HAND GESTURE RECOGNITION FOR HUMAN-ROBOT COOPERATION IN MANUFACTURING APPLICATIONS

Stanislaw Hozyn
Faculty of Mechanical and Electrical Engineering
Polish Naval Academy, Smidowicza Street 69, Gdynia 81-127, Poland
E-mail: s.hozyn@amw.gdynia.pl

KEYWORDS
human-robot cooperation, gesture recognition, computer vision, machine learning, deep learning, convolutional neural network.

ABSTRACT

Human-robot cooperation plays an increasingly important role in manufacturing applications. Together, humans and robots display an exceptional skill level that neither can achieve independently. For such cooperation, hand gesture communication using computer vision has been proven to be the most suitable due to the low cost of implementation and flexibility. Therefore, this work focuses on the hand gesture classification problem in view of human and robot collaboration. To facilitate collaboration, six of the most common gestures applicable in manufacturing applications were selected. The first part of the research was devoted to creating an image dataset using the proposed acquisition system. Then, pre-trained neural networks were designed and tested. In this step, the feature extraction approach was adopted, which utilises the representations learned by a previous network to extract meaningful features. The results suggest that all developed pre-trained networks attained high accuracy (above 98,9%). Among them, the VGG19 demonstrated the best performance, achieving accuracy equal to 99,63%. The proposed approach can be easily adapted to recognise a more extensive or different set of gestures. Utilising the proposed vision system and the developed neural network architectures, the adaptation demands only acquiring a set of images and retraining the developed networks.

INTRODUCTION

Hand gesture recognition constitutes an essential field of nonverbal communication that can be applied in many areas, such as clinical and health, sign language, robotic control, home automation or augmented reality (Oudah et al., 2020). In the beginning, hand gesture recognition was realised using various sensors attached to the glove, for example, curvature sensors, angular displacement sensors, flex sensors, optical fibre transducers or accelerometers sensors. Those sensors detected hand movements or finger bendings, further analysed by a computer connected to the glove with wires. Even though such systems achieved high accuracy, they were characterised by many limitations relevant to a prohibitively expensive manufacturing cost and a low comfort of use. These flaws resulted in the development of cost-effective and convenient techniques based on computer vision methods (Xia et al., 2019).

Traditional computer vision utilises carefully hand-designed features to understand the nature of an analysed scene. Many algorithms were developed to design features using edge detection (Żak & Hożyń, 2013b), image future detection (Bay et al., 2006), texture recognition (Hożyń, 2021), image segmentation (Zalewski & Hożyń, 2020), or particle image velocity (Piskur, 2022). They were deployed in various applications, such as robotics systems (Hożyń, 2020), autonomous surface vehicles (Praczyk et al., 2019), underwater platforms (Jurczyk et al., 2020), obstacle detection (Kot, 2022) or stereo vision (Żak & Hożyń, 2013a). The difficulty with this approach is that it demands long trial-and-error processes to select the most promising feature, which the computer vision engineer must next tune.

Deep Learning (DL) introduced a new approach where the neural network is responsible for discovering patterns in analysed images and automatically finding the most descriptive features. This approach allowed AlexNet (Krizhevsky et al., 2017), constructed by Krizhevsky et al., to win the ImageNet Large Scale Visual Recognition Challenge in 2012, proving that automatically learned features can exceed manually designed ones. After that, some disruptive network architectures, such as GoogleNet/InceptionV1 to V4 (Szegedy et al., 2015), ResNet (He et al., 2016), VGG16 or VGG19 (Simonyan & Zisserman, 2015) were proposed. They were trained to distinguish 1000 classes using ImageNet dataset (Jia Deng et al., 2009), which includes 1.5 million images divided into 20 000 categories. The obtained Top-1 accuracies of the above methods are between 71% and 79%, constituting outstanding results and making them especially suitable as pre-trained models in many computer vision applications.

