Analysis of Facial Features for Trust Evaluation in Industrial Human-Robot Collaboration

Campagna, Giulio; Chrysostomou, Dimitrios; Rehm, Matthias

Published in:
20th IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO 2024)

DOI (link to publication from Publisher):
10.1109/ARSO60199.2024.10557748

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

Analysis of Facial Features for Trust Evaluation in Industrial Human-Robot Collaboration*

Giulio Campagna[1], Dimitrios Chrysostomou[2], Matthias Rehm[1]

***Abstract*— The advent of Industry 5.0 marks a significant transition towards a collaborative partnership between humans and robots, exploiting their respective capabilities and features to enhance the manufacturing process. This increased cooperation necessitates a secure environment and, in this context, trust becomes a pivotal factor influencing the quality of human-robot interactions. To ensure safety and workload balance, it is essential to have a reliable and timely measure of trust in robots. This study explores the use of facial features to identify potential correlations with human trust levels. To this purpose, a chemical industry scenario was developed where a cobot assisted the human handing over a beaker and pouring chemicals. The analysis employed Deep Learning models, specifically Convolutional Neural Networks (CNNs), to explore the relationship between facial expressions and trust levels. The results of the investigation revealed an accuracy rate of 78.61% for the handing task, and an accuracy of 73.35% for the pouring task. Nevertheless, the findings highlight the importance of implementing sensor fusion algorithms to improve the accuracy and robustness of trust evaluation towards robots.**

I. Introduction

The concept behind Industry 5.0 fosters human-centric manufacturing, driven by technologies that empower workers through improved transparency and collaboration with intelligent machines [1]. Trust and intention prediction play a pivotal role in realizing this vision by enabling fluent interactions between humans and collaborative robots as team members. According to Muir and Moray, trust can be defined as the operator's confidence in the machine's competence, ensuring that the operator believes the system effectively performs its tasks [2]. In this regard, appropriate trust levels facilitate collaboration, while mismatches can lead to under-utilization of automation or hazardous situations with safety implications [3]. Consequently, the ability to assess trust levels during interactions is critical for adapting robot behaviors to user trust and enhancing transparency. However, conventional trust measurement often relies on post-task questionnaires [4], providing only retrospective evaluations of overall trust without capturing real-time trust dynamics [5]. Furthermore, subjective trust perceptions do not always align with actual behavior [6] and can be susceptible to well-known biases, including selective memory, recency effects, and rationalization [7].

To address these limitations, researchers have explored the use of sensors and machine learning techniques to infer trust from behavioral cues. For example, studies have investigated the relationship between trust and physiological signals such as heart rate and skin conductance [8], as well as body movements and gestures [9]. These approaches have shown promising results in predicting trust levels, but they require specialized sensors and complex data processing algorithms.

In recent years, there has been growing interest in using facial expressions as a means of assessing trust in human-robot interactions (HRI) [10]. Facial expressions are a rich source of information about a person's emotional state and have been shown to be closely linked to trust during collaborative tasks [11] and human-robot handover tasks [12].

Convolutional neural networks (CNNs) have emerged as a powerful tool for facial expression recognition and trust assessment in HRI [13]. CNNs are a type of deep learning architecture that are well-suited to facial expression recognition tasks and they have been shown to achieve state-of-the-art performance on various facial expression recognition benchmarks, outperforming traditional machine learning methods [14]. Several studies have applied CNNs to trust assessment and related emotion recognition in HRI. For example, Shen et al. introduced multi-modal feature fusion for clearer understanding of human personality traits [15] and Jaiswal and Nandi built a real-time emotion detection system to explore the range of emotions that could be tracked by facial expressions [16].

