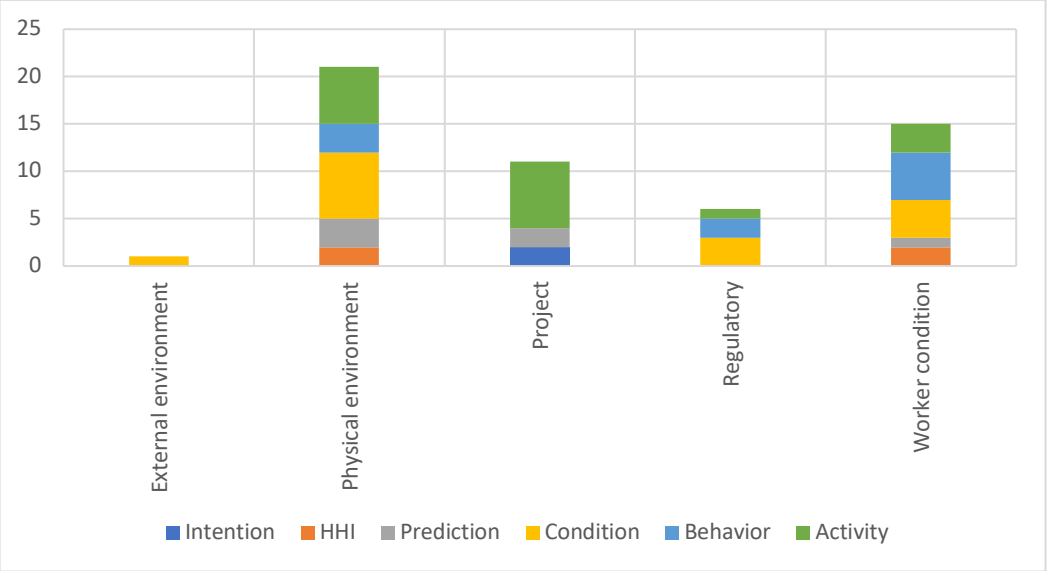

Figure 19: Contextual Factors Used for Different High-Level Tasks.

Figure 19 is a breakdown of different contextual factors, mapping to different high-level tasks, and Figure 20 is a breakdown of the context representation approaches used for the contextual factors. This paragraph discusses the results combining both charts, as the individual charts directly depict that larger values are the most used factors and approaches, respectively. The physical environment is the most utilized context across tasks, and its representation widely varies across the different modeling approaches. Descriptions are provided for the physical, project, and regulatory contexts. Hierarchical models are primarily utilized for project task context modeling. Yet, some attempts are made to capture the physical environment and worker conditions.

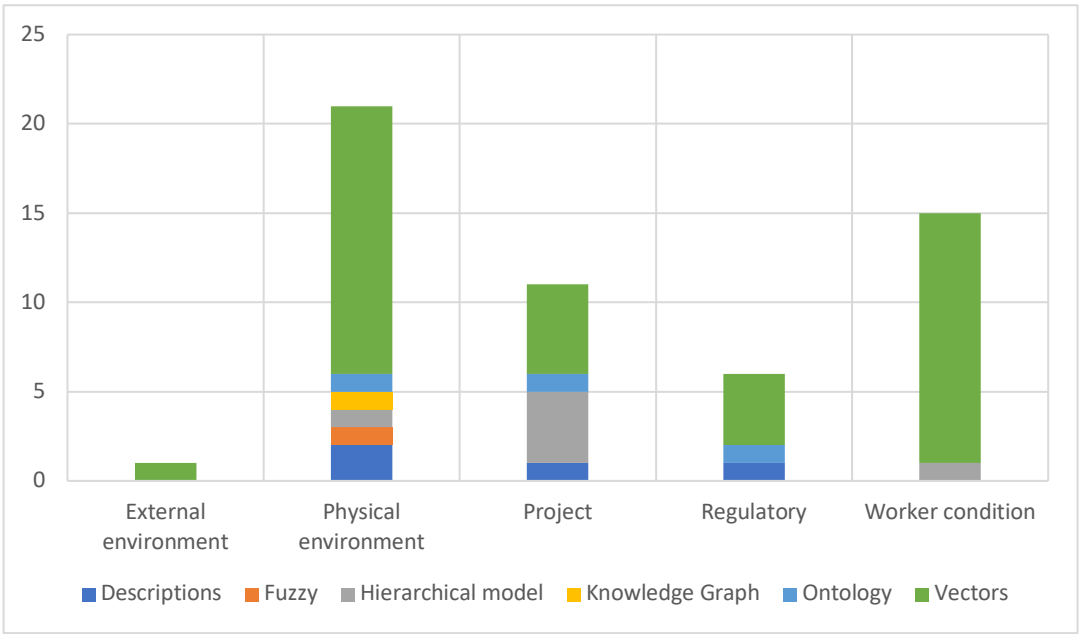

Figure 20: Contextual Factors Incorporation Using Different Approaches.

4.2.5 Quantitative Analysis

Of the seventy-one papers, thirty-seven reported accuracies for mid-level tasks, and twelve reported accuracies for high-level tasks. Other metrics mentioned at the mid-level are precision, recall, F1-score, error reduction, and BLEU scores. Other metrics like CIDEr-D, ROGUE-L, and SPICE are also at the high-level tasks. The BLEU, CIDEr-D, ROGUE-L, and SPICE are metrics used to measure the text-based approaches quantitatively.

Apart from accuracy, few papers touched on the other metrics to quantitatively suggest any observations. Natural language processing-based metrics like BLEU are mentioned in only three papers (Liu *et al.*, 2020; Zhai, Wang and Zhang, 2023; Zhong *et al.*, 2023). Hence, they are not further presented and discussed.

Table 2: Metrics of Reviewed Works.

S No	Metric	Minimum	Maximum	Average	Papers reporting the metric
	Mid-Level tasks				
1	Accuracy	57.00%	99.50%	86.09%	37
2	Precision	62.40%	95.00%	79.89%	9
3	Recall	66.70%	92.00%	80.05%	5
4	F1-Score	78.87%	78.87%	78.87%	1
	High-Level tasks				
1	Accuracy	71.70%	99.60%	88.09%	12
2	Precision	69.85%	98.26%	81.51%	5
3	Recall	70.50%	97.80%	83.18%	4
4	F1-Score	79.00%	96.78%	89.59%	3

Due to the paucity of other metrics used, this section focused on comparing the accuracy between different works for the quantitative evaluation.

Of the seventy-one papers, except for one paper that proposes to use scene graphs and ten documents that did not mention the details of the datasets, others mentioned using datasets in the form of frames or clips. Twenty-four papers used Frames for the dataset, which ranges from 110 frames to 38176 frames. Thirty-five papers use Clips for the dataset, ranging between 4 clips to 63900 clips, ranging between 0 to 20 seconds each, or using 100 to 200 frames per clip in most cases.

In the works studied in the review, only eleven papers were presented using the frame dataset and the accuracy of the mid-level tasks. We observed that two to ten classes were used at most. A 99.5% accuracy is observed for four classed datasets with 9695 datapoints, which can be mentioned as the best benchmark, with nearly 2400 frames per action class (Han, Lee and Peña-Mora, 2014). The following best is with only two classes and 2569 datapoints, reaching 1200 frames per action class and 98% accuracy with the latest graph neural network (Liu and Jiao, 2022). Only 26 papers report the accuracies with clips as datasets. Within this, most (five) papers used seven classes of datasets. And they reached an average accuracy of 89.5%. The best-performing works used four to seven classes of actions and combinations of clip lengths between 2 seconds to 10 seconds, continuous videos, and predominantly 5-second clips. The dataset ranges from just 72 clips of 4s each for four classes to around 8000 for 5s clips for seven classes. Although a few applications use data without clipping for evaluations, these numbers give a sense of the dataset requirements.

Most papers start with pre-trained models for pose and object detection tasks. However, a few also started with action recognition datasets like Kinetics-400 and HMDB; Image recognition datasets like ImageNet, Flickr8K, and MS COCO; and Task-specific datasets like Le2i falling humans dataset, WiderFace, and CelebFaces+ datasets. Three valuable datasets within the construction community are available (Yang, Shi and Wu, 2016; Roberts *et al.*, 2020; Tian *et al.*, 2022), which can be utilized for developing new action understanding networks and training approaches.

5. DISCUSSION

Table 3: Overall Taxonomy With Details Under Each Level.