Pre-trained models establish a convenient approach to deep learning. They are previously trained on large datasets, which makes their spatial hierarchy of features effective as a generic model for various classification tasks. Two ways of deploying pre-trained models can be distinguished: feature extraction and fine-tuning. Feature extraction utilises previously trained convolutional layers as a feature detector. In this instance, the dense layers are only modified to develop a new classifier. Fine-tuning, apart from changing the classifier, also trains some convolutional layers that enable them to recognise a modified set of features (Szymak et al., 2020). Pre-trained models have been utilised in many manufacturing

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

applications, such as object detection (Liu & Liu, 2018), object classification (Hożyń, 2023), system monitoring (Deng et al., 2020) or fault diagnosis (Di Maggio, 2022). Hand gesture recognition based on computer vision constitutes a very promising approach in manufacturing applications (Neto et al., 2019). In (Sheikholeslami et al., 2017) the authors explored the efficacy of robot hand configuration in expressing instructional gestures for human-robot interaction. They addressed recognition confidence measures for the gestures that humans and robots express using different hand configurations. Multimodal data fusion and multiscale parallel convolutional neural networks for human-robot interaction were presented by Gao et al. (Gao et al., 2021). The devised method was applied to a seven-degree-of-freedom bionic manipulator to achieve robotic manipulation with hand gestures.

A real-time human-robot interaction framework based on hand gesture detection was presented in (Mazhar et al., 2019). The framework facilitated a real-time safe human-robot collaboration using static hand gestures and 3D skeleton extraction. Nuzzi et al. (Nuzzi et al., 2019) utilised a Faster R-CNN object detector to find the accurate position of the hands in RGB images in view of a smart hand gesture recognition set up for Collaborative Robots. They acquired four different small datasets with different characteristics to evaluate the performances in various situations. The developed system achieved good performances that could lead to real-time human-robot interaction with a low inference time.

Even though some researchers have addressed the hand gesture recognition problem, this task is always closely related to the character of the individual implementation. It means that the number of gestures, meanings, and image acquisition conditions should always be adjusted to the application's demands. Therefore, the motivation for the present work was to devise a flexible and accurate approach to hand gesture recognition for human-robot cooperation in manufacturing applications. To achieve it, a vision system suitable for industrial applications was designed. Using this system, the database comprising six gestures was created. Then, the baseline model structure was proposed. The baseline model allows evaluation of the created dataset and constitutes a benchmark for more complex models. Finally, the most promising pre-trained models were selected and tested using feature extraction. For this purpose, five structures of dense classifiers were devised and evaluated.

The proposed solution differs from the previous approaches since it focuses on both: hand detection and database creation. It facilities a convenient way of various dataset development and modification as well as a straightforward detector development based on pre-trained neural networks. In this way, it delivers a simple methodology which can be easily adopted in similiar applications.The encouraging results demonstrate that the devised method classifies the specified gestures with high accuracy.

The paper is structured as follows: Section 2 summarises the proposed method. In Section 3, the attained results for different network structures are presented, and in Section 4, the obtained results are concluded.

METHODS

The research goal was to devise a gesture recognition method for human-robot cooperation. For this purpose, the technique based on image classification using neural networks was adopted. It demands an extensive image database of exemplary gestures for training, validation and testing neural networks. Therefore, the vision system was designed to acquire the necessary images in the first step. Then, the database was created, and a baseline model of a convolution neural network was generated. The baseline model served as a starting point for designing neural network architectures and testing the usability of the organised image dataset. Those steps enabled to design of convolutional neural network architectures. Their programming implementation was executed utilising publicity available TensorFlow 2 and Keras library.

Vision System

The developed vision system comprises the Alvium 1800 U-050 industrial camera (Fig. 1) and a PC-based image processing unit. It enables acquiring images at 117 frames per second at 0.5 MPixs resolution. The PC-based image processing system utilises Intel Core i7-6700HQ CPU 2.6 GHz and 32 GB RAM. Its programming implementation employs Vimba C++ API, delivered by Allied Vision Company. The OpenCV and Qt libraries were also adopted for image processing and graphical user interface (GUI) development.