Building on our prior work presented in [17], our current study investigates the use of CNNs for real-time trust assessment based on facial expressions in a chemical industrial scenario. Our findings show that the CNN architecture can assess trust levels in real-time, providing valuable insights for adapting robot behaviors to user trust and enhancing transparency. This study presents two primary technical contributions:

- Developing an innovative data-driven framework that integrates facial gestures images with state-of-the-art deep learning algorithms to categorize trust levels using implicit reactions and responses.
- Conducting comprehensive training and robust evaluation of the proposed deep learning framework with annotated computer vision datasets under controlled conditions to map facial gestures to trust levels reliably.

II. Methodology

As previously mentioned in Section I, the envisioned scenario is based on the chemical industry environment developed in [17]. In this framework, a collaborative robot delivered a beaker containing a chemical to the human

* The work described in this paper was funded by the Independent Research Fund Denmark, grant number 1032-00311B.

[1] are with the Technical Faculty of IT and Design, Aalborg University, 9000 Aalborg, Denmark. {gica, matthias}@create.aau.dk.

[2] is with the Faculty of Engineering and Natural Sciences, Aalborg University, 9220 Aalborg, Denmark. dimi@mp.aau.dk.

operator, followed by the robot taking another beaker with a different chemical and pouring its contents into the beaker held by the human. This task involved two distinct levels of robot performance: *high* and *low performance*.

In the high-performance trials, the robot executed the task with seamless precision, following well-planned trajectories, and avoided any non-ergonomic behaviors that might discomfort the user. Conversely, in the low-performance condition, the robot exhibited hazardous trajectories, including coming too close to the human operator and giving the impression of pouring chemicals onto the operator's hand. Consequently, this created an uncomfortable environment, leading to heightened levels of anxiety in the user.

In line with the findings presented in [17], the experiment was designed with two distinct conditions: *high trust* and *low trust*, corresponding to the high and low performance of the robot, respectively. Each participant completed the task four times, twice with the high-performance robot and twice with the low-performance robot. Notably, the robot employed different path planning for each trial, making its actions unpredictable.

As a result, this approach facilitated the collection of facial expression images with automatic labeling, categorizing them as either high or low levels of trust.

A. Experimental Setup

The experimental apparatus concerned the following two principal components: the Universal Robots UR10-CB3-Series Robot[1] and the Azure Kinect DK Camera[2].

The UR10 Robot, featuring six rotational joints, is purpose-built for collaborative tasks with human operators, leveraging its exceptional precision and reliability. It was outfitted with an OnRobot RG6 gripper, which is a versatile two-fingered gripper with the capacity to deliver a stroke of up to 150 mm. While the path planning and grasping phases were pre-determined, the participants were explicitly informed that the robot operated with dynamic and autonomous behavior, thus carrying the possibility of occasional malfunctions. The Azure Kinect DK Camera comprises various components. For the proposed research, its key feature is its 1920x1080 pixel RGB camera capturing 30Hz color video with a 75° horizontal, 65° vertical field of view.

To conclude, both participants and assistants wore laboratory coats, gloves, and safety glasses to ensure their safety. The chemicals used in the experiment were baking powder in the human-held beaker and water in the robot-held beaker. As a result, the only product generated during the reaction phase was carbon dioxide, ensuring a safe experimental environment. The actual chemical composition was revealed at the conclusion of the experiment, in accordance with the participants' prior knowledge of the potential hazards associated with these substances. The setup of the experimental scenario is illustrated in Fig. 1.

[1]https://www.universal-robots.com/cb3/

[2]https://azure.microsoft.com/en-us/products/kinect-dk

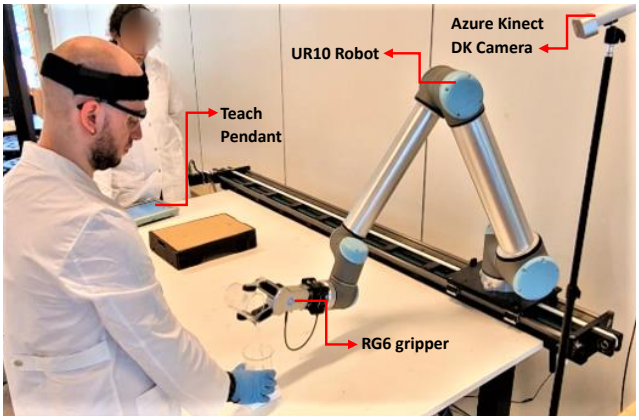


Fig. 1: The setup for the chemical industry scenario.

