DATA-DRIVEN APPROACH TO HOLISTIC SITUATIONAL AWARENESS IN CONSTRUCTION SITE SAFETY MANAGEMENT

by

Jiannan Cai

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Lyles School of Civil Engineering

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Hubo Cai, Chair
Lyles School of Civil Engineering

Dr. Dulcy Abraham
Lyles School of Civil Engineering

Dr. Mary Comer
School of Electrical and Computer Engineering

Dr. Phillip Dunston
Lyles School of Civil Engineering

Dr. Ayman Habib
Lyles School of Civil Engineering

Approved by:
Dr. Dulcy Abraham

Dedicated to my beloved family

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my Ph.D. advisor, Professor Hubo Cai, for his generous mentorship and tremendous supports during my four-year Ph.D. study at Purdue. I am always impressed and inspired by his vision and creativity in research. I enjoyed every discussion with Professor Cai, which is so beneficial to develop my logical and critical thinking capabilities and my communication skills. I appreciate that Professor Cai is always ready to answer questions and solve problems for students, and I am thankful for his efforts in improving our research skills. He revises my manuscripts sentence-by-sentence and provides constructive suggestions. He encourages and guides us to think independently and critically. I am grateful to his supports in preparing for my future career, including his mentorship in improving my teaching skills and guidance on developing research proposals. I also value the opportunity to be exposed to interdisciplinary research topics and advanced technologies in Professor Cai's group. I have learned and grown a lot both academically and professionally and built a solid foundation for my future academic career, which I owe a large extent to my advisor's mentorship and supports.

It is my great honor to have Professor Dulcy Abraham, Professor Phillips Dunston, Professor Ayman Habib, and Professor Mary Comer in my dissertation committee. Professor Abraham is always encouraging and supporting and has provided many constructive suggestions in conducting research and preparing research proposals. I treasure the experience of working with her in the student organization, where her enthusiasm, rigorousness, and always being well-organized influence me significantly and help develop my leadership. I also appreciate the advice and supports from Professor Abraham in my job search. I thank Professor Dunston for always reviewing my research documents carefully and providing detailed and constructive comments for me to improve academic writing, and I appreciate his support in my job search. I would also like to thank Professor Habib for the careful review of my dissertation and the valuable suggestions. It was in his class where I built technical foundation for my research. My Ph.D. research benefits a lot from his in-depth explanation and instruction on photogrammetry and camera systems. I thank Professor Comer for serving on my committee and bringing in new perspectives from a different discipline. My understanding of probability theory is deepened owing to her class.

I am grateful to my past and current colleagues. I thank Dr. Shuai Li for his mentorship, guidance, and help in developing research ideas and improving academic writing from the beginning of and throughout my Ph.D. study. I would not have conducted my research so smoothly and had so fruitful outcomes without his help. I thank Dr. Chenxi Yuan for being extremely patient, responsible, and resourceful when answering my questions and solving technical issues for me. I treasure the memory of being labmate with Xin Xu and am thankful for his kindness, patience, and encouragement in every conversation both academically and personally during this long journey. I thank Yuxi Zhang for helping me collect data and editing manuscripts, and sharing many research tasks. I enjoy every discussion on research ideas and thoughts with Yuxi. I thank JungHo Jeon for being responsive and active in our collaborative project and publication, and his help in conducting experiment in my research. I thank Liu Yang for preparing data, editing manuscripts, and conducting experiment for my research.

I would also like to thank all staff in the Lyles School of Civil Engineering, especially Ms. Jennifer Risky, for their professional, timely, and cardinal assistance, which makes the study process smooth and pleasant. I thank current and past members of Civil Engineering Graduate Student Advisory Council. It was such a wonderful experience to work with them and take the efforts to make all graduate students feel connected and supported.

Finally, I express my deepest gratitude and love to my family, for their endless love, support, understanding, and encouragement over the years. I thank my husband for being a true friend and partner and appreciate his encouragement and inspiration along my study. Without all these people, I would not have become who I am, and they are the most precious treasure in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADE	Average displacement error
ANN	Artificial neural network
BIM	Building information model
BLE	Bluetooth low energy
CNN	Convolutional Neural Network
D	Down
DT	Decision tree
E	East
FDE	Average final displacement error
FM	Fragmentation
FN	False negative
FOV	Field of view
FP	False positive
GCP	Ground control point
GPS	Global positioning system
GT	Ground truth
H	Horizontally
HOG	Histogram of orientated gradients
ID	Identity
IDSW	Identity switch
IMU	Inertial measurement unit
LSTM	Long short-term memory
MDP	Markov decision process
MOT	Multi-object tracking
MOTA	Multiple object tracking accuracy
MOTP	Multiple object tracking precision
MSE	Mean squared error
N	North
NE	Northeast
NW	Northwest
OSHA	Occupational Safety and Health Administration
RFID	Radio-frequency identification
RNN	Recurrent neural network
RSSI	Received signal strength indication
RTLS	Real-time locating systems
S	South
SE	Southeast
seq2seq	Sequence-to-sequence

SIFT	Scale-invariant feature transform
SLAM	Simultaneous localization and mapping
SVM	Support vector machine
SW	Southwest
U	Up
UAV	Unmanned aerial vehicle
UGV	Unmanned ground vehicle
UWB	Ultra-wideband
W	West

ABSTRACT

The motivation for this research stems from the promise of coupling multi-sensory systems and advanced data analytics to enhance holistic situational awareness and thus prevent fatal accidents in the construction industry. The construction industry is one of the most dangerous industries in the U.S. and worldwide. Occupational Safety and Health Administration (OSHA) reports that the construction sector employs only 5% of the U.S. workforce, but accounts for 21.1% (1,008 deaths) of the total worker fatalities in 2018. The struck-by accident is one of the leading causes and it alone led to 804 fatalities between 2011 and 2015. A critical contributing factor to struck-by accidents is the lack of holistic situational awareness, attributed to the complex and dynamic nature of the construction environment. In the context of construction site safety, situational awareness consists of three progressive levels: perception – to perceive the status of construction entities on the jobsites, comprehension – to understand the ongoing construction activities and interactions among entities, and projection – to predict the future status of entities on the dynamic jobsites. In this dissertation, holistic situational awareness refers to the achievement at all three levels. It is critical because with the absence of holistic situational awareness, construction workers may not be able to correctly recognize the potential hazards and predict the severe consequences, either of which will pose workers in great danger and may result in construction accidents. While existing studies have been successful, at least partially, in improving the perception of real-time states on construction sites such as locations and movements of jobsite entities, they overlook the capability of understanding the jobsite context and predicting entity behavior (i.e., movement) to develop the holistic situational awareness. This presents a missed opportunity to eliminate construction accidents and save hundreds of lives every year. Therefore, there is a critical need for developing holistic situational awareness of the complex and dynamic construction sites by accurately perceiving states of individual entities, understanding the jobsite contexts, and predicting entity movements.

The overarching goal of this research is to minimize the risk of struck-by accidents on construction jobsite by enhancing the holistic situational awareness of the unstructured and dynamic construction environment through a novel data-driven approach. The research rationale is such that with enhanced holistic situational awareness, the site dynamics can be accurately perceived, the jobsite contexts in terms of working groups and ongoing activities can be correctly

understood, and future states can be reliably predicted. These capabilities will enable the proactive detection of the potential collision hazards, based on which early warnings and instructions can be provided to involved entities for them to take proactive actions to prevent struck-by accidents. Towards that end, three fundamental knowledge gaps/challenges have been identified and each of them is addressed in a specific objective in this research.

The first knowledge gap is the lack of methods in fusing heterogeneous data from multimodal sensors to accurately perceive the dynamic states of construction entities. The congested and dynamic nature of construction sites has posed great challenges such as signal interference and line of sight occlusion to a single mode of sensor that is bounded by its own limitation in perceiving the site dynamics. The research hypothesis is that combining data of multimodal sensors that serve as mutual complementation achieves improved accuracy in perceiving dynamic states of construction entities. This research proposes a hybrid framework that leverages vision-based localization and radio-based identification for robust 3D tracking of multiple construction workers. It treats vision-based tracking as the main source to obtain object trajectory and radio-based tracking as a supplementary source for reliable identity information. It was found that fusing visual and radio data increases the overall accuracy from 88% and 87% to 95% and 90% in two experiments respectively for 3D tracking of multiple construction workers, and is more robust with the capability to recover the same entity ID after fragmentation compared to using vision-based approach alone.

The second knowledge gap is the missing link between entity interaction patterns and diverse activities on the jobsite. With multiple construction workers and equipment co-exist and interact on the jobsite to conduct various activities, it is extremely difficult to automatically recognize ongoing activities only considering the spatial relationship between entities using pre-defined rules, as what has been done in most existing studies. The research hypothesis is that incorporating additional features such as attentional cues better represents entity interactions and advanced deep learning techniques automates the learning of the complex interaction patterns underlying diverse activities. This research proposes a two-step long short-term memory (LSTM) approach to integrate the positional and attentional cues to identify working groups and recognize corresponding group activities. A series of positional and attentional cues are modeled to represent the interactions among entities, and the LSTM network is designed to (1) classify whether two entities belong to the same group, and (2) recognize the activities they are involved in. It was found that by leveraging both positional and attentional cues, the accuracy increases from 85% to 95% compared with cases using positional cues alone. Moreover, dividing the group activity recognition task into a two-step cascading process improves the precision and recall rates of specific activities by about 3%-12% compared to simply conducting a one-step activity recognition.

The third knowledge gap is the non-determining role of jobsite context on entity movements. Worker behavior on a construction site is goal-based and purposeful, motivated and influenced by the jobsite context including their involved activities and the status of other entities. Construction workers constantly adjust their movements in the unstructured and dynamic workspace, making it challenging to reliably predict worker trajectory only considering their previous movement patterns. The research hypothesis is that combining the movement patterns of the target entity with the jobsite context more accurately predicts the trajectory of the entity. This research proposes a context-augmented LSTM method, which incorporates both individual movement and workplace contextual information, for better trajectory prediction. Contextual information regarding movements of neighboring entities, working group information, and potential destination information is concatenated with movements of the target entity and fed into an LSTM network with an encoder-decoder architecture to predict trajectory over multiple time steps. It was found that integrating contextual information with target movement information can result in a smaller final displacement error compared to that obtained only considering the previous movement, especially when the length of prediction is longer than the length of observation. Insights are also provided on the selection of appropriate methods.

The resulting holistic situational awareness of the dynamic construction site consists of the following information.

- The positional states of construction entities are continuously perceived by fusing heterogeneous data.
- Jobsite context including working groups and ongoing group activities are recognized by exploiting the entity interaction patterns over a period of observations.
- The trajectories of entities are predicted given their current states as well as the jobsite context.

The results and findings of this dissertation will augment the holistic situational awareness of site entities in an automatic way and enable them to have a better understanding of the ongoing jobsite context and a more accurate prediction of future states, which in turn allows the proactive detection of any potential collisions. This augmented capability can be implemented to a system approach for struck-by prevention, where vision and radio systems can be used to collect data on entity states. The sensory data will be transmitted to a central server to perform analysis using algorithms developed in this dissertation, including state perception, jobsite context comprehension, and trajectory prediction. Based on the resulting holistic situational awareness, a proactive and context-aware struck-by prevention mechanism can be devised to provide early warnings when collision risk is high as well as plan optimal path for site entities to actively adjust their behavior to avoid potential collision. Such information and guidance can be communicated to field crews in different formats through mobile devices. By doing these, the jobsite entities are augmented with holistic and ubiquitous situational awareness to prevent struck-by accidents.

This newly enhanced capacity of holistic situational awareness is possible to be extended to prevent other types of accidents, such as fall accidents and electrocutions. Besides, it has the great potential to contribute to automatic construction progress monitoring and control. By integrating holistic situational awareness and construction plans and representations such as building information models (BIMs), one can easily tell whether construction entities are at the right place doing the right tasks with the right partners, which in turn facilitates the active control of construction operation to ensure productivity. Furthermore, on future construction jobsites where human workers and robots are expected to collaborate ubiquitously, the methods created in this dissertation research are promising to be adopted in human-robot collaboration and empower the robots with automatic situational awareness to adaptively adjust their behavior to effectively and efficiently collaborate with human workers.

1. INTRODUCTION

The construction industry is one of the most dangerous industries. In 2018, the construction sector employed only 5% of the US workforce (U.S.Bureau of Labor Statistics, 2018), but it accounted for 21.1% (1008 deaths) of the total worker fatalities (OSHA, 2018). The struck-by hazard is one of the leading causes and it alone led to 1017 fatalities from 2012 to 2018 (U.S.Bureau of Labor Statistics, 2018). A critical contributing factor to struck-by accidents is the lack of holistic situational awareness due to the complex and dynamic nature of the construction environment. This long-standing and pressing problem requires an effective solution to enhance the holistic situational awareness of the construction site to proactively prevent struck-by accidents and save hundreds of lives every year. This chapter provides an overview of this research.

1.1 Background

The construction site is dynamic and complex in nature. Various activities are performed simultaneously with numerous resources (e.g., equipment, workers, and materials) in shared working spaces (D. Fang & Wu, 2013). The unstructured and dynamic site conditions make it extremely difficult for jobsite entities to sufficiently perceive their surroundings, which is an immediate cause for many struck-by accidents. Between 1990 and 2007, 659 fatalities in the U.S. construction industry were caused by failures in perceiving the site dynamics when people are occluded by obstructions or blind spots (Hinze & Teizer, 2011). From 2003 to 2010, on average 53% of fatal accidents were struck-by-vehicle or equipment overturns and collisions (U.S.Bureau of Labor Statistics, 2013). In addition, during the construction process, workers are routinely challenged to make their own decisions when confronted with new problems and situations (D. Fang & Wu, 2013; Li et al., 2015). They may expose themselves to a potential hazard due to insufficient safety knowledge and inadequate prediction of consequences even though they have perceived the current status (Bohm & Harris, 2010; Rundmo, 2001). In such a case, situational awareness, which includes the perception, comprehension, and prediction of jobsite status, is essential for effective decision-making to ensure construction safety.

A widely adopted model of situational awareness was developed by Endsley (1995), who defined situational awareness as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future”. Such a definition includes three hierarchical levels: perception – to perceive the status, attributes, and dynamics of relevant elements in the environment, comprehension – to synthesize the states and understand the functionality of each element, and projection – to predict the future behavior of the elements in the environment. In this research, holistic situational awareness refers to the achievement at all three levels.

In the current practice, safety training and site monitoring are the two major measures to improve situational awareness. Before conducting construction tasks, workers are trained to learn the proper working procedures, identify potential hazards, and use safety devices. The training is effective and necessary in enhancing worker safety knowledge and correcting their safety attitude such that the worker is equipped with the ability to correctly perceive and recognize potential hazards. However, workers may not follow the safe practice on the site because of fatigue, distraction, and schedule pressures. On the other hand, site spotters are assigned to monitor the ongoing construction operations, especially for those involving interactions between workers and heavy machines. Workers and machine operators are alerted by site spotters regarding their surroundings when they are focusing on construction tasks without sufficient situational awareness. Nevertheless, the site monitoring is labor-intensive and error-prone as one spotter can only monitor a limited number of entities within a small range of areas with the performance heavily relying on experience and skills as well as the viewing point. Alternatively, with the advances in sensing technology and data analytics, both practitioners and researchers have acknowledged the potential and shown increasing interest in coupling multimodal sensors with advanced data analytics to automatically enhance worker situational awareness in order to ensure construction site safety (T. Cheng & Teizer, 2012; Hwang, 2012).

1.2 Problem Statement

Achieving holistic situational awareness of the construction site requires 1) accurate perception of the states of individual entities, 2) correct understanding/interpretation of the jobsite context, and 3) reliable prediction of entity movements in the near future. While existing studies have been successful, at least partially, in enhancing the perception of site dynamics regarding real-time states of entities such as locations, they overlooked pursuing the capabilities of understanding the jobsite context and of predicting the behavior (i.e., movement) of the entities to realize holistic situational awareness, a missed opportunity to eliminate construction accidents and save hundreds of lives every year.

Construction sites are typically complex and unstructured, consisting of numerous construction resources involved in various activities (as shown in Figure 1.1), making it extremely difficult to achieve holistic situational awareness. Three main problems and challenges that hinder the achievement of holistic situational awareness have been identified as follows.

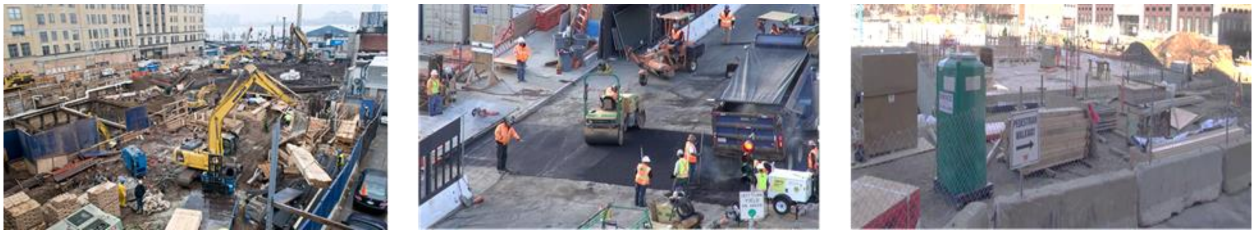


Figure 1.1 Examples of complex and unstructured construction sites

1. Lack of method in fusing heterogeneous data from multimodal sensors to accurately perceive the dynamic states of construction entities. The congested and dynamic nature of construction sites has posed great challenges such as signal interference and line of sight occlusion to state perception. It is difficult to accurately perceive site dynamics using a single mode of sensor. For instance, the visual sensor requires line-of-sight and is highly sensitive to illumination and occlusion. The radio-based sensor is significantly influenced by signal interference caused by obstacles on the jobsite. Therefore, multisensory data in varying formats and levels of accuracy must be integrated as mutual complementation to achieve improved accuracy in perceiving entities' states.
2. The link between entity interaction patterns and diverse activities on the jobsite is missing. Construction entities interact with each other to accomplish assigned tasks, formulating several working groups. It is difficult to interpret the jobsite context only from the states of individual entities, and thus, their interactions must be fully exploited and incorporated to comprehend the diverse activities on the jobsite.
3. Lack of scientific understanding of construction entities' behavior within the jobsite context. Worker behavior on a construction site is goal-based and purposeful, motivated and influenced by the jobsite context including their involved activities and the status of other entities. Workers constantly adjust their movements in the unstructured and dynamic workspace,

making it extremely challenging to reliably predict worker trajectory only considering their previous movement patterns. Therefore, the jobsite context must be integrated with the movement patterns of the target in order to more accurately predict the entity trajectory.

1.3 Review of Related Studies and Knowledge Gaps

Many research studies have attempted to develop and enhance the situational awareness of construction workers in order to improve site safety performance. This section reviews the related studies and highlights the knowledge gaps.

1.3.1 Knowledge gap in state perception of construction entities

Various sensing technologies and corresponding algorithms have been developed to enhance perception - the first level of situational awareness - by automatically monitoring and perceiving the states (e.g., location and motion) of site entities. For instance, imaging sensors (e.g., video cameras, depth cameras) are used to detect and localize construction workers and equipment (Memarzadeh et al., 2013; M. W. Park & Brilakis, 2016; Zhu et al., 2016b) as well as identify their postures (Ding et al., 2018; Y. Yu et al., 2017; H. Zhang et al., 2018). Radio-based sensors (e.g., radio-frequency identification (RFID), ultra-wideband (UWB), Bluetooth low energy (BLE) beacons) are applied for proximity detection (J. W. Park et al., 2017; Teizer et al., 2010) and 3D localization (H.-S. Lee et al., 2011; Topak et al., 2018). Inertial sensors can be used to estimate both posture (Valero et al., 2017; Yan et al., 2017) and location of construction entities (M. Ibrahim & Moselhi, 2016).

Most studies only leverage a single mode of sensor that is bounded by its own limitations. For instance, visual sensors may capture rich contextual information and they do not require attachment to the entities, but they are highly sensitive to illumination and occlusion. Radio-based sensors do not require line of sight, but their accuracy is significantly influenced by the obstacles on the construction site. Inertial sensors are relatively accurate in the short term, but subject to noise and drift errors. It is well acknowledged that a data fusion approach that integrates heterogeneous data obtained from different sensors has the attribute of mutual complementation and can improve the accuracy and confidence of the monitoring results. For instance, Chen et al. (2018) proposed a multisource fusion framework that combines indoor localization results obtained from BLE beacons and inertial measurement unit (IMU). Papaioannou et al. (2017) created a novel framework that fuses radio-, inertial-, and vision-based sensors to track construction workers on the jobsite. However, their method used a single camera and cannot provide reliable 3D location information in a large area.

Despite these pilot studies (Chen et al., 2018; Papaioannou et al., 2017), more research efforts are needed in fusing heterogeneous data obtained from different sensors to complement each other and obtain more accurate trajectories. For instance, vision-based and radio-based localization have their own advantages and limitations and has great potential to be combined to improve the accuracy and reliability of localization for both indoor and outdoor applications. Vision-based localization is accurate, but its performance will significantly decrease when it fails in detecting the target; radio-based localization is less accurate but reliable in object detection and identification. Novel methods are needed to integrate the strengths of these two approaches to realize long-term and robust tracking of construction entities. This knowledge gap will be addressed in this study by developing novel data fusion methods to integrate visual and radio-based sensors for accurate perception of positional states of construction entities.

1.3.2 Knowledge gap in construction activity and context recognition

Many studies have been dedicated to automating the interpretation of construction activity and jobsite context using information extracted from sensory data. There are typically two types of features used to infer the site context. The first type is the features extracted or computed from raw sensory data, such as spatial-temporal features of visual data (Golparvar-Fard et al., 2013; Gong et al., 2011; J. Y. Kim & Caldas, 2017; H. Luo et al., 2018), and time- or frequency- domain features of IMU data (Akhavian & Behzadan, 2015, 2016; Hyunsoo Kim et al., 2018). These features usually serve as inputs of machine learning classifiers to recognize activities of construction entities (i.e., workers and equipment). The other type of features is at a higher level, computed from the perceived entities' states, such as the location and proximity information obtained from the visual detection and tracking results (H. H. Kim et al., 2018; J. Kim et al., 2018; J. Yang et al., 2011). These features are typically used in a rule-based activity/context reasoning algorithm, which analyzes the construction operation based on the spatial-temporal relationship of entities.

Most existing studies (Ding et al., 2018; Golparvar-Fard et al., 2013; Gong et al., 2011) focus on the activity or action recognition of individual entities, such as traveling and working of workers, and moving and dumping of machines. These studies rely on features of individual entities, but neglect the dynamic interaction and collaboration among different entities. For the few studies on entity interactions (H. H. Kim et al., 2018; J. Kim et al., 2018), they mainly rely on the spatial-temporal relationship between entities through hand-crafted rules. Moreover, construction images/videos were simplified to only contain entities involved in a single activity, excluding all irrelevant entities (X. Luo et al., 2018). In reality, however, many workers and machines co-exist and collaborate to accomplish different tasks. For those entities that are spatially close, not all of them are collaborating on a single activity. Similarly, some entities interact and collaborate on a specific task even though they are not physically next to each other.

In such cases, worker's attentional information (e.g., the head direction) can serve as additional features to facilitate the interpretation of jobsite context. To address this knowledge gap, this research leverages both the positional and attentional relationships of multiple entities to represent their dynamic interaction and devises LSTM networks to automatically learn the linking between dynamic interaction patterns and diverse construction activities.

1.3.3 Knowledge gap in trajectory prediction of construction entities

To realize holistic situational awareness and further achieve proactive prevention of struck-by hazards, a critical step is the accurate prediction of entities' behavior, especially the prediction of trajectory. Conventionally, tracking filters are used to predict the future steps in a trajectory (Hermes et al., 2009; T. Liu et al., 1998; Prévost et al., 2007). For instance, the Kalman filter is applied to predict the trajectory using a Gaussian distribution with accumulated uncertainty. However, this approach often results in physically impossible locations (e.g., behind walls, within obstacles). Particle filters incorporate more sophisticated constraints and non-Gaussian distributions, but it degrades into random walks of feasible motion over large time horizons (Ziebart et al., 2009). Some researches (Karasev et al., 2016; Kitani et al., 2012; Rudenko et al., 2018; Ziebart et al., 2009) adopted planning-based approaches, where entities are treated as intelligent agents who actively plan their path to achieve a goal. The problem is formulated as a path planning or optimal control task, such as the Markov decision process (MDP). One main drawback is that the planning-based approach relies heavily on prior knowledge, and it still uses hand-crafted features to model states and reward functions that are specific to particular settings. Recently, data-driven approaches have been increasingly used given that they do not require explicitly modeling movement dynamics and their ability to be generalized to various scenarios. Long short-term memory (LSTM) network is the most widely used deep learning model for human trajectory prediction (Alahi et al., 2016; Saleh et al., 2018; Syed & Morris, 2019; Xue et al., 2018).

In the construction domain, Zhu et al. (2016a) proposed a novel Kalman filter to predict the movements of workers and mobile equipment using positions obtained from multiple video cameras. Although the methods achieved a submeter accuracy when predicting the next-step movement in 0.03s, the accuracy decreases to about 1.5m when predicting movement in 1.5s. Instead of using conventional tracking filters, Dong et al. (2018) and Rashid et al. (2018) modeled worker movements as a Markov process to predict their trajectories based on historical records. More recently, Kim et al. (2019) and Tang et al. (2019) attempted to predict the construction entity trajectory through a data-driven approach given the advances in deep learning techniques.

Most existing studies only consider individual movements while predicting a worker’s trajectory, which is insufficient to capture worker behavior under different scenarios. In reality, multiple entities co-exist on the construction site, forming various working groups to accomplish different activities. Workers’ behavior will be influenced by each other and the specific activities they are involved in. Jobsite contextual information such as entity interactions and ongoing activities must be incorporated in order to better predict entity movements, which, however, has been overlooked by existing studies in the construction domain. To address this knowledge gap, this research incorporates jobsite contextual information regarding entity interactions and involved activities to create a context-augmented deep learning model for worker trajectory prediction on dynamic and unstructured construction sites.

1.4 Research Goal and Objectives

The *overarching goal* of this research is to minimize the risk of struck-by accidents on construction jobsites by enhancing the holistic situational awareness of the unstructured and dynamic construction environment through a novel data-driven approach. The research rationale is such that with enhanced holistic situational awareness, the site dynamics can be accurately perceived, the ongoing activities can be correctly understood, and future states can be reliably predicted. These capabilities will enable the proactive detection of the potential collision hazards, based on which early warnings and instructions can be provided to involved entities for them to take proactive actions to prevent struck-by accidents. Three specific objectives are formulated to achieve the goal. Figure 1.2 illustrates the overview of the research. Entities' 3D positional states are first obtained by fusing vision- and radio-based sensing data. Second, positional and attentional cues are integrated to comprehend the jobsite context regarding working group and group activity. Finally, states of individual entities and the jobsite contextual information are integrated to predict entity trajectory. It is noted that in the first objective, the 3D trajectory of construction workers in world coordinate system is obtained using a hybrid stereo vision system and radio-based system. In the second and third objectives, due to the constraint of data availability, 2D construction videos are used to train and test the proposed deep learning models, where the positional and attentional cues are represented in 2D, and the worker trajectory is predicted on 2D image plane. The overview of the technical approaches for the three objectives is introduced as follows.

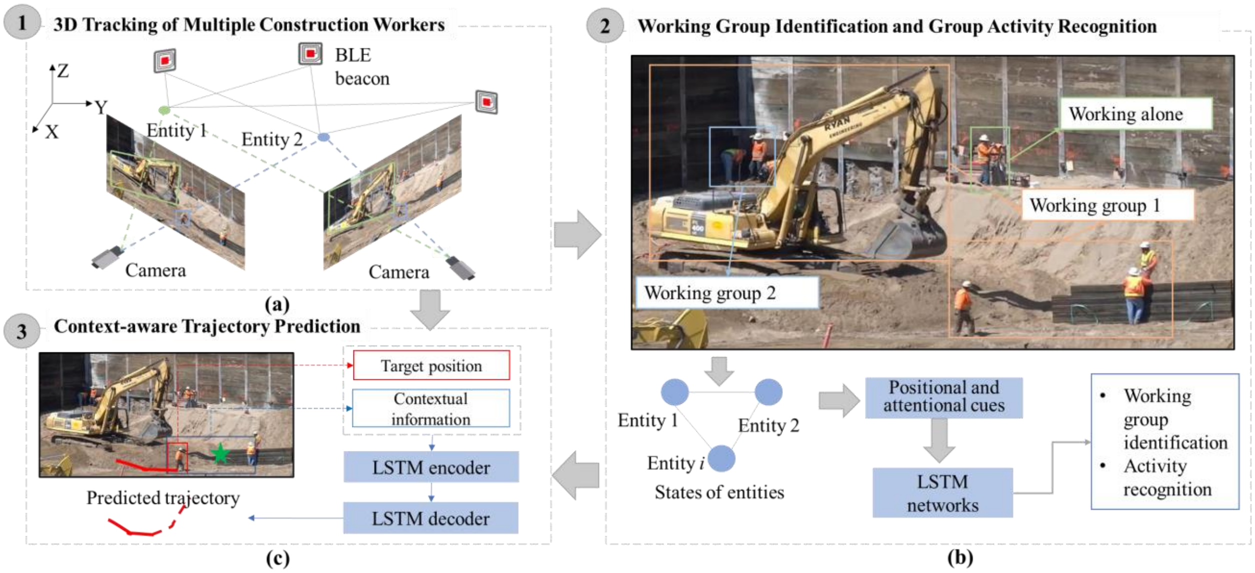


Figure 1.2 Research overview and technical approaches for three objectives: (a) 3D tracking of multiple construction workers, (b) working group identification and group activity recognition, (c) context-aware trajectory prediction.

- The first objective is to develop a novel framework that integrates imaging data and radio data to achieve robust 3D tracking of multiple construction entities (see Figure 1.2 (a)). The hypothesis is that integrating heterogeneous data from multi-modal sensors perceives the positional states of construction entities with improved accuracy and robustness. The research questions here are twofold: what sensors to use to obtain the positional information of the construction entities and how to combine heterogeneous information to improve the accuracy and robustness? In this objective, a hybrid system comprising a stereo-camera system and a radio-based system is used to track multiple construction workers. Vision-based tracking is treated as the main source to extract the object trajectory. Radio-based identification and localization results are used as a supplementary source to augment anonymous visual tracks with identity information and to correct errors (e.g., false positives) in vision-based object detection, resulting in ID-linked 3D trajectories. In addition, a searching algorithm is created to recover possible missed detections in one camera view from the corresponding observations in a second camera view. Two indoor experiments were conducted to validate the proposed method, where a stereo vision system and a radio-based (i.e., BLE beacon) system were used to obtain 3D trajectories of workers in the 3D world coordinate system. The accomplishment of this objective will achieve the accurate and robust perception of positional states of construction entities.

- The second objective is to identify construction working groups and recognize corresponding group activities by exploiting the interactions among entities leveraging positional and attentional cues (see Figure 1.2 (b)). The hypothesis is that integrating position- and attention-based cues facilitate the jobsite context reasoning. The two research questions posed here are: 1) what features are critical to reason out the context information and 2) what are the entity interaction patterns presented in the typical activity? In this objective, the spatial and attentional states of individual entities are represented numerically. Mathematical models are created to compute the spatial and attentional cues between two entities. Finally, a two-step long short-term memory (LSTM) network is devised to identify working groups and recognize group activities. Two sets of construction videos—one hospital construction project on the publicly-available website and one teaching building project taken by the author on Purdue campus, are used to validate the newly created method. Manual annotations regarding the 2D spatial and attentional states are used to compute the 2D positional and attentional cues proposed in this study. The group/non-group information and the corresponding construction activities for each pair of construction entities are manually labeled to provide ground truth labels for supervised learning when train and test the proposed two-step LSTM model. The accomplishment of this objective will enable the comprehension of jobsite context under general construction scenarios.

- The third objective is to create context-aware algorithms to predict the trajectories of entities (see Figure 1.2 (c)). The hypothesis is that the movement of construction entities is better predicted by integrating their previous states and jobsite contextual information. The twofold research question posited here is: 1) what contextual features are critical to worker trajectory prediction and 2) what is the connection between worker movements and jobsite contexts with their future trajectory? In this objective, an LSTM model augmented by the context information is proposed, which incorporates both individual movement and workplace contextual information. Contextual information regarding movements of neighboring entities, working group information, and potential destination information will be concatenated with movements of the target entity and fed into an LSTM network with an encoder-decoder architecture to enable the sequence-to-sequence prediction. The method is validated using videos collected on three construction sites—one hospital construction project and two teaching building construction sites. Visual data were pre-processed to extract entity positions and contextual features, which are then used as inputs to train and test the proposed method. The trajectory prediction is performed on the 2D image plane. The accomplishment of this objective will enable a more accurate trajectory prediction of construction workers on the unstructured and dynamic jobsites.

The expected outcomes are novel methods and new knowledge to enhance the holistic situational awareness in terms of three aspects, detailed as follows.

- *A new algorithm that fuses vision-based tracking and radio-based identification to enable accurate and robust perception of positional states of construction workers.* Workers' location information is augmented with identity information, which can be used to recover missed detection and avoid ambiguity in the dynamic and unstructured construction site. Moreover, the integration of multi-modal sensors that serve as mutual complementation can overcome the limitation of each type of sensor.
- *Identified critical features that capture the interactions among entities, and the corresponding deep learning model for jobsite context comprehension.* By representing the entity interactions with generic features identified in this dissertation, the jobsite context including both

construction working groups and ongoing group activities, under general construction scenarios, can be effectively recognized.

- *Identified critical contextual features and corresponding context-augmented deep learning model for accurate trajectory prediction of construction workers.* The newly created method allows the consideration of contextual information in predicting future states on construction sites. Insights are also provided on the selection of appropriate methods for effective and efficient trajectory prediction of construction entities.

The above three objectives are interrelated and each of them corresponds to one level of situational awareness. They together formulate the framework to enhance holistic situational awareness on the construction jobsite, as shown in Figure 1.2. This dissertation starts with state perception – entities' states are perceived from multimodal sensors, which becomes the input for Objective 2. Jobsite context is interpreted with working groups identified and corresponding activities recognized based on perceived states from individual entities. Finally, Objectives 1 and 2 serve as the foundation of Objective 3, where worker's trajectory is predicted by incorporating both worker movements and contextual information including relationships with neighboring entities and involved activities. This newly developed capacity in enhancing the holistic situational awareness of dynamic construction jobsites can be further leveraged to develop pro-active, context-aware control systems for struck-by prevention. In the system, the risk of potential collision can be estimated based on the predicted trajectory of construction entities, and early warnings can be provided to involved entities to avoid struck-by accidents.