Figure 1: Alvium 1800 U-050 industrial camera

Database description

The dataset was created using the designed vision system. It incorporates 4483 images divided into seven classes: OK (2), STOP (5), FORWARD (0), BACKWARD (6), RIGHT (1), LEFT (4), and NOGESTURE (3) (Fig. 2).

Figure 2: Examples of the classified gestures

Since the primary assumption was that the camera could be located at various distances from a human operator in the range of 0.5 to 3.5 meters, the images were divided into three folders: 0.5-1.5 meters, 1.5-2.5 meters and 2.5-3.5meters. It was to ensure uniform distribution of the images in the train, validation and test sets. Irregular distribution could affect the training process since, for example, a more significant number of images from shorter distances could be located in a train set and, consequently, a smaller number in the validation sets.
The dataset was divided into train, validation and test sets for each folder separately. The adopted procedure assumed that 80% of the dataset is devoted to training/validation steps, while 20% to a test step. The training/validation dataset was divided into 75% training data and 25% validation data.

Baseline Model

A baseline model was established to evaluate the created dataset and to provide a reference point for developing more complicated neural network structures. It represents the most straightforward architecture that achieves statistical power. In the presented approach, the baseline model consists of two convolutional layers and one dense layer (Fig. 3).

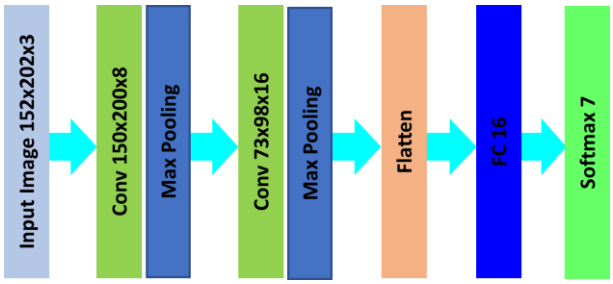

Figure 3: Baseline model

The convolutional layers are followed by pooling layers to downsample feature maps allowing successive convolutions to analyse increasingly more oversized windows. After the convolutional block, a flatten layer is inserted that enables employing dense layers, which output signals are utilised by a final softmax layer. A detailed description of the adopted architecture is depicted in Table 1.

Table 1: Baseline model architecture

Input (152 x 202 RGB image)
conv3-8 *maxpool (2x2)*
conv3-16 *maxpool (2x2)*
FC-16
Softmax-7

The size of output images from the camera is 608x808x3 pixels (in RGB format). It is too large to process in neural networks; therefore, it was reduced to 152x202x3 pixels. Additionally, the pixel values were normalised between 0 and 1.
The first experiments showed that the model overfits and presents a high variance. To mitigate this effect, data augmentation was considered. For this purpose, an in-place data augmentation was deployed, ensuring the model does not process the same images at the training time. Based on the carried-out tests, the following data augmentation coefficients were established and utilised for further experiments:

- Rotation range = 25;
- Width shift range = 0.2;
- Height shift range = 0.2;
- Shear range = 0.1;
- Zoom range = 0.1;
- Horizontal trip = True;
- Vertical flip = True.

The baseline model was evaluated by deploying classification accuracy since each class contains a similar number of images. Therefore, the learning curves were considered to analyse the performance of tested networks. In Figures 4 and 5, the accuracy and loss curves for the baseline model are presented. They suggest that the model can learn patterns from the provided

dataset and that the adopted network architecture is appropriate for the designed classification task.

