2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
June 28-30, 2023. Seattle, Washington, USA

Virtual Reality System using Explainable AI for Identification of Specific Expert Refinery Inspection Skills

Hiroki Takeuchi[1], Ryota Takamido[2], Shinji Kanda[2], Yasushi Umeda[2], Hajime Asama[1], Seiji Kasahara[3], Seigo Fukumoto[3], Sunao Tamura[3], Toshiya Kato[3], Masahiro Korenaga[3], Akinobu Sasamura[3], Misaki Hoshi[3], Jun Ota[2]

***Abstract*— In this study, we propose a virtual reality system for identifying expert-specific skills in a visual inspection task in a refinery by using an eXplainable Artificial Intelligence (XAI) technique. Most previous studies have applied statistical analysis such as t-tests to the mean value of the experimental data, and there is a consequent lack of specificity in the results (i.e., when and where expert skill appears within a long inspection duration). It is thus difficult to provide feedback based on the most important part of the collected experts' data to the novices. To address this issue, we introduce a Convolutional Neural Network (CNN) with Class Activation Map (CAM) technique, an XAI method, to analyze the experimental data of experienced and novice field operators, and identify the most significant contributors for classifying expert and novice behavior for 120 seconds inspections. The resulting model can classify field operators as expert or novice with an accuracy of 99.1% on average, and visualize the classification criteria as a heat map for each experimental trial. Based on those results, we propose a virtual reality training system for learning expert inspection skills by referencing the CNN results. The contribution of our study is the proposition of a new analytical framework, as well as a training system beyond the limitations of conventional statistical analysis.**

I. Introduction

Daily field patrol is essential for safe and stable refinery operations. During the field patrol process, experts move around the operation areas and attempt to identify any potential and/or revealed anomalies in the environment. The oversight of anomalies can cause serious incidents; hence, it is important to increase the probability of anomaly detection by improving the inspection skill of field operators.

Many previous studies have investigated the differences in inspection behavior between experts and novices. This is often done through simulated inspection task experiments, where some dependent variables, such as gazing at targets and head positioning, are measured and analyzed [1,2].

However, since most previous studies applied statistical analyses such as the t-test to the mean value of the dependent variables for detecting expert–novice differences, there is a lack of specificity in the treatment of the experts' inspection behaviors. For example, even if the results show that experts tend to position their heads at a lower level than do novices, when investigating head position, it is not clear when the experts instigate a lower gaze position, nor what prompts them to do so during a long inspection. In other words, conventional statistical analyses are unable to determine which part of the experimental data indicates "know-how" which novices should learn to improve their skills. In addition to identifying the mean characteristics of experts, it is also important to identify the most important specific motion and behavior patterns. No previous studies have analyzed such a specific part of the inspection process; hence, a new analytical framework is required to address the above issue.

II. Related Work

A. Experimental studies investigating inspection skills.

As mentioned above, previous studies have attempted to classify expert inspection skills using the novice-expert paradigm, in which both novices and experts performed the same experimental inspection task, and researchers measured and compared their inspection behavior [3,4]. Most previous studies have dealt with visual inspection, such as finding illegal materials in baggage checks [5], conducting experiments by measuring and comparing the eye behavior of experts and novices using an eye tracker device [6,7]. As a result, they obtained some insights into the experts' specific behaviors during inspection tasks. For example, experts are gazing at important (high-risk) equipment for longer durations [8], and employing a more systematic and stable search order [9] than novices.

Takamido et al. [10] recently clarified the importance of head positioning behavior (motor behavior) in addition to information on the target of the gaze (perceptual behavior). Specifically, they used a virtual reality (VR) model, representing a section of a real refinery environment, and measured both the head position and gaze position data of expert and novice field operators. As a result, they revealed that experts position their heads differently for the effective detection of different anomalies (e.g., lower head positions are considered effective for leakage inspection).

However, both these insights into eye behavior and head positioning were achieved by applying statistical analysis to the mean value among all experimental trials. Therefore, as highlighted above, there continues to be a lack of specificity in the investigation of expert inspection skills.

