University of Pardubice

Faculty of Electrical Engineering and Informatics

APPLICATION OF DEEP NEURAL NETWORKS IN IMAGE PROCESSING

Dissertation

2024 Dominik Štursa

Declaration (in Czech):

Práci s názvem *Application of Deep Neural Networks in Image Processing* jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne

Dominik Štursa

Acknowledgements

I would like to express my deepest gratitude to my wife, Tereza, for her unwavering support and patience throughout this journey. Her encouragement and understanding have been invaluable. I am also profoundly grateful to my parents for their continuous support throughout all stages of my education. Their belief in me and their constant encouragement have been a source of great strength. I extend my heartfelt thanks to my supervisor, Petr Doležel, for his guidance and mentorship. His assistance in both my professional and personal development has been essential in completing this dissertation. I would also like to thank Charlie for his invaluable help with revisions and proofreading.

Furthermore, I would like to acknowledge the support of the Cotutelle and SGS programs provided by the University of Pardubice, which made it possible to carry out this work in collaboration with the GICAP group and Bruno B. Zanon, whose contributions and insights have enriched this dissertation.

ANNOTATION

This thesis deals with the application of deep neural networks in the processing of static two-dimensional image data. The aim is to investigate the extraction and classification of key features of images and apply this knowledge for object recognition, classification, localization and detection in image data. The research focuses on testing the hypothesis of the ability of deep neural networks to efficiently process and interpret image data, with an emphasis on designing and optimizing appropriate neural network architectures. The work also includes the development of a methodology for generalizing object detection by transforming features into probabilistic maps. The outcomes of this work also include a set of applications where the proposed transformation methods can be effectively utilized, such as tracking individuals in public spaces or the precise detection of key points on objects for robotic grasping and manipulation.

KEYWORDS

Deep neural networks, image processing, object detection, probabilistic maps.

NÁZEV

Aplikace hlubokých neuronových sítí při zpracování obrazu.

ANOTACE

Tato práce se zabývá aplikací hlubokých neuronových sítí v oblasti zpracování statických dvourozměrných obrazových dat. Cílem je zkoumat extrakci a klasifikaci klíčových vlastností obrazů a aplikovat tyto poznatky pro rozpoznávání, klasifikaci, lokalizaci a detekci objektů v obrazových datech. Výzkum se zaměřuje na ověření hypotézy o schopnosti hlubokých neuronových sítí efektivně zpracovávat a interpretovat obrazová data, přičemž klade důraz na návrh a optimalizaci vhodných architektur neuronových sítí. Práce dále obsahuje rozvoj metodiky pro zobecnění detekce objektů transformací vlastností do pravděpodobnostních map. Výsledkem práce je také soubor aplikací, kde lze efektivně využít navržené transformační metody, jako je sledování osob ve veřejném prostoru nebo přesná detekce klíčových bodů na objektech pro robotické uchopování a manipulaci.

KLÍČOVÁ SLOVA

Hluboké neuronové sítě, zpracování obrazu, detekce objektů, pravděpodobnostní mapy.

Table of Contents

List of Figures and Tables

List of Abbreviations

Adam	Adaptive Moment Estimation
ASP U-Net	Attention Squeeze Parallel U-Net
CNN	Convolutional Neural Network
FPS	Frames per Second
FN	False Negative
FP	False Positive
gIoU	Generalized Intersection over Union
HOG	Histogram of Oriented Gradients
IoU	Intersection over Union
KNN	k-Nearest Neighbors
LBP	Local Binary Patterns
LoG	Laplacian of Gaussian
MAE	Mean Absolute Error
MSER	Maximally Stable Extremal Regions
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
R-CNN	Region-based Convolutional Neural Network
RUSBoost	Random Undersampling Boost
SSD	Single Shot MultiBox Detector
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VGG	Visual Geometry Group
YOLO	You Only Look Once

Objectives

The main objective of this dissertation was to develop and apply a method based on deep neural networks for specific image processing applications. The work was focused on detection, localization and classification of people and objects in image data, which presented significant challenges and opportunities in various application domains. The goal was to develop a robust algorithm capable of analyzing image data and providing reliable results in achievable time. To achieve this goal, it was necessary to define and implement metrics to evaluate the performance of the proposed methods. These metrics included classical indicators, but also specific metrics focused on detection and classification tasks in the context of image processing. Comparison of the proposed solutions with classical and state-of-the-art methods played a key role in assessing their effectiveness.

The selected problem was specified and restricted to static data, i.e., images taken at a well-defined and fixed angle. These constraints also included a low number of object classes with defined variance of the object distance from the sensor. A general sensor system scheme was developed to provide suitable data for these specific tasks. The camera system was placed over the sensed area at a defined distance, which minimized perspective distortion and ensured consistent sizes of objects of the same class in different images.

In the initial stages of the research, the detection problem was limited to a single class of objects, namely the detection of people. This approach allowed focusing on optimizing the algorithms for one specific class, which facilitated tuning and evaluation. Gradually, the research was extended to multiple object classes, increasing the complexity and practical applicability of the system. The thesis consists of a total of six thematically related publications, in which the author was a co-author, that mapped the solution procedure of the detection problem presented in this dissertation.

Article 1 (Annex 1) was focused on the classification of people in image data, investigating different combinations of feature extraction techniques and classification algorithms. This research was followed by the Article 2 (Annex 2), dedicated to testing different settings of the best performance methods of the first paper in comparison with conventional convolutional neural network topologies.

Paper 3 (Annex 3) presented an innovative approach of transforming objects into a centroid using a segmentation neural network. This approach used an encoder-decoder scheme and allowed more accurate head detection, based on transforming the input image into a probabilistic map of head occurrence. Article 4 (Annex 4) was focused on extending the idea of paper 3, to the detection of persons by predicting the pixel-accurate centroid positions using fully convolutional networks.

Article 5 (Annex 5) extended the idea of detection to other objects of non-symmetric shapes and to detect selected parts of them. In particular, a new system for grasping point detection for industrial robots using a custom-designed fully convolutional ASP U-Net was presented.

Introduction

Currently, there is a growing effort to interpret the world around us based on the perceptual systems that people possess. One of such systems is the visual system, which provides humans with up to 80% of the information about their surroundings (Šajdíková, 2018). Due to the importance of this system for human perception, image processing methods, specifically digital image processing methods, are intensively studied. This interest in image processing is derived mainly from two main application areas. The first area is by improving the interpretation of image data, and thus improving human understanding of this data. The second area involves the direct processing of image data for further analysis or use; especially in autonomous systems that rely exclusively on machine perception. Machine perception is certainly related to the capabilities of the perceptual system and the types of signals that such a system can detect. From this perspective, it is necessary to define how image data is represented by signals. The image data is understood as a two-dimensional signal that is realized at each point by the amplitude distribution of the individual color components (Chakravorty, 2018). The actual observation of colors depends on the wavelength of the electromagnetic radiation that falls on the sensor of the perceptual system.

In the case of human perception (through the visual apparatus), electromagnetic radiation can be observed exclusively in the visible spectrum, which corresponds to wavelengths from approximately 400 nm to 800 nm (Bohren, 2006). Machine vision systems have a much wider range of observable wavelengths, with technologies allowing for the observation of infrared, ultraviolet, or X-rays, for example (Gonzalez, 2002). Depending on the technological capabilities, it is also common to measure other variables, such as magnetic properties, and then display them as two-dimensional image data, as seen in magnetic resonance imaging (Hoult, 1997). For digital image data processing it is essential to capture these variables. In addition to the sensors themselves, it is also necessary to define the region of interest, which can be a single image point (pixel), a two-dimensional region of pixels (image), or a multidimensional representation of the image data. This multidimensional representation may depend on time, where it includes a sequence of individual frames (video), or space, where the intensity of individual image pixels describes the distance from the sensor (depth maps or point clouds).

In addition to capturing image data, it is also important to interpret it and extract further insights from it. When digital devices are used, such a process is referred to as digital image processing (Silva, 2005). Digital image processing methods are now widely used in human daily life. A common example of the implementation of image processing methods is the mobile phone, which commonly integrates methods for various brightness correction, color temperature, noise, or automatic pixel correction in low light conditions (Thabet, 2015). In addition to these direct methods, there are also built-in methods for detecting or locating people or objects in the image, or identifying the user through facial recognition (Guillaume, 2015).

The use of image processing methods can be classified according to the application domain or the requirements for accuracy, speed, or reliability of processing. Industrial applications require relatively high standards of accuracy and reliability in image processing, for example, when dealing with the problem of selecting an object from a storage space (bin picking) or moving objects between different spaces (pick and place) (Carvalho, 2012). In

addition to robotics, another important application area is area-of-interest monitoring. This includes remote sensing, airspace surveillance systems and detection, counting and tracking of people. These applications range from optimizing the use of transport links to analyzing trajectories and improving public spaces with regard to bottlenecks and evacuation safety.

Person detection is the first step in any tracking or counting system. Accurately tracking the number of passengers entering and exiting vehicles is essential for public transport safety, passenger movement forecasting, transport planning, vehicle utilization monitoring, station management and operations, and cost optimization (Olivo, 2020; Siebert, 2020). Camera systems are increasingly being installed in private and public spaces for security and surveillance reasons. As image acquisition becomes more affordable, more image processing methods are being tested. Systems for tracking, counting, or detecting people are often implemented using computer vision techniques and video processing algorithms. Various image and video processing methods are widely used, including methods based on shape features, pattern learning, or area estimation (Wu, 2014). Often, an additional sensing system is added to standard cameras to provide information in the form of depth maps which improve the overall accuracy of the detection system (Fu, 2012).

The ethical implications of installing a detection system in public spaces should also be considered and consequently the identification of persons (especially faces) should be avoided. This problem can be solved naturally by installing sensing systems orthogonally over the monitored area. Given the focus of this dissertation, this sensing principle will be described in detail in the following chapter.

This dissertation will address the following areas:

- Accurately defining the initial detection problem and obtaining the necessary data to perform experiments in the problem area.
- Description and application of standard image processing methods for detecting people in image data and presentation of results leading to further testing.
- Presentation of the hypothesis, the methodology for its testing and a series of experiments along with their results.
- Comparison of results with other competing approaches based on convolutional neural networks.
- Presentation of the extension of the method to detect more types of shapes and objects.
- Discussion and summary of the results, including considerations on future directions for improving the proposed methods.

The work presented in this dissertation is built upon previously published articles in which all detailed information can be found. These articles are summarized and critically discussed in the relevant sections of the dissertation.

1 Problem Definition and Data Collection

There is a wide range of research possibilities in the field of object detection in image data. For the possibility of partial development, it is necessary to specify the selected problem in more detail. For the purposes of this dissertation, the detection problem was restricted to static data (evaluated from an image), captured at a well-defined and constant angle, with a small number of classes of objects detected, and with only a relatively low variable distance of the objects from the sensor (camera system). For these tasks, the camera system is positioned directly above the captured area at a defined distance, resulting in only a small perspective distortion. The detection problem is principally limited to a single class of objects, specifically the detection of people. For person detection, two possible configurations were used based on the needs of possible applications.

In the context of these conditions, the applications relate to the assessment of indoor environments such as shopping malls, airports, office buildings and public transport areas where the movement of people on flat surfaces or stairways needs to be monitored. Consequently, two primary configurations for person detection were considered.

1.1 Flat-Floor Configuration

This configuration involves detecting people on a flat surface, such as an interior floor. The vertical positioning of the camera over a flat surface offers a clear and undistorted view of objects, which facilitates detection and subsequent image processing. This configuration is suitable for methods based on sliding windows (Arrora, 2020), where detection is performed using cutout classification combined with localization based on the position of the cutout in the image. A schematic representation of the configuration of the system for detecting people on a flat floor is shown in Figure 1.