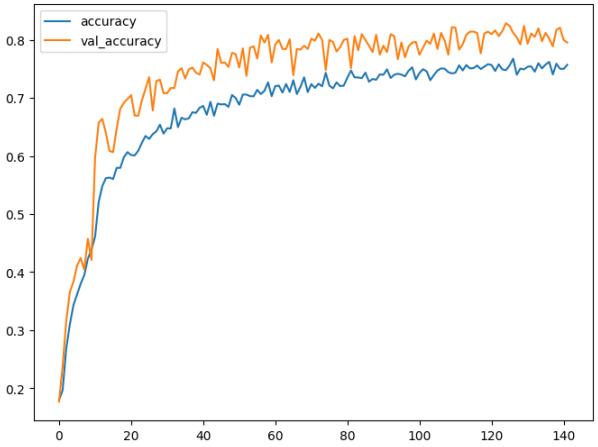

Figure 4: Accuracy of the baseline model

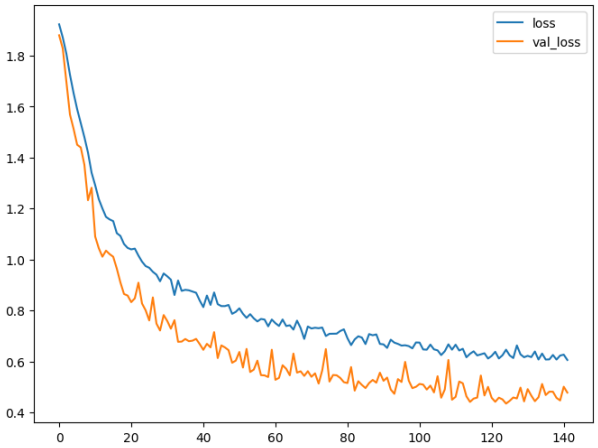

Figure 5: Loss of the baseline model

The classification results are summarised in a confusion matrix (Fig. 6). They suggest that the baseline model could not adequately differentiate mostly forward and left gestures as well as stop and forward ones. It stems from the convolutional base being too simple to discern enough useful features. The classifier is also uncomplicated and comprises only one small convolutional layer which hinders the correct interpretation of patterns in the images.

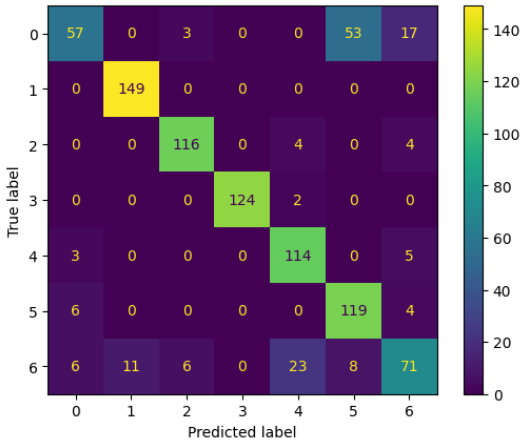

Figure 6: Confusion matrix of baseline model performance

Pre-trained Networks

The following pre-trained networks were adopted for the experiments:

- VGG16;
- VGG19;
- ResNet50;
- Xception;
- InceptionV3.

Their convolutional layers were utilised while the dense layers were replaced with new classifiers. To find the most suitable classifier, five architectures were designed and tested for each network (Table 2).

Table 2: Tested classifiers

Class. 1	Class. 2	Class. 3	Class. 4	Class. 5
FC-16	FC-32	FC-64	FC-128	FC-256
FC-16	FC-32	FC-64	FC-128	FC-256

The deployment of pre-trained networks demanded image normalisation. Therefore, the image processing step was utilised. It facilitated image conversion between RGB and BGR colour spaces and set the pixel values between the -1 and 1 range.

RESULTS

To validate the proposed models, numerous experiments were performed. Firstly, the pre-trained models with the defined classifiers were evaluated utilising learning curves and accuracy rates. Then, comparative analyses of the most promising model were conducted to assign the best architecture for the gesture classification task.
Since the deep learning models are characterised by their stochastic nature, each network was trained and validated ten times for different train and validation splits. Additionally, checkpoint and early stopping mechanisms were deployed to reduce training time. They facilitated saving networks' weights for the highest achieved accuracy and stopping the learning procedure if the

network did not improve its performance during fifteen epochs.
The results of the first part of the experiments suggest that most networks can learn using even the simplest classifier (Class. 1). In this case, only VGG16 and InceptionV3 did not distinguish patterns in the images (see Table 3).