Motion Detailing		Low-Level Feature Extraction			Mid-Level Action Detection		
Schema	**Level of Detail**	**Model Availability**	**Feature Choice**	**Feature Engineering**	**Task**	**Network Type**	**Learning Approach**
• Posture • Elementary Motion • Action • Activity • Process	• Individual Body Parts • Whole Body • Locality • Environment	• **Model-Free** • Indirect Model • Direct Model	• Motion • Trajectories • Interest points • **Pose** • Shape • Depth • Others (Texture, Gait) • Body part features • **Local features** • Frame features	• Feature Reduction • **Feature Enrichment** • **Feature Conversion**	• **Action Recognition** • Temporal Action Localization • Spatio-Temporal Action Localization	• **Convolutional Neural Network** • Recurrent Neural Network • Graph Neural Network • Attention Network	• **Supervised Learning** • Unsupervised Learning • Transfer Learning • Machine Learning • **Deep Learning** • Reinforcement Learning

High-Level Action Understanding					
Knowledge Representation	**Context Incorporation**		**Action Interpretation**		
	Context Type	**Context Incorporation Method**	**Interpretation Network Types**	**Interpretation Method**	**Interpretation Task**
• Key-Value • Markup scheme • Semantic web • Even-based representation • **Vector Representation** • Object-oriented approach • Ontology • Knowledge graph • Finite Automata • Semantic Descriptions	• Workspace Environment (e.g., layout) • External Environment (e.g., Weather) • Regulatory • Social • **Project** • **Condition of worker**	• Explicit modeling • **Implicit modeling** • Mixed modeling • Additional inputs	• **Convolutional Neural Network** • **Graph Based Neural Network** • Causal analysis • Physics constraints • Syntactic analysis	• **Rule-based** • Logic-based • Causal analysis • Physics constraints • Syntactic analysis • Reasoning under uncertainty • Probabilistic relationships • Fuzzy logic • Pattern Searching • Machine learning • **Deep learning** • Exploratory learning	• **Activity Recognition** • **Behavior Recognition** • **Condition Recognition** • Human-Human Interaction Recognition • Human-Object Interaction Recognition • Intention Recognition • Trajectory Prediction • Action Prediction

Worker action understanding represents a foundational capability necessary for fully realizing the potential of Construction 5.0 technologies. This research reviewed the current state of the art and presented a structured taxonomy, addressing the complex action hierarchies driven by organizational goals. This taxonomy presents a

unified approach for various use cases and technologies, realizing the goal of Construction 5.0. The hierarchical taxonomy (Motion detailing, Low-level Feature Extraction, Mid-level Action Recognition, and High-Level Action understanding) addresses the need for a unified taxonomy for different use cases. The hierarchical schema (Posture, Elemental Motion, Action, Activity, Process) presented as part of the motion detailing step provides a structured means of categorizing the worker motions and matching them with technological advances. A summary of all the elements within each part of the taxonomy is presented in Table 3. Under each part of the taxonomy, the most preferred elements in construction literature are highlighted in bold in the table.

5.1 Current State-of-the-Art and Scope for Improvement

Overall, there is growing interest in applying action understanding to safety and productivity, while other use cases, such as quality and human-robot collaboration, are in the exploratory phase. Notably, safety applications are preferred in high-level tasks based on the ABCs of workers – Activity, Behavior, and Condition Recognition. Contextual factors such as work environment and worker conditions have also emerged as critical inputs across different applications. A few worker-specific elements, such as health conditions, are targets for applications and contextual factors for other applications. Intention recognition and project information as context have been explored in human-robot collaboration but have been helpful for different applications like safety and productivity. High-level tasks reveal increasing adoption of vector representations for implicit context modeling. Nevertheless, the limited use of representations such as semantic descriptions suggests the unmet need to bridge the semantic gap.

Table 4: Mapping of Potential Application Use-Cases to Construction 5.0 Values.

Applications	Human-Centricity	Sustainability	Resilience
Documentation / Monitoring	**Self-review logs**: Action records and analysis for worker feedback.	**Digital archives**: Visual logs to reduce paper use and store detailed information.	**Forensic playback**: Identify and replay events to reconstruct incidents and events.
Health	**Ergonomic alerts**: Unsafe postures or movement detection for feedback and alerts.	**Wellness tracking**: Monitoring activity patterns to prevent long-term strain on workers.	**Overexertion alert**: Early warnings from detected signs of fatigue or excessive effort to avert injuries.
Human-Robot Collaboration	**Intent capture**: Worker intention recognition to improve robot interactions that are helpful to workers.	**Task efficiency**: Identify, plan, and share actions between worker and robot to enhance energy efficiency and reduce redundant work.	**Adaptive fallback**: Detection of anomalous interactions, either from robot or worker, to trigger backup safety measures.
Productivity	**Performance profiling**: Analysis of individual task actions to assess workers' efficiency while maintaining privacy.	**Inefficiency flags**: Detection of redundant or non-value-added actions that waste time or materials across trades and processes.	**Sequence disruption**: Early alerts when task flows break down, preventing workflow interruptions.
Quality	**Procedure Compliance:** Verify that correct methods and steps are followed to maintain work standards and to give feedback.	**Resource efficiency**: Monitoring actions to ensure optimal use of materials and energy, reducing wastage.	**Fault precursor**: Identification of early signs of errors in task execution that may lead to quality issues.
Safety	**Unsafe actions**: Detection of unsafe and abnormal behaviors that compromise personal safety and the safety of others.	**Hazard prevention**: Recognition of hazardous behaviors in waste material handling to avoid environmental pollution.	**Crisis detection**: Real-time detection of emergency conditions for fast response.
Skill	**Skill profiling**: Analysis of action patterns to assess individual competencies and tailor personal training.	**Competence mapping**: Evaluation of processes to align worker skills for better use of resources.	**Training gaps**: Identification of areas for skill development.

Action recognition is the commonly used mid-level task without temporal and spatial localization, crucial for several meaningful site applications. Few works achieve localization by manually placing bounding boxes before processing for action recognition. The taxonomy's inclusion of localization tasks within the mid-level presents the opportunity for automating and developing applications utilizing the time and location of workers for purposes like productivity or safety. There is an increasing trend of using pose and local features, with older methods like interest points declining in usage. This shift is beneficial for faster processing and covering more ground on

construction sites. Simultaneously, enriching and converting the features are increasingly preferred, suggesting richer feature information is needed.

Quantitative analysis shows accuracy as the most common metric for mid-level and high-level tasks. Frame-based and Clip-based datasets worked equally well in the mid-level action recognition task. However, results show the underuse of quantitative evaluations beyond accuracy, particularly for high-level tasks, a lack of comprehensive assessment across the pipeline, and ablation studies made by removing some components and evaluating the output metrics. Secondly, the datasets in the reviewed works are typically built from various sources, randomly split in 70:30 ratios for training and testing. Considering the large parameter size of deep learning methods that are currently popular, unseen data not used in training, for example, taken from a different project site, needs to be used as a validation dataset to evaluate the generalizability of the models. Finally, following a standard schema will enable the evaluation of different approaches for newer applications.

Table 4 attempts to map application use cases to the construction 5.0 values. Applications that are found from the results are also integrated within the use cases. These different use cases present a wide variety of tasks that can be adopted within the action understanding tasks to improve automation efforts in the industry.

In summary, achieving the need for a unified taxonomy, the current work presented a hierarchical taxonomy, categorizing four steps useful for developing applications relevant to different use cases. While the presented taxonomy provides a critical structural contribution, it also highlights significant gaps and future directions that must be addressed to fully realize the potential of action understanding to Construction 5.0. The advancements are discussed in the following subsection in terms of future directions.

5.2 Future Directions

To simplify the discussion, the future directions are categorized logically based on the taxonomy for algorithmic advancements and based on the schema for technological advancements. The algorithmic advancements focus strictly on computer vision-related methods but also present the vision-language connection, which is promising in the current large language model-based research. The technological advancements categorize the technologies around the worker into three categories – on-body devices, near-body agents, and ambient systems – connecting them to the schema presented.

One commonly discussed aspect in both categories is worker privacy, a critical concern for managing worker trust and organizational expectations. Identifying individuals allows for targeted training and evaluation but raises the risks of over-surveillance, bias, and misuse. Conversely, complete anonymization can limit personalized interventions and security-related measures. A balanced approach is crucial for the ethical adoption of action-understanding-based applications aligning with the value of human-centricity. As an additional benefit, targeted data collection reduces costs, making adopting technology sustainable for a cost-sensitive industry like construction.