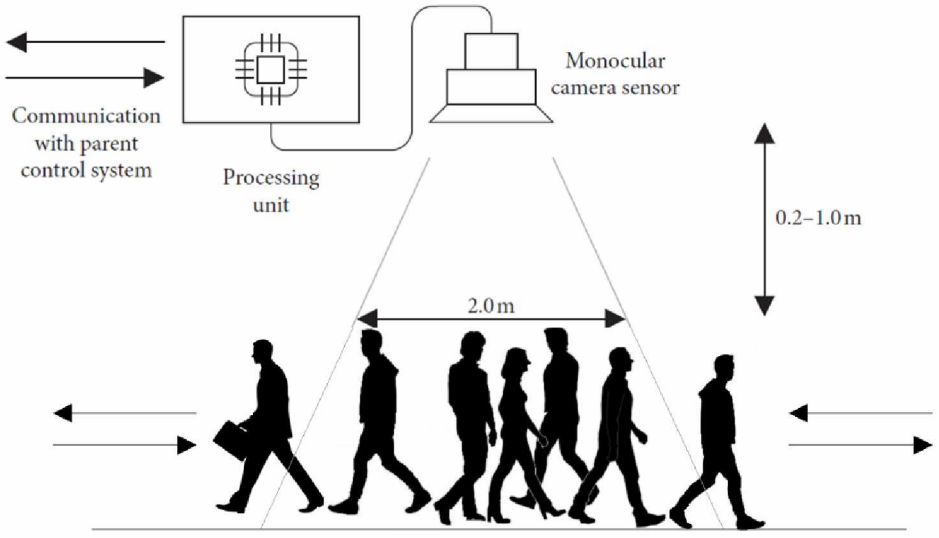


Figure 1 – Architecture of the system for tracking people on the flat surface (side view).

1.2 Staircase Configuration

This scenario is more complex due to the greater variability of distances from the sensing system. This applies to public transport access staircases as well as staircases in buildings, where the distance variability naturally increases with the size of the staircase. Detection on stairs is not ideally performed using sliding window methods due to their computational complexity with variable window size (Ferrari, 2008; Bosch, 2008). Instead, more general methods, usually based on artificial neural networks, are recommended. A schematic representation of the system configuration for detecting people on a defined staircase is shown in Figure 2.

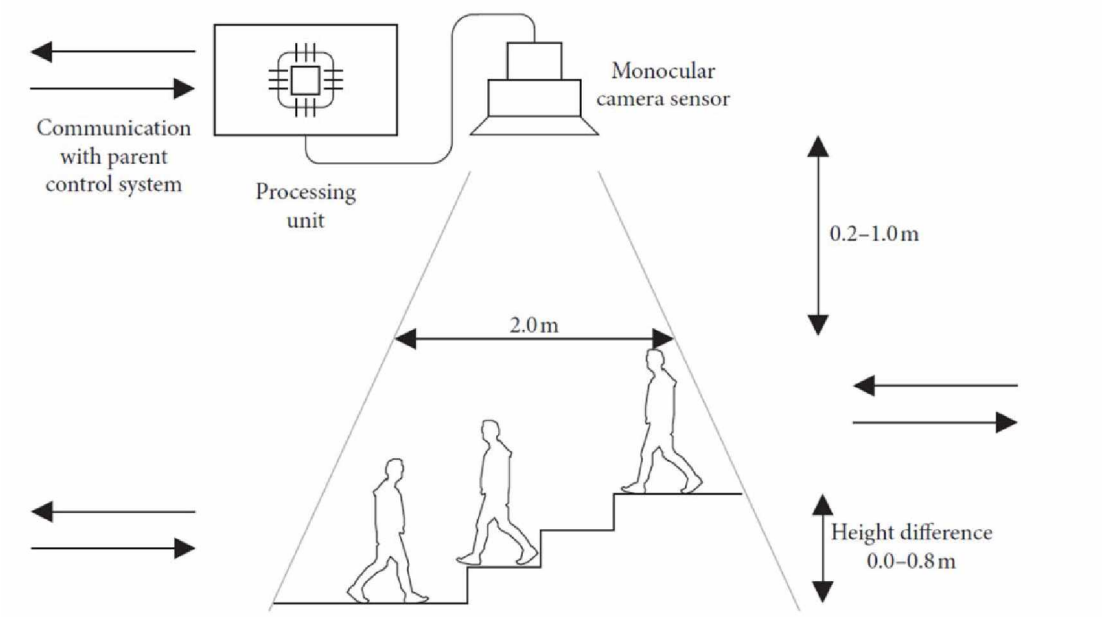


Figure 2 – Architecture of the system for tracking people on the staircase (side view).

However, to achieve greater consistency of observed head sizes within the acquired image data, imaging from greater distances using a camera system with a narrower field of view was also considered. This approach minimized the variation in head sizes, allowing the use of sliding window approach for processing. Overall, this resulted in several datasets for the person classification and detection problem, which are described in the following section.

1.3 Data Collection

Several datasets were created in total. The first group of datasets was created for testing standard approaches based on sliding window methods with normalized window sizes. These included three datasets of object images suitable for evaluating object classification methods.

The first dataset was created in an indoor public space under various lighting conditions. Video sequences of people walking on a staircase were captured using a monocular camera. Frames with significant movement of the individuals' head positions between consecutive frames were then selected. From the selected frames, uniform size images containing heads and other objects (not heads) were progressively cropped. In total, 736 original object images of normalized size were obtained. The second dataset was created in the staircase areas of public

transportation. From several video recordings taken in the doorway areas of a tram, a total of 6 020 images of normalized size were obtained. These images were divided into two classes of approximately equal size. Examples of object images from mentioned datasets are shown in Figure 3.

Figure 3 – Examples of object images.

For the second group of datasets, the images were captured in the same locations and under the same conditions as the first group. However, during the processing method, visible parts of the heads were labeled in the entire images, and the positions of the head labels were recorded. This approach made these datasets suitable for testing detection methods. In total, 1 173 images were captured on a staircase in a public building, along with an additional 7 000 images from various areas, other staircases within the same public building, and from boarding corridors of trams. The aim was to ensure high variability in background, lighting conditions, and staircase inclines to enhance the robustness of the detection system. Images from various locations are illustrated in Figure 4.

Figure 4 – Samples of images from individual locations.

Overall, the acquisition was performed using two hardware configurations. In some cases, a combination of a Basler acA2500-60uc industrial color camera (Basler, 2020) with a

high quality Computar M3514-MP lens (Computar, 2020) was used. This offered good light sensitivity and therefore only minimal image quality degradation. In other cases, a RealSense D435 stereo camera was used. However, only images from one of the stereo sensor color cameras were used for dataset purposes. All datasets created are described in Table 1.

Table 1 – Details of the individual datasets

Dataset	Type	Frames count	Acquisition HW
Article 1	Classification (cut-outs)	736	Basler acA2500-60uc
Article 2	Classification (cut-outs)	6 020	Basler acA2500-60uc
Article 3	Detection (Full frames)	1 173	RealSense D435
Article 4	Detection (Full frames)	7 000	RealSense D435

2 Image Classification Methods

This section discusses the traditional image classification methods used in person detection using the sliding window technique. Initially, standard approaches that include various combinations of feature extraction techniques and classification algorithms were evaluated. This evaluation provided insight into the effectiveness of different methodologies in extracting relevant information from image data and classifying it accurately. Following this, the research was extended by comparing these traditional methods with more advanced techniques, namely convolutional neural networks (CNNs). The aim of the comparison was to determine the most effective combinations of traditional approaches and to assess their performance compared to CNNs. This section describes the methodology and results presented in Papers 1 and 2 (Annexes 1 and 2).

2.1 Traditional Image Processing Methods

Image classification is one of the key tasks in computer vision, which involves assigning a class or category to a given image based on its content. Traditional image classification methods usually rely on feature extraction, which is then used for classification through various machine learning algorithms. This approach involves several key steps including image preprocessing, feature extraction, and finally classification. A block diagram of a traditional classification system is shown in Figure 5.

Figure 5 – Block diagram of a traditional classification system.

In this approach, for the data collection section, the input to the detector is a grayscale normalized image cropped from a real RGB video frame. The output of the detector is the object class. Since the detector is designed to distinguish only two classes (Head vs. Non-Head), the classification problem can be reduced to a binary classification problem. An illustration of the functionality for both classes is shown in Figure 6.

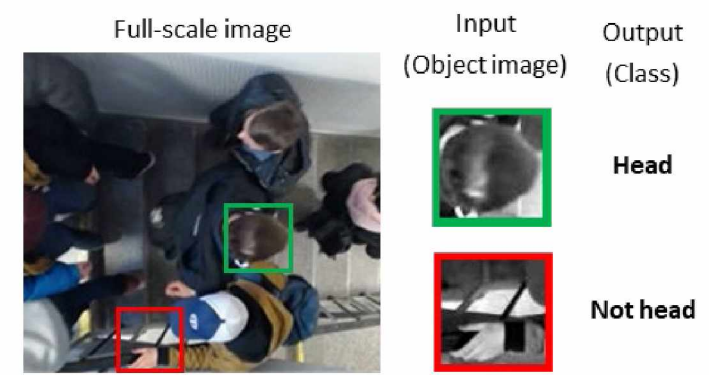


Figure 6 – Person detector functionality. (Stursa, 2020b)

To preprocess the object image, grayscale conversion was used along with contrast normalization. Well-established and well-known feature extraction techniques were selected for testing, specifically feature extraction techniques based on edge and curve detection algorithms, blob detection algorithms, binary local patterns, histograms of oriented gradients, and pixel intensities. Most of the extractors can be defined by various tunable parameters, where for the purpose of validating the given methods, these parameters were set according to the recommended heuristics. Since some extractors return a variable number of extracted features, imprinting into a new white image was chosen to normalize them. Specifically, ten feature extractors were tested.

Specifically, these were the following edge detectors: the Canny, Sobel, Prewitt, and Roberts' edge detectors, together with the Laplacian of Gaussian, descriptions of which can be found in summary articles (Parker 2010; Lim 1990). In addition, blob detectors have been used, namely Binary robust invariant scalable key points (KAZE) (Alcantarilla 2012) and Maximally Stable Extremal Regions (MSER) (Nister 2008). In terms of other extractors, Histograms of oriented gradients (HOG) (Dalal 2005) and Local Binary Patterns (LBP) (Ojala 2002) were used. Last, pixel intensities themselves were also chosen, which is basically image information without more sophisticated feature extraction. The specific parameter settings can be found in the original paper (Stursa, 2020b).

For the purpose of classification, established and widely validated classification techniques were selected, including decision trees, support vector machines (SVM), and nearest neighbor algorithms. The decision tree group included decision trees, classification trees, and regression trees, namely: Fine, Medium, Boosted, Bagged, and RUS Boosted Trees, as described in Breiman's book (Breiman, 1984). In addition to trees, the Support Vector Machine (SVM) family was represented. This included Linear, Quadratic, Medium Gaussian, and Coarse Gaussian SVM classifiers, as detailed in Su's work (Su, 2013). The final classifier chosen was Cosine Nearest Neighbor, known for its good accuracy in low-dimensional spaces (Weinberger 2009). Thus, a total of ten classifiers were tested. The specific classifier parameter settings can be found in the original paper (Stursa, 2020b).