Table 3: Accuracy of the compared models

	Class. 1	Class. 2	Class. 3	Class. 4	Class. 5
VGG16	0,165	0,512	0,848	0,981	**0,987**
VGG19	0,988	0,992	**0,994**	0,988	0,991
ResNet50	0,657	0,987	0,982	**0,991**	0,979
Xception	0,973	0,982	0,981	0,981	**0,992**
InceptionV3	0,165	0,96	0,972	**0,984**	0,983

The higher complexity of the models slightly increases the performance. Only the models with small accuracy for simpler classifiers exhibited rising performance (VGG16, InceptionV3). However, they finally achieved a lower accuracy than VGG19 for the medium complex classifier. Generally, VGG19 reached the highest accuracy (0,994) for Classifier 3.
The obtained learning curves for this model are presented in Figures 7 and 8.

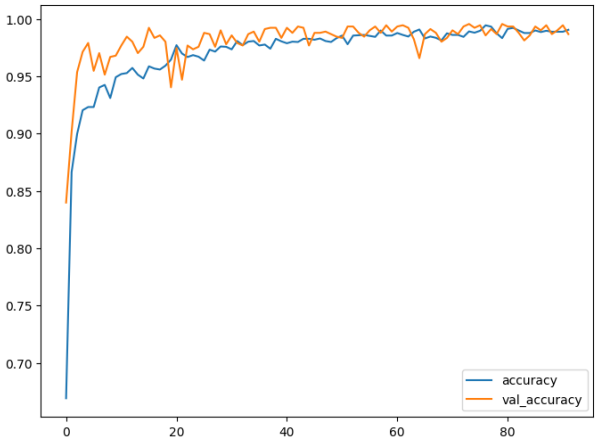

Figure 7: Accuracy of VGG19 (Class. 3)

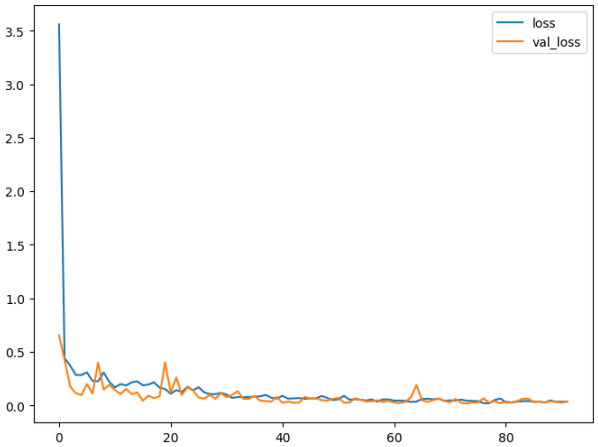

Figure 8: Loss of VGG19 (Class. 3)

They suggest that overfitting and variance are not present during the learning process and that the validation images are easier to analyse by the network. This is because data augmentation is only implemented for training input data. Consequently, since each image during the training stage is slightly distorted, while the validation stage remains images unchanged, the validation data is easier to interpret.
Based on the results from the first step, the second step was devoted to finding the most promising structure among the distinctive networks. Therefore, the following models were considered for further research: VGG16 (Class. 5), VGG19 (Class. 3), ResNet50 (Class. 5), Xception (Class. 5) and InceptionV3 (Class. 5). Each model was trained three times using the whole data (training and validation sets) and assessed on the test set. Apart from accuracy, the training and execution times were also considered. The training time was measured during the training step, while the execution time was calculated as the time needed to predict the classes for all images in the training dataset (3578 images).

Table 4: Performance of the compared models

	Accuracy	Training time (s)	Execution time (s)
VGG16	0,993	2200	8,48
VGG19	0,996	2220	10,54
ResNet50	0,991	1886	7,23
Xception	0,991	3616	16,49
InceptionV3	0,989	3480	15,32

The results indicate that all the pre-trained models highly accurately classify hand gestures (above 0,989). The best one, VGG19, achieved an accuracy equal to 0,996 (Table 4). The fastest network was ResnetNet50 attaining 1886 seconds during training and 7,23 seconds during the execution test. The confusion matrix of the VGG19 performance shows that the network incorrectly classified only four gestures from the test set (Fig. 9).