1.5 Research Contributions

This novel and original research is expected to establish a roadmap towards realizing holistic situational awareness of the construction site through a data-driven approach. By integrating artificial intelligence with construction domain knowledge, the knowledge underlying heterogeneous data that reflect various aspects of construction entities is better exploited, which enables the true understanding of construction scenarios in an automatic way. Specifically, this research identified critical features that are unique in the construction domain to capture entity interactions and created a generic model to represent them numerically. By establishing the relationship between entity interaction patterns with construction working groups and group activities, this research enables the comprehension of complex jobsite context on dynamic and unstructured workspaces. This research also identified critical contextual features that will influence worker movements and innovatively incorporate contextual information into the prediction of future worker states. It has great potential to contribute not only to improved site safety performance by avoiding struck-by accidents, but also to automatic progress monitoring and control to ensure productivity, as well as to safe and efficient human-robot collaboration on future construction scenarios.

In addition, in each objective, new algorithms or methods are created to overcome the technical challenges, with specific contributions listed below.

1. This study creates a hybrid framework that leverages vision-based localization and radio-based identification for robust 3D tracking of multiple construction workers. Instead of directly fusing locations extracted from two approaches, the proposed framework strategically integrates these two methods, using vision-based tracking as the main source to obtain object trajectory and radio-based tracking as a supplementary source for reliable identity information. The newly created method significantly improves the overall accuracy for 3D tracking of multiple construction workers compared with the vision-based systems alone.
2. This study pioneers in incorporating attentional cues into the understanding of construction jobsite context and proposes a two-step long short-term memory (LSTM) approach that integrates the positional and attentional cues to identify working groups and recognize corresponding group activities. The proposed two-step process, i.e., working group identification followed by the activity recognition, allows the differentiation of group-relevant and non-relevant entities, making it capable of addressing complex group activities under general construction scenarios, where multiple entities co-exist on the job site.
3. This study creates a context-augmented deep learning model for worker trajectory. It not only considers spatial interaction between the target and neighboring entities, but also innovatively incorporates the semantic relationship between entities (i.e., whether or not within a working group) and the long-term goal of the target (i.e., the potential destination). The context-augmented method outperforms the position-based prediction with less final displacement error, especially for long-term prediction when prediction time is no less than observation time.

1.6 Research Significance

The results and findings of this dissertation will augment the holistic situational awareness of site entities in an automatic way and enable them to have a better understanding of the ongoing jobsite context and a more accurate prediction of future states to avoid potential accidents. It can be implemented as a key element in order to prevent struck-by accidents. Vision and radio systems can be used to collect data on entity states, which will be transmitted to a central server to perform analysis using algorithms developed in this dissertation, including state perception, jobsite context comprehension, and trajectory prediction. This holistic situational awareness can be leveraged to develop a proactive and context-aware control system, such as an adaptive path planning mechanism based on the predicted trajectory of jobsite entities. Then such information and guidance can be communicated to field crews in different formats through mobile devices. For instance, in addition to early warnings in sounds and vibration, for operators in equipment, one can visualize the site condition including the movements of their surrounding entities, and also see the planned trajectories for them. Workers can be provided tailored information on their nearby hazards through voice and visualization using mobile devices or augmented reality devices. By these assistive systems, jobsite entities are augmented with holistic and ubiquitous situational awareness to prevent struck-by accidents.

It is possible to extend this newly enhanced capacity of holistic situational awareness to prevent other types of accidents, such as falls and electrocutions. For instance, 291 workers fell to a lower level during construction in 2013, accounting for 35% of the total deaths in that year (OSHA, 2015). Electrocution is also among the “fatal four” of the industry that cause most of the fatalities. Such tragedies could be avoided if workers were early informed and possessed adequate situational awareness when they were close to or stepping into the hazard areas. Besides, this capability has the great potential to contribute to automatic construction progress monitoring and control. By integrating holistic situational awareness and construction plans and representations such as building information models (BIMs), one can easily tell whether construction entities are at the right place doing the right tasks with the right partners, which in turn facilitate the active control of construction operations to ensure productivity. Furthermore, given the fact that construction robots and autonomous machines have been increasingly introduced in construction projects, human workers and robots are expected to collaborate ubiquitously in the future construction jobsite. The automatic situational awareness developed in this research is promising to be adopted in human-robot collaboration on construction sites and to empower robots to adaptively adjust their behavior to effectively and efficiently collaborate with human workers.

1.7 Dissertation Organization

This dissertation is organized into five chapters and follows the "multiple publications" formats. Each of the Chapters 2, 3, and 4 has its own introduction, literature review, methodology, implementation and results, and conclusion sections. Significant portions of these chapters have been published or submitted for review and publication in peer reviewed journals. Chapter 1 introduces the background, highlights the problem statement and limitations in related studies, and discusses the research objectives, contributions, and significance.

Chapter 2 presents a hybrid framework for 3D tracking of multiple construction workers by combining vision-based localization and radio-based identification. ***This work was previously published by ASCE Journal of Computing in Civil Engineering, 2020, Jiannan Cai and Hubo Cai, "Robust Hybrid Approach of Vision-Based Tracking and Radio-Based Identification and Localization for 3D Tracking of Multiple Construction Workers" (J. Cai & Cai, 2020). This is the pre-production version, with permission from American Society of Civil Engineers (ASCE)." This material may be found at [DOI = 10.1061/(ASCE)CP.1943-5487.0000901]. Table titles and figure captions have been modified to maintain the form of the dissertation.***

Chapter 3 presents a two-step LSTM method for identifying construction working groups and corresponding activities by integrating both positional and attentional cues. ***This work was previously published in Automation in Construction (J. Cai et al., 2019). This chapter is re-printed with permission from Vol 104, Jiannan Cai, Yuxi Zhang, and Hubo Cai, "Two-step long short-term memory method for identifying construction activities through positional and attentional cues", 102886, Copyright Elsevier (2019). Table titles and figure captions have been modified to maintain the form of the dissertation.***

Chapter 4 presents a deep learning model augmented by construction contextual information for worker trajectory prediction. ***This work is under review in Advanced Engineering Informatics, 2020, Jiannan Cai, Yuxi Zhang, Liu Yang, Hubo Cai, and Shuai Li. "A Context-Augmented Deep Learning Approach for Worker Trajectory Prediction on Unstructured and Dynamic Construction Sites". Table titles and figure captions have been modified to maintain the form of the dissertation.***

Finally, Chapter 5 concludes the dissertation by summarizing the findings and discussing the directions and visions for future work.

2. ROBUST HYBRID APPROACH OF VISION-BASED TRACKING AND RADIO-BASED IDENTIFICATION AND LOCALIZATION FOR 3D TRACKING OF MULTIPLE CONSTRUCTION WORKERS

In this chapter, a hybrid framework that fuses results obtained from vision-based tracking and radio-based identification and localization is proposed for 3D tracking of multiple construction workers. The proposed method treats vision-based tracking as the main source to extract the object trajectory. Radio-based identification and localization results are used as a supplementary source to augment anonymous visual tracks with identity information and correct errors (e.g., false positives) in vision-based object detection, resulting in ID-linked 3D trajectories. In addition, a searching algorithm is introduced to recover possible missed detections in one camera view from the corresponding observations in the other view by applying a sliding window to search for regions with the most similar appearance along the epipolar line. Two indoor experiments were conducted to validate the newly created method, where a stereo vision system and a radio-based (i.e., BLE beacon) system were used to obtain 3D trajectories of workers in the 3D world coordinate system. The results show that the new approach for fusing vision- and radio-based results increases the overall accuracy from 88% and 87% to 95% and 90%, compared to using the vision-based approach alone. The integration of radio-based identification is much more robust than using the vision system alone as it allows the recovery of the same entity ID after the trajectory is fragmented and results in fewer fragmentations that last longer than 0.2s.

This work was previously published by ASCE Journal of Computing in Civil Engineering, 2020, Jiannan Cai and Hubo Cai, "*Robust Hybrid Approach of Vision-Based Tracking and Radio-Based Identification and Localization for 3D Tracking of Multiple Construction Workers*" (J. Cai & Cai, 2020). This is the pre-production version, with permission from American Society of Civil Engineers (ASCE)." This material may be found at [DOI = 10.1061/(ASCE)CP.1943-5487.0000901]. Table titles and figure captions have been modified to maintain the form of the dissertation.

2.1 Introduction

Information regarding construction entity identity and real-time location reveals where specific construction resources are at any given time and thus, is a critical prerequisite to the context-aware jobsite safety management. It enables the identification of unauthorized personnel in restricted hazardous zones (Q. Fang et al., 2018a) and the communication of personalized and dynamic information (D. Liu et al., 2018, 2016; Papaioannou et al., 2017), e.g., informing workers of the type and location of hazards around them as they move on the jobsite. Reliable and continuous location information facilitates the analysis of spatial-temporal relationships among entities, the generation of dynamic workspaces (X. Luo et al., 2019), and the identification of abnormities such as the proximity to potential hazard (Jeelani et al., 2019) and collisions between workers and heavy equipment (Liang et al., 2019). It is also an important component of positional cues that can be used to exploit the interaction patterns among entities and recognize the corresponding activities on the jobsite. Therefore, there is a critical need to achieve continuous and robust tracking of construction entities with reliable identity information.

Existing studies on automatically tracking real-time locations of workers and equipment on the construction site fall into two major categories: sensor-based approach and vision-based approach. In the sensor-based approach, the global positioning system (GPS) is widely used for outdoor tracking and is currently embedded in most construction equipment. However, GPS is not reliable in the indoor environment or in crowded urban areas (Chen et al., 2018; Su et al., 2014). Radio-based technologies, such as UWB and RFID, are suitable for both indoor and outdoor tracking (Li et al., 2016) and they provide reliable identity information. However, radio-based technologies require attaching tags on objects and are complicated for deployment. Besides, their localization accuracy in complex and dynamic spaces is low due to the multipath error. Recently, Bluetooth low energy (BLE) technology has emerged as an alternative radio-based localization method because of its low energy consumption and low cost—inexpensive beacons are attached at fixed locations and smartphones carried by workers are leveraged as signal receivers (J. W. Park et al., 2017). Nevertheless, it still suffers from relatively large localization errors.

With the advancement in computer vision and the availability of construction surveillance videos, vision-based tracking has gained increasing attention in safety and productivity management tasks. Earlier studies (M. W. Park & Brilakis, 2016; Zhu et al., 2016b, 2017) focused on extracting two dimensional (2D) pixel coordinates of construction entities by integrating object detection models and tracking algorithms. A recent focus is on acquiring object trajectories in three dimensional (3D) world coordinates. Konstantinou and Brilakis (2018) and Lee and Park (2019) matched entities across multiple camera views to obtain 3D coordinate, while Yong et al. (2019) created a tracking method using online learning from an RGB-D camera. Vision-based tracking can achieve high localization accuracy, but it heavily depends on the performance of object detection which is significantly affected by environmental conditions such as occlusion and illumination. Furthermore, most vision-based tracking is anonymous and, therefore, subject to identity (ID) switch and fragmentation errors when multiple workers are in close proximity or occluded. As a result, long-term and robust vision-based tracking of multiple workers remains a challenge.

To overcome the above challenges and achieve more robust 3D tracking of multiple construction workers, this study proposes a hybrid framework that integrates vision-based tracking and radio-based identification and localization. Instead of directly fusing locations extracted from these two approaches, the newly created method treats vision-based tracking as the main source to obtain the object trajectory. In addition to 3D location, stereo vision provides complementary views to recover possible missed detections due to occlusions in individual views. Radio-based identification and localization is used as a supplementary source to provide reliable identity information to exclude false detections when the vision-based approach fails to correctly detect objects.

2.2 Review of Related Studies

Related studies in vision-based, radio-based, and multisource fusion-based tracking have been reviewed and are summarized as follows.

2.2.1 Vision-based tracking

A few methods have been developed to track construction workers and equipment through vision-based approaches due to its ease of deployment, low cost, and non-intrusiveness. Vision-based tracking can be grouped into two categories: 2D tracking that obtains object trajectory in terms of 2D coordinates on an image plane and 3D tracking that extracts object trajectory in the 3D world coordinate system.

In 2D tracking, a monocular camera is used, and the target objects are represented by 2D pixel coordinates in the image plane. Yang et al. (2010) proposed a tracking scheme based on machine learning which used a pre-trained appearance model to detect construction workers and a parameterized appearance feature function to uniquely estimate each worker such that multiple workers are tracked at the same time. Zhu et al. (2016b) presented a method to track mobile entities on a construction site using particle filters. However, it only tracks one object and requires manual initialization. Park and Brilakis (2016) and Zhu et al. (2017) developed hybrid methods that integrated detection and tracking processes to maintain both high recall and precision in tracking multiple workers and equipment. More recently, Roberts and Golparvar-Fard (2019) proposed a deep learning-based method for object detection and tracking based on Convolutional Neural Networks (CNNs). A common drawback for 2D vision-based tracking is that the output is limited to the 2D pixel coordinate system and the depth information is lost. However, in construction safety management, 3D location in the world coordinate system is needed to effectively avoid collisions.

3D vision-based tracking approaches use two or more cameras to reconstruct the 3D trajectories. Earlier works (Brilakis et al., 2011; M.-W. Park et al., 2011; Yuan et al., 2016; Zhu et al., 2016a) first obtained object locations on individual camera views and then recovered the 3D location through triangulation using the calibration information of the stereo vision system. One limitation of these studies is that the same entity across two camera views is matched based on epipolar geometry alone, leading to errors when multiple objects are along the same epipolar line. To overcome this limitation, recent studies (Konstantinou & Brilakis, 2018; Y. J. Lee & Park, 2019; B. Zhang et al., 2018) have developed more sophisticated models that leverage multiple cues such as epipolar geometry, appearance model, moving direction, motion patterns, and SIFT point features, to match entities across camera views for 3D tracking of construction resources. However, these approaches rely on reliable 2D vision-based tracking that eventually depends on the performance of object detection on individual camera views. Once the vision-based detection fails, there is no supplementary information to reconstruct the accurate 3D trajectories.

Recently, monocular cameras have been used to recover 3D location information. For instance, the simultaneous localization and mapping (SLAM) technique has been used to localize mobile targets while mapping the environment in real-time (Asadi et al., 2019; Jeelani et al., 2019). However, it typically requires mounting cameras on each mobile target with initial location and is subject to drift errors. In addition, Son et al., (2019b) developed a 2D vision-based system to estimate the 3D distance between the heavy machine and construction worker based on perspective transformation. Yan et al., (2019) created a novel method to estimate the 3D distance between workers based on view-invariant relative 3D joint point (R3DJP) and joint angle features (H. Zhang et al., 2018) and trained classifiers using a single 2D camera. However, these studies are mainly focused on proximity analysis instead of continuous tracking of multiple workers.

2.2.2 Radio-based tracking

Radio-based technologies such as RFID and BLE represent another type of sensors for tracking construction resources. The tracking of construction resources is mainly based on the radio signals transmitting between tags and receivers. Cai et al. (2014) proposed a novel algorithm that combines the boundary condition method and trilateration concept to estimate 3D locations of construction resources using RFID and achieved an average accuracy of 2.48m. Su et al. (2014) proposed an enhanced boundary condition algorithm that incorporates the RFID tag-reader angle and the reader geometric configuration and increased the accuracy to 1.54m. Costin and Teizer (2015) leveraged the contextual information in BIM to increase the accuracy of indoor location obtained based on multilateralization techniques using passive RFID and achieved an accuracy of 1.66m. Park et al. (2016) tracked worker location using a set of static BLE beacons distributed in known locations. Topak et al. (2018) assessed the feasibility of using BLE technology for indoor localization and achieved an accuracy of 70% at a precision of 1.8m using a fingerprinting method. Zhao et al. (2019) applied BLE technology for real-time tracking of workers and evaluated the accuracy influence from the deployment of BLE beacons, but their approach can only estimate worker presence within a rough area without exact 3D locations.

The main drawback of the radio-based approach is the relatively low localization accuracy (over 1m for RFID and BLE technology) compared with the vision-based approach (submeter level), which hinders the adoption of radio systems alone for applications that require high localization accuracy such as site safety management. However, as radio-based technologies are combined with tags or mobile devices on the targets, they provide perfect identity information through unique IDs which is critical for context-aware site safety management.

2.2.3 Multisource fusion-based tracking

It is well recognized that a data fusion approach that integrates heterogeneous data obtained from different sensors can improve the accuracy and confidence of tracking results. Table 2.1 lists related studies on object tracking using the multisource fusion approach, with fusion method and limitation discussed for each study. In general, most studies that include vision-based tracking are in 2D and cannot provide reliable 3D location information, and for those tracking in 3D, the radio-based approach is mainly used for localization, resulting in relatively large errors. Moreover, most existing studies treat different data sources equally and directly rely on noisy radio measurements (received signal strength or estimated location) for matching without effective methods to compensate for the possible errors in the radio-based approach in order to improve robustness.

Table 2.1 Related studies on object tracking via multisource fusion approach

Study	**Application Domain**	**Sensor**	**Fusion Method**	**Limitation**
Mohebbi et al. (2017)	3D indoor multi-object tracking (MOT) (general)	Passive infrared motion sensors + radio sensors (i.e., BLE beacons)	Each type of sensor data is processed separately and generates sensor-specific confidence maps, which are then merged into a single set of maps for location estimation for each target	The average localization accuracy is only 1.8 m due to directly leveraging locations estimated from two sensors with relatively low accuracy.
Jung et al. (2010)	2D indoor MOT (general)	Single 2D camera + accelerometer sensors	Matches anonymous visual tracks with entity ID obtained from wearable sensors by comparing velocity from the vision-based approach and accelerometers from the sensor via a correlation metric	The method may fail when objects have similar velocities, which are very common on construction sites. It cannot provide 3D location information.
Mandelic et al., (2013)	2D indoor MOT (general)	Multiple 2D cameras + radio sensors (i.e., UWB)	Matches anonymous visual tracks with radio-based identifications by minimizing the overall distance	Only distance is considered in the matching procedure, subject to errors and fluctuations since radio-based localization is time-step independent and does not consider movement continuity across time steps. The vision-based localization is based on occupancy map, which only considers ground plane and cannot estimate 3D location.

Table 2.1 continued

Yu and Ganz (2010)	2D outdoor MOT (general)	Single 2D camera + radio sensors (i.e., RFID)	Matches generated visual tracklets with radio measurement by calculating the likelihood of received signal strength given the distance between tracklet and RFID readers	The fusion of vision and radio is based on generated tracklets from a window of observations rather than a frame-by-frame manner.
Papaioannou et al., (2015)	2D indoor MOT (domain)	single camera + radio sensors (i.e., Wi-Fi)	Matches generated visual tracklets with footprint of received radio signals to augment anonymous tracks with ID	
Chen et al. (2018)	3D indoor localization (construction)	Inertial measurement unit (IMU) and radio sensors (i.e., BLE beacons)	Combines indoor localization results obtained from BLE beacons and IMU using the Kalman filter as the fusion core	The main focus is to improve the localization accuracy of a single entity and not applicable to MOT
Papaioannou et al., (2017)	2D outdoor MOT (construction)	Single 2D camera + radio sensors (i.e., Wi-Fi and BLE beacons) + IMU	Mathes visual tracks with radio-based identification based on received signal strength (i.e., for specific visual detection, its received radio signal should match the predicted radio measurements at the same location), and uses IMU to update the state of visual tracks.	Once a visual track is assigned to an entity ID, it can only be updated with the measurement of the same ID, which is sensitive to errors from radio measurements, i.e., if the track is assigned to an incorrect ID, it cannot be corrected. It cannot provide reliable 3D location using a single camera approach.

In fact, vision-based and radio-based technologies both have advantages and limitations and should be strategically integrated into mutual complementation to further improve the accuracy and reliability of tracking for both indoor and outdoor applications. Vision-based localization is accurate, but the tracking performance significantly decreases when failing in detecting the target. On the other hand, radio-based technology is less accurate but reliable in object detection and identification. This study aims to solve the above limitations by integrating radio-based identification with vision-based localization such that reliable identification provides supplementary information when the vision-based approach fails to correctly detect the targets without decreasing the localization accuracy.

2.3 Problem Formulation

In this study, we track multiple workers using a stereo camera system and radio-based system. Two stationary cameras are located in the environment with an overlapping field of view (FOV).

Each worker carries a mobile device (i.e., smartphone) to receive radio signals transmitted by a set of BLE beacons attached to fixed locations, as illustrated in Figure 2.1.

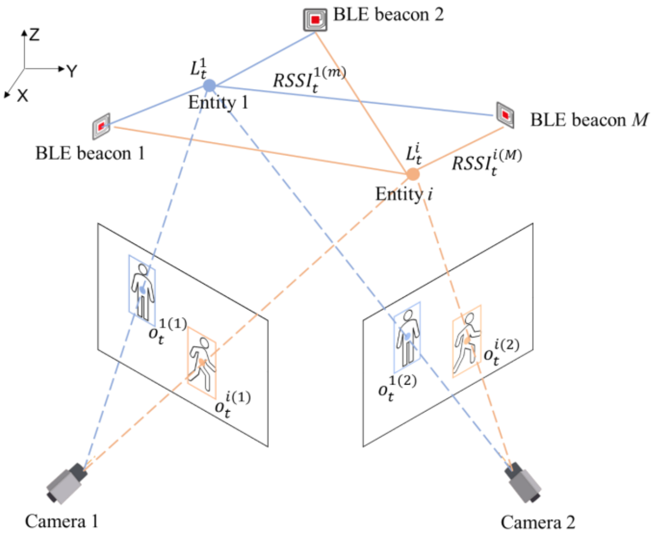


Figure 2.1 Vision-radio system set up

At each time step *t*, the system receives a collection of camera observations, denoted as $O_t^{(k)} = \{o_t^{1(k)}, o_t^{2(k)}, ..., o_t^{i(k)}, ...\}$, where $o_t^{i(k)}$ represents the bounding box of the *i*-th detected object in the *k*-th camera view (*k* = 1,2). Note that as the object detection is not perfect, it is possible that not all observations are real workers (referred to as false positive). Similarly, there may be some workers not detected (referred to as false negatives).

Meanwhile, mobile devices receive signals from BLE beacons, the measurement of which is denoted as $R_t = \{r_t^1, r_t^2, ..., r_t^j, ...\}$, where $r_t^j = (RSSI_t^{j(1)}, RSSI_t^{j(2)}, ..., RSSI_t^{j(M)})$ represents the received signal strength of the *j*-th device (i.e., *j*-th worker) from each beacon, and *M* is the number of beacons. Given R_t, the 3D world coordinates of workers can be estimated using radio-based localization algorithms, resulting in $L_t^{radio} = \{l_t^1, l_t^2, ..., l_t^j, ...\}$, where $l_t^j = (x_t^{j(radio)}, y_t^{j(radio)}, z_t^{j(radio)}, j)$ is the ID-linked coordinates in 3D world space for the *j*-th worker. For the radio system, although the localization accuracy is typically lower than the vision-based system, the ID linked with each measurement is very reliable, which is the motivation of this study in fusing vision-based localization and radio-based identification to track multiple workers.

As a result, the problem is as follows: given time-series anonymous camera detections in both camera views $O_{1:t}^{(k)}$ (k=1, 2) and ID-linked locations from the radio-based system $L_{1:t}^{radio}$ (the subscript, 1:*t*, indicates the time-series coordinates from time step 1 to *t*), estimate the trajectories ($L_{1:t}$) of all workers in the 3D world coordinate system.

2.4 Methodology

Figure 2.2 illustrates the overall framework for 3D tracking of multiple construction workers. The main processes with novel contributions are highlighted with the gray background color. The proposed framework consists of two modules—2D tracking by detection and identification and 3D tracking by entity matching and identification. The first module outputs ID-linked 2D tracks on each camera view by matching 2D vision-based tracking with radio-based localization (projected onto the image plane), where a 2D track stores the time-series pixel coordinates for the same object. The second module takes the output of the 2D tracking module and outputs the ID-linked 3D tracks by matching vision-based location obtained from entity matching and triangulation with radio-based localization, where a 3D track stores the time-series world coordinates for the same object. The 2D pixel coordinates of workers detected frame-by-frame from both camera views, and the 3D world coordinates with entity IDs obtained using radio-based localization serve as inputs. The stereo camera system is calibrated at the beginning. Note that 2D vision-based object detection, radio-based localization, and camera calibration are outside the scope of this study, and are performed using existing methods, which will be briefly introduced in this section for clarity.

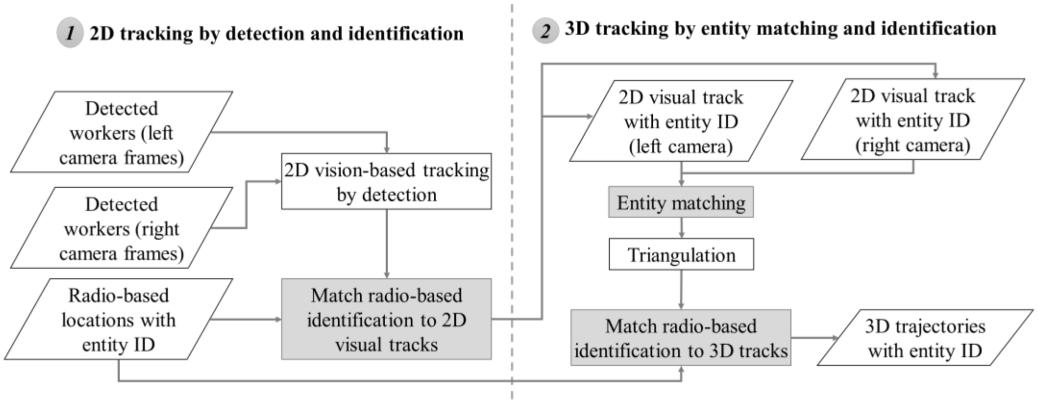

1 2D tracking by detection and identification
2 3D tracking by entity matching and identification
Detected workers (left camera frames)
Detected workers (right camera frames)
Radio-based locations with entity ID
2D vision-based tracking by detection
Match radio-based identification to 2D visual tracks
2D visual track with entity ID (left camera)
2D visual track with entity ID (right camera)
Entity matching
Triangulation
Match radio-based identification to 3D tracks
3D trajectories with entity ID

Figure 2.2 Hybrid framework for 3D tracking of multiple workers

2.4.1 System inputs

In this study, the tracking process is initialized and updated through frame-by-frame worker detection. The 2D pixel coordinates of detected workers, the 3D world coordinates for each worker obtained from radio-based localization, and the calibrated stereo camera system parameters are independent of the proposed tracking framework, and thus, are treated as system inputs.

2D vision-based worker detection

The 2D vision-based worker detection takes a 2D image as input and outputs a set of bounding boxes of workers, the central coordinates of which represent the 2D pixel coordinates of workers. This study applied a state-of-the-art object detection framework—faster R-CNN (Ren et al., 2017), and used a pre-trained ResNet-50 network (He et al., 2016) on a COCO dataset for human detection. Note that the proposed tracking framework is still valid for other object detection methods (e.g., Memarzadeh et al., 2013; M. W. Park & Brilakis, 2012).

Radio-based localization

This study uses BLE beacons for radio-based localization. The working principle is that a set of BLE beacons are attached to fixed locations with known coordinates, transmitting radio signals, and smartphones are carried on moving workers, receiving signals from these beacons. The location of the moving worker (i.e., the smartphone) is estimated using known locations of beacons and the distances between the smartphone and different beacons estimated based on the received signal strength indication (RSSI).

Distance estimation

In general, the received strength of a radio signal attenuates as the distance between transmitter (i.e., BLE beacons) and receiver (i.e., smartphones) increases. Their relationship is typically modeled using a log-distance path loss model (Zhuang et al., 2016), formulated in Equation 2.1

$$RSSI(d) = RSSI(d_0) - 10\gamma \log_{10}(\frac{d}{d_0}) + X_\sigma \tag{2.1}$$

where $RSSI(d)$ represents the RSSI at the distance d between transmitter and receiver; $RSSI(d_0)$ represents the RSSI at the reference distance d_0, where $d_0 = 1\text{m}$ is commonly used; γ is the path-loss index; and X_σ is Gaussian random noise with zero mean. In Equation 2.1, $RSSI(d)$ is the observation, while $RSSI(d_0)$ and γ are system parameters that may vary with environmental conditions and the type and configuration of transmitters and receivers, and can be calibrated using a series of known distances and RSSIs. As a result, the distance d can be derived from Equation 2.1.

Location estimation

This study adopted a weighted path loss algorithm developed by Zou et al. (2014) to estimate the target location given its distance between different beacons. The target coordinates are calculated as the weighted average of known coordinates of beacons, with the weight of each beacon inversely proportional to the estimated distance between the target and the beacon, formulated in Equation 2.2 and 2.3

$$w_i = \frac{1}{d_i} / \left(\sum_{i=1}^{N} \frac{1}{d_i} \right) \tag{2.2}$$

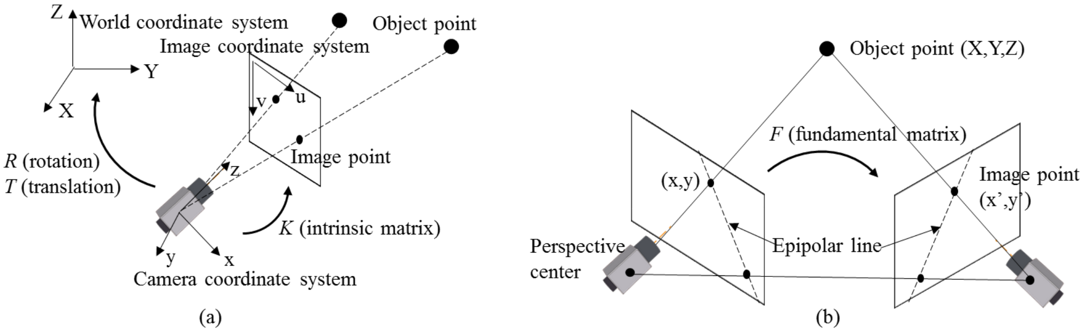
$$\left(x_{\text{target}}, y_{\text{target}} \right) = \sum_{i=1}^{N} w_i (x_i, y_i) \tag{2.3}$$

where d_i is the distance between the target and the i-th beacon obtained through Equation 2.1, (x_i, y_i) are known coordinates of the i-th beacon, and $\left(x_{\text{target}}, y_{\text{target}} \right)$ are estimated coordinates of the target. As the radio-based localization is independent of the proposed tracking framework, other localization algorithms (e.g., Thaljaoui et al., 2015; Zhuang et al., 2016) can also be used.

It is noted that this study only considers x and y coordinates when matching location obtained from radio-based and vision-based approaches and assumes all workers are on the same ground level since the absolute z-value in radio-based localization is less accurate than x and y values and incorporating it will decrease the accuracy of the matching process. This simplification is reasonable because (1) on construction sites, we usually track entities on the same vertical level to identify potential collisions; (2) in most radio-based localization studies (e.g., Costin & Teizer, 2015; Topak et al., 2018; Zhuang et al., 2016), only x and y coordinates are estimated, and z is described by the corresponding floor/ground level; (3) for applications that involve multiple floor levels, the floor level where the target is located can be estimated via the gateways installed on each floor level (Jianyu Zhao et al., 2019).

Stereo camera calibration

A stereo camera system is adopted to fuse with the radio-based system and determine the 3D location of the object, which needs to be carefully calibrated so that the image coordinate system can be correlated to the world coordinate system. Two cameras are first calibrated separately, which outputs three matrices (and vectors): (1) intrinsic matrix (K) that defines the camera coordinate system, (2) rotation matrix (R), and (3) translation vector (T) that define the position and orientation of the camera with respect to the world coordinate systems, as illustrated in Figure 2.3(a).



(a) (b)

Figure 2.3 Camera calibration: (a) single camera calibration, (b) epipolar geometry of stereo camera

The calibration is based on the pinhole camera model (Z. Zhang, 2000), represented as $w[x\ y\ 1] = [X\ \ Y\ \ Z\ \ 1]\begin{bmatrix} R \\ T \end{bmatrix} K$, where $[x\ y\ 1]$ is the homogeneous image point, $[X\ \ Y\ \ Z\ \ 1]$ is the homogeneous 3D object point, w is a scale factor, K is the intrinsic matrix, and $\begin{bmatrix} R \\ T \end{bmatrix}$ is the extrinsic rotation and translation with respect to a 3D coordinate system. In this study, the Matlab computer vision toolbox (MathWorks, 2019) is first used to solve for K, R, and T. Note that in this process, R and T refer to a local 3D coordinate system defined by a chessboard that is used in camera calibration, different from the world coordinate system. To estimate the camera pose (R and T)

with respect to the world coordinate system, the same model described above is used except that $\left[X\ Y\ Z\ 1\right]$, *R*, and *T* are all in the world coordinate system and *K* is known matrix. The model is solved using the perspective-three-point algorithm described by Gao et al. (2003). After calibration of individual cameras, the image (or pixel) coordinates of an object point with known world coordinates can be acquired, which are used to match 2D vision-based location and 3D radio-based location on the 2D image plane.

The stereo camera is calibrated to establish the epipolar geometry, which outputs a fundamental matrix *F* that correlates corresponding points in stereo images by satisfying $X^{'T}FX = 0$, where X is the homogenous image coordinates for points in one image, and $X^{'}$ is the homogenous image coordinates for corresponding points in the other image, as shown in Figure 2.3(b). Given 2D image coordinates in both camera views for the same object point, its 3D world coordinates can be recovered through triangulation.

To obtain the extrinsic parameters (*R* and *T*) of a camera, theoretically, a minimum of three non-collinear ground control points (GCPs) with known world coordinates are needed. To compute the fundamental matrix in stereo camera calibration, typically eight corresponding points across camera views are used based on the normalized eight-point algorithm (Hartley & Zisserman, 2003). Increasing the number of GCPs will improve the calibration accuracy due to the increased redundancy. For instance, Lee and Park (2019) used 30 GCPs for a 30 × 35m construction site.

2.4.2 Module 1 – 2D tracking by detection and identification

In Module 1, 2D visual tracks for individual camera views are first obtained by associating vision-based detections across frames, resulting in anonymous visual tracks. Then, the visual tracks are matched with radio-based locations projected onto the 2D image plane, resulting in ID-linked 2D visual tracks for each camera view.

