A Dynamic Graph-based Time Series Analysis Framework for On-site Occupational Hazards Identification

Shi Chen[1], Feiyan Dong[1] and Kazuyuki Demachi[1]

[1]Department of Nuclear Engineering and Management, School of Engineering, The University of Tokyo, Japan

shichen@g.ecc.u-tokyo.ac.jp, dongfeiyan@g.ecc.u-tokyo.ac.jp, demachi@n.t.u-tokyo.ac.jp

Abstract -

Different factors combined invariably cause construction fatalities at any time, most of which could be avoided if workers followed the on-site regulatory rules. However, compliance of regulatory rules is not strictly enforced among workers due to all kinds of reasons, even after prior education and training. To address the difficulties of on-site safety management, this paper proposes a graph-based time series analysis framework to dynamically integrate visual and linguistic information for on-site occupational hazard identification. A vision-based scene information understanding approach is introduced to process on-site images via a combination of deep learning-based object detection and individual detection, together with a novel dynamic graph structure to represent time-series information for integrated reasoning of hazards identification. As a case study, the hazards of grinder operation were successfully identified in the experiments with high accuracy.

Keywords -

Construction Safety; Occupational hazards identification; Deep Learning; Graph; Time series analysis

1 Introduction

The construction industry is one of the fields with the highest number of occupational accidents. According to the United States' Bureau of Labor Statistics (BLS), the number of construction fatalities in the U.S. has increased from 924 to 1,066 between 2015 and 2019 [1]. Similarly, in Japan, there were 1,522 construction fatalities between the period 2015-2019 [2].

As the construction is extremely vulnerable to the interference of various subjective and objective factors, once any sudden problem occurs it will pose a potential threat to the life and property safety of the on-site construction workers. Grinding is a commonly used operation on construction sites to produce smooth surfaces, and can also be used to fabricate workpieces such as smoothing welds and performing finishing operations on workpiece surfaces. When using a grinder, the grinding disc generates high rotational speeds, which can be hazardous if the operator lacks expertise and the operation does not follow on-site regulatory rules. Eye injuries would be a possible serious consequence. 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 [3]. 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 [4]. 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 [5]. Additionally, fine dust and particles, gases and vapors can be produced when using a grinder. Silica dust from bricks can cause lung and airway diseases such as emphysema, bronchitis, silicosis, and may increase the risk of cancer. Personal protective equipment (PPE), such as respirators or dust masks, are used to controls these hazards [6]. On the other hand, 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 [7].

However, the construction workers do not precisely follow the on-site safety regulations due to various reasons, even after prior education and training. Therefore, the development of an automated on-site occupational hazards identification system is needed to address the increasing importance of safety management, which is capable of automatically carrying out dynamic identification of occupational hazards and effectively preventing various accidents.

To this end, this paper proposes a graph-based time series analysis framework to dynamically integrate visual and linguistic information for on-site occupational hazards identification. (1) A vision-based scene information understanding approach is introduced to process on-site images via a combination of deep learning-based object detection and individual detection. (2) An automated reasoning is developed to encode regulatory information into graph structure and perform occupational hazards identifi-

cation based on graph structure analysis between extracted scene information and regulatory information. The proposed model was able to identify the hazards of grinder operation with high accuracy in the experiments.

2 Related works

Deep learning-based object detection algorithms have shown remarkable performance on most visual tasks in the architecture, engineering, and construction (AEC) industry, and there has been a significant amount of research on vision-based automatic occupational hazards identification approaches using object detection [8, 9, 10]. Fang et al. [8] proposed an object detection-based method using Faster R-CNN to automatically detect construction workers' NHU. A total of 81,000 image frames were collected from various construction sites as a training dataset and the bounding boxes that surround workers in the images were annotated as the ground truth to train the model. Wu et al. [9] deployed Single Shot Multibox Detector (SSD) with presented reverse progressive attention (RPA) for NHU identification. A benchmark dataset GDUT-HWD was created by downloading Internet images retrieved by search engines to train the SSD-RPA model. In contrast to [8], only the head regions of workers were annotated as the ground truth. Nath et al. [10] introduced and tested models built on YOLOv3 architecture to verify PPE (hard hat and vest) compliance of workers. Three approaches were verified concerning different classifiers (e.g., decision tree, VGG-16, ResNet-50, Xception, or Bayesian).