2.1.1 Dataset Preparation

As mentioned in the data collection chapter, a total of 736 original object images of normalized size were obtained from video footage for the purpose of testing image data classification. These images were approximately evenly distributed between the two classes (Heads, Not Heads). To expand the dataset, data augmentation was performed by rotating the images by 90, 180, and 270 degrees. After augmentation, the dataset comprised of 2 944 images, which were further divided into training and testing sets. The training set contained 1 065 images of the 'Head' class and 1 144 images of the 'Not Head' class, while the testing set included 355 images of the 'Head' class and 380 images of the 'Not Head' class.

2.1.2 Experiments with Standard Methods

All images of the objects in the dataset have been preprocessed. Various feature extraction techniques were then applied to the dataset, resulting in ten groups of data. Each of the classification techniques was then learned and applied to each of these groups. In total, all

the combinations of extractors and classifiers resulted in 100 individual solutions for finding the object class. All the extractors and classifiers used are listed in Table 2.

Table 2 – List of used extractors and classifiers. Adapted from (Stursa, 2020b).

Feature extraction methods		**Classification methods**	
Canny edge detector	KAZE detector	Fine Tree	Coarse Gaussian SVM
Sobel edge detector	MSER detector	Medium Tree	Cosine KNN
Prewitt edge detector	HOG features	Linear SVM	Boosted Trees
Roberts edge detector	LBP features	Quadratic SVM	Bagged Trees
LoG edge detector	Pixel intensities	Medium Gaussian SVM	RUSBoosted Trees

Evaluation Metrics

Standard metrics were used for evaluation, based on the correct assignment of predicted and actual classes. These assignments result in four possible outcomes:

- True Positive (TP) – correctly classified positive images.
- False Positive (FP) – negative images incorrectly classified as positive.
- True Negative (TN) – correctly classified negative images.
- False Negative (FN) – positive images incorrectly classified as negative.

From the counts of these groups (TP, FP, TN, and FN), obtained on the testing set, standard metrics such as accuracy (1), recall (2), precision (3), and F1-score (4) were determined. The definitions of these metrics are provided in the following equations.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 - score = \frac{2}{recall^{-1} + precision^{-1}} \tag{4}$$

In addition to these evaluations, the performance of each feature extraction method was also evaluated relative to the least efficient (longest lasting) method.

2.1.3 Results and Discussion

For all combinations, the evaluation of the primary parameters (TP, FP, TN, FN) was performed on a testing dataset. From these parameters, accuracy, recall and precision values were calculated to evaluate the results. These metrics are presented in Table 3 – Accuracy, Recall and Precision results. Taken from (Stursa, 2020b).. For clarity, the success rates in the table are highlighted by color shading: lower values are shaded in light yellow, and higher values approaching 100% are shown in blue.

Table 3 – Accuracy, Recall and Precision results. Taken from (Stursa, 2020b).

Accuracy										
Classifier / Extractor	**Canny**	**Sobel**	**Prewitt**	**Roberts**	**LoG**	**MSER**	**KAZE**	**LBP**	**PI**	**HOG**
Fine Tree	56.8%	70.7%	73.6%	74.0%	59.4%	65.9%	58.2%	83.7%	74.7%	76.9%
Medium Tree	59.0%	55.2%	58.3%	58.2%	52.4%	68.5%	60.1%	78.7%	75.1%	77.3%
Linear SVM	67.1%	78.9%	80.8%	82.6%	81.5%	78.3%	70.7%	85.6%	55.4%	93.9%
Quadratic SVM	67.7%	78.9%	81.9%	81.1%	79.1%	77.9%	70.4%	89.9%	85.1%	94.0%
Medium Gaussian SVM	70.7%	80.4%	80.4%	82.9%	81.8%	79.8%	71.7%	90.9%	85.7%	96.2%
Coarse Gaussian SVM	56.7%	78.0%	78.7%	79.5%	78.4%	77.4%	70.8%	85.9%	64.0%	94.8%
Cosine KNN	82.1%	82.6%	82.3%	82.2%	82.1%	73.9%	68.9%	87.1%	80.7%	95.5%
Boosted Trees	63.3%	62.5%	65.4%	64.9%	63.7%	76.6%	67.8%	87.9%	81.7%	92.8%
Bagged Trees	59.0%	79.5%	81.8%	82.3%	70.8%	78.1%	65.9%	89.0%	87.9%	93.3%
RUSBoosted Trees	60.3%	64.3%	67.1%	67.5%	57.3%	75.4%	62.8%	80.3%	79.9%	83.0%
Recall										
Classifier / Extractor	**Canny**	**Sobel**	**Prewitt**	**Roberts**	**LoG**	**MSER**	**KAZE**	**LBP**	**PI**	**HOG**
Fine Tree	46.2%	83.1%	84.5%	85.4%	75.2%	71.0%	51.8%	82.8%	76.6%	76.1%
Medium Tree	41.7%	92.1%	95.5%	96.9%	86.2%	68.7%	55.2%	69.3%	84.5%	76.6%
Linear SVM	58.6%	84.2%	87.6%	89.3%	83.1%	76.6%	70.1%	87.6%	47.9%	94.6%
Quadratic SVM	61.4%	83.1%	86.5%	85.4%	80.8%	76.1%	68.7%	94.1%	90.4%	94.4%
Medium Gaussian SVM	69.3%	80.8%	83.1%	85.6%	79.4%	77.5%	79.2%	95.8%	92.4%	96.9%
Coarse Gaussian SVM	13.5%	90.4%	91.8%	90.1%	82.8%	67.6%	75.2%	89.9%	82.3%	96.9%
Cosine KNN	71.8%	83.4%	84.8%	84.8%	78.6%	62.5%	46.2%	95.2%	77.5%	93.5%
Boosted Trees	34.6%	90.4%	91.3%	95.2%	80.8%	82.3%	62.0%	87.0%	93.2%	92.4%
Bagged Trees	33.8%	72.7%	77.2%	79.7%	59.2%	76.9%	59.2%	89.0%	87.0%	92.4%
RUSBoosted Trees	49.9%	86.2%	91.0%	92.4%	80.3%	82.8%	57.7%	69.6%	89.0%	83.1%
Precision										
Classifier / Extractor	**Canny**	**Sobel**	**Prewitt**	**Roberts**	**LoG**	**MSER**	**KAZE**	**LBP**	**PI**	**HOG**
Fine Tree	56.4%	65.4%	68.3%	68.6%	55.9%	63.0%	57.3%	83.3%	72.5%	76.1%
Medium Tree	60.9%	52.0%	53.8%	53.7%	50.4%	66.8%	59.2%	83.7%	70.1%	76.4%
Linear SVM	68.6%	75.1%	76.2%	77.9%	79.5%	77.9%	69.4%	83.4%	54.3%	92.8%
Quadratic SVM	68.3%	75.6%	78.3%	77.7%	76.9%	77.6%	69.5%	86.3%	80.9%	93.3%
Medium Gaussian SVM	69.7%	79.1%	77.8%	80.2%	82.2%	79.9%	67.7%	86.7%	80.8%	95.3%
Coarse Gaussian SVM	80.0%	71.5%	71.8%	73.4%	75.0%	82.5%	67.8%	82.4%	59.1%	92.7%
Cosine KNN	88.9%	81.1%	79.8%	79.6%	83.3%	79.0%	81.2%	81.3%	81.6%	97.1%
Boosted Trees	76.4%	57.0%	59.1%	58.4%	59.1%	72.8%	68.3%	87.8%	74.9%	92.7%
Bagged Trees	64.2%	82.7%	83.8%	83.0%	75.0%	77.6%	66.5%	88.3%	87.8%	93.7%
RUSBoosted Trees	60.8%	58.8%	60.6%	60.7%	53.9%	71.0%	62.3%	87.0%	74.4%	81.9%

The results presented in all three tables revealed some interesting findings. In particular, the combination of the HOG feature extractor and the SVM mean Gaussian classifier achieved the best accuracy, recall and precision. Only LBP and HOG extractors provided accuracies above 90%, while most of the remaining extractors achieved accuracies above 80%. However, the Medium Tree and RUS Boosted Tree classifiers performed consistently worse in terms of accuracy across all feature extractors.

In terms of relative computation time, the edge detectors were found to perform similarly to each other. The most computationally demanding method was MSER detection. The best performance was offered by the LBP feature extractor. The HOG feature extractor was found to be seven times more computationally intensive than LBP.

2.1.4 Conclusions

In paper 1, a set of feature extraction techniques combined with a set of classifiers for person detection was presented, proposed and tested. The results show that feature extraction using histogram of oriented gradients (HOG) in combination with classifiers based on support vector machines (SVMs) is an effective solution to this problem. It is clear that not only the accuracy but also the computational time needs to be optimized in order to develop a suitable tool for monitoring the flow of people in real applications. The performance of feature extraction is directly affected by the size of the feature vectors, and minimizing their size could lead to higher performance.

2.2 Classical and Neural Network-Based Approaches

This section builds on the findings presented in the previous sections, where the analysis concludes that the histogram of oriented gradients (HOG) and support vector machine (SVM) approach is worthy of further investigation. In addition, experiments with convolutional neural networks (CNNs) were conducted for comparison. As before, classification of the people in the images was considered for this research. The proposed system focused on the use of cameras in the visible spectrum, emphasizing both efficiency and anonymity of the occupants by placing the sensing cameras orthogonally to the ground plane. Two basic recognition techniques have been investigated: the CNNs and the HOG approach along with SVM. CNNs are widely recognized for their high accuracy in object recognition tasks, but they are associated with significant computational requirements. In contrast, the HOG approach together with SVM provides a more balanced solution that offers efficiency and reasonable performance, which is particularly suitable for real-time applications where computational resources are limited (Stursa, 2020a).

To address the classification problem, a comprehensive analysis of these methods has been performed to compare their performance and time complexity in practical settings. Experimental evaluations, based on an extended dataset, further elucidate the strengths and limitations of each approach. The aim was to identify the most practical and efficient solutions for implementation in public transport systems, following on from the work carried out in the previous sections.

Given the specific focus on public transport applications, a staircase configuration was used (see 1.2) that considers the specific parameters related to the placement and imaging capabilities of camera systems in public transport vehicles. In this context, the distance of the heads from the sensor was considered to range from 0.2 to 1 meter, with a sensing area of 2.4×2 meters. As discussed in the data collection section, normalized images of objects, including those containing heads and other objects or body parts (the 'Not Head' class), were cropped for use with standard methods and convolutional networks.

In terms of processing using conventional computer vision methods, the process of object classification in images consists of three sequential steps: image preprocessing, feature extraction, and classification. In this case, the HOG descriptor and SVM classifier were employed, with the process illustrated in Figure 7.

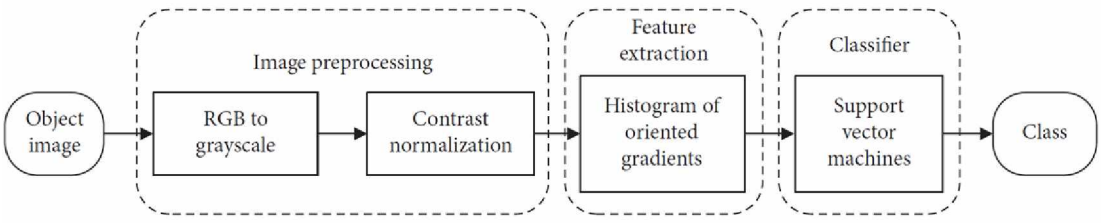


Figure 7 – Classification systems with HOGs and SVM. Taken from (Stursa, 2020a).

