博士論文

# Study on Automated Occupational Hazards Identification in Construction and Decommissioning Sites of Nuclear Facilities based on Scene Graph Representation
# （シーングラフ表現に基づく原子力施設の建築および廃止措置現場における労働災害識別自動化の研究）


指導教員

出町　和之　准教授

学籍番号　　37-177300

陳　実

# Study on Automated Occupational Hazards Identification in Construction and Decommissioning Sites of Nuclear Facilities based on Scene Graph Representation

A dissertation submitted to
The University of Tokyo
In partial fulfillment of the requirements
For the degree of

*Doctor of Philosophy*

In

Nuclear Engineering and Management

By

*Shi Chen*

# Abstract

Construction sites are one of the most perilous environments where many potential hazards may occur. Decommissioning of nuclear facilities is an invasive process that presents industrial and chemical hazards as well as radiological ones, and indeed the non-radiological hazards generally represent the higher overall risk to workers. Besides, decommissioning the Fukushima Daiichi Nuclear Power Station (NPS) is a type of work that has never been done before. The entire site was contaminated with radioactive materials from the accident, and radiation dose levels were not low. At present, a multitude of Tokyo Electric Power Company (TEPCO) workers, manufacturers of nuclear reactors, construction companies, and their contractors are engaged in the decommissioning project of Fukushima Daiichi NPS and are consequently exposed to various health risks. However, compliance of regulatory rules is not strictly enforced among workers due to all kinds of reasons. Conventional on-site occupational safety monitoring, which relies heavily on on-site/off-site observers, is not sufficient to ensure the safety of workers due to human factors and human errors. Consequently, an automated on-site occupational hazards identification system is urgently needed.

Therefore, the objective of this work is to propose an regulatory-image inference model, which enjoys both perceptual and reasoning capabilities, to process regulatory rule sentences and images for on-site occupational hazards identification, and develop a robust and efficient real-time automated system to help to facilitate the safety monitoring work of workers to ensure the compliance of regulatory rules.

The first part of the main matter describes the framework of the regulatory-image inference model based on scene graph representation to drive the de-

velopment of this work, which is composed by (a) regulatory information representation module, (b) image information representation module, and (c) automated reasoning module for on-site occupational hazards identification.

In the second part, a regulatory information extraction approach is proposed based on Natural Language Processing (NLP) techniques and ontology modeling, together with an original hierarchical scene graph structure to address the complex representing relationships of requirement types. Based on the proposed approach, a novel automated regulatory rules processing system has been developed for regulatory information representation.

Thirdly, in contrast to commonly used object detection-based on-site image information representation approaches, this work originally adopts object detection together with individual detection using geometric relationship analysis. Specifically, it provides a solution for multi-hazard identification regarding viewpoint changes of on-site cameras and different individual postures of on-site workers.

Lastly, taking advantage of the scene graph structure, the automated reasoning module of the proposed regulatory-image inference model performs the integration of processed regulatory and image information. Additionally, a novel system has been developed based on the proposed model with the capabilities to handle on-site occupational hazards identification effectively.

The performance of the developed on-site occupational hazards identification system was experimentally evaluated on the validation dataset. The validating results indicate that the developed system is capable of identifying the hazards with high precision and recall rate while ensuring real-time performance to meet the industrial requirements.

# Acknowledgment

Foremost, I would like to offer my most sincere gratitude to my supervisor, Assoc. Prof. Kazuyuki Demachi for guiding me throughout my two years of master's course and three years of doctoral course. His patience and kindness helped me to overcome the difficulties and develop my research ideas. This thesis would not have been possible without his unwavering support in his most busy moments.

I would like to thank Prof. Naoto Kasahara for providing such insightful comments and incredibly helps, as well as Dr. Takuya Sato and Dr. Masakazu Ichimiya for generously sharing their experience and suggestions in improving my work.

The present work owes much to the discussions with the members of IIU Corporation and Taisei Corporation. Prof. Kenzo Miya (IIU Corporation) offered key insights and valuable ideas in improving the originality of this work, which I am dearly thankful for. I am also very grateful to Dr. Yuji Ijiri, Mr. Kyohei Nishiyama, Mr. Haruo Nagamine, and Mr. Kai Shinohara (Taisei Corporation) for generously sharing their experience and suggestions for on-site management. I would like to thank Mr. Manabu Tsunokai, Mr. Takahiro Kamata, and Mr. Seiichi Tanaka (IIU Corporation) for kind support in the demonstration experiments and software development.

I would also like to send my best regards to the Ph.D. defense committee that comprises Prof. Mitsuru Uesaka, Prof. Shunichi Suzuki (The University of Tokyo), and Dr. Masano Hori (Japan Atomic Energy Agency) for their inspiring discussions and pertinent comments to help improve the quality of this work.

Thank you very much to my laboratory colleagues Mr. Masaki Sudo, Mr.

Takaaki Hirano, and Mr. Satoshi Yasuda, for performing experiments and sharing their ideas with me. Particularly, this work much benefited from Mr. Daisuke Miki's detailed and helpful advice. I cannot forget the long nights and discussions with Mr. Jinqi Lyu and Ms. Maoxin Tang. Without them the experience would not have been nearly as enjoyable. I am also indebted to our secretaries Ms. Ritsuko Mitsubayshi and Ms. Mariko Wada, for their kind help and support in my daily life. Specifically, I would like to thank Ms. Linlin Ni for her continuous support.

I am grateful for the funding of this work from Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research Grant No. 19K05324 and the support of the World-leading Innovative Graduate Study Program Co-designing Future Society (WINGS CFS).

Lastly, I would like to reserve my utmost gratitude to my mother and father in China, who supports me wholeheartedly from far away. They are the foundation that propels me to pursue my dreams and without them I would not be the person I am today.

-Shi Chen

# Contents

# List of Figures

# List of Tables

# Nomencalture

| | |
|---|---|
| PPE | Personal Protective Equipment |
| NPS | Nuclear Power Station |
| TEPCO | Tokyo Electric Power Company |
| NHU | Non-hardhat-use |
| AEC | Architecture, Engineering and Construction |
| NPU | Non-ppe-use |
| NLP | Natural Language Processing |
| CART | Classification and Regression Tree |
| POS | Part-of-speech |
| BIM | Building Information Modeling |
| BFS | Breadth-first Search |
| CNN | Convolutional Neural Network |
| PAF | Part Affinity Field |
| ROI | Region of Interest |
| NMS | Non-maximum Suppression |
| IOU | Intersection Over Union |
| Adam | Adaptive Moment Estimation |
| ISF | Interim Storage Facility |

# Chapter 1

# Introduction

In this chapter, the current situation of occupational safety in the construction industry and decommissioning sites of nuclear facilities is introduced as the background of this work, followed by the literature survey and the statement of the problem regarding on-site occupational safety monitoring. Lastly, the scope of this work is given.

## 1.1 Occupational safety in construction industry

Construction work is much more dangerous than most other occupations, where many potential hazards may occur. According to the United States' Bureau of Labor Statistics (BLS), the number of construction fatalities in the United States has gradually increased from 933 to 1013 between 2014 and 2017 [1]. Similarly, 306 fatal construction accidents occurred in Japan, which represented 34% of all fatal accidents in 2018 (Figure 1.1). Thus, the Ministry of Health, Labor and Welfare (MHLW) of Japan is aiming to reduce construction fatalities by at least 15% (relative to the 2017 level) by 2022 [2].

The combination of different factors always causes the construction fatalities, and the majority of these fatalities could be prevented if workers wore appropriate personal protective equipment (PPE), e.g., hard hats, gloves, body harness [3] and followed on-site regulatory rules.

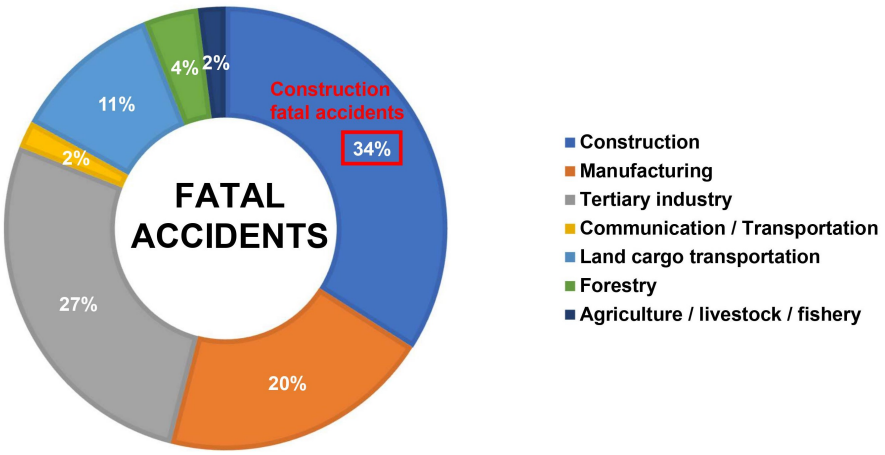



**Figure 1.1: Fatal accidents in Japan (CY2018).**

(1) Head injuries: The consequences of head injuries caused by falling from height or being stuck by vehicles and other moving plants and equipment are one of the most serious of all construction accidents. A total of 2,210 construction fatalities occurred in the United States because of traumatic brain injury (TBI), which represented 25% of all construction fatalities during 2003 and 2010 [4]. The Occupational Safety and Health Administration (OSHA) in the United States stipulated that workers working in areas where there is a possible danger of head injury from impact, or from falling or flying objects, or from electrical shock and burns shall be protected by hard hats [5].

(2) Eye injuries: Construction-related occupational eye injuries are an important cause of vision loss. According to the National Institute for Occupational Safety and Health (NIOSH), an average of 2,000 United States workers require medical treatment for job-related eye injuries every day [6]. The majority of construction-related eye injuries are preventable. The reasons cited for the majority of eye injuries include the

non-wearing of available eye protection or wearing of inappropriate eye protection for the current task [7].

(3) Lung and airway diseases: Fine specks of dust and particles, gases, and vapors can be produced when using machine tools, and silica dust from bricks can cause lung and airway diseases such as emphysema, bronchitis, and silicosis, and may increase cancer risks. PPE, such as respirators or dust masks, are used to controls these hazards [8]. OSHA indicates the workers shall be ensured to wear eye or face protection when exposed to eye or face hazards from flying particles, molten metal, liquid chemicals, acids or caustic liquids, chemical gases or vapors, or potentially injurious light radiation [9].

(4) Hands and forearms injuries: Improper handling grinder can be a dangerous power tool, hands and forearms injure results when the workers using the grinder loses control of it. OSHA indicates the workers shall use two hands to operate the grinder. One hand should grip the handle and dead-man switch (if provided), while the other hand supports the weight of the tool [10].

(5) Falls from height: Construction workers who are six feet or more above lower levels are at risk for serious injury or death if they should fall. According to BLS, there were 320 fatal falls to a lower level out of 1,008 construction fatalities in 2018, which indicates fall accidents are a substantial burden and an impediment to accomplishing occupational safety in the construction industry [1].

## 1.2 Occupational safety in decommissioning sites of nuclear facilities

Great East Japan Earthquake occurred on March 11, 2011, causing damage to the electric power supply lines to Fukushima Daiichi Nuclear Power Station (NPS), and the following tsunami caused substantial destruction of the operational and safety infrastructure on the site. The combined effects

led to the loss of off-site and on-site electrical power, which resulted in the loss of cooling functions at the operating reactors and the spent fuel pools. Large amounts of radioactive materials were released, and the problem of radioactive contamination has severely affected the lives of people and shocked many countries throughout the world [11]. Units 1 to 4 of Fukushima Daiichi NPS were damaged during the disaster, and all reactor units of the plant were brought to cold shutdown in December 2011. Units 5 and 6 were permanently shut down in December 2013. In April 2014, Tokyo Electric Power Company (TEPCO) formed the Fukushima Daiichi Decontamination & Decommissioning Engineering Company, in partnership with the Japanese Government and key contractors, to implement the decommissioning project, which is expected to take up to 40 years for completion.

Decommissioning of nuclear facilities is an invasive process that presents industrial and chemical hazards as well as radiological ones, and indeed the non-radiological hazards generally represent the higher overall risk to workers [12]. Besides, decommissioning the Fukushima Daiichi NPS is a type of work that has never been done before. The entire site was contaminated with radioactive materials from the accident, and radiation dose levels were not low. At present, a multitude of TEPCO workers, manufacturers of nuclear reactors, construction companies, and their contractors are engaged in the decommissioning project of Fukushima Daiichi NPS and are consequently exposed to various health risks. Based on the progress of measures to reduce environmental radiation dose, the Fukushima Daiichi NPS site has been divided into three zones according to contamination levels from March 8, 2016, and workers are indicated to wear appropriate PPE for each zone (as illustrated in Figure. 1.2 [13]):

(1) G Zone: in addition to G zone in Figure. 1.2, a partial area of the common pool building 2nd, 3rd floors are also covered. General work uniforms are required.

(2) Y Zone: within the yellow dotted line of Y zone in Figure. 1.2, works involving contamination such as works on concentrated salt water, etc. are performed. Patrols and on-site surveys at the time of work planning


**Figure 1.2: PPE requirements for each zone in Fukushima Daiichi NPS.**

etc. shall be equipped for G zone. Other than Figure. 1.2, in case of working on high concentration dust work (building dismantling, etc.), on transfer tank of concentrated brine, etc. in G zone, Y zone will be set temporarily. Coveralls are required.

(3) R Zone: in Units 1 to 3 reactor building, turbine building of Units 1 to 4, the area surrounding residence water in the peripheral building. Anorak and full-face masks are required.

TEPCO indicates that each zone has different PPE requirements that should be adhered to. Decommissioning workers moving from a higher-level contamination zone to a lower-level contamination zone are required to remove their PPE in changing rooms.

However, as TEPCO mentioned in the safety report of Fukushima Daiichi NPS, the number of occupational injuries increased by 23.5% compared to 2017, which including two serious injuries (more than 14 days off). TEPCO indicated it is necessary to review and devise efforts for the control of occupational injuries that occurred in the Fukushima Daiichi NPS decommissioning project.

## 1.3 Statement of the problem

Nonetheless, the construction and decommissioning workers do not precisely follow the on-site safety regulations due to all kinds of reasons, even if they have been previously educated and trained. And even though the signs and partitions mark the R zone and Y zone areas in Fukushima NPS, intrusions into these areas may occur within only a few minutes of carelessness. Besides, TEPCO has contracted various tasks to more than 20 companies (primary contractors), and each of these has outsourced parts of tasks to multiple layers of subcontractors. This complex structure may hinder the consistent implementation of occupational health rules [14]. Thus, on-site occupational safety monitoring is considered as an important part of safety management, which needs to cover the following items:

(1) Individual behaviors monitoring to avoid improper behaviors which may potentially cause accidents (e.g., use a single hand when operating a grinder);

(2) Proper PPE use (e.g., hard hats, body harness) monitoring.

Conventional on-site occupational safety monitoring is carried out by on-site/off-site observers and relies heavily on the "eye of human", which is not sufficient to protect workers since human factors and human errors. Thus, how to transfer responsibility from "eye of human" to "eye of technology", an automated on-site occupational hazards identification system which of the effectiveness of designs in reducing human factors error and is able to carry out occupational hazards identification to predict the prevention of all kinds of accidents accurately, needs to be taken into consideration.

In 2015, the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) of Japan announced to start integration of construction and ICT, "i-Construction" [15], which attempt to improve the productivity of construction sites and provide support to workers. To realize a world-leading "Super Smart Society" (Society 5.0), the Japanese government and construction industry are currently working as one to achieve ICT-driven technological innovation. However, seldom works have been conducted in Japan for the concern of on-site worker occupational safety enhancement using ICT-driven technology since the development of "eye of technology" can face the difficulties posed by the complicated individual postures of workers in multiple hazards identification, together with a significant amount of regulatory rules to address. Furthermore, the developed system needs to be capable of meeting the industrial requirements of real-time processing.

## 1.4 Related works

At present, several approaches have been investigated for automated occupational hazards identification [16–24] which mainly focus on automatic identification of proper PPE use and can be divided into two categories: sensor-based approaches and vision-based approaches.


**Figure 1.3: A passive RFID portal for on-site PPE management proposed by Kelm et al.**

### 1.4.1 Sensor-based approaches

Sensor-based detection primarily relies on remote locating and tracking techniques, such as radio frequency identification (RFID) and wireless local area networks (WLANs). Kelm et al. [16] designed a mobile RFID portal for checking PPE compliance of personnel (Figure 1.3). The RFID readers were located at the construction site entrance, and therefore only those who enter the construction site are checked, while workers in other areas are not. Additionally, the tagging of PPE with a worker's identification card only indicates that the distance between the worker and PPE is close but unable to identify whether the PPE is being worn, held, or has been placed on the ground. Barro-Torres et al. [17] introduce a novel Cyber Physical System (CPS) to monitor how PPE is worn by workers in real-time. Rather than being located at the construction site entrance, their sensors were integrated into the clothing of workers for constant monitoring. However, same as [16], this approach is not possible to identify whether a worker is wearing a hard hat or is just close to it. As an improvement, Dong et al. [18] developed the real-time location system (RTLS) and virtual construction for a worker's location tracking to decide whether the worker should wear a hard hat and to transmit a warning (Figure 1.4). To identify whether a hard hat was being worn, a pressure sensor was placed in the hard hat, and then the pressure information was transmitted via Bluetooth for monitoring. However, the implementation of the system (pressure sensor placement for each hard hat) is time and cost consuming. Generally, existing sensor-based approaches

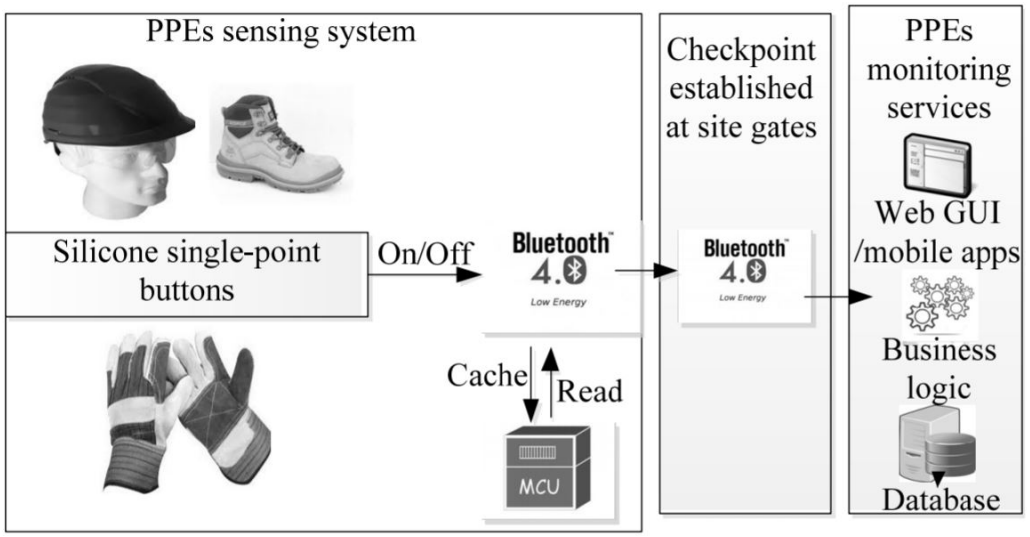



**Figure 1.4: A real-time location system for on-site PPE management proposed by Dong et al.**

relying on physical tags or sensors employed in PPE have difficulties in identifying whether any individuals on the construction sites are wearing PPE or not. Besides, the practical use of the tags or sensors will lead to high costs with massive productions.

### 1.4.2 Vision-based approaches

Vision-based approaches are nonintrusive and less device-intensive because of the wide application of surveillance cameras on construction sites. Shrestha et al. [19] use edge detection algorithms to recognize the edge of objects inside the upper head region where a hard hat may be recognized. This approach also relies on the recognition of facial features, where workers who turn their face away from the cameras cannot be recognized. Park et al. [20] proposed a vision-based non-hardhat-use (NHU) detection approach that detects both a human body and a hard hat simultaneously in each video frame using background subtraction and the histogram of oriented gradients (HOG) features (Figure 1.5). The detected human body region and hard hat region are then matched for the detection of NHU. However, the workers with various postures (e.g., crouching down, bending, and sitting) or occlusion

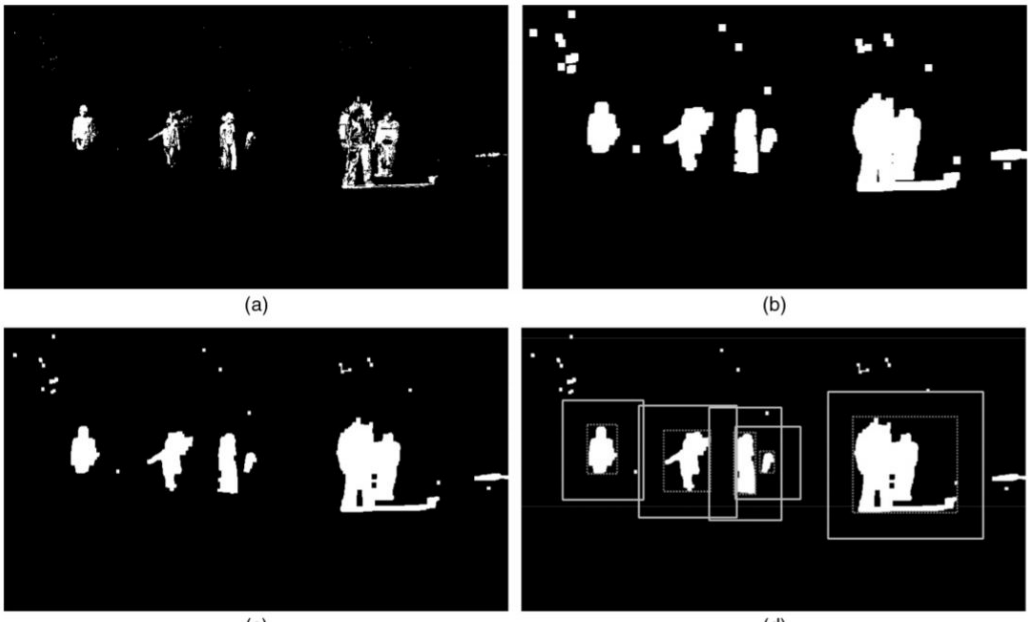



**Figure 1.5: A HOG-based approach for on-site PPE management proposed by Park, et al. (a) background subtraction; (b) dilation; (c) erosion; (d) rectangle fitting**

could not be successfully detected. Additionally, the proposed approach relies on background subtraction, which makes it unable to monitor the workers when they just stand at the site without any movement. In general, these approaches rely heavily on hand-crafted features to detect individuals on construction sites. Consequently, they may fail in the cases of complicated scenes with weather variability, different viewpoints, and occlusions.

Recently, deep learning-based object detection methods have shown remarkable performance on most visual tasks in the architecture, engineering and construction (AEC) industry. Fang et al. [21] proposed an approach to detect construction workers' NHU based on Faster R-CNN automatically. A total of 81,000 image frames were collected from various construction to train the Faster R-CNN model. The worker-of-interest (WOI) in the image was annotated as the ground truth for training. In the inference phase, the NHU workers were detected, and the rest were considered the background. However, WOI is not robustness to be applied to multiple non-ppe-use (NPU) identification. Wu et al. [22] deployed a Single Shot Multibox

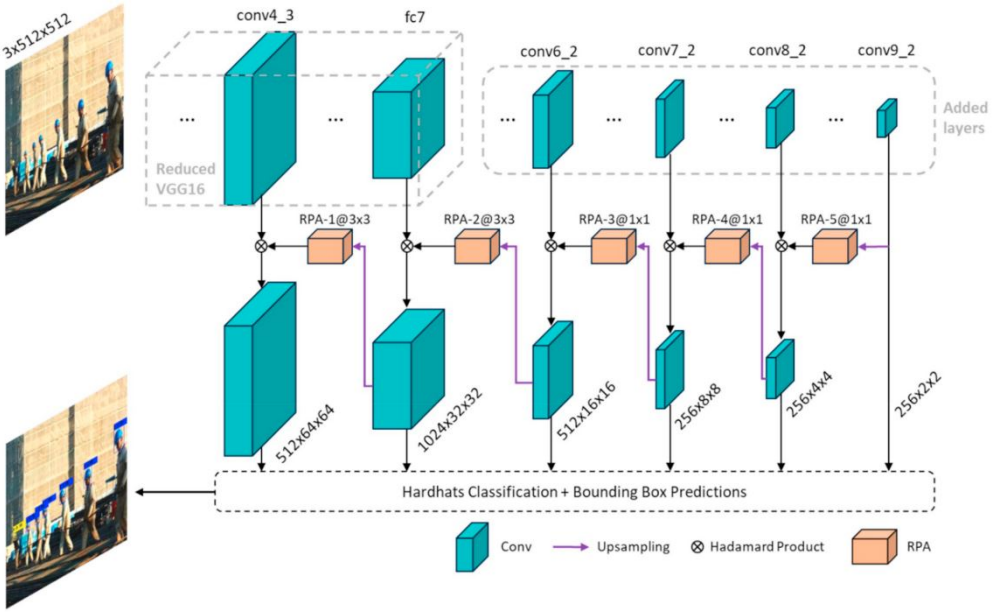



**Figure 1.6: Architecture of one-stage hardhat wearing detection model proposed by Wu et al.**

Detector (SSD)-based model combined with the presented reverse progressive attention (RPA) to propagate context information back to bottom layers discriminately (Figure 1.6). A benchmark dataset GDUT-HWD was generated by downloading Internet images retrieved by search engines to train the SSD-RPA model. Same as [21], this approach could not meet the requirements for multi-task learning on an object detection task. Nath et al. [23] introduced and tested models built on YOLOv3 architecture to verify PPE (hardhat and vest) compliance of workers. Three approaches were verified concerning different classifiers (e.g., decision tree, VGG-16, ResNet-50, Xception, or Bayesian). However, this approach is susceptible to occlusion, poor illumination, and blurriness. Xiong et al. [24] developed an Automated Hazards Identification System (AHIS) to evaluate the operation descriptions generated from site videos against the safety guidelines extracted from the textual documents with the assistance of the ontology of construction safety. Two types of significant hazards, i.e., failing to wear a hardhat and walking beneath the cane, were successfully identified. However, this work re-

quires manual effort to extract regulatory information and encode them in a computer-processable format, which can be time-consuming, costly, and error-prone. In general, even though enjoy the remarkable visual recognition ability of deep learning-based object detection methods, these approaches are still struggling with the difficulties of multi-hazard identification, different individual postures, and automated regulatory information extraction. Besides, none of the approaches mentioned in this section address individual behaviors monitoring.

## 1.5 Objectives of this work

In this section, the scope of this thesis on automated occupational hazards identification in construction and decommissioning sites of nuclear facilities will be detailed. Notably, the literature survey has indicated four major problems needing urgent solutions for academic and engineering purposes.

Firstly, existing regulatory rules in AEC domains are mostly documented with natural language sentences which need further processing for automated information understanding. However, seldom works have been done for automated regulatory information extraction and representation [25–29], most of which were focusing on rules classification for knowledge management. Besides, the state-of-the-art vision-based on-site occupational hazards identification approaches [21–23], are all case-based solutions which provide no interface to regulatory information processing. Xiong et al. [24] explored the possibility of processing regulatory information to indicate the potential hazards for on-site vision-based identification, however, this work provided no solution for automated regulatory information processing.

Secondly, the surveyed vision-based on-site occupational hazards identification approaches [21–24] took advantage of the end-to-end deep learning-based object detection model to estimate the individual who is using PPE or not using PPE in the obtained image. However, these works were focusing only on the single-hazard identification task (e.g., NHU) by a binary classification, i.e., "individual using PPE" and "individual not using PPE". When it comes to multi-hazard identification tasks, e.g., n-classes proper

PPE use identification, the outputs would be increased to $2^n$, which may lead to the difficulty for training data preparation and reduction of detecting performance.

Thirdly, when preparing training data, both [21–24] annotated regions of individual together with PPEs as an object for detection. In such case, the object detection performance may also be affected by viewpoint changes of on-site surveillance cameras and different individual postures of on-site workers because of the lack of the training samples.

Additionally, considering the real situation for on-site implementation, the feasibility and usability of the proposal are required, e.g., real-time processing. Besides, the best-known disadvantage of deep learning models is their "black box" nature: you don't know how or why your model came up with a certain output, which brings unexplainable and uncertainty for industrial application.

In order to address the aforementioned limitations, the work presented in the current thesis is devoted to the proposal and development of a unified model for occupational hazards identification with the concern of both perceptual and reasoning capabilities to automatically identify multi-hazard, which covers both proper PPE use and individual behaviors identification, in construction and decommissioning sites of nuclear facilities.

(1) As the first step toward the scope, a regulatory rules processing approach is developed to automatically extract and represent the key regulatory information that is intended to indicate the potential occupational hazards which need to be identified. Its development satisfactorily solves the first problem for providing an automated regulatory information processing solution.

(2) The second part of the scope is made to develop an image scene information understanding approach to process obtained on-site images. It considers the combining of deep learning-based object detection and individual detection model using geometric relationships and is able to address multi-hazard tasks for both proper PPE use and individual behaviors identification under different individual postures with an inter-

pretable explanation. The development of the image scene information understanding approach achieves the goal of automated image processing with a settlement of both the second and the third problems.

(3) Additionally, an automated reasoning approach is implemented to provide the integration of the processed regulatory and image information and perform hazards identification.

(4) Specifically, attempts are made to improve the robustness and efficiency of the approach to meet the industrial requirements for real-time processing in various environmental conditions, as it is figured out by the last problem.

## 1.6 Outline of this thesis

The main matters of this thesis are organized as follows.

In Chapter 2, as the fundamental of this work, the information representation structure for regulatory rules and on-site image processing are discussed. Based on the information representation structure, an regulatory-image interface model is proposed to drive the development of this work. The proposed model consists of (a) regulatory information representation module for regulatory rules processing, (b) image information representation module for on-site image processing, and (c) automated reasoning module for hazards identification.

In Chapter 3, a Natural Language Processing (NLP) and ontology-based regulatory information extraction approach is proposed, together with a novel hierarchical scene graph structure, to form the regulatory information representation module. Based on the proposed approach, a novel automated regulatory rules processing system is developed.

In Chapter 4, the proposed image information representation module, which deploys geometric relationship analysis to perform the combination of deep learning-based object detection and individual detection model to conduct scene graph representation, is detailed. At present, the proposed approach is able to process four types of individual-object relationships: (a)

individual-head protection PPE, (b) individual-grinder, (c) individual-glove, and (d) individual-body harness.

In Chapter 5, the proposed automated reasoning module to integrate the processed regulatory and image information for hazards identification is detailed. Based on the proposed approach, a novel real-time on-site hazards identification system has been developed.

Chapter 6 describes the experiments to evaluate the performance of the proposed regulatory-image inference model for on-site hazards identification. Firstly, ten construction/decommissioning regulatory rules are selected to validate the performance of the proposed regulatory information representation approach. Subsequently, the image datasets are created to train the object detection model and certify the performance of the proposed model in image information representation and hazards identification. Furthermore, the on-site hazards identification and computational efficiency analysis results using the developed real-time on-site hazards identification system are reported.

Finally, principal conclusions and future perspectives are summarized in Chapter 7

# Chapter 2

# Regulatory-image interface

This work aims at proposing and developing a novel approach for automated compliance checking to identify on-site occupational hazards, which is implemented to process on-site images and regulatory rule documents automatically. To this end, a model, which enjoys both perceptual and reasoning capabilities, needs to be proposed and implemented as the interface of regulatory rule sentences and images. The regulatory-image interface takes regulatory rule sentences and the on-site images as inputs, which are processed for information understanding and transformed into the same data structure, and outputs the identified occupational hazards. Based on the above concepts, the sketch of the regulatory-image interface model is illustrated in Figure 2.1, which consists of the regulatory information representation module for regulatory rule sentence processing and image information representation module for on-site image processing. Finally, the outputs from the previous modules are feed into the automated reasoning module to perform the occupational hazards identification. In this chapter, the structure to represent the information of sentences and images is discussed, followed by the details of the proposed regulatory-image interface model.

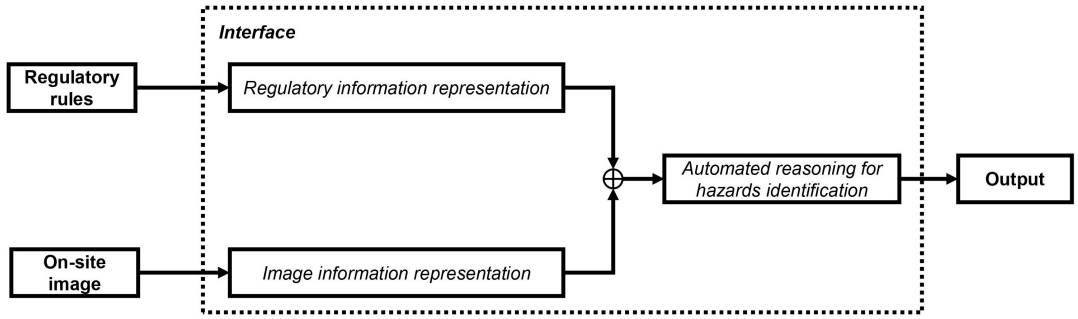



**Figure 2.1: The sketch of the proposed regulatory-image interface model.**

## 2.1 The information representation structure

Considering the automated compliance purpose using the processed outputs of the regulatory information representation module and the image information representation module, the structure to represent the information of sentences and images is fundamental to establish the regulatory-image interface. The requirements of the information representation structure are as follows:

(1) Easy to visualize and represent the richness of entities and relationships between entities in the image or sentences;

(2) Robust to be extended and updated to meet additional requirements from new regulatory rules;

(3) Relatively low time and space complexity.

To meet these requirements, four different structures (ledger, decision tree, word embedding, and scene graph) have been reviewed and validated in this works and are introduced in the following sections.