Recently, deep learning-based pose estimation algorithms have achieved impressive results in unconstrained environments, showing the potential for worker detection in complex on-site environments. Compared to object detection-based approaches, human skeletons provide more fine-grained information about a person (e.g., location and visibility), especially in the case of occlusion. Considering such benefits, several efforts are also exploring the integration of object recognition and pose estimation for occupational hazards identification [11, 12]. Chen et al. [11] introduced a vision-based approach to detect the proper use of multi-class radiation PPE in nuclear power plants via a combination of deep learning-based object detection and pose estimation using Euclidean distance between bounding boxes of detected PPE and the neck geometric relationships analysis. Xiong et al. [12] presented an extensible pose-guided anchoring framework aimed at multi-class PPE compliance detection. A pose estimator was deployed to detect individuals and provide joint-level anchors for guiding the localization of different PPE items.

Vision-based approaches have been widely used to automatically identify occupational hazards as introduced above. However, as regulation rules may change at any time in practical engineering, current vision-based approaches will be significantly reduced in practicality due to their inability to adapt to adjustments in practice. A unified model integrating visual and linguistic information would enable the automatic and effective identification of hazards in compliance with regulatory rules even when changes are made to them. Several explorations to this end have already been carried out [13, 14, 15]. Xiong et al. [13] 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 crucial hazards, i.e., failing to wear a hard hat and walking beneath the cane, were successfully identified. Fang et al. [14] integrated computer vision algorithms with ontology models to develop a knowledge graph that consists of an ontological model for hazards, knowledge extraction, and knowledge inference for hazard identification, which can automatically identify falls from heights hazards in varying contexts from images. As a previous exploration of this work, we provided a novel solution to identify improper use of PPE by the combination of deep learning-based object detection and individual detection using geometric relationships analysis and presented a hierarchical scene graph structure that enables the conditional reasoning for automated hazards identification to address different requirements in each zone of construction sites [15].

3 Methodology

In this section, the proposed dynamic graph-based framework for automated occupational hazards identification is described in detail.

3.1 Scene information extraction

We propose a novel scene understanding approach employing scene graph as the basic notion of information representation structure to extract visual information from images, the base framework of which has been presented in our previous work [15]. An entity extractor for processing images obtained from on-site surveillance cameras is developed. For each image, individual(s) are detected, together with their body joint positions, using OpenPose [16]. Meanwhile, objects (e.g., PPE, tools) are recognized and localized by training an object detection model based on YOLOv4 [17].

We first associate the detected objects with the detected individuals, with the aim of providing a prior knowledge for individual-object relationship analysis and reducing computational complexity. A weighted bipartite graph is constructed to represent the detected entities and we perform individual-object association as a minimum weighted matching in bipartite graphs. Subsequently, we

analyze the individual-object relationship on the associated individual-object pairs M, which address two types of objects in this paper:

- Head protection PPE (hard hats, safety glasses, and dust masks).
- Grinder

3.1.1 Head protection PPE

For each associated individual and head protection PPE $\{i^*, j^*\} \in M$ their relationship is identified by by measuring a distance. We take advantage of the Euclidean distance among detected neck keypoint (body parts 1 in Figure 1) and hip keypoints (body parts 8 and 11 in Figure 1) of i^* is as a dynamic reference threshold, which will keep changing synchronously when the distance between the individual and the camera changes:

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

where γ 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, γ is set to 0.8, 0.7, 0.6, 0.6, respectively.

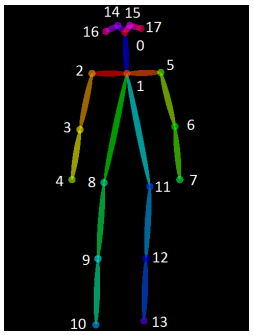


Figure 1. Output format of OpenPose.

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 1) 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:

$$c_{i^* \leftrightarrow j^*} = \begin{cases} \text{“}wear\text{”} & ,if\ d_h(i^*, j^*) < \beta_{i^* \leftrightarrow j^*} \\ N/A & ,otherwise \end{cases} \quad (2)$$

where

$$d_h(i^*, j^*) = \sqrt{\left(x_{i^*}^{(1)} - x_{j^*}\right)^2 + \left(y_{i^*}^{(1)} - y_{j^*}\right)^2} \quad (3)$$

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., $(person, wear, hard\ hat)$ or not.