2D vision-based tracking by detection

In this study, the 2D tracks in each camera view are initialized with detection results, i.e., $Tr_1^{(k)} = O_1^{(k)}$. Besides, each track is associated with a Kalman filter using the constant velocity model to predict its possible location in the next time step. At each time step *t* (*t*>1), the potential locations of the objects are predicted from the tracks at the previous step as well as detected by the object detector at the current time. Theoretically, the predicted location and detected location for the same target should be very close considering the continuity of movement.

The current track for each entity can be updated by associating detections with predicted tracks, which is formulated as a linear assignment problem, illustrated as follows,

$$\begin{aligned}
&\min \sum_{\substack{i=1,2,\ldots,I \\ j=1,2,\ldots,J}} c_{ij} m_{ij} \\
&\text{subject to } \sum_{j=1,2,\ldots,J} m_{ij} = 1 \text{ for } i = 1,2,\ldots,I, \\
&\qquad\qquad \sum_{i=1,2,\ldots,I} m_{ij} = 1 \text{ for } j = 1,2,\ldots,J, \\
&\qquad\qquad\quad m_{ij} = 0,1 \text{ for } i = 1,2,\ldots,I;\ j = 1,2,\ldots,J
\end{aligned}$$

where m_{ij} refers to the assignment between track *i* and detection *j*. If $m_{ij} = 1$, detection *j* is assigned to track *i*; if $m_{ij} = 0$, detection *j* is not assigned to track *i*. c_{ij} refers to the cost of assigning detection *j* to track *i*.

The objective is to find the optimal assignment between tracks and detections with the minimum total cost, subject to the constraints that each track can only be assigned with at most one detection and each detection can only be assigned to at most one track. In this study, given a pair of predicted track and detection, i.e., $(tr_{t(predict)}^{i(k)}, o_t^{j(k)})$, their assignment cost is the combination of their distance on the image plane and the dissimilarity of their appearance model (color histogram of the target bounding box in this study), denoted as $c_{ij} = dis\left(tr_{t(predict)}^{i(k)}, o_t^{j(k)}\right) + w_{color} dis_{color}\left(tr_{t(predict)}^{i(k)}, o_t^{j(k)}\right)$. Mahalanobis distance (Mahalanobis, 1936) is used to calculate the distance between the detected location and predicted location which incorporates the uncertainty of the Kalman filter. The Bhattacharyya distance (Battacharyya, 1943) is used to compute the dissimilarity of two histograms. w_{color} indicates the weight of two distances and is set experimentally ($w_{color} = 1.4$ in this study). The assignment problem is solved using the Munkres algorithm (Munkres, 1957).

After the above detection-to-track assignment, the locations of 2D tracks are updated by assigned detected locations with its associated Kalman filter updated by the new observations. For unassigned detections, new tracks are created at the detected locations. For unassigned tracks, a conventional approach is to reserve the tracks for a certain number of frames by updating the location using the Kalman filter with its status changed to “invisible”. After a track is invisible for a long period, it is terminated. This approach works well for linear movement but is not reliable when objects make sudden changes in direction if the termination threshold is relatively large. In this study, the unassigned tracks are updated with the predicted location for only a small number of frames (e.g., 3 in this study), then the location remains unchanged. The rationale is that in the proposed method, 2D tracks are linked with device IDs (the linking procedure will be discussed in the next section), increasing the confidence that the tracks refer to real targets present in the scene instead of being false positives. Therefore, it is reasonable to maintain their locations for a relatively long time (20 frames in this study) until their detections are recovered. Alternatively, if the radio-based system is accurate in localization, the locations of these tracks can also be updated using the radio-based localization result alone (Mandeljc et al., 2013).

Match radio-based identifications to 2D tracks on image plane

As vision-based 2D tracks are mainly updated by the frame-by-frame detections, possible false positives may cause the incorrect link between the tracks and detections, which will propagate over time and eventually cause the tracking to fail. Such errors can be corrected by introducing ID-linked radio localization results. The rationale is that although the radio-based localization accuracy of the individual target might be relatively low, the overall spatial configuration of multiple targets should be similar to that obtained in visual-based detection since they are localized in the same environment at the same time (Mandeljc et al., 2013). Therefore, it is reasonable to assign unique IDs to 2D tracks based on the overall spatial configuration—the optimal combination of radio-based and vision-based locations are determined by minimizing the total distance (or cost) instead of minimizing that for a specific pair, which is formulated as a linear assignment problem. In the matching process, each radio-based location can only be matched to at most one vision-based location, and vice versa, as illustrated in Figure 2.4. As a result, for the cases where the vision-based approach incorrectly detects workers (i.e., false positives) or fails to detect workers (i.e., false negatives), the matching process results in unassigned visual location (Figure 2.4(b)) or unassigned radio location (Figure 2.4(c)). The optimal assignment is determined via Munkres algorithm (Munkres, 1957).

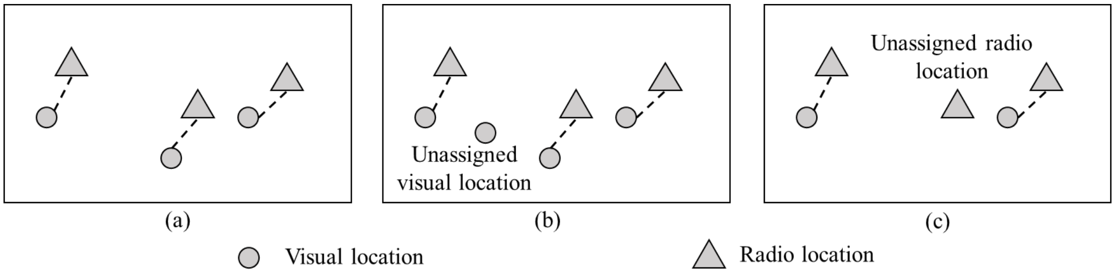


Figure 2.4 Matching between radio-based and vision-based locations: (a) one-to-one correspondence, (b) false positive of visual approach, (c) false negative of visual approach.

In this study, the assignment cost between radio-based identification and 2D tracks is computed as $c_{ij} = dis\left(tr_t^{i(k)}, l_t^{j(k)}\right) + w_{id} Penalty_t^{id} + w_{pred} Penalty_t^{pred}$, consisting of three components: (1) $dis\left(tr_t^{i(k)}, l_t^{j(k)}\right)$ is the pixel distances between tracks and radio-based localization results. The radio-based localization results are projected onto the 2D image plane using the calibrated camera parameters. And the mid-bottom point of the bounding box is used to calculate the distance as it is assumed to be on the same ground plane with radio-based locations. (2) $w_{id} Penalty_t^{id}$ is the penalty on assigning a different ID from the previous ID to the 2D tracks. At time step *t*, if the assigned ID is different from that assigned in time *t-1*, $Penalty_t^{id} = 1$, otherwise, $Penalty_t^{id} = 0$. The magnitude of the penalty is reflected by the weight w_{id} that is determined experimentally ($w_{id} = 1$ in this study). Inherently, the track tends to maintain the same ID given the continuity of the movement. As the radio-based localization may fluctuate due to the dynamic environment, it is not reliable to use the distance alone to determine the assignment. Therefore, this penalty is introduced to ensure stable and continuous tracks and mitigate the negative impact of the inaccurate radio-based localization. (3) $w_{pred} Penalty_t^{pred}$ is the penalty for assigning an ID to invisible tracks. At time step *t*, if the ID is assigned to an invisible track, $Penalty_t^{pred} = 1$, otherwise, $Penalty_t^{pred} = 0$. And the magnitude is reflected by the weight, w_{pred}, which is determined experimentally ($w_{pred} = 1$ in this study). As discussed in the previous section, if a track is not associated with detection results, it will be reserved for a period of time with status changed to invisible. A track that has corresponding detections is more reliable to be the real track compared to the invisible one. Therefore, when assigning IDs, a penalty is introduced to the invisible tracks.

After the assignment, the track IDs are updated with the assigned device (i.e., entity) IDs. The ID for an unassigned track is denoted as “0”. A 2D track is terminated when its ID equals 0, indicating that it does not refer to a real target. In addition, a track is also terminated if it is continuously invisible for too long (20 frames in this study), even if it is assigned with a device ID. In this way, the false positives in 2D visual tracking can be effectively reduced. For the unassigned radio-based location, it will be kept and addressed in Module 2. The rationale is that since the localization accuracy of the radio-based approach is not high, projecting the 3D location onto the 2D image plane will increase the error. Therefore, it is not reliable to correct the false negative in 2D visual tracking (i.e., unassigned radio-based location) directly using the projected radio-based location. Instead, the false negatives will be addressed in Module 2 by (1) recovering the false negative using the supplementary camera view, and (2) creating new tracks for the unassigned 3D location after matching vision and radio results in 3D.

2.4.3 Module 2 – 3D tracking by entity matching and identification

Module 1 outputs 2D visual tracks associated with entity IDs for both camera views, serving as the inputs of Module 2. In Module 2, 2D visual tracks in two camera views that correspond to the same entity are first matched in order to extract 3D vision-based location. Then, the vision-based location is matched with radio-based identification in the 3D world coordinate system, resulting in ID-linked 3D location. It is noted that the entity IDs obtained in Module 1 are not directly used for entity matching due to the possible error when projecting 3D radio-based location onto the 2D image plane. The main purpose of “Match radio-based identification to 2D tracks” in Module 1 is to exclude the false positives in visual tracking and provide more reliable inputs for Module 2. As a result, the final 3D location and corresponding ID are determined in Module 2.

Entity matching across camera views

To obtain the 3D locations of moving workers, the 2D tracks that correspond to the same person need to be matched across camera views. Previous studies (Konstantinou & Brilakis, 2018; Y. J. Lee & Park, 2019) have created methods for entity matching based on multiple cues, such as epipolar geometry, movement patterns, and appearance models of two entities. Despite the achievement, they match the individual entity using the best candidate according to the criteria rather than achieving a globally optimal solution. Moreover, the previous methods focus on matching entities that have already been detected without explicitly dealing with the case when the entity is occluded and undetected in one view but successfully detected in the other view.

To overcome these limitations, this study finds the global optimal solution for the entity matching through linear assignment, where the assignment cost is the combination of distance between the bounding box center and the epipolar line and the dissimilarity between the color histograms of two bounding boxes. Moreover, this study proposes a searching procedure to recover the missed detection in one view based on the detection in the other view, illustrated in Figure 2.5. In Figure 2.5, green bounding boxes refer to the matched entities in two camera views, and the red bounding box refers to the entity that is detected in one view but missed in the other view. For the unmatched entities (i.e., the red rectangle Figure 2.5(a)), a sliding window is applied along its corresponding epipolar line on the other camera view to find the area which has the most similar appearance to the target entity. The size of the sliding window is set to be equal to the bounding box of the target entity. By doing this, the possible error caused by the false negative in one camera view can be corrected. Then the 3D world coordinates of all matched entities are estimated via triangulation.

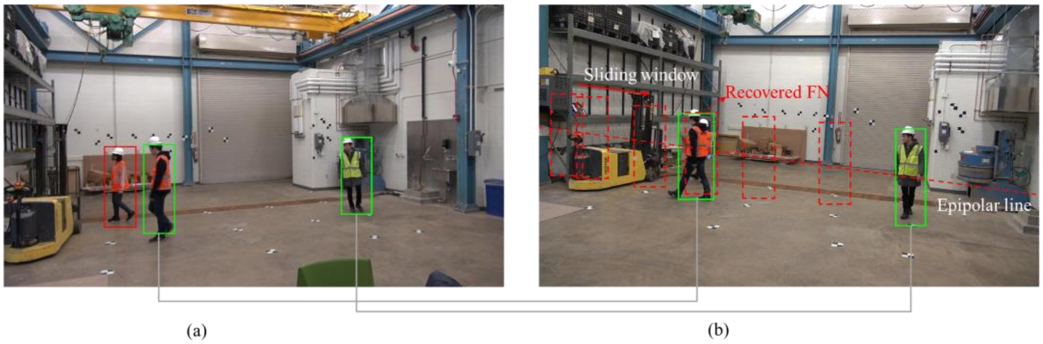


Figure 2.5 Search for missed detections across camera views: (a) left camera view, (b) right camera view.

Match radio-based identifications to 3D tracks

In each time step *t*, the obtained 3D vision-based locations are matched with radio-based locations by minimizing the total distance, resulting in 3D locations augmented with device IDs. Due to the uniqueness of the device ID, it is used to update the 3D tracks, i.e., the location of the

3D track is updated using the location with the same ID. For ID-linked 3D locations that do not have a corresponding 3D track, a new one will be created at current locations. For tracks that do not have corresponding ID-linked locations, they are reserved for a period of time before being terminated. The resulting time-sequential ID-linked 3D tracks are the output of the system. Each 3D track is associated with two 2D tracks that are assigned with the same ID and are further used to determine the penalty on assigning different IDs at the next time step. In this way, the proposed method works as a recursive system that at each time step, the tracks are updated not only based on the current vision-based detections and radio-based identifications but also considering the assigned ID transferred from the previous time step to avoid ID switches or losses due to the inaccurate detection and to ensure the robustness of the tracking process.

2.5 Implementation and Results

The proposed tracking framework is implemented in two indoor experiments. The tracking performance is compared to those obtained using other three tracking methods to demonstrate the advantages of the proposed method. The experiments and evaluation metrics are described, and the results are analyzed in this section.

2.5.1 Experimental setting

To evaluate the performance of the proposed framework, experiments were conducted in a laboratory that contains metal structures and equipment with three workers moving around simultaneously. Figure 2.6 illustrates the layout of the laboratory, where two cameras and six BLE beacons are installed about 2m above the ground. This setting presents a challenging environment as the metal materials will interfere with the radio signals and the confined space will cause occlusions as workers move around. Two cameras record videos at 24fps with a resolution of 1920 x 1080. Six BLE beacons transmit signals at 10Hz. Before the experiments, all mobile devices (i.e., smartphones) are automatically synchronized with network time, and the cameras are manually synchronized, which ensures all devices are synchronized before collecting data. The evaluation of the proposed framework was conducted in an offline manner, with videos and received radio signals preprocessed and manually synchronized using timesteps generated during data collection, similar to the method in Mandelic et al., (2013). Specifically, radio signals were interpolated to be synchronized with videos.

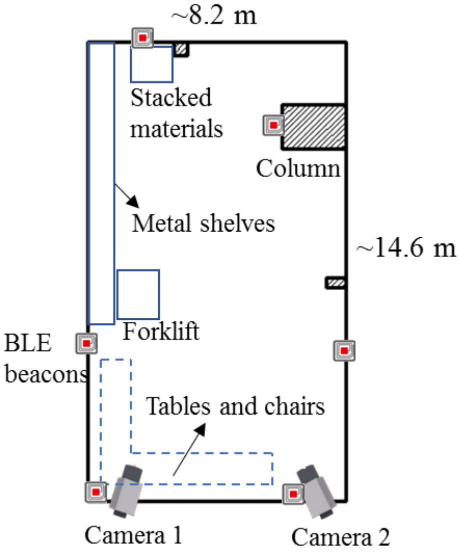


Figure 2.6 Testbed layout

The intrinsic parameters of each camera were calibrated using a 10 x 7 chessboard with 97mm square size. Twelve ground control points (GCPs) were used to correlate the camera to the world coordinate system and obtain the fundamental matrix between two cameras. The coordinates of GCPs were manually measured using a laser ranger with respect to a pre-defined 3D world coordinate system. The localization accuracy for the stereo vision system is considered as the distance between the estimated point location via triangulation and the measured point location, the average of which for the 12 GCPs is 0.04m. In practice, the land survey results can also be used to generate control points for camera calibration. For the radio system, each mobile device was calibrated separately to compute the parameters in Equation 2.1 by measuring the RSSIs with varying distances between the mobile device and BLE beacon based on the method described in Thaljaoui et al. (2015). The localization accuracy for the radio-based system is considered as the discrepancy between the estimated distance (from beacon to device) and the actual distance, which for the three mobile devices are 0.67m, 0.80m, 0.92m. The difference in localization accuracy further proves the advantage in localization by the stereo vision system.

Two experiments were performed. In the first experiment, three workers moved along pre-defined routes in the middle of the space for about 1.5 min, and this experiment is used to determine the optimal parameters (e.g., w_{color}, w_{ID}) in the proposed tracking algorithm. The second experiment lasts for 4 min, where three workers simulated construction operations with two of them compacting the ground in the middle region of the space and the other one transporting materials from one corner of the room to its diagonal corner. In this study, the ground truth of the trajectories was obtained following the method described in Mandeljc et al. (2013). Specifically, the helmet centers of workers were manually annotated frame by frame and their 3D coordinates were reconstructed using calibration information, the resulting trajectories are shown in Figure 2.7, where the red line represents worker 1, the blue line represents for worker 2, and the green line represents worker 3.

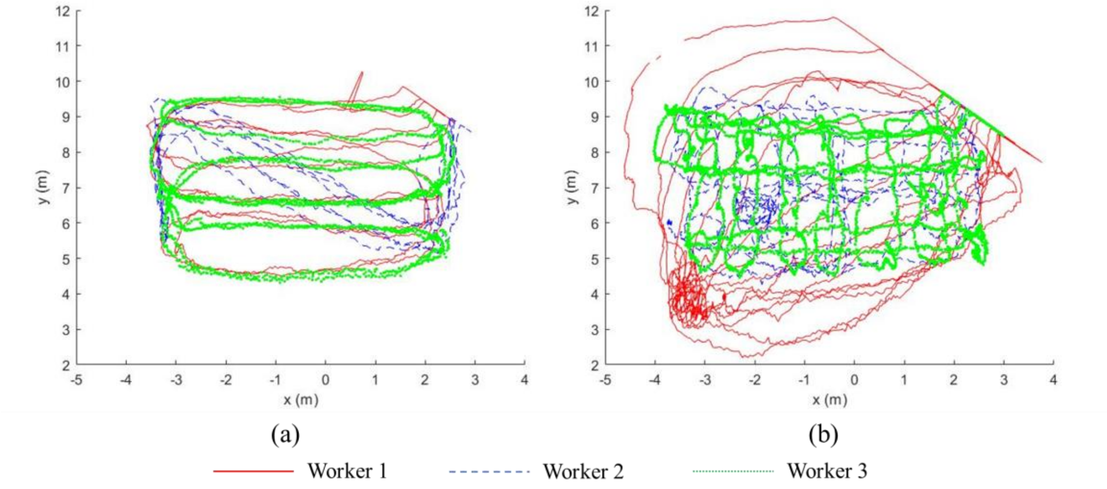


Figure 2.7 Ground truth trajectories in two experiments: (a) Experiment 1, (b) Experiment 2.

2.5.2 Evaluation Metrics

This study performs a systematic evaluation of the proposed framework. The evaluation metrics include false positive (FP), false negative (FN), identity switch (IDSW), fragmentation (FM), multiple object tracking accuracy (MOTA), and multiple object tracking precision (MOTP). These metrics are selected based on the benchmarking of the multi-object tracking problem (Milan et al., 2016) as defined in the computer vision community.

FP and FN are two common indicators to quantify the performance of the tracker-to-target assignment. Given a frame, if the output track does not correspond to an actual target, it is counted as an FP; if an actual target is missed by any output track, it is counted as an FN. In the evaluation phase, the output tracks are assigned to ground-truth targets by minimizing the global distance with a threshold of the maximum distance to be assigned (1 m in this study). The threshold was set empirically. A larger threshold will lead to less FNs but more FPs, while a smaller threshold will lead to more FNs but less FPs. Readers are referred to Milan et al. (2016) and Mandelic et al. (2013) for details. Ideally, a good tracking algorithm should result in as few FPs and FNs as possible.

IDSW and FM are another two important indicators to quantify the quality of the tracker. As multiple object tracking is a temporal problem, it is expected that a track corresponds to the same target all the time. Hence, an IDSW is counted when a ground truth target *i* is matched to track *j* at time *t* but it was not matched to the same track at time *t*-1. An FM is counted when a ground truth trajectory changes its status from tracked to untracked and then resumes to be tracked at a later point. Ideally, the number of IDSWs and FMs is expected to be as few as possible. In addition to these classic definitions, this study also measures the length of each IDSW and FM, and evaluates the number of IDSWs and FMs that last for more than 5 frames (about 0.2s) since such scenarios reflect more severe defects of the tracks compared to those lasting for a very short period. The numbers of IDSWs and FMs that last for more than 0.2s are used to evaluate the robustness of the tracking method. The smaller the number is, the more robust the method is because it indicates the capability of the tracking method to recover after experiencing possible errors. For FM, we also count the number of FMs that resume the same ID after the tracker is recovered.

MOTA is the most widely used metric that measures the overall accuracy and is treated as a primary index to evaluate the tracking performance in this study. It integrates three error sources, i.e., FN, FP, and IDSM, and is computed as $MOTA = 1 - \sum_t \left(FN_t + FP_t + IDSW_t \right) / \sum_t GT_t$, where t is the frame index and GT is the number of ground truth targets. The MOTA varies from 0 to 1, and the score increases as the performance improves. MOTP measures the overall localization precision, denoted as $MOTP = \sum_{t,i} d_{t,i} / \sum_t c_t$, where c_t refers to the number of matches in frame t and $d_{t,i}$ refers to the distance between track *i* with its assigned ground truth object at frame *t*.

2.5.3 Results

The proposed tracking framework consists of three parts with major contribution that differentiates the newly created method from other tracking methods: (1) the integration of radio-based identification in 2D tracking, (2) the searching for missed detections across camera views, and (3) the integration of radio-based identification in 3D tracking, as highlighted in Figure 2.2. This section compares the performance of tracking methods with and without the above processes to demonstrate the advantage of the proposed method.

The performance of four tracking approaches was compared and Table 2.2 lists the detailed characteristics of each tracking method. In Table 2.2, Approach 1 is the proposed tracking approach and Approach 4 is the conventional 3D tracking method based on entity matching across camera views. Approach 2 integrates the radio-based identification both in 2D and 3D, while Approach 3 only integrates the radio-based identification in 3D. Figure 2.8 shows the top-down view of tracking results for Approach 1 and 4 in both case studies, where solid lines represent the ground truth (GT) trajectories and dots represent the tracking results. From Figure 2.8(a), (c), despite some outliers, the results obtained from the proposed method align well with the GT, with each trajectory distinguished by entity ID. In contrast, although the results obtained from the vision-based approach generally align with GT, the tracking results cannot be separated into individual people due to the anonymous nature. Actually, once the tracks are lost for a while in Approach 4, a new track ID will be assigned after the tracking is recovered, making it difficult to track a specific person for the long term. It is noted that the top-down views for Approach 2 and 3 are similar to those of Approach 1 as all of them integrate vision and radio-based approaches.

Table 2.2 Tracking scenarios

Approach	**Integration of radio in 2D**	**Integration of radio in 3D**	**Search for missed detections across camera views**
1	Yes	Yes	Yes
2	Yes	Yes	No
3	No	Yes	No
4	No	No	No

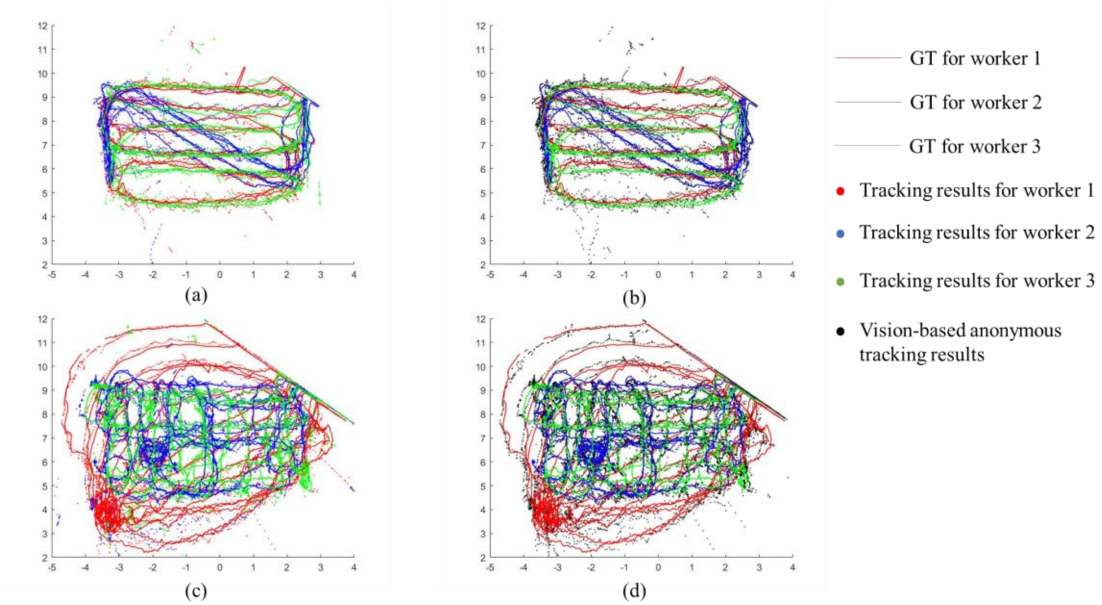


Figure 2.8 Top-down view of tracking results: (a) Case Study 1-Approach 1, (b) Case Study 1-Approach 4, (c) Case Study 2-Approach 1, (d) Case Study 2-Approach 4

Table 2.3 and Table 2.4 list the tracking performance of each approach for two cases, where the fusion approach leads to higher MOTA in both cases compared to the vision-based tracking approach. The detailed performance of each method is discussed and compared in the following sections.

Table 2.3 Tracking performance for Case 1 (~1.5 min, 2088 frames)

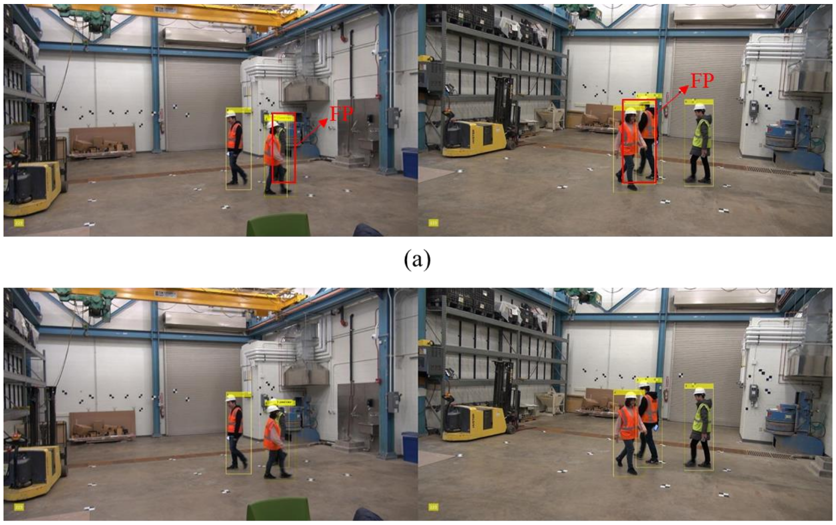
Approach	MOTA	FN	FP	FM	FM (>5)	FM (resumed)	IDSW	IDSW (>5)	MOTP (m)
1	**0.9483**	124	124	53	5	50	76	43	0.166
2	**0.9483**	139	111	54	6	52	74	41	0.164
3	0.94	138	138	62	6	56	100	52	0.166
4	0.8811	151	543	24	8	1	51	41	0.161

Table 2.4 Tracking performance for Case 2 (~4 min, 5044 frames)

Approach	MOTA	FN	FP	FM	FM (>5)	FM (resumed)	IDSW	IDSW (>5)	MOTP (m)
1	0.8953	356	1184	75	21	63	170	78	0.2
2	**0.9141**	505	753	75	16	65	145	65	0.194
3	0.8827	503	1203	100	11	82	210	103	0.209
4	0.8704	598	1406	75	20	23	112	92	0.188

Evaluation of the overall proposed framework

Conventionally, vision-based 3D trajectories of multiple objects are obtained through triangulation of matched entities across camera views. The newly created method is a holistic framework that integrates radio-based identification in both 2D and 3D tracking. The comparison of the performance (Approach 1 vs. Approach 4) clearly shows the advantage of the proposed method. As shown in Table 2.3 and Table 2.4, integrating radio-based identification with vision-based localization significantly increases the overall tracking accuracy compared to using vision-based localization alone for both cases, the improvement of which can also be explained by other metrics. The proposed method substantially reduces the number of FNs and FPs as the incorporation of radio-based detection provides reliable identification information that can effectively correct vision-based detections. Figure 2.9 illustrates an example of the correction of FPs using the proposed method. When two people are approaching, vision-based tracking (Approach 4) results in false positives (highlighted in red boxes) due to failure in correctly detecting workers, which are eliminated in the proposed method because the tracks that are not matched with radio-based identifications are excluded. The correction of FNs will be discussed in the next section.



(a)

(b)

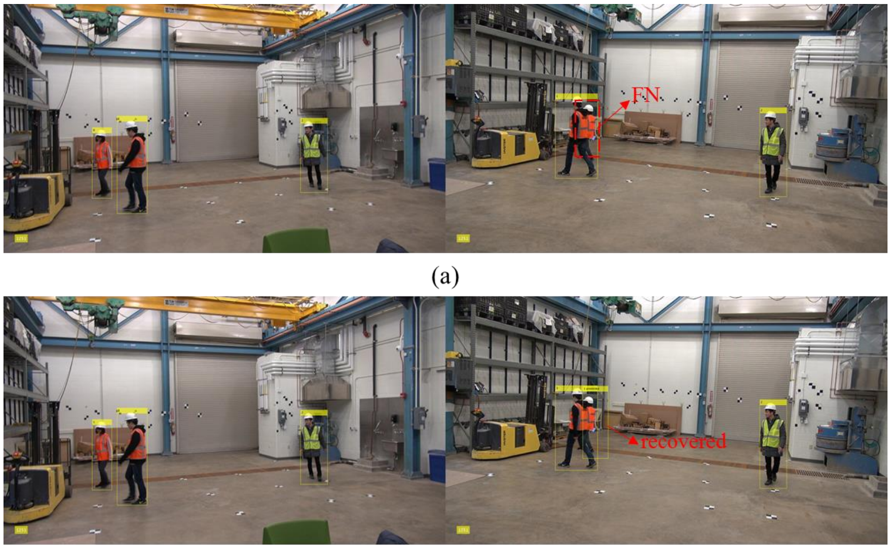
Figure 2.9 Example for the correction of false positives: (a) FP in Approach 4 (left view vs. right view), (b) corrected FP in Approach 1 (left view vs. right view).

Although the number of FMs in Approach 1 is larger than that in Approach 4 in Case 1, very few of them last for longer than 5 frames (0.2s). In other words, most FMs in the proposed method are recovered in a very short time period, which can be regarded as slight fluctuations. In such a sense, the proposed method is more robust. Furthermore, almost all FMs in Approach 1 are resumed with same ID (50 out of 53 and 63 out of 75), compared with Approach 4 (1 out of 24 and 23 out of 75), presenting a distinct advantage of the proposed method: due to the unique identification information provided by the radio system, the trajectory of the same target can be recovered even after it is untracked for a while, leading to more stable and robust trackers. Although the total number of IDSW is slightly larger in Approach 1, the number of those lasting for over 5 frames is compatible in the two approaches in Case 1 and is much fewer for Approach 1 in Case 2. For MOTP, as the proposed method does not directly leverage the localization result of the radio system, the average distance between tracking results and ground truth trajectories is almost the same in the two scenarios.

It is noted the overall accuracy is lower in Case 2 than in Case 1. It is because the second case presents a more challenging yet common scenario on construction sites: different groups of workers share the workspace and they have to adjust their movements to collaborate with co-workers as well to avoid conflicts with other entities. As a result, it is more likely to have occlusions in the congested area and workers can move in and outside the field of view from time to time. Furthermore, the parameters are fine-tuned to optimize the performance of Case 1, which may not be the optimal parameters for Case 2. In this way, we further proved that the proposed method leads to higher accuracy in different datasets.

Evaluation of the searching for missed detections

The comparison between Approach 1 and 2 evaluates the performance of the searching procedure for recovering missed detections across camera views. Figure 2.10 illustrates an example of recovering a false negative caused by occlusion using the proposed searching procedure across the camera view. For Case 1, the metrics in these two approaches are almost the same, except for the FPs and FNs. The proposed searching procedure successfully reduces the number of FNs despite some more FPs. However, it is argued that a smaller FN is favorable in construction application especially for safety management as it is more likely to cause collisions when we fail to track some entities on the site. Besides, as the state-of-the-art detection model used in this study is very accurate with over 0.95 precision and recall rates, it detects most of the targets in both camera views, resulting in less significant improvement of the proposed searching procedure. For Case 2, however, the searching procedure slightly lowers the tracking accuracy. The possible reason is that it uses the color histogram as the only cue in matching candidates with the target objects, which may lead to mismatches when the target is severely occluded. Future study will consider multiple cues to improve the robustness of the searching algorithm.



(a)

(b)

Figure 2.10 Example of recovered false negative: (a) FN in Approach 2 (left view vs. right view), (b) recovered FN in Approach 1 (left view vs. right view).

Evaluation of the integration of radio-based identification in 2D

In the proposed method, the radio system is integrated both in 2D and 3D tracking, and the comparison between Approach 2 and 3 illustrates the necessity in incorporating the radio system in 2D. In both cases, the incorporation of radio system in 2D almost improves performance in all aspects. It is because the tracking starts with 2D and by correcting FPs caused by vision-based detection errors with radio-based identification in individual camera views, the resulting 2D tracks are more accurate, which improves the accuracy of the downstream entity matching, leading to more accurate 3D tracks.

2.6 Contributions

The contribution of this research is in three aspects. First, this study strategically fuses two modes of sensing systems (i.e., vision and radio) by leveraging the advantages in two systems as a mutual complement and mitigating the limitations in the individual systems. Specifically, by augmenting visual tracks with entity IDs obtained in radio-based identification, the major drawbacks in the visual tracking system, fragmentation and ID switch caused by object detection error, are effectively corrected. For instance, in Case 1, only 9% of fragmentation in the proposed method lasts for more than 0.2s, compared to 33% in vision-based approach; and 56.6% of ID switch in the proposed method lasts for more than 0.2s, compared to 80.4% in vision-based approach. Meanwhile, as the location is updated primarily using vision-based results, the negative impact from the low-accurate radio-based localization is minimized: the resulting localization precision is in the sub-meter level, compatible with the vision-based approach, much higher than the radio-based approach (meter level).