### 2.1.1 Ledger

A ledger used in hazards identification is a log-centric database that provides non-relational mechanisms for storage and retrieval of immutable records. As demonstrated in Figure 2.2, the records stored in the ledge database are the occupational hazards scenarios manually created from regulatory documents and described by different attributes:

(1) *Area*: working area or workspace of obtained on-site images or described in regulatory rule sentences (e.g., indoor);

(2) *Action condition*: action taken by the individuals in obtained on-site images or described in regulatory rule sentences (e.g., operating grinder, operating drill);

(3) *Perceived information*: proper/improper use of PPE/tool information perceived by the vison-based system or described in regulatory rule sentences (e.g., not use safely glasses);

(4) *Other information*: provides supplemental information for detailed description (e.g.,human body);

(5) *Hazardous*: indicates whether the record describes a hazardous scenario or not.

The detected information from on-site images is also structured as a record with similar attributes (*Area*, *Action condition*, *Perceived information*, and *Other information*) as the ledger. Following this, occupational hazards identification is performed by retrieving the records in the ledger to find whether a record with similar attributes is available or not.

However, the storage and management of records with different attributes require high space complexity. Besides, performing hazards identification using a high time complexity record retrieval is a time-consuming and inefficient approach.

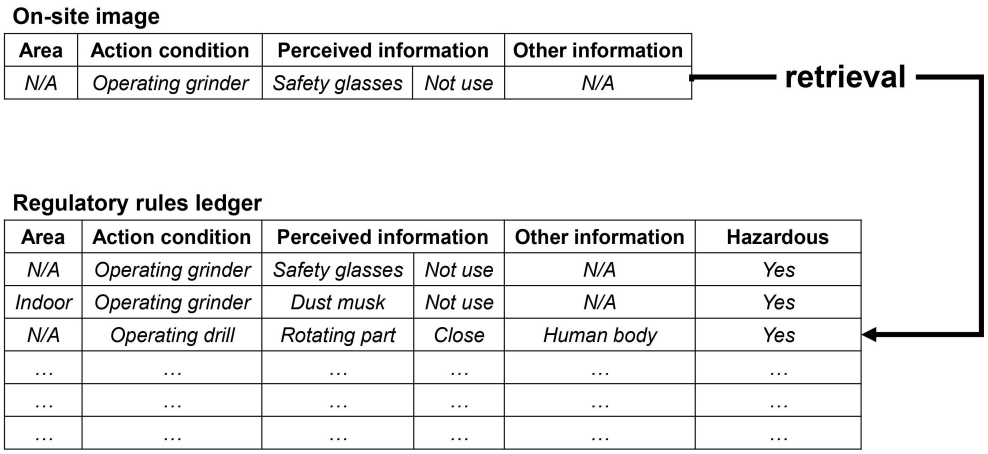



**Figure 2.2: An example of the ledger-based approach for hazard identification.**

### 2.1.2 Decision tree

The decision tree algorithm uses a tree structure for layer-by-layer reasoning to achieve the final classification. The decision tree consists of the following elements:

(1) *Root node*: it represents the entire sample and is further divided into more homogeneous sets;

(2) *Internal node*: a sub-node which splits into further sub-nodes;

(3) *Leaf node*: it represents the result of the decision tree and does not split further.

As an inductive learning algorithm, the decision tree is focusing on how to transform seemingly disordered and messy known instances into a tree model that can predict unknown instances through some technical means, Each path from the *root node* to the *leaf node* represents a decision rule. When performing decision making, a particular attribute value is used to judge at the internal nodes of the tree for which branch node to enter, and the decision result is obtained until reaching the leaf node.

Classification and regression tree (CART) algorithm is an effective non-parametric classification and regression method to create a decision tree. The CART algorithm was first proposed by Breiman et al. [30] and has been widely used in the field of statistics and data mining technology. It constructs prediction criteria in a completely different way from traditional statistics. It is given in the form of a binary tree and is easy to deploy and interpret. Specifically, the CART algorithm employs *Gini* index as a metric for classification tasks, which provides an indication of how "pure" the *leaf nodes* are (how mixed the training data assigned to each node is). During the learning process, in each created node $L_q$ a particular subset $S_q$ of the training dataset $S$ is processed. If either the list of available attributes in the node contains only one element or all elements of set $S_q$ are of the same class, the node is identified as a *Leaf node* and the split is stopped. Otherwise, an attribute with the lowest *Gini* index is chosen for a split. The *Gini* index of $S_q$ is given by:

$$Gini(S_q) = 1 - \sum_{k=1}^{K} p(k|S_q)^2 \tag{2.1}$$

where $K$ is the number of attributes and $p(k|S_q)$ is the probability of attribute $k$ appearing in $S_q$:

$$p(k|S_q) = \frac{n_k}{N} \tag{2.2}$$

where $N$ is the number of the elements of $S_q$ and $n_k$ is the number of elements belonging to attribute $i$.

As a previous exploration of this work, a decision tree was created using the CART algorithm based on a manually created regulatory rules dataset to perform occupational hazards identification (Figure 2.3).

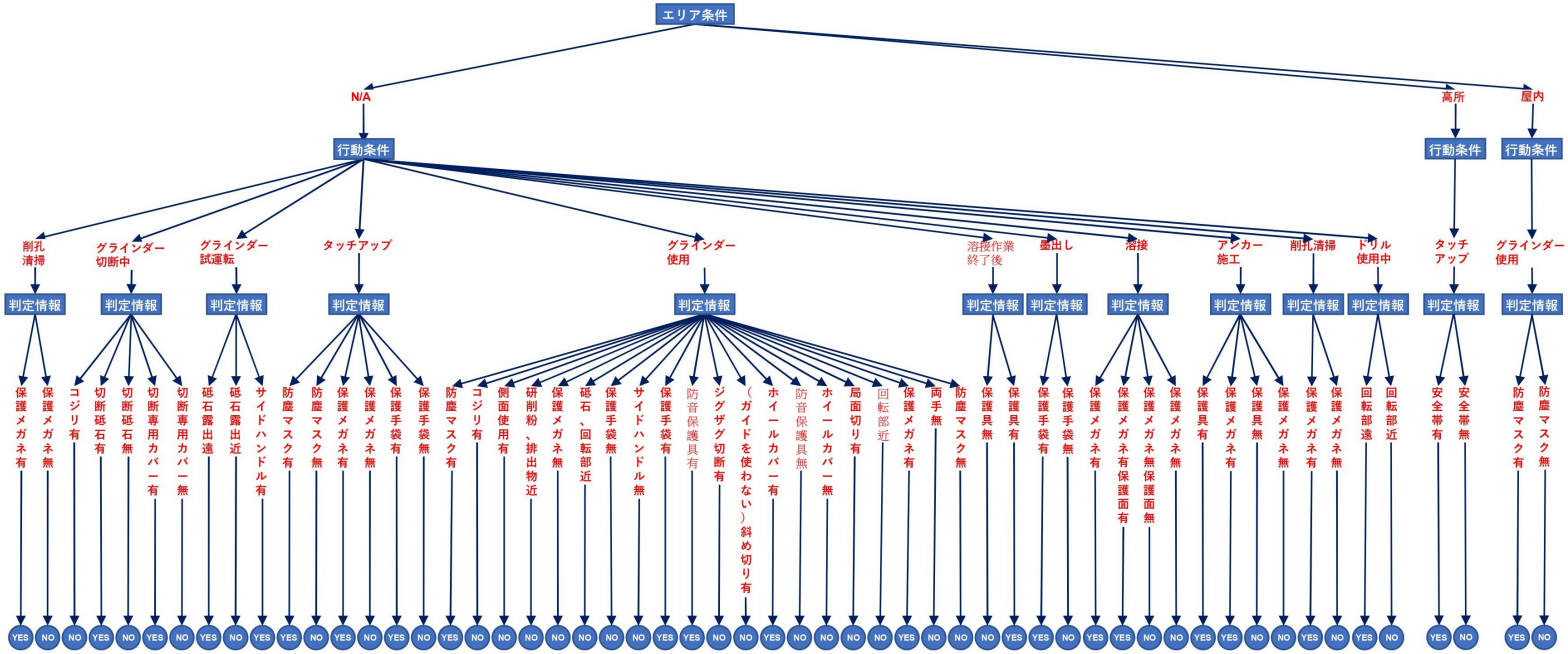



**Figure 2.3: An example of the decision tree-based approach for hazard identification.**

However, the manually created dataset is required for training a decision tree, while both positive and negative samples are necessary, and it can be computationally expensive to train (at each node, each candidate splitting field must be sorted before its best split can be found). Besides, a decision tree is prone to errors due to the overfitting of the training dataset containing a relatively small number of training examples.

### 2.1.3 Word embedding

Word embedding is a mathematical embedding from a high-dimensional space whose dimension is the number of all words into a continuous vector space with much lower dimensions, and each word or phrase is mapped as a vector on the real field. That is, by performing word embedding, uncalculable and unstructured words can be transformed into calculable and structured vectors.

To perform word embedding, Word2Vec [31] is a framework for learning semantic knowledge from a large amount of text corpus to make semantically similar words extremely close in the space through an embedded space based on the Distributional Hypothesis: "words that occur in the same contexts tend to have similar meanings" [32], and it is widely used in word embedding for NLP. Skip-gram algorithm, which uses the distributed representation of the input word to predict the context, is employed as the architecture to train a Word2Vec model by using a simple neural network with one hidden layer (Figure 2.4 [33]). A one-hot vector is used to represent the input word, while a one-hot vector representing the output word is feed into the output layer with a softmax regression classifier.

As illustrated in Figure 2.5, to implement a word embedding-based approach for on-site occupational hazard identification, the sentences from regulatory rules and occupational hazards scenarios are embedded in the word vector space that makes each rules representing as a directed line constructed by the word vectors. Subsequently, the detected information from the on-site images is also structured as a directed line in the vector space. Finally, by comparing the distances between the line of on-site images and the lines of

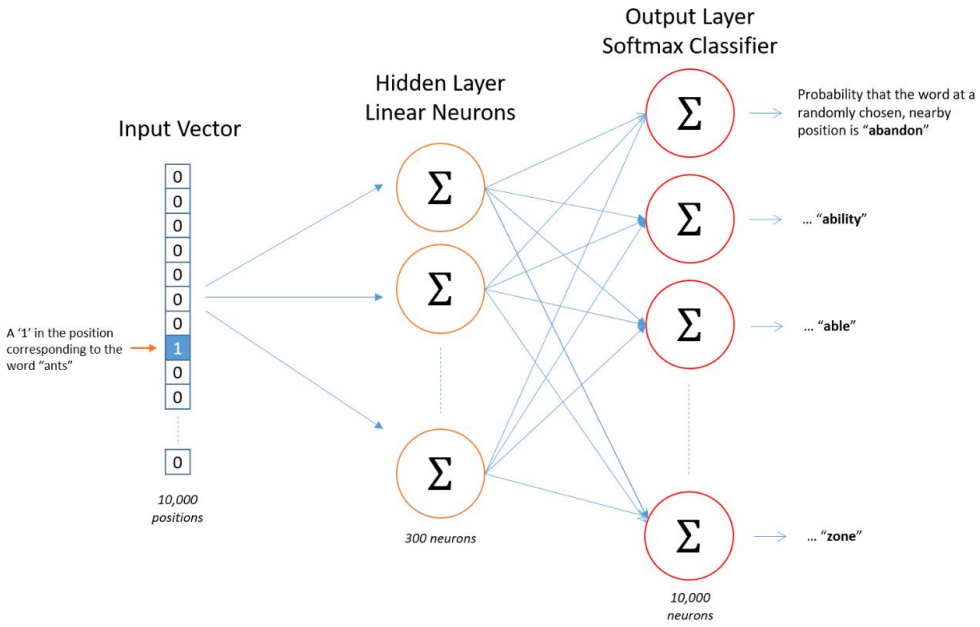



**Figure 2.4: Neural network architecture of skip-gram algorithm.**

the regulatory rules, the hazards can be identified.

However, word embedding is limited in handling words with multiple meanings, which are conflated into a single representation (a single vector in the semantic space). That is, polysemy and homonymy are not appropriately handled by word embedding. Besides, word embedding is actually an "anti-Occam's Razor" approach for performing occupational hazards identification: it makes a simple task even more complicated. Due to these limitations, the word embedding-based hazards identification approach is not shown satisfactory performance in the current experiments of this work.

### 2.1.4 Scene graph

Images are more than a collection of entities and their attributes; they represent the relationships among interconnected entities. Also, the requirements in regulatory rules can be represent based on the entities and their relationships. Recently, graph-structured methods have developed to address structured representations of visual scenes [34–39]. Scene graph, which is built on a simple directed graph structure, is a detailed and formal represen-

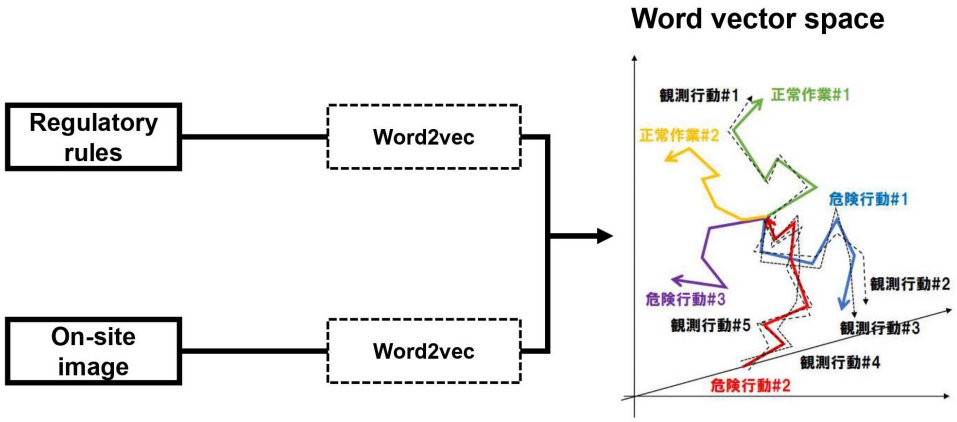



**Figure 2.5: An example of the word embedding-based approach for hazard identification.**

tation for image compositions, where encodes entities as vertices connected via pairwise relationships as edges (Figure 2.6 [24]). Given a scene graph $G(V, E)$, $V$ is the set of vertices to represent the entities in an image and $E = \{\{\mu, \nu\} : (\mu, \nu) \in V^2, \mu \neq \nu\}$ is the set of edges to represent the relationships between the entities. The notion of scene graph visualizes and represents the richness of objects and relationships between objects that can exist in an intuitive graph-structure. Besides, scene graph representation is robust to be updated or extended, which is useful for meeting new requirements from new regulatory rules. Another advantage of scene graph representation is its efficiency since graph structure has relatively low space complexity for storage and low time complexity for retrieval, which is crucial for real-time processing.

Accordingly, enjoying the merits of robust representation and computing efficiency, the scene graph is deployed as the basic notion for information representation structure in this work to form an regulatory-image interface model to encode on-site images and regulatory rules and further perform the reasoning for occupational hazards identification.


**Figure 2.6: An example of the scene graph-based approach for hazard identification.**

## 2.2 Proposed regulatory-image interface model

Based on the concepts of scene graph representation, the modules in the sketch of the proposed regulatory-image interface model shown in Figure 2.1 can be updated:

(1) *Regulatory information representation* module for regulatory rules processing. An NLP-based automated regulatory information extraction approach is proposed, together with a novel hierarchical scene graph structure, which takes ontology concept into consideration and represents not only the relations between the entities but also the status of the specific entities in a hierarchical structure, that enables the conditional reasoning for automated hazards identification to meet both obligation and prohibition regulatory rules. The regulatory information representation approach is detailed in Chapter 3.

(2) *Image information representation* module for on-site image processing. A novel solution to automatically identify the scene information from on-site images by the combining of deep learning-based object detection and individual detection model, which is more effective and robust than

traditional object detection-based approach for multi-hazard identification (needs only $n$ outputs for n-class hazard identification task) and with viewpoint changes and different individual postures (thanks to the pre-trained human pose estimation model deployed for individual detection). Subsequently, a geometric relationship analysis-based approach is proposed, which gives interpretable explanations and outputs to perform the combination of object detection and individual detection model. Specifically, individual-object association is performed using minimum weighted matching in bipartite graphs, and individual-object relationship analysis is identified by analyzing the geometric relationships of the individual's keypoints and the detected objects, which is represented as semantic phrases to further construct scene graphs. Accordingly, a constructed scene graph contains all key information of the obtained on-site image. The image information representation approach is detailed in Chapter 4.

(3) *Automated reasoning* module performs occupational hazards identification, which takes the output scene graphs from the previous modules as inputs. Based on the on-site image scene graph, the relevant regulatory rules are first extracted from the regulatory hierarchical scene graph to construct a relevant rules scene graph, from which the prohibition and obligation regulatory rules subgraph are extracted. Furthermore, automated reasoning for occupational hazard identification is performed based on the graph isomorphism analysis. The automated reasoning approach is detailed in Chapter 5.

## 2.3 Summary

In this chapter, the information representation structure to construct the outputs from the regulatory information representation module and the image information representation module is introduced. Based on the information representation structure, the proposed regulatory-image interface model is illustrated (Figure 2.7)

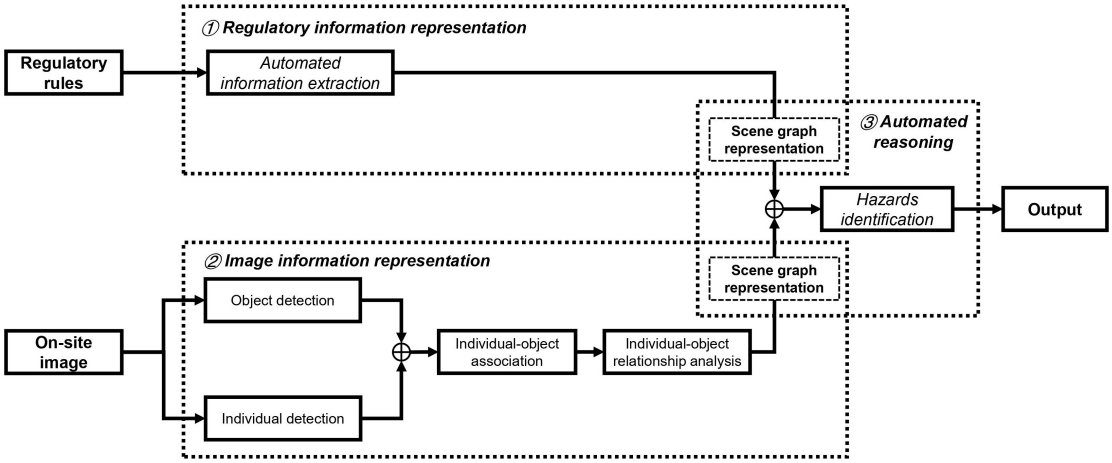



**Figure 2.7: An illustration of the proposed regulatory-image interface model.**

# Chapter 3

# Regulatory information representation

Existing regulatory rules in AEC domains are mostly documented with natural language sentences that need further processing for information understanding. However, given a large number of regulatory documents, the variability of their provisions in terms of formatting and semantics, and the large amount and complexity of the information they describe, the manual process of regulatory compliance checking is time-consuming, costly, and error-prone [40]. To address this gap, a semantic analysis approach for regulatory information extraction based on NLP and ontology is proposed, together with a novel hierarchical scene graph structure, to form a regulatory information representation module for automated regulatory rules processing. The pipeline of the proposed regulatory information representation module is illustrated in Figure. 3.1

## 3.1 Preprocessing

Regulatory rules are first preprocessed to prepare the raw text for further analysis. In the proposed approach, preprocessing consists of tokenization and morphological analysis.

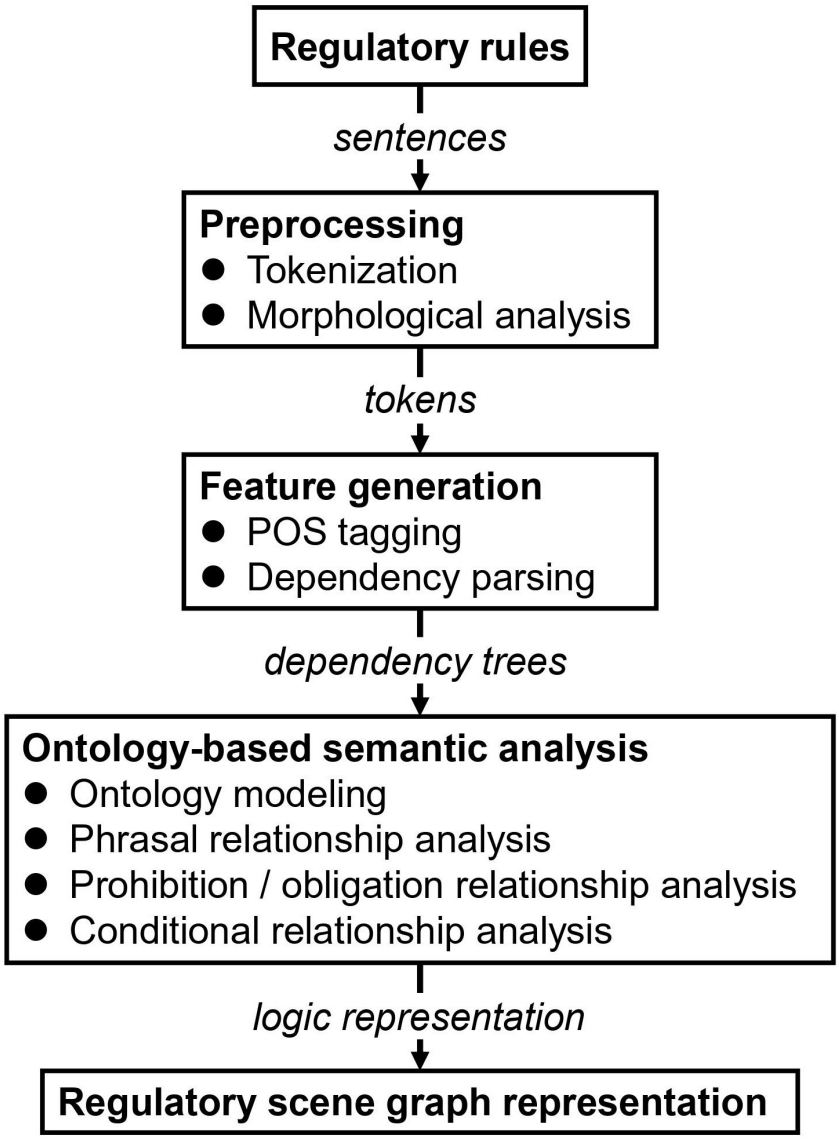



**Figure 3.1: Generic pipeline of the proposed regulatory information representation approach.**

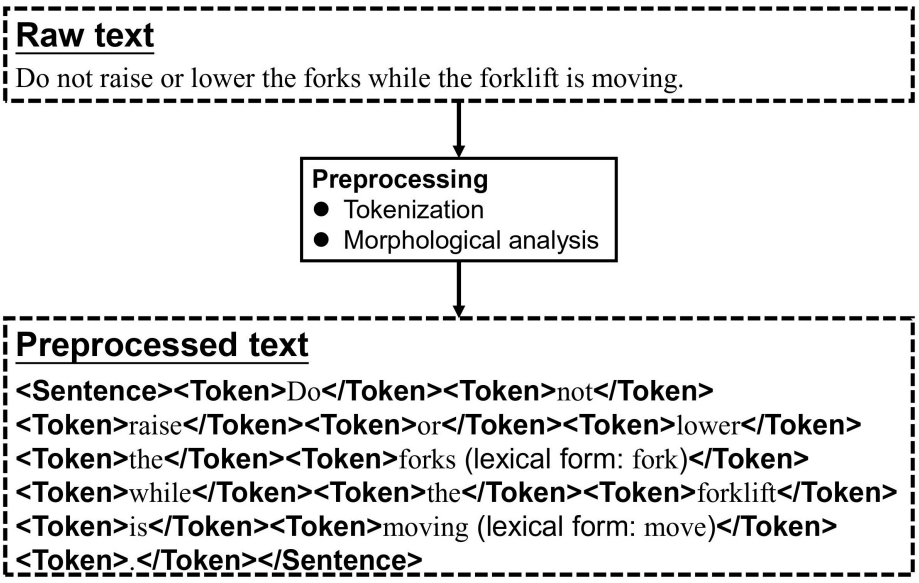



**Figure 3.2: An example of text preprocessing.**

### 3.1.1 Tokenization

Tokenization is the process of splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as sentences or words. Each of these smaller units is called tokens [41]. The regulatory rules are divided into tokens, where a token is a single word, a number, a punctuation mark, a whitespace characters, or a symbol. This process aims to prepare the text for further unit-based processing and conducted based on parsing the text according to common delimiters (i.e., white spaces and punctuations) with disambiguation consideration (e.g., "," as a delimiter in a number instead of punctuation). As shown in Figure 3.2, given the raw text from regulatory documents, boundaries of the token were recognized and labeled out using the "<token>" (i.e., starting of a token) or "</token>" (i.e., ending of a token) tags.

### 3.1.2 Morphological analysis

Basically, words are built up of minimal meaningful elements called morphemes and there are two types of morphemes:

(1) Stems: usually the lexical form of a word;

(2) Affixes: e.g., *-ed*, *-s*, *un-*, *-ly*.

Morphology is the study of composition and structure of words and their relationship to other words in the same language [42]. The morphological analysis aims to recognize the different forms of a word and to map them to the lexical form of that word in a dictionary [43]. In the proposed approach, morphological analysis is performed to map various nonstandard forms of a word (e.g., the plural form of a noun, the past tense of a verb) to its lexical form (e.g., the singular form of a noun, the infinitive form of a verb). As an example, in Figure 3.2, "forks" and "moving" were mapped to their lexical forms "fork" and "move", respectively.

## 3.2 Feature generation

The preprocessed tokens are further processed to generate syntactic features to describe the text. The proposed feature generation approach consists of Part-of-speech (POS) tagging and dependency parsing. Feature generation is performed using *spaCy* [44], which is an open-source library for advanced NLP in Python and supports over 50+ languages.

### 3.2.1 POS tagging

A POS is a category of words that have similar grammatical properties. POS tagging is the process of marking up a word in a text (corpus) as corresponding to a particular POS. Common POS tags and definitions are listed in Table 3.1. As an example in Figure 3.3, "raise" and "forks" were tagged as VERB and NOUN, respectively.

**Table 3.1: POS tags and descriptions**

| POS tag | Description |
|---|---|
| ADJ | adjective |
| ADP | adposition |
| ADV | adverb |
| AUX | auxiliary |
| CCONJ | coordinating conjunction |
| DET | determiner |
| INTJ | interjection |
| NOUN | noun |
| NUM | numberal |
| PART | particle |
| PRON | pronoun |
| PROPN | proper noun |
| PUNCT | punctuation |
| SCONJ | subordinating conjunction |
| SYM | symbol |
| VERB | verb |
| X | other |

### 3.2.2 Dependency parsing

Modern Dependency Grammar can be traced back primarily to the work of Lucien Tesnière [45]. Dependency syntax postulates that syntactic structure consists of relations between lexical items. In linguistics, the *head* of a phrase is the word that determines the syntactic category of that phrase and the other elements of the phrase modify the head (*head*'s *dependents*) [46], that is, a word depends on another either if it is a complement or a modifier of the latter. Robinson et al. [47] formulated four axioms to govern the well-formedness of dependency structures:

(1) In a sentence, one and only one word is independent, and the word is called the *root*;

(2) All others depend directly on some word;

(3) No word depends directly on more than one other, that is, if a word $A$ depends directly on another word $B$, it must not depend on a third word

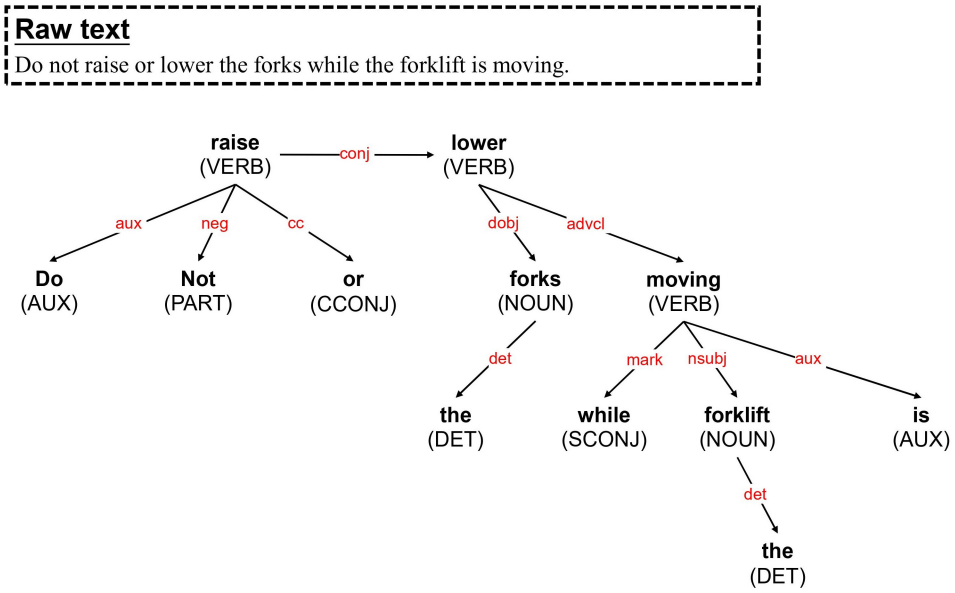



**Figure 3.3: An example of feature generation.**

$C$;

(4) If word $A$ depends directly on word $B$ and some word $C$ intervenes between them (in the linear order of the string), then $C$ depends directly on $A$ or $B$ or some other intervening word.

The fourth axiom is often called the requirement of *projectivity* and disallows crossing edges in dependency trees [48]. The output dependency tree of dependency parsing is illustrated in Figure 3.3, where each arrow connects a head to a dependent and is typed with the name of grammatical relations (dependency labels). For example, "lower" is the head of the phrase "lower forks", which is modified by its direct object (labeled as *dobj*) "forks". "raise" is the *root* of the sentence and not depend on others.

## 3.3 Ontology-based semantic analysis

In the proposed regulatory information representation approach, both syntactic (POS tags, dependency relations) and semantic features are generated. Semantic features are generated using an ontology-based semantic analysis method.

Ontology is a philosophical concept at the earliest. From the perspective of philosophy, ontology is a systematic explanation or explanation of objective existence, and it is concerned with the abstract nature of objective reality. In 1993, Gruber [49] gave one of the most popular definitions of ontology. That is, "An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general. And it is a different sense of the word than its use in philosophy." As a refinement of Gruber's definition Feilmayr and Wöß [50] stated: "An ontology is a formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity." More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.

Common components of ontologies include:

(1) *Instances*: the Instances in an ontology are the basic components that may include concrete objects such as people, animals, and vehicles, as well as abstract individuals such as numbers and words.

(2) *Classes*: sets, collections, concepts, classes in programming, types of objects, or kinds of things. As an example, "person" is the class of all people or the abstract object that can be described by the criteria for being a person.

(3) *Attributes*: aspects, properties, features, characteristics, or parameters that objects (and classes) can have.

(4) *Relations*: typically, a relation is of a particular class that specifies in what sense the object is related to the other object in the ontology. As an example of the relation *is-instance*, a lion (instance) is-instance of animals (class)

Nowadays, many studies have been performed for the application of ontology to the AEC industry for knowledge management in Building Information Modeling (BIM) and knowledge management [25–27, 51–54], which demonstrates the usefulness of ontology. Considering the scope of this work, a novel ontology-based model is proposed to perform semantic analysis for regulatory information extraction, which is equipped with the ability to parse the phrasal relationship and the requirement type described in the regulatory rules.

