Proceedings of the INMM 64th Annual Meeting

Malicious behaviors identification in nuclear security based on visual relationships extraction and knowledge reasoning

Zhan Li[*]
The University of Tokyo

Xingyu Song
The University of Tokyo

Shi Chen
The University of Tokyo

Kazuyuki Demachi
The University of Tokyo

ABSTRACT

Intrusion and sabotage to nuclear facilities pose serious consequences to society safety and economic loss, or which the protection against human malicious behaviors is also necessary and critical. However, commonly employed physical protection systems usually rely on manual monitoring and extensive sensor deployment, which proves to be easily missed and costly. To this end, works have been conducted to introduce deep learning-based object detection and action recognition models to automatically perform identification of human malicious behaviors. However, such approaches allow the identification of only malicious behaviors with relatively obvious features, such as carrying of malicious weapons or directly aggressive actions. For human behaviors with ambiguous features or non-unique intents, i.e., the reasoning result would vary according to the situation, a unified model with comprehensive reasoning capabilities turns out to be necessary and applicable. In this paper, a novel human malicious behaviors identification approach based on visual relationships extraction and knowledge reasoning is proposed. First of all, it should be noted that the main entities in surveillance videos, i.e., objects, humans, and actions, are detected or predicted using deep learning-based calculation models. Subsequently, the visual relationships in videos are extracted by temporal-spatial relationships analysis, and are then converted to knowledge units. Finally, human malicious behaviors identification is performed based on knowledge reasoning. In order to complete this task, a nuclear security-specific knowledge base is pre-generated according to the statistical information of training dataset, which contains features of typical knowledge unit element sequences annotated as malicious behaviors in nuclear security. Therefore, in the testing phase, the reasoning process could be carried out by checking the existence of items in the knowledge base. With the implementation of the proposed approach, a preliminarily identification of human malicious behaviors in four scenarios, i.e., fence climbing, wire net cutting, weapons holding, and normal status, has been conducted.

INTRODUCTION

With the advancement and progress of nuclear engineering technology, a numerous number of people have been benefited from the clean nuclear power. However, after the occurrence of Chernobyl and Fukushima Daiichi nuclear accident [1-2], priority has been given to the safety and radiation prevention capabilities of nuclear power plants. Apart from this, some man-made malicious behaviors are also conducted including sabotage, invasion and theft [3-6]. Comparing to system safety which might suffer from natural disasters and physical system damage, the identification of human malicious behaviors is more likely to be demanding and challenging, due to the complex spatial interactivity and temporal causality of human behaviors.

For identification of human malicious behaviors, the application of deep learning technology never fails to be effective, while the fundamental information in surveillance camera could be obtained automatically and quickly. However, much related research mainly focuses on the overall analysis process of videos [7-8], lacking detailed analysis and reasoning process for each individual character. Besides, the analysis of human-object interaction [9] and temporal action sequences [10] also turns out to be inadequate. In this paper, a novel calculation framework for identification of human malicious behaviors is proposed, as shown in Fig.1.

In Fig.1, it could be easily seen that the proposed framework is mainly comprised by computer vision module and reasoning module. In computer vision module, the key information including objects, temporal actions and human-object interaction actions would be extracted by some typical deep learning models. Then, in reasoning module, the extracted information would be compared with the elements in the pre-constructed knowledge base so that the final reasoning result could be output.