5.2.1 Algorithmics Advancements

In low-level features, the increasing use of feature enrichment and conversion suggests taking a look at low-level features such as Interest points (Li *et al.*, 2017) and Poselets (Tian *et al.*, 2023). With the increase in model availability for representing the human body like SMPL (Chu *et al.*, 2020), extending features from the two-dimensional perspective of the image frame is becoming quite valuable for applications like ergonomic analysis. 3D positioning using methods like objective knowledge (Shen *et al.*, 2021), scene analysis (Shen *et al.*, 2023), camera calibration using epipolar geometry (Assadzadeh *et al.*, 2021), and homography transformations (Fang, Li, *et al.*, 2020) are found useful in this direction. The problems of varying perspectives, occlusions, and low-level errors can be offset by regressing human mesh models like SMPL over the image frames instead of mapping image features to 3D. Biomechanical models, like 3DSSPP and OpenSim (Yu *et al.*, 2017; Li *et al.*, 2019), used for musculoskeletal disorders analysis, can be extended to action understanding, particularly in applications related to health and safety. In case of a lack of data, model-based action datasets can be created by generating synthetic data (Neuhausen, Herbers and König, 2020; Kim *et al.*, 2022, 2023) or by repurposing larger datasets (Tian *et al.*, 2022). Simple methods like utilizing the biological regularities within human motion can also present interesting low-level features or partial model inputs (Noceti *et al.*, 2017). Instead of modeling individuals, automaton approaches like Finite state machines (Martinez *et al.*, 2021) can also be expanded and integrated with building models. Models of humans and the environment can be used to build digital twins or move information to a metaverse and

extract relevant features from simulations. Finally, the schema can be tailored to specific applications, such as the task-oriented schema for Human-Robot Collaboration applications (Pan and Yu, 2024a).

In the mid-level action detection methods, considering that the industry faces a lack of appropriate data for each application, adopting techniques like Weakly supervised learning, Unsupervised learning, and Self-supervised learning techniques need to be verified in more detail. Within the typically used transfer learning, different approaches can be adopted to solve specific issues (Ray and Kolekar, 2024). For instance, unsupervised transfer learning can be adopted to identify anomalous behavior, and transductive transfer learning can be used to adapt a previously trained model to site-specific processes with fewer data. Replacing some learning approaches with interpretation approaches like logic, causal analysis, physical constraints, and probabilistic alternatives like fuzzy logic can be explored to reduce data dependency.

In high-level action interpretation methods, particularly the context types, human attention as context based on head and body orientations has shown considerable utility (Cai, Zhang and Cai, 2019), beyond the regularly used bounding box-based positional context (localization). Other useful contextual features currently identified are workplace factors like proximity and congestion and worker conditions like working height and leading postures (Xu and Wang, 2023); external environmental factors like humidity and temperature (Moohialdin *et al.*, 2023); worker conditions like expertise (Ryu *et al.*, 2022). Other contexts like project context (the work and workplace details) need to be exploited further, as these are the workers' most relevant task guiding factors. Instead of manually capturing and inputting the contextual information, it can be accessed through Building Information Models and site data (Xu *et al.*, 2021) and digital twins (Pal *et al.*, 2023), integrating into existing virtual construction processes. Construction tasks inherently present a hierarchical breakdown structure, and explicit knowledge modeling approaches like knowledge graphs and ontologies are preferable as they can capture these structures while maintaining human and machine readability. Applications related to quality and health need structured approaches to utilise high-level task outputs.

The vision-language connection provides avenues for many hitherto unexplored applications in construction by connecting the language-related methods in the high-level tasks, building upon directly from the low-level features or mid-level action information. Older methods in vision-language connection focused on generating words from visual patterns in top-down and bottom-up approaches (Wang, Zhao and Yuan, 2014). Newer methods have bi-directional capabilities like generating descriptions from images and generating and reasoning over images, among others (Mogadala, Kalimuthu and Klakow, 2021). Although using pose data for descriptions (Chen, Dong and Demachi, 2023) can provide useful information, generating descriptions for action data caters to applications like worker skill analysis. Documenting the visual process is essential for business purposes, as well as for safety records, process descriptions (Ren and Zhang, 2021), and method statements. When documentation is available before fieldwork, action understanding can help cross-verify the field, reducing the workload of field supervisors. As construction activities in the real world vary beyond the instructions documented, unsupervised techniques like topic modeling will help identify commonalities across different worker actions (Pal *et al.*, 2021). Apart from the video captioning techniques identified in the reviewed literature, Scene description (Pereira *et al.*, 2023) and Video Description (Aafaq *et al.*, 2019) techniques can provide detailed accounts of the activities in a scene or a video, providing plausible explanations for incidents and accidents. Specific to the construction activities, considered to have a hierarchical structure, syntactic approaches provide the modularity needed for analysis at different steps (Astolfi *et al.*, 2021). Furthermore, syntactic descriptions, semantic graphs, and knowledge graphs represent human actions in more detail (Wu *et al.*, 2022). In some interesting reverse approaches, given an image and some description, models are developed in construction research to identify the bounding boxes (Liu *et al.*, 2022), and given an image, the features are enriched with semantic details like object, status, action, and activity (Zeng and Hartmann, 2023). In summarizing the methods and their relation to Construction 5.0, exploiting the vision-language connection helps maintain human-centricity and meet the needs of supervisors and organizations. Reducing the need for manual supervision at multiple locations simultaneously allows supervisors to focus on critical work tasks, improving the resilience of process control. Reducing the manual analysis reduces site inefficiencies like rework and resource misutilization, improving organizational sustainability.

The recent progress in generative large vision-language models (alternatively called foundation models) enables simultaneous documentation and visual feature processing while following the hierarchical structures. These models have shown emergent capabilities in reasoning, making them apt for high-level interpretation tasks. Research is needed on how well these models can be utilized in the field since these generative models often face

the issue of hallucinating in their responses. Thus, it is still necessary for construction researchers to identify appropriate context, establish the knowledge representation, and integrate the knowledge into these large models to improve their correctness in interpretations.

Differential privacy is another promising research avenue, enabling tailored privacy settings based on application needs – for instance, revealing identities for security applications while protecting them for productivity applications, which are more prone to misuse. It also allows workers to self-declare privacy preferences, ensuring consistency across site applications that may require identity information. Clearly defining and declaring the privacy aspects in newly developed applications is crucial for industry and academia to make informed decisions on their adoption. For example, a skill assessment application needs to declare the use of identity information right from the research stage, facilitating smoother field implementation. Algorithmic methods like face anonymization, cartooning, and encryption (Jung, 2020) can enable differential privacy.

5.2.2 Integration with Technologies

The current technologies that enable workers can be broadly categorized into - On-body or Body-worn devices, Near-person agents, and Ambient systems. Due to the proximity to humans and the level of detail available, these technologies can provide information under specific parts of the schema, as shown in Figure 21, and the applications can be tailored to these schema parts for the technologies. Other technologies and sensors will be necessary for applications that require information beyond the parts of the schema from which technology can collect information. This section discusses the research directions for improving action understanding in these three categories with respect to the schema.

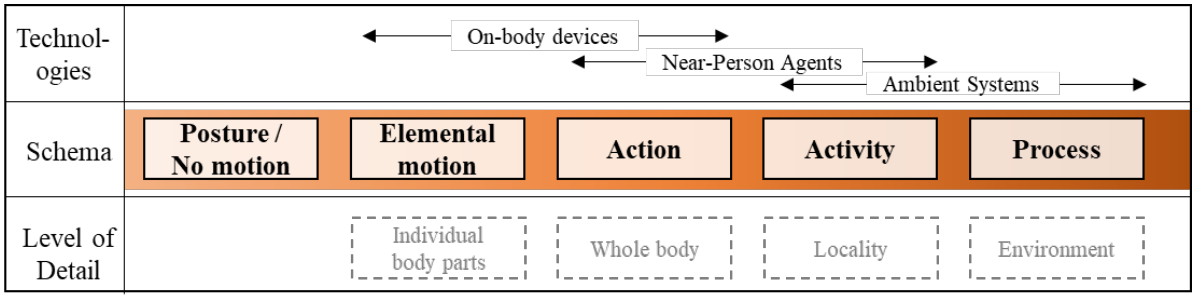

Figure 21: Existing Technologies Mapped to Presented Schema.