The second approach was classification using convolutional neural networks. Convolutional neural networks (CNNs) are a specialized type of artificial neural network that utilize a discrete two-dimensional convolution operation in certain layers, instead of the traditional matrix multiplication. This unique feature gives them the name "convolutional." In a convolutional layer, the network sequentially performs convolution on individual sections of the previous layer using a convolutional kernel, gradually generating a new layer known as the feature map. Subsequent layers are applied to these feature maps to select the most relevant parameters and act as feature extractors. As a result, when using convolutional networks, the need for preprocessing and individual feature extraction can be eliminated. Then, in CNNs designed for classification, multilayer feed-forward networks are used to classify the extracted features. This reduces the total number of steps required. This classification process using convolutional neural network is shown in Figure 8.

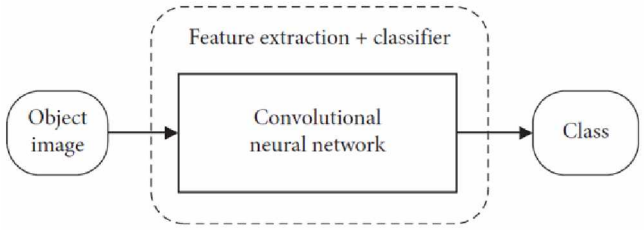


Figure 8 – Classification system based on CNN. Taken from (Stursa, 2020a).

2.2.1 Dataset preparation

As mentioned in the data collection section, a total of 6 020 original object images of normalized size were extracted from the video data for the purpose of testing image classification. These images were approximately equally divided between the two classes. To ensure robustness, data were collected at different locations and under different lighting conditions. Specifically, 1 720 images were taken in outdoor areas with strong natural light. In addition, images were taken indoors with natural light, with 1 700 images taken in sufficient lighting and 1 400 in low lighting. Finally, 1 200 images were taken indoors with artificial lighting.

The dataset was then typically split for training and testing purposes. The training set consisted of 2 008 images of the 'Head' class and 2 012 images of the 'Not Head' class, while the testing set consisted of 1 000 images of the 'Head' class and 1 000 images of the 'Not Head' class. Since the classification performance of the classifiers depends on the composition of the training sets, a random split of the training set into training and validation subsets was further performed with a ratio of 85:15.

2.2.2 Experimental Procedure

The aim of the experiments was to evaluate the accuracy and performance of two different passenger recognition methodologies: the classical HOG and SVM approach and modern convolutional neural networks (CNNs). The main objective was to assess the classification accuracy, time complexity and practical applicability in real-time scenarios.

HOG and SVM: Traditional Approach

The HOG descriptor encodes local shape information from regions in the image into a feature vector (Dalal, 2005). The descriptor has five parameters: number of bins, orientation binning, cell size, number of cells in blocks, and number of overlapping cells between adjacent blocks. Since cell size significantly affected the final performance of image recognition systems (Skrabanek, 2017), the effect of this parameter on the classification performance of HOG based passenger recognition systems was investigated. Specifically, the effect of cell sizes of [6, 6], [8, 8], ..., and [16, 16] pixels can be seen in Figure 9. For the other parameters, a conservative setting was used, which can be found in the original paper (Stursa, 2020a).

Figure 9 – Images with highlighted HOG features for different cell sizes. Taken from (Stursa, 2020a).

Training an SVM classifier involves solving an optimization problem to find the hyperplane that maximizes the difference from the training data (source). In cases where the data is not linearly separable, it needs to be transformed into a linearly separable problem using an appropriate kernel function. For strongly nonlinear problems, the choice of kernel function is crucial. With this in mind, the effect of different kernels on the performance of a passenger recognition system was tested. Specifically, the focus was on the best performing linear, Gaussian radial basis function (RBF), and polynomial kernel functions with polynomial kernels of order 2 and 3. The system has been trained and verified on the corresponding subsets (see 2.2.1). This training and validation process was repeated hundreds of times for every possible combination of kernel function and cell size. A detailed setting can be found in the original paper (Stursa, 2020a).

Validation was performed on the corresponding validation subsets using a loss function defined as the sum of the misclassified observations in the following equation (5), in which n is the number of images in the training subset, I is the indicator function, y_j and $\widehat{y_j}$ are an actual and a predicted class of the j-th object image.

$$E_{SVM} = \sum_{j=1}^{n} I\{\widehat{y_j} \neq y_j\} \tag{5}$$

Convolutional Neural Networks

To verify the classification capability of CNN, the most-commonly used state-of-the-art architectures were considered. The main disadvantage of state-of-the-art deep convolutional network architectures is their high computational complexity. Passenger recognition can be implemented using a CNN of a suitable custom architecture. Due to the importance of low computation time, the performance of five CNN architectures with different complexity was tested.

The simplest architecture, Net1, consisted of one convolutional layer, one max-pooling layer, and two fully connected layers. Classification was performed using the softmax function. In the second simplest architecture, Net2, the single convolutional and max-pooling layers were replaced by a sequence of layers. Thus, in total, Net2 contained a sequence of 2 convolutional layers with a max-pooling layer, where this sequence was inserted 2 times in a row leading to two fully connected layers. Both architectures used the ReLU activation functions in the convolutional and fully connected layers. To mitigate overfitting, dropout layers were added after each max-pooling layer and the first fully connected layer in both networks. The remaining three architectures considered were the well-known LeNet-5 (Lecun, 1998), AlexNet (Krizhevsky, 2017), and VGG-16 (Simonyan, 2015) networks, sorted by complexity. LeNet-5, the pioneering CNN, has a relatively simple architecture. AlexNet is one of the most cited deep CNNs with numerous industrial and technical applications. VGG-16 is a very deep convolutional network with 13 convolutional layers and 3 fully connected layers. Despite its depth, the VGG-16 can process data in adequate time.

All these networks were trained from scratch with randomly initialized weights according to a normal distribution. In addition, transfer learning was applied to the AlexNet and VGG-16 networks to test potential performance improvements, where for both architectures the last three layers of the pre-trained networks were fine-tuned. Due to the stochastic nature of the training process, training was repeated hundreds of times for each network and training strategy. At each training, the training set was randomly divided into training and validation subsets as mentioned in 2.2.1 . Training was performed in batch mode for 100 epochs with batches of 32 images. The data in the training subsets were randomly shuffled for each epoch. The adaptive moments estimation method was used as the optimizer. The choice of optimizer and hyperparameter settings resulted from a pilot study. All specific parameters can be found in the original paper (Stursa, 2020a).

For learning, the binary cross-entropy loss function (6) was minimized, in which n is the number of images in the training subset, y_j and $\widehat{y_j}$ are an actual and a predicted class of the j-th object image.

$$E_{CNN} = -\frac{1}{n}\sum_{j=1}^{n} \widehat{y_j} \cdot \ln(y_j) + (1 - y_j) \cdot \ln(1 - \widehat{y_j}) \tag{6}$$

2.2.3 Results and Discussion

For all methods used, the resulting loss functions were evaluated and consisted of 100 experiments for each method. Box plots were used to present the validation results, where the center lines represent the medians of the loss functions, the box edges indicate the 25th and 75th percentiles, and the whiskers indicate variability outside the upper and lower quartiles. Data were grouped by training methods and strategies (x-axis), with values on the y-axis corresponding to loss function values. These results for the convolutional networks tested are shown in the following Figure 10, in which TL stands for transfer learning.

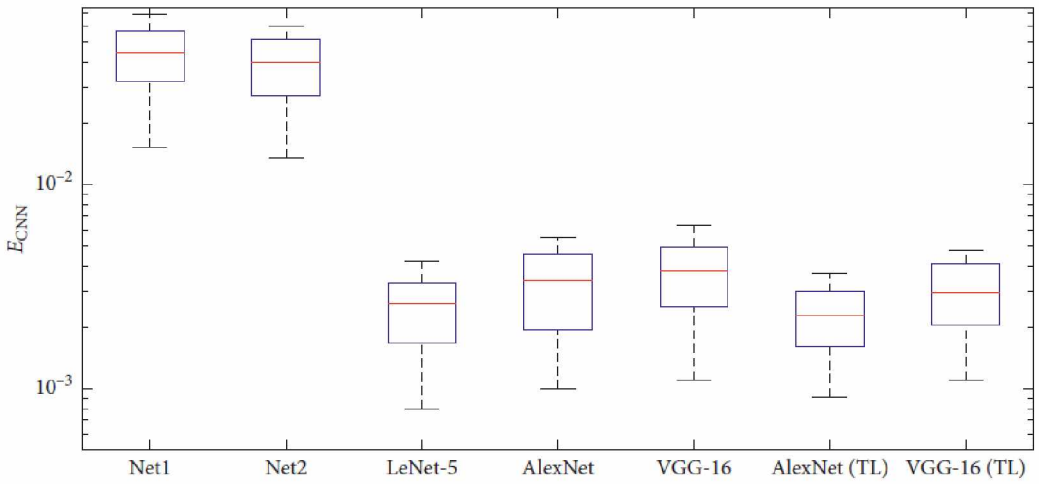


Figure 10 – Loss function values shown in box plots. Taken from (Stursa, 2020a).

The methods with the best results were further used to obtain the primary parameters (TP, FP, TN, FN) on the test dataset. From these parameters, the values for precision, recall, accuracy, and F1-score were calculated to evaluate the results. The numerical results for accuracy, precision, recall, and F1-score for the best-trained convolutional networks, along with the three best-performing and two worst-performing HOG and SVM, are presented in Table 4. To highlight the values, a color scale gradient from light yellow to deep blue is used, with the highest values indicated by the deepest blue.

Table 4 – Selected results of tested approaches. Taken from (Stursa, 2020a).

Classifier	Accuracy	Precision	Recall	F1-score
Net1	0.949	**0.950**	0.948	0.949
Net2	0.953	0.947	0.961	0.954
LeNet-5	**0.956**	0.946	0.966	**0.956**
AlexNet	0.947	0.921	**0.977**	0.948
VGG_16	0.928	0.903	0.958	0.93
SVM with RBF kernel function, cell size [6, 6] px	0.949	**0.957**	0.941	0.949
SVM, polynomial degree = 3, cell size [10, 10] px	**0.959**	**0.957**	**0.961**	**0.959**
SVM, polynomial degree = 3, cell size [12, 12] px	0.95	**0.957**	0.942	0.949
SVM with linear kernel function, cell size [14, 14] px	0.919	0.925	0.913	0.919
SVM, polynomial degree = 3, cell size [14, 14] px	0.921	0.92	0.923	0.922

Box plots of the error functions for HOG and SVM, along with all the results for precision, recall, accuracy, and F1 score, can be found in the original article (Stursa, 2020a). Since the computational complexity of each method was also tested, Figure 11 shows a comparison of the relative time with the worst method along with the F1-score as a robust indicator for evaluating the accuracy of person detection.