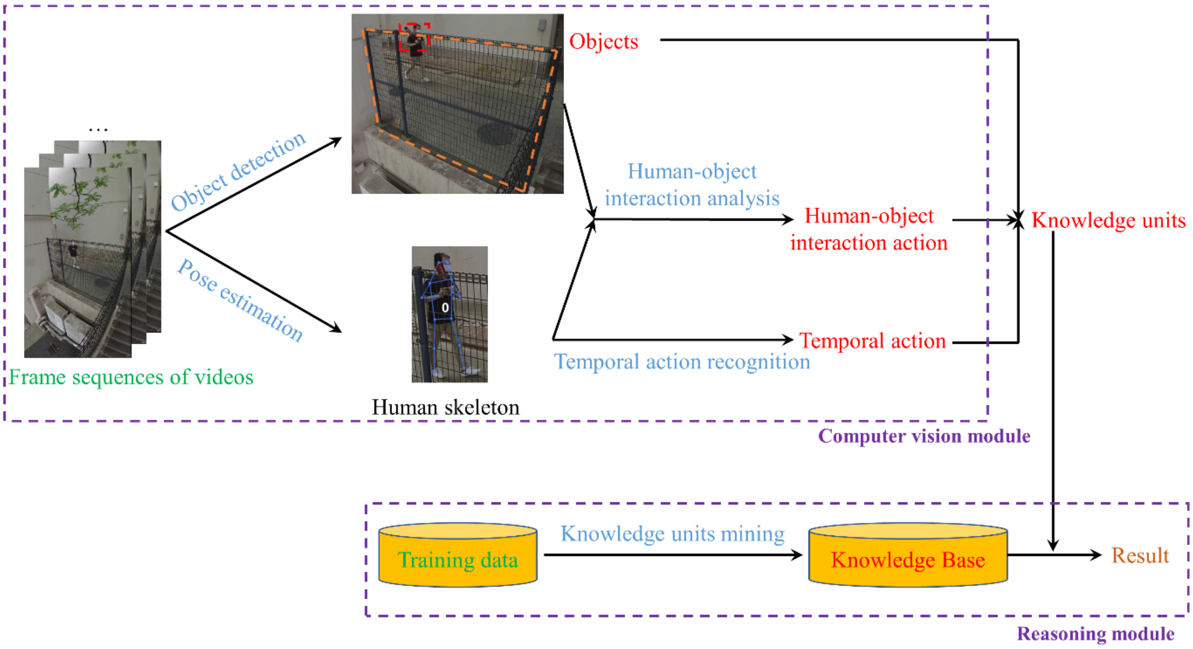


Fig.1 Framework for identification of human malicious behaviors

COMPUTER VISION MODULE

The computer vision module aims to extract three types of information from the input frame sequences of videos, which are objects, temporal actions and human-object interaction actions. Additionally, as shown in Fig.1, it should be noted that there are totally four calculation stages (blue texts) in this module, while three of them are typical deep learning models and the rest one is based on the geometry analysis.

• *Object detection*. The stage of object detection is used for extracting all the nuclear security-related objects occurred in input frame sequences. With the application of the typical model YOLOX [11-12], the coordinates of object bounding boxes with its corresponding categories could be obtained.

• *Pose estimation*. During the procedure of pose estimation, a batch of skeleton joint coordinates would be extracted for each person. In this research, the COCO-format human pose topology [13] is chosen as the skeleton structure (as shown in Fig.2), while the *MMPose* toolbox [14] is applied to extracting joint coordinates. The output information of human skeletons would be then utilized for the subsequent process, which are temporal action recognition and human-object interaction analysis.

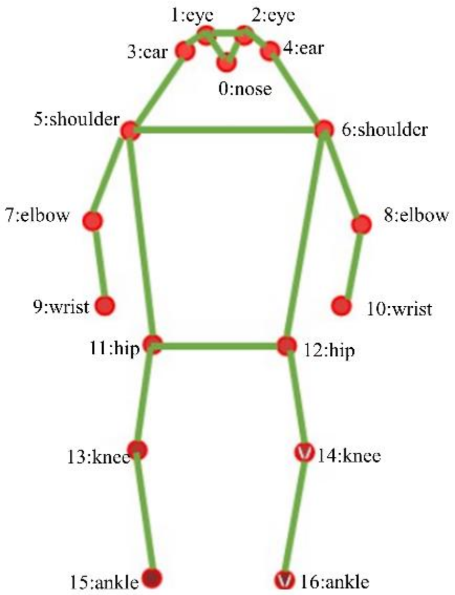


Fig.2 Construction of human pose topology [13]

• *Temporal action recognition*. Based on the output information of human skeleton by pose estimation, temporal action recognition tends to determine the category of temporal action in current frame for each human. The type of temporal actions only depends on the skeleton information of humans, while the interaction analysis with other objects is not considered. In this research, a simple two-branch graph neural network with *SAGEconv* operator [15] and global mean pooling operator is applied to solve this task, as shown in Fig.3.