Second, the proposed method integrates 2D and 3D tracking into a holistic framework. By incorporating radio-based identification in both 2D and 3D tracking processes, anonymous 2D and 3D visual tracks in each time step are associated via unique IDs, which are further carried to next time steps while updated by new observations. In this way, the possible detection and matching errors in each time step are mitigated and the continuity of the tracks is ensured. As a result, the proposed method increased the overall tracking accuracy by 8% and 3% for two case studies compared to the vision-based approach.

Third, the proposed searching procedure in entity matching across camera views allows for the recovery of missed detections in individual views due to occlusion, which reduced false negatives by 10.8% and 29.5% for two case studies compared to the method without the searching procedure. It can be extended to multiple cameras with even larger redundancy, resulting in more accurate and robust 3D tracking.

2.7 Conclusions and Discussion

This chapter presents a new hybrid framework that fuses vision-based tracking and radio-based identification and localization results for accurate and robust 3D tracking of multiple construction workers. Vision-based tracking is treated as the main source to extract the trajectory.

Radio-based identification and localization results used as a supplementary source, which are first matched with anonymous trackers obtained in single-camera views to provide identity information and correct false detections, and then associated with 3D locations in the world coordinate system obtained from stereo vision, resulting in ID-linked 3D trajectories. In addition, a searching algorithm is introduced to recover possible missed detections in one camera view from the corresponding observations in the other view by applying a sliding window to search for regions with the most similar appearance along the epipolar line.

The newly created method has been validated using two indoor experiments. The results show that the proposed method significantly improves the overall multiple object tracking accuracy. The proposed method resulted in a MOTA of 0.95 and 0.9 for two cases, compared to 0.88 and 0.87 obtained using conventional vision-based 3D tracking. In two cases, the proposed method reduced false negatives by 20% and 40%, and false positives by 77% and 15.8% respectively. The result suggests a more substantial improvement for Case 1, a relatively simple setting where all workers move within the field of view along pre-defined routes. Its efficacy is also validated in Case 2, a more challenging setting where workers simulate construction operations in different groups interacting with each other in a common workspace while moving in and out of view constantly. Moreover, the integration of radio-based identification allows the recovery of the same entity ID after the trajectory is fragmented—about 95% and 84% of fragmented trajectories are resumed with the same ID. It also ensures the tracking robustness—only 9% and 28% of fragmentations last for more than 0.2s.

The processing time for faster-RCNN-based visual detection is about 0.1s/frame using an NVIDIA GeForce GTX1060 6GB GPU. The processing time for the hybrid tracking is about 0.26s/frame using an Intel Core i7 CPU. Therefore, the processing time for current implementation is about 3fps, slightly faster than that (2.1fps) of a stereo vision-based tracking proposed in Lee and Park (2019). Although not achieving real-time, the processing time can be further reduced by code optimization, using a faster object detection framework (e.g., YOLO and SSD), and using more powerful processors. Moreover, it is argued that given the relatively slow speed of worker movement and the far-field construction videos, it is not necessary to track workers at a high frame rate (e.g., 24fps) since the difference in position between two frames will be small. Therefore, after appropriate optimization, the proposed method can be used for tracking construction workers.

As a demonstration of the proposed multi-worker tracking method, this study uses two fixed cameras in indoor experiments to provide the full coverage of the area of interest. In practice, when monitoring construction operations on the complex jobsites, more cameras are needed to ensure the desired coverage. To determine the optimal camera placement, one needs to simultaneously maximize camera coverage and minimize the total cost while satisfying various constraints, which has been an active research area and explored by many studies (Ahn et al., 2016; Altahir et al., 2017; Jian Zhao et al., 2013). In the construction domain, Yang et al., (2018) created an optimization method to find the optimal camera placement on construction sites. Kim et al. (2019) have created a novel framework to determine the optimal number, types, locations, and orientations of fixed cameras that maximize visible coverage and minimize total costs considering unique conditions of construction jobsites, such as power accessibility, facilities and occlusions, and work zones. In their case study, the optimal camera number of a 70 x 30m construction site is three. However, this number is site-specific depending on many factors discussed above.

In the proposed method, manual survey is needed to determine the locations of GCPs for camera system calibration. Although the land survey results can be used to generate control points, the dynamic changes of layout on construction sites pose great challenges in determining the locations for GCPs. To reduce manual work and ensure the practical feasibility, three possible solutions are recommended. First, when selecting GCPs, point locations that do not change frequently (e.g., on floor plane or other permanent structures) are preferred. Second, the designed site layout or building information models (BIM models), especially 4D models, can be leveraged to extract component locations and guide the dynamic deployment of GCPs to adapt to site changes. Third, as stereo vision mainly requires matching points cross camera views to establish epipolar geometry, visual feature descriptors can be used to automatically generate matching feature points, which will significantly reduce manual work. For instance, scale-invariant feature transform (SIFT) can be used to generate feature points in multiple camera views in order to establish epipolar geometry, as described in Lee and Park (2019).

When deploying the BLE beacon system, the number of beacons can be estimated considering the signal range of beacons (depending on the transmitting power) and the possible signal attenuations (Topak et al., 2018). For instance, for the beacons used in this study, its maximum range is about 40m with the power of -4 dBm, 15m with the power of -12 dBm, and can be at most 200m with the power of 10 dBm. The minimum coverage of beacons should be within the maximum signal range. In this study, the transmitting power of -4 dBm was selected to provide a maximum range of 40m considering the signal attenuation in the testbed. Park et al., (2017) also used a similar density (12 beacons for 240 m^2) in a construction safety monitoring application. Increasing the density of beacons will improve the localization accuracy. For instance, Chen et al., (2018) tested different numbers of beacons in an area of 274 m^2 and found 37 beacons results in the highest accuracy compared with 12 beacons and 8 beacons. Therefore, the optimal number of beacons should be determined considering signal range, signal attenuations, and desired accuracy.

There remain some limitations that deserve further research efforts. First, as a demonstration, only six BLE beacons are used in the indoor experiments and the mobile devices are of different quality levels, resulting in about 2m accuracy in the radio system, which may affect the tracking performance when integrating with a vision system. To further improve the tracking performance, more beacons can be deployed to improve the radio-based localization accuracy. Second, as the searching algorithm uses the color histogram as the only cue to recover missed detections, it may not work well when a target is completed colluded by other entities. In such a case, one can remove this process from the tracking framework and only implement the remaining parts (integrating radio-based identification in both 2D and 3D) to avoid errors caused by the mismatch of entities. Third, since the main purpose and scope of this study is the creation and evaluation of the hybrid tracking algorithm, the tracking process was performed in an off-line manner and the time-synchronization was manually conducted. In the future, the algorithm will be further optimized and a real-time tracking system will be devised based on the newly created hybrid tracking framework with all devices automatically synchronized via a Network Time Protocol server.

3. TWO-STEP LONG SHORT-TERM MEMORY METHOD FOR IDENTIFYING CONSTRUCTION ACTIVITIES THROUGH POSITIONAL AND ATTENTIONAL CUES

This chapter presents a two-step classification approach – working group identification followed by activity recognition, leveraging both positional and attentional cues, to recognize complex interactions and their involved entities from videos that contain different activities with multiple entities. The spatial and attentional states of individual entities on 2D images are represented numerically, and the corresponding 2D positional and attentional cues between two entities are computed. Long short-term memory (LSTM) networks are designed to (1) classify whether two entities belong to the same group, and (2) recognize the activities they are involved in. Two sets of construction videos—one hospital construction project on the publicly-available website and one teaching building project taken by the author on Purdue campus, are used to validate the newly created method. Manual annotations regarding the spatial and attentional states are used to compute the positional and attentional cues proposed in this study. The group/non-group information and the corresponding construction activities for each pair of construction entities are manually labeled to provide ground truth labels for supervised learning when training and testing the proposed two-step LSTM model. It was found that by leveraging both positional and attentional cues, the accuracy increases from 85% to 95% compared with cases using positional cues alone. Moreover, identifying working groups before recognizing ongoing activities enables the exclusion of group-irrelevant entities and thus, improves the performance.

This work was previously published in Automation in Construction (J. Cai et al., 2019). This chapter is re-printed with permission from Vol 104, Jiannan Cai, Yuxi Zhang, and Hubo Cai, "*Two-step long short-term memory method for identifying construction activities through positional and attentional cues*", 102886, Copyright Elsevier (2019). Table titles and figure captions have been modified to maintain the form of the dissertation.

3.1 Introduction

Construction entities (including both workers and equipment) interact with each other, constituting working groups to accomplish assigned tasks. Recognizing ongoing activities and involved working groups is important as it enables the comprehension of jobsite context, which in turn allows the interpretation of workers' intentions, the prediction of their movements, and the detection of inappropriate interactions that are counterproductive and may cause disastrous consequences such as struck-by accidents. This type of accident has led to 804 fatalities from 2011 to 2015 (X. S. Dong et al., 2017). Since construction activities are goal-orientated and they determine the movement patterns of involved entities, information on construction activities and their working groups enables context-aware movement prediction that is expected to be accurate and reliable. Consequently, improper interactions and potential conflicts are detected in advance to prevent struck-by accidents.

A number of methods have been developed to automatically recognize the actions of individual entities from images/videos (Golparvar-Fard et al., 2013; Khosrowpour et al., 2014; H. Luo et al., 2018), but little attention has been paid on the interactions between two entities over time (H. H. Kim et al., 2018; J. Kim et al., 2018). The few studies on entity interactions relied on the spatial-temporal relationship between entities through hand-crafted rules. Moreover, construction images/videos were simplified to only contain entities involved in a single activity, excluding all irrelevant entities (X. Luo et al., 2018). In reality, however, many workers and machines co-exist and collaborate to accomplish different tasks. For those entities that are spatially close, not all of them are collaborating on a single activity. Therefore, there is a critical knowledge gap – methods are needed to identify working groups and recognize corresponding activities using images/videos that contain many entities collaborating on various tasks.

Aiming at accurately identifying working groups and recognizing corresponding activities with a specific focus on those involving human workers, this chapter presents a two-step long short-term memory (LSTM) approach that integrates the positional and attentional cues to first identify working groups and then recognize activities. The positional cues refer to the spatial relationship between entities, such as the distance and relative direction. The attentional cues are the features that model an entity's visual attention (e.g., the head pose) and the attentional exchange between two entities (e.g., the gaze exchange between two entities).

3.2 Review of Related Studies

This section reviews studies related to vision-based construction object detection and tracking, construction activity recognition, as well as the studies on attention-based group activity analysis in the computer vision community, and discusses the knowledge gaps.

3.2.1 Vision-based construction object detection and tracking

Vision-based object detection and tracking have been an active research area in the construction domain during the past 10 years. Table 3.1 lists related studies on vision-based object detection. Among these studies, earlier works (Memarzadeh et al., 2013; M. W. Park & Brilakis, 2012; Rezazadeh Azar & McCabe, 2011, 2012) used visual features such as histogram of orientated gradients (HOG) to describe objects and applied traditional machine learning algorithms such as support vector machine (SVM) to detect target objects. Recently, the deep learning-based approach has been increasingly adopted to detect workers and equipment with various appearances and postures (W. Fang et al., 2018b, 2018c; Son et al., 2019a). Table 3.2 lists related studies on vision-based object tracking, grouped into two types: 2D tracking that obtains object trajectory in terms of 2D coordinates on the image plane; and 3D tracking that extracts object location in the 3D world coordinate system. These studies have proven the great potential of extracting real-time states of construction objects from images and videos and formed the technical premise of this study.

Table 3.1 Related studies on vision-based object detection

Target object(s)	**Features**	**Detection method**	**Literature**
Workers, excavators, and trucks	HOG and color histogram	SVM, kNN	(Memarzadeh et al., 2013; M. W. Park & Brilakis, 2012)
Trucks	Haar-HOG, Blob-HOG	SVM	(Rezazadeh Azar & McCabe, 2011)
Excavators	HOG	Part-based object recognition model	(Rezazadeh Azar & McCabe, 2012)
Workers and excavators	Deep CNN models	Faster R-CNN	(W. Fang et al., 2018b, 2018c; Son et al., 2019a)

Table 3.2 Related studies on vision-based object tracking

Type	**Tracking method**	**Literature**
2D	Particle filter-based tracking with manually initialized objects	(Zhu et al., 2016b)
	Integrating detection with tracking to address occlusion challenge	(M. W. Park & Brilakis, 2016; Zhu et al., 2017)
	Machine learning-based tracking by comparing the target with pretrained appearance model	(J. Yang et al., 2010)
3D	Integrating 2D tracking using triangulation based on stereo vision	(Brilakis et al., 2011; M.-W. Park et al., 2011; Zhu et al., 2016a)
	Matching entities cross camera views based on various criteria (e.g., epipolar constraints, appearance model, moving direction)	(Konstantinou & Brilakis, 2018; Y. J. Lee & Park, 2019)
	Tracking and detecting excavators based on kinematic shape and key node features using stereo cameras	(Yuan et al., 2016)

3.2.2 Construction activity recognition

Aggarwal and Ryoo (2011) categorized human activity into four levels based on the complexity: gestures, actions, interactions, and group activities. This categorization can be applied to the construction context. For instance, “raising arm” of a worker and “swinging” of an excavator are gestures; “laying bricks” and “excavating” are actions; “an excavator is loading dirt onto a dump truck” is an interaction; and “a fleet of excavators and trucks are moving the dirt” is a group activity. In this study, group activities and interactions, as defined by Aggarwal and Ryoo (2011), are the focus and refer to a construction activity that involves multiple entities working with each other.

Table 3.3 summarizes related studies on construction activity recognition. The studies are divided into three major groups: motion-based, audio-based, and vision-based approaches, among which a vision-based approach is the focus of this study as visual data provide rich contextual information and can be used to exploit interactions among multiple entities without attaching any sensors on the objects. From Table 3.3, most existing studies focused on the first two levels of activities, i.e., gesture and action recognition for individual entities, while the analysis of interactions and group activities among multiple entities remains a challenge (H. Luo et al., 2018). Few studies (H. H. Kim et al., 2018; J. Kim et al., 2018) have exploited the interaction between excavators and dump trucks in the earthmoving operations. A common limitation in this stream of studies is the simplification of the videos by cleaning them to contain only one activity and its involved entities, excluding all irrelevant entities. This simplification is far from reality—the co-existence of multiple working groups collaborating on several activities on a typical construction site.

Table 3.3 Related studies on construction activity recognition

Type	Sensors	Activity type	Features	Recognition method	Literature
Motion-based	Accelerometer	**Actions** (laying brick, filling joints)	Time-domain and frequency-domain features	Naïve Bayes, decision tree (DT), artificial neural network (ANN)	(Joshua & Varghese, 2010)
	Accelerometer, gyroscope, GPS	**Actions** (excavator's dumping, scooping; worker's sewing, hammering, etc.)		ANN, DT, kNN, SVM	(Akhavian & Behzadan, 2015, 2018)
Audio-based	Audio recorders	**Actions** (Major and minor actions of heavy machines)	Time-frequency features	SVM	(C.-F. Cheng et al., 2019; C. F. Cheng et al., 2017)
Vision-based	Cameras	**Posture** for excavator	Locations of visual markers attached on excavator	Calculating angles between different components of excavator	(Feng et al., 2018; Lundeen et al., 2016)
		Actions (excavator's swing, excavating; worker's transporting, bending, etc.)	Bag-of-video-feature-words model	Bayesian network model	(Gong et al., 2011)
		Actions (excavator's digging, dumping, and hauling; truck's filling, moving, dumping.)	Spatial-temporal features, HOG	SVM	(Golparvar-Fard et al., 2013)
		Actions (walking, transporting, actions on the ladder)	Deep learning-based methods (CNN, LSTM networks)		(Ding et al., 2018; H. Luo et al., 2018)
		Interactions between excavators and dump trucks	spatial-temporal relationship between equipment	pre-defined rules	(H. H. Kim et al., 2018; J. Kim et al., 2018)
		Group activity (Leveling land, placing concrete, etc.)	object locations and classes	Semantic and spatial relevance network	(X. Luo et al., 2018)

Luo et al. (2018) attempted to overcome this limitation via a two-step method to first detect construction objects and then recognize diverse construction activities using a predefined semantic and spatial relevance network. While it works with a diverse group of construction activities, its precision and recall rates are relatively low, possibly because 1) the use of still images causes the loss of temporal information, leading to the difficulty in detecting prolonged activities and transitive states, 2) the manually defined relevance networks and corresponding activity patterns are inadequate for dynamic construction job sites, and 3) the neglect of important cues (e.g., attentional cues) other than spatial proximity limited the capability of fully exploiting the interactions among entities.

3.2.3 Attention-based group activity analysis

Group activity analysis has been an active research area in the computer vision community for many applications such as video surveillance, human detection, and path prediction. Many studies have found that attention-based cues are critical to understanding the interactions within and the context of a group. These studies typically use a human's head pose as the approximation of their visual attention due to its visibility in the low-resolution videos. For instance, Ba and Odobez (2011) inferred visual focus of attention and interaction with others from the head pose information obtained from videos of meetings. Chamveha et al. (2014) incorporated attention-based cues, such as the gaze exchange and joint attention of pedestrians, and position-based cues, including the displacement and difference in velocity, to discover the social groups among pedestrians. Pereira et al. (2017) used position- and attention-based cues to classify individual behaviors of "exploring, interested, distracted, and disorientated" as well as group behaviors of "equally interested, balanced interests, unbalanced interests, and chatting". Qin and Shelton (2016) proved that the social grouping information, multi-target tracking, and head pose estimation can be coupled together to gain better performance in multi-target tracking and head pose estimation. These studies prove the great potential in leveraging attentional cues to capture interactions among multiple people and recognize group activities, which motivates the study presented in this paper—incorporating attentional cues to better identify construction working groups and recognize corresponding group activities.

3.2.4 Knowledge gaps

The review of related studies reveals three knowledge gaps in construction group activity analysis. First, lack of methods have been created to identify construction working group. In the congested and dynamic construction scene, not all co-existing entities are interacting with each other and collaborating on the same activities. The existence of irrelevant entities will hinder the accurate recognition of an ongoing activity. Therefore, it is critical to identify the working groups and involved entities so that the ongoing activities can be recognized only considering the relevant entities. Second, attention-based cues presented in the interactions among entities are neglected. Most studies analyze the interactions relying on position-based cues such as the distance and relative movement between two entities, which is inadequate as not all entities that are spatially close are in the same working group. The attention-based cues are a missing opportunity to enhance the identification of construction working groups and the understanding of ongoing activities. Third, interaction patterns among multiple entities are not sufficiently learned. Most approaches use hand-crafted rules to determine the interactions based on the spatial-temporal relationship between two entities, which might be effective for simple interactions with repeated cycles but is inadequate for complex group activities with spatial-temporal patterns that are difficult to pre-design.

To overcome these gaps, this study aims at (1) identifying both attentional and positional cues among workers and/or equipment, (2) computing and representing them as numerical features, and (3) deploying an LSTM-based machine learning model to learn the temporal dependency of the features in order to better identify working groups and corresponding group activities.

3.3 Methodology

A new two-step LSTM-based method has been developed to identify working groups and recognize the corresponding group activities on construction sites using positional and attentional cues as features. Figure 3.1 illustrates the overall framework. Step 1 focuses on working group identification. The spatial and attentional states of individual entities are represented numerically and the corresponding positional and attentional cues between two entities are computed. The computed cues are constructed into time-sequential features and fed into an LSTM-based, binary classification model to determine whether two entities belong to the same working group. Step 2 involves an LSTM-based, multi-classification model to recognize specific activities. Note this method is built on the premise that the real-time states of entities (e.g., location, velocity, head pose, etc.) can be acquired from visual data, as proven in many references (e.g., (Konstantinou & Brilakis, 2018; Memarzadeh et al., 2013; M. W. Park & Brilakis, 2016; Raza et al., 2018; Zhu et al., 2017)).

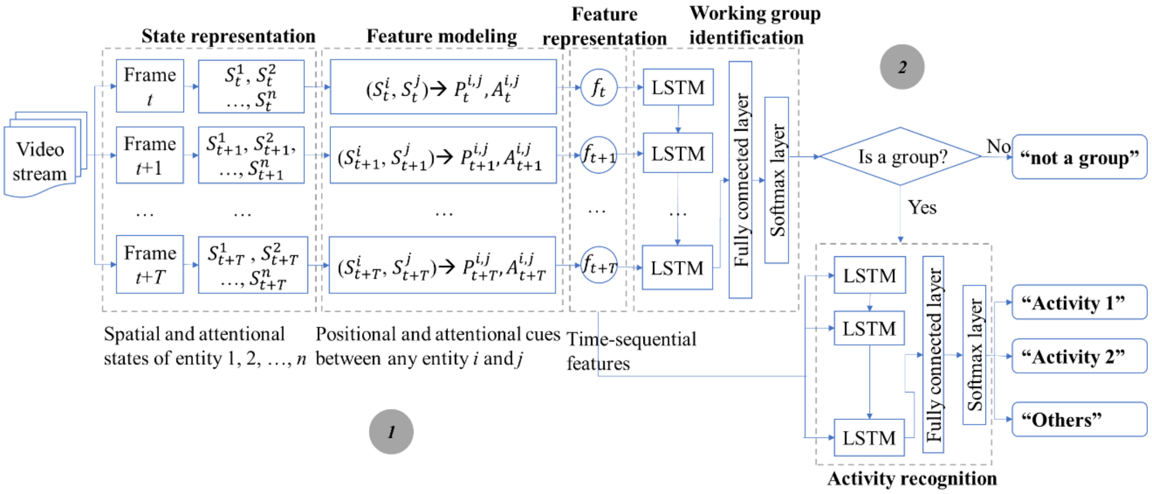

State representation
Feature modeling
Feature representation
Working group identification
Video stream
Frame t
Frame t+1
Frame t+T
$S_t^1, S_t^2, \ldots, S_t^n$
$S_{t+1}^1, S_{t+1}^2, \ldots, S_{t+1}^n$
$S_{t+T}^1, S_{t+T}^2, \ldots, S_{t+T}^n$
$(S_t^i, S_t^j) \rightarrow P_t^{i,j}, A_t^{i,j}$
$(S_{t+1}^i, S_{t+1}^j) \rightarrow P_{t+1}^{i,j}, A_{t+1}^{i,j}$
$(S_{t+T}^i, S_{t+T}^j) \rightarrow P_{t+T}^{i,j}, A_{t+T}^{i,j}$
f_t
f_{t+1}
f_{t+T}
LSTM
Fully connected layer
Softmax layer
Is a group?
No
"not a group"
Yes
"Activity 1"
"Activity 2"
"Others"
Activity recognition
Spatial and attentional states of entity 1, 2, …, n
Positional and attentional cues between any entity i and j
Time-sequential features
1
2

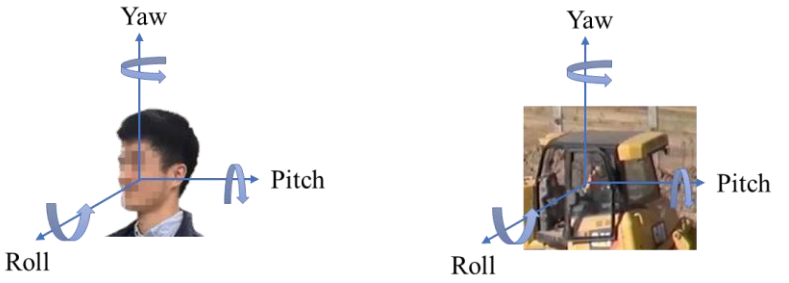
Figure 3.1 Two-step LSTM-based method for activity recognition

The main departing point of this method is the use of LSTM networks for learning the patterns of positional and attentional cues presented in the interactions among multiple objects, rather than using predefined, hand-crafted rules and patterns. This design captures the temporal dependency among features undergoing complex interactions. This newly developed method is the first of its kind in incorporating attentional cues into the construction domain as the supplementary of positional cues to more accurately interpret visual construction data for working group identification and activity recognition.

3.3.1 Spatial and attentional states

Both spatial and attentional states are used to describe entities. An entity's spatial state refers to its real-time position, typically measured by the central coordinates of its bounding box for 2D images; an entity's attentional state refers to the direction of its visual attention, captured by head pose, body orientation, and body pose.

Figure 3.2 illustrates the use of yaw, pitch, and roll to model the head pose for workers and equipment (noting that the main cab of the equipment is regarded as the "head"). Traditionally, the head yaw alone is regarded as an approximation of the visual attention in the horizontal direction (Chamveha et al., 2014; Qin & Shelton, 2016; Raza et al., 2018). However, on construction sites, the visual attention in the vertical direction, captured by head pitch, is also critical to infer working groups and activities. For instance, workers tend to watch horizontally when transporting materials and look down to the ground when paving roads or pouring concrete. Therefore, both head yaw and pitch are considered in this study.



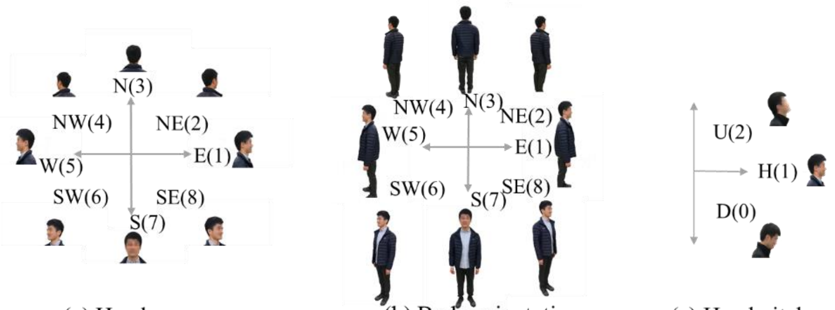
(a) Head pose for worker (b) Head (cab) pose for equipment

Figure 3.2 Illustration for head pose.

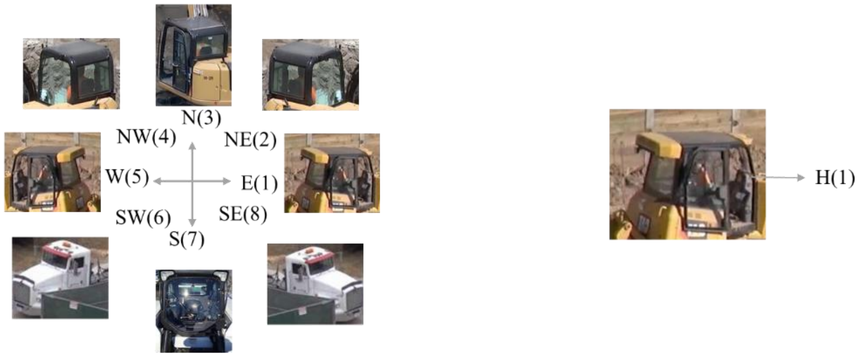
In addition to head pose, body orientation and body pose are included to describe attentional states. For workers, although body orientation is usually consistent with the head yaw, the inconsistent cases provide strong cues on visual attention. When the head yaw is different from body orientation, it is most likely that the entity is directly interacting or focusing on the objects in the direction of head yaw (Ozturk et al., 2011). The body pose (i.e., bend or standing) also affects the worker's attention because when workers bend, they are looking at the ground and their head pose points to the entities with which they are directly interacting. For equipment, it is treated as rigid objects and therefore, the body orientation is considered to be identical to the head yaw.

Given spatial and attentional states at time step t, an entity i is represented as $S_t^i = \{P_t^i, H_t^i, bo_t^i, bp_t^i\}$, where $P_t^i=(x_{\min t}^i, y_{\min t}^i, x_{\max t}^i, y_{\max t}^i)$ is the boundary 2D coordinates of the bounding box, with which the 2D location of the entity is derived as $x_t^i = (x_{\min t}^i + x_{\max t}^i)/2$ and $y_t^i = (y_{\min t}^i + y_{\max t}^i)/2$; $H_t^i = (yaw_t^i, pitch_t^i)$ represents the head yaw and pitch; bo_t^i represents the body orientation; and bp_t^i represents the body pose categorized in three classes: standing – 1, bending – 2, and not applicable (for equipment) – 3.

The head pose and body orientation are categorized further and each category is represented with a numerical value. Specifically, the worker's head yaw and body orientation are categorized into eight bins: north (N) – 1, south (S) – 2, east (E) – 3, west (W) – 4, northeast (NE) – 5, northwest (NW) – 6, southeast (SE) – 7, and southwest (SW) – 8, as shown in Figure 3.3(a) and (b). The head pitch is categorized into three bins: looking up (U) – 2, looking horizontally (H) – 1, and looking down (D) – 0, as shown in Figure 3.3(c). For equipment (considered as rigid objects), the body orientation is identical to the head yaw (Figure 3.3d) and the head pitch always remains horizontal (Figure 3.3e). The numerical values assigned to the orientation bins imply the similarity among the orientations. Such a design incorporates the uncertainty in the estimated orientation and is suitable for processing low-resolution videos of dynamic, complex, and often congested construction sites.



(a) Head yaw (b) Body orientation (c) Head pitch



(d) Head yaw and body orientation of equipment (e) Head pitch of equipment

Figure 3.3 Head pose and body orientation

Figure 3.4 illustrates an example image of the cooperation of a worker and a bulldozer at time *t*. The spatial and attentional state of worker *i* at time *t* is denoted by $S_t^i = \{(37,116,128,372),(1,1),1,1\}$, indicating the worker is located at (82.5, 244) (82.5 = (37+128)/2; 244 = (116+372)/2) with head and body facing east, looking horizontally and standing still. The state of equipment *j* at time *t* is denoted by $S_t^j = \{(436,10,830,372),(6,1),6,0\}$, indicating it is located at (633,191) and facing southwest.

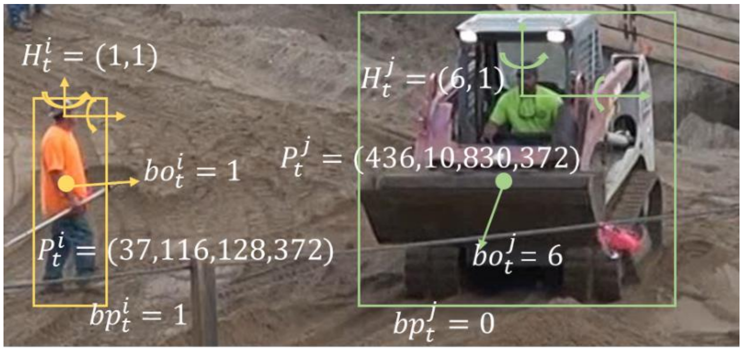


Figure 3.4 Spatial and attentional states of individual entities

3.3.2 Modeling positional and attentional cues

The positional and attentional cues refer to the positional and attentional relationship between two entities, computed using the spatial and attentional states of individual entities. These cues are regarded as critical features for working group identification and construction activity recognition.

Positional cues

The positional cues are measured considering both distance and direction between two entities based on their locations. The distance relationship between two entities is modeled topologically. The directional relationship is represented using a project-based model (Isli, 2003) that divides relative direction into eight zones: east, northeast, north, northwest, west, southwest, south, and southeast. Taking time steps into consideration, differences in moving direction and speed between two entities and the difference between the moving direction (of one entity) and the relative direction (between two entities) are computed as the additional measures for the positional cues. The resulting five measures, namely, distance relationship, directional relationship, difference in speed, difference in moving direction, and difference between moving direction and relative direction, are described in detail as follows.

Distance relationship

Distance relationship is modeled as topological relationship between bounding boxes of two entities, measuring the proximity between two entities. It is represented as $P_1^{i,j}$, where *i* and *j* refer to two entities and subscript 1 indicates it is the first positional cue measurement. The 9-Intersection model (Egenhofer & Herring, 1992) is adopted to describe the topological relationship. Figure 3.5 illustrates eight common topological relationships: “disjoint,” “meet,” “overlap,” “covers,” “covered by,” “contains,” “inside,” and “equal.” “Disjoint” is for two entities that are relatively far away, and it is further labeled as very close, close, medium, far, or very far; “meet” is for two entities that are next to or touch each other; “overlap” means that two entities are spatially close, indicating it is likely that they belong to the same group, and “cover/covered by,” “equal,” and “contain/inside” are special cases. Topological relationships connected by a solid line are topological neighbors – they can be converted directly to each other through spatial transformation. The change in the topological relationship between two entities over time reveals the change of their distance relationship as time progresses.

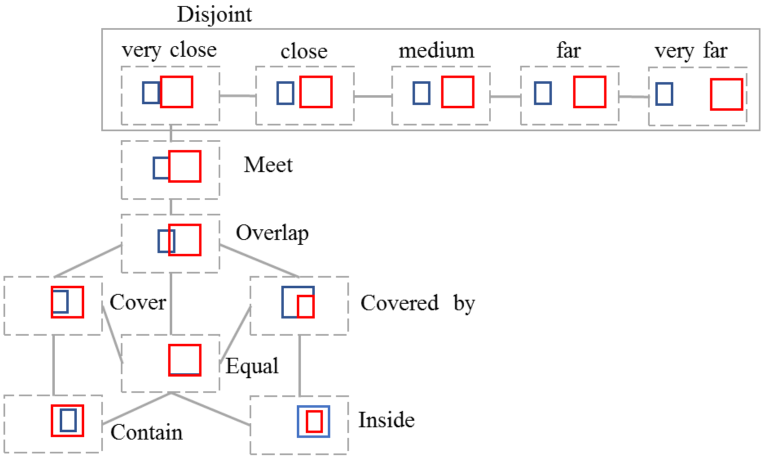


Figure 3.5 Topological relationships

The five categories of “disjoint” describe the degree of proximity, based on the ratio of the distance between two entities to the average size of the entities. The degree of disjoint relation is determined using Equation 3.1, where $w^{i,j}$ represents the average size of the two entities, computed as $w^{i,j} = (w^i + w^j)/2$ (where w^i is the width of the bounding box of entity *i*), and $d^{i,j}$ is the distance between the bounding box centers of the two entities *i* and *j*, computed as $d^{i,j} = \sqrt{(x^i - x^j)^2 + (y^i - y^j)^2}$ (where x^i and y^i are the center coordinates for entity *i*).

$$\text{Degree of disjoint} = \begin{cases} \text{very far}, & \text{if } d^{i,j} \geq 9w^{i,j} \\ \text{far}, & \text{if } 7w^{i,j} \leq d^{i,j} < 9w^{i,j} \\ \text{medium}, & \text{if } 5w^{i,j} \leq d^{i,j} < 7w^{i,j} \\ \text{close}, & \text{if } 3w^{i,j} \leq d^{i,j} < 5w^{i,j} \\ \text{very close}, & \text{if } d^{i,j} < 3w^{i,j} \end{cases} \quad (3.1)$$

The topological distance refers to the number of links in the shortest path between topological relationships in the neighborhood graph (Nabil et al., 1996) (shown in Figure 3.5). For instance, there is only one step for "meet" and "overlap" to be converted to each other, and thus, their topological distance is 1. For "cover" to be transformed to "disjoint", the shortest path is "cover" – "overlap" – "meet" – "disjoint", hence, the distance between them is 3. Table 3.4 illustrates the topological distance matrix for all topological relationships. The topological distance captures the similarity between topological relationships - a smaller value indicates a stronger similarity. It forms the basis to numerically represent the topological relationships.