B. Procedure

A group of 20 participants was involved in the study. Specifically, 10 males and 10 females were selected with diverse age (M=29.1, SD=7.54).

The study adhered to the Declaration of Helsinki and underwent thorough ethical review, gaining approval from the institutional review board. Participants received a printed consent form detailing the research objectives, tasks, methodology, and associated risks prior to the experiment.

Afterwards, the assistant supported the participant in putting on the protective equipment. To ensure an unobstructed view, the camera was appropriately positioned at a suitable height and angle. It was securely mounted and placed at a one-meter distance from the participants. This distance was chosen because increased distance would lead to decreased facial resolution, potentially compromising face matching performance, as elaborated in [18]. Moreover, adequate and consistent lighting was set up, as poor lighting can introduce undesirable elements like noise, shadows, or overexposure.

The participant's involvement spanned a total of 30 minutes, which included the introductory explanation. Over this duration, participants engaged in the tasks four times, thereby experiencing both high and low-performance modes of the robot.

C. Data Collection

Data collection was specifically focused on capturing participants' facial expressions during their reactions for both the high and low robot performance, encompassing both the beaker-handling and pouring stages. The captured 2D RGB images featured 1980x1020 pixels resolution, and were collected at a rate of 30 Hz.

D. Data Pre-Processing

The initial phase revolved around deriving the participants' facial expressions from the original RGB images, a task that demanded the utilization of a *face detection* algorithm. To this end, the pre-trained **Multi-Task Cascaded Convolutional Networks** (MTCNN) model was utilized. The

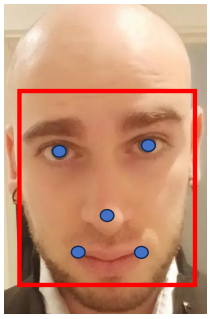


Fig. 2: The facial landmarks in each individual bounding box.

architecture of MTCNN is described in [19]. Each image is resized to various scales to construct an image pyramid, serving as the input for the subsequent three-stage cascaded network, as described in the following. In the first phase (*The Proposal Network*), a fully convolutional network (FCN) is employed to generate a set of candidate face regions (bounding boxes) that may possibly enclose faces. This stage efficiently filters out non-face regions, reducing computational load in subsequent stages. The second stage (*The Refine Network*) concerns the refinement of the previously generated bounding boxes. A CNN is applied to eliminate false positives and optimize the size and positioning of the bounding boxes for improved alignment with the actual faces in the image. Lastly, in the third phase (*The Output Network*), *facial landmark detection* is conducted. This stage identifies key facial landmarks, including the eyes, nose, and corners of the mouth (both left and right), within each bounding box (refer to Fig. 2). MTCNN was selected for face detection for several reasons, including its exceptional accuracy in identifying faces within images, resilience in handling diverse lighting conditions, its multi-stage approach for refining face detection, efficient computation, and its outstanding generalization ability across various face sizes.

Using the bounding box positions, the regions containing the detected faces were extracted, resulting in cropped face images. To ensure a balanced trade-off between computational efficiency, information preservation, and model performance, the images were resized to 200x200 pixels using *bilinear interpolation* — a resampling method that facilitates smooth resizing. The initial dataset comprised 9849 samples for the handing task and 7944 samples for the pouring task. The images were labeled as *high trust* or *low trust* depending on the robot's performance in the several trials. Images collected during high-performance trials were labeled as *high trust*, while those from low-performance trials were labeled as *low trust*. Subsequently, the images underwent standardization. Given that the dataset size for both tasks was relatively small, employing *data augmentation* techniques was crucial for enhancing the model's performance. As a result, online data augmentation methods were applied, encompassing, for example, flips and rotations (ranging from 0 to 40 degrees) to mitigate overfitting. Choosing the right techniques is crucial, as an improper selection may lead to the loss of facial details (e.g., extreme zooming). The augmented images were also standardized for consistency.