On-body devices can capture elemental motions due to their proximity and can extend to action recognition. Sensors and smart wearables (Calvetti *et al.*, 2020) and exoskeletons (Zhu, Dutta and Dai, 2021; Perera *et al.*, 2023) fall under the on-body devices. Sensors and smart wearables alone do not contribute to the worker's actions but can passively collect information and be utilized with other technologies. Exoskeletons are wearable mechatronic devices that assist and augment human capabilities. For such systems, intention recognition must be close to real-time for better performance and feedback. Different sensor systems enable recognizing the motion intentions that travel from the brain to the relevant body parts at different scales of the motion being created (Li *et al.*, 2023). A detailed evaluation is essential for understanding motion-level intentions using vision. Egocentric or first-person view camera-based datasets like EPIC-KITCHENS (Damen *et al.*, 2018) are useful in understanding actions by focusing on the movement of hands. Understanding actions at such proximity can help capture the worker's intentions for low-level motions. Simple gesture recognition and localization can enable safety applications (Rabbi and Jeelani, 2024). Real-time feedback for improving productivity, safety, health, and quality is also possible by using predictive and proactive applications while accommodating individual work styles. Evaluating individuals' performance, work quality, skill level, and training needs is possible with action evaluation techniques (Lei *et al.*, 2019).

Near-person agents can be robots that do the task themselves or machines that humans operate to achieve the tasks (You, Zhou and Ding, 2023). These systems interact with humans to augment their capabilities and replace humans in riskier tasks. Human-robot interactions are classified into coexistence, cooperation, and collaboration (Hentout et al., 2019) and represent the increasing complexity of such interactions. Action understanding is essential for robots for complex interactions like working together safely and efficiently. As both these agents can only be at some distance from the worker for safety reasons, they are limited in capturing the elemental motions but gain the

capability to observe the locality surrounding the worker. Though it enables capturing actions and activities, these systems have limited capability in understanding the whole processes that can span multiple locations within the field. Progress monitoring (Martinez *et al.*, 2021) and abnormal operation detection (Lin, Chen and Hsieh, 2021) can be implemented for workers, as done on machines in the cited works. Worker well-being can also be evaluated through emotion, fatigue, and stress recognition based on worker condition as contextual input. Beyond trajectory prediction, as done in the existing works, future research can focus on evaluating worker conditions and predicting their behavior for different applications in safety, as well as in productivity and quality.

Unlike machines, robots can be developed to recognize and imitate different gestures, actions, and emotions (Ji *et al.*, 2020), as observed in other industries like entertainment, healthcare, and education. Three steps are needed for these agents to imitate human actions - action recognition, action synthesis, and task-level planning (Krüger *et al.*, 2007). Beyond the action recognition proposed in the schema, it is suggested that these agents need to recognize the actions and movements of the observed agent (worker). Action synthesis is understanding the effects of the actions on the environment. Task-level planning is understanding how to act to cause the same effect on the environment using its end effectors (robot's arms, for example), which might differ from that of the observed agent (worker's arms). The concept of action understanding mentioned in the current work encompasses action recognition and action synthesis. Task-level planning is considered to be a different cognitive function and is not included since it does not involve the perception part, but is an internal simulation considering its own effectors. Action recognition is extensively discussed in the preceding content. Action synthesis consists of object state, affordances, and function understanding. These ideas are not at all explored in the construction domain, probably due to the limited research into robotics in the past. Affordance and function are essential concepts that can be integrated into the context modeling step of the taxonomy proposed. Functionality is the possible set of tasks that can be performed with an object. Affordances are the possible set of actions an environment allows and possible use cases of objects. Knowing the affordances enables agents to understand and interact with the environment effectively (Hassanin, Khan and Tahtali, 2021). These two steps will enable agents to recognize the activities better and predict future actions. In addition, the provision of valid functionality of objects enables the agents to identify anomalous usage of objects, such as the wrong usage of PPEs. Knowing the affordances also enables the agents to work autonomously with tools and explore creative solutions similar to those of their human counterparts.

There are no strict examples of ambient systems in the construction industry literature. Alternative terms used for these systems are Cyber-Physical Systems, Ambient Intelligence, Ubiquitous Computing, and Pervasive Computing, among others (Rocher *et al.*, 2020). These systems are predominantly developed, focusing on everyday life (Cook, Augusto and Jakkula, 2009) by sensing, reasoning, and acting upon the real world. Construction literature did not specifically utilize the terms of ambient systems, yet many systems developed can be considered industrial applications of ambient systems. Computer vision applications using far-field cameras placed over tower cranes can be considered part of ambient systems. These systems typically capture limited detail and resolution. While this problem limits the capturing of the elemental motion and actions of individuals, the systems take on the role of capturing the activities and processes within the environment effectively. Extending these systems for action recognition is a point of active research in construction literature (Luo *et al.*, 2019). Due to the need to collect as much information as possible from the environment, the placement of cameras is also an active research topic (Kim *et al.*, 2018; Yang *et al.*, 2018; Kim *et al.*, 2019; Chen *et al.*, 2021; Tran *et al.*, 2022). Stabilizing the inputs using features from videos (Kim *et al.*, 2019), Improving the scene illumination (Chen and Yu, 2023), and View invariance setup (Yan, Zhang and Li, 2019) are some interesting directions for establishing ambient systems. Techniques like image super-resolution can also be adopted to improve the detail, but have not been observed in the literature.

Due to the large volume of data that is possible to collect, ambient systems have many possible applications in construction. Social signal processing, proposed to understand actions in groups of people, uses human-human interactions as a basis. The interactions include behavioral cues like physical appearance, gesture and posture, face and eye behavior, vocal behavior, space, and environment. Such understanding can identify workplace social networks and aspects like team cohesiveness and leadership (Beyan, Vinciarelli and Bue, 2023). Further, crowd action analysis can help in public safety (emergency evacuations), anomalous individual identification, and crowd behavior understanding (Cristani *et al.*, 2013). The workspaces can be divided according to the crowd's actions in that locality (Luo *et al.*, 2019), identifying different work zones and catering to the needs of the workers.

In summary, with different levels of detail captured by different technologies, the technologies face a hard limitation in their application across the presented schema. Table 4 maps the discussed technologies and the schema elements to the high-level tasks and application use cases. Table 3 presented earlier maps the application use cases to the Construction 5.0 values, and hence, they are not discussed here specific to each technology. The specified use cases and value-related applications are not comprehensive, but they present interesting research directions relevant to the industry.

Table 5: Technology Categories Mapped to Application Use-cases.

Technology	Motion Detailing: Schema	High-Level Tasks	Application Use-cases
On-Body Devices: Exoskeletons	Elemental motion, Action	Intention recognition, Gesture recognition, Predictive and proactive action recognition, Action evaluation	Monitoring, Health, Safety, Productivity, Human-Robot Collaboration, Quality, and Skill.
Near-Person Agents: Robots	Action, Activity	Activity recognition, Behavior recognition, Condition recognition, Intention recognition, Acion synthesis, Affordance and functionality recognition	Productivity, Human-Robot Collaboration, Health, Safety, Quality.
Ambient Sytems: Remote Monitoring Systems	Activity, Process	Activity recognition, Process recognition, Social signal processing, Crowd action analysis, Work zone classification, Process evaluation	Monitoring, Health, Safety, Productivity.

The last step of the proposed schema - Process and its understanding would help identify the practices adopted in the field for various aspects like waste management, environmental friendliness, sustainability, and resilience. Monitoring a larger environment will help apply techniques like location-based management systems, theft monitoring, and safety hazards. Video retrieval (Ramezani and Yaghmaee, 2016) is necessary to identify relevant videos over many surveillance camera footage. Video classification and summarization through key frame selection and video skimming (Sabha and Selwal, 2023) will help compress large volumes of data from multiple cameras. From this compressed data, evaluating processes becomes more manageable.

In cases where algorithmic privacy preservation is not trusted, technological solutions can be applied to protect privacy at the source of data collection. In place of high-definition RGB videos that allow for the visual identification of workers, extremely low-resolution images (Yang *et al.*, 2024) and thermal cameras (Wu *et al.*, 2023) have shown considerable success. Additionally, technologies like thermal cameras work in low-light conditions, improving the resilience of the technology-based processes for action understanding.

Moving beyond the RGB camera, vision technologies with additional capabilities are also useful for action understanding. LIDAR, Stereo Vision, and RGB-D can provide depth information, Event cameras can provide motion information, and thermal cameras can provide thermal information. These technologies add helpful information for the low-level features mentioned in the schema if provided with the additional cost and technical expertise.

The ABCs of activity recognition, predominantly used for safety applications in literature, can also be utilized for worker health and well-being use-cases beneficial for achieving human-centricity and resilience. In the current work, an activity combines several actions and is part of a process. The workers do activities in line with the organizational goals. However, behavior is related to the worker's mood and other internal factors dictating their interactions with others and the work elements. For example, the same activity of masonry can be achieved by workers with quality-oriented behavior or very unsafe behavior, or passive behavior without interacting with other workers. Context in our work is related to several factors surrounding the activity and the worker. Activity-aware, Behavior-aware, and Context-aware computing (Favela, 2013) paradigms are already much-researched areas in the Internet of Things and sensor research. These systems are focused on providing individual-level feedback to users based on their application focus (Miranda, Viterbo and Bernardini, 2022). For example, Activity-aware systems can monitor workers' tasks to suggest productivity improvements. Activity-aware systems are also shown to improve motion prediction (Heravi *et al.*, 2024). Behavior can arise from several personal factors (Dávila-Montero *et al.*, 2021) without the control of the individuals. Behavior-aware systems can recognize unsafe practices of an individual or a group and alert a supervisor to potential hazards. Context-aware systems can observe

environmental and site conditions and promote well-being by adjusting workloads for workers. Since capturing these details with vision alone is extremely difficult, additional sensors or data sources must be in place to expand the action understanding into new paradigms.