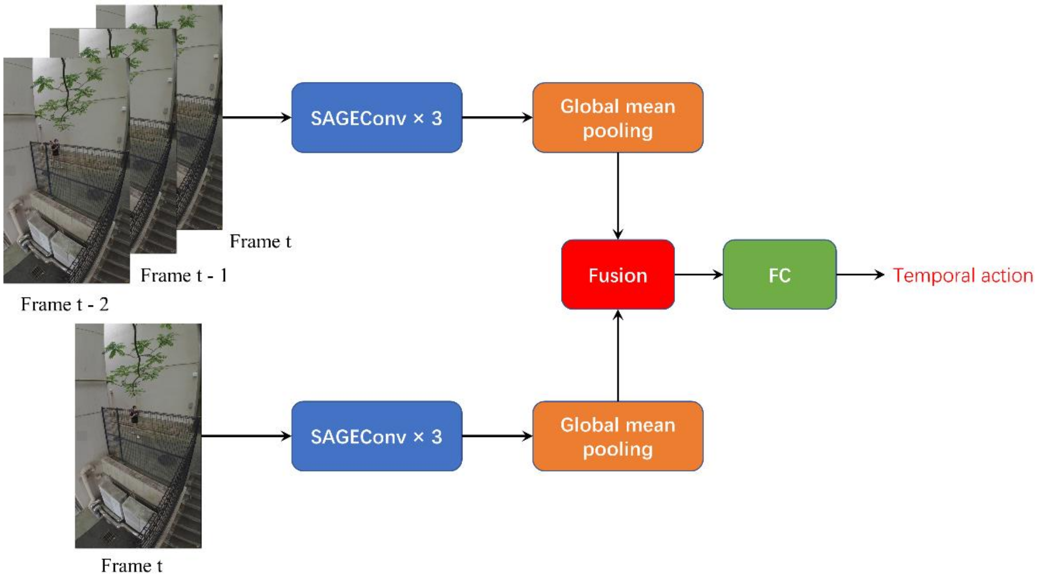


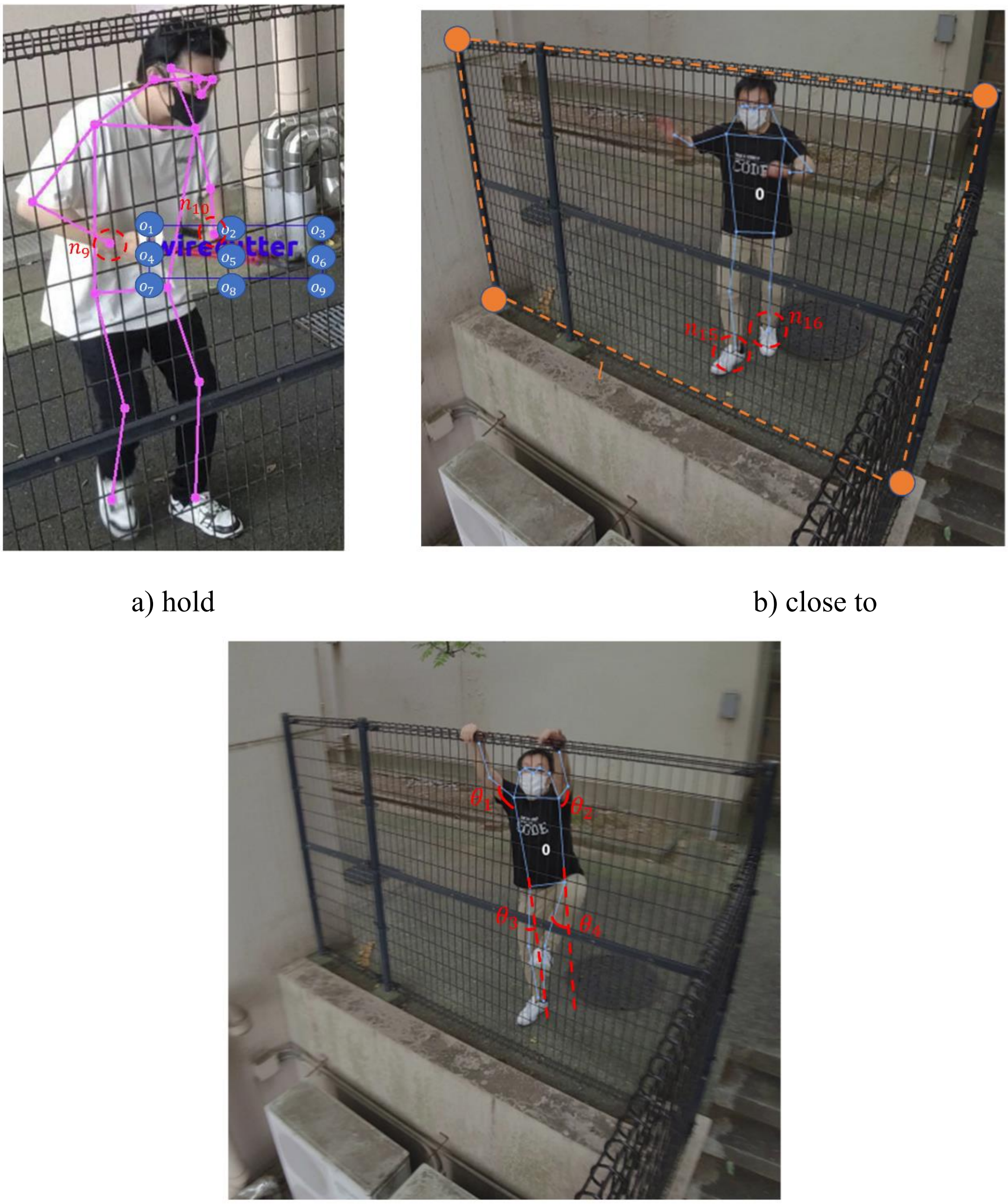
Fig.3 Construction of two-branch graph neural network