III. Experimental Results

The subsequent analysis delves into investigating the correlation between facial gestures and the trust levels of human operators, framed as a binary classification problem: *high trust* and *low trust*.

Given their inherent capability to capture spatial patterns within images, **2D-CNNs** stand as a well-suited choice for the analysis of facial gesture images. The deep learning algorithms were implemented on Tensorflow platform. In the context of both the Handing and Pouring datasets, the data were divided into training, test, and validation sets with participant-based splitting approach. Specifically, the training set encompassed data from 12 participants, constituting 60% of the dataset, while the test set was composed of data from 4 participants, representing 20%, and the remaining 4 participants were allocated to the validation set, also making up 20%. This partitioning strategy ensured the model's evaluation on unseen data. In the following, the three different architectures used for 2D-CNNs are presented with the relative classification results. To improve the training efficiency and convergence of the models, two optimization algorithms were employed - Adaptive Moment Estimation (*Adam*) and Stochastic Gradient Descent (*SGD*). Regarding Adam, a *learning rate* of 0.001 was utilized to regulate the step size of weight updates, while *beta 1* and *beta 2* parameters were set to 0.9 and 0.999, respectively. These parameter values influenced the exponential decay rates for the first and second moments of gradients, thereby enhancing the model's ability to adapt its learning rates dynamically. Regarding SGD, a *learning rate* of 0.001 was utilized, and a *momentum factor* of 0.9 was added to exploit previous gradients, facilitating faster convergence. The 2D-CNN models were compiled with the *binary cross-entropy loss function* and underwent training for 20 epochs, with a batch size of 32. The *input* of the model was a 4D tensor (*batch size*, *height*, *width*, *depth*) where the last three dimensions represented the dimensions of the image 200x200x3 (200x200 pixels and 3 channels, i.e red, green and blue).

With reference to **Architecture 1**, a 2D-CNN was implemented with a series of layers designed to process and extract features from the input data. To begin with, a *convolutional layer* was used, employing *32 filters*, each with a *3x3 kernel size*, and Rectified Linear Unit (*ReLU*) activation. Following this, a *max-pooling layer* with a *2x2 pooling window* was applied to reduce spatial dimensions and ease computational load. To prevent overfitting, a *dropout layer* was introduced with a *0.25 dropout rate*, randomly deactivating 25% of neuron outputs during training. The network then incorporated another *convolutional layer* with *64 filters* and a *3x3 kernel*, followed by *ReLU activation* to capture more complex patterns in the data. Subsequently, a second *max-pooling layer* with a *2x2 pooling window* was employed. Another *dropout layer* with a *0.25 rate* further enhanced