These sensors can also feed into different contextual information that prove helpful for action understanding. External environmental context from additional input sensors (Ma *et al.*, 2023) is found useful in the reviewed literature. Motion sensors and altimeters (Khan *et al.*, 2022), Pressure and electroencephalography sensors (Xiahou *et al.*, 2023), and Electrodermal and electroencephalography sensors (Mehmood *et al.*, 2023) are shown to be helpful in action understanding. However, with different types of sensors, it is necessary to consider the match between applications and sensors, types of devices housing the sensors, positioning and orientation of the sensors, sampling rate, and application domain (Khan and Ghani, 2021). Activity recognition methods vary across sensor technologies (Ariza-Colpas et al., 2022), but the taxonomical hierarchy from the current work can be utilized.

5.3 Limitations

The current work focuses on technology-related issues while developing the taxonomy and has limited discussion over social and ethical issues like explainability and fairness, practical issues like installation feasibility and serviceability, and compute availability for large-scale or real-time applications pointed out as potential use cases. Considering the exponential growth of current automation technologies like language models and robotic agents, the potential applications presented can be considered quite limited since we primarily refer to the successful applications from past literature while synthesizing the directions. Beyond the algorithmic and technological directions, the current work presented no discussion related to system integration and implementation lifecycle. Despite several research studies, action understanding is still in a nascent phase. Significant challenges remain to be answered in the domain of artificial intelligence, with several higher-order issues in interpretation (Rodríguez *et al.*, 2014) as well as lower-order issues arising from the use of 2D RGB cameras (Jegham *et al.*, 2020; Pareek and Thakkar, 2021). Future research needs to take up such challenges in developing action-understanding-based applications.

6. CONCLUSION

In conclusion, worker action understanding is a foundational capability necessary for different technologies surrounding the workers, with use-cases in safety, productivity, and many more aspects of the processes. Through a double review, literature from computer vision and construction automation domains are combined to successfully develop a taxonomy of four levels essential for action understanding. Through the hierarchical taxonomy, critical issues like industry-specific action hierarchies and use cases and the semantic gap problem have been identified and addressed. The current state of the art is analyzed in the construction automation literature based on the taxonomy suggesting the predominant focus on safety and productivity applications, utilization of skeletal pose and bounding box as feature vectors, deep learning methods for the action recognition task, and utilization of project tasks and workplace environment as contextual factors for activity recognition, again through deep learning methods. Two future directions are presented, covering the algorithmic advancements relevant to the taxonomy and technology integrations relevant to the schema, which is a part of the presented taxonomy. The use of advanced low-level features and models, better learning approaches at mid-level tasks, and additional contextual information at high-level tasks are discussed in relation to the taxonomy. The potential applications using the vision-language connection, which are increasingly becoming relevant with the large language models and subsequent developments, are also discussed in the algorithmic advancements. The technological advancements are categorized as on-body devices, near-body agents, and ambient systems, and the potential applications are presented. In both directions, technologies that can provide privacy preservation are also discussed. The relevance of action understanding to the core values of Construction 5.0 is discussed throughout the discussion.

REFERENCES

Aafaq, N. et al. (2019) 'Video Description: A Survey of Methods, Datasets, and Evaluation Metrics', ACM Comput. Surv., 52(6). Available at: https://doi.org/10.1145/3355390.

Abu-Bakar, S.A.R. (2019) 'Advances in human action recognition: an updated survey', IET IMAGE PROCESSING. Available at: https://doi.org/10.1049/iet-ipr.2019.0350.

Afsar, P., Cortez, P. and Santos, H. (2015) 'Automatic visual detection of human behavior: A review from 2000 to 2014', Expert Systems with Applications. Available at: https://doi.org/10.1016/j.eswa.2015.05.023.

Aggarwal, J.K. and Park, S. (2004) 'Human motion: modeling and recognition of actions and interactions', Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Available at: https://doi.org/10.1109/TDPVT.2004.1335299.

Aggarwal, J.K. and Ryoo, M.S. (2011) 'Human activity analysis: A review', ACM Computing Surveys, 43(3). Available at: https://doi.org/10.1145/1922649.1922653.

Alsakka, F. et al. (2023) 'Computer vision applications in offsite construction', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2023.104980.

Arashpour, M., Ngo, T. and Li, H. (2021) 'Scene understanding in construction and buildings using image processing methods: A comprehensive review and a case study', JOURNAL OF BUILDING ENGINEERING. Available at: https://doi.org/10.1016/j.jobe.2020.101672.

Assadzadeh, A. et al. (2021) 'Automatic far-field camera calibration for construction scene analysis', Computer-Aided Civil and Infrastructure Engineering. Available at: https://doi.org/10.1111/mice.12660.

Astolfi, G. et al. (2021) 'Syntactic Pattern Recognition in Computer Vision: A Systematic Review', ACM Comput. Surv., 54(3). Available at: https://doi.org/10.1145/3447241.

Baskaran, P. and Adams, J.A. (2023) 'Multi-dimensional task recognition for human-robot teaming: literature review', Frontiers in Robotics and AI. Available at: https://doi.org/10.3389/frobt.2023.1123374.

Beddiar, D.R. et al. (2020) 'Vision-based human activity recognition: a survey', Multimedia Tools and Applications. Available at: https://doi.org/10.1007/s11042-020-09004-3.

Beyan, C., Vinciarelli, A. and Bue, A.D. (2023) 'Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey', ACM Comput. Surv., 56(5). Available at: https://doi.org/10.1145/3626516.

Blakemore, S.-J. and Decety, J. (2001) 'From the perception of action to the understanding of intention', Nature Reviews Neuroscience, 2(8), pp. 561–567. Available at: https://doi.org/10.1038/35086023.

Bonci, A. et al. (2021) 'Human-Robot Perception in Industrial Environments: A Survey', SENSORS, 21(5). Available at: https://doi.org/10.3390/s21051571.

Cai, J., Zhang, Y. and Cai, H. (2019) 'Two-step long short-term memory method for identifying construction activities through positional and attentional cues', AUTOMATION IN CONSTRUCTION. Available at: https://doi.org/10.1016/j.autcon.2019.102886.

Calvetti, D. et al. (2020) 'Worker 4.0: The Future of Sensored Construction Sites', BUILDINGS, 10(10). Available at: https://doi.org/10.3390/buildings10100169.

Chen, C., Zhu, Z. and Hammad, A. (2022) 'Critical Review and Road Map of Automated Methods for Earthmoving Equipment Productivity Monitoring', JOURNAL OF COMPUTING IN CIVIL ENGINEERING, 36(3). Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0001017.

Chen, S., Dong, F. and Demachi, K. (2023) 'Hybrid visual information analysis for on-site occupational hazards identification: A case study on stairway safety', Safety Science. Available at: https://doi.org/10.1016/j.ssci.2022.106043.

Chen, X. et al. (2021) 'BIM-based optimization of camera placement for indoor construction monitoring considering the construction schedule', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2021.103825.

Chen, X. and Yu, Y. (2023) 'Image Illumination Enhancement for Construction Worker Pose Estimation in Low-light Conditions', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Available at: https://doi.org/10.1007/978-3-031-25082-8_10.

Chu, W. et al. (2020) 'Monocular Vision-Based Framework for Biomechanical Analysis or Ergonomic Posture Assessment in Modular Construction', Journal of Computing in Civil Engineering. Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000897.

Cook, D.J., Augusto, J.C. and Jakkula, V.R. (2009) 'Ambient intelligence: Technologies, applications, and opportunities', Pervasive and Mobile Computing, 5(4), pp. 277–298. Available at: https://doi.org/10.1016/j.pmcj.2009.04.001.

Cristani, M. et al. (2013) 'Human behavior analysis in video surveillance: A Social Signal Processing perspective', Neurocomputing. Available at: https://doi.org/10.1016/j.neucom.2011.12.038.