[1] Department of Precision Engineering, School of Engineering, The University of Tokyo, Japan.
[2] Research into Artifacts, Center for Engineering (RACE), School of Engineering, The University of Tokyo, Japan.
[3] Engineering & Capital Planning Department, ENEOS Corporation, Japan.

B. *Explainable AI for identifying specific expert skills*

One approach that may address the above problem is the introduction of eXplainable Artificial Intelligence (XAI) to analyze the data obtained from experiments. XAI is an analytical technique in machine learning that visualizes the input data and identifies the part of the input data with the highest significance, using this with making predictions or classifications [11]. Although there are many methodologies for visualizing the significance of each part of the input data, a Class Activation Map (CAM) [12] is often used to analyze human movement [13]. Compared to other XAI methods, a CAM has the advantage of providing an explanation that is faithful to the original model because it does not build a surrogate model [14].

For example, Fawaz et al. [15] used XAI to explain the reason for classification of expert–novice motion in a surgery simulator by using a CAM based on the 76-dimensional position data of the surgery simulator arm. In addition, Zhang et al. [16] collected six-dimensional position and angle data from four manipulators of a surgical robotic arm. The data were differentiated 0–3 times, and 96-dimensional data were used as the input to determine the skill level of a surgery. In this process, a CAM was used to provide a visual feedback based on the model classification. However, these studies utilized only unimodal information (position information). For application to visual inspection tasks, this method must be modified such that both motion and eye movement information are considered key features for expert–novice classification [8–10].

III. Method

A. *Overview of the proposal method*

Based on the above background, this study aimed to develop a virtual reality system for identifying the specific inspection behavior of experienced field operators using an XAI technique with two different motion information resources (head position and angle) and gaze information. Figure 1 presents an overview of the proposed method. The process was divided into two parts: the training phase; and the feedback phase. In the training phase, the gaze and head position data are collected in the virtual environment, and pre-processing, such as normalization, is performed to prepare the data for input into the Convolutional Neural Network (CNN) architecture. Then, the model, using a CNN with CAM, is trained to classify the input data as either "experienced" or "novice", with six experienced and three novice field operators, and the accuracy of the classification is evaluated.

If the model shows a classification accuracy of over 90%, which is higher than that shown (67%) when all data are classified as experienced, then, we construct a feedback system that automatically displays the input head position data with the color that shows the percentage of importance to classification between the experienced and novices based on the CAM results.

Finally, we propose a use case of this system in novice training, and present a possible platform for this application. Novice field operators can perform inspection tasks in the VR environment; and the model can identify the most significant deviations in their behavior compared to that of the experienced field operators. This information is fed back to them, enabling them to identify which specific parts of their inspection process should be improved.

B. *Data for construction of the proposal system*

We used data collected in a previous study [10] for the proposed classification exercise. In [10], six experienced field operators with more than four years of experience and three novices without any practical inspection experience performed the same simulated inspection task in a virtual environment representing a section of a refinery. In the experiment, they moved around the virtual environment and attempted to identify the arbitrarily assigned anomalies within an inspection time of