3.1.2 Grinder

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

(1) “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 1) of i^* to the position (x_{g^*}, y_{g^*}) of g^* is calculated (Figure 2):

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

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

$$h_{i^* \leftrightarrow g^*} = \begin{cases} 1 & ,if\ d_l(i^*, g^*) < \beta_{i^* \leftrightarrow g^*}\ or\ d_r(i^*, g^*) < \beta_{i^* \leftrightarrow g^*} \\ 0 & ,otherwise \end{cases} \quad (5)$$

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

$$\beta_{i^* \leftrightarrow g^*} = \max(w_{g^*}, h_{g^*}) \quad (6)$$

Figure 2. Relationship identification strategies to address the rule “Always use two hands when operating a grinder”.

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:

$$
\begin{aligned}
&if\ d_{l\leftrightarrow r}(i^*) < \beta_{i^*\leftrightarrow g^*}:\\
&c_{i^*\leftrightarrow h^*_{i^*}}, h^*_{i^*}, c_{h^*_{i^*}\leftrightarrow g^*} = \text{“}use\text{”}, \text{“}two\ hands\text{”}, \text{“}operate\text{”}\\
&otherwise:\\
&c_{i^*\leftrightarrow h^*_{i^*}}, h^*_{i^*}, c_{h^*_{i^*}\leftrightarrow g^*} = \text{“}use\text{”}, \text{“}single\ hand\text{”}, \text{“}operate\text{”}
\end{aligned} \tag{7}
$$

where $h^*_{i^*}$ indicates the hands status (e.g., two hands, single hand) when operating a grinder while $c_{i^*\leftrightarrow h^*_{i^*}}$ and $c_{h^*_{i^*}\leftrightarrow g^*}$ create 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 2)

(2) "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 1) of i^* to the position (x_{g^*}, y_{g^*}) of the bounding box of g^* is calculated:

$$
d_n(i^*, g^*) = \sqrt{\left(x^{(1)}_{i^*} - x_{g^*}\right)^2 + \left(y^{(1)}_{i^*} - y_{g^*}\right)^2} \tag{8}
$$

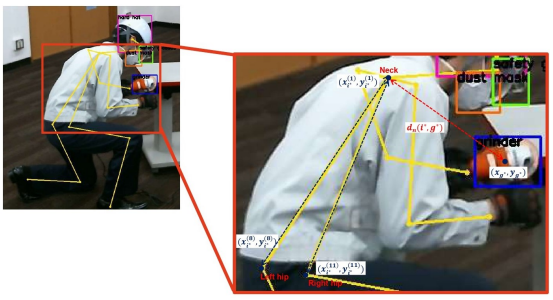


Figure 3. Relationship identification strategies to address the rule "Never operate a grinder near face".

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

$$
c^{face}_{i^*\leftrightarrow g^*} = \begin{cases} \text{“}near\text{”} & ,if\ d_n(i^*, g^*) < \beta_{i^*\leftrightarrow g^*} \\ N/A & ,otherwise \end{cases} \tag{9}
$$

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

$$
\begin{aligned}
\beta^{face}_{i^*\leftrightarrow g^*} = \max(&\sqrt{(x^{(1)}_{i^*} - x^{(8)}_{i^*})^2 + (y^{(1)}_{i^*} - y^{(8)}_{i^*})^2},\\
&\sqrt{(x^{(1)}_{i^*} - x^{(11)}_{i^*})^2 + (y^{(1)}_{i^*} - y^{(11)}_{i^*})^2}) \cdot \gamma
\end{aligned} \tag{10}
$$

and $c^{face}_{i^*\leftrightarrow g^*}$ indicates the connection to create the relationship between the face of i^* and g^* as a semantic phrase $(face, c^{face}_{i^*\leftrightarrow g^*}, g^*)$ (e.g., the relationship $(face, near, grinder)$ is created in Figure 3)

3.1.3 Scene information representation

Based on the semantic phrases established by individual-object relationship analysis, we generate a scene graph for image information representation for each captured frame from on-site surveillance cameras. An example is illustrated in Figure 4: on-site image (Figure 4(a)) is identified and transformed to semantic phrases triplets (Figure 4(b)), and coded in a scene graph $G(V, E)$ (Figure 4(c)), where V is the set of vertices to represent the objects in the semantic phrase triplets and $E = \{\{\mu, \nu\} : (\mu, \nu) \in V^2, \mu \neq \nu\}$ is the set of edges to represent the relationships in the semantic phrase triplets.