### 3.3.1 Ontology modeling

Commonly, the regulatory rule sentences are formed with the entities, together with their attributes and the relations between each entity. The three base classes in the proposed ontology model are *Entity*, *Status*, and *Relation* (Figure 3.4)

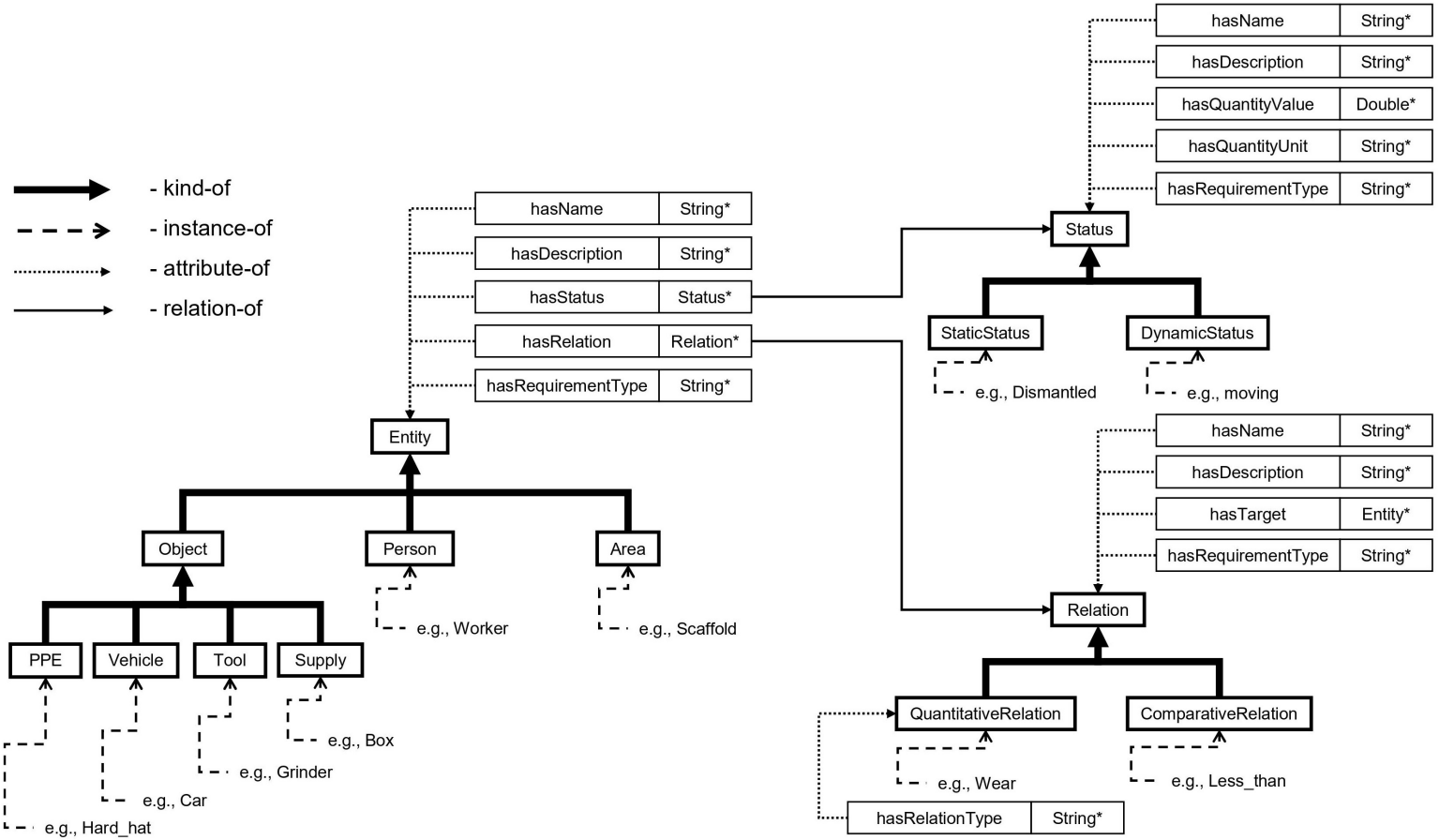



**Figure 3.4: The framework of the proposed ontology model.**

**Table 3.2: The attributes of the class *Entity*.**

| Name | Type | Description |
|---|---|---|
| hasName | String | The name of entity. |
| hasDescription | String | The description of entity. |
| hasStatus | List | The (dynamic or static) status of entity. |
| hasRelations | List | The relations connected to the entity. |
| hasRequirementType | String | The requirement type attached to entity in the rule sentence. |

### Entity

*Entity* is the base class to model all kinds of entities described in the regulatory rules. The attributes of *Entity* is shown in Table 3.2. Specifically, the attribute *hasStatus* of the class *Entity* is a list used to describe the status of the entity, which refers to an instance of the class *Status*:

$$\forall hasStatus.Status \sqsubseteq Status \tag{3.1}$$

Subsequently, the attribute *hasRelation* of the class *Entity* refers to a list of instances of the class *Relation* to create a relationship between the instances of *Entity*.

$$\forall hasRelations.Relation \sqsubseteq Relation \tag{3.2}$$

Another attribute of *Entity* is *hasRequirementType* which is used to specify the type of requirement: obligation (i.e., "must..."), prohibition (i.e., "must not..."), or condition (i.e., "if...then...").

Additionally, three subclasses of *Entity* are defined in ontology which inherit all attributes of the base class to further describe the entities in regulatory rule sentences:

(1) *Object*: on-site inanimate objects, which are further classified into four subclasses: *PPE* (e.g., *hard hat*), *Vehicle* (e.g., *car*), *Tool* (e.g., *grinder*),

and *Supply* (e.g., *box*).

$$object \sqsubseteq entity \tag{3.3}$$

$$PPE \sqsubseteq object \tag{3.4}$$

$$Vehicle \sqsubseteq object \tag{3.5}$$

$$Tool \sqsubseteq object \tag{3.6}$$

$$Supply \sqsubseteq object \tag{3.7}$$

(2) *Person*: e.g., *worker*.

$$Person \sqsubseteq entity \tag{3.8}$$

(3) *Area*: the place where a *Person* is working on or a *Object* is set to (e.g., *construction site*).

$$Area \sqsubseteq entity \tag{3.9}$$

**Status**

The class *Status* is used to describe the status of an instance of *Entity* and classified into *StaticStatus* (e.g., Dismantled) and *DynamicStatus* (e.g., mov-

**Table 3.3: The attributes of the class *Status*.**

| Name | Type | Description |
|---|---|---|
| hasName | String | The name of status. |
| hasDescription | String | The description of status. |
| hasQuantityValue | Double | The quantity value specified in the status. |
| hasQuantityUnit | String | The quantity unit specified in the status. |
| hasRequirementType | String | The requirement type attached to status in the rule sentence. |

ing).

$$StaticStatus \sqsubseteq Status \tag{3.10}$$

$$DynamicStatus \sqsubseteq Status \tag{3.11}$$

As shown in Table 3.3, the attributes *hasQuantityValue* and *hasQuantityUnit* are used to specify the detatiled quantity value in the a status (e.g., *5m*).

### Relation

The class *Relation* is used to define the relationship between two instances of *Entity* and classified into *QuantitativeRelation* (e.g., Wear) and *ComparativeRelation* (e.g., Less_than).

$$QuantitativeRelation \sqsubseteq Relation \tag{3.12}$$

$$ComparativeRelation \sqsubseteq Relation \tag{3.13}$$

**Table 3.4: The attributes of the class *Relation*.**

| Name | Type | Description |
|---|---|---|
| hasName | String | The name of relation. |
| hasDescription | String | The description of relation. |
| hasHeads | List | The heads connected to the relation. |
| hasTargets | List | The targets connected to the relation. |
| hasRequirementType | String | The requirement type attached to relation in the rule sentence. |

As shown in Table 3.4, the attribute *hasHeads* and *hasTargets* are used to create the relations between the certain instances of *Entity* to other instances of Entity.

$$\forall hasHeads.Head \sqsubseteq Entity \tag{3.14}$$

$$\forall hasTargets.Target \sqsubseteq Entity \tag{3.15}$$

Besides, for the class *QuantitativeRelation*, the attribute *hasRelationType* is used to specify the type of quantitative relation: modification (e.g., Move), possession (e.g., Wear), or loacting (e.g., On).

### Gazetteer lists

A gazetteer is a set of lists containing names of specific entities and relations. It groups any set of terms based on their commonality, based on which automated information extraction can be performed using gazetteer lists [55]. To enhance the information extraction ability of the proposed ontology model, a gazetteer is used to provide a set of term lists as follows, which provides a reference for the instantiation of the ontology:

(1) *PPE List*: hard hat, safety helmet, safety glasses, dust mask, full face mask, glove, welding glove, heavy-duty rubber glove, insulated glove, body harness, work shoes, boots, safety toed footwear;

(2) *Vehicle List*: car, forklift, crane, dump truck, bulldozer, front Loader, grader, trencher;

(3) *Tool List*: grinder, saw, circular saw, drill, hammer, crowbar, drill machine, polisher, vibrator, trowel;

(4) *Supply List*: ladder, power line, barrel, box, brick, block, guardrail, midrail, toeboard, load, fork;

(5) *Person List*: worker, construction worker, operator, observer, ordinary person;

(6) *Area List*: work area, height, heavy equipment, scaffold, construction site;

(7) *Dynamic Status List*: welding, cutting, grinding, nailing, concrete work, moving;

(8) *Static Status List*: erected, moved, dismantled, altered, height, width, volume;

(9) *Comparative Relation List*: less than, greater than, equal to, at least, at most;

(10) *Modification Relation List*: drive, move, use, equip, provide, support, raise, lower, handle, contact;

(11) *Possession Relation List*: has, wear;

(12) *Locating Relation List*: on, from, over, enter, around;

(13) *Quantity Unit List*: foot, m, cm, kg, g, ton;

As an example, for the regulatory rule "Wear a hard hat on a construction site", the items "hard hat" and "construction site" are included in the *PPE list* and *Area list*, respectively, and the items "wear" is included in the *Possession Relations list*.

### 3.3.2 Phrasal relationship parsing

Based on the modeling of ontology, the phrasal relationship analysis for the dependency tree, which is the output of the feature generation stage (section 3.2), is conducted. The pipeline of the proposed phrasal relationship parsing is introduced in Algorithm 1 and the Breadth-first Search (BFS) algorithm [56] is deployed to traverse all tokens in the dependency tree $T$ to parse their relationships. To drive the BFS traversal step, a queue, which is a data structure that allows "first in, first out" item insertion and removal, is used to keep track of the unvisited tokens. The traversal is processing until the token queue $Q$ is empty.

**Algorithm 1** BFS-based algorithm for phrasal relationship analysis.

**Input:**

The dependency tree, $T$;
The token queue, $Q$;

**Output:**

The topological set of the ontology model, $S$;

$Q.enqueue(T.root)$
**while** $Q$ is not empty **do**
    $h \Leftarrow Q.dequeue()$
    **for all** $d$ such that $d \in h.dependents$ **do**
        $Q.enqueue(d)$
        **if** $d.dep$ is $compound \vee d.dep$ is $amod$ **then**
            $h \Leftarrow merge(d, h)$
        **else if** $d.dep$ is $nsubj$ **then**
            $S.insert(QuantitativeRelation(h))$
            $Entity(d).hasRelations \Leftarrow QuantitativeRelation(h)$
            $S.insert(Entity(d))$
        **else if** $d.dep$ is $nsubjpass \vee d.dep$ is $dobj$ **then**
            $S.insert(Entity(d))$

18: $QuantitativeRelation(h).hasTargets(Entity(d))$
19: $S.insert(QuantitativeRelation(h))$
20: $Entity(person).hasRelations \Leftarrow QuantitativeRelation(h)$
21: $S.insert(Entity(person))$
22: **else if** $d.dep$ is $prep$ **then**
23: $S.insert(QuantitativeRelation(h))$
24: $S.insert(QuantitativeRelation(d))$
25: $QuantitativeRelation(d).hasRequirementType(Condition)$
26: **else if** $d.dep$ is $conj$ **then**
27: $C.insert(h, d)$
28: **else if** $d.dep$ is $neg$ **then**
29: $QuantitativeRelation(h).hasRequirementType(Prohibition)$
30: $S.insert(QuantitativeRelation(h))$
31: **else if** $d.dep$ is $advcl$ **then**
32: $Status(d).hasRequirementType(Condition)$
33: $S.insert(Status(d))$
34: **else**
35: $Skip$;
36: **end if**
37: **end for**
38: **end while**
**return** $S$;

At the start, the *root* token of $T$ is enqueued [1] into the token queue $Q$ (line 5 in Algorithm 1). $Q$ contains the route along which the algorithm is currently searching. For each step, a toekn is dequeued [2] from $Q$ (line 7 in Algorithm 1) while all its *dependents* are enqueued into $Q$ (line 9 in Algorithm 1). Subsequently, the syntactic relation (dependency label) connecting the *dependent* to its *head* is obtained and the phrasal relationship parsing is performed as the following patterns:

(1) If the dependency label of the *dependent* is *compound* (either a noun

[1] Adds an item to the end of a queue.
[2] Removes and returns the item at the beginning of a queue.

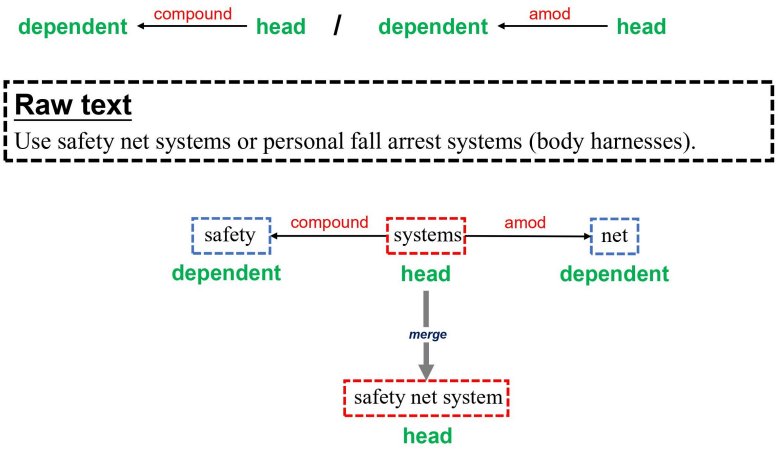



**Figure 3.5: An example of the phrasal relationship pattern (1).**

modifying the head of noun phrase, a number modifying the head of quantifier phrase, or a hyphenated word (or a preposition) modifying the head of the prepositioanl phrase) or *amod* (adjectival modifier: an adjective or an adjective phrase that modifies the meaning of another word), then the *dependent* needs to be merged together with the *head* to form a new token (line 11 in Algorithm 1). An example is illustrated in Figure 3.5, the compound *dependent* "safety" and the adjectival modifier *dependent* "net" are merged together with the *head* "system" to form a new token "safety net systems".

(2) If the dependency label of the *dependent* is *nsubj* (nominal subject: a non-clausal constituent in the subject position of an active verb), which usually indicates an action taken by an individual in regulatory rules, then the *head* and the *dependent* are instanced as *QuantitativeRelation* and *Person*, respectively. Besides, the attribute *hasRelations* of the instance of the *dependent* is set to the instance of the *head* (line 14 in Algorithm 1). An example is illustrated in Figure 3.6, the *head* "wear" is identified as an *QuantitativeRelation* and its nominal subject "Operators" is identified as an *Person*.

(3) If the dependency label of the *dependent* is *nsubjpass* (nominal passive

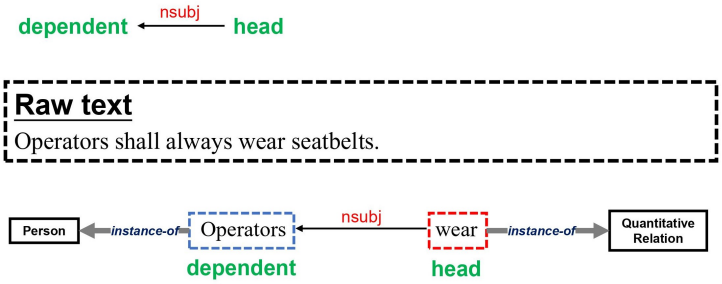



**Figure 3.6: An example of the phrasal relationship pattern (2).**

subject: a non-clausal constituent in the subject position of a passive verb), then the *head* and the *dependent* are instanced as *QuantitativeRelation* and *Entity*, respectively. Additionally, to provide more detailed description of the *dependent* token, further searching is performed in the Gazetteer lists to identify whether the *dependent* token is available in *PPE List*, *Vehicle List*, *Tool List*, *Supply List*, or *Area List* to instance *dependent* as a subclass of *Entity*. Besides, the attribute *hasRelations* of the instance of the *dependent* is set to the instance of the *head*. Subsequently, a passive object needs to be added into the phase by the instantiation of a *Person* in the ontology model (line 20 in Algorithm 1). An example is illustrated in Figure 3.7, the *head* "worn" is identified as an *QuantitativeRelation* and its nominal passive subject "Safety glasses" is identified as an *Entity* (further instanced as *PPE* by searching the Gazetteer lists).

(4) If the dependency label of the *dependent* is *dobj* (direct object: a noun phrase that is the accusative object of a (di)transitive verb), then the *head* and the *dependent* are instanced as *QuantitativeRelation* and *Entity*, respectively. Additionally, to provide more detailed description of the *dependent* token, further searching is performed in the Gazetteer lists to identify whether the *dependent* token is available in *PPE List*, *Vehicle List*, *Tool List*, *Supply List*, or *Area List* to instance *dependent* as a subclass of *Entity*. Subsequently, a subject needs to be added into the

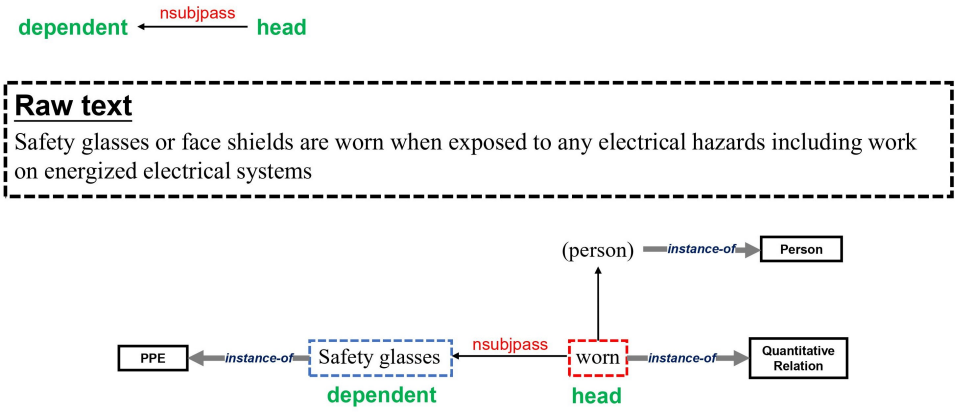



**Figure 3.7: An example of the phrasal relationship pattern (3).**

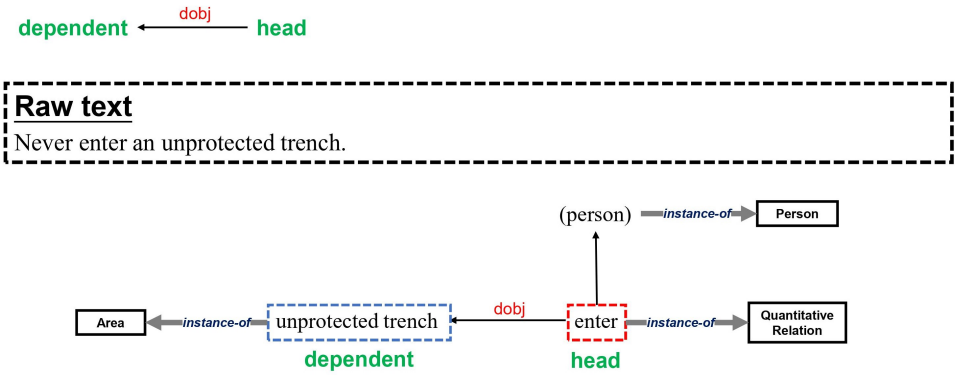



**Figure 3.8: An example of the phrasal relationship pattern (4).**

phase by the instantiation of a *Person* in the ontology model (line 20 in Algorithm 1). An example is illustrated in Figure 3.8, the *head* "enter" is identified as an *QuantitativeRelation* and its direct object "unprotected trench" is identified as an *Entity* (further instanced as *Area* by searching the Gazetteer lists).

(5) If the dependency label of the *dependent* is *prep* (prepositional modifier: any prepositional phrase that modifies the meaning of its head), then the classes of the *head* and the *dependent* in the ontology model are both identified as *QuantitativeRelation*. And *prep* in regulatory rules usually

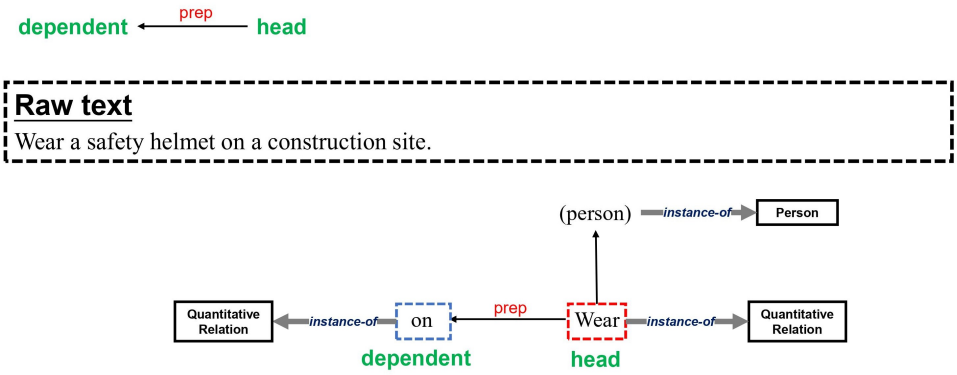



**Figure 3.9: An example of the phrasal relationship pattern (5).**

indicates a conditional clause is followed, thus *hasRequirementType* attribute of the instance of the *dependent* is set to "Condition" (line 25 in Algorithm 1). Subsequently, if the *head* had no nominal subject, a passive object needs to be added into the phase by the instantiation of a *Person* in the ontology model. An example is illustrated in Figure 3.9, the *head* "Wear" and its prepositional modifier "on" are identified as an *QuantitativeRelation*.

(6) If the dependency label of the *dependent* is *conj* (conjunct: a dependent of the leftmost conjunct in coordination), then the *head* and its conjunct are put into a container $C$ in which all instances share their attributes (line 27 in Algorithm 1). An example is illustrated in Figure 3.10, the *head* "raise" and its conjunct "lower" share their attributes.

(7) If the dependency label of the *dependent* is *neg* (negation modifier: an adverb that gives negative meaning to its head), then the *hasRequirementType* attribute of the instance of the *head* is set to "Prohibition" which represents a prohibition relationship in a regulatory rule (line 29 in Algorithm 1). An example is illustrated in Figure 3.11, the *head* "enter" has a negative modifier "Never", thus the *hasRequirementType* attribute of "enter" is set to "Prohibition".

(8) If the dependency label of the *dependent* is *advcl* (adverbial clause mod-

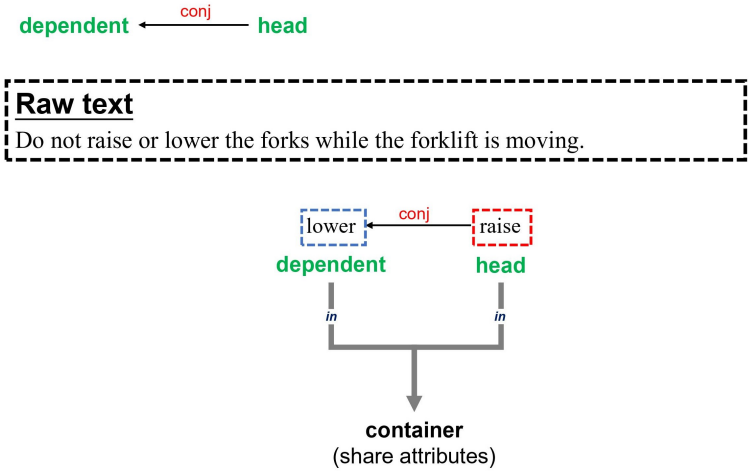



**Figure 3.10: An example of the phrasal relationship pattern (6).**

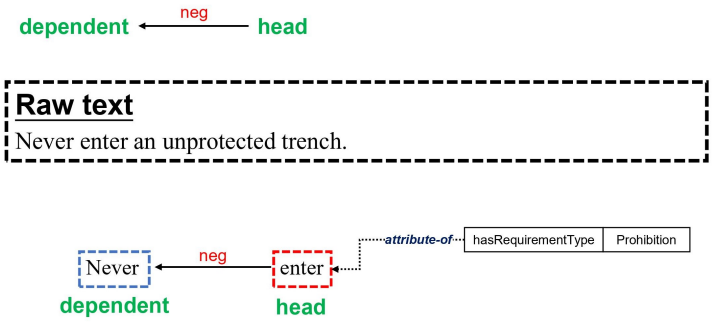



**Figure 3.11: An example of the phrasal relationship pattern (7).**

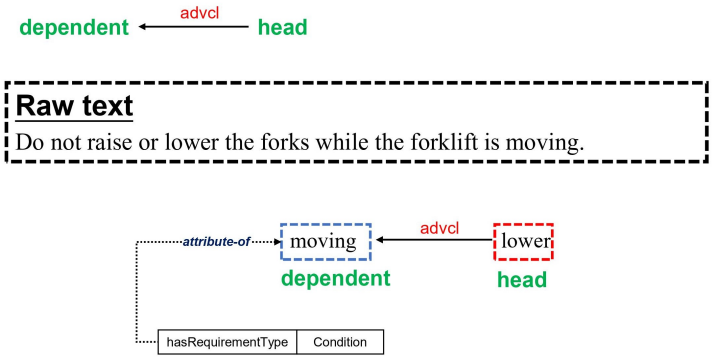



**Figure 3.12: An example of the phrasal relationship pattern (8).**

ifier: a clause that acts like an adverbial modifier), which in regulatory rules usually indicates a conditional clause is followed, then the *hasRequirementType* attribute of the instance of the *head* is set to "Condition" which represents a condition relationship (line 32 in Algorithm 1). An example is illustrated in Figure 3.12, the *head* "lower" has a negative modifier "moving", thus the *hasRequirementType* attribute of "moving" is set to "Condition".

(9) Other dependency labels are all ignored.

All of the instances are inserted in the topological set $S$ of the ontology model, which contains the semantic information of the regulatory rules and is used to represent the regulatory scene graph.

## 3.4 Information representation

### 3.4.1 Logic representation

The extracted regulatory information are first encoded in the logic representation with semantic phrases and logical connectives. The instances in the topological set of the ontology model are transformed to semantic phrases which are defined as a triplet (e.g., (element1, connection, element2)). Two types of triplets are defined to represent the relationship between two entities

**Table 3.5: Examples of phrases in the regulatory rules and the transformed triplets**

| Type | Phrases | Triplets |
|---|---|---|
| Entity relation | "operator wears seatbelt" | (operator, wear, seatbelt) |
| Entity status | "forklift is moving" | (forklift, be, moving) |

or the status of an entity as shown in Table 3.5. For the semantic phrases of a entity relation, "element1" or "element2" is a instance of *Entity* in the ontology model, which can be a instance of *Person* (e.g., worker), a instance of *Object* (e.g., glove), or a instance of *Area* (e.g., construction site). "connection" is a instance of *QuantitativeRelation* which semantically connects entities with limitations. On the other hand, for the semantic phrases of a entity status, "element1" is a instance of *Entity* and "element2" is a instance of *Status* (e.g., moving) in the ontology model. "connection" in this case is set as "be".

Furthermore, four types of logical connectives are used to represent the requirement types in the regulatory rules:

(1) *If...then...* ($\rightarrow$) is used to indicate a conditional relation assigning left conditional requirement triplets (one or more) to the instructed triplets (one or more) in its right;

(2) *Negation* ($\neg$) is used to indicate a prohibition requirement type assigning to a triplet in its right;

(3) *Logical* conjunction ($\wedge$) is used to indicate a conjunction relation between the triplets;

(4) *Logical* disjunction ($\vee$) is used to indicate a disjunction relation between the triplets.

As an example, Figure 3.13(b) demonstrates the logic representation of the regulatory rules in Figure 3.13(a).

### 3.4.2 Scene graph representation

Based on the logic representation (Figure 3.13(b)), the scene graph is generated for regulatory information representation. However, even though the notion of Scene graph representation has captured researchers' attention for on-site image understanding in AEC industry (Figure 2.6, introduced by the work of Xiong et al. [24]), regulatory information representation using scene graph structure is not yet available because the pairwise relationships in traditional scene graph cannot represent complex relationships in regulatory information:

(1) Conditional relationship, e.g., use body harness when working on height;

(2) Prohibition relationship, e.g., never operate a grinder near face;

To this end, an original hierarchical scene graph structure is proposed in this work to represent the information extracted from regulatory rules. Let $\hat{G}(\hat{V}, \hat{E}, \hat{C})$ be the hierarchical scene graph of the regulatory rules, where $\hat{V}$ is the set of vertices to represent the elements in the semantic phrase triplets and $\hat{E} = \{\{\mu, \nu, s, r, t\} : (\mu, \nu) \in \hat{V}^2, \mu \neq \nu\}$ is the set of edges to represent the relations in the semantic phrase triplets ($s$ and $r$ are the connections to represent a entity relation and a entity status, respectively. $t$ is the edge property to indicate the type of the requirements: obligation rule or prohibition rule). $\hat{C} = \{k \rightharpoonup I : k \in \hat{E}, I \subseteq \hat{E}\}$ is the set of conditional relations assigning instructed triplets to conditional triplets, which are stored in a hash map. The structure of a hierarchical scene graph is visualized in Figure 3.13(c):

(1) Dots on the inner layer represent elements of the conditional triplets;

(2) Dots on the intermediate layer represent elements of the instructed prohibition triplets;

(3) Dots on the outer layer represent elements of the instructed obligation triplets;

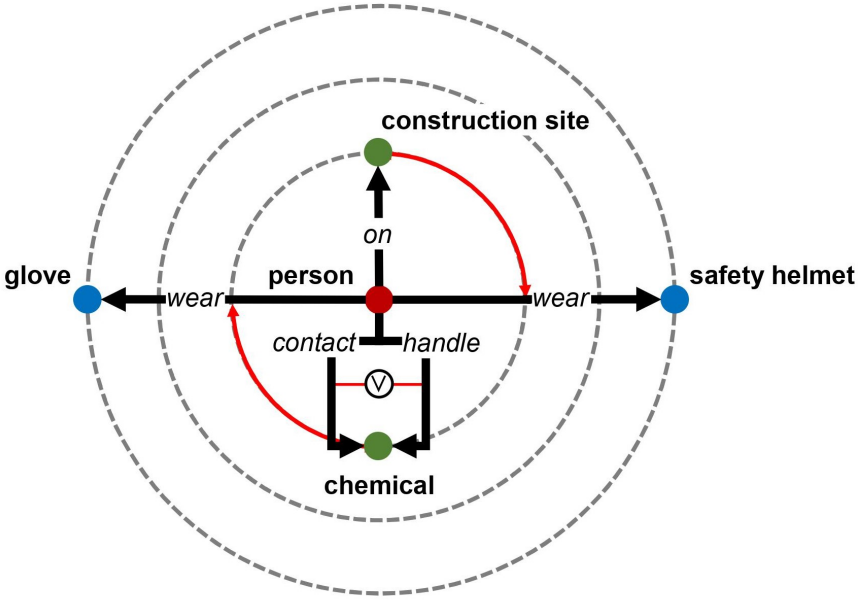
(4) The red lines in the circle denote the conditional relations between triplets.

(1): "Wear a safety helmet on a construction site.“
(2): "Wear gloves when handling or contacting chemicals."

(a)

((person,on,construction site))→()∨((person,wear,safety helmet))
((person,handle,chemical) ∨ (person,contact,chemical))→()∨((person,wear,glove))

(b)




(c)

**Figure 3.13: An example of the regulatory rules and its hierarchical scene graph. (a) The regulatory rules are decomposed and transformed to (b) the logic representation; (c) Hierarchical scene graph.**

In this example, the regulatory hierarchical scene graph is consists of the vertices $\hat{V} = \{person, construction\ \ site, chemical, safety\ \ helmet, glove\}$, the edges $\hat{E} = \{(person, construction\ site, on), (person, chemical, contact), (person, chemical, handle), (person, safety\ helmet, wear, obligation), (person, glove, wear, obligation)\}$, and the conditional relations set $\hat{C} = \{((person, construction\ site, on)) \rightharpoonup ((person, safety\ helmet, wear, obligation)), ((person, chemical, contact), (person, chemical, handle)) \rightharpoonup ((person, glove, wear, obligation))\}$

Consequently, regulatory rules documented with natural language sentences are transformed into a calculable and structured hierarchical scene graph with the ability to represent complex requirements in regulatory information.

## 3.5 System development

Based the proposed regulatory information representation approach introduced above, a novel automated regulatory rules processing system has been developed to implement the regulatory information representation module. The open-source libraries *spaCy* [44], *networkx* [57], *graphviz* [58] are deployed for feature generation, graph structure generation, visualization, respectively. Taking sentences from regulatory documents as inputs, the regulatory rules processing system creates the dependency trees and performs semantic analysis by ontology modeling. The extracted information in the ontology model is encoded in the logic representation with semantic phrases and logical connectives and finally generates a hierarchical scene graph for regulatory information representation. The experiments to validate the performance of the developed system is introduced in Chapter 6.

## 3.6 Summary

In this chapter, the proposed regulatory information representation module for automatically extracting information from regulatory documents and rep-

resented in the hierarchical scene graph is detailed, which is composed of the following stages:

(1) Preprocessing: The raw texts of regulatory rules are preprocessed by performing tokenization and morphological analysis to create the tokens of the sentences;

(2) Feature generation: The preprocessed tokens are further performed for POS tagging and dependency parsing to create the dependency trees;

(3) Ontology modeling: An ontology model is proposed to analyze the entities, attributes, and relations in the regulatory rules, which consists of three base classes: *Entity*, *Status*, and *Relation*. Phrasal relationship analysis is performed to extract key information from the dependency trees and model the ontology;

(4) Scene graph representation: Based on the transformed logic representation from the ontology model, a hierarchical scene graph is created to represent the regulatory information.

Based on the proposed approach, a novel automated regulatory rules processing system has been developed.

# Chapter 4

# Image information representation

In response to the limitations of the conventional object detection-based approaches in multi-hazard identification and with viewpoint changes and different individual postures of on-site workers, a novel solution to automatically identify the scene information from on-site images is performed, which deploys geometric relationships analysis to perform the combination of object detection and individual detection model to construct scene graphs. The pipeline line of the proposed image information representation module is illustrated in Figure. 4.1.

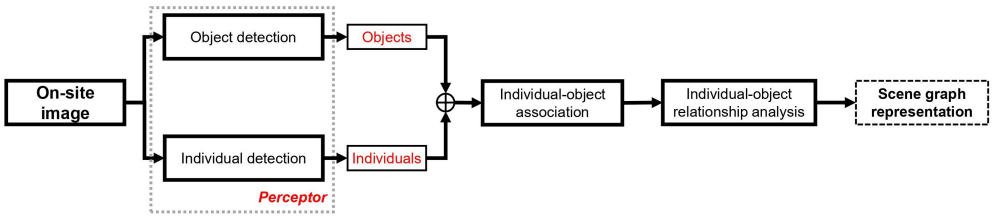



**Figure 4.1: Generic pipeline of the proposed image information representation approach.**

## 4.1 Image perception

A *Perceptor* is developed to process the images obtained by the on-site surveillance cameras. For each observed image, objects are recognized and localized by training an object detection model. Meanwhile, individual(s) are detected, together with their keypoints coordinates, using a pre-trained multi-person pose estimation model.