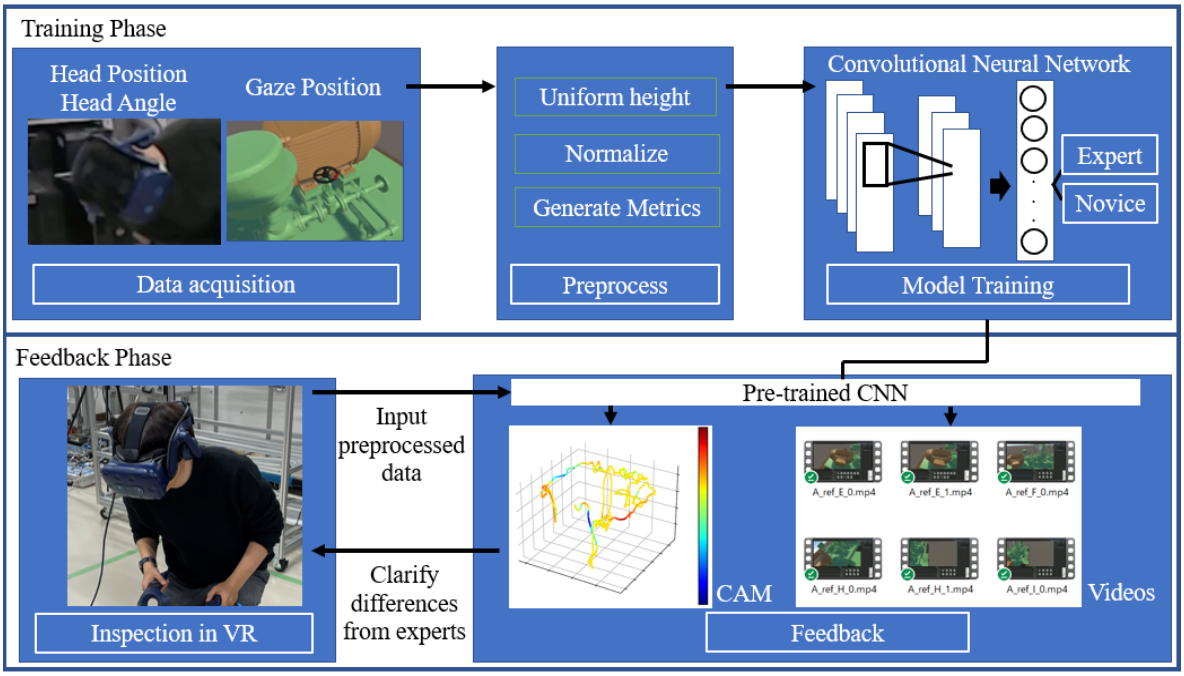

Figure 1.Overview of the virtual reality system for identifying the specific inspection movement by using Explainable-AI

120 s. The experiment was conducted ten times for each participant. Figure 2 shows the refinery model used in the experiment. The model includes typical equipment for refinery inspection tasks, such as heat exchangers, rotating machines, and pumps. These pieces of equipment are appropriate targets for performing patrol inspection tasks because they include many inspection items.

As for the training data for the CNN with CAM architecture, we used both head position data and gaze position data, which reported significant differences in some previous studies [10,17]. Specifically, the training data included two-dimensional data indicating gaze position, three-dimensional data indicating head position, three-dimensional data indicating head angle, and time differential values of each datum. In addition, to consider the interaction between motion and gaze data, six-dimensional data were added by multiplying the head and gaze position data. Thus, a total of 21 × 10800 (120 s × 90 Hz) time-series vectors were used as the input data. Using these, the CNN with CAM architecture learned the relationship between each piece of input data and the field operators' level of expertise. The z-positional data were divided by the height of each participant, and each data point was normalized. We also collected the first-person videos of the participants during the inspection activity for use in training novices.

C. CNN with CAM Model architecture

Figure 3 shows the architecture of the CNN with the CAM model used in this study. It is mainly based on the method of a previous study used for the classification of expert/novice surgeons [15]. The input data, which consisted of 21-dimensional time-series data, were convolved in two stages. First, eight filters were used in the first phase and sixteen in the second phase. Next, the data were passed to the Global Average Pooling (GAP) layer, where they were transformed into 16-dimensional scalar values, and these values were used for binary experience classification (experienced or novice). The weights in the network were then updated according to the error values. The binary cross-entropy function was used here as the error function, a ReLU function was used for the activation function, and the softmax function was the last activation function.

The presence of a GAP layer allows for CAM output. The method of creating a CAM from time-series data was introduced in [18]. In each class c, the output score S_c and importance score $I_c(t)$ used to write the CAM are expressed as follows:

$$S_c = \sum_k w_k^c \sum_t f_k(t) = \sum_t \sum_k w_k^c f_k(t), \quad (1)$$

$$I_c(t) = \sum_k w_k^c f_k(t), \quad (2)$$

where the k-th feature map is denoted by $f_k(t)$ as a function of time t, and the weight of the connection between the corresponding value at the GAP layer and the output is denoted by w_k^c.