3.2 Automated reasoning for hazards identification

3.2.1 Regulatory Information Representation

To represent regulatory Information from natural language sentences, we have proposed a textual information representation and transformation method to encode the regulatory rules into the graph structure [15]. We first decompose and transform the regulatory rules to semantic phrases which are are defined as a triplet (e.g., (object1, relation, object2)). "object1" or "object2" is subjected to a particular ontology in regulations, which can be a "personnel" (e.g., worker) or a "thing" (e.g., PPE). "Relation" semantically connects objects with limitations, such as geometric (e.g., beneath, in, on), and possession (e.g., has). Let $\hat{G}(\hat{V}, \hat{E})$ be the 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 relationships in the semantic phrase triplets (s and r are the connections to represent an entity relationship and an entity status, respectively. t is the edge property to indicate the type of the requirements: obligation rule or prohibition rule).

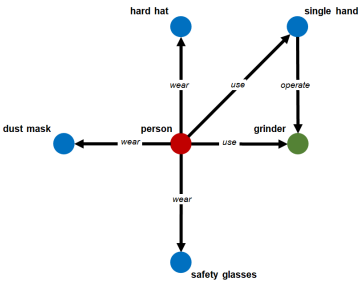
3.2.2 Frame-level reasoning for hazards identification

Frame-level reasoning for hazards identification is performed by checking compliance of prohibition and obligation regulatory rules based on graph structure analysis

(a) On-site image.

- (person, wear, hard hat)
- (person, use, grinder)
- (person, use, single hand)
- (single hand, operate, grinder)
- (person, wear, dust mask)
- (person, wear, safety glasses)

(b) Semantic phrases triplets.



(c) Scene graph $G\,(V,E)$.

Figure 4. An example of on-site image and its scene graph

between $G\,(V,E)$ and $\hat{G}(\hat{V},\hat{E})$. $\hat{G}(\hat{V},\hat{E})$ consists of both prohibition and obligation regulatory rules. Thus pruning is first performed to extract the prohibition regulatory rules subgraph $\hat{G_P}\left(\hat{V_P},\hat{E_P}\right)$ and the obligation regulatory rules subgraph $\hat{G_O}\left(\hat{V_O},\hat{E_O}\right)$ (see Figure 5).

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

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 $G \cong \hat{G_O}$ [18]. Otherwise, $G(V,E)$ is non-isomorphic to $\hat{G_O}'(\hat{V_O},\hat{E_O})$ and the violated obligation regulatory rules $H_O = \{(\mu,\nu,\tau) \in \hat{E_O},(\mu,\nu,s,r) \notin E\}$ from the on-site image are identified.

3.2.3 Sequence-level reasoning for hazards identification

Frame-level hazards identification results are subject to misidentification due to environmental or occlusion reasons, thus we dynamize the graph structure which represents the scene information to perform verification of current frame's identification results with the identification results of historical frames as sequence-level reasoning. For each node of the scene graph the set of identification results of the previous k frames is saved as "window states", which is updated dynamically in the form of "first-in-first-out". The identification results of the current frame is determined by the "window states": for each "window states" we identify the state with the majority as the final identification results of the current frame according to the "winner-takes-all" (WTA) principle (Figure 6). As an example, Figure 6(c) visualizes the sequence-level reasoning result of image sequence Figure 6(a).

4 Experiments and results

4.1 Regulatory rules

To demonstrate the validity of the proposed approach, we selected five regulatory rules to perform the experiments:

1. "Wear a hard hat."
2. "Wear a dust mask when operating a grinder."
3. "Wear a safety glasses when operating a grinder."
4. "Always use two hands when operating a grinder."
5. "Never operate a grinder near face."

As demonstrated in Figure 7, a graph $\hat{G}\left(\hat{V},\hat{E}\right)$ is generated to represent all linguistic information of these five regulatory rules.

4.1.1 Ddatasets

To create a training dataset for object detection, we collected images of hard hats, dust masks, safety glasses, and grinders from two sources: real-world images and web-mined images . Real-world images were obtained using a SONY α5000 digital camera at six different places (including construction sites, school campus, and experimental rooms) under different environmental conditions (e.g., level of illumination, visual range) and the obtained images are all 1440×1080 in resolution. Web-mined images were retrieved by search engines using a web crawler. The resolution of these collected web images ranges from 150×150 to 4896×3672. A total of 6029 images containing 12265 objects were collected and annotated as listed in Table 1.