### 4.1.1 Object detection

Deep learning-based object detection approaches have been applied to the visual tasks in the AEC industry for on-site image understanding [21–24]. Based on the different architecture of Convolutional Neural Networks (CNNs) and detection strategies, object detection approaches are broadly divided into one-stage approaches and two-stage approaches. Two-stage approaches (e.g., R-CNN family [59–61]) are all region-based, which make predictions in two stages: First, the model proposes a set of regions of interests by select search or regional proposal network, which are sparse as the potential bounding box candidates can be infinite. Subsequently, a classifier only processes the region candidates. On the other hand, as one-stage approaches, YOLO [62], and its variants [63, 64], skips the region proposal stage and runs detection directly over a dense sampling of possible locations, which makes it extremely fast in the inference phase.

In 2018, Redmon et al. proposed YOLOv3 [64] as an improvement of the YOLO family. As demonstrated in Figure 4.2, YOLOv3 uses a deep network architecture with residual blocks and a total of 53 convolutional layers, i.e., Darknet-53, for feature extraction, which has better performance and is 1.5× faster than ResNet-101 [65]. Drawing on the idea of feature pyramid networks [66] (small size feature maps are used to detect large size objects, while large size feature maps detect small size objects), YOLOv3 makes predictions at three scales, which are precisely given by downsampling the dimensions of the input image by 32, 16 and 8 respectively. This means, with an input of $416 \times 416$, the YOLOv3 model contains three output layers, each dividing the input image into $13 \times 13$ grids, $26 \times 26$ grids, and $52 \times 52$ grids,

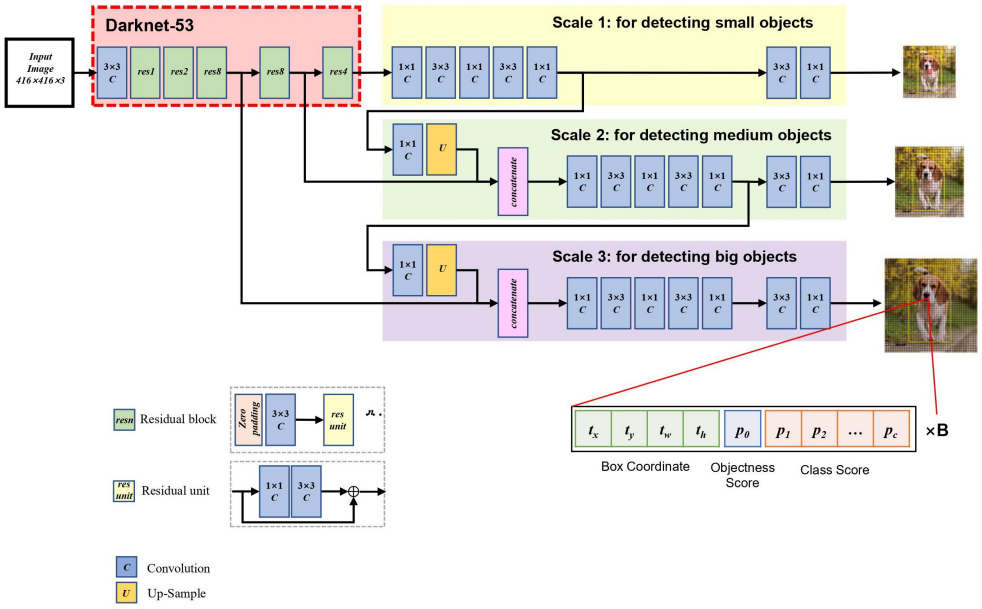



**Figure 4.2: An illustration of the architecture of the YOLOv3 model.**

respectively. This method allows YOLOv3 to get more meaningful semantic information from the upsampled features and finer-grained information from the earlier feature map. The prediction of the width and the height of the bounding box may lead to unstable gradients during training. Thus YOLOv3 deploys pre-defined default bounding boxes called anchors to predict log-space transforms. Like YOLOv2, the anchor boxes of YOLOv3 are also obtained by clustering. As demonstrated in Figure 4.3 [64], YOLOv3 predicts four coordinate values $(t_x, t_y, t_w, t_h)$ for each bounding box. For the predicted cell according to the offset of the upper left corner of the image $(c_x, c_y)$, the width and height of the bounding box $p_w$, $p_h$ can be used to predict the bounding box as follows:

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{4.1}$$

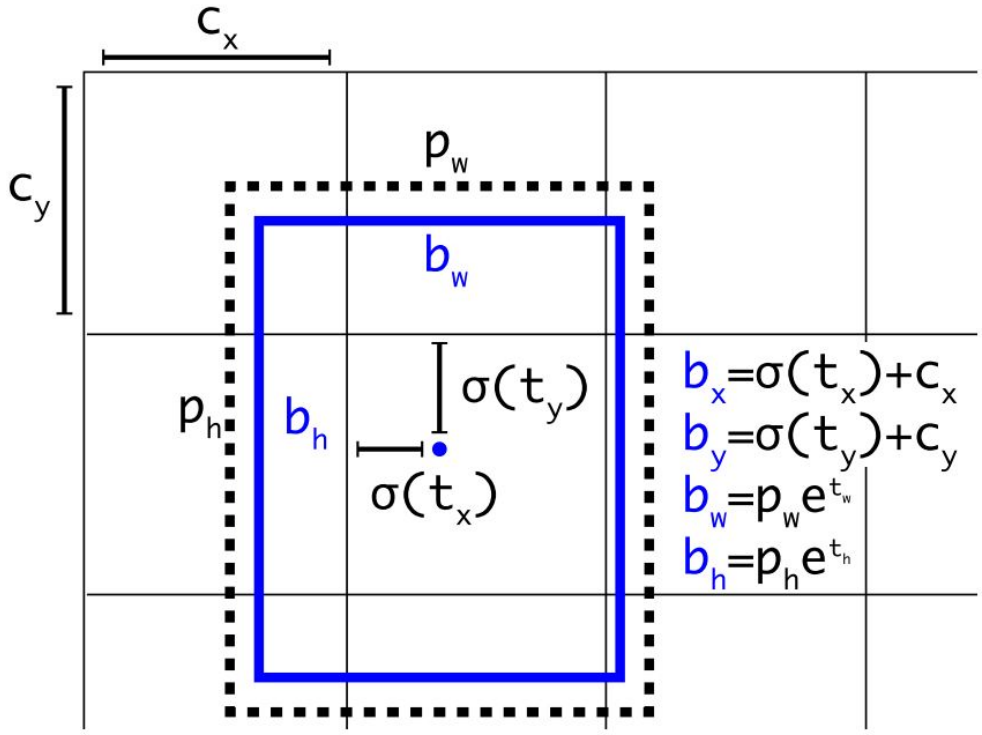



**Figure 4.3: Bounding boxes with dimension priors and location prediction.**

Each of the three output layers is associated with three anchor boxes, resulting in a total of nine anchor boxes. When training the YOLOv3 model, each grid cell in the output layers takes corresponding anchor boxes and learns how to shift and/or scale these anchor boxes so that the bounding boxes perfectly fit the objects of interest.

The prediction result of the network is a 3-d tensor that encodes bounding box, objectness score and prediction over classes:

$$N \times N \times (3 * (4 + 1 + C)) \tag{4.2}$$

where $N \times N$ is the number of the grid cells of the model and $C$ is the number of the classes to train the network on.

Besides, YOLOv3 predicts an confidence score for each bounding box using logistic regression, while YOLO and YOLOv2 uses the sum of squared er-

rors for classification terms. The Softmax classifier deployed in the YOLOv2 assumes that a target belongs to only one class, and each box is assigned to the class with the largest score. However, in some complex scenarios, a target may belong to multiple classes (with overlapping class labels), thus YOLOv3 employs multiple independent logistic classifier (using Sigmoid function) for each class rather than one softmax layer when predicting class confidence. In the training phase, binary cross-entropy loss is used as the loss function to train class prediction (both YOLO and YOLOv2 use a loss function based on the sum of squares):

$$\begin{aligned}
&\lambda_{coord} \sum_{i=0}^{N^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} [(b_x - \hat{b_x})^2 + (b_y - \hat{b_y})^2 + (b_w - \hat{b_w})^2 + (b_h - \hat{b_h})^2] \\
&+ \sum_{i=0}^{N^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} [-log(p_c) + \sum_{i=1}^{n} BCE(\hat{C}_i, C_i)] \\
&+ \lambda_{noobj} \sum_{i=0}^{N^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{noobj} [-log(1 - p_c)]
\end{aligned} \tag{4.3}$$

where $N^2$ is the number of the grid cells ($13 \times 13$ grids, $26 \times 26$ grids, or $52 \times 52$ grids), $B$ is the bounding boxes. $\mathbb{1}_i^{obj}$ denotes if object appears in cell $i$ and $\mathbb{1}_{i,j}^{obj}$ denotes that the $j_{th}$ bounding box predictor in cell $i$ is "responsible" for that prediction. In contrast, $\mathbb{1}_{i,j}^{noobj}$ denotes that the $j_{th}$ bounding box predictor in cell $i$ that don't contain objects. The parameters $\lambda_{coord}$ and $\lambda_{noobj}$ is employed to increase the loss from bounding box coordinate predictions and decrease the loss from confidence predictions for boxes that don't contain objects. Additionally, binary cross-entropy loss $BCE(\hat{C}_i, C_i)$ is given by:

$$BCE(\hat{C}_i, C_i) = -\hat{C}_i log(C_i) - (1 - \hat{C}_i) log(1 - C_i) \tag{4.4}$$

where $C_i$ is the objectness in cell $i$, i.e. confidence score of whether there is an object or not.

In summary, predictions of YOLOv3 are carried out from one single network, which can be trained end-to-end to improve accuracy. High efficiency and speed make YOLOv3 a reasonable option for real-time processing for

industrial purposes.

### 4.1.2 Individual detection

In contrast to the common on-site information representation and hazards identification approaches [21–23], an individual detection model is employed in this work to specify individual features. This strategy makes the image information representation model of this work be more robust with viewpoint changes and different individual postures.

The workers' postures are characterized by extracting the joint positions of the person in the image using OpenPose [67], which is the state-of-the-art model for detecting human body parts in images. Instead of the conventional top-down approaches which employ a person detector and performing single-person pose estimation for each detected person, OpenPose provides a bottom-up approach by detecting all body parts in the image and associating them with different person.

As demonstrated in Figure 4.4, OpenPose takes the image of size $w \times h$ as input and processes images through a two-branch multi-stage CNNs, where each stage in the first branch predicts confidence maps, and each stage in the second branch predicts Part Affinity Fields (PAFs). A set of feature map $F$ is generated by analysing the raw image using a CNN. Then, the network is divided into multiple similar stages. In each stage, there are two branches, one for confidence maps, which obtains the joint position candidates, and the other one predicts the PAF, which is a set of 2D vector fields that encode the location and orientation of limbs over the image domain to correlate the relationships between joint points to splicing into the full-body postures of an unknown number of people. The first stage takes the feature map $F$ as input and generates a set of confidence maps $S^1$ and a set of PAFs $L^1$. In the rest of the stage $t(t > 1)$, the output of the previous stage, $S^{t-1}$ and $L^{t-1}$, and the feature map $F$ will be used as input of the current stage. The loss functions of both branches at stage $t$ are

$$\begin{cases} f_S^t = \sum_{j=1}^{J} \sum_p W(p) \cdot \left\| S_j^t(p) - S_j^*(p) \right\|_2^2, \\ f_L^t = \sum_{c=1}^{C} \sum_p W(p) \cdot \left\| L_c^t(p) - L_c^*(p) \right\|_2^2. \end{cases} \tag{4.5}$$

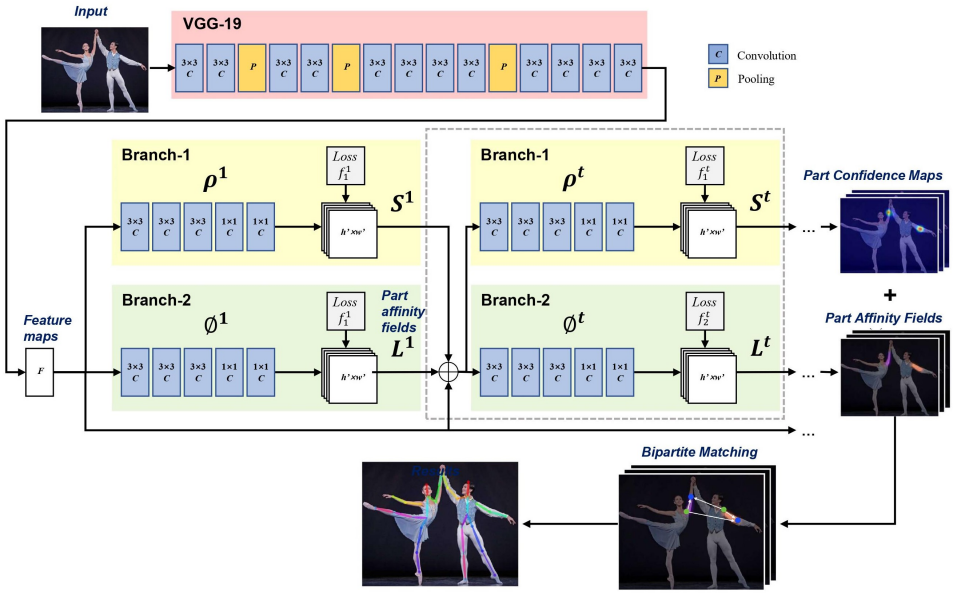



**Figure 4.4: An illustration of the architecture of the OpenPose model.**

where $S_j^*$ is the groundtruth part confidence map of the jth joint point, $L_j^*$ is the groundtruth part affinity vector field of the jth joint point, and $W$ is a binary mask. When the annotation is missing at an image location $p$, $W(p) = 0$. The complete loss function is

$$f = \sum_{t=1}^{T} \left( f_S^t + f_L^t \right) \tag{4.6}$$

Finally, the confidence maps and the affinity fields are parsed by greedy inference to output the 2D keypoints for all people in the image. OpenPose provides the positions of 18 body joints (pre-trained using COCO 2016 keypoints challenge dataset [68], see Figure 4.5).

The choice of OpenPose is motivated for its functionality on RGB images or videos taken by on-site surveillance cameras. This provides a huge benefit in comparison to the skeletal tracking capability of RGB-D devices (e.g., Microsoft Kinect [69]) which depend on depth information. Besides, in contrast

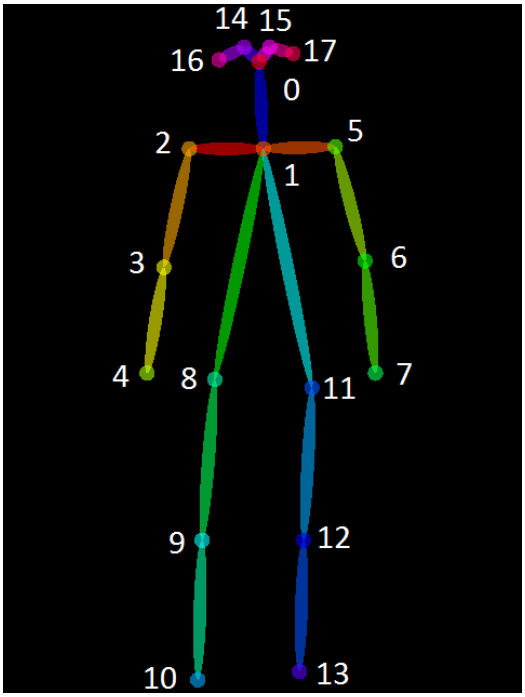



**Figure 4.5: Output format of OpenPose.**

to the top-down approaches, e.g., Mask R-CNN [70] and Alpha-Pose [71], the inference time of OpenPose is invariant to the number of people in the image and able to provide real-time performance.

To further reduce the inference time, a light-weight architecture Mobilenetv2 [72] is deployed as the feather extractor instead of VGG-19 [73] in the original paper. To achieve model acceleration, Mobilenetv2 employs a special convolutional filter called depthwise separable convolution as the replacement of the standard convolutional filter together with linear bottleneck ($1 \times 1$ convolutional layer without ReLU) to solve the problem of information loss due to nonlinear activation functions. First, a pointwise ($1 \times 1$) convolution is deployed to expand the low-dimensional input feature map to a higher-dimensional space suited to nonlinear activations. Next, a depthwise convolution is performed using $3 \times 3$ kernels to achieve spatial filtering of the higher-dimensional tensor. Finally, the spatially-filtered feature map is projected back to a low-dimensional subspace using another pointwise convolution.

## 4.2 Individual-object association

Prior to individual-object relationship analysis, the individual-object association is performed to associate detected objects to a detected individual near to. This stage aims to provide prior knowledge for individual-object relationship analysis and reduce computational complexity.

Let $H = \{H_1, H_2, ..., H_I\}$ be the set of the detected individual(s) via OpenPose, where $I$ is the number of detected individual(s) in the obtained image and $H_i = \{(x_i^{(0)}, y_i^{(0)}), (x_i^{(1)}, y_i^{(1)}), ..., (x_i^{(17)}, y_i^{(17)})\}$ represents the detected body parts of the $i_{th}$ individual (see Figure 4.5). The output of YOLOv3 are formulated as a set of object bounding boxes $B = \{B^{(1)}, B^{(2)}, ..., B^{(K)}\}$, where $K$ categories of objects is detected in the obtained on-site image. Each bounding box $B_j^{(k)} = (x_j^{(k)}, y_j^{(k)}, w_j^{(k)}, h_j^{(k)}, k),\ j \in \{1, 2, ..., J\}$ contains five elements, where $(x_j^{(k)}, y_j^{(k)})$ and $(w_j^{(k)}, h_j^{(k)}\ )$ are respectively the bounding boxes' position and size and $k \in \{1, 2, ..., K\}$ represents the class index of the object in the bounding box. For each category of detected object, an object $j^*$ is associated to an individual $i^*$ by searching the minimum Euclidean distance between bounding boxes $B$ and detected neck keypoints $H^{(1)} = \{(x_0^{(1)}, y_0^{(1)}), (x_1^{(1)}, y_1^{(1)}), ..., (x_I^{(1)}, y_I^{(1)})\}$ (body part 1 in Figure 4.5) in $H$. To this end, a weighted bipartite graph is constructed to represent the detected entities and perform individual-object association as a minimum weighted matching in bipartite graphs. Given a bipartite graph $G_e = (S \cup T, E_e)$ with weight function $w : E_e \to \mathbb{R}_+$, where $S = \{1, 2, \cdots, I\}$ and $T = \{1, 2, \cdots, J\}$ are the vertices to represent the detected individuals and objects (with category index $k$), respectively. The weight $w_e(s, t)$ of an edge $e_e = (s, t)$ is the Euclidean distance between bounding boxes $B_t^{(k)}$ and detected neck keypoints of $H_s$:

$$w_e(s,t) = \sqrt{(x_s^{(1)} - x_t^{(k)})^2 + (y_s^{(1)} - y_t^{(k)})^2} \tag{4.7}$$

A bi-adjacency matrix $A$ is associated with the graph $G_e = (S \cup T, E_e)$, where $A$ is a $J \times I$ matrix and $a_{ji} = w_e(i, j)$. The minimum weighted

matching is performed by searching the minimum value for each row:

$$M = \{\{i^*, j^*\} : i^*, j^* = \underset{i \in \{1,2,...,I\}}{\arg\min}\ a_{ji}, 1, 2, ..., J\} \tag{4.8}$$

Figure 4.6 illustrates an example of individual-object (hard hats) association: (a) Four individuals and three hard hats are detected in the on-site image. (b) The on-site image is converted to a bipartite graph $G_e = (S \cup T, E_e)$ with its bi-adjacency matrix $A$, where $S = \{S_1, S_2, S_3, S_4\}$ and $T = \{T_1, T_2, T_3\}$; (c) The minimum weighted matching is performed by searching the minimum value for each row: $M = \{\{S_1, T_1\}, \{S_2, T_2\}, \{S_3, T_3\}\}$ (red lines).

## 4.3 Individual-object relationship analysis

Based on the associated individual-object pairs, individual-object relationship analysis is performed. At present, the proposed individual-object relationship analysis approach in this work is able to process four types of object:

(1) Head protection PPE: hard hats, safety glasses, dust masks, and full-face masks;

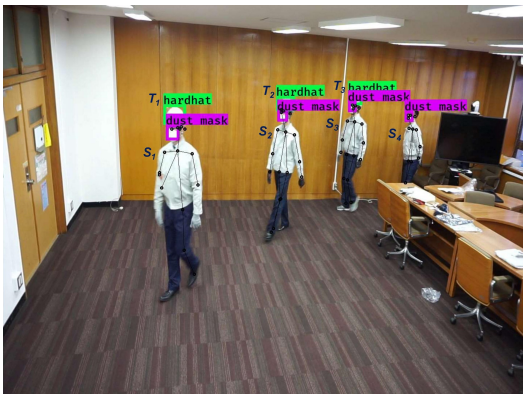
(2) Grinder: two regulatory rules have been addressed ("Always use two hands when operating a grinder", and "Never operate a grinder near face")

(3) Glove

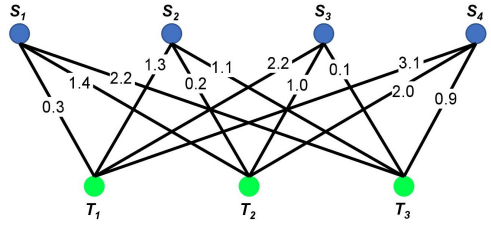
(4) Body harnesses

### 4.3.1 Head protection PPE

For each associated individual and head protection PPE $\{i^*, j^*\} \in M$ their relationship is identified by analyzing key lengths. The Euclidean distance among detected neck keypoint (body parts 1 in Figure 4.5) and hip keypoints

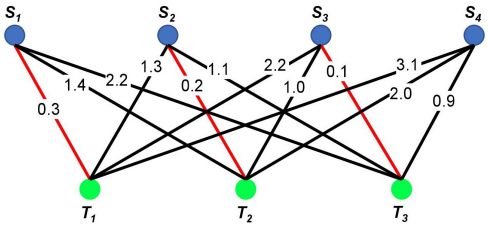



(a)




$$A = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{bmatrix} 0.3 & 1.3 & 2.2 & 3.1 \\ 1.4 & 0.2 & 1.0 & 2.0 \\ 2.2 & 1.1 & 0.1 & 0.9 \end{bmatrix} \end{array}$$

(b)




$$A = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{bmatrix} 0.3 & 1.3 & 2.2 & 3.1 \\ 1.4 & 0.2 & 1.0 & 2.0 \\ 2.2 & 1.1 & 0.1 & 0.9 \end{bmatrix} \end{array}$$

(c)

**Figure 4.6: An example of the individual-object association.**

(body parts 8 and 11 in Figure 4.5) of $i^*$ is considered as a dynamic reference threshold, which will keep changing synchronously when the distance between the individual and the camera changes:

$$\beta_{i^* \leftrightarrow j^*} = \max(\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(8)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(8)})^2}, \sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(11)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(11)})^2}) \cdot \gamma \tag{4.9}$$

where $\gamma$ is the scaling coefficient to strike the relationship analysis for different head protection PPE. For hard hats, safety glasses, dust masks, and full-face masks, $\gamma$ is set to 0.8, 0.7, 0.6, 0.6, respectively.

If the Euclidean distance between the position $(x_{j^*}, y_{j^*})$ of the bounding box of $j^*$ and detected neck keypoint (body parts 1 in Figure 4.5) of $i^*$ is smaller than the reference threshold $\beta_{i^* \leftrightarrow j^*}$, then the relationship between the $i^*$ and $j^*$ is created; otherwise, even though $j^*$ is associated with $i^*$, no relationship is created between them:

$$c_{i^* \leftrightarrow j^*} = \begin{cases} \text{“}wear\text{”} & , if\ d_h(i^*, j^*) < \beta_{i^* \leftrightarrow j^*} \\ N/A & , otherwise \end{cases} \tag{4.10}$$

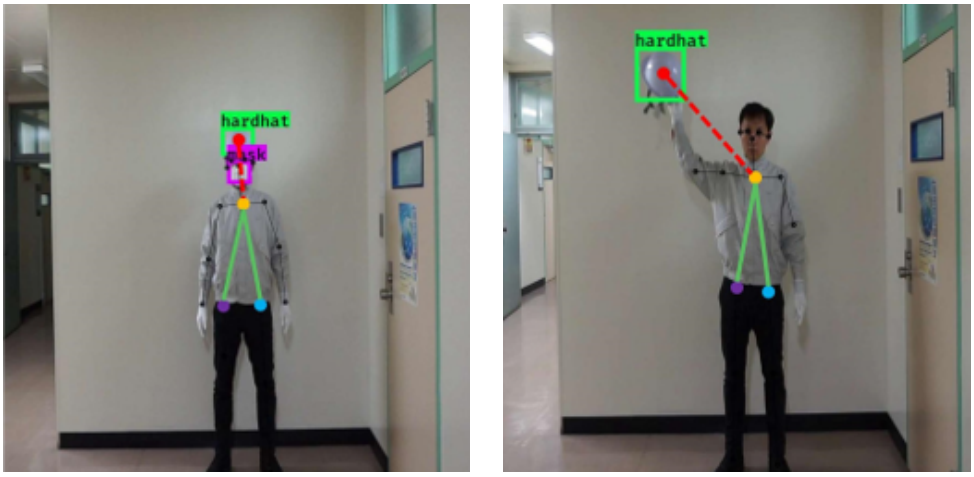
where

$$d_h(i^*, j^*) = \sqrt{\left(x_{i^*}^{(1)} - x_{j^*}\right)^2 + \left(y_{i^*}^{(1)} - y_{j^*}\right)^2} \tag{4.11}$$

and $c_{i^* \leftrightarrow j^*}$ indicates the connection to create the relationship between $i^*$ and $j^*$ as a semantic phrase $(i^*, c_{i^* \leftrightarrow j^*}, j^*)$ (e.g., $(person, wear, hard\ hat)$ in Figure 4.7(a)) or not (Figure 4.7(b)).

### 4.3.2 Grinder

Currently in this work, two regulatory rules related to grinder proper use are addressed to create a individual-grinder relationship:




(a) The relationship is created: $(person, wear, hard\ hat)$.

(b) No relationship is created.

**Figure 4.7: The individual-head protection PPE relationship identification strategies.**

**"Always use two hands when operating a grinder"**

Let $B_{g^*} = (x_{g^*}, y_{g^*}, w_{g^*}, h_{g^*})$ be the detected bounding box of a grinder $g^*$ which is associated with the detected individual $i^*$. Firstly, the Euclidean distance from the left wrist keypoint and right wrist keypoint (body parts 7 and 4 in Figure 4.5) of $i^*$ to the position $(x_{g^*}, y_{g^*})$ of $g^*$ is calculated (Figure 4.8):

$$\begin{aligned} d_l(i^*, g^*) &= \sqrt{\left(x_{i^*}^{(7)} - x_{g^*}\right)^2 + \left(y_{i^*}^{(7)} - y_{g^*}\right)^2} \\ d_r(i^*, g^*) &= \sqrt{\left(x_{i^*}^{(4)} - x_{g^*}\right)^2 + \left(y_{i^*}^{(4)} - y_{g^*}\right)^2} \end{aligned} \tag{4.12}$$

If the grinder is close enough to the wrists, then the individual is identified as holding the grinder:

$$h_{i^* \leftrightarrow g^*} = \begin{cases} 1 & , if\ d_l(i^*, g^*) < \beta_{i^* \leftrightarrow g^*}\ or\ d_r(i^*, g^*) < \beta_{i^* \leftrightarrow g^*} \\ 0 & , otherwise \end{cases} \tag{4.13}$$

where $\beta_{i^* \leftrightarrow g^*}$ is the reference threshold calculated based on the size of the


**Figure 4.8: Relationship identification strategies to address the rule "Always use two hands when operating a grinder".**

bounding box of $g^*$:

$$\beta_{i^*\leftrightarrow g^*} = \max(w_{g^*}, h_{g^*}) \tag{4.14}$$

If $h_{i^*\leftrightarrow g^*} = 1$, then relationship identification needs to be further performed to identify whether the individual $i^*$ is holding the grinder $g^*$ using single hand or two hands. It's known that when an object is holding by two hands the distance between the wrists is small. Thus, the relationship between $i^*$ and $g^*$ is identified as follows:

$$c_{i^*\leftrightarrow h^*_{i^*}}, h^*_{i^*}, c_{h^*_{i^*}\leftrightarrow g^*} = \begin{cases} \text{“}use\text{”}, \text{“}two\ hands\text{”}, \text{“}operate\text{”} & , if\ d_{l\leftrightarrow r}(i^*) < \beta_{j^*\leftrightarrow g^*} \\ \text{“}use\text{”}, \text{“}single\ hand\text{”}, \text{“}operate\text{”} & , otherwise \end{cases} \tag{4.15}$$

where $h^*_{i^*}$ indicates the hands status (e.g., two hands, single hand) when operating a grinder while $c_{i^*\leftrightarrow h^*_{i^*}}$ creates the relationship between $i^*$ and $h^*_{i^*}$, $h^*_{i^*}$ and $g^*$, as semantic phrases $(i^*, c_{i^*\leftrightarrow h^*_{i^*}}, h^*_{i^*})$, $(h^*_{i^*}, c_{h^*_{i^*}\leftrightarrow g^*}, g^*)$, respectively (e.g., the relationship $(person, use, two\ hands)$, $(two\ hands, operate, grinder)$ is created in Figure 4.8)

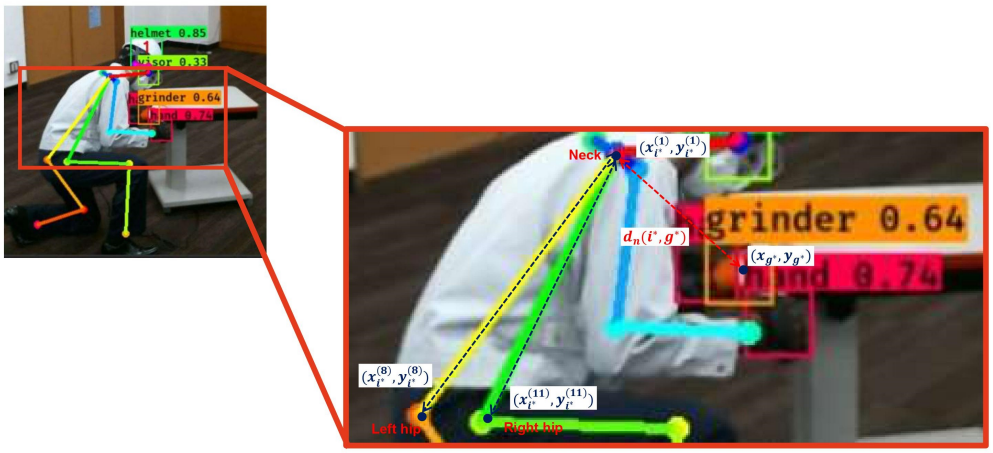



**Figure 4.9: Relationship identification strategies to address the rule "Never operate a grinder near face".**

**"Never operate a grinder near face"**

For the detected individual $i^*$ and the associated grinder $g^*$, the Euclidean distance from the neck keypoint (body part 1 in Figure 4.5) of $i^*$ to the position $(x_{g^*}, y_{g^*})$ of the bounding box of $g^*$ is calculated:

$$d_n(i^*, g^*) = \sqrt{\left(x_{i^*}^{(1)} - x_{g^*}\right)^2 + \left(y_{i^*}^{(1)} - y_{g^*}\right)^2} \tag{4.16}$$

If the grinder is close enough to the neck, then the relationship between the grinder and the face of individual is created:

$$c_{i^* \leftrightarrow g^*}^{face} = \begin{cases} \text{"}near\text{"} & , if\ d_n(i^*, g^*) < \beta_{i^* \leftrightarrow g^*} \\ N/A & , otherwise \end{cases} \tag{4.17}$$

where $\beta_{i^* \leftrightarrow g^*}$ is the reference threshold calculated based on the Euclidean distance among detected neck keypoint (body parts 1 in Figure 4.5) and hip keypoints (body parts 8 and 11 in Figure 4.5) of $i^*$ with $\gamma = 0.3$:

$$\begin{aligned} \beta_{i^* \leftrightarrow g^*}^{face} = \max(&\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(8)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(8)})^2}, \\ &\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(11)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(11)})^2}) \cdot \gamma \end{aligned} \tag{4.18}$$

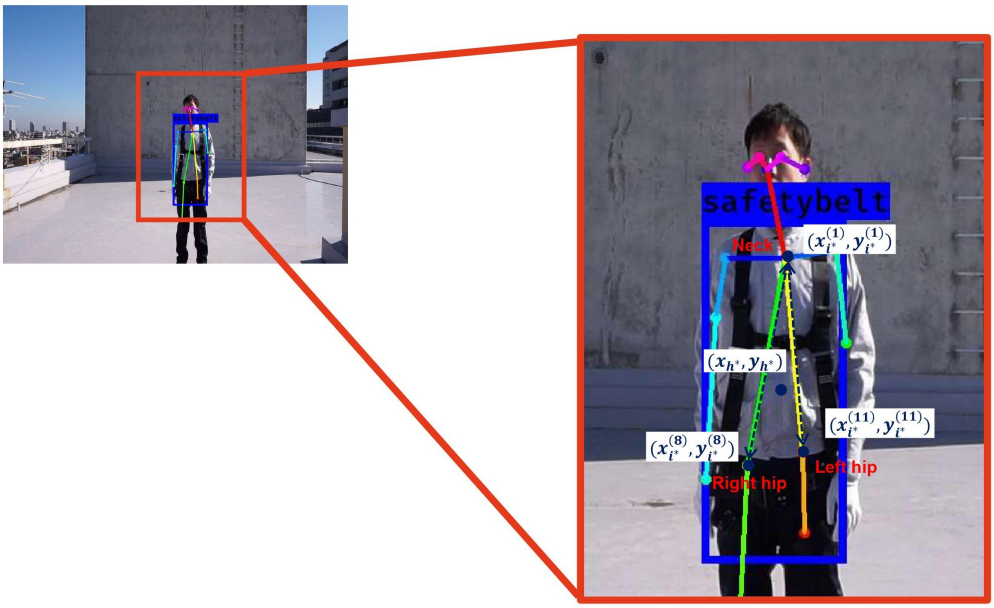



**Figure 4.10: The individual-body harnesses relationship identification strategies.**

and $c_{i^* \leftrightarrow g^*}^{face}$ indicates the connection to create the relationship between the face of $i^*$ and $g^*$ as a semantic phrase $(face, c_{i^* \leftrightarrow g^*}^{face}, g^*)$ (e.g., the relationship $(face, near, grinder)$ is created in Figure 4.9)

### 4.3.3 Body harnesses

Let $B_{h^*} = (x_{h^*}, y_{h^*}, w_{h^*}, h_{h^*})$ be the bounding box of a detected body harnesses $h^*$ which is associated with the detected individual $i^*$. Their relationship is identified by performing the geometric relationship analysis for the key point $(x_{i^* \leftrightarrow h^*}^k, y_{i^* \leftrightarrow h^*}^k)$, the geometric center of the neck keypoint (body part 1 in Figure 4.5) and hip keypoints (body parts 8 and 11 in Figure 4.5) of $i^*$, which is given as follows:

$$\begin{aligned} x_{i^* \leftrightarrow h^*}^k &= \frac{x_{i^*}^{(1)} + x_{i^*}^{(8)} + x_{i^*}^{(11)}}{3} \\ y_{i^* \leftrightarrow h^*}^k &= \frac{y_{i^*}^{(1)} + y_{i^*}^{(8)} + y_{i^*}^{(11)}}{3} \end{aligned} \tag{4.19}$$

If $(x^k_{i^* \leftrightarrow h^*}, y^k_{i^* \leftrightarrow h^*})$ is in the bounding box $B_{h^*}$, then the relationship between $i^*$ and $h^*$ is created; otherwise, even though $h^*$ is associated with $i^*$, no relationship is created between them:

$$c_{i^* \leftrightarrow h^*} = \begin{cases} \text{“}wear\text{”} & , if \quad x^k_{i^* \leftrightarrow h^*} \in \left[x_{h^*} - \frac{w_{h^*}}{2}, x_{h^*} + \frac{w_{h^*}}{2}\right] \ and \\ & \quad\ y^k_{i^* \leftrightarrow h^*} \in \left[y_{h^*} - \frac{h_{h^*}}{2}, y_{h^*} + \frac{h_{h^*}}{2}\right] \\ N/A & , otherwise \end{cases} \tag{4.20}$$

where $c_{i^* \leftrightarrow h^*}$ indicates the connection to create the relationship between $i^*$ and $h^*$ as a semantic phrase $(i^*, c_{i^* \leftrightarrow h^*}, h^*)$ (e.g., the relationship $(person, wear, body\ harness)$) is created in Figure 4.10)

### 4.3.4 Glove

Previous in this work, individual-glove relationship identification was considered to be performed based on the same approach as other objects, that is, first detect gloves using the deep learning-based object detection model, then analyze the geometric relationship between detected gloves with individuals. However, even though CNN could partly recognize different categories via texture or color, recognition of gloves and bare hands, which enjoy the same shape, still reduces the performance to some degree.