The above architecture was defined based on a previous study [15]. However, to deal with both motion and gaze information, we modified the methodology by considering the characteristics of the target task in this study (inspection) and the collected data. First, we set the seven subgroups in the first layer according to the characteristics of the data to avoid large differences in the weights of the information content of multiple modalities, because the number of dimensions is different for each category. The subgroups were set up as follows:

① Eye position subgroup consisting of two-dimensional eye position information (x, y);

② Head position sub-group consisting of three-dimensional head position information (x, y, z);

③ Head angle sub-group consisting of three-dimensional head angle information (α, β, γ);

④ Eye velocity sub-group consisting of one-dimensional eye velocity information (v_{eye});

⑤ Head position velocity subgroup consisting of three-dimensional head position velocity information (x', y', z');

⑥ Head angle velocity sub-group consisting of three-dimensional head angle information $(\alpha', \beta', \gamma')$; and

⑦ Product of gaze position (x, y) and head position (x, y, z) subgroups, yielding six-dimensional data.

By adding the last subgroup (consisting of the results of multiplying the head and gaze position data), we can consider their interaction (e.g., viewing position and body states at the time of gazing). In addition, by outputting the CAM scores along with the head position, visual feedback is expressed in a format that is easier for humans to understand. The CAMs for each person and skill level are overlaid on top of one another so that the overall trend can be seen.

D. Training of the CNN with CAM model

The data used for training and validation were split according to the two evaluation methods described below. The model was trained using an online learning method in which the input data for each person and each trial were entered individually, and the weights were updated sequentially. Each training session was conducted over a maximum of 1000 epochs, and its performance was evaluated on the test input data, which were separated from the training input data. To avoid overfitting the training data, an

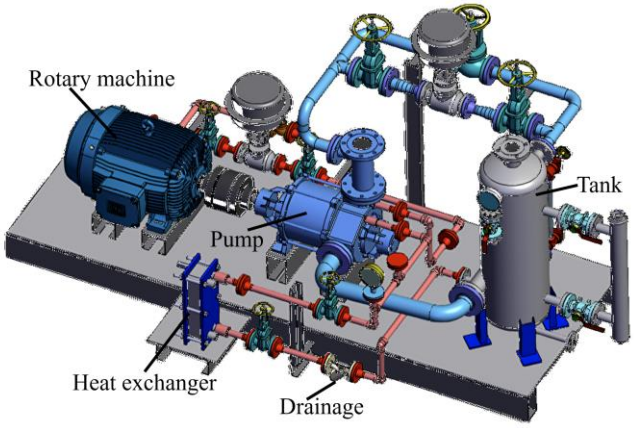

Figure 2. 3D refinery model used in VR experiments

L2 regularization term was added to the error function, and its coefficient parameter was set to 1×10^{-5}. The model with the lowest error value, i.e., the best case, is used. The GPU used to train the model was an NVIDIA RTX A5000. The Adam algorithm [19] was used to update the network weights. The initial learning rate was set as 5×10^{-4} and the parameter for adding the inertia term was set as 0.9 and 0.999, as in [19]. The initial value of He [20] was used to initialize the weights.

The trained model was evaluated using two cross validation methods: the leave-one-super-trial-out (LOSO) method; and the leave-one-user-out (LOUO) method introduced in [21]. In the former method, the i-th trial of each subject was used as the test data and the other data as the training data for cross-validation. This method confirmed whether the model correctly classified the new data when a participant performed a new trial. This was performed for 10 trials, and the sum of the predictions of all models was evaluated. In the latter case, all trials for the j-th person were used as test data, and the weights were updated using the other data. This method tested the generalization performance of the model to correctly handle the data when a new participant performed trials. This was done for the nine participants in the same manner, and the combined predictions were evaluated. The accuracy of the evaluation was calculated based on the confusion matrix.

E. Proposed of the feedback training system for novice field operators

After completing the above processes, we constructed a feedback training system for novice field operators that visualizes the different components of experienced behavior based on the output from the CAM architecture. Here, the first-person video data is used along with the data used to train the model (Figure 4(a)). In a previous study, it was shown that scrutinization of a video from the first-person view of experienced workers could improve the inspection skill of novices [22]. Hence, by observing the work of experienced field operators and comparing it with their own, novices can learn how to improve their inspection behaviors from this system. In the use case of the system, novice field operators perform the inspection task in the VR environment (Figure 2); then, the most different part of their behavior within all inspection durations is fed back to them, and they can check which specific part of their inspections should be improved (Figure 4(b)).