Table 3.4 Topological distance matrix (Nabil et al., 1996)

Distance	**disjoint**	**meet**	**overlap**	**cover**	**contain**	**covered by**	**inside**	**equal**
disjoint	0	1	2	3	4	3	4	3
meet	1	0	1	2	3	2	3	2
overlap	2	1	0	1	2	1	2	1
cover	3	2	1	0	1	2	2	1
contain	4	3	2	1	0	2	2	1
covered by	3	2	1	2	2	0	1	1
inside	4	3	2	2	2	1	0	1
equal	3	2	1	1	1	1	1	0

As Table 3.5 illustrates for this study, the topological relationships of "contain", "inside", and "equal" are treated as the baseline and assigned with a value of 0 as they are the closest distance relationships between two entities. For other relationships, their topological distances to the closest baseline topological relationship are used as the numerical values. For instance, the distance between "cover" and "contain" is 1, and thus, the numerical representation of "cover" is 1. Similarly, "overlay" is represented by 2 and "meet" by 3. For the five subcategories in "disjoint", "very close" is assigned a value of 4, "close" 5, "medium" 6, "far" 7, and "very far" 8. Table 3.5 tabulates these numerical value assignments to the eight topological relationships and the five subcategories under "disjoint".

Table 3.5 Numerical representation of distance relationship ($P_1^{i,j}$)

Relation	**Numerical value**	**Relation**	**Numerical value**
Disjoint (very far)	8	Overlap	2
Disjoint (far)	7	Cover	1
Disjoint (medium)	6	Contain	0
Disjoint (close)	5	Covered by	1
Disjoint (very close)	4	Inside	0
Meet	3	Equal	0

Directional relationship

The directional relationship between entity *i* and *j* ($P_2^{i,j}$) measures the relative direction of entity *j* with respect to entity *i* on a 2D plane. Figure 3.6 illustrates the numerical representation of $P_2^{i,j}$, where a 2D plane is divided into eight regions, centered at the position of entity *i,* and the region in which entity *j* locates indicates the directional relation between *i* and *j*. Each region is one relative direction that is represented using discrete values: east (E) – 1, northeast (NE) – 2, north (N) – 3, northwest (NW) – 4, west (W) – 5, southwest (SW) – 6, south (S) – 7, and southeast (SE) – 8. Using regions rather than the directional vector directly affords a reasonable tolerance to noises and uncertainties—the directional relationship is still correct even when the perceived position is slightly inaccurate.

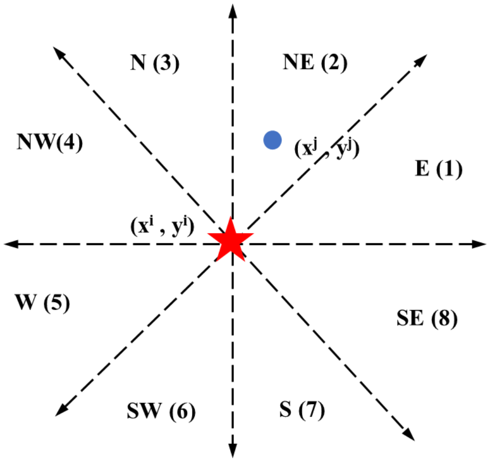


Figure 3.6 Representation of directional relationship ($P_2^{i,j}$)

Difference in speed and difference in moving direction

The differences in speed ($P_3^{i,j}$) and in moving direction ($P_4^{i,j}$) together measure the relative movement between two entities. The difference in speed (normalized to range [0, 1]) is computed using Equation 3.2, where v_t^i is the speed of entity *i* and calculated based on the positions at consecutive time steps, i.e., $v_t^i = \sqrt{\left(x_{t+1}^i - x_t^i\right)^2 + \left(y_{t+1}^i - y_t^i\right)^2}$.

$$P_3^{i,j} = abs(v_t^i - v_t^j) / max(v_t^i, v_t^j) \tag{3.2}$$

The difference in the moving direction is computed as Equation 3.3, where θ^i is the moving direction of entity *i*, represented as the numerical values in Figure 3.6. The difference between two directions refers to the shortest path between them. Therefore, when $abs(\theta^i - \theta^j)$ is greater than 4, the shortest distance is computed clockwise as $8 - abs(\theta^i - \theta^j)$; when $abs(\theta^i - \theta^j)$ is less than or equal to 4, the shortest distance is computed counterclockwise as $abs(\theta^i - \theta^j)$. For instance, the shortest path from “E” to “N” is “E” (1) – “NE” (2) – “N” (3) and the difference is calculated as 3 – 1 = 2; the shortest path from “E” to “S” is “E” (1) – “SE” (8) – “S” (7) and the difference is calculated as 8 – abs(1-7) = 2.

$$P_4^{i,j} = \min\{abs(\theta^i - \theta^j), 8 - abs(\theta^i - \theta^j)\} \tag{3.3}$$

Difference between moving direction and relative direction

The difference between the moving direction of entity *i*, θ^i, and the relative direction of entity *i* and *j*, $\varphi^{i,j}$, measures the degree of *i* moving towards *j*. It is computed using Equation 3.3, with the orientation in the equation replaced by θ^i and $\varphi^{i,j}$, where θ^i and $\varphi^{i,j}$ are numerical values as illustrated in Figure 3.6.

Attentional cues

In this study, attentional cues are measured by difference between head yaw and relative direction, difference in head yaw, difference between head yaw and moving direction, difference between head yaw and body orientation, head pitch, and head pose, described in detail as follows. All measures are applicable to both workers and equipment.

Difference between head yaw and relative direction

The difference between the head yaw of entity *i*, yaw^i, and the relative direction of entity *j* with respect to *i*, $\varphi^{i,j}$, measures the gaze exchange between two entities and is computed using Equation 3.3 with the orientations replaced by yaw^i and $\varphi^{i,j}$. This cue captures the degree of entity *i* looking at entity *j*. It can also be used to infer the intention of construction workers. For instance, if the difference between head pose and relative direction is very small, the entity *i* is more likely to interact with entity *j*.

Difference in head yaw

The difference in head yaw measures the joint attention of entity *i* and *j*, computed using Equation 3.3, with the directions in the equation being yaw^i and yaw^j. In many construction activities, entities are collaborating to operate on a common object, resulting in a small difference in head yaw.

Differences between head yaw and moving direction and difference between head yaw and body orientation

For one entity, the difference between its head yaw and body orientation, and difference between its head yaw and moving direction are strong cues inferring the change of its visual attention (Ozturk et al., 2011). For instance, if a worker is standing to the north with his head facing east, it is more likely that he is interacting with entities in the east direction. These two measures are computed as the difference between yaw^i and bo^i, and yaw^i and θ^i using Equation 3.3.

Head pitch and body pose

The head pitch and body pose of one entity also reflect its visual attention, which are special cues on construction job sites, as discussed in Section 3.3.1. In this study, these two measures are denoted by $A_5^i = pitch^i$ and $A_6^i = bp^i$.

3.3.3 LSTM-based classification

Neural networks are the base of deep learning techniques, which consist of two most commonly used variants, CNN and recurrent neural network (RNN). CNN enables the automatic extraction of features over the spatial domain via a multi-layer architecture, and widely used in computer vision tasks such as object detection (Ding et al., 2018). However, it treats each input independently and overlooks the temporal dependency among time-series data, and may not be sufficient for sequential problems such as activity recognition. On the other hand, RNN is designed to address sequential problems such as speech recognition. It performs the same operation for each element of a sequence, with the output of each element depending on the computations in previous elements. However, it is subject to the vanishing gradient problem in long-term dynamics (Donahue et al., 2017).

To address the vanishing gradient problem in classic RNNs, LSTM network was first created by Hochreiter and Schmidhuber (1997), which is capable of modeling temporal dependency among sequential features. LSTM networks have been applied to many sequential problems in computer vision such as activity recognition. They achieve better accuracy compared to traditional machine learning algorithms (Donahue et al., 2017; M. S. Ibrahim et al., 2016). In this study, LSTM networks were designed to capture the temporal dependency of positional and attentional cues between two entities, to determine whether they belong to the same group and recognize their involved activities.

To construct the input, time-sequential feature, the positional and attentional cues are concatenated into a 17-dimensional feature vector, denoted by $f_t^{i,j} = \left[P_{1t}^{i,j}, P_{2t}^{i,j}, P_{3t}^{i,j}, P_{4t}^{i,j}, P_{5t}^{i,j}, P_{5t}^{j,i}, A_{1t}^{i,j}, A_{1t}^{j,i}, A_{2t}^{i,j}, A_{3t}^{i}, A_{3t}^{j}, A_{4t}^{i}, A_{4t}^{j}, A_{5t}^{i}, A_{5t}^{j}, A_{6t}^{i}, A_{6t}^{j} \right]$. This feature vector describes the relationship between them at any given time. The time-sequential feature is constructed by chaining a series of time-variant feature vectors over a time period and denoted by $\{ f_t^{i,j}, f_{t+\Delta t}^{i,j}, f_{t+2\Delta t}^{i,j}, ..., f_{t+T}^{i,j} \}$, where t is the starting time, Δt is the sampling frequency and T is the time duration of observation. The temporal resolution Δt depends on the sampling frequency of the data used to extract the spatial and attentional states, such as the frame rate of construction videos. This time-sequential feature captures the dynamic interactions and temporal relationship between entities using both positional and attentional cues, and serves as the input to the LSTM models.

Figure 3.7 illustrates the LSTM network, which is composed of a number of LSTM cells ordered sequentially. The network takes the time-sequential feature $\{f_t^{i,j}, f_{t+\Delta t}^{i,j}, f_{t+2\Delta t}^{i,j}, ..., f_{t+T}^{i,j}\}$, simplified as $\{x_1, x_2, ..., x_n\}(n = T / \Delta t + 1)$, as input. Each feature vector is fed into its corresponding LSTM cell. All LSTM cells have the same structure that contains three gates—input gate, forget gate, and output gate—to control the flow of and modify the information in and out of the cell. At time step *t*, the feature vector x_t (the *t*th element within the time-sequential feature) is the input to the *t*th LSTM cell, and the cell state is updated through Equation 3.4.

$$\begin{cases} i_t = \delta\left(W_{xi}x_t + V_{hi}h_{t-1} + b_i\right) \\ f_t = \delta\left(W_{xf}x_t + V_{hf}h_{t-1} + b_f\right) \\ o_t = \delta\left(W_{xo}x_t + V_{ho}h_{t-1} + b_o\right) \\ g_t = \tanh\left(W_{xc}x_t + V_{hc}h_{t-1} + b_c\right) \\ c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \\ h_t = o_t \otimes \tanh(c_t) \end{cases} \tag{3.4}$$

δ is the sigmoid function, $\delta(x) = 1/$ （1+exp($-x$)）, where x_t is the input, h_t is the hidden state with *N* hidden units (*N*=17 in this study) and is also the output of this cell, c_t is the cell state, i_t, f_t, o_t are input gate, forget gate, and output gate at time *t* respectively. g_t is the input modulation that adds information to cell state. $\otimes$ represents element-wise multiplication. W_{xi}, W_{xf}, W_{xo}, W_{xc}, V_{hi}, V_{hf}, V_{ho}, W_{hc}, b_i, b_f, b_o, b_c, are the learnable parameters for each LSTM cell that control the level of information transferred from previous time steps as well as the level of information taken from the current time step.

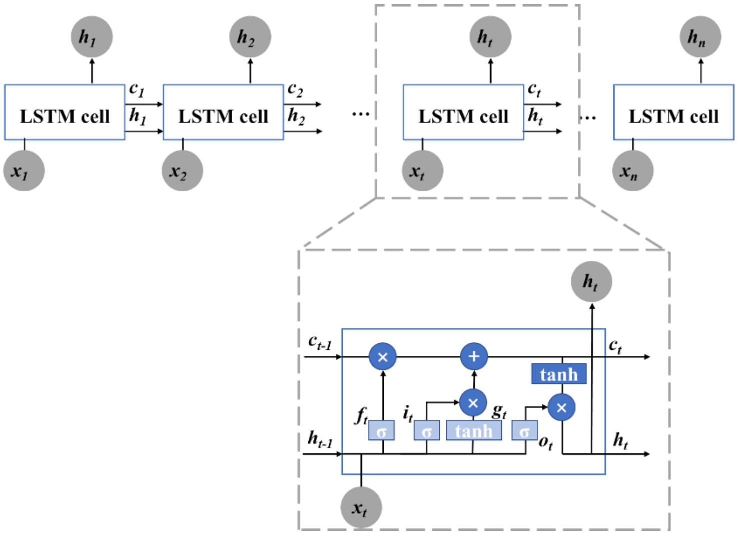


Figure 3.7 Typical architecture of LSTM network and LSTM cell

After feeding each feature vector to its corresponding LSTM cell and updating the cell states following Equation 3.4, the output of the last LSTM cell, an *N*-dimensional vector h_n, is then fed into a fully connected layer with the number of nodes equals to the number of class, n_c (e.g., $n_c = 2$ for working group identification). Finally, a softmax layer takes the output of the fully connected layer as input and computes the probability of being any particular class using a softmax function.

The network is trained by minimizing the cross-entropy loss function using *Adam* optimizer (Kingma & Ba, 2014). For working group identification, the problem is a binary classification: "1" means that the two entities are in one group, and "0" means that the two entities are not in one group. For group activity recognition, the problem is a multi-class classification: the number of classes equals the number of activities of interest.

3.4 Implementation

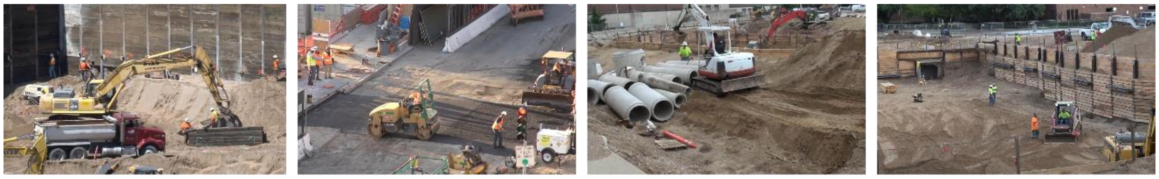
To validate the proposed framework, construction videos are used to extract the state information of individual entities, including the location, body orientation, head pose, and body pose of each entity. The positional and attentional cues are then computed from the extracted state information and used to train and test the LSTM networks. The proposed method is implemented on a workstation with 2.6GHz Intel Xeon CPU, 128GB RAM, and NVIDIA GeForce GTX 1060 6GB GPU.

3.4.1 Data description

The experiment consists of 14 videos from two sources: public-available website – YouTube (YouTube, 2019) and videos captured by the authors from construction jobsite on Purdue campus. The details of the video are illustrated in Table 3.6. The column of "included activity" in Table 3.6 indicates the activities to recognize in this study and will be discussed in Section 3.4.4. To evaluate the proposed method in general construction jobsites, the selected videos were taken from different construction scenarios with varying construction entities, working groups, and activities, and from different viewpoints with varying distances from objects ranging from 30m to 100m. Videos with two different resolutions are included, i.e., 1920x1080 and 1280x720. The average worker size in the dataset is 63x124, and average equipment size is 456x359. Since some of the videos were surveillance videos with low frame rate, all videos were downsampled to 2fps in this study. Figure 3.8 shows some sample images from the dataset.

Table 3.6 Data description

Data source	Video	Duration (s)	# of entity	# of groups	Average group size	Included activity
YouTube (2019)	1	210	11	4	2.25	spotting, road paving, others
	2	120	12	4	2.5	others
	3	30	3	1	2	others
	4	85	16	3	3.33	road paving, others
	5	45	6	2	3	road paving, others
	6	80	2	1	2	spotting
	7	75	4	2	2	spotting, others
	8	22	2	1	2	spotting
	9	65	4	1	4	spotting
	10	35	2	1	2	spotting
	11	110	2	1	2	spotting
	12	75	2	1	2	spotting
	Total	**952**	**66**	**22**	**2.5**	**spotting, road paving, others**
Taken by authors	13	52	2	1	2	spotting
	14	65	7	3	2	others
	Total	**117**	**9**	**4**	**2**	**spotting, others**



(a) (b) (c) (d)

Figure 3.8 Sample images (a-b from hospital project, c-d from teaching building project).

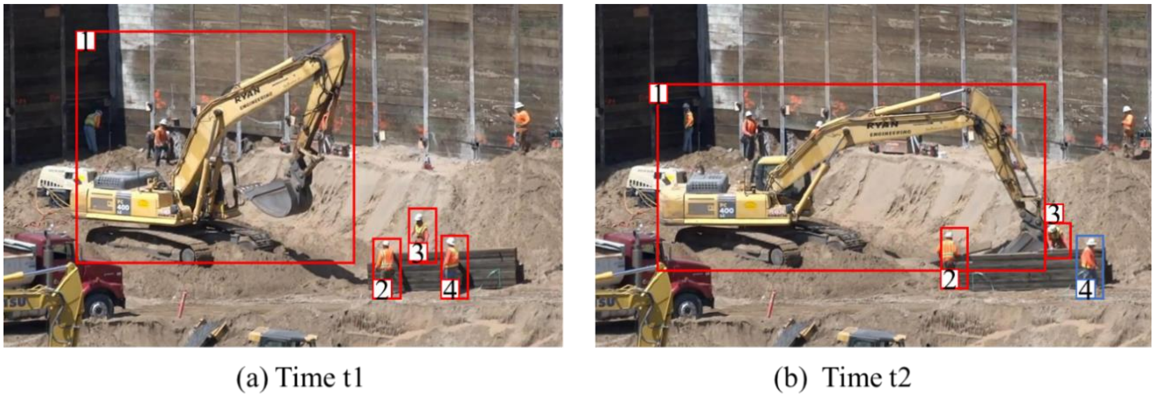
3.4.2 Data preprocessing

To obtain the positional and attentional cues presented in the construction videos, the relevant states of each entity including position, orientation, head pose, and body pose, need to be extracted. Since the focus of this study is to analyze the group activity using higher-level information, i.e., the features computed based on extracted states, manually annotated state information is used to compute the cues in order to exclude the impact from state detection. In this study, the states are manually annotated using image annotation tool, LabelImg. Specifically, in each frame, the bounding boxes were drawn around the entities and the pixel coordinates of the four corner points were extracted automatically by the annotation tool. The head pose, body orientation, and body pose were determined manually based on the representation illustrated in Figure 3.3 and annotated frame by frame. The annotation is based on the representation of spatial and attentional states described in Section 3.3.1, and performed by two researchers in the construction domain including the author. Part of the annotated data was double-checked by the author to ensure the consistency between different annotators and the annotation quality. In future study, crowdsourcing can be adopted to annotate a large amount of data and voting techniques (e.g., majority vote) can be used to ensure quality. If the annotation is domain-specific, domain experts can also be consulted to ensure the correctness.

Several studies in the computer vision domain have proven the feasibility of identifying attentional states from low-resolution images. For instance, Raza et al. (2018) have devised a CNN-based deep learning approach to identify head pose and body orientation of pedestrians. They tested their method on videos with a resolution of 640x480, and the input object size of their network is 64x64. Saleh et al. (2017) developed a multi-task learning network to recognize head pose and body posture of pedestrians with input image resolution being 227x227.

3.4.3 Working group identification

For working group identification, two entities are considered belonging to one working group if they are interacting with each other during the construction. Due to the dynamics of the construction site, the grouping results may vary with time. Therefore, the entire videos are carefully analyzed by the authors so that the working groups are accurately annotated as ground truth. Figure 3.9 shows an example of a working group and its changes over time. At time t_1, entity 1 – 4 were in the same group with the visual attention of the three workers on the excavator and the excavator moving to the workers to transport the materials. At time t_2, entity 1-3 were still in the same group and transporting the material, while entity 4 exits the group. As an example, in Figure 3.9, $G_{t1}^{2,4}=1$, and $G_{t2}^{2,4}=0$, where $G_t^{i,j}=0,1$ represents the group information between entity *i* and *j* at time *t* with 0 indicating "not a group" and 1 indicating "is a group". In this way, the proposed method is able to detect the start and finish time of the working group through continuous observation.

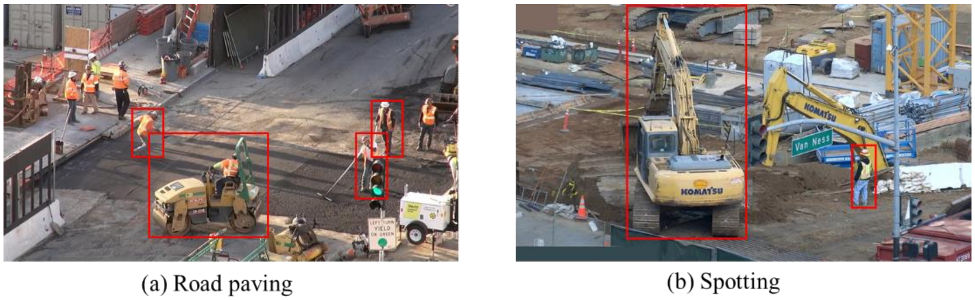


(a) Time t1 (b) Time t2

Figure 3.9 Change in working group (entities 1-4 belong to the same group at Time t1, while entity 4 exits the group at Time t2).

3.4.4 Group activity recognition

This study involves grouping entities that collaborate on one activity and recognizing the activity. Regardless of the type of activities, collaborating entities are always grouped. Whereas in activity recognition, construction activities are selected to be differentiated and explicitly classified based on the following two criteria. First, this study focuses on group activities that are cooperatively performed by workers and equipment as they typically involve complex and dynamic interactions. Second, as our motivation is to analyze the interactions in the group activities in order to prevent struck-by accidents, activities where workers are exposed to struck-by hazard are selected. In the dataset, the two activities that meet these two criteria are spotting and road paving (as shown in Figure 3.10), and therefore, these two activities are explicitly labeled and remaining group activities are labeled as “others”.


(a) Road paving

(b) Spotting

Figure 3.10 Two group activities classified in this study: (a) road paving, (b) spotting.

3.4.5 Training and test procedure

The proposed framework involves a two-step classification, i.e., for any pair of construction entities, they are first determined whether belonging to the same working group, and if so, their participated activity is further recognized. The two classifiers, referred to as group classifier and activity classifier, were first trained and evaluated separately, and then the entire process was tested using the trained classifiers. In the experiments, 5s (i.e., 10 frames) of observations were used to construct the time-sequential features described in Section 3.3.3. This time duration was then changed to assess the influence of available observations. The group/non-group information and the corresponding construction activities for each pair of construction entities were manually labeled to provide ground truth labels for supervised learning when training and testing the proposed two-step LSTM model. The annotation is based on the criteria described in Section 3.4.3 and 3.4.4. The training and test procedure are detailed as follows.

1. All videos in the dataset described in Figure 3.6 were trimmed into video clips with fixed length (e.g., 5s) that can start from an arbitrary frame of the original video. The sequential feature computed from one pair of entities in one video clip was considered as one data sample. From all available data obtained from video clips, 8,000 were randomly selected to train the

group classifier, which was then evaluated using another 2000 data that were randomly selected from the remaining dataset.

2. For the activity classifier, only data representing entities within the same group were used for training and evaluation. 8000 and 2000 data were randomly selected from all eligible data and used to train and evaluate the activity classifier. It is noted that in the experiments, the activity classifier can classify 3 types of group activities, i.e., road paving, spotting, and others.
3. After training the two classifiers, 2000 data were randomly selected to test the entire process. They were first fit to the group classifier and assigned the label “not a group” or “is a group”, and those with “is a group” label further went through the activity classifier with their group activities identified to be either road paving, spotting, or other group activities. The final performance was evaluated against the ground truth annotation.

3.4.6 Evaluation metrics

The performance of separate classifications, as well as the entire process, were evaluated quantitatively in terms of accuracy, precision, and recall, where accuracy represents the proportion of correctly classified instances among all instances; precision represents the proportion of true positive instances among all classified positive instances; recall represents the proportion of positive instances that have been correctly identified. The evaluation metrics are computed using Equation 3.5-3.7.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{3.5}$$

$$Precision = TP / (TP + FP) \tag{3.6}$$

$$Recall = TP / (TP + FN) \tag{3.7}$$

For working group identification, TP represents true positive, indicating that two entities that are within one group are correctly identified as one group. TN represents true negative, indicating that entities that are not within one group are correctly identified as not a group. FP represents false positive, indicating entities that do not belong to one group are wrongly identified as one group. FN represents false negative, indicating that entities within one group are wrongly identified as not a group. Correspondingly, accuracy measures the overall correctness in classifying two classes. Precision measures the proportion of predicted working groups being true working groups. Recall measures the proportion of true working groups being correctly recognized. Note that since group activity recognition and the entire process involve more than two classes, the precision and recall are calculated for each individual class, while accuracy is calculated for the overall performance.

3.5 Results

This section analyzes the results of group identification and activity recognition using the proposed method and compares the results with those obtained using positional cues alone. Moreover, the importance of working group identification and the influence of available observations are discussed. Finally, two additional construction videos were used to verify the efficacy of the method.

3.5.1 Performance of LSTM-based classification

Figure 3.11 demonstrates an example result of the proposed method, where blue solid boxes indicate the entities, and red dash boxes indicate the group activity recognition result. After performing the two-step LSTM-based classification, the entity pair 3 and 7 are labeled as class 2, indicating they are in the same group and performing road paving activity; the entity pair 1 and 2 are labeled as class 4, indicating they are in the same group and performing other activities (having a conversation in this case); and all other pairs of entities are classified as class 1, i.e., “not a group”.

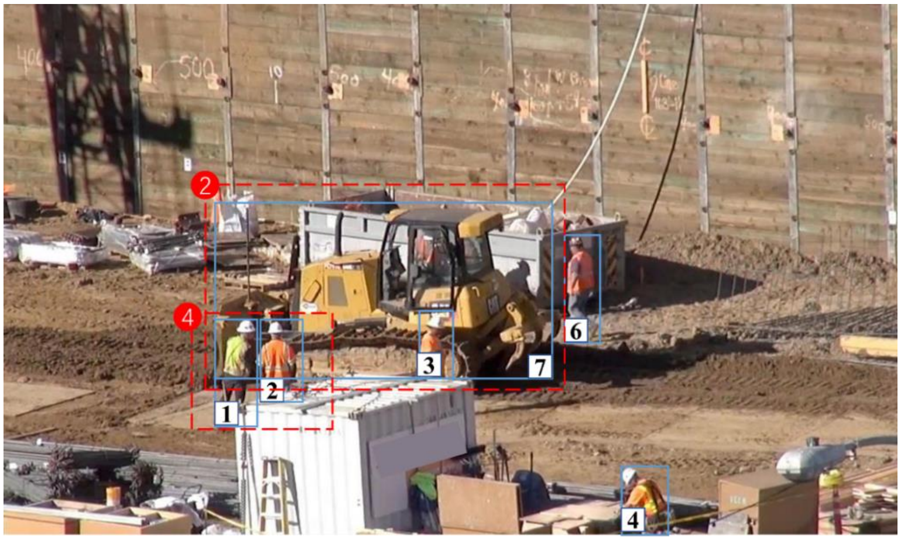
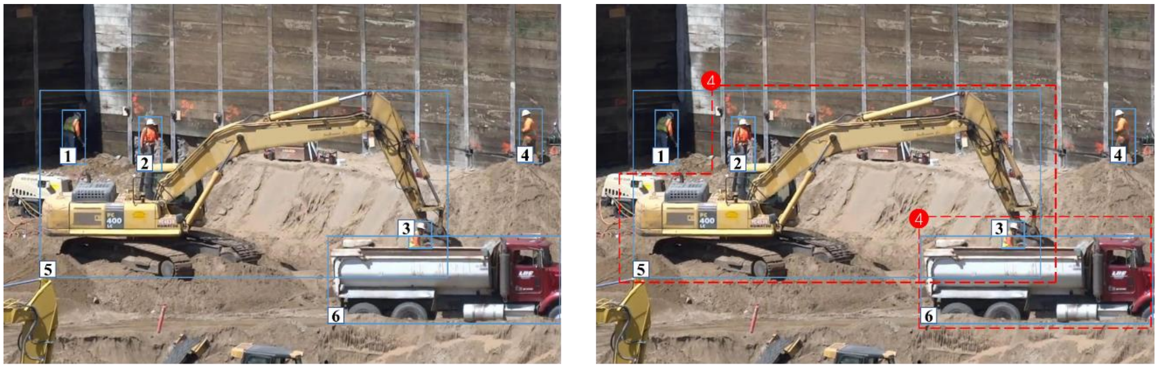


Figure 3.11 Example result of work group identification and activity recognition

Figure 3.12 illustrates the comparison of results with and without attentional cues. To perform experiments without attentional cues, an additional LSTM-based model was trained with only positional cues incorporated as inputs such that the difference between two experimental results reflects the influence of attentional cues. The ground truth is that none of the entities are working as a group. In the snapshot illustrated in Figure 3.12, the truck is very close to the excavator, but they are not interacting with each other to perform earth loading. Instead, the truck is passing through: it enters the scene from the left side, moves towards the right side, and directly leaves the scene without stopping. The ground truth is correctly identified by incorporating attentional cues (see Figure 3.12 (a)). However, in the case without attentional cues (see Figure 3.12 (b)), the entity pairs 2 and 5, and 3 and 6 were wrongly classified as working groups. This is because these two pairs are spatially closed and moving in similar directions. However, attentional cues indicate that they are not interacting with each other and therefore, when incorporated, the result aligns with the ground truth. This comparison scenario clearly demonstrates the value of using time-sequential features to exploit the evolution of entity interactions to better recognize construction activities.



(a) With attentional cues

(b) Without attentional cues

Figure 3.12 Comparison of results with and without attentional cues

Table 3.7 lists the quantitative performance of each individual task (i.e., group identification and activity recognition) as well as the integrated process. Results obtained using only positional cues were also listed as a comparison to illustrate the advantage of incorporating attentional cues. In general, the proposed method achieves over 95% accuracy in all tasks. For group identification alone, the recall is relatively low, indicating that about 91% of working groups can be correctly identified.

Table 3.7 Performance of LSTM-based classification

Task	Feature type	Class	Accuracy	Precision	Recall
Group identification	positional+ attentional	is a group	0.965	0.947	0.909
	positional	is a group	0.874	0.818	0.628
Activity recognition	positional+ attentional	road paving	0.996	1.000	0.992
		spotting		0.986	1.000
		others		0.997	0.997
	positional	road paving	0.891	0.951	0.951
		spotting		0.807	0.807
		others		0.884	0.884
Group identification + Activity recognition	positional+ attentional	not a group	0.963	0.971	0.983
		road paving		0.927	0.885
		spotting		0.948	0.892
		others		0.940	0.912
	positional	not a group	0.854	0.887	0.954
		road paving		0.725	0.446
		spotting		0.644	0.461
		others		0.729	0.631

It is found that the cases that fail to identify the groups mainly occur in road paving as well as when entities are entering or leaving a working group. Compared to spotting, the interactions in road paving are more complex and dynamic as they involve more entities with much larger span both spatially and temporally. For instance, as road paving requires the entity to move back and forth, it may result in the machine and worker moving in the opposite direction with their attention paid to the spot they are working on instead of their partners. As a result, if the observation period is short, it may classify the entities as not a group by mistake. Moreover, when an entity is entering or leaving a group as illustrated in Figure 3.9, it is very difficult to distinguish whether two entities belong to one group with limited period of observation, even for human experts. It is argued that the performance in both scenarios can be improved by incorporating longer previous observations, which is validated and discussed in Section 3.5.3.

For activity recognition alone, the proposed method achieves almost perfect results. There are two reasons for the very high accuracy, precision, and recall rates: (1) as a demonstration, the two activities selected in this study are relatively easy to distinguish, as they all involve one worker and one machine, with worker’s attention significantly different in each activity – the worker is visually focusing on the machine in the “spotting” activity, while the worker is primarily focusing on the ground and only occasionally on the machine in the “road paving” activity; (2) the activity recognition is trained and tested only on entities that are within one group. In other words, entities that are not performing the target group activity have already been filtered out in the group identification. Such result proves the importance of working group identification in group activity analysis, and it will be further analyzed in Section 3.5.2.

The integrated process (i.e., group identification + activity recognition) illustrates the performance of the proposed two-step classification. According to the analysis of two individual classifications, the performance of the integrated process mainly relies on the performance of working group identification. Since the recall rate for the working group is only 91%, the recall rates for the road paving and spotting activities are lower than other metrics.

Moreover, in all tasks, the performance of integrating positional and attentional cues is much higher than those only using positional cues in terms of accuracy, precision, and recall. Such results prove that attentional cues are critical in group activity analysis on the construction site and should not be ignored.

3.5.2 Influence of working group identification on activity recognition

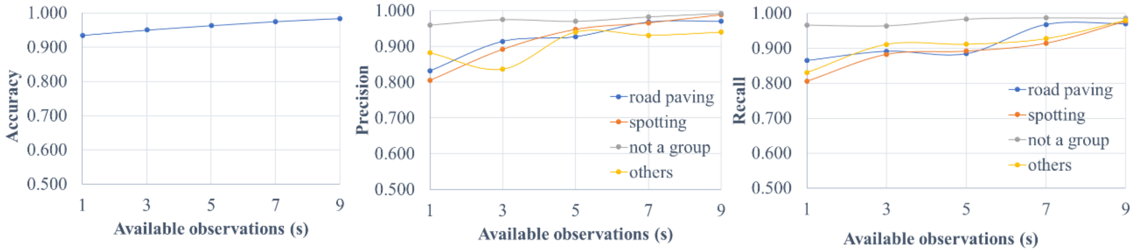
Due to the complex and dynamic nature of the construction site, multiple entities may co-exist and conduct different activities simultaneously. Dividing multiple entities into different working groups prior to activity recognition is expected to improve the performance of activity recognition. Table 3.8 compares the performance of activity recognition with and without working group identification. Although the accuracy of activity recognition is compatible, the precision and recall for the specific activities are much higher if the working group is first identified. This is because for single activity recognition without working group identification, most entities fall into the class of “others” no matter whether they are doing group activities or simply not interacting with each other. As a result, two target activities are not effectively recognized with the existence of so many “negative examples”. Hence, it is necessary and crucial to divide entities into several working groups in order to better analyze the working scenarios on the construction site.