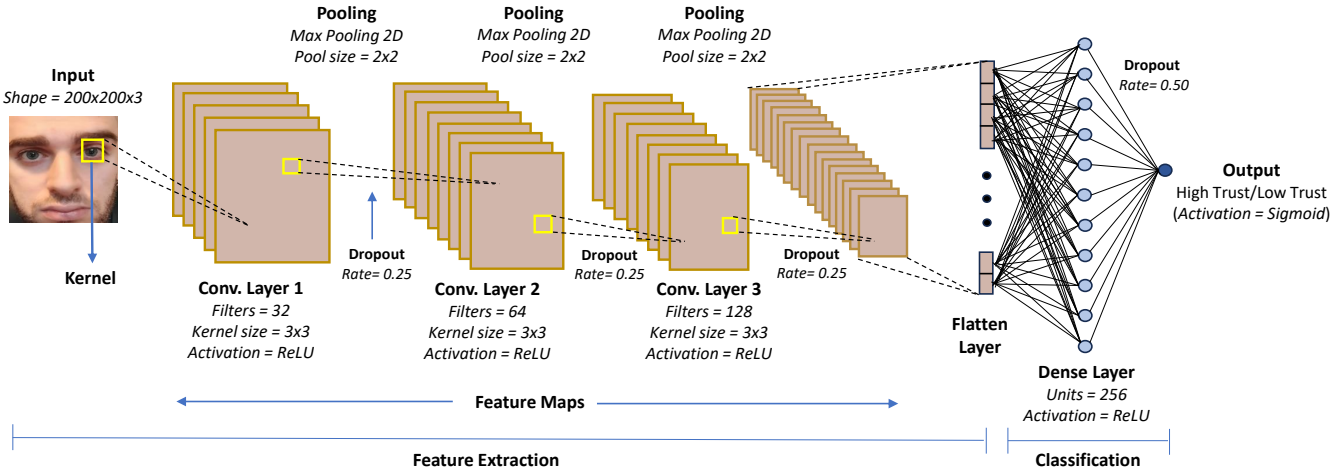


Fig. 3: The 2D-CNN Architecture 3 that provided the best accuracy solution for both handing and pouring analysis.

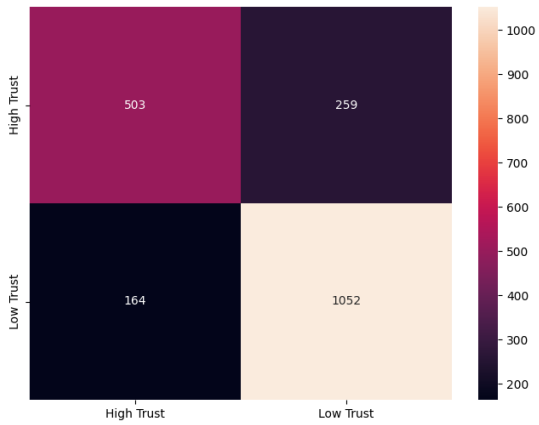
regularization. The feature maps were flattened into a one-dimensional vector using a *flatten layer*, serving as input for subsequent fully connected layers. A *dense layer* with *128 units* and *ReLU activation* was introduced to extract higher-level features. A *dropout layer* with a *0.5 rate* was included before the final *output layer*, which consisted of a *single neuron* with *sigmoid activation*. The results demonstrated an accuracy of 65.77% using Adam and 71.59% with SGD for the handing task. In the case of pouring, the 2D-CNN model achieved an accuracy of 66.53% with Adam, while it yielded 71.48% with SGD.

Concerning **Architecture 2**, the following model was developed. To begin with, a *convolutional layer* was characterized by *64 filters*, *3x3 pixel kernel*, and *ReLU activation* function. Subsequently, a *max-pooling layer* with a *2x2 pooling window* was utilized. A *dropout layer* was used with a *dropout rate* of *0.25* to mitigate overfitting. Then, the model consisted of an additional *convolutional layer*, featuring *128 filters* utilizing a *3x3 kernel* and the *ReLU activation* function. Subsequently, spatial dimensions were reduced by a *max-pooling layer* with a *2x2 pooling window*. A *dropout layer* with a *dropout rate* of *0.25*, was employed. Continuing with the description of the model, the feature maps underwent transformation into a one-dimensional vector through a *flatten layer*, preparing them for integration into the subsequent fully connected layers. Consequently, a *dense layer*, with *256 units* and utilizing *ReLU activation*, was incorporated. Before the final output layer, a *dropout layer* with a *0.5 rate* was introduced, further intensifying the regularization efforts. To conclude, the output layer comprised a *single neuron* using *sigmoid activation*, which yielded the probability output for the class. The accuracies were observed as 54.96% with Adam and 69.46% with SGD for the Handing analysis. In the case of Pouring, the accuracy reached 63.02% with Adam and 73.30% with SGD.