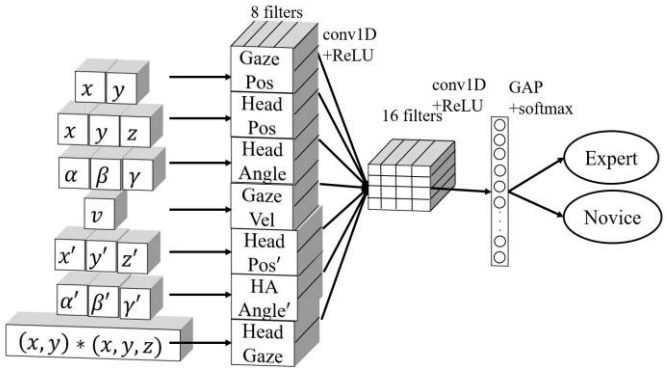

Figure 3. Overview of proposed CNN Model

The specific process is as follows. First, novice field operators perform inspection tasks using the VR system. Then, the array of scores, based on the input model and contributed by the CAM model to the decision-making process, can be obtained. The part of the data with the greatest contribution to classification over each 120 s inspection is identified, and the novice is shown "when and where" the novice–expert difference is most obvious through videos of both field operators at the corresponding time. In particular, the segment with the highest continuous importance in the time-series data is extracted, and this section of the video, including the three seconds before and after the segment, is clipped. Through this system, novices can confirm which part of their inspection behavior should be most improved by referencing the corresponding experienced field operator's behavior.

IV. Results and Discussions

Tables 1 and 2 show the results of the evaluation of the CNN for the classification between experienced (E) and novice (N) field operators. Each evaluation index exhibited a high ratio. Mean accuracy was 98.9% for LOSO, and 88.8% for LOUO. One potential reason the LOSO method had larger values than the LOUO method may be the number of participants. Since the total number of participants in this study was relatively small (nine participants), the individual differences between each group may have affected the performance.

These results suggest that the proposed method, using CNN with CAM and data on gaze position, head angle, and head position, classified the inspection behavior between experienced and novice with an accuracy of over 90%. Therefore, the visualizations and feedback on highly-contributing data for classification are likely to be useful to novices.

Figure 5 shows two examples of CAM results obtained when a novice performed inspection activities in a VR space. Once the novices' data are input, the system generates and displays the heat map of the three-dimensional head position data and the color of the heat map represents the magnitude of the

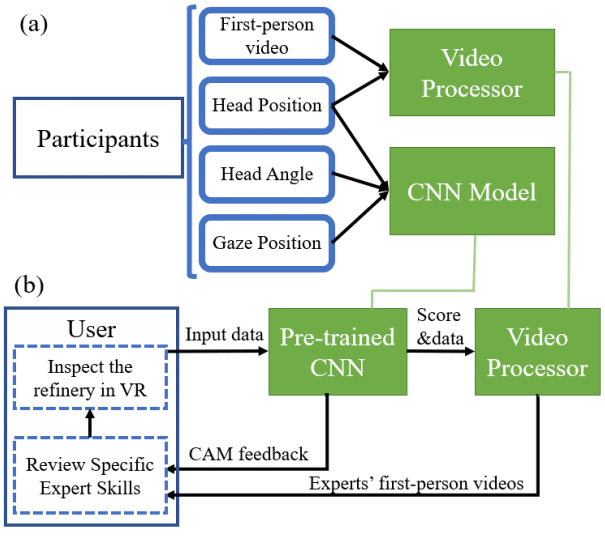

Figure 4.Overview of feedback system (a) Model training and video processing to make feedback (b) Users can get CAM and videos feedback and review the difference and experts' specific inspection skills

contribution to the classification. For large differences between the current data and those of experienced field operators, the color is near red. Therefore, novices can confirm which specific parts of their inspection behavior is different from that of experienced field operators and should therefore be improved. Since this map is presented each time the novices perform, they can improve different behaviors in each case (Figure 5).