Damen, D. et al. (2018) 'Scaling Egocentric Vision: The EPIC-KITCHENS Dataset', in. Proceedings of the European Conference on Computer Vision (ECCV), pp. 720–736. Available at: https://openaccess.thecvf.com/content_ECCV_2018/html/Dima_Damen_Scaling_Egocentric_Vision_ECCV_2018_paper.html (Accessed: 13 November 2024).

Dávila-Montero, S. et al. (2021) 'Review and Challenges of Technologies for Real-Time Human Behavior Monitoring', IEEE Transactions on Biomedical Circuits and Systems, 15(1), pp. 2–28. Available at: https://doi.org/10.1109/TBCAS.2021.3060617.

Demiris, Y. (2007) 'Prediction of intent in robotics and multi-agent systems', Cognitive Processing. Available at: https://doi.org/10.1007/s10339-007-0168-9.

European Commission et al. (2021) Industry 5.0 – Towards a sustainable, human-centric and resilient European industry. Publications Office of the European Union. Available at: https://doi.org/doi/10.2777/308407.

Everett, J.G. and Slocum, A.H. (1994) 'Automation and Robotics Opportunities: Construction versus Manufacturing', Journal of Construction Engineering and Management, 120(2), pp. 443–452. Available at: https://doi.org/10.1061/(ASCE)0733-9364(1994)120:2(443).

Fang, Q. et al. (2018) 'A deep learning-based method for detecting non-certified work on construction sites', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2018.01.001.

Fang, Q., Li, H., et al. (2020) 'A sematic and prior-knowledge-aided monocular localization method for construction-related entities', COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING. Available at: https://doi.org/10.1111/mice.12541.

Fang, W., Ding, L., et al. (2020) 'Computer vision applications in construction safety assurance', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2019.103013.

Fang, W., Love, P.E.D., et al. (2020) 'Computer vision for behaviour-based safety in construction: A review and future directions', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2019.100980.

Favela, J. (2013) 'Behavior-Aware Computing: Applications and Challenges', IEEE Pervasive Computing, 12(3), pp. 14–17. Available at: https://doi.org/10.1109/MPRV.2013.44.

Gammulle, H. et al. (2023) 'Continuous Human Action Recognition for Human-machine Interaction: A Review', ACM COMPUTING SURVEYS, 55(13 s). Available at: https://doi.org/10.1145/3587931.

Gao, R. et al. (2022) 'Review of the Application of Wearable Devices in Construction Safety: A Bibliometric Analysis from 2005 to 2021', BUILDINGS, 12(3). Available at: https://doi.org/10.3390/buildings12030344.

Guo, B.H.W. et al. (2021) 'Computer vision technologies for safety science and management in construction: A critical review and future research directions', SAFETY SCIENCE. Available at: https://doi.org/10.1016/j.ssci.2020.105130.

Guo, G. and Lai, A. (2014) 'A survey on still image based human action recognition', PATTERN RECOGNITION. Available at: https://doi.org/10.1016/j.patcog.2014.04.018.

Guo, K., Ishwar, P. and Konrad, J. (2013) 'Action Recognition from Video Using Feature Covariance Matrices', IEEE TRANSACTIONS ON IMAGE PROCESSING, 22(6), pp. 2479–2494. Available at: https://doi.org/10.1109/TIP.2013.2252622.

Han, S. et al. (2023) 'Four-Dimensional (4D) Millimeter Wave-Based Sensing and Its Potential Applications in Digital Construction: A Review', Buildings. Available at: https://doi.org/10.3390/buildings13061454.

Han, S., Lee, S. and Peña-Mora, F. (2014) 'Comparative Study of Motion Features for Similarity-Based Modeling and Classification of Unsafe Actions in Construction', JOURNAL OF COMPUTING IN CIVIL ENGINEERING. Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339.

Hassanin, M., Khan, S. and Tahtali, M. (2021) 'Visual Affordance and Function Understanding: A Survey', ACM Comput. Surv., 54(3). Available at: https://doi.org/10.1145/3446370.

Herath, S., Harandi, M. and Porikli, F. (2017) 'Going deeper into action recognition: A survey', Image and Vision Computing. Available at: https://doi.org/10.1016/j.imavis.2017.01.010.

Heravi, M. et al. (2024) 'Deep learning-based activity-aware 3D human motion trajectory prediction in construction', EXPERT SYSTEMS WITH APPLICATIONS. Available at: https://doi.org/10.1016/j.eswa.2023.122423.

Jegham, I. et al. (2020) 'Vision-based human action recognition: An overview and real world challenges', FORENSIC SCIENCE INTERNATIONAL-DIGITAL INVESTIGATION, 32. Available at: https://doi.org/10.1016/j.fsidi.2019.200901.

Ji, Y. et al. (2020) 'A Survey of Human Action Analysis in HRI Applications', IEEE Transactions on Circuits and Systems for Video Technology. Available at: https://doi.org/10.1109/TCSVT.2019.2912988.

Ke, S.-R. et al. (2013) 'A review on video-based human activity recognition', Computers. Available at: https://doi.org/10.3390/computers2020088.

Kendal, S.L. and Creen, M. (eds) (2007) 'An Introduction to Knowledge Engineering', in An Introduction to Knowledge Engineering. London: Springer, pp. 1–25. Available at: https://doi.org/10.1007/978-1-84628-667-4_1.

Khan, M. et al. (2022) 'Fall Prevention from Scaffolding Using Computer Vision and IoT-Based Monitoring', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/(ASCE)CO.1943-7862.0002278.

Khan, N.S. and Ghani, M.S. (2021) 'A Survey of Deep Learning Based Models for Human Activity Recognition', WIRELESS PERSONAL COMMUNICATIONS, 120(2), pp. 1593–1635. Available at: https://doi.org/10.1007/s11277-021-08525-w.

Kim, J. et al. (2018) 'Camera placement optimization for vision-based monitoring on construction sites', ISARC 2018 - 35th International Symposium on Automation and Robotics in Construction and International AEC/FM Hackathon: The Future of Building Things. Available at: https://doi.org/10.22260/isarc2018/0102.

Kim, J. et al. (2019) 'Systematic Camera Placement Framework for Operation-Level Visual Monitoring on Construction Jobsites', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/(ASCE)CO.1943-7862.0001636.

Kim, J. et al. (2022) 'Synthetic Training Image Dataset for Vision-Based 3D Pose Estimation of Construction Workers', Construction Research Congress 2022: Computer Applications, Automation, and Data Analytics - Selected Papers from Construction Research Congress 2022. Available at: https://doi.org/10.1061/9780784483961.027.

Kim, J. et al. (2023) 'Hybrid DNN training using both synthetic and real construction images to overcome training data shortage', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2023.104771.

Kong, Y. and Fu, Y. (2022) 'Human Action Recognition and Prediction: A Survey', International Journal of Computer Vision, 130(5), pp. 1366–1401. Available at: https://doi.org/10.1007/s11263-022-01594-9.

Koppula, H.S. and Saxena, A. (2016) 'Anticipating Human Activities Using Object Affordances for Reactive Robotic Response', IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1), pp. 14–29. Available at: https://doi.org/10.1109/TPAMI.2015.2430335.

Krüger, V. et al. (2007) 'The meaning of action:: a review on action recognition and mapping', Advanced Robotics. Available at: https://doi.org/10.1163/156855307782148578.

Kulsoom, F. et al. (2022) 'A review of machine learning-based human activity recognition for diverse applications', Neural Computing and Applications, 34(21), pp. 18289–18324. Available at: https://doi.org/10.1007/s00521-022-07665-9.

Lei, Q. et al. (2019) 'A Survey of Vision-Based Human Action Evaluation Methods', SENSORS, 19(19). Available at: https://doi.org/10.3390/s19194129.

Li, J. et al. (2024) 'A Review of Computer Vision-Based Monitoring Approaches for Construction Workers' Work-Related Behaviors', IEEE ACCESS. Available at: https://doi.org/10.1109/ACCESS.2024.3350773.

Li, L.-L. et al. (2023) 'Human Lower Limb Motion Intention Recognition for Exoskeletons: A Review', IEEE Sensors Journal, 23(24), pp. 30007–30036. Available at: https://doi.org/10.1109/JSEN.2023.3328615.

Li, X. et al. (2019) 'Automated post-3D visualization ergonomic analysis system for rapid workplace design in modular construction', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2018.11.012.

Li, Y. et al. (2017) 'Survey of Spatio-Temporal Interest Point Detection Algorithms in Video', IEEE Access. Available at: https://doi.org/10.1109/ACCESS.2017.2712789.

Lin, Z., Chen, A. and Hsieh, S. (2021) 'Temporal image analytics for abnormal construction activity identification', AUTOMATION IN CONSTRUCTION. Available at: https://doi.org/10.1016/j.autcon.2021.103572.