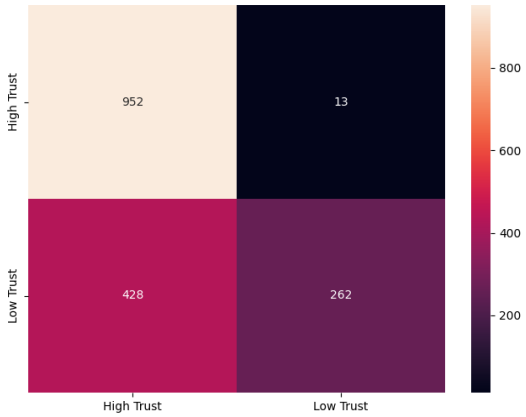
The **Architecture 3** comprised *three convolutional layers* with varying parameters, each followed by a max-pooling layer and dropout layer. For the convolutional layers, the parameters included the number of *filters 32*, *64*, and *128*, the *3x3 kernel size* for each, and the *activation function ReLU*. The *max-pooling layers* consistently used a *2x2 pooling window* size. The *dropout layers* were set with a *dropout rate* of *0.25*. After the three repetitions of the pattern (Conv2D, MaxPooling2D, and dropout layer), a *flatten layer* prepared the data for fully connected layers. The following *dense layer* had *256 units* and used *ReLU activation*, while the last *dropout layer*, just before the final output layer, had a *0.5 dropout rate*. The architecture concluded with a *dense layer* containing a *single neuron* with *sigmoid activation*, making it suitable for binary classification tasks. With this architecture, an accuracy of 58.20% was achieved with Adam and 78.61% with SGD for Handing. Regarding Pouring analysis, the accuracy reached 58.01% with Adam and 73.35% with SGD. Fig. 3 illustrates the *Architecture 3*, which, among all the architectures, achieved the highest accuracy for both handing and pouring task.

IV. Discussion

The analysis delved in the investigation of facial gestures as expression of the trust level of the human operator towards the robot's capabilities. Facial gestures, a form of non-verbal communication, offer valuable insights into the user's emotional state towards the robot and can provide real-time feedback regarding trust levels during the interaction. This would enable prompt adjustments and adaptations in the robot's behavior or decision-making to maintain a safe environment and balance workload. 2D-CNNs were employed to explore the relationship between trust categorization and facial gestures. In the following, the discussion of performance metrics results is presented. In Table I, a comprehensive summary of the results for the Handing task can be found, while Table II contains the summary for the Pouring task. Concerning the description about the different architectures, refer to Section III.



(a) Handing.



(b) Pouring.

Fig. 4: Confusion Matrix Analysis (Architecture 3 with SGD optimizer).

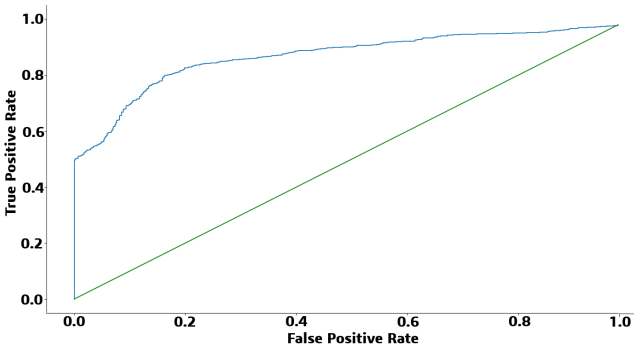


Fig. 5: AUC-ROC curve for handing task.

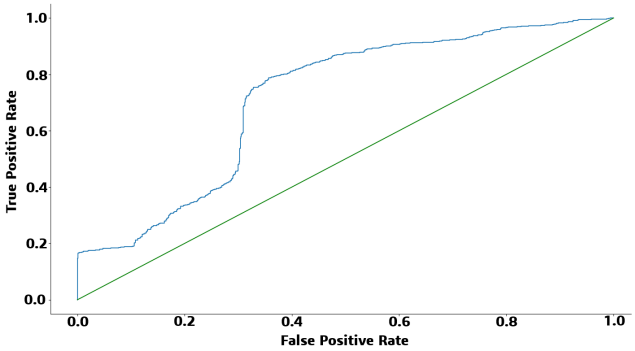


Fig. 6: AUC-ROC curve for pouring task.