Accordingly, individual-glove relationship identification is considered as a color-based skin detection task in hand ROI (Region of Interest), which can be obtained using the wrists keypoints provided by OpenPose, using HSV [1] and YCbCr [2] color space. Compared with RGB color space, HSV and YCbCr color space are capable of processing images of different light conditions, which is beneficial for the detection of different on-site environmental conditions.

As demonstrated in Figure 4.11, the ROI of hands is first extracted based on the wrists keypoints (body parts 4 and 7 in Figure 4.5) provided by

[1] H: hue, S: saturation, V: value (alum)
[2] Y: luminance, Cb: chrominance blue, Cr: chrominance red

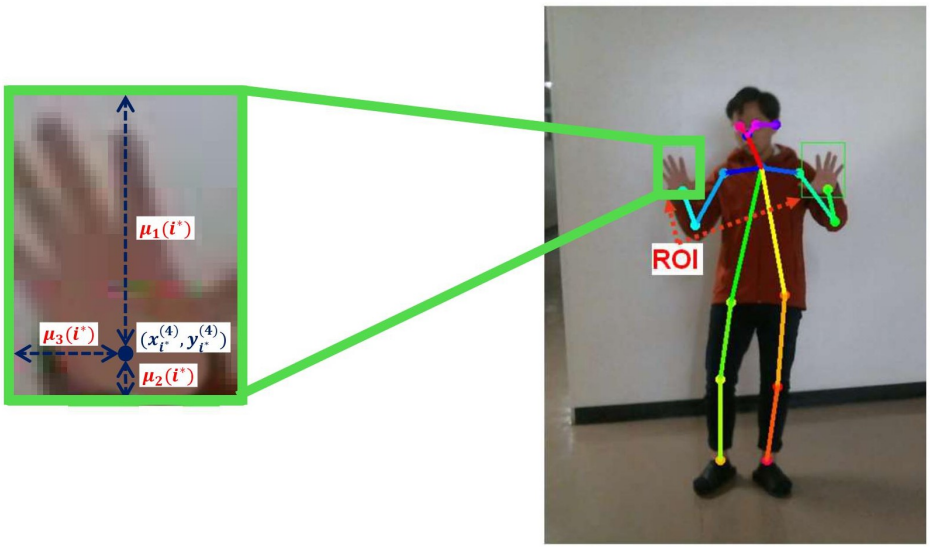



**Figure 4.11: The ROI of hands extraction based on the wrists keypoints.**

OpenPose. The ROI size is given by:

$$\mu_1(i^*) = \max(\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(8)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(8)})^2}, \sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(11)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(11)})^2}) \cdot \gamma_1 \tag{4.21}$$

$$\mu_2(i^*) = \max(\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(8)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(8)})^2}, \sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(11)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(11)})^2}) \cdot \gamma_2 \tag{4.22}$$

$$\mu_3(i^*) = \max(\sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(8)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(8)})^2}, \sqrt{(x_{i^*}^{(1)} - x_{i^*}^{(11)})^2 + (y_{i^*}^{(1)} - y_{i^*}^{(11)})^2}) \cdot \gamma_3 \tag{4.23}$$

where $\mu_1(i^*)$ and $\mu_2(i^*)$ are the vertical offsets of the fingertips direction and

wrists direction, respectively. $\mu_3(i^*)$ is the horizontal offset. $\gamma_1$, $\gamma_2$, and $\gamma_3$ are set to 0.3, 0.05, 0.1, respectively.

Subsequently, to improve the performance, the extracted ROI is converted from RGB to HSV and YCbCr color space. A skin pixel is identified whether the its HSV and YCbCr values lie in a range of predefined threshold values for each parameter and the mask matrixes of HSV and YCbCr color space are created:

$$M_{HSV}(m,n) = \begin{cases} 1 & ,if \quad H(m,n) \in [0,20] \; and \\ & \quad\;\; S(m,n) \in [30,150] \; and \\ & \quad\;\; V(m,n) \in [60,255] \\ 0 & ,otherwise \end{cases} \tag{4.24}$$

$$M_{YCbCr}(m,n) = \begin{cases} 1 & ,if \quad Y(m,n) \in [0,255] \; and \\ & \quad\;\; Cb(m,n) \in [135,180] \; and \\ & \quad\;\; Cr(m,n) \in [85,135] \\ 0 & ,otherwise \end{cases} \tag{4.25}$$

where $H$, $S$, $V$, $Y$, $Cb$, $Cr$ are matrixes assigning the ROI's values of hue, saturation, value (alum), luminance, chrominance blue, chrominance red, respectively.

Besides, element-wise product is performed to merge $M_{HSV}$ and $M_{YCbCr}$ in order to extract only the pixels corresponding to the skin color (Figure 4.12):

$$M_{merge} = M_{HSV} \odot M_{YCbCr} \tag{4.26}$$

Finally, by calculating the skin color pixel proportion in ROI and individual-glove relationship is identified:

$$c_{i^* \leftrightarrow glove} = \begin{cases} \text{“}wear\text{”} & ,if \; \frac{N_{non-zero}(M_{merge})}{N(M_{merge})} > \rho \\ N/A & ,otherwise \end{cases} \tag{4.27}$$

where $N$ and $N_{non-zero}$ are the number of all elements and non-zero ele-

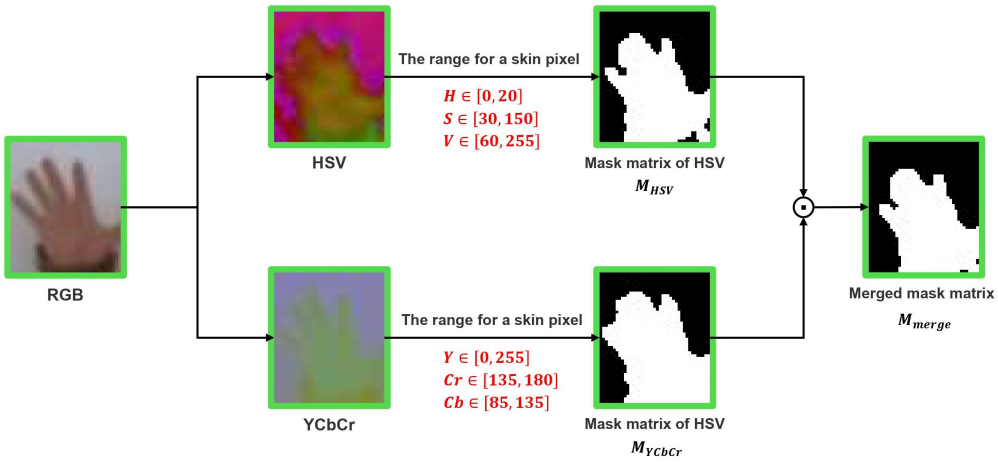



**Figure 4.12: The skin pixel extraction strategies.**

ments in $M_{merge}$, respectively. $\rho$ is the number ratio of skin pixels to identify the glove is worn, whcih is set to 0.2. $c_{i^* \leftrightarrow glove}$ indicates the connection to create the relationship between $i^*$ and glove as a semantic phrase $(i^*, c_{i^* \leftrightarrow glove}, glove)$.

### 4.3.5 Alternative identification strategy

The proposed image information representation module, based on geometric relationship analysis to perform the combination of object detection and individual detection model, can be applied to different on-site environmental conditions. To further improve the robustness of the image information representation module, an alternative identification strategy is proposed in case of the OpenPose's detection failed (e.g., at short camera-to-subject distances (<1m)).

The alternative identification strategy considers the situation that only object detection model is available and employ human face to characterize individuals in the obtained on-site images. The individual-object association and relationship identification is performed using detected face bounding boxes. Take head protection PPE as an example (Figure 4.13), given a pair of associated face and object $\{f^*, j^*\} \in M$, let $B_{f^*} = (x_{f^*}, y_{f^*}, w_{f^*}, h_{f^*})$,

$B_{j^*} = (x_{j^*}, y_{j^*}, w_{j^*}, h_{j^*})$ be the detected bounding box $f^*$ and $j^*$, respectively. If the Euclidean distance between the position $(x_{f^*}, y_{f^*})$ and $(x_{j^*}, y_{j^*})$ is smaller than the reference threshold $\beta_{f^* \leftrightarrow j^*}$, then the relationship between the $f^*$ and $j^*$ is created; otherwise, even though $f^*$ is associated with $f^*$, no relationship is created between them:

$$c_{f^* \leftrightarrow j^*} = \begin{cases} \text{“}wear\text{”} & , if\ d_f(f^*, j^*) < \beta_{f^* \leftrightarrow j^*} \\ N/A & , otherwise \end{cases} \tag{4.28}$$

where

$$d_f(f^*, j^*) = \sqrt{(x_{f^*} - x_{j^*})^2 + (y_{f^*} - y_{j^*})^2} \tag{4.29}$$

$$\beta_{f^* \leftrightarrow j^*} = \max(w_{f^*}, h_{f^*}) \tag{4.30}$$

and $c_{f^* \leftrightarrow j^*}$ indicates the connection to create the relationship between $f^*$ and $j^*$ as a semantic phrase $(f^*, c_{f^* \leftrightarrow j^*}, j^*)$ (e.g., $(person, wear, hard\ hat)$ in Figure 4.13(a)) or not (Figure 4.13(b)).

## 4.4 Information representation

Based on the semantic phrases created by individual-object relationship analysis, a scene graph is generated for image information representation for each obtained image. Given the on-site image's scene graph $G(V, E)$, where $V$ is the set of vertices to represent the objects in the semantic phrase triplets and $E = \{\{\mu, \nu, s, r\} : (\mu, \nu) \in V^2, \mu \neq \nu\}$ is the set of edges to represent the individual-object relationships in the image ($s$ and $r$ are the connections to represent a entity relation and a entity status, respectively).

In the example in Figure 4.14, the on-site image scene graph is $G\,(V, E)$ is consists of the vertices $V = \{person, hard\ hat, dust\ mask, grinder\}$, the edges $E = \{(person, hard\ hat, wear), (person, dust\ mask, wear), (person, grinder, use)\}$.


(a) The relationship is created: $(person, wear, hard\ hat)$.

(b) No relationship is created.

**Figure 4.13: The alternative strategies for individual-object relationship identification.**

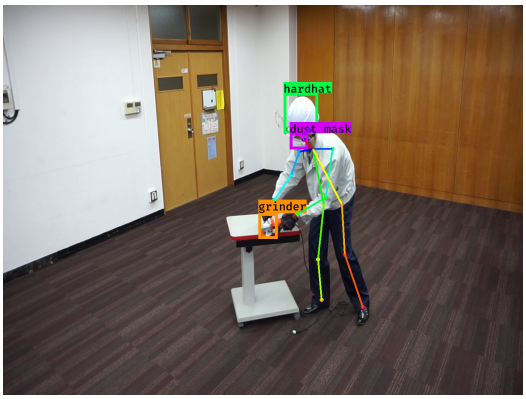
## 4.5 Summary

In this chapter, the proposed image information representation module for on-site image scene understanding and scene graph representation is detailed, which is composed of the following stages:

(1) Image perception: a *Perceptor*, which consists of a deep learning-based object detection model and a pre-trained human pose estimation model, takes on-site images as inputs to detect objects (e.g., PPE, tools) and individuals, respectively;

(2) Individual-object association: for each detected object, it is associated with a closest detected individual to it;

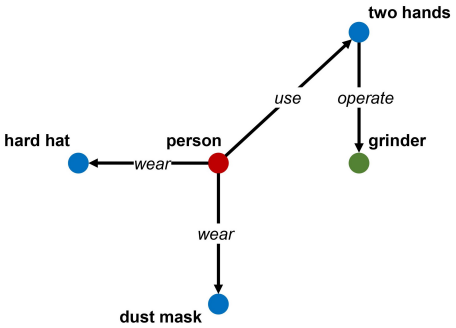
(3) Individual-object relationship analysis: for each associated individual-object pair, the relationship between them (e.g., wear, use, not proper use) is further identified based on geometric relationship analysis of the individual's keypoints and the detected object and represented as semantic phrases. Currently, the relationships between an individual and head protection PPE, grinder, glove, body harness, are covered;




(a)

- (person, wear, hard hat)
- (person, use, two hands)
- (two hands, operate, grinder)
- (person, wear, dust mask)

(b)




(c)

**Figure 4.14: An example of on-site image and its scene graph. (a) On-site image; (b) Individual-object relationships extracted from (a); (c) Scene graph $G(V, E)$**

(4) Scene graph representation: the scene information of each obtained on-site image is represented in a scene graph for further hazards identification.

# Chapter 5

# Automated reasoning for hazards identification

Equipped with the ability to automated extract and represent regulatory information as well as percept and represent on-site image information, now it is time to perform automated reasoning for on-site occupational hazards identification. Based on the on-site image scene graph, the relevant regulatory rules are first extracted from the regulatory hierarchical scene graph to construct a relevant rules scene graph which contains all regulatory rules for the situation of the on-site image. Subsequently, pruning is performed on the relevant rules scene graph to extract the prohibition regulatory rules scene graph and obligation regulatory rules scene graph. Furthermore, automated reasoning for hazard identification is performed based on the extracted scene graphs analysis.

## 5.1 Relevant regulatory rules extraction

By processing the regulatory rules in regulatory information representation module (Chapter 3), a regulatory rules hierarchical scene graph $\hat{G}(\hat{V}, \hat{E}, \hat{C})$ is created which contains the rules for different on-site situations. To specify the reasoning targets and reduce the computational complexity, the relevant regulatory rules for the scene information of the on-site image need to be

first extracted.

As processed in Algorithm 2, given the on-site image's scene graph $G(V, E)$ (Figure 5.1(a)), where $V$ is the set of vertices to represent the detected entities in the obtained on-site images and $E = \{\{\mu, \nu, s, t\} : (\mu, \nu) \in V^2, \mu \neq \nu\}$ is the set of edges to represent the individual-object relationships in the image, the conditional relationship triplets are extracted by searching a subset $V' \subseteq V$ which is the keys of the conditional relation set $\hat{C}$ in $\hat{G}(\hat{V}, \hat{E}, \hat{C})$ (Figure 5.1(b)). As the example showed in Figure 5.1, $V' = \{(person, grinder, use)\}$. Subsequently, the corresponding vertices $\hat{V}'$ and edges $\hat{E}'$ are retrieved by traversing $\hat{C}$ and a subgraph $\hat{G}'(\hat{V}', \hat{E}')$ (Figure 5.1(c)) is created from $\hat{G}$. $\hat{G}'$ contains all regulatory rules for the situation of the on-site image and is further used for hazards identification.

**Algorithm 2** Relevant regulatory rules extraction algorithm.

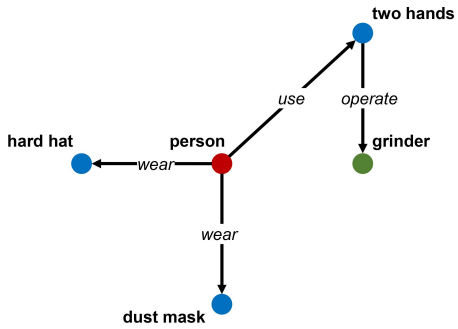
**Input:**

On-site image's scene graph, $G\,(V, E)$;
Regulatory rules hierarchical scene graph, $\hat{G}(\hat{V}, \hat{E}, \hat{C})$;

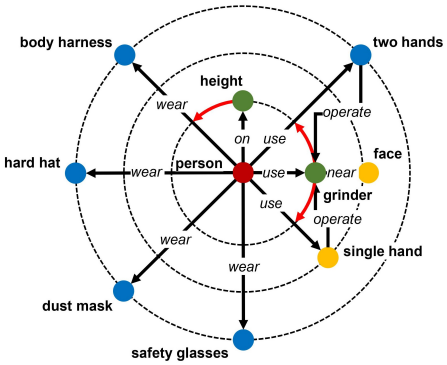
**Output:**

Relevant rules scene graph $\hat{G}'\left(\hat{V}', \hat{E}'\right)$;

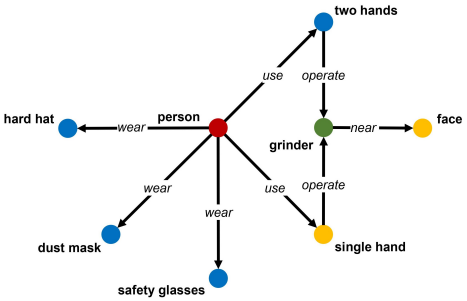
**for all** $e$ such that $e \in E$ **do**
    **if** $\hat{C}.containsKey(e)$ **then**
        $\hat{E}'.insert(e)$
        $\hat{V}'.insert(e.head, e.target)$
        $ins \leftarrow \hat{C}.get(e)$
        **for all** $i$ such that $i \in ins$ **do**
            $\hat{E}'.insert(i)$
            $\hat{V}'.insert(i.head, i.target)$
        **end for**
    **end if**
**end for**
**return** $\hat{G}'\left(\hat{V}', \hat{E}'\right)$




(a)




(b)




(c)

**Figure 5.1: (c) is the relevant rules scene graph $\hat{G}'\left(\hat{V}', \hat{E}'\right)$ extracted from (b) regulatory rules hierarchical scene graph $\hat{G}(\hat{V}, \hat{E}, \hat{C})$ and contains all regulatory rules for the situation of (a) on-site image scene graph $G\,(V, E)$.**

## 5.2 Hazards identification

Based on the extracted relevant regulatory rule scene graph $\hat{G}'\left(\hat{V}', \hat{E}'\right)$, reasoning for hazards identification is performed by checking compliance of prohibition and obligation regulatory rules.

### 5.2.1 Pruning

Pruning is the technique performing to resize the structure of a graph. $\hat{G}'\left(\hat{V}', \hat{E}'\right)$ consists of both prohibition and obligation regulatory rules. Before regulatory rules reasoning, pruning is performed on $\hat{G}'\left(\hat{V}', \hat{E}'\right)$ to extract the prohibition regulatory rules subgraph $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$ and the obligation regulatory rules subgraph $\hat{G_O}'\left(\hat{V_O}', \hat{E_O}'\right)$. The processing of pruning is demonstrated in Algorithm 3.

**Algorithm 3** Relevant regulatory rules scene graph pruning algorithm.

**Input:**

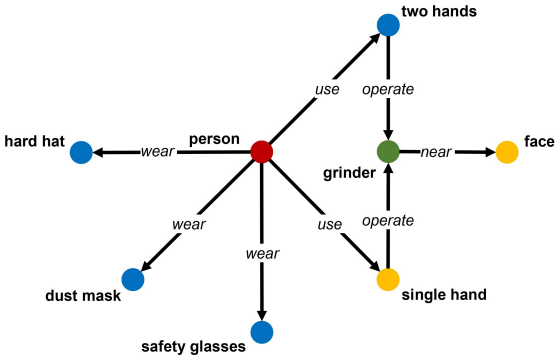
Relevant regulatory rules scene graph, $\hat{G}'\left(\hat{V}', \hat{E}'\right)$;

**Output:**

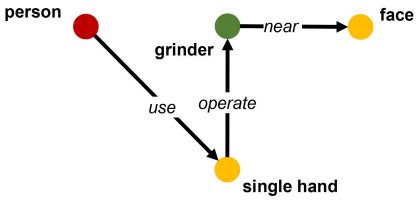
Prohibition regulatory rules subgraph, $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$;

Obligation regulatory rules subgraph, $\hat{G_O}'\left(\hat{V_O}', \hat{E_O}'\right)$;

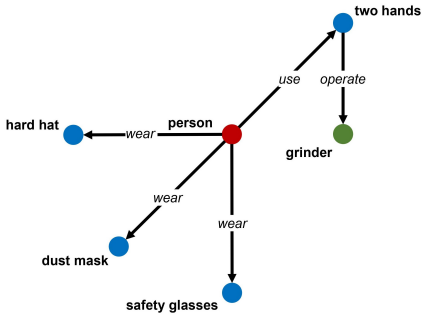
**for all** $\hat{e}'$ such that $\hat{e}' \in \hat{E}'$ **do**
    **if** $\hat{e}'.requirementType$ is *prohibition* **then**
        $\hat{E_P}'.insert(\hat{e}')$
        $\hat{V_P}'.insert(\hat{e}'.head, \hat{e}'.target)$
    **else if** $\hat{e}'.requirementType$ is *obligation* **then**
        $\hat{E_O}'.insert(\hat{e}')$
        $\hat{V_O}'.insert(\hat{e}'.head, \hat{e}'.target)$
    **end if**
**end for**
    **return** $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$, $\hat{G_O}'\left(\hat{V_O}', \hat{E_O}'\right)$




(a)




(b)




(c)

**Figure 5.2: By performing pruning on (a) the relevant regulatory rules scene graph $\hat{G}'\left(\hat{V}', \hat{E}'\right)$, (b) the prohibition regulatory rules subgraph $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$ and (c) the obligation regulatory rules subgraph $\hat{G_O}'\left(\hat{V_O}', \hat{E_O}'\right)$ are extracted.**

Figure 5.2 demonstrates an example of pruning. Given a relevant regulatory rules scene graph $\hat{G}'\left(\hat{V}', \hat{E}'\right)$, where $\hat{V}' = \{person, hard\ hat, dust\ mask, safety\ glasses, grinder, face, single\ hand, two\ hands\}$ and $\hat{E}' = \{(person, hard\ hat, wear, obligation), (person, dust\ mask, wear, obligation), (person, safety\ glasses, wear, obligation), (person, single\ hand, use), (person, two\ hands, use), (single\ hand, grinder, use, prohibition), (two\ hands, grinder, use, obligation), (grinder, face, near, prohibition)\}$, by performing pruning the prohibition regulatory rules subgraph $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$ is extracted where $\hat{V_P}' = \{person, grinder, face, single\ hand\}$ and $\hat{E_P}' = \{(person, single\ hand, use), (single\ hand, grinder, use, prohibition), (grinder, face, near, prohibition)\}$, together with the obligation regulatory rules subgraph $\hat{G_O}'\left(\hat{V_O}', \hat{E_O}'\right)$ where $\hat{V_O}' = \{person, hard\ hat, dust\ mask, safety\ glasses, grinder, two\ hands\}$ and $\hat{E_O}' = \{(person, hard\ hat, wear, obligation), (person, dust\ mask, wear, obligation), (person, safety\ glasses, wear, obligation), (person, two\ hands, use), (two\ hands, grinder, use, obligation)\}$

### 5.2.2 Prohibition regulatory rules reasoning

Prohibition regulatory rules reasoning is performed based on compliance checking between the on-site image's scene graph $G(V, E)$ and the prohibition regulatory rules subgraph $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$.

As processed in Algorithm 4, if an edge $\hat{e_P}'$ of $\hat{E_P}'$ exists in $E$, which means a prohibition entities relationship exists in the on-site image scene, then $\hat{e_P}'$ is extracted as a violated regulatory prohibition rule and the on-site image scene is hence identified as hazardous.

**Algorithm 4** Prohibition regulatory rules reasoning algorithm.

**Input:**

1: On-site image's scene graph, $G(V, E)$;

2: Prohibition regulatory rules subgraph $\hat{G_P}'\left(\hat{V_P}', \hat{E_P}'\right)$;

**Output:**

3: Violated prohibition regulatory rules; $H_P$

4:

5: **for all** $\hat{e_P}'$ such that $\hat{e_P}' \in \hat{E_P}'$ **do**
6: **if** $\hat{e_P}'$ in $E$ **then**
7: $H_P.insert(\hat{e_P}')$
8: **end if**
9: **end for**
**return** $H_P$

### 5.2.3 Obligation regulatory rules reasoning

The obligation regulatory rules reasoning for hazards identification of the on-site image is performed based on the isomorphism between $G(V, E)$ and $\hat{G_O}'(\hat{V_O}', \hat{E_O}')$.

In graph theory, an isomorphism is a mapping between two graph structures of the same type that can be reversed by an inverse mapping. $G(V, E)$ is isomorphic to $\hat{G_O}'(\hat{V_O}', \hat{E_O}')$, if there exists a bijective function $f : V \to \hat{V_O}'$ such that $\forall u, v \in V, (u, v) \in E \leftrightarrow (f(u), f(v)) \in \hat{E_O}'$, which is denoted as $G \cong \hat{G_O}'$ [74]. Otherwise, $G(V, E)$ is non-isomorphic to $\hat{G_O}'(\hat{V_O}', \hat{E_O}')$ and the violated obligation regulatory rules $H_O = \{(\mu, \nu, s, r) \in \hat{E_O}', (\mu, \nu, s, r) \notin E\}$ from the on-site image are identified. In the case of $G(V, E)$ in Figure 5.1(a) and $\hat{G_O}'(\hat{V_O}', \hat{E_O}')$ in Figure 5.2(c), $H_O = (person, safety\ glasses, use)$ which represents the hazard that the person is not wearing safety glassees when operating the grinder.

## 5.3 System development

Based on the proposed occupational hazard identification approach, together with the proposed image information representation approach, a novel real-time system, has been developed to perform on-site occupational hazard identification. The open-source libraries *OpenCV* [75] and *Tkinter* [76] are deployed for real-time image capturing/preprocessing and user interface (UI) rendering, respectively. As demonstrated in Figure 5.3, the real-time on-site occupational hazards identification system provides dual-window for raw image visualization and image information representation. The identified re-

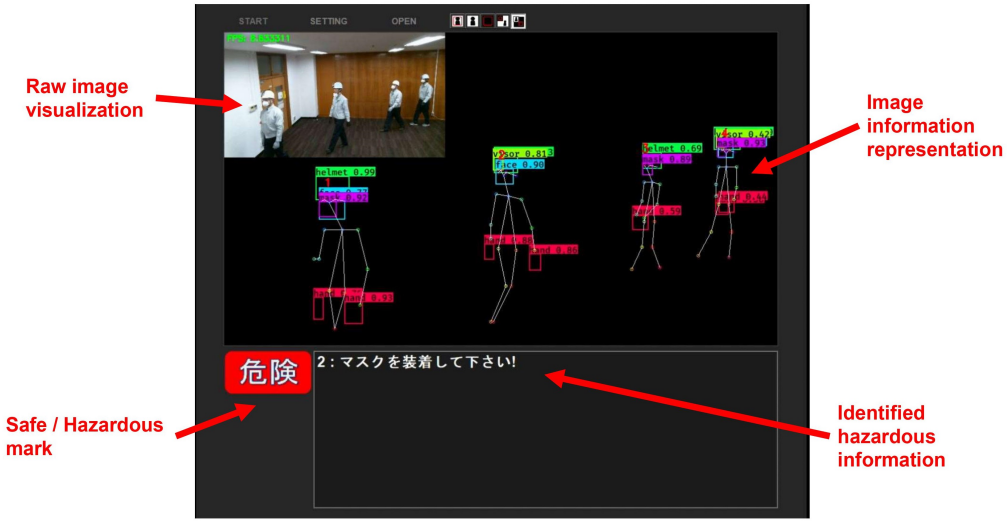



**Figure 5.3: UI of the real-time on-site hazards identification system.**

sults are shown at the bottom of the UI to indicate the hazardous information of the on-site monitoring scene. The performed of the developed system is validated in Chapter 6.

## 5.4 Summary

In this chapter, the proposed automated reasoning module for hazards identification is detailed, which consists of the following stages:

(1) Relevant regulatory rules extraction: the regulatory rules for the on-site images are first extracted from the regulatory rules hierarchical scene graph to specify the reasoning targets and reduce the computational complexity;

(2) Hazards identification: based on the extracted relevant regulatory rules, pruning is performed to separate and create the prohibition and obligation regulatory rules subgraphs. Prohibition and obligation regulatory rules reasoning are performed based on compliance checking and isomorphism analysis.

Based on the proposed approach, a novel real-time on-site hazards identification system has been developed.

# Chapter 6

# Experiments and results

To evaluate the performance of the proposed image-sentence inference model for on-site occupational hazards idenfication, the experiments are performed. Firstly, ten construction/decommissioning regulatory rules were selected to validate the performance of the developed automated regulatory information processing system. Subsequently, the image datasets were created to train the object detection model and certify the performance of the on-site occupational hazards identification system. Furthermore, the results of regulatory information representation experiments and the on-site hazards identification experiments were discussed. Lastly, computational efficiency analysis results were reported.

## 6.1 Experimental description

### 6.1.1 Regulatory rules

To demonstrate the validity of the proposed regulatory information representation approach and the developed automated regulatory rules processing system in this work (Chapter 3), ten construction/decommissioning regulatory rules related to proper PPE/grinder use on construction/decommissioning sites were selected to perform the regulatory information representation experiments:

(1) “Wear a hard hat on a construction site.”

(2) “Wear a hard hat on a decommissioning site.”

(3) “Wear a dust mask on a construction site.”

(4) “Wear a full-face mask on a decommissioning site.”

(5) “Use body harness when working on height.”

(6) “Wear gloves on a decommissioning site.”

(7) “Wear gloves when operating a grinder.”

(8) “Wear a safety glasses when operating a grinder.”

(9) “Always use two hands when operating a grinder.”

(10) “Never operate a grinder near face.”

### 6.1.2 Image datasets

#### Training dataset

To create the training dataset of object detection, images of hard hats, dust masks, full-face masks, safety glasses, body harnesses, and grinders were collected from two sources: downloading Internet images retrieved by search engines using keywords, and capturing real-world images using the webcam as listed in Table 6.1. A total of 13,893 image samples were collected and annotated using the graphical image annotation tool *LabelImg* [77]. The annotations were saved as XML files in PASCAL VOC format to train the YOLOv3 model.

#### Validation dataset

Furthermore, to create the validation dataset to validate the performance of the trained model and the developed on-site occupational hazards identification system (Chapter 4), seven volunteers were instructed to perform normal working behaviors and abnormal grinder operating behaviors while

**Table 6.1: Information of collected training dataset**

| | Number of internet image samples | Number of real-world image samples | Total |
|---|---|---|---|
| Hard hat | 1,323 | 3,076 | 4,399 |
| Dust mask | 983 | 1,222 | 2,205 |
| Full-face mask | 642 | 1,021 | 1,663 |
| Safety glasses | 116 | 2,100 | 2,216 |
| Body harness | 136 | 872 | 1,008 |
| Grinder | 356 | 2,046 | 2,402 |
| Overall | 3,556 | 10,337 | 13,893 |

wearing PPEs at different distances to the camera. As surveillance cameras are placed in various locations on construction/decommissioning sites, and the trajectory of workers is stochastic, workers and objects were captured in different resolutions in the surveillance videos. Thus, different distance conditions (3m, 5m, 7m) were considered in the experiments to validate the robustness of the proposed model. Besides, considering the impact of illumination, the images in the validation dataset were captured under different environments (e.g., high level of illumination while working outdoors and weaker level of illumination while working indoors). Additionally, there are a single worker or multiple workers in each image frame and had a variety of postures with different angles to the camera. Finally, 9,000 images (3000 images for each distance case) were randomly selected from the collected image sequences and created the validation dataset. The details are provided in Table 6.2 where positive samples refer to the individuals who are wearing PPE or operating the grinder properly and negative samples referred to the individuals who are using PPE or operating the grinder improperly.