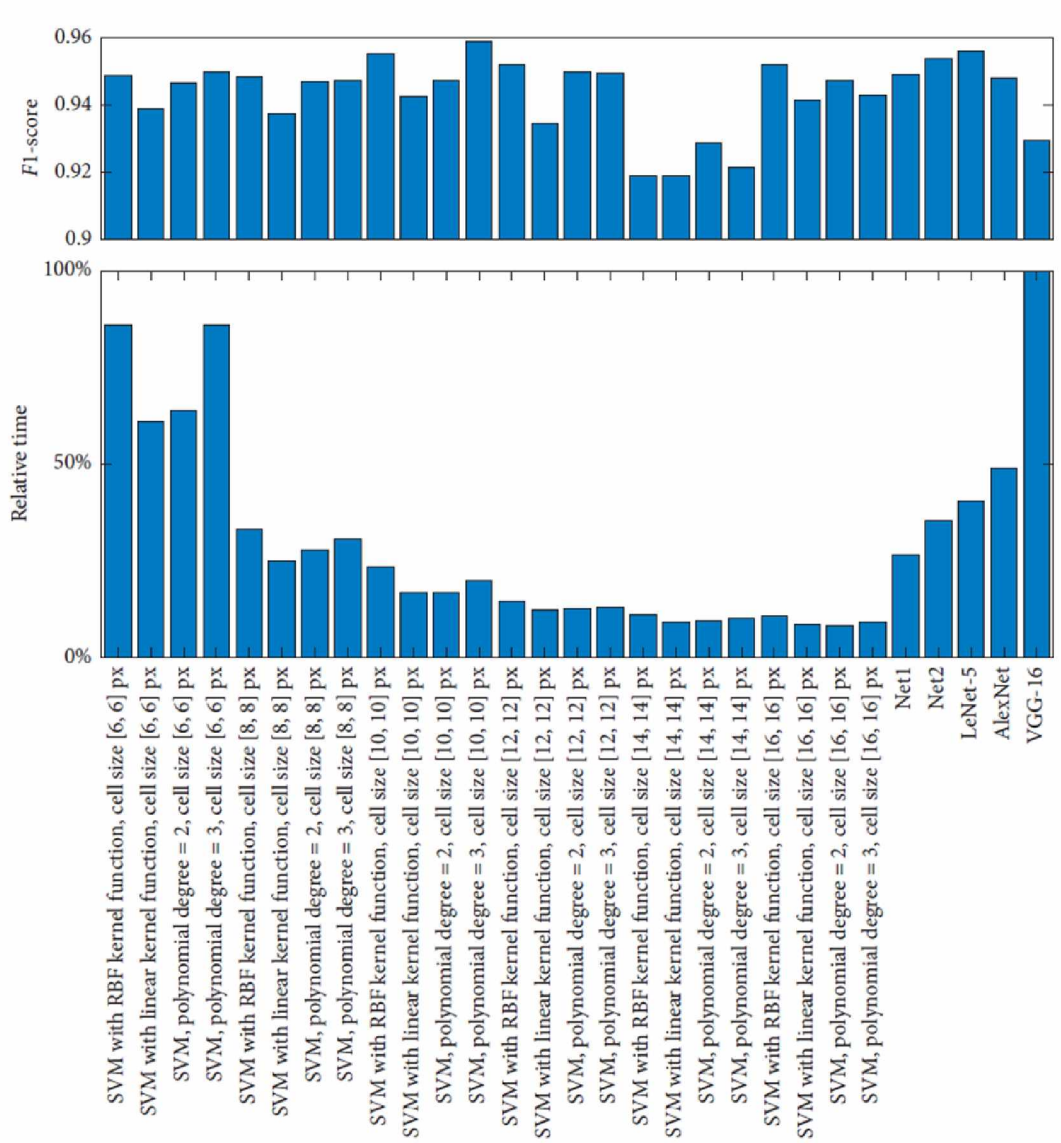


Figure 11 – Relative computational times of the passenger recognition systems in comparison with F1-score of each system. Taken from (Stursa, 2020a).

The results indicated that the HOG based system with a polynomial kernel function of third degree and cell sizes of [10, 10] px outperformed CNN based systems in most metrics, including lower computational complexity. While CNN architectures such as AlexNet and LeNet-5 demonstrated competitive classification performance, the high learning capacity of VGG-16 led to overfitting, and its high computational cost makes it less suitable for this task. The advantages of transfer learning were evident, as it resulted in lower variability and smaller loss function values compared to training from scratch.

2.2.4 Conclusion

In conclusion, although deep convolutional networks are often the first choice for developing image recognition systems, traditional computer vision methods can achieve equally good classification performance. These traditional methods, if properly designed and configured, can outperform CNN-based solutions in terms of time efficiency, which is crucial for real-world applications. Specifically, a HOG-based passenger recognition system using HOG features combined with an SVM classifier has demonstrated both time efficiency and high accuracy. However, it is important to note that these methods were validated on normalized object image sizes and in more complex applications with non-normalized cutouts, the method would suffer reduced computational efficiency.

2.3 Summary

The previous chapters explored both classical and advanced image classification methods for person detection. The research focused on traditional methods which include feature extraction and classification using various algorithms such as SVM and decision trees. Techniques like edge detection, HOG, LBP, and others were used for image features extraction. The results demonstrated that the combination of HOG and SVM with a polynomial kernel achieves high accuracy and is computationally efficient.

Additionally, traditional methods were compared with convolutional neural networks (CNNs), known for their high accuracy but also for their computational demands. The experiments showed that although CNNs provide high accuracy, traditional methods can offer comparable performance with lower computational resource requirements. The research highlighted the importance of selecting an appropriate method depending on the specific application.

It should be noted, however, that the methods investigated so far are primarily classification-based, meaning they rely on the sliding window method. This technique involves scanning the entire image in sections (windows), which are then passed to the classifier for evaluation. To achieve accurate detection, it is often necessary to create cutouts in various sizes, resulting in thousands of object images to classify. This process is computationally intensive and can significantly burden the system, creating a bottleneck for real-time applications. For these reasons, methods that allow processing of the whole image were further considered. This approach increases efficiency and allows for faster system response, which is essential in applications such as security systems and public transport monitoring.

In this dissertation, new methods, including advanced deep neural networks, are presented in detail in the following chapters. These chapters are focused on the implementation of these methods, their experimental validation, and comparison with state-of-the-art neural network-based detection approaches. A thorough analysis of their performance and practical applicability is performed to provide a comprehensive overview of their potential for real-world deployment.

3 Advanced Object Detection Techniques

The next subchapters focus on advanced methods for object detection in images, specifically the use of segmentation neural networks and one-stage and two-stage object detectors. A chapter based on Paper 3 (Annex 3) presents an innovative approach to transforming objects into centroids using a segmentation neural network that employs an encoder-decoder architecture. This approach enables more accurate head detection by transforming the input image into a map of the probability of head occurrence. Article 4 (Annex 4) extends this concept to person detection and predicts the location of center points at a pixel-accurate level using fully convolutional networks. In addition, Article 5 (Annex 5) extends this detection idea to other objects with asymmetric shapes and the detection of selected parts of them. Specifically, a new system for detecting the grip points of industrial robots was presented that uses a custom-designed fully convolutional network called ASP U-Net. Before discussing the contribution of each paper, traditional methods for object detection in images are introduced.

Conventional Approaches in CNN-Based Object Detection

Traditionally used methods for detecting objects in images using neural networks have been based on two-stage detectors. Two-stage object detectors first create a map of similar regions, known as regions of interest, and then use this map to detect objects in the image. One of the most well-known two-stage detectors is the region-based convolutional neural network (R-CNN), which uses selective search to select exactly 2,000 regions from an image and performs classification on them using a standard convolutional network (Girshick, 2014).

A more recent popular alternative is single-stage detectors. Single-stage object detectors work by directly predicting bounding boxes and their classes in a single step. This approach makes single-stage detectors generally faster, although they may achieve lower detection accuracy compared to some two-stage detectors. One of the earliest and most well-known single-stage detectors is the Single Shot MultiBox Detector (SSD), which combines the prediction of multiple differently scaled bounding boxes with classification in a single step, making it suitable for real-time detection applications (Liu, 2016). Another popular single-step detector is You Only Look Once (YOLO) (Redmon, 2016), which uses fixed-size anchor boxes as candidate regions. One major drawback of most anchor-based detectors is the need for ad-hoc heuristics to determine the number and size of anchor boxes (Law, 2020).

Object detectors based on deep neural networks typically provide object localization in the form of a bounding box around the detected instance, to which a class label is assigned (Redmon, 2017). However, other geometric shapes, such as polygons (Zhou, 2019) and circles (Nguyen, 2022), can also be used for this purpose.

3.1 Centroid-Based Person Detection

Given the visual similarity of individual heads captured from the top-down view and the trend of using segmentation methods for object detection, a transformation of the data from the input color image to a segmented image containing a mask of detected heads was considered. This approach shifted the problem from finding a bounding box to identifying approximately

circular clusters of pixels with uniform intensity. By focusing on a single object type, the transformation was further refined to a probabilistic map, where pixel intensity indicates the probability of an object belonging to a particular class. Figure 12 shows the various detection outputs discussed, including bounding boxes, clusters of pixels with the same value corresponding to a class (segmentation mask), and clusters with values that decrease progressively with distance from the object's center (probabilistic map).

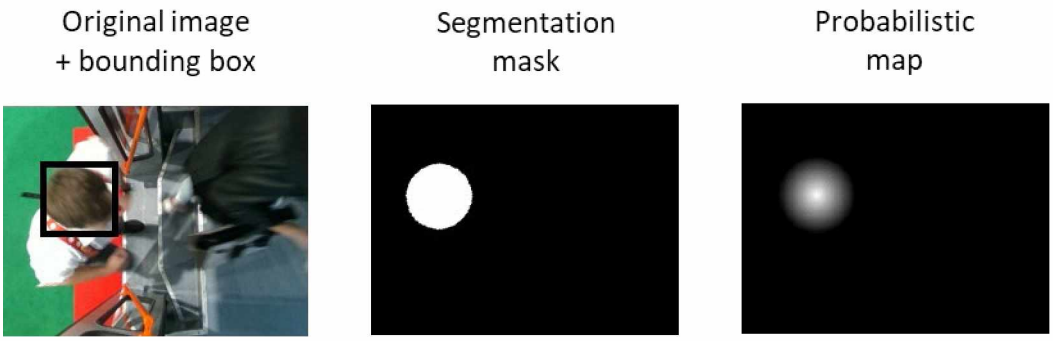


Figure 12 – Various detection outputs.

The resulting transformation of the objects into a probability map was defined by a circular pattern around the center of each head (centroid), where the pixel at the center of the head had the highest value and the values decreased with distance from the center according to the following formula (7)

$$h_x = \frac{R - d_{0x}}{R} \text{ for } d_{0x} < R, \qquad h_x = 0 \text{ for } d_{0x} \geq R, \tag{7}$$

where h_x is the current pixel value, d_{0x} is distance between current pixel and the head center, and R is the radius of the head. Due to the conversion to a probabilistic map, it is necessary to locate individual centroids to obtain the positions of individual heads. The overall scheme of the detection system is shown in Figure 13.

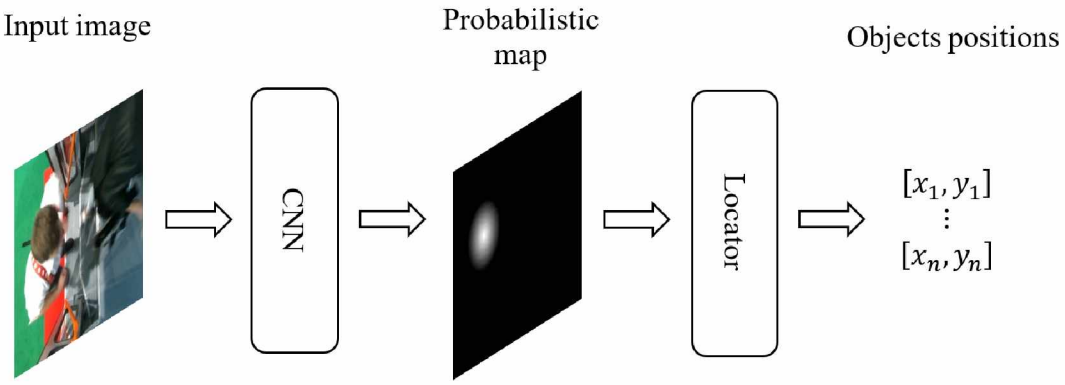


Figure 13 – Diagram of a detection system employing a transformation to a probability map.