To begin with, 2D-CNN (*Architecture 3* with *SGD* optimizer) attained a 78.61% accuracy for the Handing task, and 73.35% for the Pouring task. The SGD optimizer likely delivered improved results over Adam thanks to its inherent stochasticity that can help it escape local minima, making it a robust optimization algorithm. The confusion matrices are depicted for both Handing (Fig. 4a) and Pouring (Fig. 4b) considering the architecture that yielded the highest accuracy. To further evaluate the classification performance, *Receiver Operating Characteristic (ROC)* curves plotting the true positive rate against false positive rate were generated for each architecture. The area under the curve (*AUC*) provides a comprehensive measure of classification ability across different thresholds. For the handing task, *Architecture 3* with *SGD* demonstrated an AUC of 0.87 while it achieved an AUC of 0.72 for the pouring task. Fig. 5 and Fig. 6 depict the AUC-ROC curve for the handing and pouring tasks, respectively.

Analyzing the findings, it is evident that the participants exhibited susceptibility to emotional responses conveyed through their facial expressions when engaged in both handing and pouring task. The emotional expressiveness observed may be attributed to the heightened risk perception of potential harm. This risk is associated with the possibility of a collision with the robot during the handover task and exposure to chemicals during the pouring task. Nevertheless, the manifestation of facial expressions was not flawless. It is important to note that not all participants demonstrated maximum expressiveness. One contributing factor could be the potential influence of the physical environment within the experiment's framework. Since the study took place in a laboratory environment under the observation of researchers, it is plausible that participants demonstrated greater confidence levels than they would in real-world situations. Furthermore, participants' diverse cultural backgrounds may have influenced their responses to the robot's capabilities, subsequently affecting their trust levels and facial expressions. Throughout the experiment, a subset of participants exhibited a lack of facial expressions, irrespective of the robot's performance. Enhancing emotional responses could be improved by introducing riskier robot movements, accompanied by vocal feedback that underscores the potential for collisions, thereby likely provoking more pronounced psychological reactions from the participants.

Trust has an intricate nature, highlighting the need for a more robust analysis to capture the full spectrum of its expressions. While facial gestures can offer insights into trust, incorporating additional sensors alongside facial gestures has

TABLE I: Performance metrics relative to the three different architectures for Handing analysis.

Optimizer	Arch.	Accuracy	AUC	Precision	Recall	F1-score
Adam	Arch. 1	65.77%	0.72	0.43	0.66	0.52
	Arch. 2	54.96%	0.59	0.54	0.55	0.54
	Arch. 3	58.20%	0.61	0.56	0.58	0.57
SGD	Arch. 1	71.59%	0.80	0.70	0.66	0.65
	Arch. 2	69.46%	0.82	0.71	0.69	0.66
	Arch. 3	78.61%	0.87	0.78	0.79	0.78

TABLE II: Performance metrics relative to the three different architectures for Pouring analysis.

Optimizer	Arch.	Accuracy	AUC	Precision	Recall	F1-score
Adam	Arch. 1	66.53%	0.67	0.72	0.67	0.61
	Arch. 2	63.02%	0.71	0.72	0.63	0.54
	Arch. 3	58.01%	0.68	0.45	0.58	0.43
SGD	Arch. 1	71.48%	0.65	0.72	0.71	0.70
	Arch. 2	73.30%	0.67	0.80	0.74	0.70
	Arch. 3	73.35%	0.72	0.80	0.73	0.70

the potential to enrich the understanding of trust, especially considering its intricate and multifaceted nature.