Furthermore, to address compliance with these five regulatory rules, a demo video on the operation of the grinder

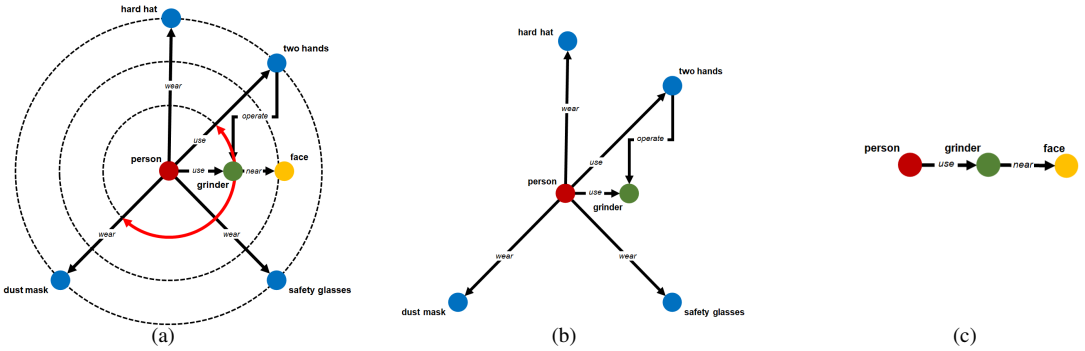


Figure 5. (b) $\hat{G_O}\left(\hat{V_O}, \hat{E_O}\right)$ and (c) $\hat{G_P}\left(\hat{V_P}, \hat{E_P}\right)$ are the obligation and prohibition regulatory rules subgraph extracted from (a) regulatory rules graph $\hat{G}(\hat{V}, \hat{E}, \hat{C})$.

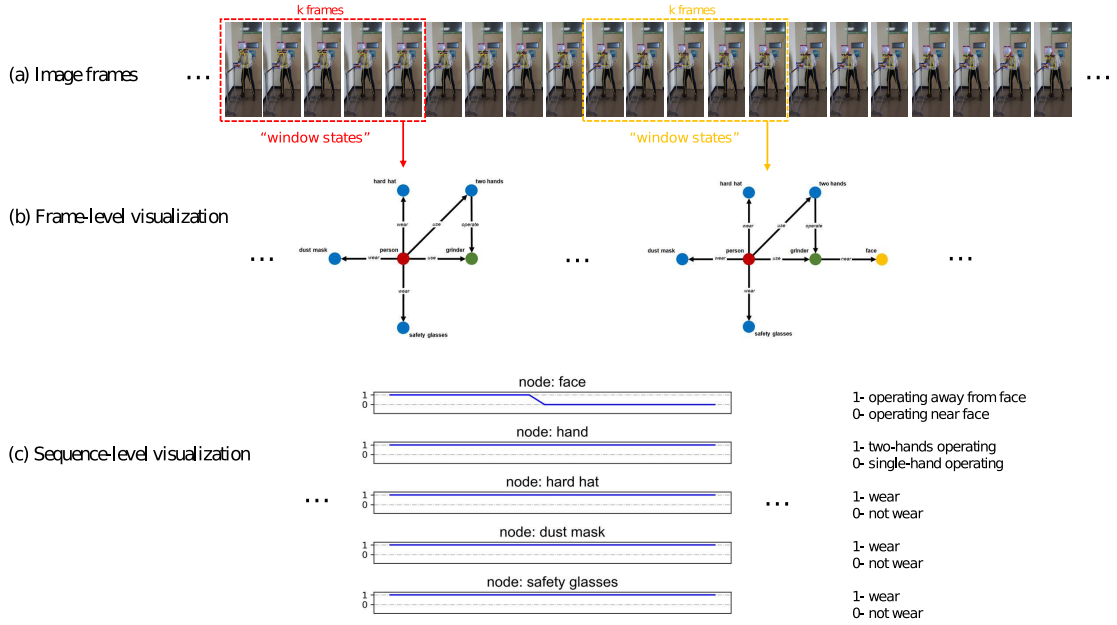


Figure 6. Sequence-level reasoning for hazards identification

was captured as the testing dataset. The demo video consists of 2745 frames and includes both normal and hazardous operations (one-handed and near-face operations).