Figure 6 shows an example of the use case of the feedback training system constructed in this study. When a novice inspects a drainage area, the inspection is performed from eye level in a standing position (Figure 6 (a)). However, the CAM result points to that behavior as a major contributor to the prediction result of novice (Figure 6 (b)). Thus, the novice can learn that it is possible to improve skills by changing that behavior. This specific knowledge of inspection skills cannot be extracted by conventional statistical analysis techniques; hence, we consider that one of the contributions of this study is the proposition of a new analytical framework that focuses on the specific skills of experienced field operators with high classification accuracy.

Additionally, in this system, novices can also review videos of the corresponding inspection time from the first-person view of the experienced field operators (Figure 6 (c)). In this way, the novice can reflect on what was different between his own movements and those of the expert. Specifically, the novice can know that the expert inspects the drainage area from a lower angle. While the previous study only used the gaze position data for the training system, the method of this study can provide multi-variable feedback including both head position and gaze

TABLE 1 Confusion matrix (LOSO)

		Actual	
		E	N
Predicted	E	59	1
	N	0	29

TABLE 2 Confusion matrix (LOUO)

		Actual	
		E	N
Predicted	E	50	10
	N	0	20

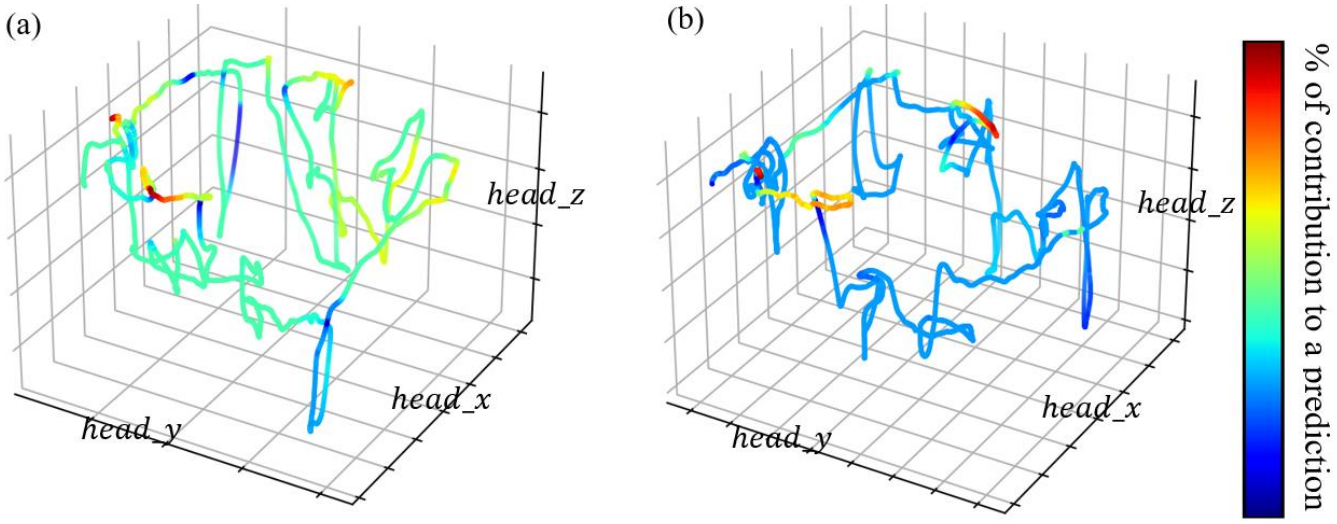

Figure 5. Two examples of results: input data is from the same novice participant (a) the fifth performance (b) the eighth performance.