This presented transformation was implemented with fully convolutional networks using an encoder-decoder scheme. For comparison with high-performance methods, testing was conducted with several topologies from the YOLO family. Methods of both types will be further described in the following section.

3.1.1 Methods

In the process of testing the transformation method, several iterations of experiments were conducted to create probabilistic maps using modified CNNs for object classification. The modifications to the CNN topologies for detection consisted of adding a "decoding" part after the standard classifiers. The classifiers tested included two custom architectures (Net1 and Net2) and two well-known architectures AlexNet, and LeNet-5, which were separately attached to the same decoding part. The decoding part was designed as a combination of a feedforward neural network (FFNN) and a convolutional neural network (CNN), and it converted the extracted information into an output schematic image in the form of a probability map. (Stursa, 2021b). The complete overview of the architectures used in the encoder-decoder method is shown in Figure 14.

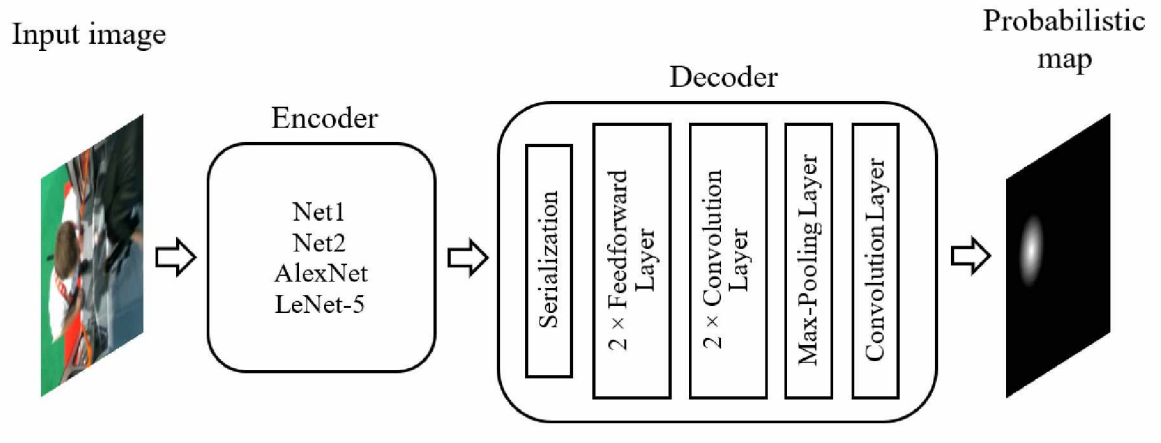


Figure 14 – Diagram of the Encoder-Decoder Approach

The encoder topologies were chosen based on previous experience and have already been described (see Convolutional Neural Networks). In addition to custom networks based on the encoder-decoder scheme, the U-Net topology was also used. U-Net inherently utilizes the same principle and is specifically designed for image data segmentation (Ronneberger, 2015).

Comparisons were further made with three topologies from the YOLO family using different backbone networks to ensure variability in testing. Specifically, the second version of YOLO (YOLOv2) was tested with the following backbone networks: SqueezeNet (Iandola, 2016), ShuffleNet (Zhang, 2018) and MobileNetV2 (Sandler, 2018) . These networks were selected due to their excellent performance at the time of testing. Detailed description can be found in original paper (Stursa, 2021b).

3.1.2 Experimental Procedure

Dataset Creation

Person detection using the described methods required specific datasets. These datasets were created keeping in mind that both methods are based on neural networks, with each dataset containing a series of input-output pairs for use in supervised learning. The inputs for both methods were two-dimensional matrices with three layers representing RGB images. The staircase configuration was considered (see Staircase Configuration). Due to the differences between the tested methods, two types of outputs were prepared. The output of the YOLO architectures is assumed to be an annotated image. Therefore, the annotation of the image was done using the Image Labeler tool in MATLAB. The output from Image Labeler was then modified into the appropriate structure required for YOLO training. A special training set was prepared for the proposed method. Specifically, the output images were created using the Image Labeler data by applying equation (7) (Stursa, 2021b).

Experimental Setup

Overall, 1 173 frames were extracted from the captured video (see 1.3). These frames were size-normalized, making them ready for input into both methods. Then, a corresponding expected output was created for each frame. The datasets were divided into two groups in a 3:1 ratio. The first group, consisting of a total of 881 input-output pairs, was randomly selected from the dataset for training the neural network. The second group, containing the remaining 292 pairs, was reserved for testing.

The YOLO architecture is well-known. The pre-trained model was conducted using modifications on the specific data according to the authors' recommendations (Redmon, 2017). For the topologies considered in the new approach, experimental parameter settings and training were required. Each of the five topologies was trained 10 times due to the stochastic nature of the training process. For comparison, the best-performing models were selected based on the minimum value of the total mean squared error. All specific parameters can be found in the original paper (Stursa, 2021b). To further illustrate the configuration and complexity of the neural network topologies used in the experiments, Table 5 provides a detailed comparison of the depth, size, and number of parameters for each network. These parameters showed the computational demands and potential performance of each architecture.

Table 5 – Relative sizes of models. Taken from (Stursa, 2021b).

Architecture	Backbone	Depth	Size (MB)	Parameters (millions)
YOLOv2	squeezenet	18	4.6	1.24
	shufflenet	50	6.3	1.4
	mobilenetv2	53	13	3.5
Encoder -Decoder	AlexNet	12	227	61
	Net1	8	285	24.9
	Net2	10	535	46.8
	LeNet	8	153	13.4
	U-Net	24	355	31

Evaluation Metrics

As this method differs from pure classification methods, it is necessary to add evaluation criteria. For detection methods, in addition to the accuracy of the predicted class, it is also necessary to determine whether the position of the predicted object is sufficiently close to the actual position of the object in the image (ground truth). This is usually evaluated using the intersection over union (IOU) metric, which assesses the overlap between the predicted and ground truth bounding boxes. A representation of this process is shown in Figure 15.

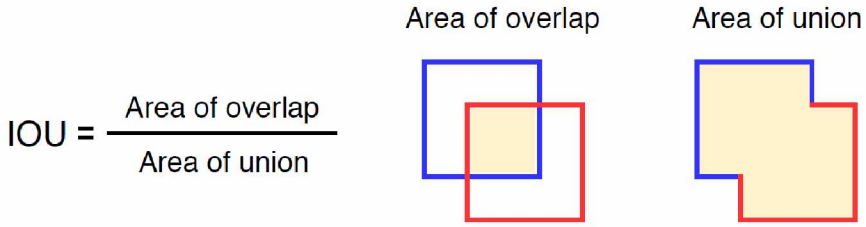


Figure 15 – Illustrative description of the IOU. Taken from (Stursa, 2021b)

Based on the thresholding of this IOU value and class matching, the classification into the primary parameters (TP, FP, TN, FN) is determined and other metrics such as precision, and recall can be evaluated (see 2.1.2).

3.1.3 Results and Discussion

The best topology of each structure was tested on a test dataset. Subsequently, IOU (accuracy), precision, and recall values were calculated with a defined threshold of 0.75. The resulting values of all selected metrics evaluated over the test set are summarized in Table 6. (Stursa, 2021b)

Table 6 – Resulting values of all the selected metrics. Taken from (Stursa, 2021b).

Metric	AlexNet	LeNet	Net1	Net2	U-Net	YOLOv2 (squeezenet)	YOLOv2 (shufflenet)	YOLOv2 (mobilenet.v2)
IOU	0.755	0.115	0.122	0.165	0.908	0.737	0.791	0.78
Precision	0.829	0.118	0.126	0.314	0.949	0.942	0.936	0.954
Recall	0.887	0.195	0.178	0.183	0.96	0.949	0.833	0.902

The results obtained in the previous table clearly showed that U-Net is the most accurate detection technique in terms of IOU, accuracy, and recall. However, the other architectures (LeNet, AlexNet, Net1, Net2) used as encoders could not outperform the YOLOv2 architecture, which is a widely accepted standard for object detection using deep learning. Moreover, Table 5 clearly showed that the number of learning parameters and the memory required to store the detector were large in the case of U-Net. As a result, the detectors used in the YOLOv2 approach are simpler and arguably more computationally efficient. (Stursa, 2021b)

3.1.4 Conclusion

Considering the results, it can be seen that the memory size and computational complexity for topologies based on encoder-decoder scheme should be optimized. Furthermore, it is clear that U-Net achieved significantly better results than the other topologies investigated. In this section, the proposed person detection method based on a deep convolutional neural network using transformation into probabilistic map was verified. However, this work was only one step in the development of a comprehensive and robust person flow monitoring system. Future work included optimization of the neural network architecture, computational testing, and of course, testing under operational conditions.

Overall, the method of transforming the input image into to a probabilistic map demonstrated significant potential. Then, the principle of transforming an image into a probability map using image segmentation was subjected to further investigation, according to the original Article 4 (Annex 4). This principle is described in the following section.

3.2 Pixel-Accurate Person Detection

In the original paper (Stursa, 2022), a centroid-based person detection technique was proposed that focuses on orthogonal scanning of a scene with variable head-to-object distance to match the staircase configuration (see 1.2). This technique introduced an efficient approach to transform scene images into localization maps, where the positions and sizes of the persons' heads were encoded into gradient ellipses that provided centroid positions for each head. The localization maps also accounted for heads partially represented in the scene. In addition to the presented method itself, topologies from the YOLO family as well as CenterNet (Duan, 2019), were used to compare the performance of the detection problem. Moreover, a new metric suitable for evaluating both centroid and bounding box predictions was proposed to account for inaccuracies in head positioning as well as false positive and false negative detections.

3.2.1 Methods

Bounding Box-based Object Detection

As mentioned, two methods that use bounding rectangles to predict the detected object were used for comparison. First, the YOLOv2 architecture using several different backbone networks were used, namely GoogLeNet (Szegedy, 2016), MobileNet-v2 (Sandler, 2018), and SqueezeNet (Iandola, 2016). The input is an expected grayscale image, and the output is a list of bounding box predictions $\hat{b}$ consisting of a 5-tuple of elements according to the following formula (8)

$$\hat{b}_i = (\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i, \hat{c}_i), \tag{8}$$

where $\hat{x}_i$ and $\hat{y}_i$ are x and y coordinates of the prediction of left-top rectangle corner respectively, $\hat{w}_i$ and $\hat{h}_i$ are width and height of the predicted rectangle respectively, and $\hat{c}_i$ is the prediction of the class of the object.

CenterNet architecture with two different backbone networks, namely ResNet101 (Zhang, 2022) and EfficientDET D0 (Tan, 2020), was selected as the second topology using a

bounding box. Compared to YOLO, these topologies use different bounding boxes consisting of 7-tuples defined by equation (9)

$$\hat{b}_i = (\hat{x}_i, \hat{y}_i, \hat{\hat{x}}_i, \hat{\hat{y}}_i, \hat{\hat{\hat{x}}}_i, \hat{\hat{\hat{y}}}_i, \hat{c}_i), \tag{9}$$

where $\hat{\hat{x}}_i$ and $\hat{\hat{y}}_i$ are x and y coordinates of the prediction of right-bottom rectangle corner respectively, and $\hat{\hat{\hat{x}}}_i$ and $\hat{\hat{\hat{y}}}_i$ are x and y coordinates of the prediction of the object centroid.