V. Conclusion

The study aimed to explore the link between trust and facial expressions of human operators while interacting with robots. In the presented research, it was simulated a chemical industry setting where a robotic arm assisted users by delivering and dispensing chemicals into a beaker. The methodology employed to investigate the correlation between facial gestures and trust levels involved the utilization of 2D-CNNs. The obtained accuracy resulted in 78.61% for the handover task and 73.35% for the pouring task. However, facial gestures alone is not sufficient in delivering the requisite precision for accurately categorizing trust, primarily due to their inherent complexity and susceptibility to misinterpretation. As result, a more advanced approach is imperative, which involves the implementation of sensor fusion algorithms. Leveraging data from diverse sensors has the potential to significantly augment the detection of a proper trust response from human operators. Consequently, in future endeavors, we will focus on the development of sensor fusion-based models designed to categorize trust efficiently and conduct a more nuanced analysis of trust evolution over time. The ultimate goal is to enable the robot to adapt its behavior in response to the current trust level, thereby ensuring both safety and workload balance.

ACKNOWLEDGMENT

The authors of the article would like to express their gratitude to Mahed Dadgostar, a student from Aalborg University, for his dedication and support in this research project.

References

[1] J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis, and L. Wang, "Industry 5.0: Prospect and retrospect," *Journal of Manufacturing Systems*, vol. 65, pp. 279–295, 2022.

[2] B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[3] K. Hald, M. Rehm, and T. B. Moeslund, "Human-robot trust assessment using top-down visual tracking after robot task execution mistakes," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 892–898.

[4] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, 2016.

[5] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in human-robot interaction*. Elsevier, 2021, pp. 3–25.

[6] M. Adamik, K. Dudzinska, A. J. Herskind, and M. Rehm, "The difference between trust measurement and behavior: Investigating the effect of personalizing a robot's appearance on trust in hri," in *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 880–885.

[7] B. Leichtmann, V. Nitsch, and M. Mara, "Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement," *Frontiers in Robotics and AI*, vol. 9, p. 838116, 2022.

[8] K. Hald, M. Rehmn, and T. B. Moeslund, "Human-robot trust assessment using motion tracking & galvanic skin response," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 6282–6287.

[9] L. Onnasch and C. L. Hildebrandt, "Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 1, pp. 1–24, 2021.

[10] L. Cominelli, F. Feri, R. Garofalo, C. Giannetti, M. A. Meléndez-Jiménez, A. Greco, M. Nardelli, E. P. Scilingo, and O. Kirchkamp, "Promises and trust in human–robot interaction," *Scientific reports*, vol. 11, no. 1, p. 9687, 2021.

[11] W. Mou, M. Ruocco, D. Zanatto, and A. Cangelosi, "When would you trust a robot? a study on trust and theory of mind in human-robot interactions," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 956–962.

[12] T. Faibish, A. Kshirsagar, G. Hoffman, and Y. Edan, "Human preferences for robot eye gaze in human-to-robot handovers," *International Journal of Social Robotics*, vol. 14, no. 4, pp. 995–1012, 2022.

[13] H. Zhu, C. Yu, and A. Cangelosi, "Explainable emotion recognition for trustworthy human–robot interaction," in *Proc. Workshop Context-Awareness Hum.-Robot Interact. Approaches Challenges ACM/IEEE HRI, Sapporo, Japan*, 2022.

[14] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[15] Z. Shen, A. Elibol, and N. Y. Chong, "Multi-modal feature fusion for better understanding of human personality traits in social human–robot interaction," *Robotics and Autonomous Systems*, vol. 146, p. 103874, 2021.

[16] S. Jaiswal and G. C. Nandi, "Robust real-time emotion detection system using cnn architecture," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11 253–11 262, 2020.

[17] G. Campagna and M. Rehm, "Analysis of proximity and risk for trust evaluation in human-robot collaboration," in *32nd IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2023.

[18] U. Park, H.-C. Choi, A. K. Jain, and S.-W. Lee, "Face tracking and recognition at a distance: A coaxial and concentric ptz camera system," *IEEE transactions on information forensics and security*, vol. 8, no. 10, pp. 1665–1677, 2013.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.