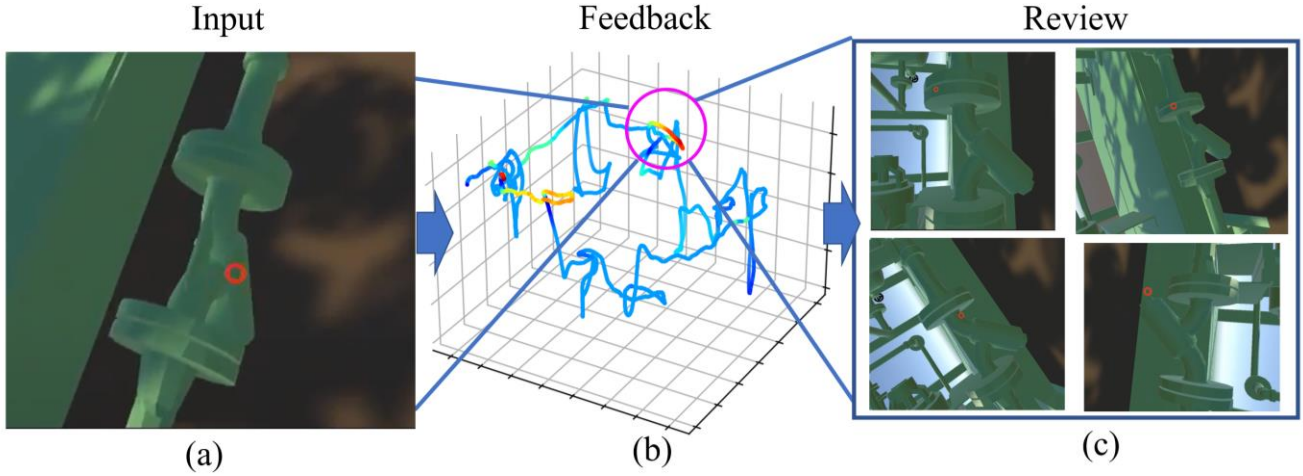

Figure 6. The use case of our feedback system: (a) Novice see the drainage area from eye level (b) Novice can get feedback from Class Activation Map (c) First-person videos of experts show that the experts are looking at the drainage area from a lower perspective.

position variables using XAI for multivariate analysis, which is difficult to achieve using conventional techniques. However, this study has some limitations. First, the number of participants was limited; hence, future work should be done to verify the performance of the CNN with the CAM model using a larger database. Second, based on the differences between the results obtained from LOSO and LOUO, the variance between individuals significantly affects the model learning owing to the small number of participants. Moreover, if the number of experienced field operators increases, we can also extract more common (general) features of their skills. Another limitation is that the validity of the feedback from the CAM was not evaluated. The evaluation methods can be based on both quantitative evaluation metrics, such as deletion and insertion [23], and qualitative feedback from experienced operators.

In addition, we only proposed a new feedback training system with XAI, and thus the training effects of the system must still be verified experimentally. Furthermore, to compare the training effects of the proposed method and the conventional method, we only verified the importance of specific knowledge of experienced skills. Finally, because we used a virtual environment for the experiment and feedback system, the fidelity of the VR system must still be verified in the future.

V. Conclusion

This study aims to construct a virtual reality system for identifying expert skills in a visual inspection task in a refinery using Explainable AI (XAI) techniques. Since most previous studies applied statistical analysis, such as t-tests, to the mean value of the experimental data, there is a lack of specificity in the existing body of research, and it is difficult to provide feedback on the most important part of the collected data to novices. To address the above issue, we introduced a CNN with a Class Activation Map (CAM) technique, a type of XAI, to analyze the experimental data of experienced and novice field operators and identify the most important data contributing to classification of experienced vs. novice behavior for 120 s inspections. Using this method, field operators were classified with a mean accuracy of 99.1%, and the classification criteria were successfully visualized as a heat map for each experimental trial. Based on these results, we proposed a training system for training novices in "experienced" inspection skills. Although several issues need to be addressed in the future, we believe that the new analytical framework for identifying specific expert skills proposed in this study has the potential to solve the specificity problems of this research area and enable effective inspection training.

References

[1] S. Brams et al., "The relationship between gaze behavior, expertise, and performance: A systematic review," *Psychol. Bull.*, vol. 145, no. 10, pp. 980–1027, Oct. 2019.