### 6.1.3 Evaluation metrics

Precision and recall were adopted to evaluate the performance of the proposed approach:

$$Precision = \frac{TP}{TP + FP} \tag{6.1}$$

**Table 6.2: Information of collected validation dataset**

| Distance (m) | Number of images | Categories | Number of positive samples | Number of negative samples |
|---|---|---|---|---|
| | | Hard hat | 2,664 | 1,407 |
| | | Dust mask | 1,929 | 1,993 |
| | | Full-face mask | 526 | 792 |
| 3 | 3,000 | Safety glasses | 866 | 1,684 |
| | | Body harness | 934 | 1,616 |
| | | Glove | 200 | 200 |
| | | Grinder | 100 | 100 |
| | | Hard hat | 3,065 | 1,526 |
| | | Dust mask | 2,260 | 2,211 |
| | | Full-face mask | 533 | 799 |
| 5 | 3,000 | Safety glasses | 634 | 2,454 |
| | | Body harness | 1,332 | 1,756 |
| | | Glove | 200 | 200 |
| | | Grinder | 100 | 100 |
| | | Hard hat | 3,075 | 1,594 |
| | | Dust mask | 2,452 | 2,077 |
| | | Full-face mask | 519 | 783 |
| 7 | 3,000 | Safety glasses | 413 | 2,741 |
| | | Body harness | 1,425 | 1,729 |
| | | Glove | 200 | 200 |
| | | Grinder | 100 | 100 |

**Table 6.3: Defination of TP, FP, and FN**

| | Predicted | Ground truth |
|---|---|---|
| TP | Proper use (safe) | Proper use (safe) |
| FP | Proper use (safe) | Improper use (hazardous) |
| FN | Improper use (hazardous) | Proper use (safe) |

$$Recall = \frac{TP}{TP + FN} \tag{6.2}$$

where $TP$ (true positive) is defined as the number of correct identification of individuals who are properly using PPE/grinder , $FP$ (false positive) is the number of wrong identification of individuals who are properly using PPE/grinder, while $FN$ (false negative) is the number of the ground truth not identified as defined in Table 6.3.

## 6.2 Object detection model training

As introduced in Section 4.1.1, YOLOv3 deploys nine anchors to predict log-space transforms. To obtain anchors for the YOLOv3 models, all the annotated bounding boxes in the training dataset were clustered into nine groups using k-means clustering ($k = 9$). For an image of size $416 \times 416$, YOLOv3 model predicts $((52 \times 52) + (26 \times 26) + (13 \times 13)) \times 3 = 10,647$ bounding boxes. To optimize the predicted results, boxes were first filtered based on their objectness score, and the bounding boxes having scores below the threshold 0.5 were ignored. Subsequently, Non-maximum Suppression (NMS) was performed to select the optimized bounding box when several boxes overlap with each other and detect the same object using the Intersection Over Union (IOU) metric. As shown in Figure 6.1, IOU represents the percentage of overlap between two boxes, e.g., the ground-truth box ($G$) and the predicted box ($P$), and is calculated as follows:

$$IOU = \frac{intersection}{union} = \frac{G \cap P}{G \cup P} \tag{6.3}$$

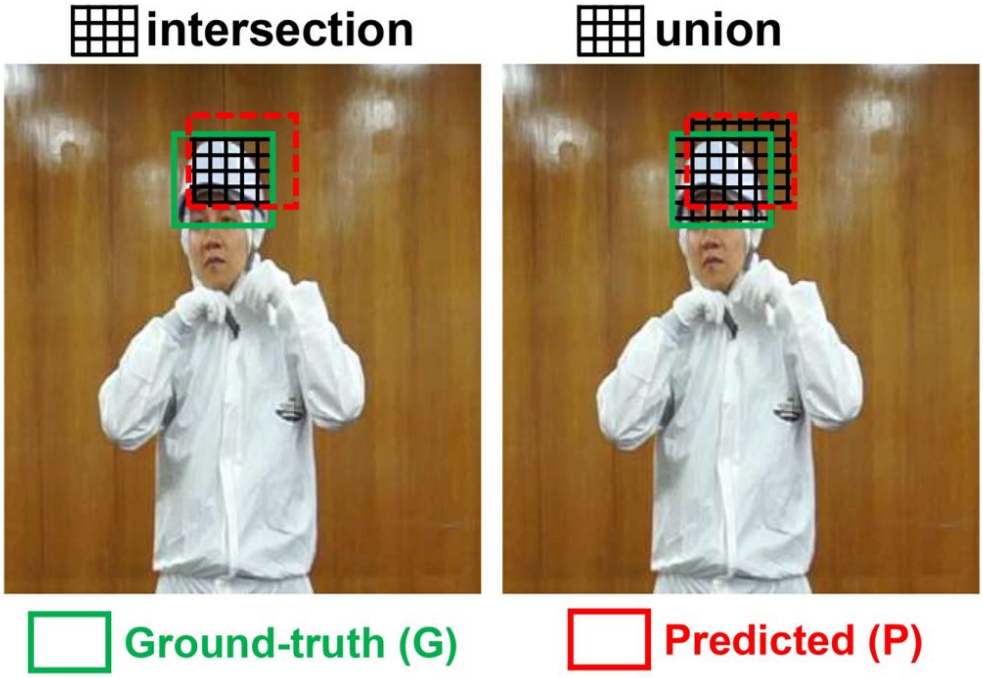



**Figure 6.1: Calculation of IOU.**

Thus, NMS was performed following the subsequent steps:

(1) Select the bounding box that has the highest score.

(2) Compute its overlap IOU with all other bounding boxes and remove boxes that overlap it more than the threshold (= 0.45).

(3) Go back to step (1) and iterate until there are no more bounding boxes with a lower score than the current selected bounding box.

Finally, all the bounding boxes that have a large overlap with the selected bounding boxes were removed, and only the optimized bounding boxes remain.

The YOLOv3 model was built using TensorFlow [78] and initialized based on pre-trained weights on the ImageNet dataset [79]. Training of YOLOv3 was performed in two stages. All convolutional layers were first frozen up to the last convolutional block in Darknet-53, and the model was trained with

frozen layers to get a stable loss in 50 epochs. Subsequently, all convolutional layers of Darknet-53 proceeded to unfreeze to perform fine-tuning in 50 epochs. The learning rate is schedule is as follows: for the first stage the model was trained with a learning rate of $1e-3$; for the second stage the model was trained with a learning rate started at $1e-4$, then "ReduceLROnPlateau" metric was employed to reduce the learning rate to avoid diverges due to unstable gradients. "ReduceLROnPlateau" is a callback that monitors a quantity, and if no improvement is seen for a "patience" number of epochs, the learning rate is reduced. Adaptive Moment Estimation (Adam) optimizer was adopted to adjust the learning rate during optimization automatically. $\alpha$ (initial learning rate), $\beta_1$ (exponential decay rate for the first moment estimates), $\beta_2$ (exponential decay rate for the second-moment estimates) and $\xi$ (a very small number to prevent any division by zero in the implementation) were set to $1e-3$, 0.9, 0.99, $1e-8$, respectively. Besides, a batch size of 8 is used throughout training.

## 6.3 Results

### 6.3.1 Regulatory information representation results

The regulatory information representation experiments were performed on the selected construction/decommissioning regulatory rules (Section 6.1.1). The developed automated regulatory rules processing system (Section 3.5) processed the regulatory rules following the steps below:

(1) Firstly, the dependency trees, with both POS tags and dependency labels, were created as the outputs of the preprocessing (Section 3.1) and feature generation (Section 3.2) stages as shown in Figure 6.2 - Figure 6.11.

(2) Subsequently, semantic analysis was performed by ontology modeling to automatically extract the entities, together with their attributes and relations between them. The topologies were automated generated and demonstrated in Figure 6.12 - Figure 6.21.

(3) The extracted information in ontology model was encoded in the logic representation with semantic phrases and logical connectives (Table 6.4) which correctly demonstrated the extracted entities and requirements from regulatory rules.

(4) Finally, based on the logic representation, the hierarchical scene graph $\hat{G}\left(\hat{V}, \hat{E}, \hat{C}\right)$ is generated for regulatory information representation as demonstrated in Figure 6.22, where $\hat{V} = \{person, construction\ site, decommissioning\ site, height, grinder, hard\ hat, gloves, body\ harness, dust\ mask, full{-}face\ mask, safety\ glasses, face, two\ hands\}$, the edges $\hat{E} = \{(person, construction\ site, on), (person, decommissioning\ site, on), (person, height, on), (person, grinder, use), (person, hard\ hat, wear, obligation), (person, gloves, wear, obligation), (person, body\ harness, use, obligation), (person, dust\ mask, wear, obligation), (person, full{-}face\ mask, wear, obligation), (person, safety\ glasses, wear, obligation), (person, face, near, prohibition), (person, use, two\ hands, obligation), (two\ hands, operate, grinder)\}$, and the conditional relations set $\hat{C} = \{((person, construction\ site, on)) \rightharpoonup ((person, hard\ hat, wear, obligation), (person, dust\ mask, wear, obligation)), ((person, decommissioning\ site, on)) \rightharpoonup ((person, hard\ hat, wear, obligation), (person, full{-}face\ mask, wear, obligation)), ((person, height, on)) \rightharpoonup ((person, body\ harness, use, obligation)), ((person, grinder, use)) \rightharpoonup ((person, gloves, wear, obligation), (person, safety\ glasses, wear, obligation), (person, face, near, prohibition), (person, use, two\ hands, obligation), (two\ hands, operate, grinder))\}$. $\hat{G}\left(\hat{V}, \hat{E}, \hat{C}\right)$ contains all regulatory information of the target regulatory rules which demonstrated the effective of the developed system on automated regulatory information extraction and representation. $\hat{G}\left(\hat{V}, \hat{E}, \hat{C}\right)$ was further used to perform the hazard identification experiments.