Advanced Centroid-based Object Detection

This section expands on the idea from the previous chapter of assigning importance to each pixel separately. Additionally, the asymmetry of heads, as well as heads with coverings that increase their length, were considered. Due to this, a transformation into a centroid with an ellipsoidal spread was explored. For the purposes of centroid-based detection, topologies based on the U-Net architecture and reduced variant working with smaller image sizes were used; given previous testing (Stursa, 2021b).

When converted to probabilistic maps, false detections of objects graphically similar to human heads often occurred. Therefore, in addition to modifying the centroid detection method, the method was extended with a module that evaluates false detections and refines the final localization. This extended method is shown in Figure 16.

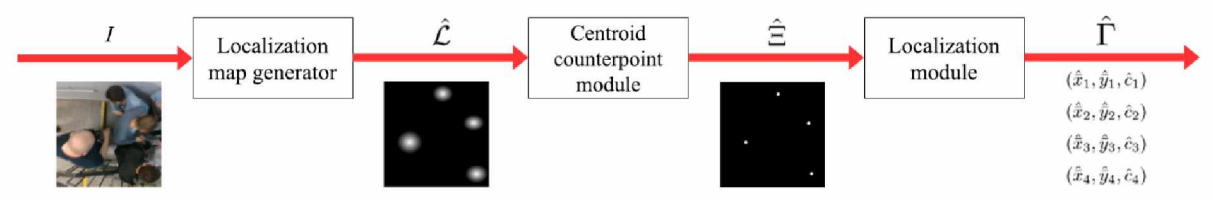


Figure 16 – Extended centroid-based object detector. Taken from (Stursa, 2022).

Figure 16 shows several apparent variables representing the individual outputs. The total output of the system is a prediction of a list of individual centroids $\hat{\Gamma}$, where each centroid γ is defined by the following equation (10)

$$\gamma_i = (x_i, y_i, c_i), \tag{10}$$

where x_i and y_i are pixel coordinates of the i-th centroid and c_i is the object class.

The localization map generator assigns a value corresponding to the probability of the object's occurrence to each pixel of the input image I. The location map $\mathcal{L}$ is defined by the following equation (11)

$$\mathcal{L}(x, y, c) = \begin{cases} \lambda(x, y, c), & \text{for } (x, y) \text{ of } c_{\text{th}} \text{ class object,} \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where $\lambda : \mathcal{L} \rightarrow (0, 1\rangle, \mathcal{L}(x, y, c) = 1$ indicates presence of the centroid of an object of the c-th class at the location (x, y), and the values decrease towards 0 with increasing distances of the elements from their centroids. The predicted localization map $\hat{\mathcal{L}}$ is further processed by

the Centroid counterpoint module to obtain a predicted centroid map $\hat{\Xi}$, where the functionality of this module will be further detailed. The centroid map Ξ is then defined by equation (12)

$$\Xi(x,y,c) = \begin{cases} 1, & \text{if a centroid of } c_{\text{th}} \text{ class is at (x,y) coordinates,} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The predicted centroid map $\hat{\Xi}$ is then converted by the localization module into a predicted list of individual centroids $\hat{\Gamma}$ defined as follows (13)

$$\hat{\Gamma} = \left(\frac{h_i}{h_\Xi}, \frac{w_i}{w_\Xi}, 1\right) \odot \underset{(x,y,c)\in S_\Xi}{\arg\max}\{\hat{\Xi}(x,y,c)\}, \tag{13}$$

where $\odot$ denotes the product of elements and $S_\Xi = X_\Xi \times Y_\Xi \times C_\Xi$, h_i and w_i are height and width of an original image I, and h_Ξ and w_Ξ are height and width of a centroid map.

The Centroid counterpoint module was introduced to evaluate false detections and refine the final localization. This module emphasizes centroids and suppresses false detections through a series of operations. Initially, each layer of the localization map is processed using a maximum filter with a kernel size of $h_K \times w_K$, resulting in a refined map $\widehat{M}$ defined by equation (14)

$$\widehat{M}(x,y,c) = \max_{(s,t)\in S_{xy}} \{\hat{\mathcal{L}}(s,t,c)\}, \tag{14}$$

where S_{xy} is a set of spatial coordinates in a rectangular sub-window of size $h_K \times w_K$, centered at point (x, y). Then, to highlight the local maxima, the predicted refined map $\widehat{M}$ is compared with the predicted localization map $\hat{\mathcal{L}}$. Where the values are equal, the resulting pixel is retained; otherwise, it is set to zero. The resulting map is labeled as $\widehat{M_1}$. The map $\widehat{M_1}$ contains the centroids as well as local maxima caused by noise of background. To suppress irrelevant regions in the map $\widehat{M_1}$, the mask $\hat{\Omega}$ (15) is used.

$$\hat{\Omega}(x,y,c) = \begin{cases} 1, & \text{if } \hat{\mathcal{L}}(x,y,c) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

To suppress artifacts in the map $\widehat{M_1}$ caused by the local maxima filter, each layer of the mask $\hat{\Omega}$ is extended by a rectangular structuring element of a defined size starting at its center, which creates an extended mask $\hat{\Omega}_\ominus$. Applying an exclusive disjunction between the extended mask $\hat{\Omega}_\ominus$ and the $\widehat{M_1}$ map creates a new $\widehat{M}_\Omega$ map. This map is used to identify centroids among the maxima highlighted in the map by considering their values in the predicted localization map $\hat{\mathcal{L}}$. Each element of the predicted localization map $\hat{\mathcal{L}}$ associated with a centroid must be greater than or equal to a threshold value t_m, where $t_m \in (0, 1\rangle$. This operation results in the centroid map prediction $\hat{\Xi}$ defined by following equation (16)

$$\hat{\Xi}(x,y,c) = \begin{cases} 1, & \text{if } \widehat{M}_\Omega(x,y,c) = 1 \wedge \hat{\mathcal{L}}(x,y,c) \geq t_m, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The setting of t_m depends on the problem and must reflect the quality of the location map predictions (Stursa, 2022). To better understand these individual processes, the principle of the module is illustrated in Figure 17.

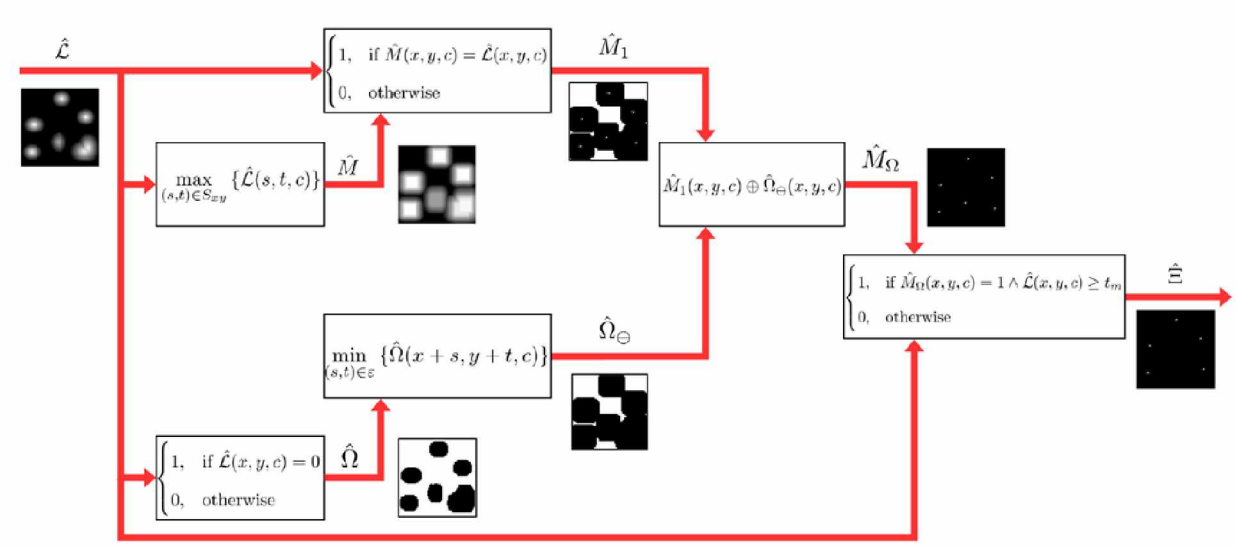


Figure 17 – Diagram of the processes in centroid counterpoint module. Taken from (Stursa, 2022).

3.2.2 Experimental Procedure

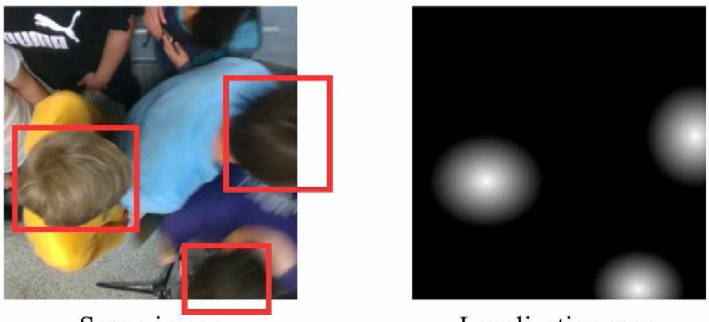
Dataset Creation

In order to evaluate the effectiveness of the proposed person detection techniques, a large dataset containing 8000 annotated images was created, which are characterized by high variability in terms of lighting conditions, background complexity, image quality, and scene height profile. Images were captured at eight different locations using an orthogonally positioned RealSense D435 camera. The acquired videos captured adults walking in a variety of environments, including stairwells, hallways, and public transportation boarding areas. The dataset was divided into a training set, a normal test set, and a blind test set. The training set consisted of 6000 images from seven different locations, which were used for both training and validation of the detectors. The normal test set D_E consisted of 1000 images from the same seven locations. The blind test set D_B consisted of 1000 images from the eighth location, which were used to evaluate the real-world capabilities of the proposed detection techniques. All images in the datasets were resized to 288 × 288 pixels (Stursa, 2022).

Annotation Process

The annotation process involved creating bounding boxes and centroid maps for the images. For detection based on bounding boxes, each head in the images was bounded by a rectangular boundary. For centroid-based detection, head positions were represented as gradient ellipses that provided the location of centroids and also accounted for heads partially visible in the scene. To create centroid maps, rectangles were drawn around the heads, which also estimated the shapes of the heads protruding from the image edges. These rectangles were then used to create gradient ellipses, with regions of the ellipse filled with values that decreased linearly from the centroid towards the edge. This method allowed the creation of a detailed

localization map for each image (Stursa, 2022). An example of this annotation process for location maps is shown in Figure 18.



Scene image Localization map

Figure 18 – Image annotation example. Taken from (Stursa, 2022)

Experimental Setup

The experimental setup compared two primary detection approaches: bounding box-based detection and centroid-based detection. The architectures for transforming input images into localization maps used the U-Net architecture and were trained from scratch while minimizing the binary cross-entropy function. A normal distribution with zero mean and a standard deviation of 0.05 was used for initialization. During training of the reduced U-Net, the maps were scaled to 72 × 72 pixels. Both U-Net variants rescaled the image and location map values to the range [0, 1]. In the centroid counterpart module, the threshold was set to 0.65 and the maximum $h_K \times w_K$ filter size was set to 10 × 10 pixels. These values were determined experimentally, and the filter size was chosen based on the most common head sizes in the images.