4.2 Implementation Details

We build the YOLOv4-CSP model using PyTorch and initialize the model based on the pre-trained weights on the MSCOCO 2017 object detection dataset [19]. We train the model for 100 epochs by stochastic gradient descent (SGD), and throughout training we use a batch size of 8, a momentum of 0.9 and a decay of 0.0005. The learning rate is initialized to $1e{-}2$ and is decreasing following the cosine function. All experiments are performed on a machine with Intel Core i7-7820X (8 cores, 3.6GHz), 32GB DDR4 SDRAM RAM, NVIDIA GeForce GTX 1080 Ti GPU (11GB of GDDR5X memory and 3584 CUDA cores).

4.3 Experimental results

The sequence-level hazards identification results are reported in Figure 8, where the values of windows size k are considered to be 1 (without windows states analysis), 10, 20. When using window states for time-series analysis, sporadically misidentified frames can be well corrected. As shown in Figure 8(b) and Figure 8(c), the proposed approach achieves significant improvements for hazardous operation of grinder near face (95.47% $\rightarrow$ 96.28% $\rightarrow$ 96.36%), improper operation of grinder with single hand (97.74% $\rightarrow$ 98.79% $\rightarrow$ 98.26%), and proper use of dust

Table 1. Information of collected training dataset

	Number of real-world image samples	Number of web-mined image samples	Total
Hard hat	3003	1322	4325
Dust mask	3099	186	3285
Safety glasses	2180	115	2305
Grinder	2005	355	2360
Overall	10287	1978	12265

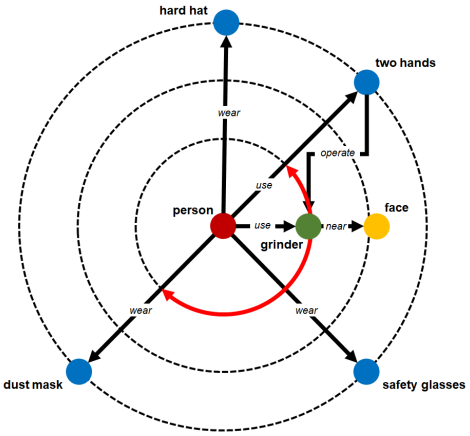


Figure 7. The graph generated for regulatory information representation.

masks (96.04% → 97.94% → 97.98%) and safety glasses (84.24% → 85.74% → 85.78%) [1], which demonstrates the effectiveness of the proposed approach for on-site occupational hazards identification.

5 Conclusion

This paper proposes a graph-based time series analysis framework to dynamically integrate visual and linguistic information for on-site occupational hazard identification. Firstly, a vision-based scene information understanding approach is introduced to process on-site images via a combination of deep learning-based object detection and individual detection. Subsequently, a novel dynamic graph structure to represent time-series information for integrated reasoning of hazards identification using window states analysis. The experimental results demonstrate that the proposed approach can effectively identify the hazards of grinder operation and facilitate improved safety inspection and supervision. Further extensions of this work will be investigated to improve the performance of visual information representation by introducing monocular 3D entity extraction.

[1]When operating the grinder near face, both dust masks and safety glasses are not visible in the frame image due to occlusion, therefore these frames are not included in the accuracy calculation for the proper use of dust masks and safety glasses

References

[1] Bureau of Labor Statistics. Industries at a glance: construction. https://www.bls.gov/iag/tgs/iag23.htm. Accessed: 5 July 2021.

[2] Japan Industrial Safety and Health Association. Osh statistics in japan. https://www.jisha.or.jp/english/statistics/index.html. Accessed: 5 July 2021.

[3] National Institute for Occupational Safety and Health. Eye safety. https://www.cdc.gov/niosh/topics/eye/. Accessed: 5 July 2021.

[4] Andrew L Dannenberg, Leonard M Parver, Ross J Brechner, and Lynn Khoo. Penetrating eye injuries in the workplace: the national eye trauma system registry. *Archives of Ophthalmology*, 110(6):843–848, 1992.