Liu, H. et al. (2020) 'Manifesting construction activity scenes via image captioning', AUTOMATION IN CONSTRUCTION. Available at: https://doi.org/10.1016/j.autcon.2020.103334.

Liu, J. et al. (2022) 'Detection and location of unsafe behaviour in digital images: A visual grounding approach', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2022.101688.

Liu, W. et al. (2021) 'Applications of computer vision in monitoring the unsafe behavior of construction workers: Current status and challenges', Buildings. Available at: https://doi.org/10.3390/buildings11090409.

Liu, Y. and Jebelli, H. (2022) 'Intention Estimation in Physical Human-Robot Interaction in Construction: Empowering Robots to Gauge Workers' Posture', Construction Research Congress 2022: Computer Applications, Automation, and Data Analytics - Selected Papers from Construction Research Congress 2022. Available at: https://doi.org/10.1061/9780784483961.065.

Liu, Y. and Jiao, S. (2022) 'Application of ST-GCN in unsafe action identification of construction workers', China Safety Science Journal. Available at: https://doi.org/10.16265/j.cnki.issn1003-3033.2022.04.005.

Luo, X. et al. (2019) 'Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning', COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING. Available at: https://doi.org/10.1111/mice.12419.

Ma, A. et al. (2023) 'Personalized Edge-Cloud Collaborative High-Fall-risk Monitoring Strategy for Grid Workers at Heights', 2023 IEEE Sustainable Power and Energy Conference, iSPEC 2023. Available at: https://doi.org/10.1109/iSPEC58282.2023.10403027.

Martinez, P. et al. (2021) 'A vision-based approach for automatic progress tracking of floor paneling in offsite construction facilities', AUTOMATION IN CONSTRUCTION. Available at: https://doi.org/10.1016/j.autcon.2021.103620.

Mehmood, I. et al. (2023) 'Multimodal integration for data-driven classification of mental fatigue during construction equipment operations: Incorporating electroencephalography, electrodermal activity, and video signals', Developments in the Built Environment. Available at: https://doi.org/10.1016/j.dibe.2023.100198.

Meng, Z. et al. (2020) 'Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis', ELECTRONICS, 9(9). Available at: https://doi.org/10.3390/electronics9091357.

Minh Dang, L. et al. (2020) 'Sensor-based and vision-based human activity recognition: A comprehensive survey', Pattern Recognition. Available at: https://doi.org/10.1016/j.patcog.2020.107561.

Miranda, L., Viterbo, J. and Bernardini, F. (2022) 'A survey on the use of machine learning methods in context-aware middlewares for human activity recognition', Artificial Intelligence Review. Available at: https://doi.org/10.1007/s10462-021-10094-0.

Mogadala, A., Kalimuthu, M. and Klakow, D. (2021) 'Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods', J. Artif. Int. Res., 71, pp. 1183–1317. Available at: https://doi.org/10.1613/jair.1.11688.

Moohialdin, A.S.M. et al. (2023) 'Proximity Activity Intensity Identification System in Hot and Humid Weather Conditions: Development and Implementation', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/JCEMD4.COENG-13332.

Moragane, H.P.M.N.L.B. et al. (2024) 'Application of computer vision for construction progress monitoring: a qualitative investigation', Construction Innovation. Available at: https://doi.org/10.1108/CI-05-2022-0130.

Morshed, M.G. et al. (2023) 'Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities', Sensors. Available at: https://doi.org/10.3390/s23042182.

Neuhausen, M., Herbers, P. and König, M. (2020) 'Using synthetic data to improve and evaluate the tracking performance of construction workers on site', Applied Sciences (Switzerland). Available at: https://doi.org/10.3390/app10144948.

Noceti, N. et al. (2017) 'Exploring Biological Motion Regularities of Human Actions: A New Perspective on Video Analysis', ACM Trans. Appl. Percept., 14(3). Available at: https://doi.org/10.1145/3086591.

Onofri, L. et al. (2016) 'A survey on using domain and contextual knowledge for human activity recognition in video streams', EXPERT SYSTEMS WITH APPLICATIONS. Available at: https://doi.org/10.1016/j.eswa.2016.06.011.

Pal, A. et al. (2023) 'Automated vision-based construction progress monitoring in built environment through digital twin', Developments in the Built Environment. Available at: https://doi.org/10.1016/j.dibe.2023.100247.

Pal, R. et al. (2021) 'Topic-based Video Analysis: A Survey', ACM Comput. Surv., 54(6). Available at: https://doi.org/10.1145/3459089.

Pan, Z. and Yu, Y. (2024a) 'Learning multi-granular worker intentions from incomplete visual observations for worker-robot collaboration in construction', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2023.105184.

Pan, Z. and Yu, Y. (2024b) 'Learning multi-granularity task primitives from construction videos for human-robot collaboration', Computing in Civil Engineering 2023: Data, Sensing, and Analytics - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2023. Available at: https://doi.org/10.1061/9780784485224.081.

Paneru, S. and Jeelani, I. (2021) 'Computer vision applications in construction: Current state, opportunities & challenges', AUTOMATION IN CONSTRUCTION, 132. Available at: https://doi.org/10.1016/j.autcon.2021.103940.

Pareek, P. and Thakkar, A. (2021) 'A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications', ARTIFICIAL INTELLIGENCE REVIEW, 54(3), pp. 2259–2322. Available at: https://doi.org/10.1007/s10462-020-09904-8.

Pentland, A. (2007) 'Social signal processing', IEEE Signal Processing Magazine, 24(4), pp. 108–111. Available at: https://doi.org/10.1109/MSP.2007.4286569.

Pereira, A. et al. (2023) 'From a Visual Scene to a Virtual Representation: A Cross-Domain Review', IEEE Access, 11, pp. 57916–57933. Available at: https://doi.org/10.1109/ACCESS.2023.3283495.

Perera, S. et al. (2023) 'Exoskeletons for Manual Handling: A Scoping Review', IEEE Access. Available at: https://doi.org/10.1109/ACCESS.2023.3323249.

Rabbi, A.B.K. and Jeelani, I. (2024) 'Computer Vision-Based Automatic Emergency Notification System: Interpreting Construction Workers' Hand Gestures', Computing in Civil Engineering 2023: Resilience, Safety, and Sustainability - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2023. Available at: https://doi.org/10.1061/9780784485248.056.

Ramezani, M. and Yaghmaee, F. (2016) 'A review on human action analysis in videos for retrieval applications', Artificial Intelligence Review. Available at: https://doi.org/10.1007/s10462-016-9473-y.

Ray, A. and Kolekar, M.H. (2024) 'Transfer learning and its extensive appositeness in human activity recognition: A survey', EXPERT SYSTEMS WITH APPLICATIONS, 240. Available at: https://doi.org/10.1016/j.eswa.2023.122538.

Ren, R. and Zhang, J. (2021) 'Semantic Rule-Based Construction Procedural Information Extraction to Guide Jobsite Sensing and Monitoring', Journal of Computing in Civil Engineering. Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000971.

Roberts, D. et al. (2020) 'Vision-Based Construction Worker Activity Analysis Informed by Body Posture', JOURNAL OF COMPUTING IN CIVIL ENGINEERING. Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000898.

Rocher, G. et al. (2020) 'Overview and Challenges of Ambient Systems, Towards a Constructivist Approach to their Modelling'. arXiv. Available at: https://doi.org/10.48550/arXiv.2001.09770.

Rodríguez, N.D. et al. (2014) 'A survey on ontologies for human behavior recognition', ACM COMPUTING SURVEYS, 46(4). Available at: https://doi.org/10.1145/2523819.

Ryu, J. et al. (2022) 'Automatic clustering of proper working postures for phases of movement', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2022.104223.

Sabha, A. and Selwal, A. (2023) 'Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions', MULTIMEDIA TOOLS AND APPLICATIONS, 82(21), pp. 32635–32709. Available at: https://doi.org/10.1007/s11042-023-14925-w.

Sargano, A.B., Angelov, P. and Habib, Z. (2017) 'A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition', Applied Sciences (Switzerland). Available at: https://doi.org/10.3390/app7010110.

Sedmidubsky, J. et al. (2021) 'Content-Based Management of Human Motion Data: Survey and Challenges', IEEE ACCESS, 9, pp. 64241–64255. Available at: https://doi.org/10.1109/ACCESS.2021.3075766.

Seo, J. et al. (2015) 'Computer vision techniques for construction safety and health monitoring', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2015.02.001.