• *Human-object interaction analysis*. Although there are some existing deep learning models for this task [16-17], they are slightly immature with relatively low accuracies. Therefore, the stage of human-object interaction analysis is conducted by simple geometry analysis, which turns out to be appliable and useful [18-19]. The objective of this stage is to detect and determine the category of human-object interaction actions, which depends on the spatial configuration of human skeletons and objects. In this paper, as shown in Fig.4, only three types of human-object interaction actions are considered, which are hold, close to and climb, respectively.

(1) *Hold*. As shown in Fig.4a, the existence of action 'hold' would depend on the distance between wrist joints and key points of objects. The wrist joints in human skeleton are denoted as n_9 and n_{10}, referring to the human pose topology in Fig.2. On the other hand, nine key points of objects are chosen in our research ($o_1 - o_9$ in Fig.4a), containing four corner points, four side midpoints and one center point of object bounding box. If the minimum distance among the wrist joints and key points of objects was lower than threshold σ_1 (chosen as the smaller value of half-width and half-height of bounding box), then the interaction action could be determined as hold.

(2) *Close to*. In our research, this action is only related to the object 'wire net'. In order to detect the action 'close to', the distance between ankle joints (n_{15} and n_{16}) and the bottom borders of objects (*l*) would be calculated, as shown in Fig.4b. If the minimum distance value is lower than threshold σ_2 (chosen as half the length of human legs), the action 'close to' could be determined.

(3) *Climb*. When the action 'close to' is detected in current frame or last frame, it should be further judged for the existence of 'climb'. Therefore, as similar to the action 'close to', the action 'climb' is only related to the object 'wire net'. In this process, four key angles in human skeleton should be calculated ($\theta_1 - \theta_4$ in Fig.4c). If the maximum value between θ_1 and θ_2 is larger than $\frac{\pi}{6}$ and the maximum value between θ_3 and θ_4 is larger than $\frac{\pi}{4}$, the human-object interaction action 'climb' could be determined.



a) hold

b) close to

c) climb

Fig.4 Samples of three human-object interaction actions

REASONING MODULE

As mentioned in the previous section, with the calculation procedures in computer vision module, the information of objects, temporal actions and human-object interaction actions could be extracted. For reasoning the identification result of human behaviors in an explainable and practical way, a knowledge base would be pre-constructed which is comprised of multiple knowledge units. Hereafter, in the application phase, if the knowledge units extracted from the input frame sequences include any element in knowledge base, then the testing video would include human malicious behaviors. In our research, there are totally three types of knowledge units, as stated in follows.

- *Temporal actions (TA)*. For example, 'stand'.

- *Human-portable object interaction actions with corresponding objects(HPOIA)* . The portable objects are defined as those whose coordinates would change during the time. Therefore, 'hold wire cutter' could be regarded as HPOIA, since the coordinates of wire cutter would usually be changed as the carrier moves.

- *Human-background object interaction actions with corresponding objects(HBOIA)*. Different from portable objects, the coordinates of background objects are always fixed in the whole video. For example, 'climb wire net' is a typical type of HBOIA.

Based on the definition of three knowledge units, the statistical information of them in training dataset would be calculated so that all the malicious knowledge units (i.e., the knowledge units usually lead to malicious results) would be extracted. This procedure is denoted as 'knowledge units mining' in Fig.1, while the detailed framework of it would be seen in Fig.5.