Table 3.8 Comparison of activity recognition with and without group identification

Task	Class	Accuracy	Precision	Recall
Group identification + Activity recognition	not a group	0.963	0.971	0.983
	road paving		0.927	0.885
	spotting		0.948	0.892
	others		0.940	0.912
Activity recognition	road paving	0.971	0.925	0.854
	spotting		0.820	0.804
	others (including entities not within a group)		0.982	0.989

3.5.3 Influence of available observations

As the interaction between entities evolves with time, the longer previous observations are available, the better the pattern of their interaction can be revealed, and in turn, the better the working groups and corresponding group activity can be identified. This study assessed the performance of the integrated process with available observation varying from 1s to 9s, shown in Figure 3.13. Specifically, Figure 3.13 (a) illustrates the overall accuracy in correctly identifying working groups and corresponding group activities. Figure 3.13 (b)-(c) illustrate the precision and recall rates for each individual class.



(a) Accuracy

(b) Precision

(c) Recall

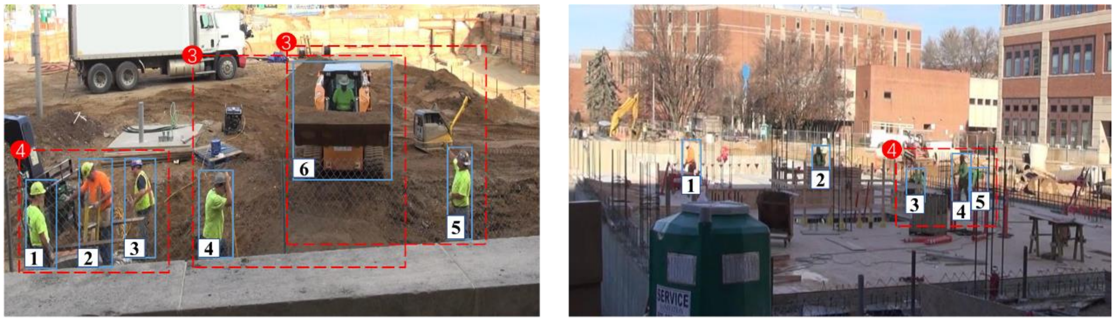
Figure 3.13 Performance (accuracy, precision, and recall) with respect to available observations

From Figure 3.13, the overall accuracy and the precision and recall for all classes increase as the length of available observations increases. It is reasonable because the longer previous observations are available, the better the dynamic interactions among entities can be revealed, leading to both higher precision and recall. However, on the construction site, longer previous observations require more accurate and reliable detection of individual entities' real-time states, which are not always available. Therefore, the duration of previous observation should be carefully selected by considering both target accuracy and the available sensing technologies.

3.5.4 Verification on additional scenarios

Two additional scenarios in building construction were used to verify the efficacy of the proposed method. The first scenario involves site work, as shown in Figure 3.14 (a), where worker 1, 2, and 3 are checking the manhole collaboratively, and worker 4 and 5 are spotting bulldozer 6 to unload the earth. The method successfully identified the group formed by 1, 2, and 3 and classified the activity as class 4 ("others"). Regarding the second group, the interactions between 4 and 6, and 5 and 6 were correctly identified and classified as class 3 ("spotting"), where 4 and 5 were determined as not in a group. This is because the proposed method is based on pairwise relationship, and 4 and 5 do not interact with each other directly, although both of them interact with 6.

The second scenario involves formwork and rebars, as shown in Figure 3.14(b), where worker 3 and 4 are tying rebars in one group, while 1 and 2 are setting formworks and 5 is tying rebars separately. This method successfully identified that worker 1 and 2 were working alone, and 3 and 4 were in the same group doing "other" activity. However, the system incorrectly classified 3 and 5 into one group. A possible reason is that worker 3 is facing worker 5 and they were spatially close in the 2D image plane, while their distance in the third dimension—the direction perpendicular to the image plan is ignored by using 2D cues. Such error could be mitigated by introducing 3D information in future research.



(a) Scenario 1: site work

(b) Scenario 2: formwork and rebars

Figure 3.14 Construction activity recognition in sample phases of building construction

3.6 Contributions

This study pioneers in incorporating attentional cues into the understanding of the construction jobsite context. It has three main contributions. First, a novel method has been created to numerically represent the states of individual entities and mathematically compute the positional and attentional cues that are regarded as critical features to recognize construction working groups and ongoing group activities. While videos were the single data format in this study, the identified features are at a higher level and not constrained to visual data. For instance, the positional cues can be extracted using real-time locating systems (RTLS) such as GPS and ultra-wideband (UWB), and the attentional cues can be inferred using inertial measurement unit (IMU) and eye-tracking technologies. Second, the proposed two-step process, i.e., working group identification followed by the activity recognition, allows the differentiation of group-relevant and non-relevant entities, making it capable of addressing complex group activities under general construction scenarios, where multiple entities co-exist on the job site. Third, this study adopts LSTM networks to model the temporal dependency among features, which allows the capture of complex and dynamic patterns of interactions on the construction site.

3.7 Conclusions

Construction sites involve multiple workers and equipment interacting with each other and conducting different activities simultaneously, making automatically recognizing diverse ongoing activities extremely challenging. This study proposes a two-step framework that decomposes the activity recognition into two cascading tasks – working group identification and group activity recognition. Novel methods are created to mathematically represent the spatial and attentional states of individual entities and compute the positional and attentional cues based on the pairwise relationship between two entities, which are further constructed as time-sequential features to identify working groups and corresponding activities. Given any pair of entities, LSTM networks are used to (1) classify whether they belong to the same groups, and (2) recognize the corresponding activity they are performing. Experiments were conducted using videos from a hospital construction project that are available online and videos from an ongoing teaching building project taken by the authors on campus. The results show the proposed framework achieves over 95% accuracy in correctly identifying the working groups and recognizing the activities. The performance obtained by integrating positional and attentional cues is much higher than that obtained using positional cues alone. Moreover, dividing the group activity recognition task into a two-step cascading process obtained better performance than simply conducting a one-step activity recognition. The newly created method was also tested on two additional construction scenarios, which further verified the efficacy of the method.

There remain some limitations that deserve further research efforts. First, the positional and attentional cues are computed based on manually annotated states of individual entities. Future study will focus on automating the entire process of state detection, group identification, and activity recognition. Second, due to data availability, the state information and corresponding cues are represented in 2D images, which may be sensitive to camera viewpoints. However, it is argued that by using the cues computed from pairs of entities that are observed from the same viewpoints, this impact is effectively mitigated. Besides, the incorporation of videos from different viewpoints also ensures the diversity and representativeness of the dataset, which in turn improves the generalizability of the proposed approach. Future research will extend the positional and attentional states and corresponding cues into 3D to further improve the robustness of the method. Third, this study only explicitly labels road paving and spotting while making all other group activities as “others”. Future study will extend this method to more settings as more construction videos are collected.

4. A CONTEXT-AUGMENTED DEEP LEARNING APPROACH FOR WORKER TRAJECTORY PREDICTION ON UNSTRUCTURED AND DYNAMIC CONSTRUCTION SITES

In this chapter, an LSTM model augmented by the context information is proposed, which incorporates both individual movement and workplace contextual information. Contextual information regarding movements of neighboring entities, working group information, and potential destination information is concatenated with movements of the target entity and fed into an LSTM network with an encoder-decoder architecture to enable the sequence-to-sequence prediction, i.e., a sequence of estimated positions is generated from a sequence of observations. The method is validated using videos collected on three construction sites—one hospital construction project and two teaching building construction sites. Visual data are pre-processed to extract entity positions and contextual features, which are then used as inputs to train and test the proposed method. The trajectory prediction is performed on the 2D image plane. The accuracy of prediction achieved by this method is 8.51 pixels in terms of final displacement error, with an observation time of 3s and prediction time of 5s. It was found that integrating contextual information with target movement information can result in a smaller final displacement error, especially when the length of prediction is longer than the length of observation. Compared with the conventional method that predicts position only one step ahead, the proposed method predicts trajectories over multiple steps following a sequence-to-sequence architecture and consequently, eliminates the error accumulation issue.

This work is under review in Advanced Engineering Informatics, 2020, Jiannan Cai, Yuxi Zhang, Liu Yang, Hubo Cai, and Shuai Li. "*A Context-Augmented Deep Learning Approach for Worker Trajectory Prediction on Unstructured and Dynamic Construction Sites*". Table titles and figure captions have been modified to maintain the form of the dissertation.

4.1 Introduction

The construction industry is one of the most dangerous industries: it employs only 5% of the US workforce (U.S.Bureau of Labor Statistics, 2018) but accounts for 21.1% (1008 deaths) of the total worker fatalities in 2018 (OSHA, 2018). The struck-by accident is a major cause, leading to 804 worker fatalities (18%) in construction from 2011 to 2015 (X. S. Dong et al., 2017). It is also a single leading cause for non-fatal injuries, accounting for 34% of cases of injuries from 2011 to 2015 (US Department of Labor, 2016). To prevent struck-by accidents, previous studies (Marks & Teizer, 2013; Teizer et al., 2010; Teizer & Cheng, 2015) focused on determining the proximity between workers and equipment using sensing technologies and comparing the proximity to predefined thresholds to detect struck-by hazards. Low detection accuracy and reliability are the main challenges attributed to the difficulty in predicting the future movements of jobsite entities while considering the uncertainties of their movements on the unstructured and dynamic construction sites. For instance, warning systems can raise 59% false alarms due to the uncertainty in proximity analysis (Ruff, 2006). As a result, workers may lose confidence in and ignore the alarms, which hinders the efficacy of struck-by prevention systems. According to Luo et al. (2017), the estimated response rate of proximity warning systems for generic hazards is about 0.528. Under such a situation, accurate prediction of worker trajectory provides additional information and is critical to achieving a proactive and informative struck-by prevention system.

Existing studies have created a few methods to predict trajectories of construction resources. Zhu et al. (2016a) proposed a novel Kalman filter to predict the movements of workers and mobile equipment using positions obtained from multiple video cameras. Dong et al. (2018) and Rashid et al. (2018) modeled the worker movements as a Markov process to predict their trajectories based on historical records. However, one main challenge in trajectory prediction of construction entities is the low accuracy over large time horizons because of two interrelated reasons. First, it is insufficient to only consider the previous movements of individual entities when predicting their future trajectories. Since multiple entities co-exist on the construction site, forming various working groups to accomplish different activities, their behavior will be influenced by each other and the specific activities they are involved in. To accurately predict worker trajectory, such contextual information must be incorporated. Second, due to the complex and dynamic jobsite context, it is not adequate to capture the worker movement using a pre-defined model with hand-crafted features that may only fit particular scenarios.

A few recent studies (D. Kim et al., 2019; Tang et al., 2019) attempted to predict the construction entity trajectory through a data-driven approach given the advances in deep learning techniques. Despite the promise of deep learning, the rich contextual information regarding working groups and involved activities on construction jobsites have not been fully exploited to better predict worker's trajectory under various construction scenarios. Towards that end, this study proposes a long short-term memory (LSTM)-based, context-augmented deep learning model that integrates both individual movement information and contextual information, including movements of neighboring entities, working group information, and potential destination information. In addition, the proposed method adopts a sequence-to-sequence (seq2seq) neural network architecture that allows the elimination of error accumulation in prediction trajectories over multiple time steps.

4.2 Review of Related Studies

In this section, related studies on proximity-based struck-by prevention and trajectory prediction are reviewed and their limitations are outlined.

4.2.1 Related studies on proximity-based struck-by prevention

Struck-by accident is one of the leading causes of construction fatalities and has attracted increasing research interest. Many studies developed prevention mechanisms to provide alerts when workers and equipment are too close to each other. Most of them compare the proximity information detected via various real-time locating systems (RTLS) with a pre-defined threshold and provide early warnings when the distance is less than the threshold (Marks & Teizer, 2013; Teizer et al., 2010; Teizer & Cheng, 2015). To adapt to different working states of equipment, Vahdatikhaki and Hammad (2015) proposed a method that generates the smart working zone of equipment to prevent the struck-by accident considering pose, geometry, and speed of the equipment. In addition to the analysis of the pairwise relationship, Wang and Razavi (2018) assessed the struck-by risk at a network level and integrated the proximity, blind spot information, and velocity into the decision on risk level.

A major limitation of these studies is that the decisions are made, and early warnings are provided from a deterministic perspective. Due to the sensing errors and the subjectivity in determining the threshold, such approaches will generate numerous false alarms, resulting in less confidence of the site personnel in the warnings. To reduce the false alarm, Kim et al. (2015) leveraged a fuzzy inference approach considering the proximity and crowdedness (i.e., number of entities). Wang and Razavi (2016a, 2016b) developed different rules for different working scenarios given proximity, direction, and speed information. Edrei and Isaac (2017) integrated the inaccuracy of the tracking system and created statistical zones based on the probability of struck-by hazards. Despite the great research efforts, current approaches detect struck-by hazards and take actions “just” before potential accidents might happen with limited prediction ability, which has a large chance of interrupting normal operation and making incorrect warnings. Therefore, there is a critical need for accurate prediction of worker trajectory, which paves the way for a proactive and informative struck-by prevention mechanism.

4.2.2 Related studies on trajectory prediction

Different approaches in trajectory/intention prediction

Trajectory/intention prediction is an essential yet challenging task in the computer vision community and has been increasingly studied in applications such as pedestrian behavior analysis due to the emergence of autonomous vehicles. There are typically three types of approaches in trajectory/intention prediction, i.e., model-based, planning-based, and data-driven approaches.

Model-based approaches explicitly model the movement dynamics as mathematical models. Conventionally, tracking filters are used to predict the future steps in a trajectory (Hermes et al., 2009; T. Liu et al., 1998; Prévost et al., 2007). For instance, the Kalman filter is applied to predict the trajectory using a Gaussian distribution with accumulated uncertainty. However, this approach often results in physically impossible locations (e.g., behind walls, within obstacles). Koojj et al. (2019) modeled pedestrian movement as a Switching Linear Dynamical System which considers different motion states, e.g., moving and stop. The Kalman filter is applied for prediction in the moving state, and the position remains unchanged in the stop state.

Most studies on trajectory prediction in the construction domain fall in this category. For instance, Dong et al. (2018) and Rashid et al. (2018) modeled the worker movements using a hidden Markov Model and predict their trajectories based on historical records. Zhu et al. (2016a) adopted the Kalman filter to predict the movements of workers and mobile equipment from positions obtained from stereo cameras. One main drawback is that the model-based approach relies on simplified dynamics models and hand-crafted states with parameters estimated from historical records/observations, which may only fit particular scenarios and simple movements. Moreover, it treats entities as objects and only considers movement patterns, which works well in the short-term prediction, but may degrade into random walks over large horizons (Ziebart et al., 2009).

Planning-based approaches treat entities as intelligent agents who actively plan their motion/path to achieve a goal. The problem is formulated as a path planning or optimal control task, such as the Markov decision process (MDP). The optimal policy is determined by maximizing some inherent reward functions. For instance, Ziebart et al. (2009) and Kitani et al. (2012) modeled the goal-directed human trajectories using a maximum entropy inverse optimal control method and incorporated environment features in the cost function to determine the optimal path an entity will select. Karasev et al. (2016) modeled pedestrian behavior as a jump-Markov process and the goal as a hidden variable. The reward function is formulated from the semantic map of the environment. Then, the trajectory is predicted by obtaining the optimal strategy that maps the goal to actions. Rudenko et al. (2018) integrate the MDP and social force model, where MDP is used for long-term prediction and social force is used to update the short-term states. One main drawback is that the planning-based approach relies heavily on prior knowledge, and it still uses hand-crafted features to model states and reward functions that are specific to particular settings.

Recently, with the advances in deep learning techniques, the data-driven approach has been increasingly used given that it does not require explicitly modeling movement dynamics and that it can be generalized to various scenarios. The problem is usually formulated as a time-series classification or regression problem. For instance, Völz et al. (2016) predicted pedestrian intention of crossing or not crossing the crosswalks using three data-driven models, i.e., deep neural network, LSTM, and Support Vector Machine (SVM). Saleh et al. (Saleh et al., 2018) predicted lateral positions of pedestrians using three stacked layers of LSTM networks.

Context-aware prediction

Traditionally, only past movements of individual entities are used as inputs to predict future trajectory, which is insufficient to capture human behavior under different scenarios, especially when human behavior is influenced by the environment. Recent studies in the computer vision community have recognized the significance of context information and considered various contextual features to predict pedestrian trajectory and intention on the road. For instance, Kooij et al. (2019) found that incorporating pedestrian situational awareness, situation criticality, and spatial layout of the environment increases the prediction accuracy. Alahi et al. (2016) created a social-LSTM model and proved that the pedestrian trajectory can be better predicted by incorporating the interaction among multiple pedestrians. Xue et al. (2018) and Syed and Morris (2019) incorporate the occupancy map and scene features in the trajectory prediction.

Very few studies have incorporated the contextual information in trajectory prediction in the construction domain. Kim et al. (2019) applied a hyper-parameter tuned Social Generative Adversarial Network to predict trajectories of construction entities in 5s. Tang et al. (2019) developed an LSTM network that integrates entity type (i.e., worker and equipment) and occupancy maps of the construction site to prediction entity trajectory in up to 2s. Despite these pilot studies, the trajectory is predicted only in one specific job setting with entities conducting a specific activity. There remains a critical need to exploit the contextual cues that are effective to predict the entity trajectory under general construction jobsite scenarios. To close this gap, this study proposes an LSTM-based, context-augmented model that integrates both individual movement information and contextual information, including movements of neighboring entities, relationship with neighboring entities (i.e., within one group or not in one group), and potential destination, to accurately predict trajectory of construction workers.

4.3 Methodology

In this study, a context-aware LSTM-based method has been designed to accurately predict worker trajectories using visual data that contain rich contextual information. Entity movement and contextual information are incorporated in the LSTM-based seq2seq neural network for trajectory prediction. Figure 4.1 illustrates the overall framework. This method consists of two major steps: Step 1—contextual information formulation and Step 2—LSTM-based seq2seq trajectory prediction.

In the first step, contextual information regarding the interaction between the entity and its nearest neighbor, and the involved construction activity is considered. Specifically, the contextual information is represented by three features, the neighbor position, the relationship with the neighbor (i.e., group/not a group), and the distance from potential estimation. In Chapter 3, it was found that the interactions among construction entities can be modeled using positional and attentional cues and further used to reason about the construction working group and corresponding group activity. This forms the technical foundations to formulate the contextual features in this study. In the second step, the above features are concatenated and fed into an LSTM encoder that encodes the information regarding both entity movements and jobsite context during the observation time. The encoded information is then fed into an LSTM decoder that generates a sequence of estimated positions during the prediction period. In this way, the proposed method takes into account the construction job contextual information and avoids the error accumulation when predicting trajectory over multiple time steps.

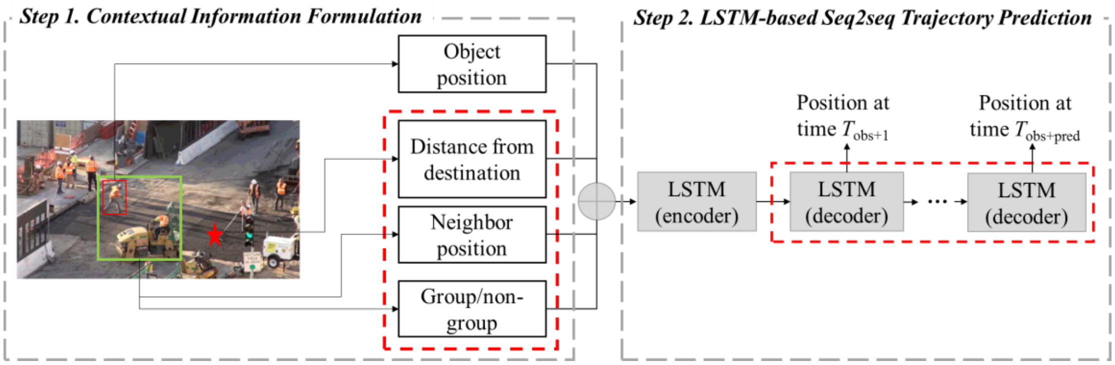

Step 1. Contextual Information Formulation
Object position
Distance from destination
Neighbor position
Group/non-group
Step 2. LSTM-based Seq2seq Trajectory Prediction
Position at time T_{obs+1}
Position at time $T_{obs+pred}$
LSTM (encoder)
LSTM (decoder)
...
LSTM (decoder)

Figure 4.1 Context-aware LSTM method for construction worker trajectory prediction

4.3.1 Problem formulation

In this study, the entity position is captured by the mid-bottom point of its bounding box on the 2D image plane. As a result, at any time step *t*, the i^{th} entity on the jobsite is represented by its pixel coordinates on the image plane, i.e., (x_t^i, y_t^i), where the superscript refers to the i^{th} entity, subscript refers to the time step *t*, and *x* and *y* represent the 2D pixel coordinates. The objective is to predict the entity positions from time step T_{obs+1} to $T_{obs+pred}$ based on the observation of site dynamics, including both the positions of all entities and the jobsite contexts from time step 1 to time step T_{obs}. Different from previous studies (Alahi et al., 2016; Tang et al., 2019) which only observe entity positions and implicitly incorporate the interactions among entities using hidden states learned from deep neural networks, this study explicitly models the contextual information (including entity interaction and involved activity) on the jobsite based on the methods developed in Chapter 3. Note that it is assumed the visual data are first preprocessed to obtain entity positions and contextual features, consistent with most of the related studies (Alahi et al., 2016; D. Kim et al., 2019; Tang et al., 2019; Xue et al., 2018).

4.3.2 Contextual information formulation

Construction entities (including both workers and equipment) interact with each other, constituting working groups to accomplish assigned tasks. It is expected that the worker's behavior will be influenced by other entities as well as the involved construction activity. The rationale is that construction workers tend to avoid obstacles to prevent potential collisions, while staying close to their co-workers or group members to conduct the activity collaboratively. Meanwhile, worker's movement is typically within the workspace specified by their involved activity, which indicates their potential destination. The specific contextual features considered in this study include neighbor position, group relationship with neighbor, and distance to potential destination.

Neighbor position

It is not uncommon that the positions of other entities in the scene are incorporated to reflect their interactions with the target entity when predicting its trajectory. A conventional approach is to construct an occupancy map of the scene or within a certain area of the target entity to represent the existence of other entities (Alahi et al., 2016; Tang et al., 2019). A main drawback is that if the grid size is large, resulting in coarse occupancy map, the dynamic changes of entity positions cannot be effectively reflected, especially when entity movement is not substantial across consecutive time steps, such as on construction sites; if the grid size is small, resulting in fine occupancy map, only a few grids will be occupied by entities, which leads to very sparse occupancy map, i.e., most values are zero.

In contrast, this study directly uses neighbor position information as one contextual feature. Note that, only the position of entity's nearest neighbor is considered in order to ensure the same dimensional features in different scenarios. It is reasonable as an entity is more likely to be affected by others who are spatially closer to them. Because position information of all entities (from 1 to N) are observed at each time step, the position of the nearest neighbor for entity i can be easily denoted as $\left(x_t^j, y_t^j\right), j = \arg\min \left\| \left(x_t^i - x_t^k, y_t^i - y_t^k\right) \right\|, k \in 1 \ldots N, k \neq i$.

Group relationship with neighbor

In addition to the neighbor position, the relationship between an entity and its neighbor in terms of whether or not they belong to the same working group also influences entity movement.

For instance, workers tend to avoid entities that are not in the same group to prevent potential conflict, while they tend to have similar movement patterns with their co-workers. However, such scenarios are not differentiated, and the group information has been overlooked in current studies.

The group relationship between an entity and its nearest neighbor is considered as a second contextual feature: if they belong to the same working group, the feature value is 1, otherwise, it is 0. The group information can be obtained using the method created in Chapter 3, as illustrated in Figure 4.2.

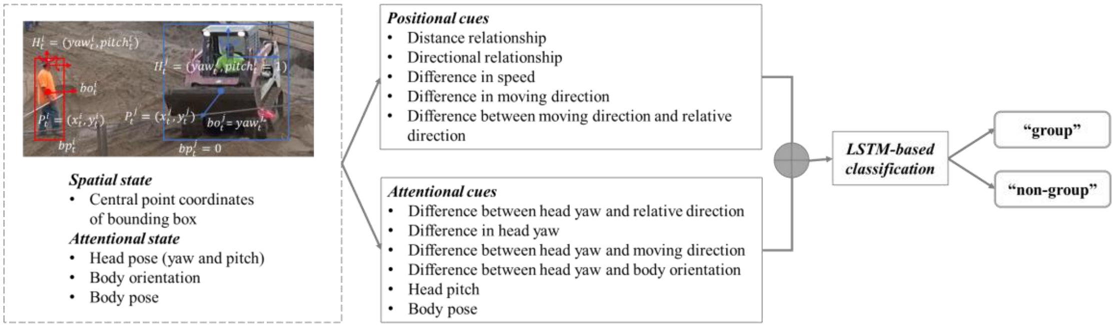

$H_t^i = (yaw_t^i, pitch_t^i)$
$H_t^j = (yaw_t^j, pitch_t^j = 1)$
bo_t^i
$P_t^i = (x_t^i, y_t^i)$
$P_t^j = (x_t^j, y_t^j)$
$bo_t^j = yaw_t^j$
bp_t^i
$bp_t^j = 0$
Spatial state
• Central point coordinates of bounding box
Attentional state
• Head pose (yaw and pitch)
• Body orientation
• Body pose
Positional cues
• Distance relationship
• Directional relationship
• Difference in speed
• Difference in moving direction
• Difference between moving direction and relative direction
Attentional cues
• Difference between head yaw and relative direction
• Difference in head yaw
• Difference between head yaw and moving direction
• Difference between head yaw and body orientation
• Head pitch
• Body pose
LSTM-based classification
"group"
"non-group"

Figure 4.2 Construction working group identification

In Chapter 3, an LSTM-based method for working group identification was created by integrating positional and attentional cues between construction entities. First, spatial and attentional states of construction entities from construction videos are represented as numerical values. The spatial state refers to an entity's real-time position on the image plane and the attentional state refers to the direction of an entity's visual attention, captured by head pose, body orientation, and body pose. Then, positional and attentional cues are computed from the spatial and attentional states of two entities to model their interaction. Five positional cues (i.e., distance relationship, directional relationship, difference in speed, difference in moving direction, and difference between moving direction and relative direction), and six attentional cues (i.e., difference between head yaw and relative direction, difference in head yaw, difference between moving direction, difference between head yaw and body orientation, head pitch, and body pose) are modeled as critical features. Finally, LSTM-based classification is performed to determine whether two entities belong to the same working group based on the time-series positional and attentional cues.

Distance to potential destination

On construction sites, worker behavior is goal-based and purposeful, motivated by their involved activities. It is expected that the worker will inherently move towards the potential destination, and thus, distance between worker's current position and the potential destination is treated as a third contextual feature, denoted as $\left(\Delta x_t^i, \Delta y_t^i\right) = \left(\left|x_t^i - x^{dest}\right|, \left|y_t^i - y^{dest}\right|\right)$. Note that it is assumed the destination is time-invariant during a short period of time, and the distance to the destination is used as a contextual feature to incorporate the temporal dynamics. This study simplifies the destination as prior knowledge to examine its influence on worker trajectory prediction. In practice, the potential destination can be inferred from the involved activity and the corresponding workspace, where ongoing activity can be automatically learned from visual data and workspace can be acquired from site layout or a building information model.

4.3.3 LSTM-based sequence-to-sequence (seq2seq) trajectory prediction

LSTM network (Hochreiter & Schmidhuber, 1997) is a typical recurrent neural network (RNN) and can be used to model temporal dependency among sequential features. It has been successfully applied to many sequential problems such as natural language translation and activity recognition. Figure 4.3 illustrates a typical LSTM network that takes time-series features $\{x_1, x_2, ..., x_n\}$ as input. The LSTM network consists of several cells ordered sequentially, each of which has the same structure with three gates, i.e., input gate, forget gate, and output gate, to control the information flow within the cell. At time step *t*, the cell state is determined by both the input of the current time step and the output from the previous time step, updated using Equation 4.1.

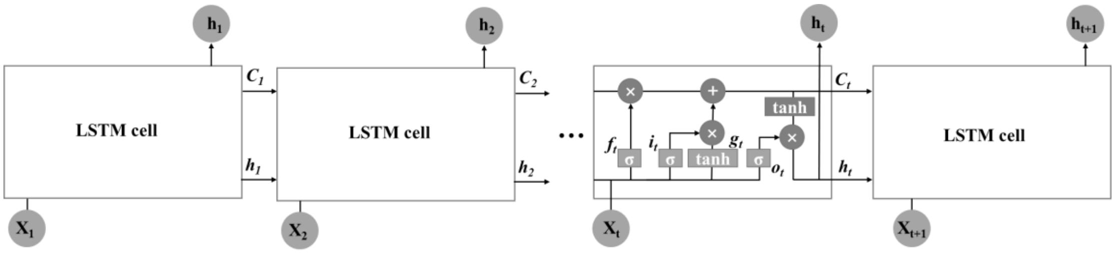


Figure 4.3 Typical structure of LSTM network

$$\begin{cases} i_t = \delta\left(W_{xi}x_t + V_{hi}h_{t-1} + b_i\right) \\ f_t = \delta\left(W_{xf}x_t + V_{hf}h_{t-1} + b_f\right) \\ o_t = \delta\left(W_{xo}x_t + V_{ho}h_{t-1} + b_o\right) \\ g_t = \tanh\left(W_{xc}x_t + V_{hc}h_{t-1} + b_c\right) \\ c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \\ h_t = o_t \otimes \tanh(c_t) \end{cases} \tag{4.1}$$

Where x_t is the input, i_t, f_t, o_t are input gate, forget gate, and output gate at time t respectively. h_t is the hidden state with N hidden units (N=25 in this study) and is also the output of this cell, and c_t is the cell state. g_t is the input modulation that adds information to cell state. δ is the sigmoid function and $\otimes$ represents element-wise multiplication. W_{xi}, W_{xf}, W_{xo}, W_{xc}, V_{hi}, V_{hf}, V_{ho}, W_{hc}, b_i, b_f, b_o, b_c, are the learnable parameters for each LSTM cell that control the level of information transferred from previous time steps as well as the level of information taken from the current time step.

Recently, LSTM network has been widely used in data-driven trajectory prediction. A conventional approach (Alahi et al., 2016; Saleh et al., 2018) is that the observations from time step 1 to T_{obs} are fed into the LSTM network (as shown in Figure 4.3) and the position in the next time step T_{obs+1} is estimated using the output of the last LSTM cell. Then, the estimated position at time T_{obs+1} is used as input along with observations from time 2 to T_{obs}, to predict for time T_{obs+2}, which happens recursively till $T_{obs+pred}$. Under such a case, the model only predicts one step each time and the predicted result is used as inputs recursively in order to generate a sequence of positions over multiple time steps. This practice leads to large error accumulation.

To solve this problem, this study adopts the LSTM encoder-decoder architecture, which allows the generation of a sequence with arbitrary length from a given sequence and was first introduced in machine translation tasks (Sutskever et al., 2014). Figure 4.4 illustrates the proposed model. In the method, the entity position during observation time and the corresponding contextual features (discussed in Section 4.3.2) are concatenated into time-series feature vectors and fed in LSTM encoder. The encoder outputs an encoded vector (i.e., the hidden state of the final encoder LSTM cell) that encapsulates the information from the observed movements and jobsite context. The encoded vector is used to initialize the states in LSTM decoder which allows the integration of previous information for better prediction of future trajectory. The hidden state of each LSTM cell in the decoder is considered as the output of the corresponding time step, which is further fed into a dense layer with two nodes. The dense layer essentially performs a linear regression, resulting in estimated positions from time T_{obs+1} to $T_{obs+pred}$.

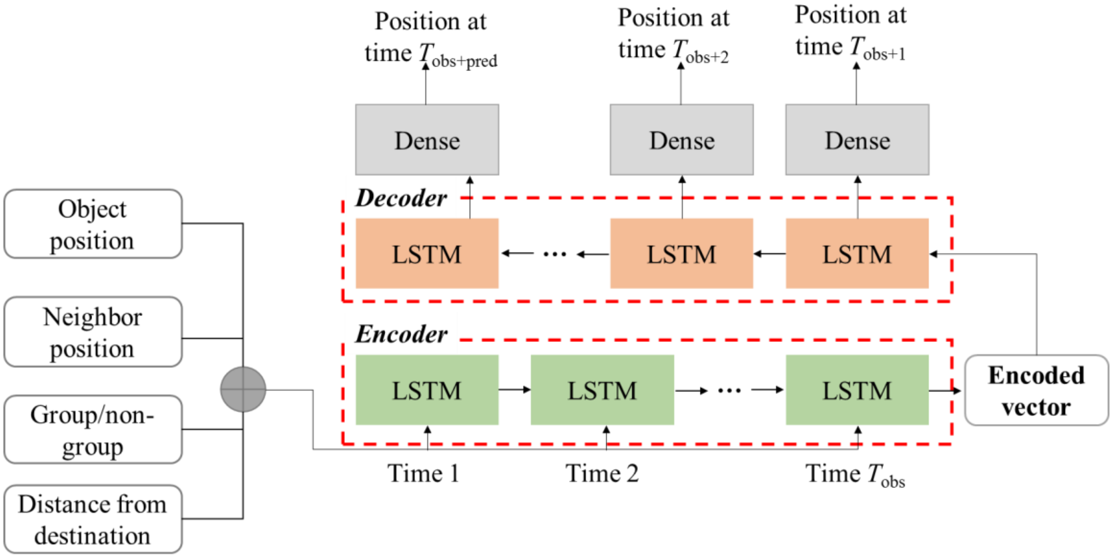


Figure 4.4 Context-aware LSTM-based seq2seq model

Similar to Saleh et al. (2018), the network is trained by minimizing one of the most commonly used loss functions, i.e., mean squared error (MSE) loss function (T. Lee, 2007), using *Adam* optimizer (Kingma & Ba, 2014). The MSE is computed as $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2$, where N is the size of training data, $\hat{Y}_i$ and Y_i are the predicted and actual i^{th} trajectory.

4.4 Experiments

The dataset used to test the proposed method is introduced and the implementation details are described. Two evaluation metrics are also explained to assess the prediction performance.