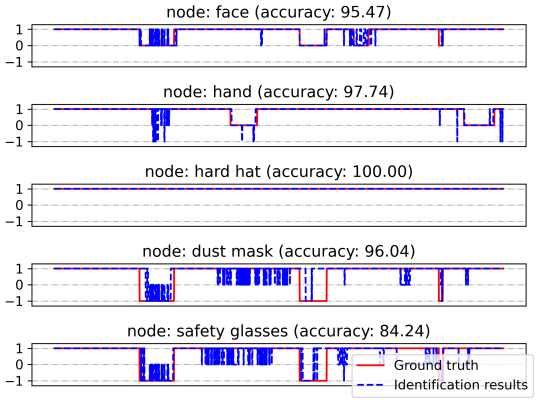
[5] Occupational Safety & Health Administration. Eye and face protection. https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.133, . Accessed: 5 July 2021.

[6] WorkSafe Division 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. Accessed: 5 July 2021.

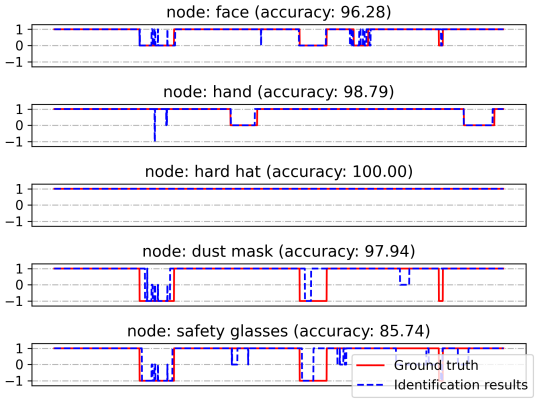
[7] Occupational Safety & Health Administration. Angle grinder safety. https://www.osha.gov/sites/default/files/2018-12/fy15_sh-27664-sh5_Toolbox_Angle_Grinder.pdf., . Accessed: 5 July 2021.

[8] Qi Fang, Heng Li, Xiaochun Luo, Lieyun Ding, Hanbin Luo, Timothy M Rose, and Wangpeng An. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Automation in Construction*, 85:1–9, 2018.

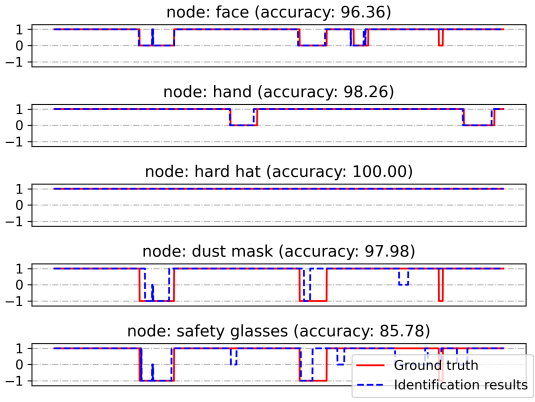
[9] Jixiu Wu, Nian Cai, Wenjie Chen, Huiheng Wang, and Guotian Wang. Automatic detection of hardhats worn by construction personnel: A deep learn-



(a) k=1 (w/o windows states analysis)



(b) k=10



(c) k=20

Figure 8. The sequence-level hazards identification results (In the vertical axis, 1 indicates a safe state, 0 indicates a hazardous state, and -1 indicates N/A).

ing approach and benchmark dataset. *Automation in Construction*, 106:102894, 2019.

[10] Nipun D Nath, Amir H Behzadan, and Stephanie G Paal. Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 112:103085, 2020.

[11] Shi Chen and Kazuyuki Demachi. A vision-based approach for ensuring proper use of personal protective equipment (ppe) in decommissioning of fukushima daiichi nuclear power station. *Applied Sciences*, 10 (15):5129, 2020.

[12] Ruoxin Xiong and Pingbo Tang. Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction*, 130: 103828, 2021.

[13] Ruoxin Xiong, Yuanbin Song, Heng Li, and Yuxuan Wang. Onsite video mining for construction hazards identification with visual relationships. *Advanced Engineering Informatics*, 42:100966, 2019.

[14] Weili Fang, Ling Ma, Peter ED Love, Hanbin Luo, Lieyun Ding, and Ao Zhou. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. *Automation in Construction*, 119:103310, 2020.

[15] Shi Chen and Kazuyuki Demachi. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Automation in Construction*, 125:103619, 2021.

[16] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[17] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13029–13038, 2021.

[18] Shu-Ming Hsieh, Chiun-Chieh Hsu, and Li-Fu Hsu. Efficient method to perform isomorphism testing of labeled graphs. In *International Conference on Computational Science and Its Applications*, pages 422–431. Springer, 2006.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.