In Fig.5, it could be easily seen that there are mainly two sub-processes in knowledge units mining. Above all else, in the sub-process of 'count the number of malicious types', the numbers of all possible malicious types would be recorded, including the normal samples, malicious samples as well as total samples in training dataset. In our research, there are totally three types of malicious types, which are self-malicious, co-occurrence malicious and sequential malicious, and the definition of them could also be seen in Fig.5. On the other hand, in the sub-process of 'judgement of malicious units', there are two additional stages. In the stage of preliminary judgement, the ratio of malicious samples for each element in each malicious type would be calculated. If the value of this ratio is larger than the corresponding threshold (as shown in Fig.5, for three malicious types, they are 1, α_1 and α_2, respectively), the unit would be determined as malicious knowledge unit. Hereafter, in the stage of filtering, the redundant elements in 'co-occurrence malicious' and 'sequential malicious' would be deleted, as mentioned in Fig.5. Finally, the knowledge base, which contains all the malicious knowledge units in three malicious types, would be output.

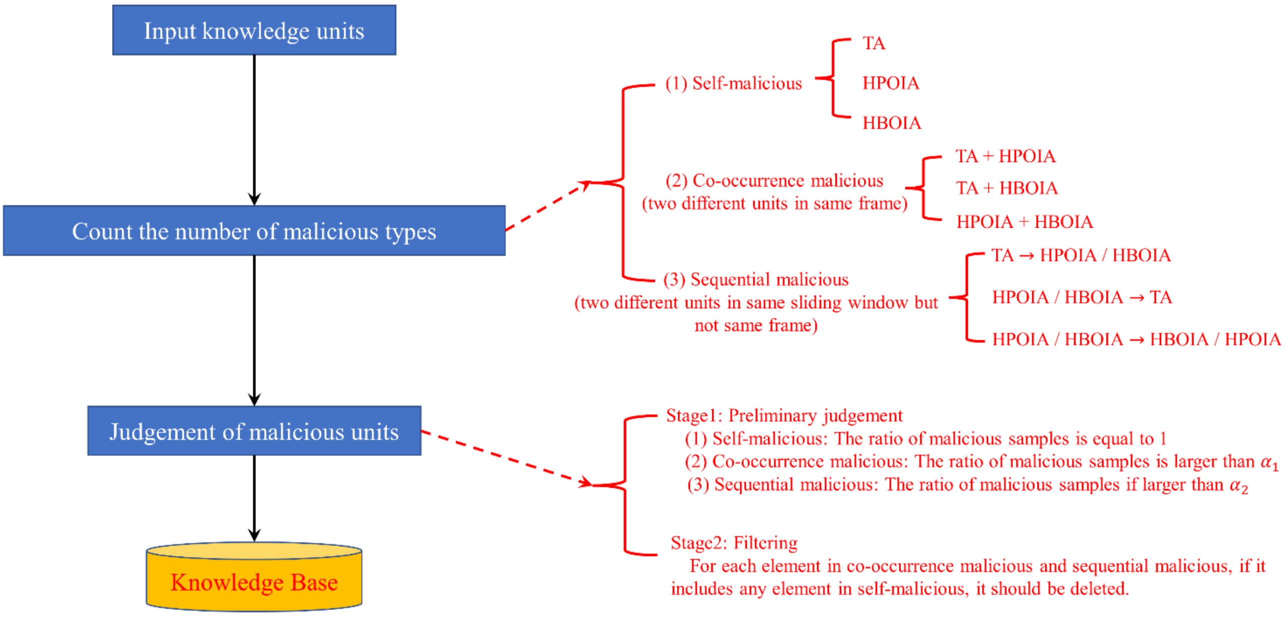

Input knowledge units
Count the number of malicious types
Judgement of malicious units
Knowledge Base
(1) Self-malicious
TA
HPOIA
HBOIA
(2) Co-occurrence malicious
(two different units in same frame)
TA + HPOIA
TA + HBOIA
HPOIA + HBOIA
(3) Sequential malicious
(two different units in same sliding window but not same frame)
TA → HPOIA / HBOIA
HPOIA / HBOIA → TA
HPOIA / HBOIA → HBOIA / HPOIA
Stage1: Preliminary judgement
(1) Self-malicious: The ratio of malicious samples is equal to 1
(2) Co-occurrence malicious: The ratio of malicious samples is larger than α_1
(3) Sequential malicious: The ratio of malicious samples if larger than α_2
Stage2: Filtering
For each element in co-occurrence malicious and sequential malicious, if it includes any element in self-malicious, it should be deleted.