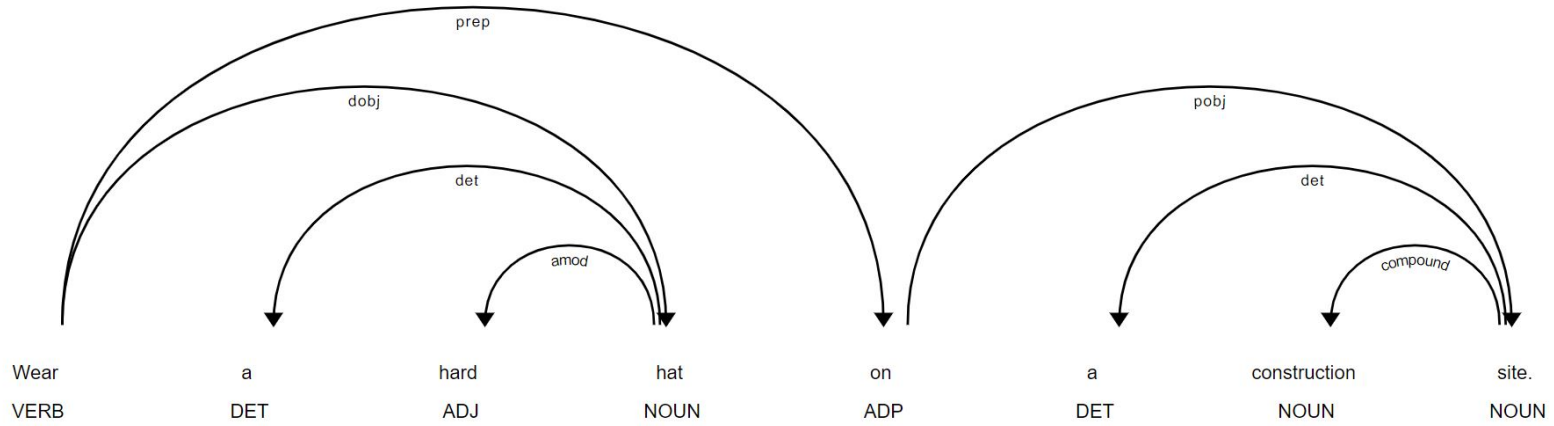



**Figure 6.2: The dependency tree of regulatory rule (1): "Wear a hard hat on a construction site."**

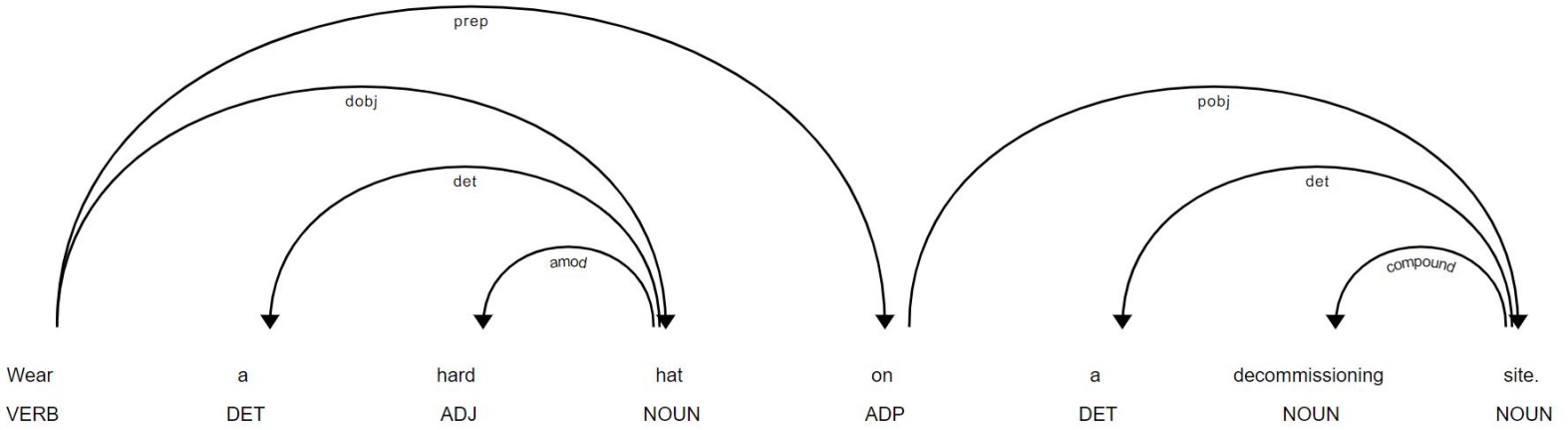



**Figure 6.3: The dependency tree of regulatory rule (2): "Wear a hard hat on a decommissioning site."**

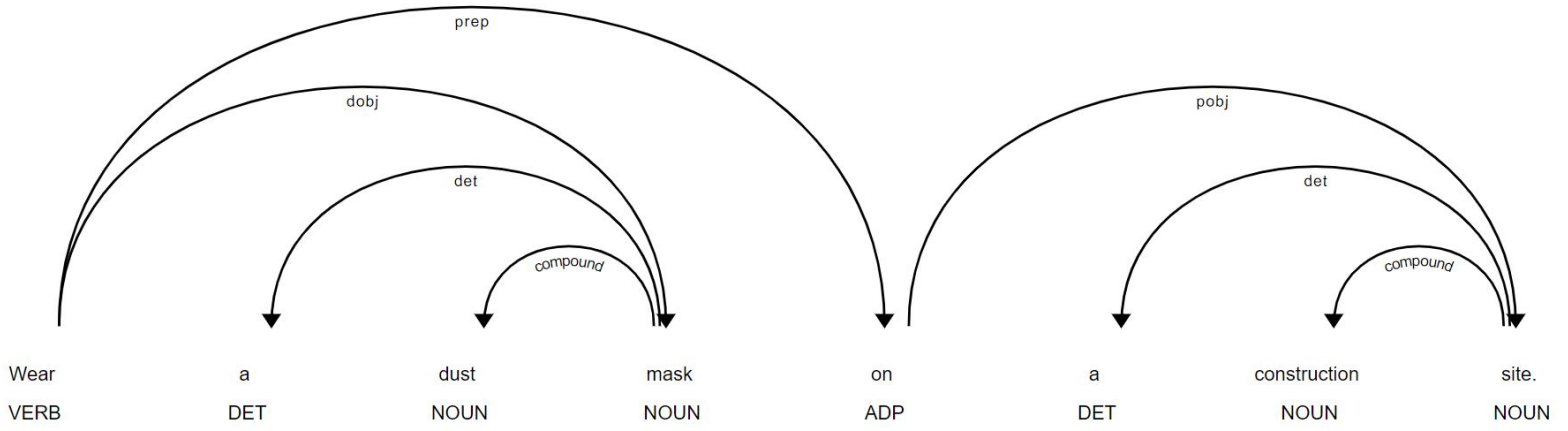



**Figure 6.4: The dependency tree of regulatory rule (3): "Wear a dust mask on a construction site."**

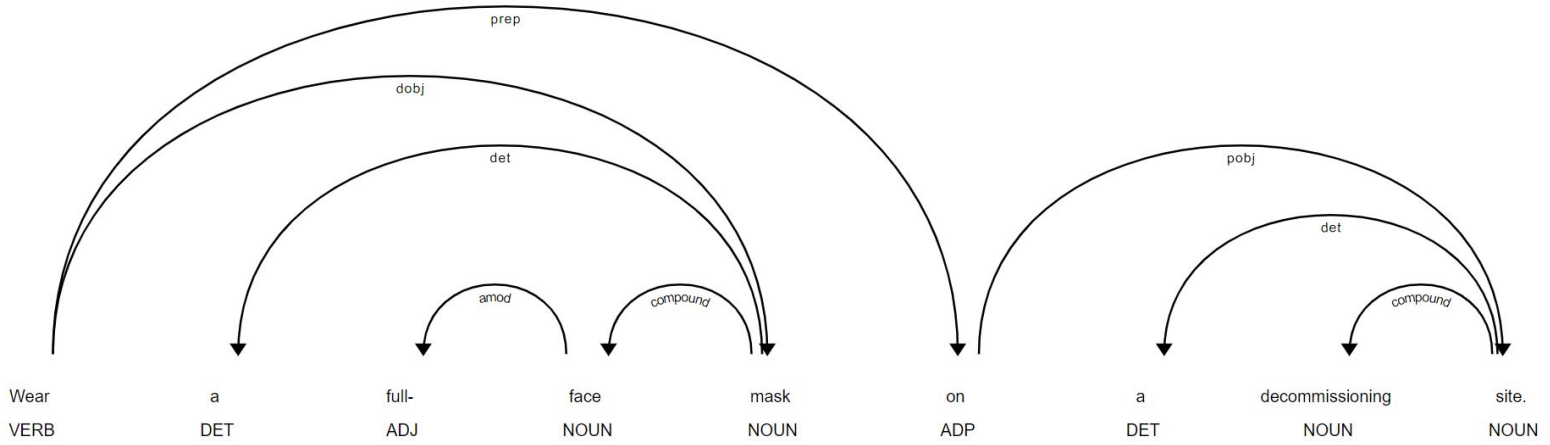



**Figure 6.5: The dependency tree of regulatory rule (4): "Wear a full-face mask on a decommissioning site."**

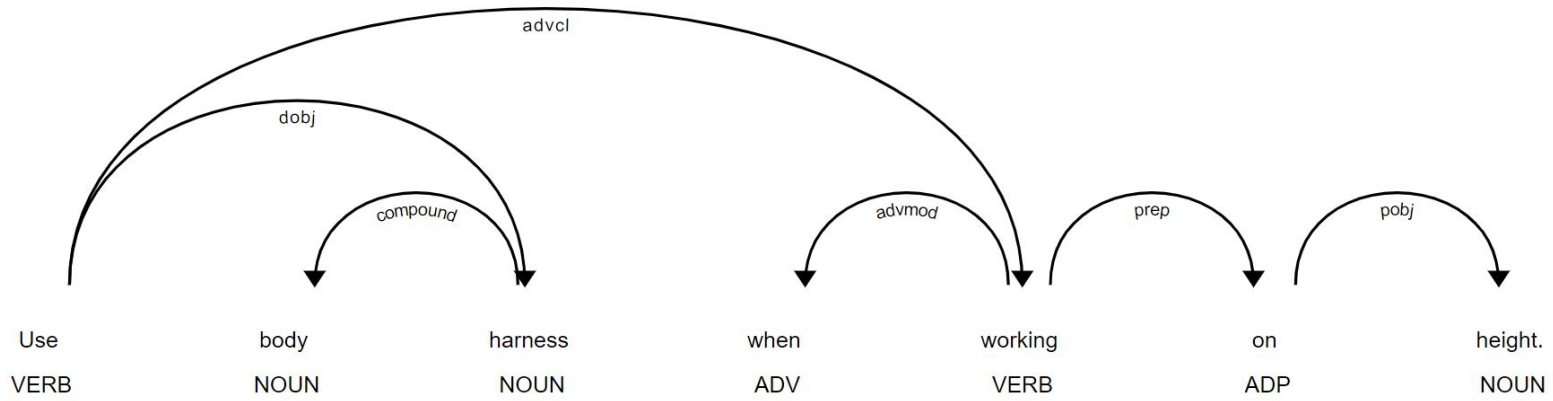



**Figure 6.6: The dependency tree of regulatory rule (5): "Use body harness when working on height."**

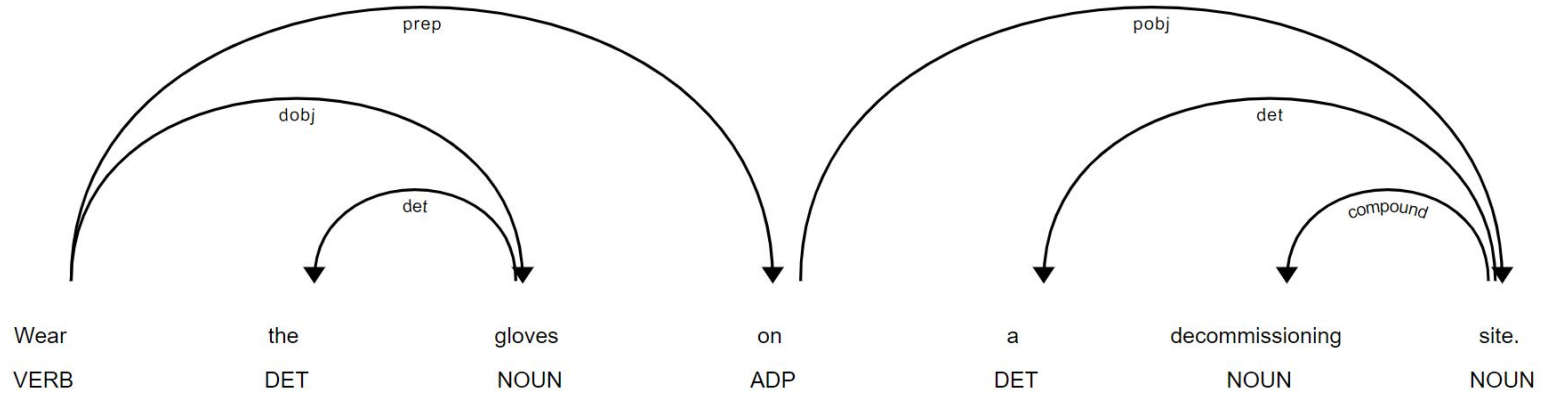



**Figure 6.7: The dependency tree of regulatory rule (6): "Wear gloves on a decommissioning site."**

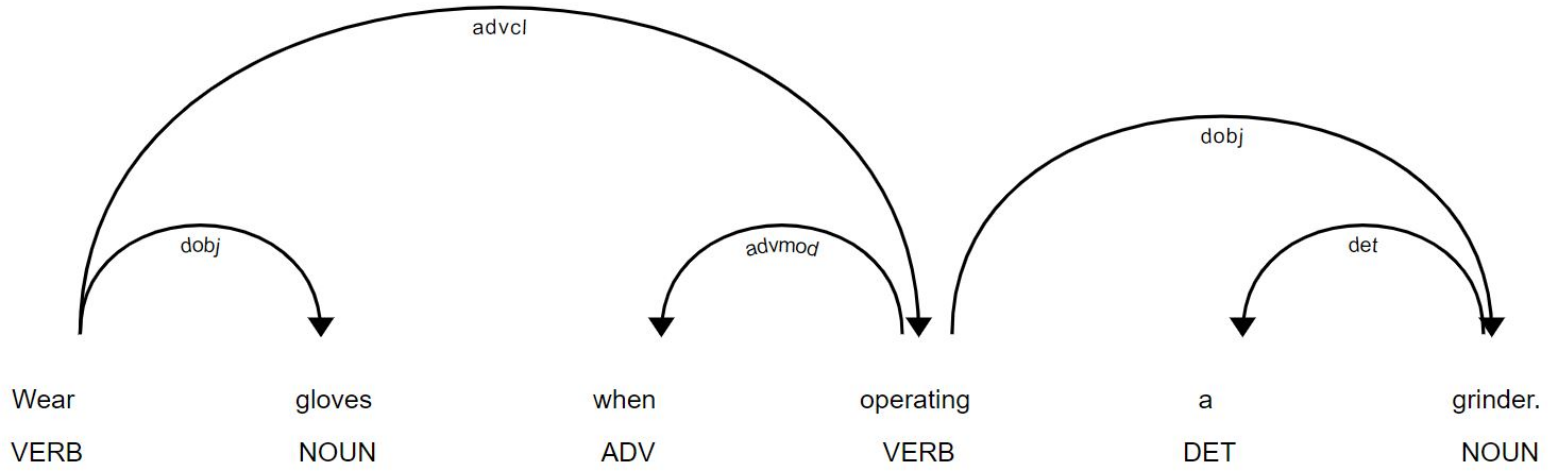



**Figure 6.8: The dependency tree of regulatory rule (7): "Wear gloves when operating a grinder."**

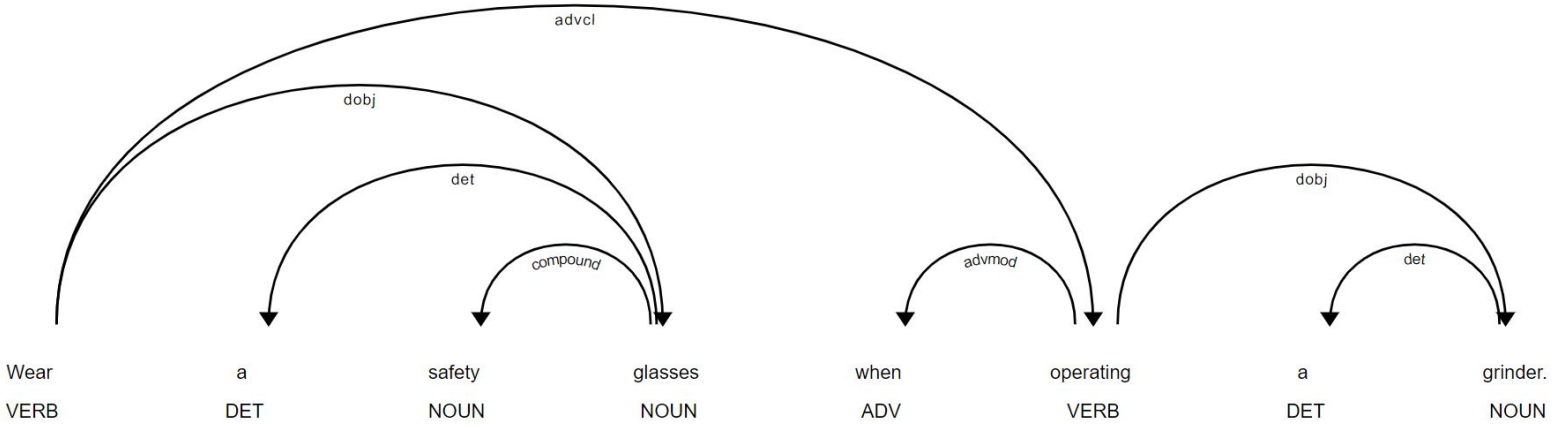



**Figure 6.9: The dependency tree of regulatory rule (8): "Wear a safety glasses when operating a grinder."**

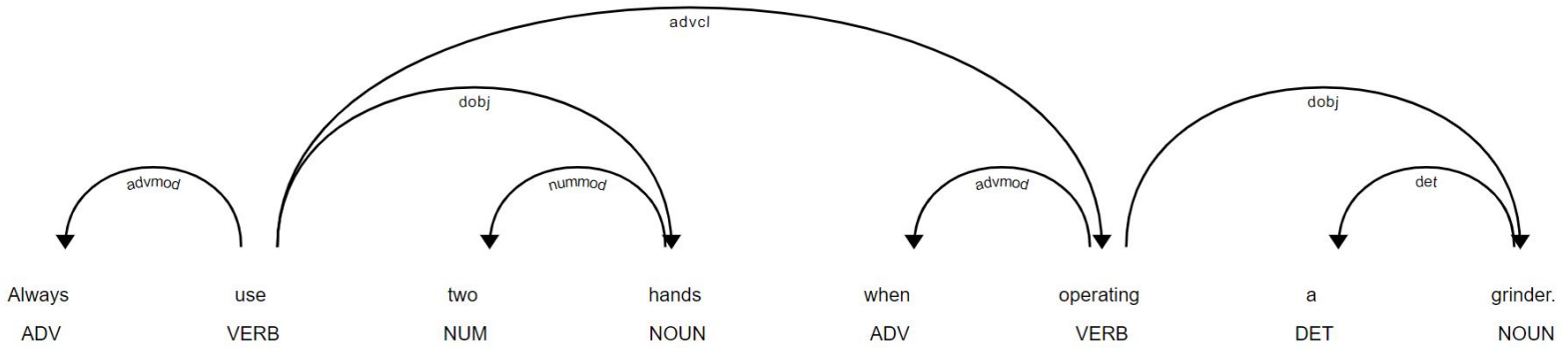



**Figure 6.10: The dependency tree of regulatory rule (9): "Always use two hands when operating a grinder."**

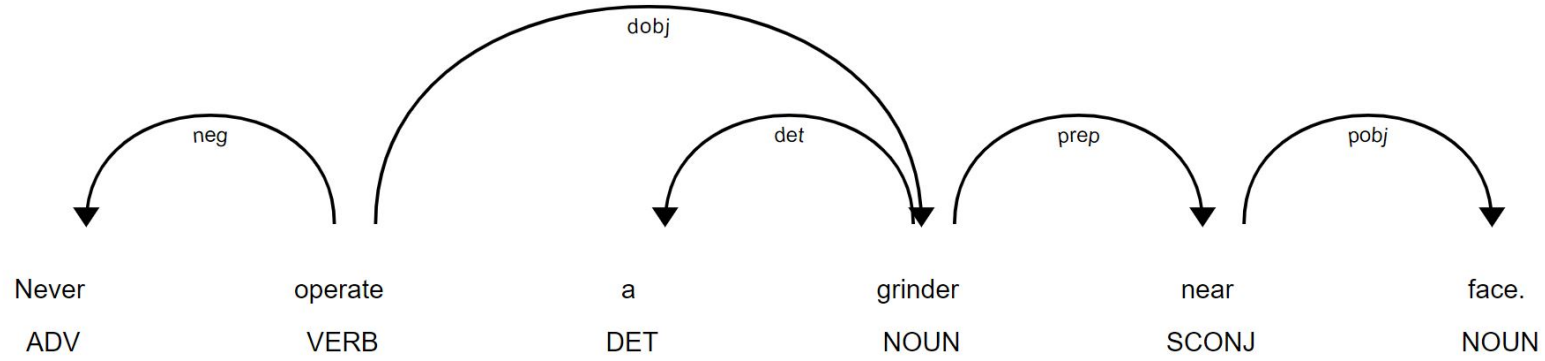



**Figure 6.11: The dependency tree of regulatory rule (10): "Never operate a grinder near face."**

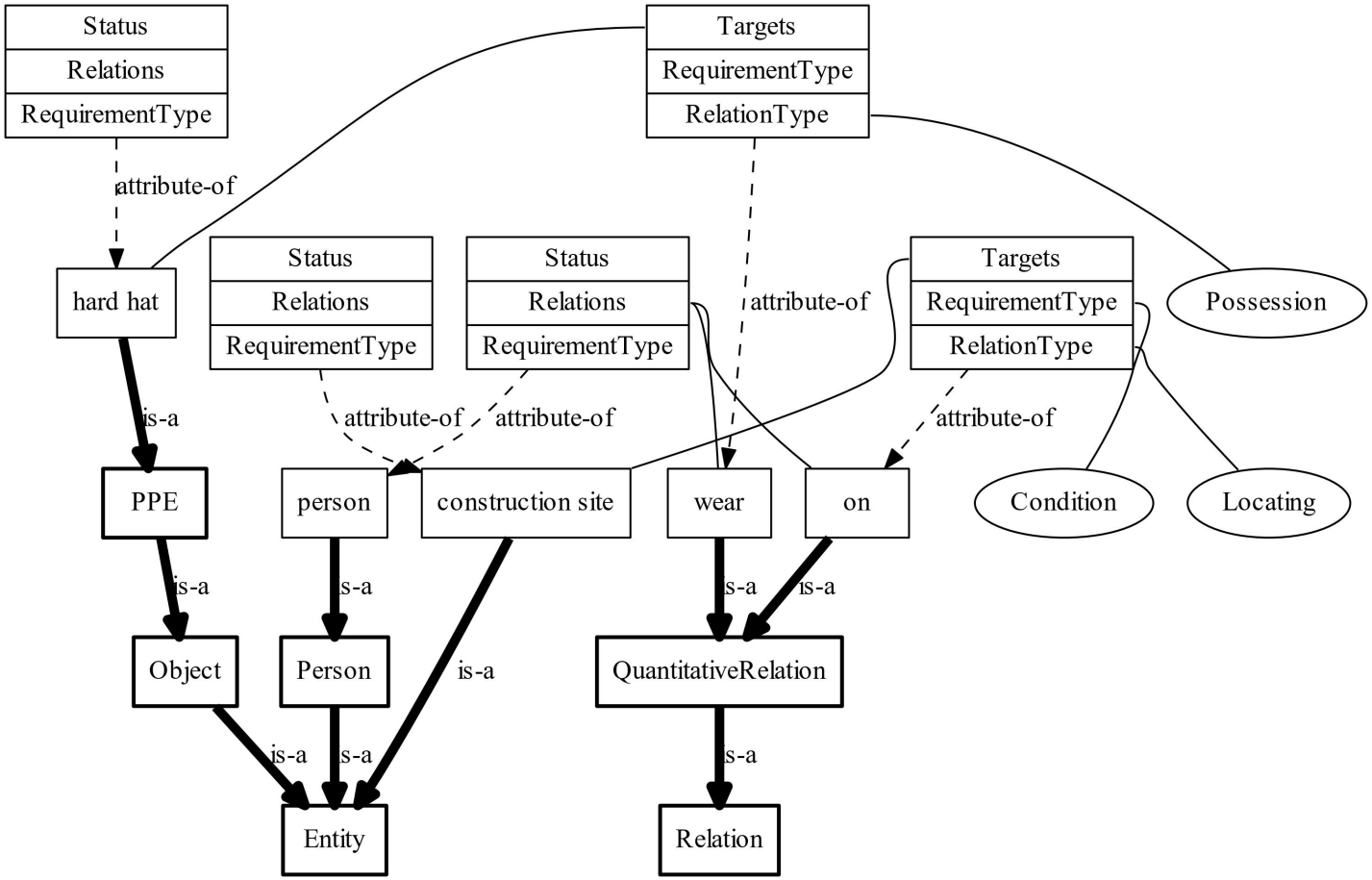



**Figure 6.12: Ontology modeling for regulatory rule (1): "Wear a hard hat on a construction site."**

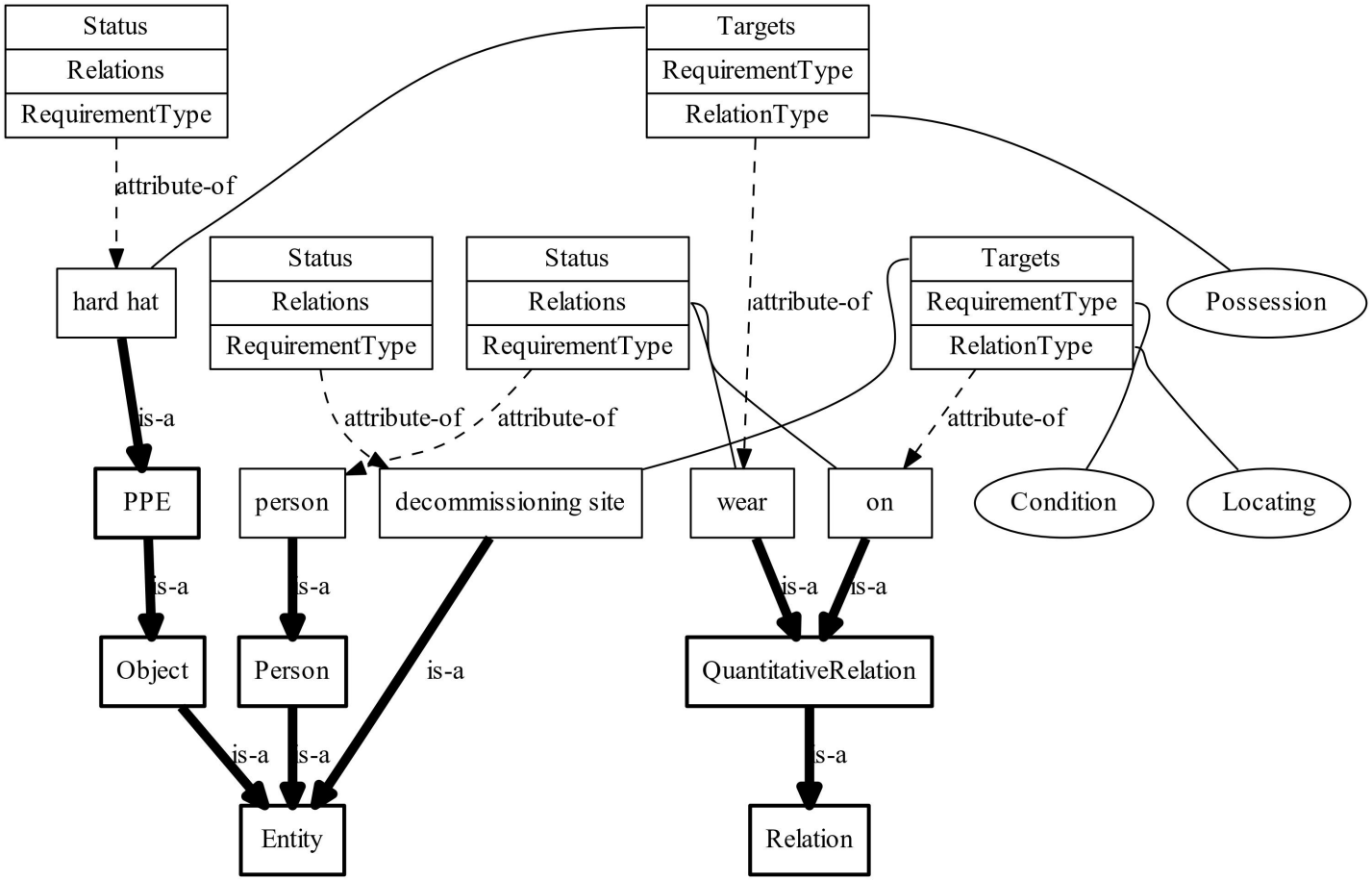



**Figure 6.13: Ontology modeling for regulatory rule (2): "Wear a hard hat on a decommissioning site."**

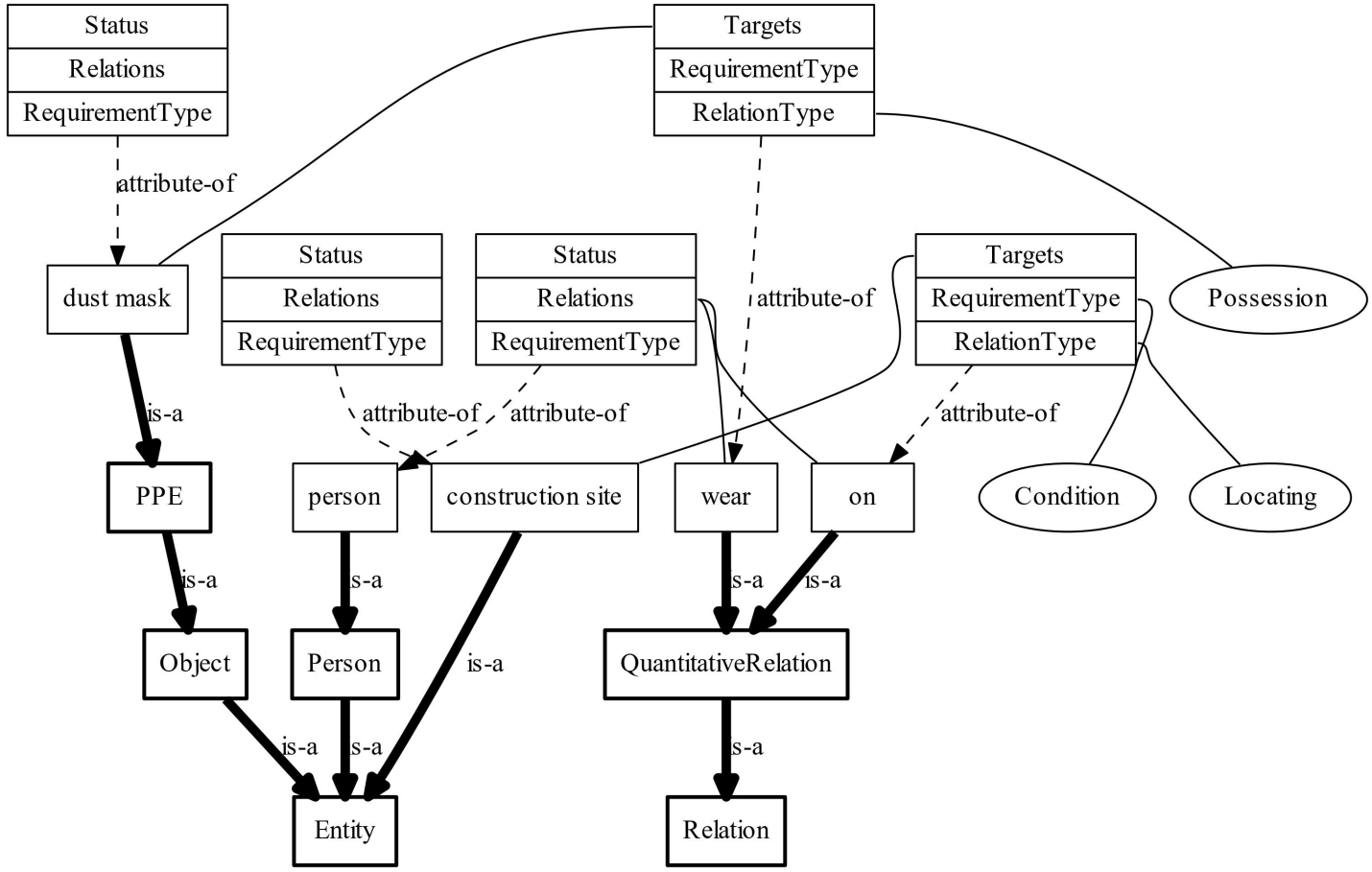



**Figure 6.14: Ontology modeling for regulatory rule (3): "Wear a dust mask on a construction site."**

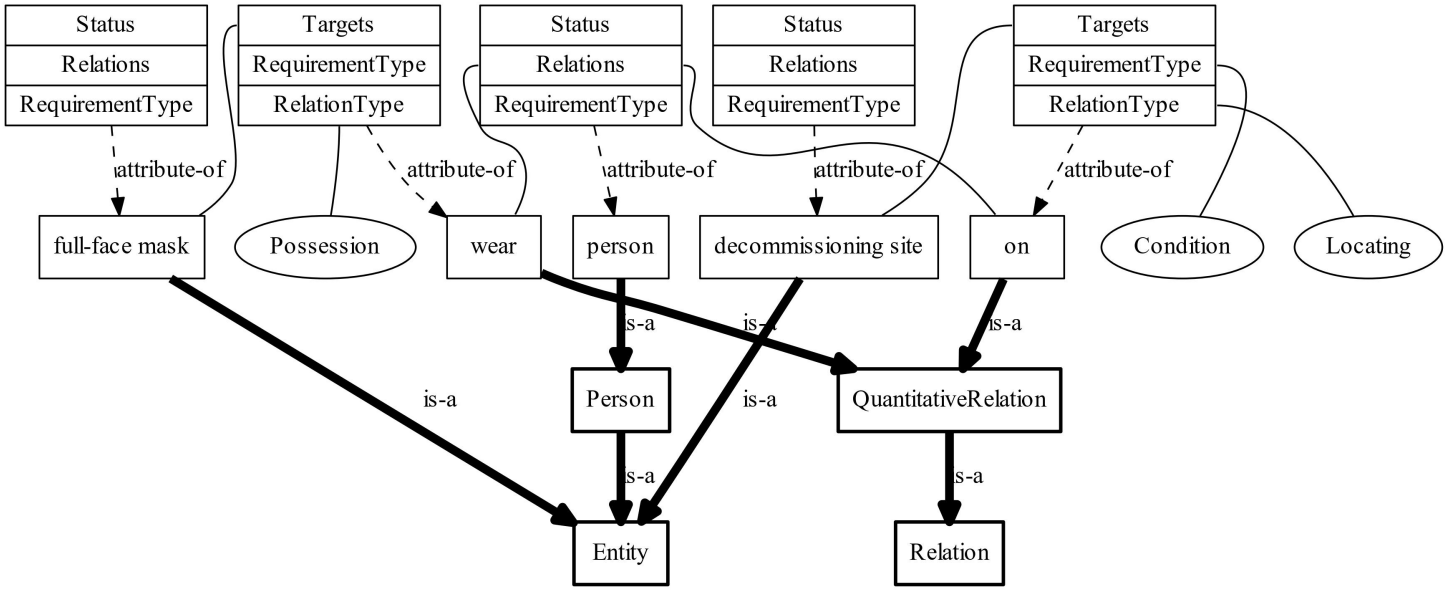



**Figure 6.15: Ontology modeling for regulatory rule (4): "Wear a full-face mask on a decommissioning site."**

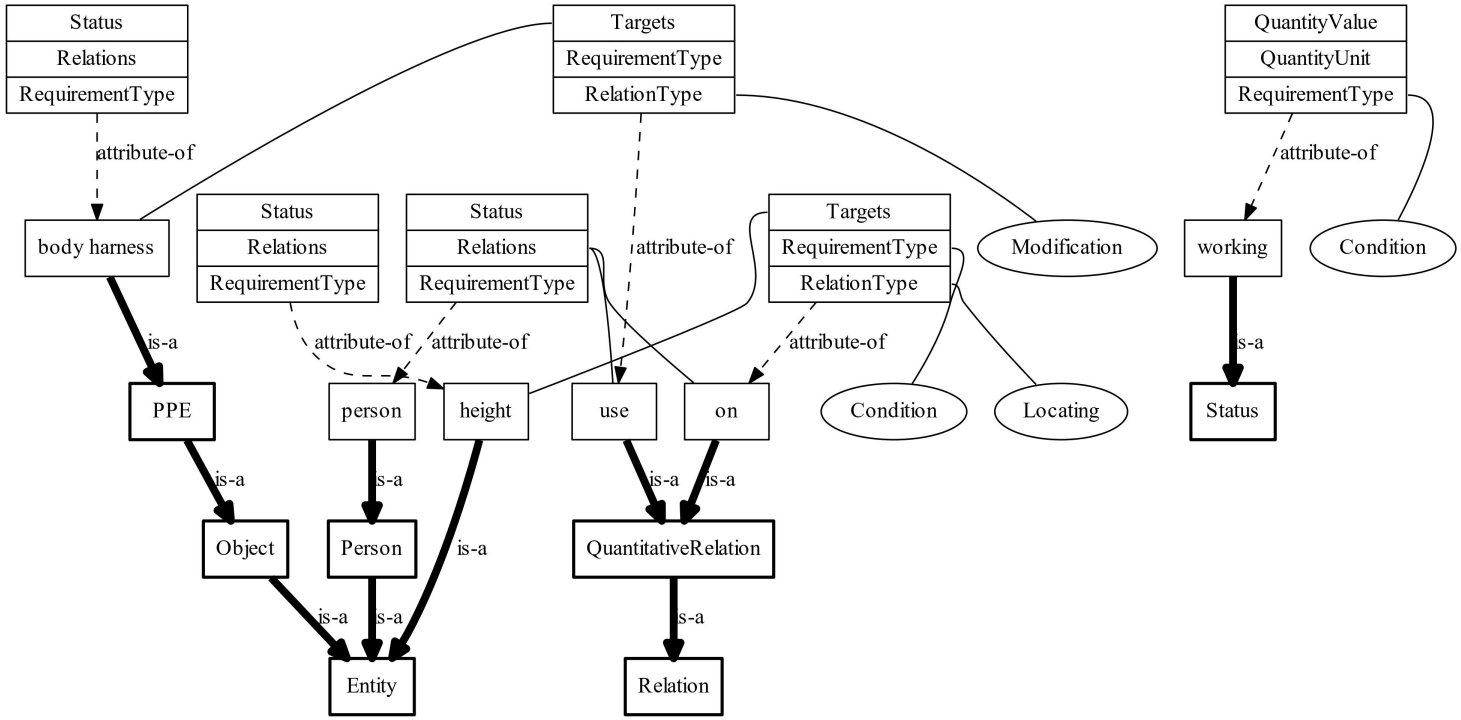



**Figure 6.16: Ontology modeling for regulatory rule (5): "Use body harness when working on height."**

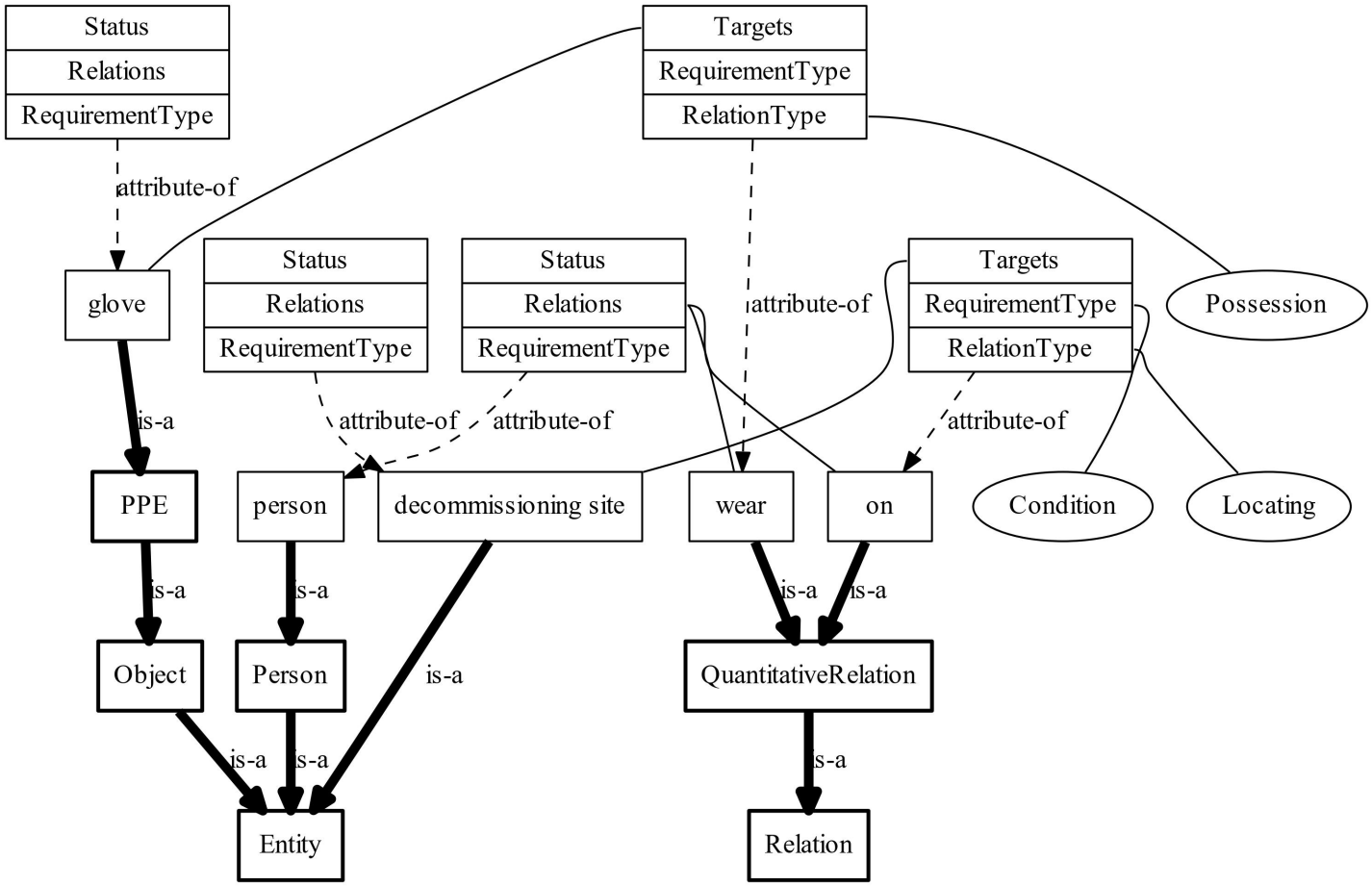



**Figure 6.17: Ontology modeling for regulatory rule (6): "Wear gloves on a decommissioning site."**

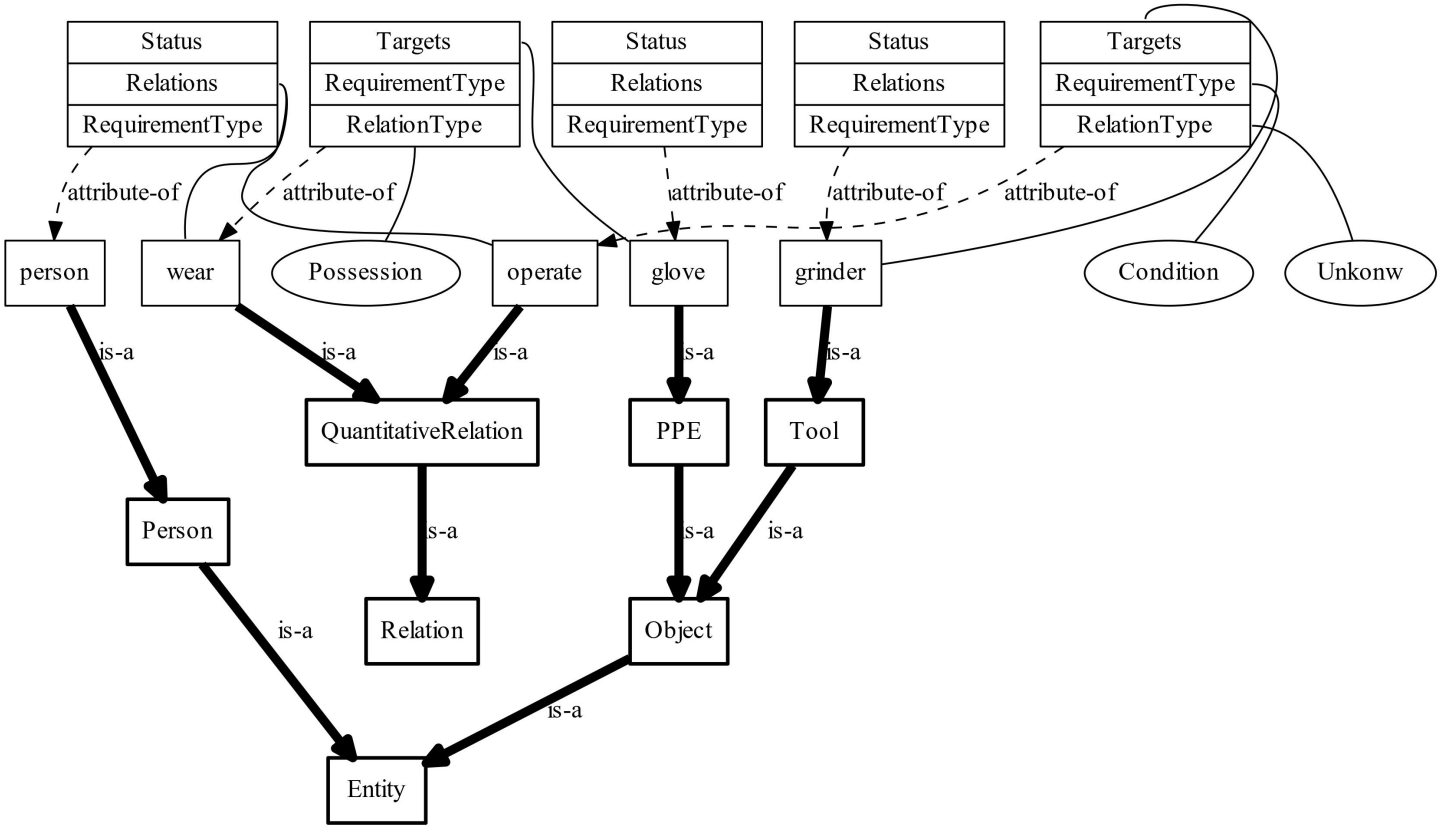



**Figure 6.18: Ontology modeling for regulatory rule (7): "Wear gloves when operating a grinder."**

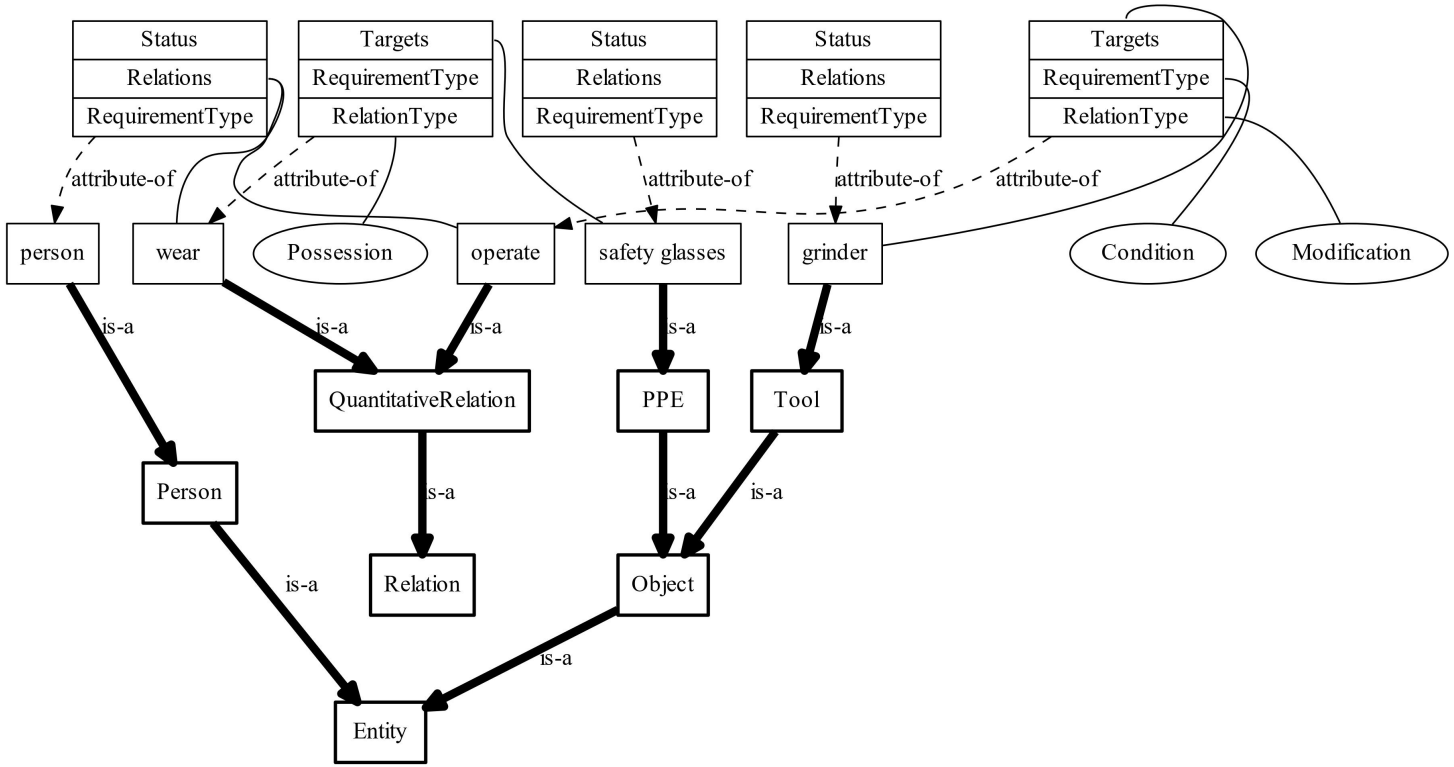



**Figure 6.19: Ontology modeling for regulatory rule (8): "Always use two hands when operating a grinder."**

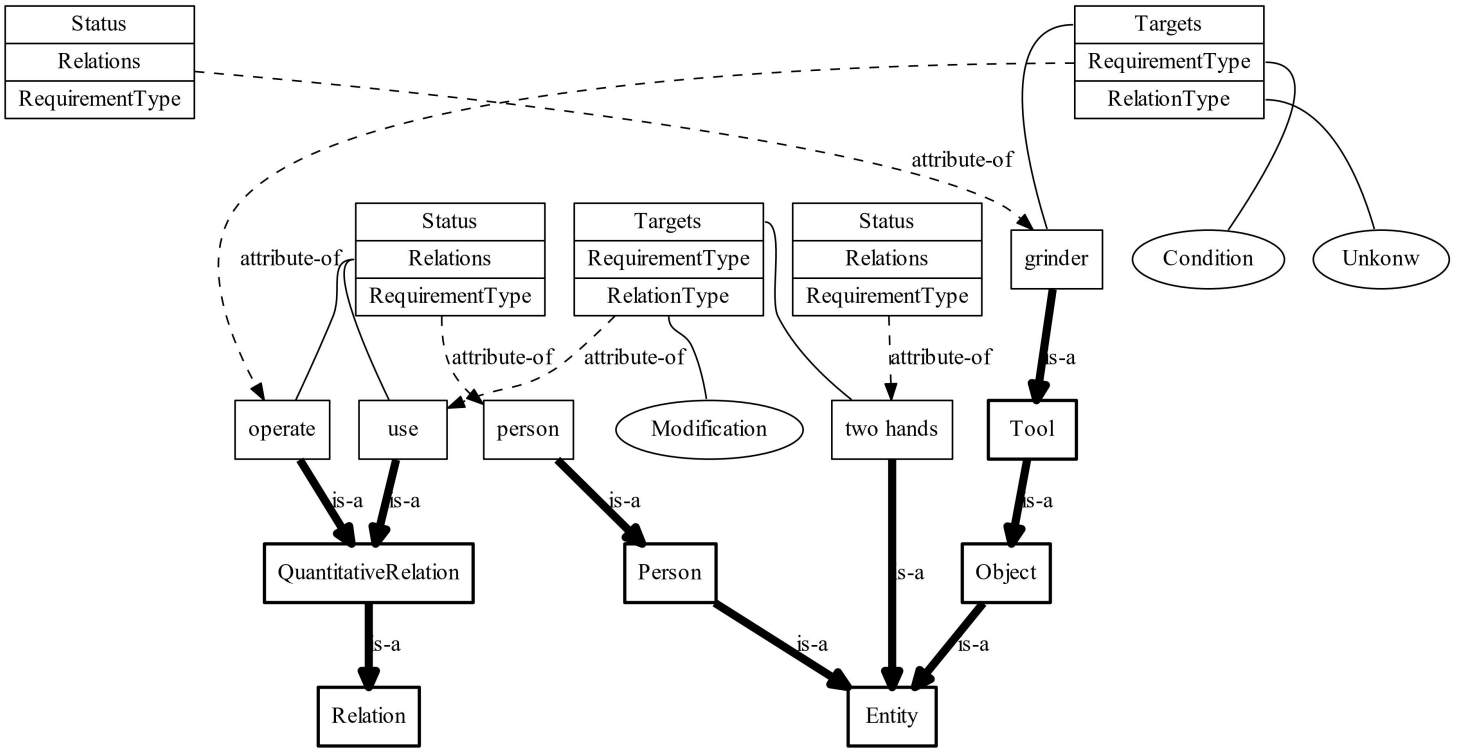



**Figure 6.20: Ontology modeling for regulatory rule (9): "Always use two hands when operating a grinder."**

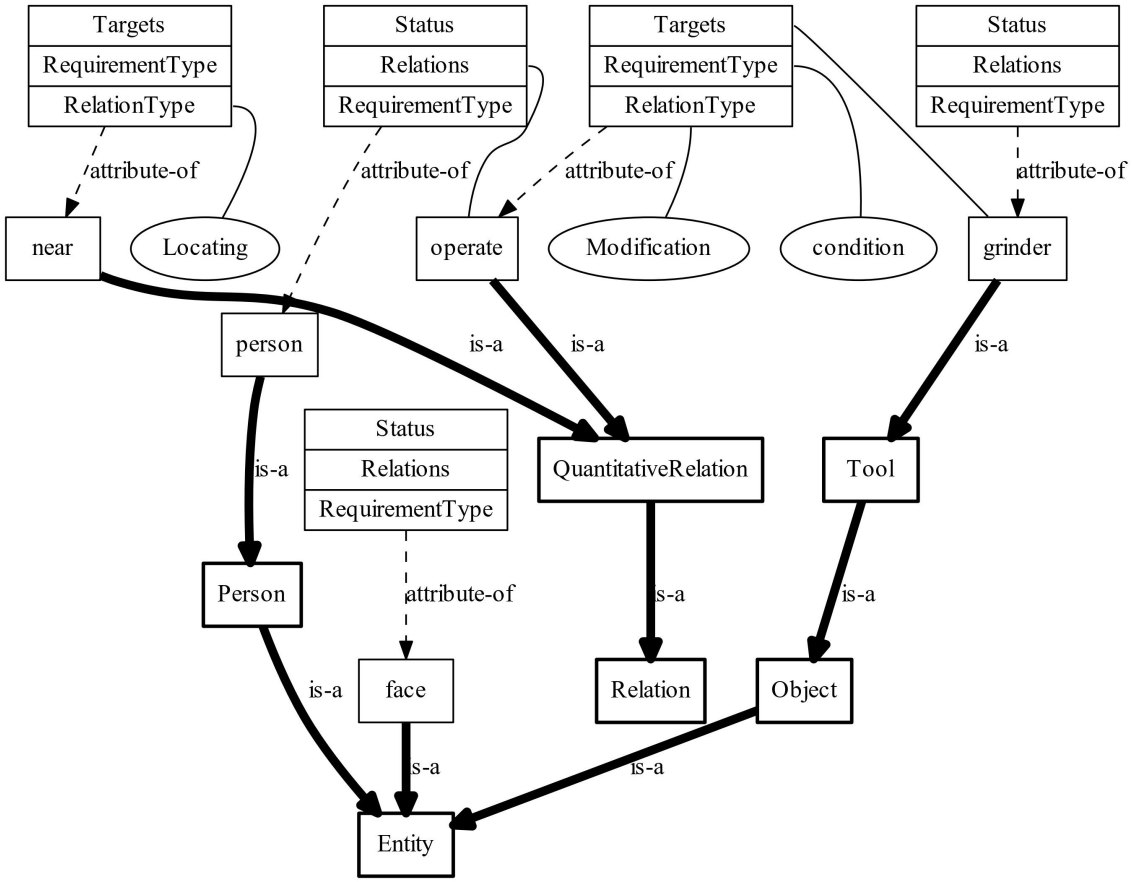



**Figure 6.21: Ontology modeling for regulatory rule (10): "Never operate a grinder near face."**

**Table 6.4: Logic representation of regulatory rules.**

| No. | Regulatory rules | Logic representation |
|---|---|---|
| 1 | Wear a hard hat on a construction site. | $((person, on, construction\ site)) \rightarrow () \vee ((person, wear, hard\ hat))$ |
| 2 | Wear a hard hat on a decommissioning site. | $((person, on, decommissioning\ site)) \rightarrow () \vee ((person, wear, hard\ hat))$ |
| 3 | Wear a dust mask on a construction site. | $((person, on, construction\ site)) \rightarrow () \vee (person, wear, dust\ mask))$ |
| 4 | Wear a full-face mask on a decommissioning site. | $((person, on, decommissioning\ site)) \rightarrow () \vee ((person, wear, full - face\ mask))$ |
| 5 | Use body harness when working on height. | $((person, on, height)) \rightarrow () \vee ((person, use, body\ harness))$ |
| 6 | Wear gloves on a decommissioning site. | $((person, on, decommissioning\ site)) \rightarrow () \vee ((person, wear, glove))$ |
| 7 | Wear gloves when operating a grinder. | $((person, operate, grinder)) \rightarrow () \vee ((person, wear, glove))$ |
| 8 | Wear a safety glasses when operating a grinder. | $((person, operate, grinder)) \rightarrow () \vee ((person, wear, safety\ glasses))$ |
| 9 | Always use two hands when operating a grinder. | $((person, operate, grinder)) \rightarrow () \vee ((person, use, two\ hands) \wedge (two\ hands, operate, grinder))$ |
| 10 | Never operate a grinder near face. | $((person, operate, grinder)) \rightarrow (\neg(person, operate, grinder) \wedge ((grinder, near, face))$ |

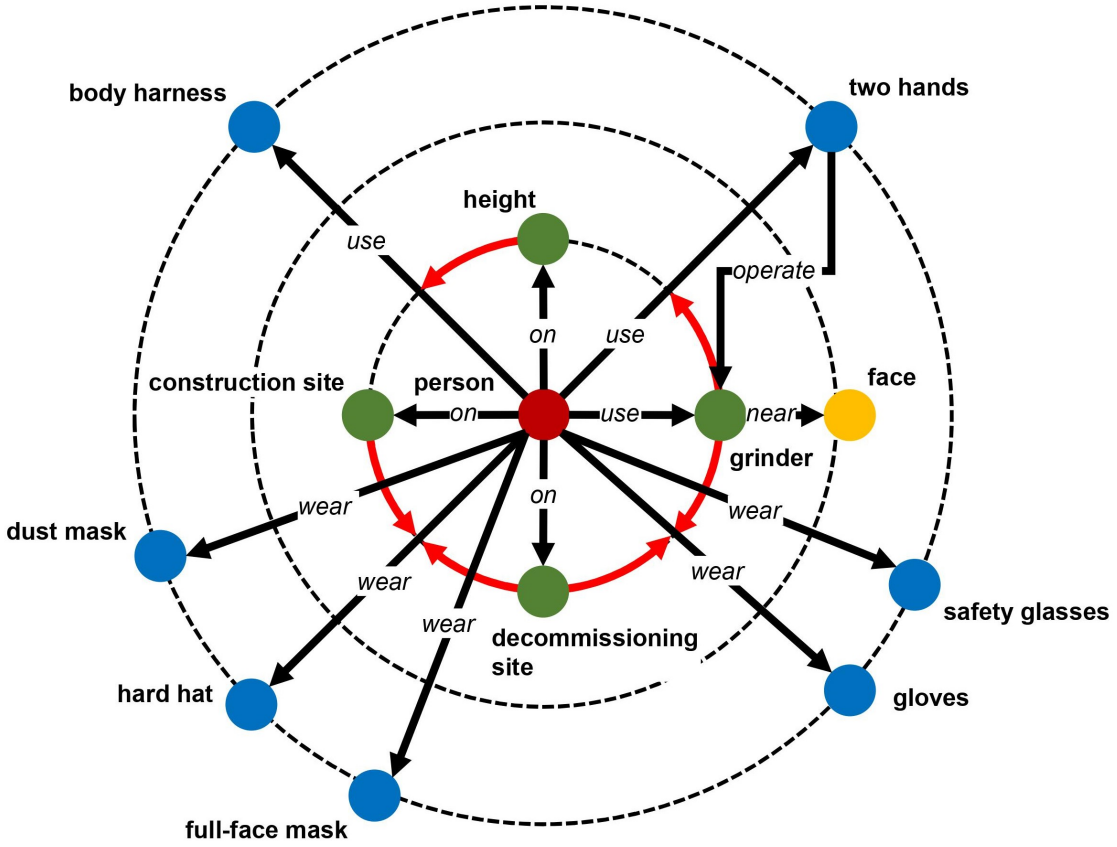



Figure 6.22: The hierarchical scene graph generated for regulatory information representation.