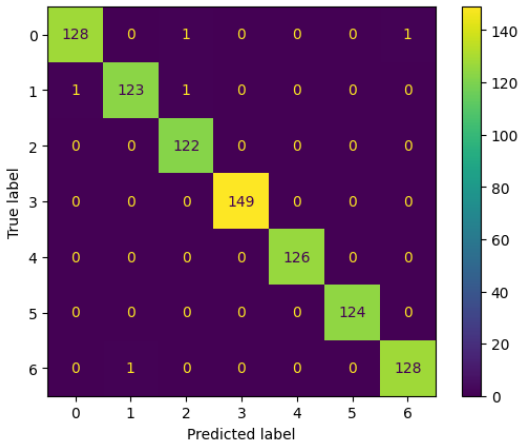

Figure 9: Confusion matrix of the VGG19 performance

CONCLUSIONS

This article has addressed the problem of hand gesture recognition for human-robot cooperation. It describes all necessary steps, such as image acquisition, database creation and neural network development.

The image acquisition system was designed in the first step of the experiments. Based on it, the database was created. The database was evaluated using the baseline model, which also served as a benchmark for more complicated pre-trained neural network structures.

As pre-trained networks, the most promising architectures were selected. Then, their classifiers were replaced with the proposed ones, and the new structures were trained. Based on the obtained results, the best architectures were appointed for the final comparison. It showed that the VGG19 model achieved the highest accuracy, equal to 0,996. The other models obtained slightly lower accuracy. Among the tested networks, the fastest was ResNet50 which executed almost twice the time faster as the slowest InceptionV3.

The proposed solution can be easily deployed in similar applications devoted to hand gesture recognition in industrial applications. However, it is only limited to gesture classification. In some cases, the applications should also be capable of localising a human operator's hand. For this purpose, object detection methods should be employed. Therefore, future work will focus on applying object detection techniques to human-robot cooperation based on hand gesture recognition.

REFERENCES

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded Up Robust Features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 3951 LNCS* (pp. 404–417). https://doi.org/10.1007/11744023_32

Deng, Z., Li, Y., Zhu, H., Huang, K., Tang, Z., & Wang, Z. (2020). Sparse stacked autoencoder network for complex system monitoring with industrial applications. *Chaos, Solitons & Fractals*, *137*, 109838. https://doi.org/10.1016/j.chaos.2020.109838

Di Maggio, L. G. (2022). Intelligent Fault Diagnosis of Industrial Bearings Using Transfer Learning and CNNs Pre-Trained for Audio Classification. *Sensors*, *23*(1), 211. https://doi.org/10.3390/s23010211

Gao, Q., Liu, J., & Ju, Z. (2021). Hand gesture recognition using multimodal data fusion and multiscale parallel convolutional neural network for human–robot interaction. *Expert Systems*, *38*(5). https://doi.org/10.1111/exsy.12490

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hożyń, S. (2020). An automated system for analysing swim-fins efficiency. *Nase More*, *67*. https://doi.org/10.17818/NM/2020/3.9

Hożyń, S. (2021). A review of underwater mine detection and classification in sonar imagery. *Electronics (Switzerland)*, *10*(23). https://doi.org/10.3390/electronics10232943

Hożyń, S. (2023). Convolutional Neural Networks for Classifying Electronic Components in Industrial Applications. *Energies*, *16*(2), 887. https://doi.org/10.3390/en16020887

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvprw.2009.5206848

Jurczyk, K., Piskur, P., & Szymak, P. (2020). Parameters Identification of the Flexible Fin Kinematics Model Using Vision and Genetic Algorithms. *Polish Maritime Research*, *27*(2), 39–47. https://doi.org/10.2478/pomr-2020-0025

Kot, R. (2022). Review of Obstacle Detection Systems for Collision Avoidance of Autonomous Underwater Vehicles Tested in a Real Environment. *Electronics*, *11*(21), 3615. https://doi.org/10.3390/electronics11213615

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Liu, C., & Liu, S. (2018). Tiny Electronic Component Detection Based on Deep Learning. *2018 IEEE 3rd International Conference on Cloud Computing and Internet of Things (CCIOT)*, 341–345. https://doi.org/10.1109/CCIOT45285.2018.9032521