Fig.5 Framework of knowledge units mining

EXPERIMENTS AND RESULTS

In order to verify the feasibility and practicality of the proposed framework in Fig.1, a self-shot dataset is utilized. In this dataset, four typical scenarios in the field of nuclear security are included, which are normal status, fence climbing, wire net cutting and weapon holding. Among them, 51 videos (17 min 47 s) are used for training while the rest 37 videos (9 min 39 s) are used for testing.

For evaluation of computer vision module, the metrics of weighted precision (*wPre*) and weighted recall (*wRec*) are chosen [20], while the results could be seen in Table 1. In Table 1, it could be seen that the results for the tasks of object detection and human-object interaction analysis turn out to be acceptable and practical. However, for the task of temporal action recognition, both the values of *wPre* and *wRec* are approximately 0.57, which needs to be further advanced in the future work.

On the other hand, for evaluation of reasoning module, there are still three undetermined parameters in Fig.5, which are α_1, α_2 and the size of sliding windows *SW* [21]. However, it should be obvious that these two values should be relatively large, otherwise the malicious results could be seen as accidental events. The parameter analysis procedure of reasoning module could be seen in Table 2. In Table 2, it could be easily known that the output knowledge base could be perfectly correct when α_1=α_2=0.7 (in fact, if these values are larger than 0.7, the output would still be perfect). Besides, with the increase of SW, the value of precision might be decreased while the value of recall is still unchanged. Finally, when using the perfect output knowledge base for identification of human malicious behaviors, the precision and recall for reasoning results are 0.76 and 1.00, respectively. In this way, the proposed calculation framework in Fig.1 turns out to be practical and promising that the actual malicious behaviors could be perfectly detected.

Table 1 Evaluation results of computer vision module

Task	Categories	*wPre*	*wRec*
Object detection	wire cutter, launcher	0.9834	0.7500
Temporal action recognition	stand, walk, squat	0.5757	0.5756
Human-object interaction analysis	hold, close to, climb	0.9425	0.7839

Table 2 Parameter analysis of reasoning module

Experiment	SW	α_1	α_2	Knowledge Base	Precision	Recall
#1	3	0.6	0.6	Hold launcher Climb wire net Hold wire cutter + close to wire net Squat + hold wire cutter Hold wire cutter → stand Hold wire cutter → close to wire net Close to wire net → hold wire cutter	0.7143	1.0
#2	3	0.7	0.7	Same to reference solution	1.0	1.0
#3	4	0.6	0.6	Hold launcher Climb wire net Hold wire cutter + close to wire net Stand → hold wire cutter Hold wire cutter → stand Hold wire cutter → close to wire net Close to wire net → hold wire cutter	0.7143	1.0

#4	4	0.7	0.7	Same to reference solution	1.0	1.0
#5	5	0.6	0.6	Hold launcher Climb wire net Stand + hold wire cutter Hold wire cutter + close to wire net Squat + hold wire cutter Stand → hold wire cutter Hold wire cutter → stand Hold wire cutter → close to wire net Close to wire net → hold wire cutter	0.5556	1.0
#6	5	0.7	0.7	Same to reference solution	1.0	1.0
Reference solution				Hold launcher Climb wire net Hold wire cutter + close to wire net Hold wire cutter → close to wire net Close to wire net → hold wire cutter		

CONCLUSIONS

For solving the problems of human malicious behaviors identification in nuclear security, a novel and comprehensive calculation framework is proposed, which contains computer vision module and reasoning module. In computer vision module, three typical deep learning models and one geometry analysis process are utilized to extract the key information of input videos, including objects, temporal actions and human-object interaction actions. Then, in reasoning module, the identification result would be determined by referring to the elements in pre-constructed knowledge base.

The dataset in this research is comprised by multiple self-shot videos, containing four typical scenarios related to nuclear security, which are normal status, fence climbing, wire net cutting and weapon holding. The results show that the performance for the tasks of object detection and human-object interaction analysis prove to be effective. However, the *wPre* and *wRec* for temporal action recognition are only 0.57, which still need to be advanced. On the other hand, for reasoning module, when α_1=α_2=0.7, the output knowledge base is completely equal to the reference solution. Based on this knowledge base, the values of precision and recall for reasoning results achieve 0.76 and 1.00, respectively.