### 6.3.2 Image information representation & hazard identification results

Quantitative results of hazard identification results on the validation dataset are reported in Table 6.5.

(1) The precision and recall rate of individual-hard hat relationship identification declined only slightly from 3m to 5m, and even the hard hats were quite small in far-field (7m) images, the precision and recall rate remained higher than 92%.

(2) For individual-full-face mask relationship identification, the precision rate remained 100% while the recall rate declined only slightly as illumination distance increased.

(3) The precision rate of individual-dust mask relationship identification remained greater than 90%, but the recall rate decreased from 79.47% to 65.58% as the distance between the camera and workers increases since it is difficult to recognize the dust masks from the side view in far-field images.

(4) Individual-safety glasses relationship is considered to be the most challenging in the experiments because the safety glasses in the images of the validation dataset are transparent, and the recognition will be seriously affected by the level of the illumination. Even so, the precision and recall rate from 3m to 5m remained higher than 99% and 80%, respectively.

(5) For individual-body harness relationship identification, the precision rate in 3m and 5m are near 100% and remained higher than 83% in far-field (7m) images. But the recall rate is decreased to less than 80% since it is difficult to recognize the body harness from the side view in far-field images.

(6) The precision and recall rate of individual-glove relationship identification from 3m to 5m remained higher than 76% and 81%, respectively. But the precision and recall rate is decreased to less than 70% and 75% since wrists localization error of OpenPose in far-field (7m) images.

(7) For individual-grinder relationship identification, the precision, and recall in 3m are above 75% and 78%, respectively, but the performance is decreased in the cases of 5m and 7m since it is difficult to detect an individual-grinder relationship without their grinder visible and body occlusion can also lead to false negatives.

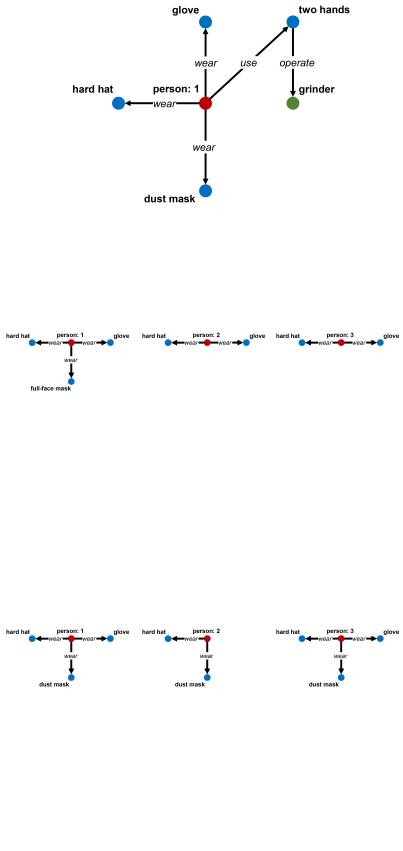
The overall precision and recall rates are 94.22% and 85.45%, respectively, which demonstrates the robustness of the proposed model in hazards identification at different distances.

Figure 6.23, Figure 6.24, and Figure 6.25 qualitatively illustrate the image information representation examples and hazards identification results on the validation dataset at the distance of 3m, 5m, and 7m, respectively. The proposed image information representation module detected the individuals and objects in the obtained images under different environmental conditions with different individual postures (Figure 6.23(a), Figure 6.24(a), Figure 6.25(a)) and visualizes intuitive scene graphs (Figure 6.23(b), Figure 6.24(b), Figure 6.25(b)). Furthermore, the proposed automated reasoning module identified the hazard from the obtained images (Figure 6.23(c), Figure 6.24(c), Figure 6.25(c)).

**Table 6.5: Image information representation results under different distance**

| Categories | Distance (m) | TP | FP | FN | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|
| | 3 | 2,606 | 35 | 58 | 98.67 | 97.82 |
| Hard hat | 5 | 2,946 | 45 | 119 | 98.50 | 96.12 |
| | 7 | 2,841 | 201 | 234 | 93.39 | 92.39 |
| | 3 | 524 | 0 | 2 | 100.00 | 99.62 |
| Full-face mask | 5 | 527 | 0 | 6 | 100.00 | 98.87 |
| | 7 | 514 | 0 | 5 | 100.00 | 99.04 |
| | 3 | 1,533 | 43 | 396 | 97.27 | 79.47 |
| Dust mask | 5 | 1,706 | 66 | 554 | 96.28 | 75.49 |
| | 7 | 1,608 | 163 | 844 | 90.80 | 65.58 |
| | 3 | 771 | 1 | 95 | 99.87 | 89.03 |
| Safety glasses | 5 | 510 | 3 | 124 | 99.42 | 80.44 |
| | 7 | 269 | 225 | 144 | 54.45 | 65.13 |
| | 3 | 866 | 1 | 68 | 99.88 | 92.72 |
| Body harness | 5 | 1,050 | 1 | 282 | 99.90 | 78.83 |
| | 7 | 1,110 | 213 | 275 | 83.90 | 80.14 |
| | 3 | 173 | 43 | 27 | 80.09 | 86.50 |
| Glove | 5 | 163 | 50 | 37 | 76.53 | 81.50 |
| | 7 | 149 | 64 | 51 | 69.95 | 74.50 |
| | 3 | 78 | 26 | 22 | 75.00 | 78.00 |
| Grinder | 5 | 71 | 30 | 29 | 70.30 | 71.00 |
| | 7 | 55 | 22 | 45 | 71.43 | 55.00 |
| Overall | | 20,070 | 1,232 | 3,417 | 94.22 | 85.45 |


(a) Image perception results




(b) Scene graph representation

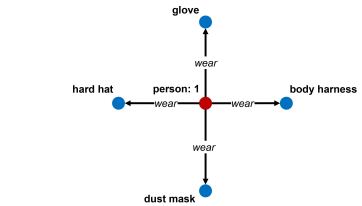
**person: 1**
*Hazardous: - not wearing safety glasses*

**person: 1**
*Safe*
**person: 2**
*Hazardous: - not wearing full-face mask*
**person: 3**
*Hazardous: - not wearing full-face mask*

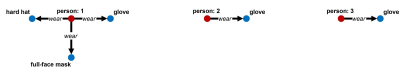
**person: 1**
*Safe*
**person: 2**
*Hazardous: - not wearing gloves*
**person: 3**
*Safe*

(c) Hazards identification results

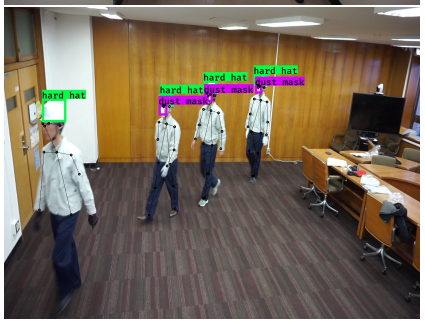
**Figure 6.23: Image information representation examples at the distance of 3m.**

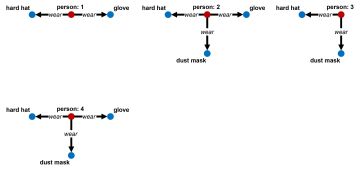




**person: 1**
*Safe*

**person: 1**
*Safe*
**person: 2**
*Hazardous: - not wearing hard hat*
*- not wearing full-face mask*
**person: 3**
*Hazardous: - not wearing hard hat*
*- not wearing full-face mask*



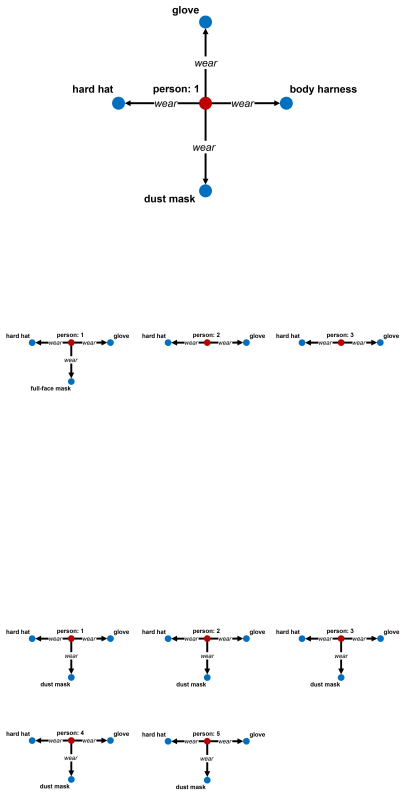



**person: 1**
*Hazardous: - not wearing dust mask*
**person: 2**
*Safe*
**person: 3**
*Hazardous: - not wearing gloves*
**person: 4**
*Safe*

(a) Image perception results (b) Scene graph representation (c) Hazards identification results

**Figure 6.24: Image information representation examples at the distance of 5m.**


(a) Image perception results




(b) Scene graph representation

**person: 1**
*Safe*

**person: 1**
*Safe*
**person: 2**
*Hazardous: - not wearing full-face mask*
**person: 3**
*Hazardous: - not wearing full-face mask*

**person: 1**
*Safe*
**person: 2**
*Safe*
**person: 3**
*Safe*
**person: 4**
*Safe*
**person: 5**
*Safe*

(c) Hazards identification results

**Figure 6.25: Image information representation examples at the distance of 7m.**

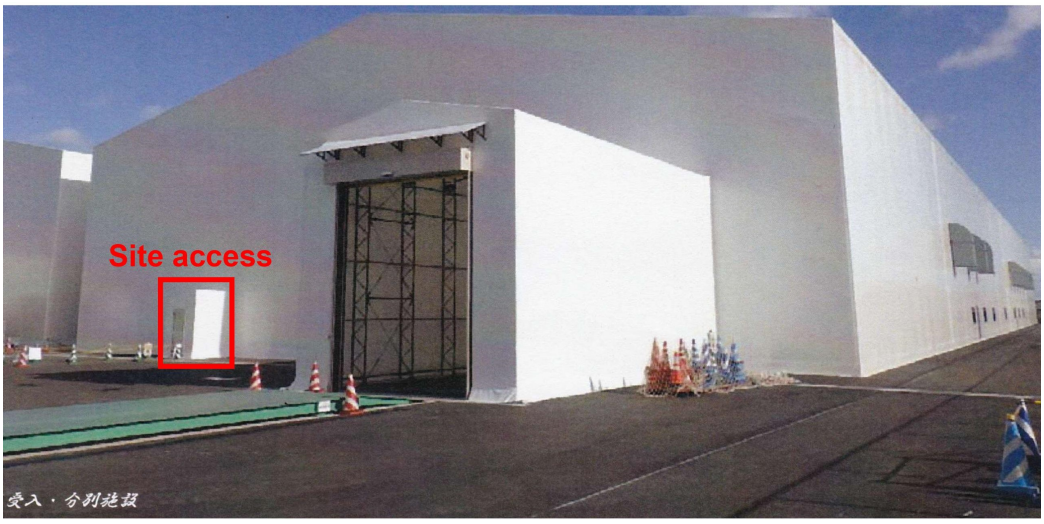



**Figure 6.26: Site access of the soil separation/storage facility in Futaba, Fukushima.**

### 6.3.3 On-site hazards identification results

To further validate the robustness of the proposed approach, on-site validation experiments were performed using the real monitoring data of the Interim Storage Facility (ISF) for soil separation and storage in Futaba, Fukushima [80]. ISF is carrying out soil, and waste derived of decontamination activities (soil and waste is removed, specified wastes ($> 100,000\ Bq/kg$) are stored) and the workers in ISF are required to wear appropriate PPE (hard hat, dust mask, and gloves) for radiation protection.

One surveillance camera is set in the site access of ISF (Figure 6.26) for security purposes, from which images were captured to perform the on-site validation experiments. The recorded data of the surveillance camera indicated the workers did not always precisely follow the safety regulations, and by using the developed on-site hazards identification system, these hazards were identified (examples shown in Figure 6.27).


(a)


(b)

**Figure 6.27: On-site validation results on the real monitoring data of ISF in Futaba, Fukushima.**

## 6.4 Computational efficiency analysis

To meet the industrial requirements of real-time processing, computational efficiency analysis experiments were also performed. Computational efficiency analysis results are presented in Table 6.6. Based on the fast processing speed of YOLOv3 and the metric of using a light-weight architecture for the feather extractor of OpenPose, the inference time of the proposed approach in this work outperforms other state-of-the-art approaches while preserving high-quality results. It was able to run at about 7.95 FPS in a machine with a GeForce GTX 1080 Max-Q with 8GB of GDDR5X memory and 2560 CUDA cores, and it indicates that the proposed approach is more effective compared to the Faster R-CNN approach adopted by Fang et al. [21] and SSD-RPA approach proposed by Wu et al. [22].

**Table 6.6: Computational efficiency analysis results.**

| Approach | Input size | FPS |
|---|---|---|
| Faster R-CNN-based | $300 \times 500$ | 4.88 |
| SSD-RPA-based | $304 \times 304$ | 3.22 |
| This work | $416 \times 416$ | 7.95 |

# Chapter 7

# Concluding Remarks

## 7.1 Principal conclusions

Construction and decommissioning sites are one of the most perilous environments where many potential hazards may occur. To avoid the occurrence of occupational hazards, workers are required to follow the on-site regulatory rules. However, compliance of regulatory rules is not strictly enforced among workers due to all kinds of reasons. Conventional on-site occupational safety monitoring is not sufficient to ensure the safety of workers due to human factors and human errors. Consequently, an automated on-site occupational hazards identification system is urgently needed.

The goal of this thesis is to propose an regulatory-image inference model to process images and regulatory rules sentence for on-site occupational hazards identification and develop a robust and efficient real-time automated system to meet industry requirements. The main contents and results of this work are summarized as follows.

In chapter 1, the essential requirements and difficulties regarding automated on-site occupational hazards identification are stated. Additionally, the state-of-the-art works made attempts to on-site occupational hazards identification are reviewed. Both merits and limitations of these works are discussed which have figured out four problems to be addressed: (a) the needs of automated regulatory information extraction and representation, (b) the

model to perform multi-hazard identification task, (c) the solution to avoid impacts from viewpoint changes of the on-site surveillance cameras and different individual postures, and (d) the requirements of real-time processing and reliability for industrial applications.

As a point of departure, Chapter 2 reviews four candidate structures to represent information extracted from regulatory rules and on-site images. Taking advantage of scene graph structure, the framework of an regulatory-image inference model is proposed to drive the development of this work, which is constructed by (a) regulatory information representation module, (b) image information representation module, and (c) automated reasoning module for on-site hazards identification.

In chapter 3, a regulatory information extraction approach is proposed based on NLP techniques and ontology modeling. Subsequently, to address the limitation of the conventional scene graph in representing complex relationships of types of requirements, an original hierarchical scene graph structure is proposed for regulatory information representation. Based on the proposed approach, a novel automated regulatory rules processing system has been developed.

Chapter 4 presents the proposed image information representation approach. It adopts YOLOv3 and OpenPose for detecting objects and individuals, respectively. Meanwhile, geometric relationship analysis is originally implemented to combine deep learning-based object detection and individual detection model with interpretable explanations and outputs. To provide prior knowledge and reduce computational complexity, a method based on minimum weighted matching in bipartite graphs is proposed to associate detected objects with individuals. Present work in this thesis is able to cover four types of individual-object relationship processing: (a) individual-head protection PPE, (b) individual-grinder, (c) individual-glove, and (d) individual-body harness. (a), (b), and (d) is identified based on key length analysis of associated objects and individuals. (c) is realized based on a color-based skin detection algorithm. Based on the proposed approach, the scene information of each obtained on-site image is represented in a scene graph for further hazards identification.

In chapter 5, an automated reasoning approach is proposed. It integrates the proposed regulatory and image information and deploys graph structure analysis for hazards identification. Additionally, a novel system is developed based on the proposed approach for real-time occupational hazards identification.

Chapter 6 describes the experiments to validate the robustness and efficiency of the proposed approach. The validity of the developed automated regulatory rules processing system was demonstrated in the experiments of processing ten selected construction/decommissioning regulatory rules. Subsequently, 13,893 images of hard hats, dust masks, full-face masks, safety masks, body harness, and grinder were collected to train the object detection model. Furthermore, a validation dataset was created considering the impacts of illumination, viewpoint changes of cameras, and different individual postures under various distances (3m, 5m, 7m). The performance of the developed on-site occupational hazards identification system was experimentally evaluated on the validation dataset. The validating results indicate that the developed system was capable of identifying the hazards with high precision (94.22%) and recall rate (85.45%) while ensuring real-time performance (7.95 FPS on average).

In particular, the following *originalities and achievements* are thought to be the contribution of the present work.

(1) As the framework of this work, the regulatory-image inference model drives the pipeline of this work from regulatory rules/on-site image processing to on-site hazards identification. To the author's best knowledge, it is the first model to address both regulatory and image information in AEC domains.

(2) The regulatory information extraction approach originally deploys NLP-based grammatical structure analysis with the ontology concept to extract key information from regulatory rules in AEC domains. Regulatory rules relating to the proper individual behaviors or the safe operation of equipment can be well addressed.

(3) The hierarchical scene graph presents a novel structure extending the

conventional scene graph to represent conditional and prohibition relationships in regulatory information.

(4) The image information representation approach originally deploys geometric relationship analysis to perform the combination of object detection and individual detection model with interpretable explanations and outputs. This is meaningful for the the feasibility and usability on industrial applications. Specifically, it provides a solution for multi-hazard identification regarding viewpoint changes of on-site cameras and different individual postures of on-site workers and some complex scenes with multiple individual can also be represented in a scene graph for further relationship analysis.

(5) Robustness and efficiency of the developed on-site occupational hazards identification system were experimentally evaluated for real-time on-site processing.

(6) Compared with on-site current safety monitoring system carried out by the manully effects, the development of this work provides an automated solution that helps facilitate the task of safety monitoring. Meanwhile, the developed system in this work can be easily deployed on construction or decommissioning sites and can be used by users without a technical background.

## 7.2 Perspectives

Based on the novel proposal of the informatica interface model of image (on-site) and sentence (regulatory rules), the "Eye of Technology" for on-site occupational safety monitoring was preliminary realized as an improvement of current "i-Construction" framework. It has excellent prospects and extensive application possibilities.

Further extensions of this work are to be investigated in the following directions considering the improvements for regulatory and image information representation.

(1) It is expected in the proposed regulatory information representation approach for its contribution to the advancement of automated regulatory information processing in different languages other than English and complex sentence structures. Recent advanced NLP models (e.g., BERT [81], XLM [82], MASS [83], XLNet [84]) are suggested to be considered to further realize this scope.

(2) The experimental results demonstrated the performance of the proposed approach under 3m, 5m, 7m, while indeitification performance on far-field situation (>7m) is also believed to be covered by the implementation of high-resolution surveillance cameras. The ROI of individuals can be first extracted based on the keypoints detected by OpenPose to reduce computational complexity to ensure real-time processing.

(3) Although the proposed image information representation approach responses proved successful in individual-object relationship analysis, the performance may still be affected by object invisibility and individual occlusion. For the application of the proposed image information representation approach to industrial implementation, further improvements in robustness are desirable. 3D individual detection and object detection are suggested to be implemented. Recently, many vision-based 3D multi-person pose estimation approaches [85–87] have shown noticeable performance, while 3D object detection approaches [88–90] are mainly focusing on and applied the 3D car detection (for autonomous driving) and the training dataset for detecting 3D position of common objects is rare. To this end, a new training dataset for object detection and 3D orientation estimation, which provides accurate 3D bounding boxes for construction object such as hard hats, masks and grinders, needs to be created. Based on these prior knowledges, a new 3D individual-object interaction model can be proposed and implemented for robust image information representation.

(4) Functional extensions for on-site application possibilities are also suggested to be explored. There are still many on-site safety monitoring

requirements needs urgent solutions, e.g., proper use identifications for multiple PPE in Fukushima Daiichi NPS and caugut-in accident avoidance. Additionally, the application possibility on accident evacuation is also expected to be investigated based on the deployment of smart glasses (e.g., Vuzix M-Series [91]) and IoT platform (e.g., Microsoft Azure IoT [92]).

# Bibliography


[1] U. B. of Labor Statistics, "Construction: Naics 23." https://www.bls.gov/iag/tgs/iag23.htm. 2020.

[2] L. The Japanese Ministry of Health and Welfare, "The 13th occupational safety & health program." https://www.mhlw.go.jp/content/11200000/000341159.pdf. 2018.

[3] O. S. . H. Administration, "Worker safety series: construction." https://www.osha.gov/Publications/OSHA3252/3252.html. Accessed: 2020-05-20.

[4] S. Konda, H. M. Tiesman, and A. A. Reichard, "Fatal traumatic brain injuries in the construction industry, 2003- 2010," *American journal of industrial medicine*, vol. 59, no. 3, pp. 212–220, 2016.

[5] O. S. . H. Administration, "Head protection.." https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.100. 2012.

[6] N. I. for Occupational Safety and Health, "Eye safety." https://www.cdc.gov/niosh/topics/eye/. 2013.

[7] A. L. Dannenberg, L. M. Parver, R. J. Brechner, and L. Khoo, "Penetrating eye injuries in the workplace: the national eye trauma system registry," *Archives of Ophthalmology*, vol. 110, no. 6, pp. 843–848, 1992.

[8] W. D. Government of Western Australia, Department of Commerce, "Guide to using dust masks in construction work."

https://www.commerce.wa.gov.au/sites/default/files/atoms/files/guide_to_using_dust_mask.pdf. 2019.

[9] O. S. . H. Administration, "Eye and face protection.." https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.133. 2016.

[10] O. S. . H. Administration, "Angle grinder safety.." https://www.osha.gov/sites/default/files/2018-12/fy15_sh-27664-sh5_Toolbox_Angle_Grinder.pdf. 2018.

[11] Y. Amano, "The fukushima daiichi accident: Report by the director general," *Vienna: International Atomic Energy Agency*, 2015.

[12] *Safety Assessment for Decommissioning*. No. 77 in Safety Reports Series, Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY, 2013.

[13] "Efforts to improve working environment and reduce radiation exposure at fukushima daiichi nuclear power station," tech. rep., Tokyo Electric Power Company, 2016.

[14] K. Mori, S. Tateishi, and K. Hiraoka, "Health issues of workers engaged in operations related to the accident at the fukushima daiichi nuclear power plant," in *Psychosocial Factors at Work in the Asia Pacific*, pp. 307–324, Springer, 2016.

[15] T. The Japanese Ministry of Land, Infrastructure and Tourism, "i-construction." https://www.mlit.go.jp/tec/i-construction/index.html. Accessed: 2020-05-21.

[16] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M. J. Khazaee, A. M. Costin, M. Helmus, and J. Teizer, "Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites," *Automation in construction*, vol. 36, pp. 38–52, 2013.

[17] S. Barro-Torres, T. M. Fernández-Caramés, H. J. Pérez-Iglesias, and C. J. Escudero, "Real-time personal protective equipment monitoring system," *Computer Communications*, vol. 36, no. 1, pp. 42–50, 2012.

[18] S. Dong, Q. He, H. Li, and Q. Yin, "Automated ppe misuse identification and assessment for safety performance enhancement," in *ICCREM 2015*, pp. 204–214, 2015.

[19] K. Shrestha, P. P. Shrestha, D. Bajracharya, and E. A. Yfantis, "Hardhat detection for construction safety visualization," *Journal of Construction Engineering*, vol. 2015, 2015.

[20] M.-W. Park, N. Elsafty, and Z. Zhu, "Hardhat-wearing detection for enhancing on-site safety of construction workers," *Journal of Construction Engineering and Management*, vol. 141, no. 9, p. 04015024, 2015.

[21] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T. M. Rose, and W. An, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1–9, 2018.

[22] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," *Automation in Construction*, vol. 106, p. 102894, 2019.

[23] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in Construction*, vol. 112, p. 103085, 2020.

[24] R. Xiong, Y. Song, H. Li, and Y. Wang, "Onsite video mining for construction hazards identification with visual relationships," *Advanced Engineering Informatics*, vol. 42, p. 100966, 2019.

[25] J. Zhang and N. M. El-Gohary, "Automated information transformation for automated regulatory compliance checking in construction," *Journal of Computing in Civil Engineering*, vol. 29, no. 4, p. B4015001, 2015.

[26] J. Zhang and N. M. El-Gohary, "Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking," *Journal of Computing in Civil Engineering*, vol. 30, no. 2, p. 04015014, 2016.

[27] J. Zhang and N. M. El-Gohary, "Integrating semantic nlp and logic reasoning into a unified system for fully-automated code checking," *Automation in construction*, vol. 73, pp. 45–57, 2017.

[28] H. Chen and X. Luo, "An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing," *Advanced Engineering Informatics*, vol. 42, p. 100959, 2019.

[29] B. Zhong, X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, and W. Fang, "Deep learning-based extraction of construction procedural constraints from construction regulations," *Advanced Engineering Informatics*, vol. 43, p. 101003, 2020.

[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[32] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[33] C. McCormick, "Word2vec tutorial - the skip-gram model." `http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/`. Accessed: 2020-05-21.

[34] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceed-*

*ings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

[35] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1261–1270, 2017.

[36] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419, 2017.

[37] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE conference on computer vision and Pattern recognition*, pp. 3076–3086, 2017.

[38] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, 2018.

[39] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1969–1978, 2019.

[40] P. Boken and G. Callaghan, "Confronting the challenges of manual journal entries," *Protiviti, Alexandria, VA*, pp. 1–4, 2009.

[41] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence?: problems of tokenisation," 1994.

[42] M. Aronoff and K. Fudeman, *What is morphology?*, vol. 8. John Wiley & Sons, 2011.

[43] C. Fautsch and J. Savoy, "Algorithmic stemmers or morphological analysis? an evaluation," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 8, pp. 1616–1624, 2009.

[44] "spacy: Industrial-strength nlp." `https://spacy.io/`. Accessed: 2020-04-27.

[45] L. Tesnière, "Eléments de syntaxe structurale," 1959.

[46] J. E. Miller and J. Miller, *A critical introduction to syntax*. A&C Black, 2011.

[47] J. J. Robinson, "Dependency structures and transformational rules," *Language*, pp. 259–285, 1970.

[48] R. Debusmann, "An introduction to dependency grammar," *Hausarbeit fur das Hauptseminar Dependenzgrammatik SoSe*, vol. 99, pp. 1–16, 2000.

[49] T. R. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–221, 1993.

[50] C. Feilmayr and W. Wöß, "An analysis of ontologies and their success factors for application to business," *Data & Knowledge Engineering*, vol. 101, pp. 1–23, 2016.

[51] S.-K. Lee, K.-R. Kim, and J.-H. Yu, "Bim and ontology-based approach for building cost estimation," *Automation in construction*, vol. 41, pp. 96–105, 2014.

[52] L. Ding, B. Zhong, S. Wu, and H. Luo, "Construction risk knowledge management in bim using ontology and semantic web technology," *Safety science*, vol. 87, pp. 202–213, 2016.

[53] P. Zhou and N. El-Gohary, "Ontology-based automated information extraction from building energy conservation codes," *Automation in Construction*, vol. 74, pp. 103–117, 2017.

[54] K. Liu and N. El-Gohary, "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports," *Automation in Construction*, vol. 81, pp. 313–327, 2017.

[55] D. D. Maynard, D. K. Bontcheva, and D. H. Cunningham, "Automatic language-independent induction of gazetteer lists," 2004.

[56] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[57] A. Hagberg, D. Schult, and P. Swart, "Networkx: Python software for the analysis of networks," *Mathematical Modeling and Analysis, Los Alamos National Laboratory*, 2005.

[58] J. Ellson, E. Gansner, L. Koutsofios, S. C. North, and G. Woodhull, "Graphviz—open source graph drawing tools," in *International Symposium on Graph Drawing*, pp. 483–484, Springer, 2001.

[59] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[60] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[62] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[63] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[64] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[66] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[67] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[68] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[69] "Kinect for windows." `https://developer.microsoft.com/en-us/windows/kinect/`. Accessed: 2020-07-22.

[70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[71] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334–2343, 2017.

[72] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[74] S.-M. Hsieh, C.-C. Hsu, and L.-F. Hsu, "Efficient method to perform isomorphism testing of labeled graphs," in *International Conference on Computational Science and Its Applications*, pp. 422–431, Springer, 2006.

[75] G. Bradski and A. Kaehler, "Opencv," *Dr. Dobb's journal of software tools*, vol. 3, 2000.

[76] F. Lundh, "An introduction to tkinter," *URL: www. pythonware. com/library/tkinter/introduction/index. htm*, 1999.

[77] L. Tzutalin, "Git code (2015)."

[78] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[79] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[80] "Interim storage facility." `http://josen.env.go.jp/en/storage/`. Accessed: 2020-05-19.

[81] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[82] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[83] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation," *arXiv preprint arXiv:1905.02450*, 2019.

[84] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, pp. 5754–5764, 2019.

[85] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2500–2509, 2017.

[86] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10133–10142, 2019.

[87] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera," *arXiv preprint arXiv:1907.00837*, 2019.

[88] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," *arXiv preprint arXiv:1912.13192*, 2019.

[89] J. Lehner, A. Mitterecker, T. Adler, M. Hofmarcher, B. Nessler, and S. Hochreiter, "Patch refinement–localized 3d object detection," *arXiv preprint arXiv:1910.04093*, 2019.

[90] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7345–7353, 2019.

[91] "Vuzix m-series." `https://www.vuzix.com/products/m-series`. Accessed: 2020-08-3.

[92] "Microsoft azure iot." https://azure.microsoft.com/en-us/overview/iot/. Accessed: 2020-08-3.