4.4.1 Data description

To demonstrate the proposed method, ten construction videos were collected from three projects: a hospital construction project from the publicly-available website YouTube (YouTube,

2019) and two teaching building projects taken by authors on the campus of Purdue University and the University of Tennessee, Knoxville (UTK), respectively. The videos consist of a total of 84 workers in different construction scenarios, conducting various activities in different working groups. All videos were down-sampled to 2fps, considering the various frame rates of videos in the dataset, which is also compatible to other studies (Alahi et al., 2016; Xue et al., 2018) on pedestrian trajectory prediction using surveillance video. Note that choosing a low frame rate (e.g., 2 fps) will improve the processing time but lose some information compared to a high frame rate (e.g., 30 fps). It is reasonable to select a relatively low frame rate when the speed of the entity is not very fast, such as on construction sites. Figure 4.5 illustrates some images from the dataset.

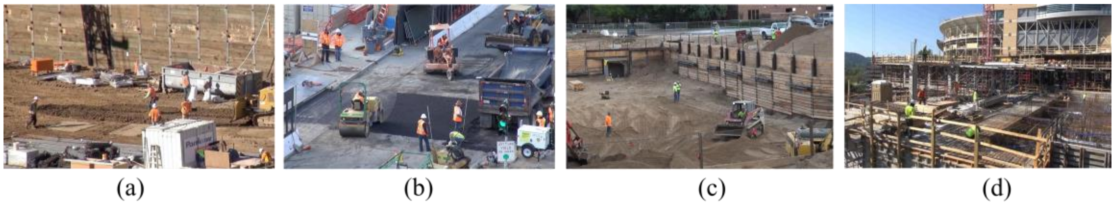


Figure 4.5 Sample images ((a)-(b) from hospital project, (c) from teaching building project on Purdue campus, (d) from teaching building project on UTK campus)

4.4.2 Data preparation

Visual data were pre-processed to extract entity positions and contextual features, which are then used as inputs to train and test the proposed method. First, all entities (workers and equipment) are manually annotated using bounding boxes with pixel coordinates of the mid-bottom points representing their positions on the images. The image annotation tool, LabelImg, was used to annotate entities on frames. Second, the nearest neighbor of each worker is identified by computing the distances between any two entities. It is noted that only workers are considered as target entities for trajectory prediction, however, the neighboring entity may include both workers and equipment. Third, two entities are considered belonging to one working group if they are interacting during the construction, and are labeled as "1". Otherwise, they are considered working independently, labeled as "0". As explained in Section 4.3.2, this information can be automatically obtained from positional and attentional cues using the method created in Chapter 3. However, to validate the influence of group information on trajectory prediction, the annotated group information is used. Finally, the potential destination of workers is determined as their final position in the scene, based on which the dynamic distance from worker to the potential destination is computed in both *x* and *y* directions. As the purpose of this study is to examine the influence of contextual information on trajectory prediction, the potential destination is simplified as prior knowledge.

As a result, a total of 241 trajectories with various lengths were obtained for 84 workers. The length of observation was set as 3s (i.e., 6 frames) and prediction length as 5s (i.e., 10 frames), consistent with relevant studies (Alahi et al., 2016; Xue et al., 2018) on pedestrian trajectory prediction. Correspondingly, the 241 trajectories were trimmed into tracks using a sliding window with a fixed length of 8s (i.e., 16 frames). To augment the dataset, the sliding window starts from every other frame of the original trajectory, resulting in 3640 tracks (tracks that are less than 16 frames were excluded).

4.4.3 Implementation details

The proposed method is implemented using Keras library on top of Tensorflow platform, on a desktop with 3.6GHz Intel i9-9900K CPU, 32GB, and NVIDIA GeForce GTX 2080 Ti GPU. The dataset is randomly split into training set (80%), evaluation set (10%), and testing set (10%). The network is trained with *Adam* optimizer, with a learning rate of 0.001, batch size of 20, and dropout of 0.5. To further prevent overfit, early stopping criterion is used. Specifically, the network is trained on the training set, and if the accuracy on the evaluation set does not increase for 100 epochs, the model will be terminated and the checkpoint that leads to the highest accuracy on the evaluation set will be saved. The trained model will then be tested on the testing set to evaluate the performance.

4.4.4 Evaluation metrics

The model is evaluated using two metrics:

(a) *Average final displacement error (FDE)*: The MSE between the final predicted location and the final actual location of all testing data, computed as $FDE = \frac{\sum_{i=1}^{N} \left\| \hat{y}_T^i - y_T^i \right\|}{N}$, where N is data size, $\hat{y}_T^i$ is the final predicted location for i^{th} data, and y_T^i is the final actual location for i^{th} data.

(b) *Average displacement error (ADE)*: The MSE over all locations of predicted trajectories and the actual trajectories, computed as $ADE = \frac{\sum_{i=1}^{N}\sum_{t=0}^{t=T}\left\|\hat{y}_t^i - y_t^i\right\|}{N \times T_{pred}}$, where T_{pred} is the prediction duration.

4.5 Results

4.5.1 Quantitative prediction results

The result of the proposed method is compared with that obtained using two other data-driven models: (1) a baseline model that recursively predicts trajectory based on object positions; and (2) a seq2seq model that predicts trajectory over multiple time steps simultaneously based on object positions. Figure 4.6 illustrates two example results of trajectory prediction. The proposed method results in the predicted trajectory (blue line) closest to the ground truth (red line). The position-based seq2seq model (yellow line) leads to a trajectory with a slightly larger discrepancy compared to the proposed method. In contrast, the position-based recursive model (green line) has the largest discrepancy from ground truth trajectory due to the error accumulation.

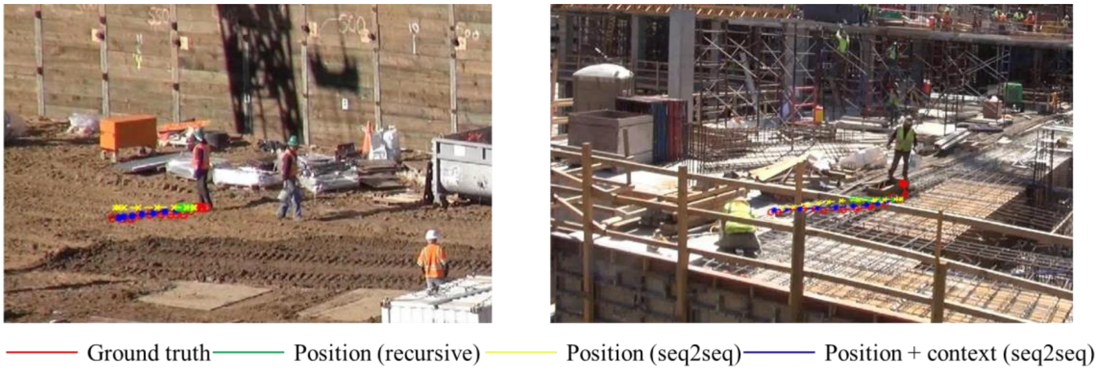


Figure 4.6 Example results of trajectory prediction

Table 4.1 lists the quantitative results from the three models. The recursive approach leads to much larger errors in both final displacement error (FDE) and average displacement error (ADE) compared to the seq2seq approach, which proves that the seq2seq model is an effective way to avoid error accumulation when predicting trajectory over multiple time steps. The context-augmented model results in smaller FDE but a slightly larger ADE compared to the position-based model. This is because by incorporating contextual information, especially the potential destination information, the model is inherently trained to adapt more to the long-term goal, rather than accurate prediction of each step. It is reasonable because the final displacement is more critical in predicting the struck-by hazard in safety management.

Table 4.1 Quantitative results from three models

Model	**FDE (pixel)**	**ADE (pixel)**
Position (recursive)	28.32	15.41
Position (seq2seq)	9.00	**8.95**
Position +Context (seq2seq)	**8.51**	9.00

4.5.2 Qualitative analysis

The results from two seq2seq models, i.e., position-based seq2seq model and context-augmented seq2seq model, are analyzed qualitatively to evaluate the impact of contextual information and identify the scenarios, under which integrating contextual information leads to better performance.

It was found that when workers walk continuously and are not involved in specific collaborating activities, contextual information does not have a significant influence and both models result in relatively accurate prediction, as shown in Figure 4.7. On the other hand, if the target is collaborating with others or involved in activities within an area, incorporating contextual information leads to better prediction (see Figure 4.8). In Figure 4.8(a), the target intends to move towards his co-worker, who is working in the left-bottom corn of the image. With contextual information, especially the position and the relationship with the nearest neighbor, the context-aware model accurately predicts the behavior of the target moving towards his neighbor, resulting in a path closer to the actual trajectory. In contrast, the position-based model only considers individual movement patterns and is more likely to end up with a near-linear trajectory, which is farther from the actual trajectory. In Figure 4.8(b), the target is conducting road paving activity with a roller and other co-workers. Although there remains some discrepancy with the actual trajectory, the context-aware model accurately predicts the trend of worker movement, whereas the position-based model predicts the movement in the opposite direction.

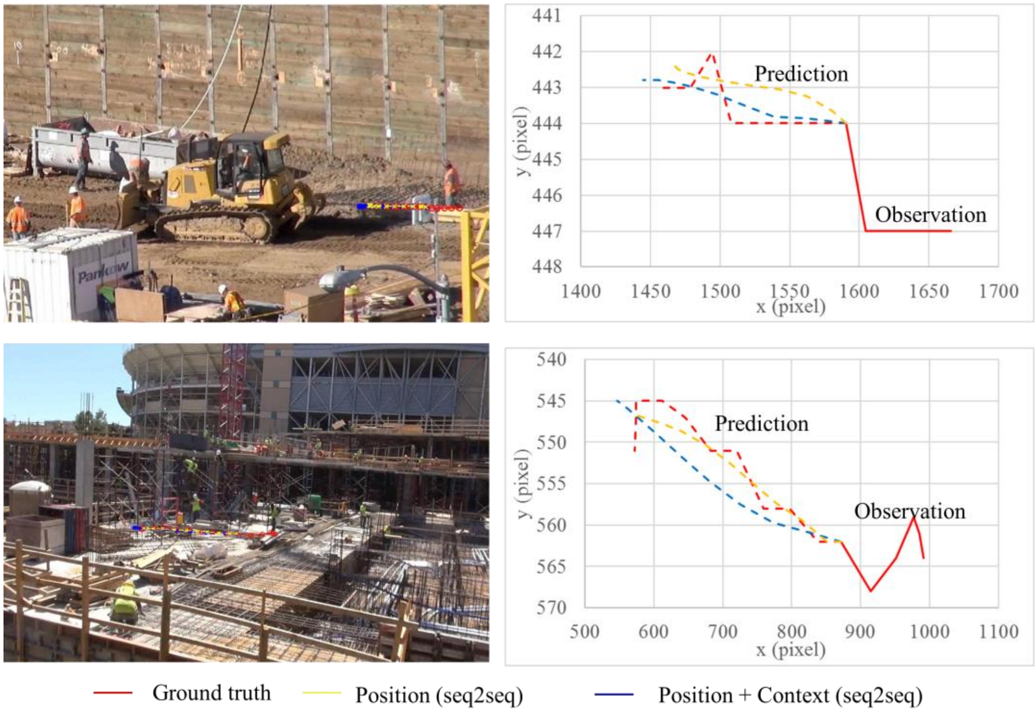
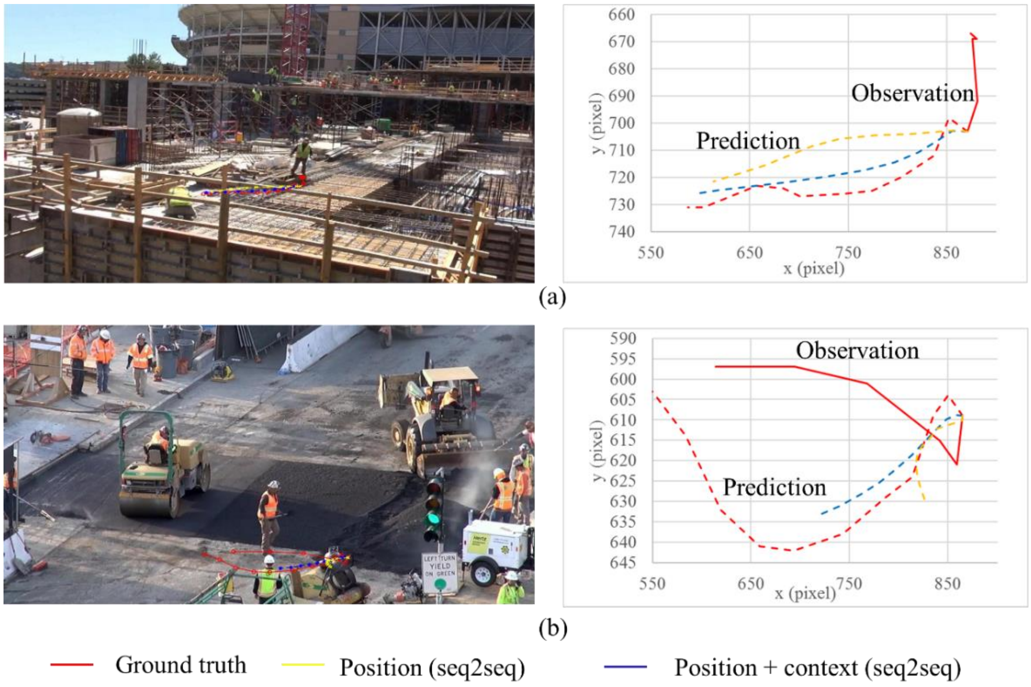


Figure 4.7 Two seq2seq models lead to similar results under moving scenarios



(a)

(b)

Figure 4.8 Context-augmented model leads to better prediction

In some cases, however, the proposed method may fail, see Figure 4.9. When the status of target significantly changes during prediction time (e.g., from stationary to moving and vice versa), the movement cannot be accurately predicted, see Figure 4.9 (a). In addition, it is also very challenging when workers are conducting activities within a limited area without substantial movement, as shown in Figure 4.9 (b).

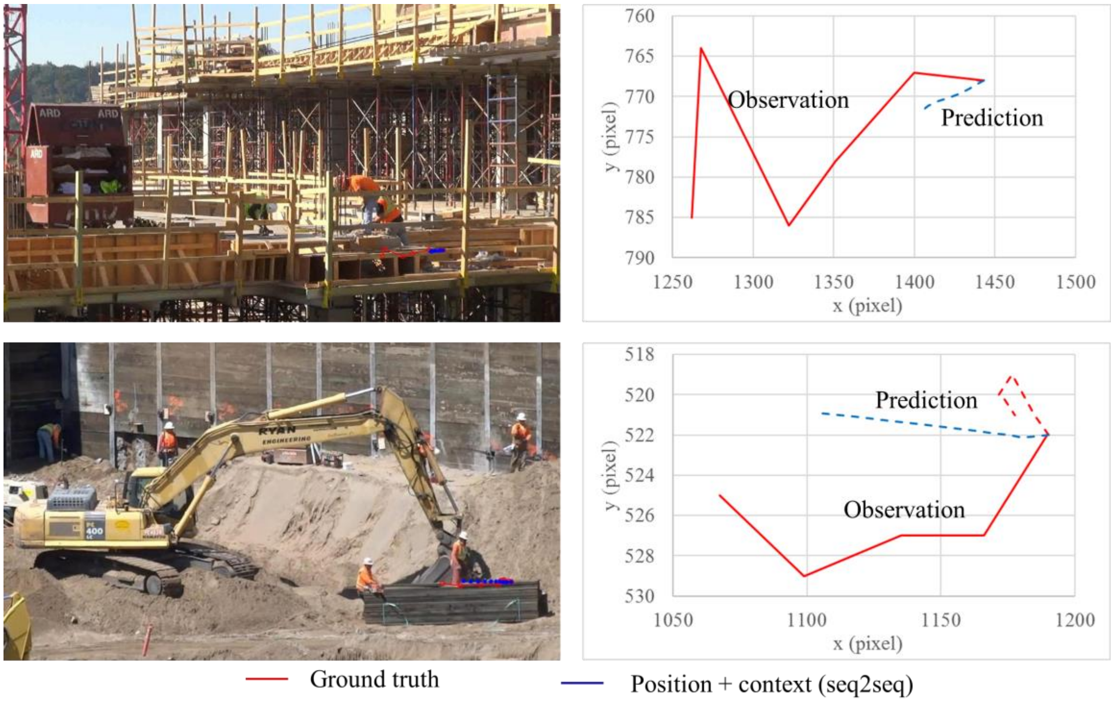
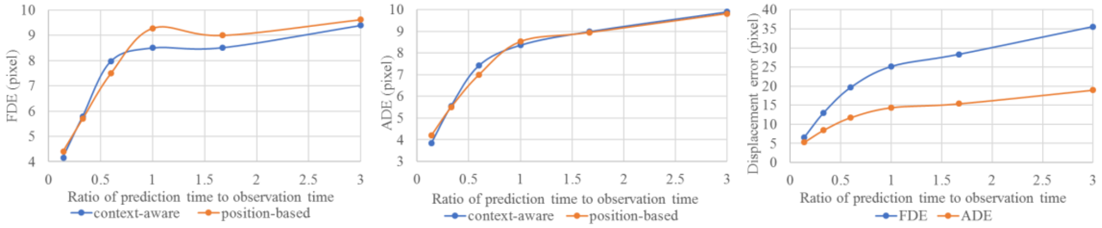


Figure 4.9 Examples when context-augmented model fails

4.5.3 Influence of prediction time

To evaluate the influence of prediction time on different methods, this study examines the prediction performance with respect to various ratios of prediction to observation length within the 8-s track prepared in the dataset. Specifically, the partition of observation time and prediction time varies as 7s/1s, 6s/2s, 5s/3s, 4s/4s, 3s/5s (used in the previous experiment), and 2s/6s. The results are illustrated in Figure 4.10. It is not surprising that both FDE and ADE increase as the ratio of prediction to observation increases for all three prediction models, which further proves the challenge in long-term trajectory prediction (i.e., when prediction time is no less than observation time). From Figure 4.10(c), the discrepancy between FDE and ADE for the position-based recursive model becomes much larger as the increase of the ratio, compared to those in two seq2seq models (Figure 4.10(a) and (b)). It proves the advantage of seq2seq architecture in mitigating the error accumulation for long-term trajectory prediction. In the comparison of position-based and context-aware seq2seq models, the FDEs for both models are compatible in short-term prediction (i.e., when the ratio is less than 1). However, the context-aware method leads to lower FDE in long-term prediction.



(a) FDE of seq2seq models under different ratios of prediction to observation

(b) ADE of seq2seq models under different ratios of prediction to observation

(c) Prediction performance of recursive model under different ratios of prediction to observation

Figure 4.10 Influence of prediction time on different models

4.6 Contributions

This study contributes to the body of knowledge in three aspects. First, the proposed context-augmented deep learning method for construction worker trajectory prediction not only considers spatial interaction between the target and neighboring entities, but also innovatively incorporates the semantic relationship between entities (i.e., whether or not within a working group) and the long-term goal of the target (i.e., the potential destination). The results show that integrating the above contextual information outperforms the position-based prediction, especially for long-term prediction when prediction time is no less than observation time. Second, an LSTM encoder-decoder architecture is adopted to form a sequence-to-sequence model, which eliminates the error accumulation in predicting trajectory over multiple time steps, compared to conventional prediction model that only predicts position one-step ahead each time. Third, an extensive qualitative analysis is conducted to identify the scenarios where incorporating contextual information is more worthwhile. It is found that context-aware model leads to better performance when workers are conducting collaborative activities. When workers move continuously with limited interactions with other entities, integrating contextual information does not have a significant impact, and both context-augmented and position-based seq2seq methods achieve relatively accurate prediction results. These findings provide valuable insights on the selection of appropriate methods for effective and efficient trajectory prediction of construction entities.

4.7 Summary and Conclusions

Predicting workers' trajectories on unstructured and dynamic construction sites has great potential to improve workplace safety. It provides rich information and is critical to pro-actively prevent struck-by accidents, which has been a major cause of construction fatalities and a single leading cause for non-fatal injuries. This study proposed an LSTM model augmented by jobsite contextual information for construction worker trajectory prediction considering both individual movement information and jobsite contextual information. The contextual information is represented as movements of neighboring entities, working group information, and potential destination information. Experiments were conducted using videos collected from three different construction projects. The results show that the newly created method leads to a smaller final displacement error than the model relying solely on target movements, especially in long-term prediction when the length of prediction is no less than that of observation. The adopted sequence-to-sequence network architecture also significantly improves the performance in both final displacement error and average displacement error by eliminating error accumulation over multiple time steps.

In addition, qualitative analysis was conducted to identify scenarios when incorporating contextual information is worthwhile. It was found that when workers are conducting collaborative activities within an area, incorporating contextual information leads to better results. The context-aware prediction model should be selected when the construction scenario involves multiple entities collaborating on group activities. Both context-aware and position-based methods lead to relatively accurate predicted trajectories when workers move continuously and are not involved in collaborating activities. However, in such case, the position-based method is favorable. Although in this study, the training time for two models is almost the same (about 3s per epoch), with more data in the future, the position-based method is expected to be less computational expensive considering the fewer features involved in training the model. Moreover, extracting contextual information involves much more complex computing process and may introduce additional errors. Both models may fail when entity states change significantly. In such case, it is not reliable to directly predict worker’s trajectory and more information (e.g., activity type, entity posture) may be needed.

This study can be extended towards several directions in future research. First, besides contextual information that reflects dynamic interactions among construction entities, static scene contexts such as site layout can also be incorporated for further improvement. Semantic scene segmentation networks such as SegNet (Badrinarayanan et al., 2017) can be leveraged for semantic scene features (Syed & Morris, 2019). Second, more activity-related information (e.g., worker poses) can be added to differentiate various states of workers whose states change significantly or who are involved in multiple working activities in a limited area without substantial movement. Third, because the available construction dataset is very limited and data annotation is time-consuming and labor-intensive, transfer learning can be explored by leveraging the public datasets in other domain (e.g., crowds datasets (Lerner et al., 2007; Pellegrini et al., 2009)) to overcome the limitation in available annotated construction datasets. Fourth, the bivariate Gaussian distribution can be used to represent the trajectory to incorporate the uncertainty associated with the predicted trajectory.

5. SUMMARY

This chapter summarizes the entire dissertation and discusses directions for future work.

5.1 Summary and Conclusions

Construction sites are unstructured and dynamic, with numerous resources (e.g., workers, equipment, and materials) co-existing and interacting constantly. Having a holistic situational awareness of the site dynamics is essential to improve construction site safety performance, such as prevention of struck-by accidents that have been a major cause of construction fatalities and non-fatal injuries. This dissertation presents a novel data-driven approach to enhancing holistic situational awareness—perception, comprehension, and prediction—of the jobsite for minimizing the risk of potential struck-by hazards. Three specific problems have been addressed in this dissertation, including 1) accurate perception of positional states of construction workers – a hybrid frame that fuses vision-based tracking and radio-based identification for multi-worker tracking, 2) jobsite context comprehension in terms of working groups and activities – a two-step LSTM method integrating positional and attentional cues, and 3) construction worker trajectory prediction – a context-augmented LSTM method incorporating both worker movements and contextual information. Chapters 2 to 4 are dedicated to these three problems respectively.

Chapter 2 presents a hybrid framework that fuses results obtained from vision-based tracking and radio-based identification and localization for 3D tracking of multiple construction workers. Compared to traditional fusion approaches that directly fuse locations extracted from these two approaches, the proposed method treats vision-based tracking as the main source to extract the object trajectory. Radio-based identification and localization results are used as a supplementary source to augment anonymous visual tracks with identity information and correct errors (e.g., false positives) in vision-based object detection. The newly created method has been validated using two indoor experiments. Results show that the new approach for fusing vision- and radio-based results increases the overall accuracy from 88% and 87% to 95% and 90%, compared to using the vision-based approach alone. The integration of radio-based identification is much more robust than using vision system alone as it allows the recovery of the same entity ID after the trajectory is fragmented and results in fewer fragmentations that last longer than 0.2s.

Chapter 3 presents a two-step classification approach—working group identification followed by activity recognition, leveraging both positional and attentional cues, to recognize complex interactions and their involved entities from videos that contain different activities with multiple entities. The spatial and attentional states of individual entities are represented numerically, and the corresponding positional and attentional cues between two entities are computed. LSTM networks are designed to (1) classify whether two entities belong to the same group, and (2) recognize the activities they are involved in. The newly created method is validated using two sets of construction videos. Identifying working groups before recognizing ongoing activities enables the exclusion of group-irrelevant entities and thus, improves the performance. Moreover, by leveraging both positional and attentional cues, the accuracy increases from 85% to 95% compared with cases using positional cues alone.

Chapter 4 presents an LSTM model augmented by the context information, which incorporates both individual movement and workplace contextual information. Contextual information regarding movements of neighboring entities, working group information, and potential destination information is concatenated with movements of the target entity and fed into an LSTM network with an encoder-decoder architecture to enable the sequence-to-sequence prediction, i.e., a sequence of estimated positions is generated from a sequence of observations. The method is validated using videos collected on construction sites. The accuracy of prediction achieved by this method is 8.51 pixels in terms of final displacement error, with an observation time of 3s and prediction time of 5s. It was found that integrating contextual information with target movement information can result in a smaller final displacement error, especially when the length of prediction is longer than the length of observation. Compared with the conventional method that predicts position only one step ahead, the proposed method predicts trajectories over multiple steps following a sequence-to-sequence architecture and consequently, eliminates the error accumulation issue.

This dissertation research enhances the holistic situational awareness of the construction site through a data-driven approach. This research identified critical features that are unique in the construction domain to capture entity interactions and created generic models to represent them numerically. By establishing the relationship between entity interaction patterns with construction working groups and group activities, this research enables the comprehension of complex jobsite contexts on dynamic and unstructured workspaces. This research also identified critical contextual features that will influence worker movements and innovatively incorporates contextual information into the prediction of future worker states.

The resulting holistic situational awareness of dynamic construction jobsites can be further leveraged to develop pro-active, context-aware control systems for struck-by prevention. In the system, the risk of potential collision can be estimated based on the predicted trajectory of construction entities, and early warnings can be provided to involved entities to avoid struck-by accidents. This newly enhanced capacity is possible to be extended to prevent other types of accidents, such as fall accidents and electrocutions. It has great potential to contribute not only to improve site safety performance by avoiding struck-by accidents, but also to automatic progress monitoring and control to ensure productivity, as well as to safe and efficient human-robot collaboration on future construction scenarios.

5.2 Future Work – Roadmap towards Harmonious Human-Robot Collaboration in Future Smart Construction

The achievement of this dissertation can be implemented to a system approach in order to prevent struck-by accidents, illustrated in Figure 5.1. Vision and radio systems can be used to collect data on entity states, which will be transmitted to a central server to perform analysis using algorithms developed in this dissertation, including state perception, jobsite context comprehension, and trajectory prediction. This holistic situational awareness can be leveraged to develop a proactive and context-aware control system, such as an adaptive path planning mechanism based on the predicted trajectory of jobsite entities. Then such information and guidance can be communicated to field crews in different formats through mobile devices. For instance, in addition to early warnings in sounds and vibration, for operators in equipment, we can visualize the site condition including the movements of their surrounding entities, and show the planned trajectory for them. For workers, we can provide tailored information on their nearby hazards through voice and visualization using mobile devices or augmented reality devices. By doing these, the jobsite entities are augmented with holistic and ubiquitous situational awareness to prevent struck-by accidents.

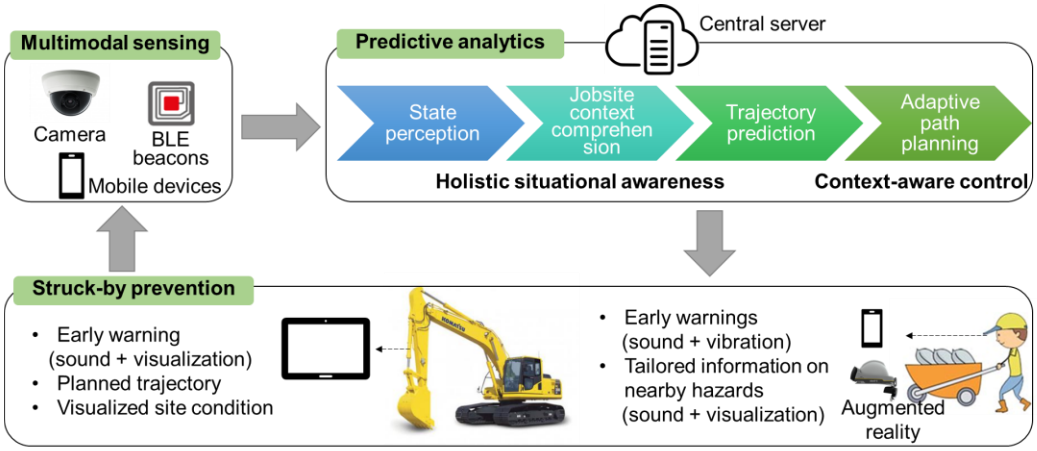


Figure 5.1 Future system implementation for struck-by prevention based on holistic situational awareness

This dissertation also establishes the roadmap towards harmonious human-robot collaboration in future smart construction. With the emergence of automation and robotics, autonomous robots have been increasingly introduced to construction projects to relieve human workers from demanding and hazardous tasks (Daeho Kim et al., 2019). Examples include mobile robots for bridge inspection (Sutter et al., 2018) and robotic excavator (ASI, 2019). It is predicted that the robot market revenue will increase from $22.7 million in 2018 to $226 million by 2025, with more than 7,000 construction robots deployed on construction sites and most being robot assistants (Sanders & Kaul, 2019). The rapid advances in construction robots and the co-existence of and interaction among workers and construction robots will bring in more challenges to the already unstructured and dynamic site, such as the difficulty in mutual understanding and harmonious collaboration between workers and robots.

Figure 5.2 illustrates the vision of future smart construction and how this research can contribute to that. With multimodal sensors throughout the jobsites including cameras, robotic systems such as unmanned aerial vehicle (UAV) and unmanned ground vehicle (UGV), wearable on the workers (e.g., motion sensors and biosensors), and embedded on the machines, the jobsite is monitored continuously in all aspects in real-time. Having the heterogeneous data, we can do predictive analysis to gain holistic situational awareness on the jobsite to facilitate the real-time decision and control, and the gained knowledge is also visualized to facilitate the collaboration and decision making. More importantly, humans and robots will gain shared and enhanced situational awareness, which will allow for their mutual understanding and effective collaboration on complex construction tasks.

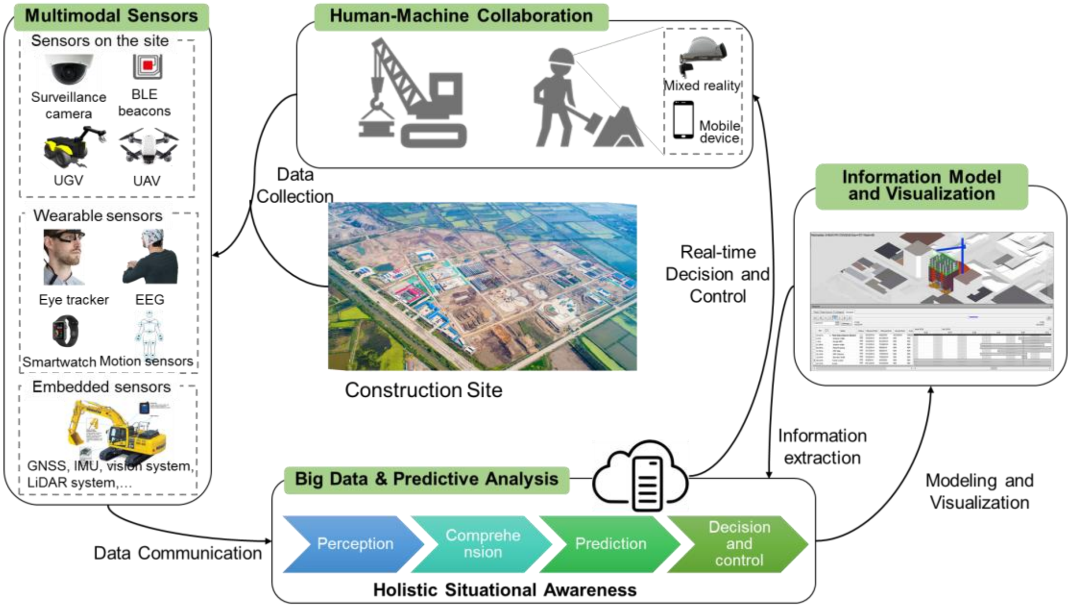


Figure 5.2 Vision on future smart construction

REFERENCES

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis. *ACM Computing Surveys*, *43*(3), 1–43. https://doi.org/10.1145/1922649.1922653

Ahn, J. W., Chang, T. W., Lee, S. H., & Seo, Y. W. (2016). Two-Phase Algorithm for Optimal Camera Placement. *Scientific Programming*, *2016*. https://doi.org/10.1155/2016/4801784

Akhavian, R., & Behzadan, A. H. (2015). Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. *Advanced Engineering Informatics*, *29*(4), 867–877. https://doi.org/10.1016/j.aei.2015.03.001

Akhavian, R., & Behzadan, A. H. (2016). Smartphone-based construction workers' activity recognition and classification. *Automation in Construction*, *71*(Part 2), 198–209. https://doi.org/10.1016/j.autcon.2016.08.015

Akhavian, R., & Behzadan, A. H. (2018). Coupling human activity recognition and wearable sensors for data-driven construction simulation. *Journal of Information Technology in Construction*, *23*, 1–15.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 961–971. https://doi.org/10.1109/CVPR.2016.110

Altahir, A. A., Asirvadam, V. S., Hamid, N. H., Sebastian, P., Saad, N., Ibrahim, R., & Dass, S. C. (2017). Modeling Multicamera Coverage for Placement Optimization. *IEEE Sensors Letters*, *1*(6), 1–4. https://doi.org/10.1109/lsens.2017.2758371

Asadi, K., Ramshankar, H., Noghabaei, M., & Han, K. (2019). Real-time image localization and registration with BIM using perspective alignment for indoor monitoring of construction. *Journal of Computing in Civil Engineering*, *33*(5). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000847

ASI. (2019). Robotic Excavators. Retrieved October 12, 2019, from https://www.asirobots.com/mining/excavator/

Ba, S. O., & Odobez, J. M. (2011). Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 101–116. https://doi.org/10.1109/TPAMI.2010.69

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Battacharyya, A. (1943). On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society*. https://doi.org/10.1038/157869b0

Bohm, J., & Harris, D. (2010). Risk perception and risk-taking behavior of construction site dumper drivers. *International Journal of Occupational Safety and Ergonomics*, *16*(1), 55–67. https://doi.org/10.1080/10803548.2010.11076829

Brilakis, I., Park, M. W., & Jog, G. (2011). Automated vision tracking of project related entities. *Advanced Engineering Informatics*, *25*(4), 713–724. https://doi.org/10.1016/j.aei.2011.01.003

Cai, H., Andoh, A. R., Su, X., & Li, S. (2014). A boundary condition based algorithm for locating construction site objects using RFID and GPS. *Advanced Engineering Informatics*, *28*(4), 455–468. https://doi.org/10.1016/j.aei.2014.07.002

Cai, J., & Cai, H. (2020). Robust Hybrid Approach of Vision-Based Tracking and Radio-Based Identification and Localization for 3D Tracking of Multiple Construction Workers. *Journal of Computing in Civil Engineering*, *34*(4), 4020021.