Mazhar, O., Navarro, B., Ramdani, S., Passama, R., & Cherubini, A. (2019). A real-time human-robot interaction framework with robust background invariant hand gesture detection. *Robotics and Computer-Integrated Manufacturing*, *60*, 34–48. https://doi.org/10.1016/j.rcim.2019.05.008

Neto, P., Simão, M., Mendes, N., & Safeea, M. (2019). Gesture-based human-robot interaction for human assistance in manufacturing. *The International Journal of Advanced Manufacturing Technology*, *101*(1–4), 119–135. https://doi.org/10.1007/s00170-018-2788-x

Nuzzi, C., Pasinetti, S., Lancini, M., Docchio, F., & Sansoni, G. (2019). Deep learning-based hand gesture recognition for collaborative robots. *IEEE Instrumentation & Measurement Magazine*, *22*(2), 44–51. https://doi.org/10.1109/MIM.2019.8674634

Oudah, M., Al-Naji, A., & Chahl, J. (2020). Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging*, *6*(8), 73. https://doi.org/10.3390/jimaging6080073

Piskur, P. (2022). Strouhal Number Measurement for Novel Biomimetic Folding Fins Using an Image Processing Method. *Journal of Marine Science and*

Engineering, *10*(4), 484. https://doi.org/10.3390/jmse10040484

Praczyk, T., Hożyń, S., Bodnar, T., Pietrukaniec, L., Błaszczyk, M., & Zabłotny, M. (2019). Concept and first results of optical navigational system. *Transactions on Maritime Science*, *8*(1). https://doi.org/10.7225/toms.v08.n01.005

Sheikholeslami, S., Moon, Aj., & Croft, E. A. (2017). Cooperative gestures for industry: Exploring the efficacy of robot hand configurations in expression of instructional gestures for human–robot interaction. *The International Journal of Robotics Research*, *36*(5–7), 699–720. https://doi.org/10.1177/0278364917709941

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*. https://doi.org/10.1109/CVPR.2015.7298594

Szymak, P., Piskur, P., & Naus, K. (2020). The Effectiveness of Using a Pretrained Deep Learning Neural Networks for Object Classification in Underwater Video. *Remote Sensing*, *12*(18), 3020. https://doi.org/10.3390/rs12183020

Xia, Z., Lei, Q., Yang, Y., Zhang, H., He, Y., Wang, W., & Huang, M. (2019). Vision-Based Hand Gesture Recognition for Human-Robot Collaboration: A Survey. *2019 5th International Conference on Control, Automation and Robotics (ICCAR)*, 198–205. https://doi.org/10.1109/ICCAR.2019.8813509

Żak, B., & Hożyń, S. (2013a). Distance Measurement Using a Stereo Vision System. *Solid State Phenomena*, *196*, 189–197. https://doi.org/10.4028/www.scientific.net/SSP.196.189

Żak, B., & Hożyń, S. (2013b). Segmentation Algorithm Using Method of Edge Detection. *Solid State Phenomena*, *196*, 206–211. https://doi.org/10.4028/www.scientific.net/SSP.196.206

Zalewski, J., & Hożyń, S. (2020). Shoreline Detection and Land Segmentation for Autonomous Surface Vehicle Navigation with the Use of an Optical System. *Sensors*, *20*(10), 2799. https://doi.org/10.3390/s20102799

STANISŁAW HOŻYŃ is an Engineering graduate from the Polish Naval Academy with a major in Electrical Systems Operation. He earned his MA in Electro-Automation from the Faculty of Marine Electrical Engineering, Gdynia Maritime University, in 2010. He served on two Polish warships consecutively. The first was as a Control Engineer for three years and the second as an Electrical Group Commander for eight years. Currently he is a PhD holder from the Polish Naval Academy. His PhD focused on computer vision for unmanned underwater vehicle control. He has been an Assistant Professor at the Department of Ship Automation since 2019. His main research interests include unmanned underwater vehicles and computer vision.