[2] K. Takayasu, K. Yoshida, T. Mishima, M. Watanabe, T. Matsuda, and H. Kinoshita, "Upper body position analysis of different experience level surgeons during laparoscopic suturing maneuvers using optical motion capture," *Am. J. Surg.*, vol. 217, no. 1, pp. 12‑16, Jan. 2019.

[3] J. Aust, D. Pons, and A. Mitrovic, "Evaluation of Influence Factors on the Visual Inspection Performance of Aircraft Engine Blades," *Aerospace*, vol. 9, no. 1, p. 18, Dec. 2021.

[4] M. H. Papesh, M. C. Hout, J. D. Guevara Pinto, A. Robbins, and A. Lopez, "Eye movements reflect expertise development in hybrid search," *Cogn Res Princ Implic*, vol. 6, no. 1, p. 7, Feb. 2021.

[5] Y. Sterchi, N. Hättenschwiler, and A. Schwaninger, "Detection measures for visual inspection of X-ray images of passenger baggage," *Atten. Percept. Psychophys.*, vol. 81, no. 5, pp. 1297–1311, Jul. 2019.

[6] K. Hirasawa, K. Maeda, T. Ogawa, and M. Haseyama, "A Trial of Fine-grained Classification of Expert-novice Level Using Bio-signals While Inspecting Subway Tunnels," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, Oct. 2021, pp. 204–205.

[7] A. K. Pradhan, K. R. Hammel, R. DeRamus, A. Pollatsek, D. A. Noyce, and D. L. Fisher, "Using eye movements to evaluate effects of driver age on risk perception in a driving simulator," *Hum. Factors*, vol. 47, no. 4, pp. 840–852, Winter 2005.

[8] R. Y. I. Koh, T. Park, C. D. Wickens, L. T. Ong, and S. N. Chia, "Differences in attentional strategies by novice and experienced operating theatre scrub nurses," *J. Exp. Psychol. Appl.*, vol. 17, no. 3, pp. 233–246, Sep. 2011.

[9] R.-J. Dzeng, C.-T. Lin, and Y.-C. Fang, "Using eye-tracker to compare search patterns between experienced and novice workers for site hazard identification," *Saf. Sci.*, vol. 82, pp. 56–67, Feb. 2016.

[10] R. Takamido et al., "Evaluation of expert skills in refinery patrol inspection: visual attention and head positioning behavior," *Heliyon*, vol. 8, no. 12, p. e12117, Dec. 2022.

[11] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey," arXiv [cs.LG], Apr. 02, 2021. [Online]. Available: http://arxiv.org/abs/2104.00950

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, Jul. 2019.

[14] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, "XCM: an explainable convolutional neural network for multivariate time series classification," Sci. China Ser. A Math., vol. 9, no. 23, p. 3137, Dec. 2021.

[15] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 9, pp. 1611–1617, Sep. 2019.

[16] D. Zhang et al., "Automatic microsurgical skill assessment based on cross-domain transfer learning," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4148–4155, Jul. 2020.

[17] S. Kurihara et al., "Skill Extraction of Experienced Oil Refinery Plant OperatorsUsing Virtual Reality System," *Seimitsu kougakukai gakujutu kouennkai kouenn ronnbunn syuuu 2022 nenndo seimitsu kougakukai syunnkitaikai* (*Proceedings of the Journal of the Japan Society of Precision Engineering Annual Conference 2022 in Spring Meeting of the Japan Society for Precision Engineering*), 2022, pp. 622–623.

[18] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 1578–1585.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv [cs.LG], Dec. 22, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," arXiv [cs.CV], pp. 1026–1034, Feb. 06, 2015. Accessed: Jan. 23, 2023. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html

[21] N. Ahmidi et al., "A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2025–2041, Sep. 2017.

[22] S. Sadasivan, J. S. Greenstein, A. K. Gramopadhye, and A. T. Duchowski, "Use of eye movements as feedforward training for a synthetic aircraft inspection task," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Portland, Oregon, USA, Apr. 2005, pp. 141–149.

[23] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," arXiv [cs.CV], Jun. 19, 2018. [Online]. Available: http://arxiv.org/abs/1806.07421