Cai, J., Zhang, Y., & Cai, H. (2019). Two-step long short-term memory method for identifying construction activities through positional and attentional cues. *Automation in Construction*, *106*, 102886. https://doi.org/10.1016/j.autcon.2019.102886

Chamveha, I., Sugano, Y., Sato, Y., & Sugimoto, A. (2014). *Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues*. 121.1-121.11. https://doi.org/10.5244/c.27.121

Chen, H., Luo, X., & Ke, J. (2018). Multisource Fusion Framework for Environment Learning–Free Indoor Localization. *Journal of Computing in Civil Engineering*, *32*(5), 04018040. https://doi.org/10.1061/(asce)cp.1943-5487.0000782

Cheng, C.-F., Rashidi, A., Davenport, M. A., & Anderson, D. V. (2019). Evaluation of Software and Hardware Settings for Audio-Based Analysis of Construction Operations. *International Journal of Civil Engineering*, 1–12. https://doi.org/10.1007/s40999-019-00409-2

Cheng, C. F., Rashidi, A., Davenport, M. A., & Anderson, D. V. (2017). Activity analysis of construction equipment using audio signals and support vector machines. *Automation in Construction*, *81*, 240–253. https://doi.org/10.1016/j.autcon.2017.06.005

Cheng, T., & Teizer, J. (2012). Modeling Tower Crane Operator Visibility to Minimize the Risk of Limited Situational Awareness. *Journal of Computing in Civil Engineering*, *28*(3), 04014004. https://doi.org/10.1061/(asce)cp.1943-5487.0000282

Costin, A. M., & Teizer, J. (2015). Fusing passive RFID and BIM for increased accuracy in indoor localization. *Visualization in Engineering*, *3*(1). https://doi.org/10.1186/s40327-015-0030-6

Ding, L., Fang, W., Luo, H., Love, P. E. D., Zhong, B., & Ouyang, X. (2018). A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Automation in Construction*, *86*, 118–124. https://doi.org/10.1016/j.autcon.2017.11.002

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 677–691. https://doi.org/10.1109/TPAMI.2016.2599174

Dong, C., Li, H., Luo, X., Ding, L., Siebert, J., & Luo, H. (2018). Proactive struck-by risk detection with movement patterns and randomness. *Automation in Construction*, *91*, 246–255. https://doi.org/10.1016/j.autcon.2018.03.021

Dong, X. S., Wang, X., Katz, R., West, G., & Bunting, J. (2017). Struck-by Injuries and Prevention in the Construction Industry. Retrieved April 12, 2019, from CPWR Quarterly Data Report website: http://www.cpwr.com/sites/default/files/publications/Quarter1-QDR-2017.pdf

Edrei, T., & Isaac, S. (2017). Construction site safety control with medium-accuracy location data. *Journal of Civil Engineering and Management*, *23*(3), 384–392. https://doi.org/10.3846/13923730.2016.1144644

Egenhofer, M., & Herring, J. (1992). Categorizing binary topological relations between regions, lines, and points in geographic databases. *University of Maine, Orono, Maine, Dept. of Surveying Engineering, Technical Report*, (July 2016), 1–28. Retrieved from https://pdfs.semanticscholar.org/b303/39af3f0be6074f7e6ac0263e9ab34eb84271.pdf

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Fang, D., & Wu, H. (2013). Development of a Safety Culture Interaction (SCI) model for construction projects. *Safety Science*, *57*, 138–149. https://doi.org/10.1016/j.ssci.2013.02.003

Fang, Q., Li, H., Luo, X., Ding, L., Rose, T. M., An, W., & Yu, Y. (2018a). A deep learning-based method for detecting non-certified work on construction sites. *Advanced Engineering Informatics*, *35*, 56–68. https://doi.org/10.1016/j.aei.2018.01.001

Fang, W., Ding, L., Luo, H., & Love, P. E. D. (2018b). Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, *91*, 53–61. https://doi.org/10.1016/j.autcon.2018.02.018

Fang, W., Ding, L., Zhong, B., Love, P. E. D., & Luo, H. (2018c). Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. *Advanced Engineering Informatics*, *37*, 139–149. https://doi.org/10.1016/j.aei.2018.05.003

Feng, C., Kamat, V. R., & Cai, H. (2018). Camera marker networks for articulated machine pose estimation. *Automation in Construction*, *96*, 148–160. https://doi.org/10.1016/j.autcon.2018.09.004

Gao, X. S., Hou, X. R., Tang, J., & Cheng, H. F. (2003). Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(8), 930–943. https://doi.org/10.1109/TPAMI.2003.1217599

Golparvar-Fard, M., Heydarian, A., & Niebles, J. C. (2013). Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, *27*(4), 652–663. https://doi.org/10.1016/j.aei.2013.09.001

Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced Engineering Informatics*, *25*(4), 771–782. https://doi.org/10.1016/j.aei.2011.06.002

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hermes, C., Barth, A., Wöhler, C., & Kummert, F. (2009). Object Motion Analysis and Prediction in Stereo Image Sequences. *8. Oldenburger 3D-Tage*.

Hinze, J. W., & Teizer, J. (2011). Visibility-related fatalities related to construction equipment. *Safety Science, 49*(5), 709–718. https://doi.org/10.1016/j.ssci.2011.01.007

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hwang, S. (2012). Ultra-wide band technology experiments for real-time prevention of tower crane collisions. *Automation in Construction*, *22*, 545–553. https://doi.org/10.1016/j.autcon.2011.11.015

Ibrahim, M., & Moselhi, O. (2016). Inertial measurement unit based indoor localization for construction applications. *Automation in Construction*, *71*, 13–20. https://doi.org/10.1016/j.autcon.2016.05.006

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). Hierarchical Deep Temporal Models for Group Activity Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1971–1980. Retrieved from http://arxiv.org/abs/1607.02643

Isli, A. (2003). Integrating cardinal direction relations and other orientation relations in Qualitative Spatial Reasoning. *Annals of Mathematics*. Retrieved from http://arxiv.org/abs/cs/0307048

Jeelani, I., Ramshankar, H., Han, K., Albert, A., & Asadi, K. (2019). Real-Time Hazard Proximity Detection—Localization of Workers Using Visual Data. In *Computing in Civil Engineering 2019: Data, Sensing, and Analytics* (pp. 281–289). American Society of Civil Engineers Reston, VA.

Joshua, L., & Varghese, K. (2010). Accelerometer-Based Activity Recognition in Construction. *Journal of Computing in Civil Engineering*, *25*(5), 370–379. https://doi.org/10.1061/(asce)cp.1943-5487.0000097

Jung, D., Teixeira, T., & Savvides, A. (2010). Towards cooperative localization of wearable sensors using accelerometers and cameras. *Proceedings - IEEE INFOCOM*. https://doi.org/10.1109/INFCOM.2010.5462059

Karasev, V., Ayvaci, A., Heisele, B., & Soatto, S. (2016). Intent-aware long-term prediction of pedestrian motion. *Proceedings - IEEE International Conference on Robotics and Automation*, *2016-June*, 2543–2549. https://doi.org/10.1109/ICRA.2016.7487409

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

Kim, D., Liu, M., Lee, S., & Kamat, V. R. (2019). Trajectory prediction of mobile construction resources toward pro-active struck-by hazard detection. *Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019*, 982–988. https://doi.org/10.22260/isarc2019/0131

Kim, Daeho, Goyal, A., Newell, A., Lee, S., Deng, J., & Kamat, V. R. (2019). Semantic Relation Detection between Construction Entities to Support Safe Human-Robot Collaboration in Construction. In *Computing in Civil Engineering 2019: Data, Sensing, and Analytics* (pp. 265–272). American Society of Civil Engineers Reston, VA.

Kim, H. H., Bang, S., Jeong, H., Ham, Y., & Kim, H. H. (2018). Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation. *Automation in Construction*, *92*, 188–198. https://doi.org/10.1016/j.autcon.2018.04.002

Kim, Hongjo, Kim, K., & Kim, H. (2015). Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects. *Journal of Computing in Civil Engineering*, *30*(4), 04015075. https://doi.org/10.1061/(asce)cp.1943-5487.0000562

Kim, Hyunsoo, Ahn, C. R., Engelhaupt, D., & Lee, S. H. (2018). Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement. *Automation in Construction*, *87*, 225–234. https://doi.org/10.1016/j.autcon.2017.12.014

Kim, J., Chi, S., & Seo, J. (2018). Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Automation in Construction*, *87*, 297–308. https://doi.org/10.1016/j.autcon.2017.12.016

Kim, J., Ham, Y., Chung, Y., & Chi, S. (2019). Systematic Camera Placement Framework for Operation-Level Visual Monitoring on Construction Jobsites. *Journal of Construction Engineering and Management*, *145*(4). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001636

Kim, J. Y., & Caldas, C. H. (2017). Vision-Based Action Recognition in the Internal Construction Site Using Interactions between Worker Actions and Construction Objects. *Proceedings of the 30th International Symposium on Automation and Robotics in Construction and Mining (ISARC 2013): Building the Future in Automation and Robotics*. https://doi.org/10.22260/isarc2013/0072

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv Preprint ArXiv:1412.6980*. Retrieved from http://arxiv.org/abs/1412.6980

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., & Hebert, M. (2012). Activity forecasting. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7575 LNCS*(PART 4), 201–214. https://doi.org/10.1007/978-3-642-33765-9_15

Konstantinou, E., & Brilakis, I. (2018). Matching Construction Workers across Views for Automated 3D Vision Tracking On-Site. *Journal of Construction Engineering and Management*, *144*(7), 04018061. https://doi.org/10.1061/(asce)co.1943-7862.0001508

Kooij, J. F. P., Flohr, F., Pool, E. A. I., & Gavrila, D. M. (2019). Context-Based Path Prediction for Targets with Switching Dynamics. *International Journal of Computer Vision*, *127*(3), 239–262. https://doi.org/10.1007/s11263-018-1104-4

Lee, H.-S., Lee, K.-P., Park, M., Baek, Y., & Lee, S. (2011). RFID-Based Real-Time Locating System for Construction Safety Management. *Journal of Computing in Civil Engineering*, *26*(3), 366–377. https://doi.org/10.1061/(asce)cp.1943-5487.0000144

Lee, T. (2007). Loss Functions in Time Series Forecasting. *University of California*, *1*(1999), 1–14. Retrieved from http://www.faculty.ucr.edu/~taelee/paper/lossfunctions.pdf

Lee, Y. J., & Park, M. W. (2019). 3D tracking of multiple onsite workers based on stereo vision. *Automation in Construction*, *98*, 146–159. https://doi.org/10.1016/j.autcon.2018.11.017

Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. *Computer Graphics Forum*, *26*(3), 655–664. https://doi.org/10.1111/j.1467-8659.2007.01089.x

Li, H., Chan, G., Wong, J. K. W., & Skitmore, M. (2016). Real-time locating systems applications in construction. *Automation in Construction*, Vol. 63, pp. 37–47. https://doi.org/10.1016/j.autcon.2015.12.001

Li, H., Lu, M., Hsu, S. C., Gray, M., & Huang, T. (2015). Proactive behavior-based safety management for construction safety improvement. *Safety Science*, *75*, 107–117. https://doi.org/10.1016/j.ssci.2015.01.013

Liang, C. J., Lundeen, K. M., McGee, W., Menassa, C. C., Lee, S. H., & Kamat, V. R. (2019). A vision-based marker-less pose estimation system for articulated construction robots. *Automation in Construction*, *104*, 80–94. https://doi.org/10.1016/j.autcon.2019.04.004

Liu, D., Chen, J., Li, S., & Cui, W. (2018). An integrated visualization framework to support whole-process management of water pipeline safety. *Automation in Construction*, *89*, 24–37. https://doi.org/10.1016/j.autcon.2018.01.010

Liu, D., Wu, Y., Li, S., & Sun, Y. (2016). A real-time monitoring system for lift-thickness control in highway construction. *Automation in Construction*, *63*, 27–36. https://doi.org/10.1016/j.autcon.2015.12.004

Liu, T., Bahl, P., & Chlamtac, I. (1998). Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks. *IEEE Journal on Selected Areas in Communications*, *16*(6), 922–935. https://doi.org/10.1109/49.709453

Lundeen, K. M., Dong, S., Fredricks, N., Akula, M., Seo, J., & Kamat, V. R. (2016). Optical marker-based end effector pose estimation for articulated excavators. *Automation in Construction*, *65*, 51–64. https://doi.org/10.1016/j.autcon.2016.02.003

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

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., & Lee, S. (2018). Recognizing Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks. *Journal of Computing in Civil Engineering*, *32*(3), 04018012. https://doi.org/10.1061/(asce)cp.1943-5487.0000756

Luo, X., Li, H., Dai, F., Cao, D., Yang, X., & Guo, H. (2017). Hierarchical Bayesian Model of Worker Response to Proximity Warnings of Construction Safety Hazards: Toward Constant Review of Safety Risk Control Measures. *Journal of Construction Engineering and Management*, *143*(6). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001277

Luo, X., Li, H., Wang, H., Wu, Z., Dai, F., & Cao, D. (2019). Vision-based detection and visualization of dynamic workspaces. *Automation in Construction, 104,* 1–13. https://doi.org/10.1016/j.autcon.2019.04.001

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, *2*, 49–55. https://doi.org/10.1145/1390156.1390302

Mandeljc, R., Kovačič, S., Kristan, M., & Perš, J. (2013). Tracking by identification using computer vision and radio. *Sensors (Switzerland), 13*(1), 241–273. https://doi.org/10.3390/s130100241

Marks, E. D., & Teizer, J. (2013). Method for testing proximity detection and alert technology for safe construction equipment operation. *Construction Management and Economics, 31*(6), 636–646. https://doi.org/10.1080/01446193.2013.783705

MathWorks. (2019). Computer vision toolbox. Retrieved June 4, 2019, from https://www.mathworks.com/products/computer-vision.html

Memarzadeh, M., Golparvar-Fard, M., & Niebles, J. C. (2013). Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction, 32,* 24–37. https://doi.org/10.1016/j.autcon.2012.12.002

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. *ArXiv Preprint ArXiv:1603.00831*.

Mohebbi, P., Stroulia, E., & Nikolaidis, I. (2017). Sensor-data fusion for multi-person indoor location estimation. *Sensors (Switzerland)*, *17*(10). https://doi.org/10.3390/s17102377

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, *5*(1), 32–38.

Nabil, M., Ngu, A. H. H., & Shepherd, J. (1996). Picture similarity retrieval using the 2D projection interval representation. *IEEE Transactions on Knowledge and Data Engineering*, *8*(4), 533–539. https://doi.org/10.1109/69.536246

OSHA. (2015). Fall Protection in Construction. Retrieved June 3, 2019, from https://www.osha.gov/Publications/OSHA3146.pdf

OSHA. (2018). Commonly Used Statistics. Retrieved March 25, 2020, from https://www.osha.gov/oshstats/commonstats.html

Ozturk, O., Yamasaki, T., & Aizawa, K. (2011). Estimating human body and head orientation change to detect visual attention direction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6468 LNCS*(PART1), 410–419. https://doi.org/10.1007/978-3-642-22822-3_41

Papaioannou, S., Markham, A., & Trigoni, N. (2017). Tracking People in Highly Dynamic Industrial Environments. *IEEE Transactions on Mobile Computing*, *16*(8), 2351–2365. https://doi.org/10.1109/TMC.2016.2613523

Papaioannou, S., Wen, H., Markham, A., & Trigoni, N. (2015). Fusion of radio and camera sensor data for accurate indoor positioning. *Proceedings - 11th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2014*, 109–117. https://doi.org/10.1109/MASS.2014.52

Park, J. W., Yang, X., Cho, Y. K., & Seo, J. (2017). Improving dynamic proximity sensing and processing for smart work-zone safety. *Automation in Construction*, *84*, 111–120. https://doi.org/10.1016/j.autcon.2017.08.025

Park, Jeewoong, Kim, K., & Cho, Y. K. (2017). Framework of Automated Construction-Safety Monitoring Using Cloud-Enabled BIM and BLE Mobile Tracking Sensors. *Journal of Construction Engineering and Management*, *143*(2). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001223

Park, JeeWoong, Kim, K., & Cho, Y. K. (2016). Framework of Automated Construction-Safety Monitoring Using Cloud-Enabled BIM and BLE Mobile Tracking Sensors. *Journal of Construction Engineering and Management*, *143*(2), 05016019. https://doi.org/10.1061/(asce)co.1943-7862.0001223

Park, M.-W., Koch, C., & Brilakis, I. (2011). Three-Dimensional Tracking of Construction Resources Using an On-Site Camera System. *Journal of Computing in Civil Engineering*, *26*(4), 541–549. https://doi.org/10.1061/(asce)cp.1943-5487.0000168

Park, M. W., & Brilakis, I. (2012). Construction worker detection in video frames for initializing vision trackers. *Automation in Construction*, *28*, 15–25. https://doi.org/10.1016/j.autcon.2012.06.001

Park, M. W., & Brilakis, I. (2016). Continuous localization of construction workers via integration of detection and tracking. *Automation in Construction*, *72*, 129–142. https://doi.org/10.1016/j.autcon.2016.08.039

Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You’ll never walk alone: Modeling social behavior for multi-target tracking. *Proceedings of the IEEE International Conference on Computer Vision*, 261–268. https://doi.org/10.1109/ICCV.2009.5459260

Pereira, E. M., Ciobanu, L., & Cardoso, J. S. (2017). Cross-layer classification framework for automatic social behavioural analysis in surveillance scenario. *Neural Computing and Applications*, *28*(9), 2425–2444. https://doi.org/10.1007/s00521-016-2282-z

Prévost, C. G., Desbiens, A., & Gagnon, E. (2007). Extended Kalman filter for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle. *Proceedings of the American Control Conference*, 1805–1810. https://doi.org/10.1109/ACC.2007.4282823

Qin, Z., & Shelton, C. R. (2016). Social Grouping for Multi-Target Tracking and Head Pose Estimation in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2082–2095. https://doi.org/10.1109/TPAMI.2015.2505292

Rashid, K. M., Datta, S., Behzadan, A. H., & Hasan, R. (2018). Risk-Incorporated Trajectory Prediction to Prevent Contact Collisions on Construction Sites. *Journal of Construction Engineering and Project Management*, *8*(1), 10–21.

Raza, M., Chen, Z., Rehman, S. U., Wang, P., & Bao, P. (2018). Appearance based pedestrians’ head pose and body orientation estimation using deep learning. *Neurocomputing*, *272*, 647–659. https://doi.org/10.1016/j.neucom.2017.07.029

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rezazadeh Azar, E., & McCabe, B. (2011). Automated Visual Recognition of Dump Trucks in Construction Videos. *Journal of Computing in Civil Engineering*. https://doi.org/10.1061/(asce)cp.1943-5487.0000179

Rezazadeh Azar, E., & McCabe, B. (2012). Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in Construction*, *24*, 194–202. https://doi.org/10.1016/j.autcon.2012.03.003

Roberts, D., & Golparvar-Fard, M. (2019). End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. *Automation in Construction*, *105*. https://doi.org/10.1016/j.autcon.2019.04.006

Rudenko, A., Palmieri, L., & Arras, K. O. (2018). Joint Long-Term Prediction of Human Motion Using a Planning-Based Social Force Approach. *Proceedings - IEEE International Conference on Robotics and Automation*, 4571–4577. https://doi.org/10.1109/ICRA.2018.8460527

Ruff, T. (2006). Evaluation of a radar-based proximity warning system for off-highway dump trucks. *Accident Analysis and Prevention*, *38*(1), 92–98. https://doi.org/10.1016/j.aap.2005.07.006

Rundmo, T. (2001). Employee images of risk. *Journal of Risk Research*, *4*(4), 393–404. https://doi.org/10.1080/136698701100653259

Saleh, K., Hossny, M., & Nahavandi, S. (2017). Early intent prediction of vulnerable road users from visual attributes using multi-task learning network. *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*. https://doi.org/10.1109/SMC.2017.8123150

Saleh, K., Hossny, M., & Nahavandi, S. (2018). Intent Prediction of Pedestrians via Motion Trajectories Using Stacked Recurrent Neural Networks. *IEEE Transactions on Intelligent Vehicles*, *3*(4), 414–424. https://doi.org/10.1109/tiv.2018.2873901

Sanders, G., & Kaul, A. (2019). *Construction and Demolition Robots*. Retrieved from https://www.tractica.com/research/construction-demolition-robots/

Son, H., Choi, H., Seong, H., & Kim, C. (2019a). Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks. *Automation in Construction*, *99*, 27–38. https://doi.org/10.1016/j.autcon.2018.11.033

Son, H., Seong, H., Choi, H., & Kim, C. (2019b). Real-Time Vision-Based Warning System for Prevention of Collisions between Workers and Heavy Equipment. *Journal of Computing in Civil Engineering*, *33*(5). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000845

Su, X., Li, S., Yuan, C., Cai, H., & Kamat, V. R. (2014). Enhanced Boundary Condition–Based Approach for Construction Location Sensing Using RFID and RTK GPS. *Journal of Construction Engineering and Management*, *140*(10), 04014048. https://doi.org/10.1061/(asce)co.1943-7862.0000889

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *4*(January), 3104–3112.

Sutter, B., Lelevé, A., Pham, M. T., Gouin, O., Jupille, N., Kuhn, M., … Rémy, P. (2018). A semi-autonomous mobile robot for bridge inspection. *Automation in Construction*, *91*, 111–119. https://doi.org/10.1016/j.autcon.2018.02.013

Syed, A., & Morris, B. T. (2019). SSeg-LSTM: Semantic scene segmentation for trajectory prediction. *IEEE Intelligent Vehicles Symposium, Proceedings*, *2019-June*, 2504–2509. https://doi.org/10.1109/IVS.2019.8813801

Tang, S., Golparvar-Fard, M., Naphade, M., & Gopalakrishna, M. M. (2019). Video-Based Activity Forecasting for Construction Safety Monitoring Use Cases. *Computing in Civil Engineering 2019: Smart Cities, Sustainability, and Resilience - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2019*, 204–210. https://doi.org/10.1061/9780784482445.026

Teizer, J., Allread, B. S., Fullerton, C. E., & Hinze, J. (2010). Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system. *Automation in Construction*, *19*(5), 630–640. https://doi.org/10.1016/j.autcon.2010.02.009

Teizer, J., & Cheng, T. (2015). Proximity hazard indicator for workers-on-foot near miss interactions with construction equipment and geo-referenced hazard areas. *Automation in Construction*, *60*, 58–73. https://doi.org/10.1016/j.autcon.2015.09.003

Thaljaoui, A., Val, T., Nasri, N., & Brulin, D. (2015). BLE localization using RSSI measurements and iRingLA. *Proceedings of the IEEE International Conference on Industrial Technology*, *2015-June*(June), 2178–2183. https://doi.org/10.1109/ICIT.2015.7125418

Topak, F., Pekeriçli, M. K., & Tanyer, A. M. (2018). Technological Viability Assessment of Bluetooth Low Energy Technology for Indoor Localization. *Journal of Computing in Civil Engineering*, *32*(5), 04018034. https://doi.org/10.1061/(asce)cp.1943-5487.0000778

U.S.Bureau of Labor Statistics. (2013). An analysis of fatal occupational injuries at road construction sites 2003–2010. Retrieved June 3, 2019, from https://stats.bls.gov/opub/mlr/2013/article/pdf/an-analysis-of-fatal-occupational-injuries-at-road-construction-sites-2003-2010.pdf

U.S.Bureau of Labor Statistics. (2018). Employment, Hours, and Earnings from the Current Employment Statistics survey (National). Retrieved March 25, 2020, from https://beta.bls.gov/dataQuery/find?st=0&r=20&fq=survey:[ce]&more=0

US Department of Labor. (2016). Employer-reported workplace injuries and illnesses-2015. Retrieved June 3, 2019, from https://www.bls.gov/news.release/archives/osh_10272016.pdf

Vahdatikhaki, F., & Hammad, A. (2015). Dynamic equipment workspace generation for improving earthwork safety using real-time location system. *Advanced Engineering Informatics*, *29*(3), 459–471. https://doi.org/10.1016/j.aei.2015.03.002

Valero, E., Sivanathan, A., Bosché, F., & Abdel-Wahab, M. (2017). Analysis of construction trade worker body motions using a wearable and wireless motion sensor network. *Automation in Construction*, *83*, 48–55. https://doi.org/10.1016/j.autcon.2017.08.001

Völz, B., Behrendt, K., Mielenz, H., Gilitschenski, I., Siegwart, R., & Nieto, J. (2016). A data-driven approach for pedestrian intention estimation. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2607–2612. https://doi.org/10.1109/ITSC.2016.7795975

Wang, J., & Razavi, S. (2016a). Two 4D Models Effective in Reducing False Alarms for Struck-by-Equipment Hazard Prevention. *Journal of Computing in Civil Engineering*, *30*(6), 04016031. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000589

Wang, J., & Razavi, S. (2018). Spatiotemporal Network-Based Model for Dynamic Risk Analysis on Struck-by-Equipment Hazard. *Journal of Computing in Civil Engineering*, *32*(2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000732

Wang, J., & Razavi, S. N. (2016b). Low False Alarm Rate Model for Unsafe-Proximity Detection in Construction. *Journal of Computing in Civil Engineering*, *30*(2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000470

Xiao, Y., Kamat, V. R., & Menassa, C. C. (2019). Human tracking from single RGB-D camera using online learning. *Image and Vision Computing*, *88*, 67–75. https://doi.org/10.1016/j.imavis.2019.05.003

Xue, H., Huynh, D. Q., & Reynolds, M. (2018). SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-Janua*, 1186–1194. https://doi.org/10.1109/WACV.2018.00135

Yan, X., Li, H., Li, A. R., & Zhang, H. (2017). Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. *Automation in Construction*, *74*, 2–11. https://doi.org/10.1016/j.autcon.2016.11.007

Yan, X., Zhang, H., & Li, H. (2019). Estimating Worker-Centric 3D Spatial Crowdedness for Construction Safety Management Using a Single 2D Camera. *Journal of Computing in Civil Engineering*, *33*(5). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000844

Yang, J., Arif, O., Vela, P. A., Teizer, J., & Shi, Z. (2010). Tracking multiple workers on construction sites using video cameras. *Advanced Engineering Informatics*, *24*(4), 428–434. https://doi.org/10.1016/j.aei.2010.06.008

Yang, J., Vela, P. A., Teizer, J., & Shi, Z. K. (2011). Vision-Based Crane Tracking for Understanding Construction Activity. *Computing in Civil Engineering (2011)*, 258–265. https://doi.org/10.1061/41182(416)32

Yang, X., Li, H., Huang, T., Zhai, X., Wang, F., & Wang, C. (2018). Computer-Aided Optimization of Surveillance Cameras Placement on Construction Sites. *Computer-Aided Civil and Infrastructure Engineering*, *33*(12), 1110–1126. https://doi.org/10.1111/mice.12385

YouTube. (2019). Hospital construction. Retrieved April 7, 2019, from https://www.youtube.com/channel/UCEKwrM78pRv8WRcKvZNtE1w

Yu, X., & Ganz, A. (2010). Global identification of tracklets in video using long range identity sensors. *Proceedings - IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2010*, 361–368. https://doi.org/10.1109/AVSS.2010.46

Yu, Y., Guo, H., Ding, Q., Li, H., & Skitmore, M. (2017). An experimental study of real-time identification of construction workers' unsafe behaviors. *Automation in Construction*. https://doi.org/10.1016/j.autcon.2017.05.002

Yuan, C., Li, S., & Cai, H. (2016). Vision-Based Excavator Detection and Tracking Using Hybrid Kinematic Shapes and Key Nodes. *Journal of Computing in Civil Engineering*, *31*(1), 04016038. https://doi.org/10.1061/(asce)cp.1943-5487.0000602

Zhang, B., Zhu, Z., Hammad, A., & Aly, W. (2018). Automatic matching of construction onsite resources under camera views. *Automation in Construction*, *91*, 206–215. https://doi.org/10.1016/j.autcon.2018.03.011

Zhang, H., Yan, X., & Li, H. (2018). Ergonomic posture recognition using 3D view-invariant features from single ordinary camera. *Automation in Construction*, *94*, 1–10. https://doi.org/10.1016/j.autcon.2018.05.033

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334. https://doi.org/10.1109/34.888718

Zhao, Jian, Yoshida, R., Cheung, S. S., & Haws, D. (2013). Approximate techniques in solving optimal camera placement problems. *International Journal of Distributed Sensor Networks*, *9*(11), 241913.

Zhao, Jianyu, Seppänen, O., Peltokorpi, A., Badihi, B., & Olivieri, H. (2019). Real-time resource tracking for analyzing value-adding time in construction. *Automation in Construction*, *104*, 52–65. https://doi.org/10.1016/j.autcon.2019.04.003

Zhu, Z., Park, M. W., Koch, C., Soltani, M., Hammad, A., & Davari, K. (2016a). Predicting movements of onsite workers and mobile equipment for enhancing construction site safety. *Automation in Construction*. https://doi.org/10.1016/j.autcon.2016.04.009

Zhu, Z., Ren, X., & Chen, Z. (2016b). Visual Tracking of Construction Jobsite Workforce and Equipment with Particle Filtering. *Journal of Computing in Civil Engineering*, *30*(6), 04016023. https://doi.org/10.1061/(asce)cp.1943-5487.0000573

Zhu, Z., Ren, X., & Chen, Z. (2017). Integrated detection and tracking of workforce and equipment from construction jobsite videos. *Automation in Construction*, *81*, 161–171. https://doi.org/10.1016/j.autcon.2017.05.005

Zhuang, Y., Yang, J., Li, Y., Qi, L., & El-Sheimy, N. (2016). Smartphone-based indoor localization with bluetooth low energy beacons. *Sensors (Switzerland)*, *16*(5). https://doi.org/10.3390/s16050596

Ziebart, B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., … Srinivasa, S. (2009). Planning-based prediction for pedestrians. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, 3931–3936. https://doi.org/10.1109/IROS.2009.5354147

Zou, H., Xie, L., Jia, Q.-S., & Wang, H. (2014). Platform and Algorithm Development for a RFID-Based Indoor Positioning System. *Unmanned Systems, 02*(03), 279–291. https://doi.org/10.1142/s2301385014400068

PUBLICATIONS

Peer-reviewed journal publications

1. **Cai, J.**, & Cai, H. (2020). Robust Hybrid Approach of Vision-Based Tracking and Radio-Based Identification and Localization for 3D Tracking of Multiple Construction Workers. *Journal of Computing in Civil Engineering*, 34(4), 04020021.
2. **Cai, J.**, Jeon, J., Cai, H., & Li, S. (2020). Fusing Heterogeneous Information for Underground Utility Map Generation Based on Dempster-Shafer Theory. *Journal of Computing in Civil Engineering*, *34*(3), 04020013.
3. **Cai, J.**, Zhang, Y., & Cai, H. (2019). Two-step long short-term memory method for identifying construction activities through positional and attentional cues. *Automation in Construction*, *106*, 102886.
4. **Cai, J**., Gao, Q., Chun, H., Cai, H., & Nantung, T. (2019). Spatial Autocorrelation in Soil Compaction and Its Impact on Earthwork Acceptance Testing. *Transportation Research Record*, 2673(1), 332-342.
5. **Cai, J**., Li, S., & Cai, H. (2019). Empirical Analysis of Capital Structure Determinants in Infrastructure Projects under Public–Private Partnerships. *Journal of Construction Engineering and Management*, 145(5), 04019032.
6. Li, S., **Cai, J.**, Feng, Z., Xu, Y., & Cai, H. (2019). Government contracting with monopoly in infrastructure provision: Regulation or deregulation?. *Transportation Research Part E: Logistics and Transportation Review*, 122, 506-523.
7. Li, S., **Cai, J.**, & Cai, H. (2019). Infrastructure Privatization Analysis: A Public-Private Duopoly Game. *Transport Policy*.
8. Li, S., Hu, D., **Cai, J.**, & Cai, H. (2019). Real option-based optimization for financial incentive allocation in infrastructure projects under public-private partnerships. *Frontiers of Engineering Management*, 1-13.

Technical report

9. **Cai, J.**, Gao, Q., Chun, H., & Cai, H. (2019). *Pavement acceptance testing: Risk-controlled sampling strategy* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2019/08). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284316918

Peer-reviewed conference proceedings

10. **Cai, J.**, Yang, L., Zhang, Y., & Cai, H. (2020). Estimating the Visual Attention of Construction Workers from Head Pose Using Convolutional Neural Network-based Multi-task Learning. In *Construction Research Congress 2020* (accepted).
11. Zhang, Y., Cai, H., & **Cai, J**. (2020). CNN-based Symbol Recognition in Piping Drawings. In *Construction Research Congress 2020* (accepted).
12. **Cai, J**, Zhang, Y., & Cai, H. (2019). Integrating Positional and Attentional Cues for Construction Working Group Identification: A Long Short-Term Memory Based Machine Learning Approach. *Computing in Civil Engineering 2019*, 35–42.
13. **Cai, J.**, Li, S., & Cai, H. (2018). Accurate Mapping of Underground Utilities: An Information

Fusion Approach Based on Dempster-Shafer Theory. In *Construction Research Congress 2018* (pp. 712-721).

Poster presentation in Transportation Research Board 97th Annual Meeting

14. **Cai, J.**, Li, S., Gao, Q., Chun, H., Nantung, T., & Cai, H. (2018). *Optimal Sampling Strategy for Acceptance Decision in Highway Construction: A Cost-Benefit Analysis Approach* (No. 18-00218).