In summary, the proposed calculation framework turns out to be practical and promising for solving the problems of human malicious behaviors identification.

REFERENCE

[1] Steinhauser, G., Brandl, A., Johnson, T.E., 2014. Comparison of the Chernobyl and Fukushima nuclear accidents: A review of the environmental impacts. Science of the Total Environment 470-471, 800-817.

[2] Laffolley, H., Journeau, C., Delacroix, J., Grambow, B., Suteau, C., 2022. Synthesis of Fukushima Daiichi Cs-bearing microparticles through molten core-concrete interaction in nitrogen atmosphere. Nuclear Materials and Energy 33, 101253.

[3] Bedi, R., 2009. Radioactive water poisons 55 at Indian nuclear plant. https://www.irishtimes.com/news/radioactive-water-poisons-55-at-indian-nuclear-plant-1.781115. Accessed date: 30 November 2009.

[4] Reuters, 2022. Russia blames attack at nuclear power station on Ukrainian saboteurs. https://www.reuters.com /world/europe/russia-blames-attack-nuclear-power-station-ukrainian-saboteurs-interfax-2022-03-04/. Accessed date: 4 March 2022.

[5] World news, 2007. Fire at Japanese nuclear power plant. https://www.theguardian.com/world/2007/jul/24/nuclear.japan. Accessed date: 24 July 2007.

[6] Bunn, M., Bunn, G., 2001. Nuclear Theft & Sabotage Priorities for Reducing New Threats. IAEA Bulletin 43(4), 20-29.

[7] Khaire, P., Kumar, P., 2022. A semi-supervised deep learning based video anomaly detection framework using RGB-D for surveillance of real-world critical environments. Forensic Science International: Digital Investigation 40, 301346.

[8] Kim, S., Hwang, S., Hong, S.H., 2021. Identifying shoplifting behaviors and inferring behavior intention based on human action detection and sequence analysis. Advanced Engineering Informatics 50, 101399.

[9] Li, Y., Liu, X., Wu, X., Huang, X., Xu, L., Lu, C., 2022. Transferable interactiveness knowledge for human-object interaction detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(7), 3870-3882.

[10] Han, Y., Zhuo, T., Zhang, P., Huang, W., Zha, Y., Zhang, Y., Kankanhalli, M., 2022. One-shot video graph generation for explainable action reasoning. Neurocomputing 488, 212-225.

[11] Hong, R., Xiao, B., Yan, H., Liu, J., Liu, P., Song, Z., 2023. Multitemporal greenhouse mapping for high-resolution remote sensing imagery based on an improved YOLOX. Computers and Electronics in Agriculture 206, 107689.

[12] Song, C.Y., Zhang, F., Li, J.S., Xie, J.Y., Yang, C., Zhou, H., Zhang, J.X., 2022. Detection of maize tassels for UAV remote

sensing image with an improved YOLOX model. Journal of Integrative Agriculture (in press).

[13] Ji, Z., Wang, Z., Zhang, M., Chen, Y., Qian, Y., 2022. 2D Human Pose Estimation with Explicit Anatomical Keypoints Structure Constraints. arXiv preprint, arXiv:2212.02163.

[14] MMPose Contributors, 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

[15] Hamilton, W.L., Ying, R., Leskovec, J., 2017. Inductive Representation Learning on Large Graphs. Conference on Neural Information Processing Systems (NIPS), 2017.

[16] Li, Q., Xie, X., Zhang, J., Shi, G., 2022. Language-guided graph parsing attention network for human-object interaction recognition. Journal of Visual Communication and Image Representation 89, 103640.

[17] Su, Z., Wang, Y., Xie, Q., Yu, R., 2022. Pose graph parsing network for human-object interaction detection. Neurocomputing 476, 53-62.

[18] Chen, S., Demachi, K., 2021. 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.

[19] Chen, S., Demachi, K., Dong, F., 2022. Graph-based linguistic and visual information integration for on-site occupational hazards identification. Automation in Construction 137, 104191.

[20] Talukder, M.S.H., Sarkar, A.K., 2023. Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning. Smart Agricultural Technology 4, 100155.

[21] Huang, H., You, Z., Cai, H., Xu, J., Lin, D., 2022. Fast detection method for prostate cancer cells based on an integrated ResNet50 and YoloV5 framework. Computer Methods and Programs in Biomedicine 226, 107184.