Seo, J.L., SangHyun; Seo, Jongwon (2016) 'Simulation-Based Assessment of Workers' Muscle Fatigue and Its Impact on Construction Operations', Journal of Construction Engineering and Management, 142(11), pp. 04016063-NA. Available at: https://doi.org/10.1061/(asce)co.1943-7862.0001182.

Sezer, O.B., Dogdu, E. and Ozbayoglu, A.M. (2018) 'Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey', IEEE Internet of Things Journal, 5(1), pp. 1–27. Available at: https://doi.org/10.1109/JIOT.2017.2773600.

Shen, J. et al. (2021) 'Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning', Computer-Aided Civil and Infrastructure Engineering. Available at: https://doi.org/10.1111/mice.12579.

Shen, J. et al. (2023) 'A self-supervised monocular depth estimation model with scale recovery and transfer learning for construction scene analysis', COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING. Available at: https://doi.org/10.1111/mice.12938.

Sherafat, B. et al. (2020) 'Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/(ASCE)CO.1943-7862.0001843.

Tian, Y. et al. (2022) 'Construction motion data library: an integrated motion dataset for on-site activity recognition', SCIENTIFIC DATA. Available at: https://doi.org/10.1038/s41597-022-01841-1.

Tian, Y. et al. (2023) 'Multiple-input streams attention (MISA) network for skeleton-based construction workers' action recognition using body-segment representation strategies', AUTOMATION IN CONSTRUCTION, 156. Available at: https://doi.org/10.1016/j.autcon.2023.105104.

Tran, S.V.-T. et al. (2022) 'Generative planning for construction safety surveillance camera installation in 4D BIM environment', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2021.104103.

Turaga, P. et al. (2008) 'Machine Recognition of Human Activities: A Survey', IEEE Transactions on Circuits and Systems for Video Technology, 18(11), pp. 1473–1488. Available at: https://doi.org/10.1109/TCSVT.2008.2005594.

Vahdani, E. and Tian, Y. (2023) 'Deep Learning-Based Action Detection in Untrimmed Videos: A Survey', IEEE Transactions on Pattern Analysis and Machine Intelligence. Available at: https://doi.org/10.1109/TPAMI.2022.3193611.

Van-Horenbeke, F.A. and Peer, A. (2021) 'Activity, Plan, and Goal Recognition: A Review', Frontiers in Robotics and AI. Available at: https://doi.org/10.3389/frobt.2021.643010.

Vrigkas, M., Nikou, C. and Kakadiaris, I.A. (2015) 'A review of human activity recognition methods', Frontiers Robotics AI. Available at: https://doi.org/10.3389/frobt.2015.00028.

Wang, D. et al. (2021) 'Real-time monitoring for vibration quality of fresh concrete using convolutional neural networks and IoT technology', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2020.103510.

Wang, H., Zhao, G. and Yuan, J. (2014) 'Visual pattern discovery in image and video data: a brief survey', WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY, 4(1), pp. 24–37. Available at: https://doi.org/10.1002/widm.1110.

Willems, R.M., Özyürek, A. and Hagoort, P. (2007) 'When Language Meets Action: The Neural Integration of Gesture and Speech', Cerebral Cortex, 17(10), pp. 2322–2333. Available at: https://doi.org/10.1093/cercor/bhl141.

Wiriyathammabhum, P. et al. (2016) 'Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics', ACM Comput. Surv., 49(4). Available at: https://doi.org/10.1145/3009906.

Woznowski, P., Kaleshi, D., et al. (2016) 'Classification and suitability of sensing technologies for activity recognition', COMPUTER COMMUNICATIONS. Available at: https://doi.org/10.1016/j.comcom.2016.03.006.

Woznowski, P., Burrows, A., et al. (2016) 'SPHERE: A Sensor Platform for Healthcare in a Residential Environment', Designing, Developing, and Facilitating Smart Cities: Urban Design to IoT Solutions. Available at: https://doi.org/10.1007/978-3-319-44924-1_14.

Wu, H. et al. (2022) 'A survey on teaching workplace skills to construction robots', EXPERT SYSTEMS WITH APPLICATIONS. Available at: https://doi.org/10.1016/j.eswa.2022.117658.

Wu, H. et al. (2023) 'Thermal image-based hand gesture recognition for worker-robot collaboration in the construction industry: A feasible study', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2023.101939.

Wurm, M.F. and Schubotz, R.I. (2017) 'What's she doing in the kitchen? Context helps when actions are hard to recognize', PSYCHONOMIC BULLETIN & REVIEW, 24(2), pp. 503–509. Available at: https://doi.org/10.3758/s13423-016-1108-4.

Xiahou, X. et al. (2023) 'A Feature-Level Fusion-Based Multimodal Analysis of Recognition and Classification of Awkward Working Postures in Construction', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/JCEMD4.COENG-13795.

Xin, W. et al. (2023) 'Transformer for Skeleton-based action recognition: A review of recent advances', Neurocomputing. Available at: https://doi.org/10.1016/j.neucom.2023.03.001.

Xu, S. et al. (2021) 'Computer Vision Techniques in Construction: A Critical Review', ARCHIVES OF COMPUTATIONAL METHODS IN ENGINEERING, 28(5), pp. 3383–3397. Available at: https://doi.org/10.1007/s11831-020-09504-3.

Xu, W. and Wang, T.-K. (2023) 'Construction Worker Safety Prediction and Active Warning Based on Computer Vision and the Gray Absolute Decision Analysis Method', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/JCEMD4.COENG-12695.

Yan, X., Zhang, H. and Li, H. (2019) 'Estimating Worker-Centric 3D Spatial Crowdedness for Construction Safety Management Using a Single 2D Camera', Journal of Computing in Civil Engineering. Available at: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000844.

Yang, J. et al. (2015) 'Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future', Advanced Engineering Informatics. Available at: https://doi.org/10.1016/j.aei.2015.01.011.

Yang, J., Shi, Z. and Wu, Z. (2016) 'Vision-based action recognition of construction workers using dense trajectories', ADVANCED ENGINEERING INFORMATICS. Available at: https://doi.org/10.1016/j.aei.2016.04.009.

Yang, M. et al. (2024) 'A teacher-student deep learning strategy for extreme low resolution unsafe action recognition in construction projects', ADVANCED ENGINEERING INFORMATICS. Available at: https://doi.org/10.1016/j.aei.2023.102294.

Yang, X. et al. (2018) 'Computer-Aided Optimization of Surveillance Cameras Placement on Construction Sites', Computer-Aided Civil and Infrastructure Engineering. Available at: https://doi.org/10.1111/mice.12385.

You, K., Zhou, C. and Ding, L. (2023) 'Deep learning technology for construction machinery and robotics', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2023.104852.

Yousefi, B. and Loo, C.K. (2019) 'Biologically-inspired computational neural mechanism for human action/activity recognition: A review', Electronics (Switzerland). Available at: https://doi.org/10.3390/electronics8101169.

Yu, Y. et al. (2017) 'An experimental study of real-time identification of construction workers' unsafe behaviors', Automation in Construction. Available at: https://doi.org/10.1016/j.autcon.2017.05.002.

Zeng, C. and Hartmann, T. (2023) 'Towards a semantic enriching framework for construction site images', eWork and eBusiness in Architecture, Engineering and Construction - Proceedings of the 14th European Conference on Product and Process Modelling, ECPPM 2022. Available at: https://doi.org/10.1201/9781003354222-46.

Zhai, P., Wang, J. and Zhang, L. (2023) 'Extracting Worker Unsafe Behaviors from Construction Images Using Image Captioning with Deep Learning-Based Attention Mechanism', Journal of Construction Engineering and Management. Available at: https://doi.org/10.1061/JCEMD4.COENG-12096.

Zhang, Y. et al. (2022) 'Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly', ADVANCED ENGINEERING INFORMATICS. Available at: https://doi.org/10.1016/j.aei.2022.101792.

Zhong, B. et al. (2019) ‘Mapping computer vision research in construction: Developments, knowledge gaps and implications for research’, AUTOMATION IN CONSTRUCTION, 107. Available at: https://doi.org/10.1016/j.autcon.2019.102919.

Zhong, B. et al. (2023) ‘Visual attention framework for identifying semantic information from construction monitoring video’, Safety Science. Available at: https://doi.org/10.1016/j.ssci.2023.106122.

Zhu, Z., Dutta, A. and Dai, F. (2021) ‘Exoskeletons for manual material handling – A review and implication for construction applications’, Automation in Construction, 122, p. 103493. Available at: https://doi.org/10.1016/j.autcon.2020.103493.

Ziaeefard, M. and Bergevin, R. (2015) ‘Semantic human activity recognition: A literature review’, Pattern Recognition. Available at: https://doi.org/10.1016/j.patcog.2015.03.006.