To train the YOLO detector on the person detection task, models pre-trained on ImageNet were used. Specifically, the layers in the GoogLeNet, MobileNet-v2, and SqueezeNet networks were replaced by the last YOLOv2 layers. In addition, since these backbone architectures expect three-channel RGB images as input, a convolutional layer with three trainable filters (3, 3) was added to transform single-channel inputs into three-channel ones. The overlap threshold at non-maximum suppression was set to 0.75, and 7 anchor fields were used, balancing performance and processing time. The widths and heights of the anchor fields were estimated using the k-means clustering algorithm and the IoU distance metric. To train the CenterNet detector for the person detection task, the ResNet-101 and EfficientDET D0 backbones were used, pre-trained on the COCO dataset. The backbones were modified to handle 288 × 288 pixels images like the YOLO detector, and the CenterNet outputs were modified for single-class detection.

Five training sessions were performed for both centroid-based and bounding box-based detectors. In each session, the localization map generators and all variants of bounding box-based detectors were trained on an identical training subset. The dataset was randomly split into training and validation subsets in a 17:3 ratio. For training, minibatches of 8 samples were used, with map generators trained for 300 epochs and bounding box-based detectors trained for 30 epochs. The models were stored and validated on a validation subset at each epoch, with

samples shuffled at each epoch. Full specifications of all parameters are provided in the original article (Stursa, 2022).

Evaluation Metrics

In order to accurately assess the performance of centroid-based object detectors, it is necessary to measure the distances between the predicted centroids and the nearest ground truth centroids. This process ensures that each prediction is paired with exactly one ground truth label and vice versa. In cases where the number of predictions does not match the number of ground truth labels, additional virtual predictions or ground truth labels are added to balance the number, and these virtual points are assigned coordinates at infinity. Overall localization error is a key metric used to evaluate detector performance. It is calculated as the average localization error over all images in the dataset. For each image, the localization error e_l is determined by summing the smallest relative distances between each ground truth centroid and its nearest predicted centroid. This process involves assigning each prediction to one ground truth label and simultaneously assigning each ground truth label to one prediction. This evaluation process is illustrated in Figure 19.

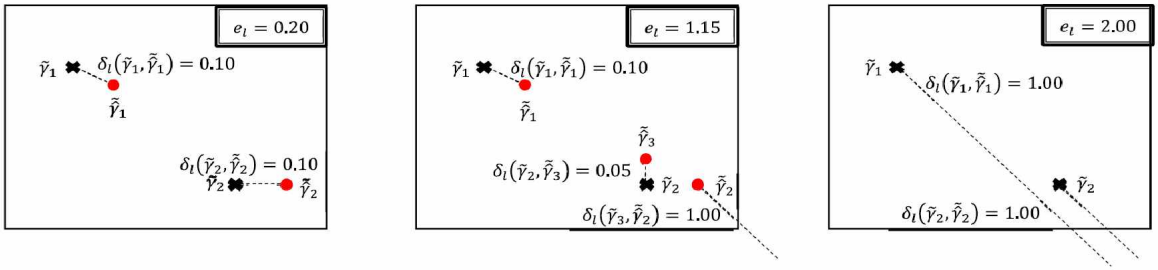


Figure 19 – Examples of calculation of image localization error. Taken from (Stursa, 2022).

The calculation involves several steps:

- For each image in the dataset, the ground truth centroids and the predicted centroids are determined.
- If the number of predictions differs from the number of ground truth labels, virtual points are added to make the counts equal.
- The relative distances between each ground truth centroid and the nearest predicted centroid are calculated.
- The sum of these distances for each frame provides the localization error for the frame.
- By averaging the localization error over all frames, the total localization error is obtained.

Mathematically, the total localization error Σe for a dataset with N frames is given by the relation (17)

$$\Sigma e = \frac{1}{N} \sum_{l}^{N} e_l, \tag{17}$$

where e_l represents the localization error for the l-th image. The relative distance between a ground truth centroid and a predicted centroid is normalized using the image dimensions. This normalization ensures that errors are consistent and comparable across images of varying sizes.

Relative inference time is another evaluation metric that measures the efficiency of the detector. This metric is defined as the ratio of the total inference time of the evaluated detector to the inference time of the baseline detector, which was chosen to be the reduced U-Net topology. The relative inference time allows comparison of the computational performance with different detectors under similar conditions.

Using these evaluation metrics - total localization error and relative inference time - it is possible to comprehensively assess the accuracy and efficiency of centroid-based object detectors. These metrics facilitate a thorough understanding of how effectively different detection methods can localize objects, while accounting for their computational complexity. A detailed description of evaluation metrics is provided in the original article (Stursa, 2022).

3.2.3 Results and Discussion

The performance of the centroid-based detectors along with competing bounding box-based detectors was evaluated using the datasets described in Dataset Creation Section, according to the experimental procedures presented in the Experimental Setup. To assess the performance of these detectors, Table 7 summarizes the resulting values of the evaluation metrics as described in Evaluation Metrics Section. A color scale was selected to highlight the best values, consistent with the other evaluation tables.

Table 7 – Evaluation results. Taken from (Stursa, 2022)

Measure	Σe	Σe	τ,–	F, FPS
Dataset	D_E	D_B	$I \times 1000$	$I \times 1000$
Full resolution U-Net	0,1472	0,3712	1,4148	8,62
Reduced U-Net	0,1352	0,3378	1,0000	12,19
CenterNet-D0	1,4659	1,7200	3,9260	3,10
CenterNet-ResNet101	1,2080	1,7090	4,0215	3,03
YOLO-GoogleNet	0,6755	1,2159	1,2074	10,10
YOLO-MobileNetv2	0,3497	1,0016	1,4378	8,48
YOLO-SqueezeNet	1,9441	1,5899	0,9355	13,03

Table 7 shows a comparison of the best models in various metrics. The overall localization errors are given for both the test dataset (D_E) and the blind dataset (D_B). The relative inference times and frame rates, which indicates the number of evaluated frames per second (FPS), are evaluated over 1000 images. Additionally, the absolute frequencies of differences between the number of ground truth labels and the number of detector predictions were demonstrated for the test dataset (Table 8) and for the blind dataset (Table 9).

Table 8 – Absolute frequencies for the test dataset D_T. Taken from (Stursa, 2022)

Absolute frequencies	<−2	−2	−1	0	1	2	>2
Full resolution U-Net	0	9	95	876	19	1	0
Reduced U-Net	0	4	84	**885**	24	3	0
CenterNet-D0	97	101	148	328	179	50	97
CenterNet-ResNet101	55	68	159	385	218	55	60
YOLO-GoogleNet	27	75	259	503	124	10	2
YOLO-MobileNetv2	1	22	172	700	97	8	0
YOLO-SqueezeNet	317	194	223	248	18	0	0

For each detector, the frequencies of multiple detections (represented by negative numbers in the first row), misses (represented by positive numbers in the first row) and correct detections (represented by zero in the first row) are displayed. The highest value of correct detections is highlighted in bold and indicates the best performing detector.

Table 9 – Absolute frequencies for the blind dataset D_B. Taken from (Stursa, 2022)

Absolute frequencies	<−2	−2	−1	0	1	2	>2
Full resolution U-Net	1	27	195	692	83	2	0
Reduced U-Net	2	14	157	**719**	104	4	0
CenterNet-D0	18	28	80	171	266	276	161
CenterNet-ResNet101	34	43	104	173	277	235	134
YOLO-GoogleNet	20	70	161	266	275	156	52
YOLO-MobileNetv2	51	96	260	343	201	40	9
YOLO-SqueezeNet	182	197	232	238	124	25	2

The results of the evaluation clearly confirm the superiority of the centroid-based person detection method over bounding box-based methods. Both variants of the centroid-based detector exhibited significantly lower localization errors on both test and blind datasets compared to the best performing YOLOv2 detector with the MobileNetv2 backbone. In particular, the reduced U-Net variant showed errors that were 2.6 to 2.9 times smaller. In addition, the reduced U-Net demonstrated excellent generalization capabilities with a high number of correct detections even on lower quality images. This variant also offered faster inference times, so that it was approximately 40% faster than YOLOv2-MobileNetv2 while maintaining accuracy. In contrast, the YOLO detectors, especially those with the SqueezeNet and GoogleNet backbones, showed a tendency for multiple detections and misdetections, contributing to higher localization error rates. CenterNet detectors exhibited similar problems with even higher errors, indicating a tendency to miss people in the images.

Reducing the map resolution in the centroid-based detector led to a slight improvement in detection performance, further improving generalization without compromising detection accuracy. Overall, these findings confirm the benefits of centroid-based detection, especially

with a reduced U-Net, in providing efficient and accurate person detection in real-time applications. Full results are available in the original article (Stursa, 2022).

3.2.4 Conclusion

The paper (Stursa, 2022) demonstrated that determining the position of head centroids using a fully convolutional network (U-Net) combined with a simple sequence of image processing operations (centroid counterpart module) is an effective method for fast and accurate detection of people in orthogonally acquired images. The centroid-based person detector not only satisfies the edge computation requirements, but also exhibits strong generalization capabilities and maintains low localization errors even on low-quality images. It performs effectively in a variety of environments, including those with significant variability in elevation profiles. The use of quarter-size localization maps instead of full-resolution maps resulted in a 40% reduction in detector inference time, with the side benefit of a slight improvement in detection performance. Furthermore, the use of bounding box-inspired annotation for dataset preparation facilitated a simplified process that allowed simultaneous annotation of images for both centroid-based and bounding box-based detection.

Based on these results, further work was done with detection based on transforming the input images into location maps. Specifically, this involved extending the original method to include multi-class localization, where the classes did not represent the objects themselves, but rather their regions of interest, which were labelled with different shaped gradients. A description of this approach is given in the following section reflecting Article 5 (Annex 5).

3.3 Detection of Significant Features in Complex Objects

Given the success of the transformation technique, its ability to encode different parts of objects into different gradient shapes was considered. These gradients represent desirable grasping points for robotic manipulation. Thus, the original paper (Stursa, 2021a) focused on the development of an efficient grasp point detection technique for robotic manipulation of complex objects. This proved to be a task of major industrial importance. Based on the successful use of the U-Net architecture in previous experiments, the network was modified to meet the specific needs of this application. The proposed method utilized a perception system based on RGB data and a fully convolutional neural network architecture, specifically an attention squeeze parallel U-Net (ASP U-Net). This neural network transformed the RGB image of the scene into a grayscale scheme where the positions of the grasping points were highlighted using gradient geometric shapes. The method was particularly effective in identifying grasping points on objects that offered multiple edges and planes suitable for manipulation by different end-effectors, such as parallel grippers and vacuum cups.

The ASP U-Net architecture was designed with computational efficiency in focus, allowing it to run on single board computers such as NVIDIA Jetson NANO while maintaining high accuracy and fast response. The performance of the network was compared to competing architectures using metrics such as generalized intersection over union and mean absolute error. The method was tested in scenarios involving complex object arrangements, including overlapping and occluded objects. The results demonstrated robustness in detecting grasping

points in challenging conditions. The results highlighted the potential of ASP U-Net for real-time industrial applications where both speed and accuracy are critical.

3.3.1 Methods

Problem Formulation

Robotic manipulation of non-trivial objects that provide different types of grasping points is a complex task. A typical industrial scenario involves multiple objects that are randomly positioned and oriented on a conveyor belt (see Figure 20). While the shapes of the objects are known in advance, their positions and orientations are not. The perceptual system of the robotic manipulator must allow for rapid decision making about the position of the grasping point while allowing for automatic replacement of the appropriate end effector. The proposed solution consists of an RGB data-driven perceptual system that detects the grasping points based on a single RGB scene image and provides all the necessary information for an end effector with 3+1 degrees of freedom.

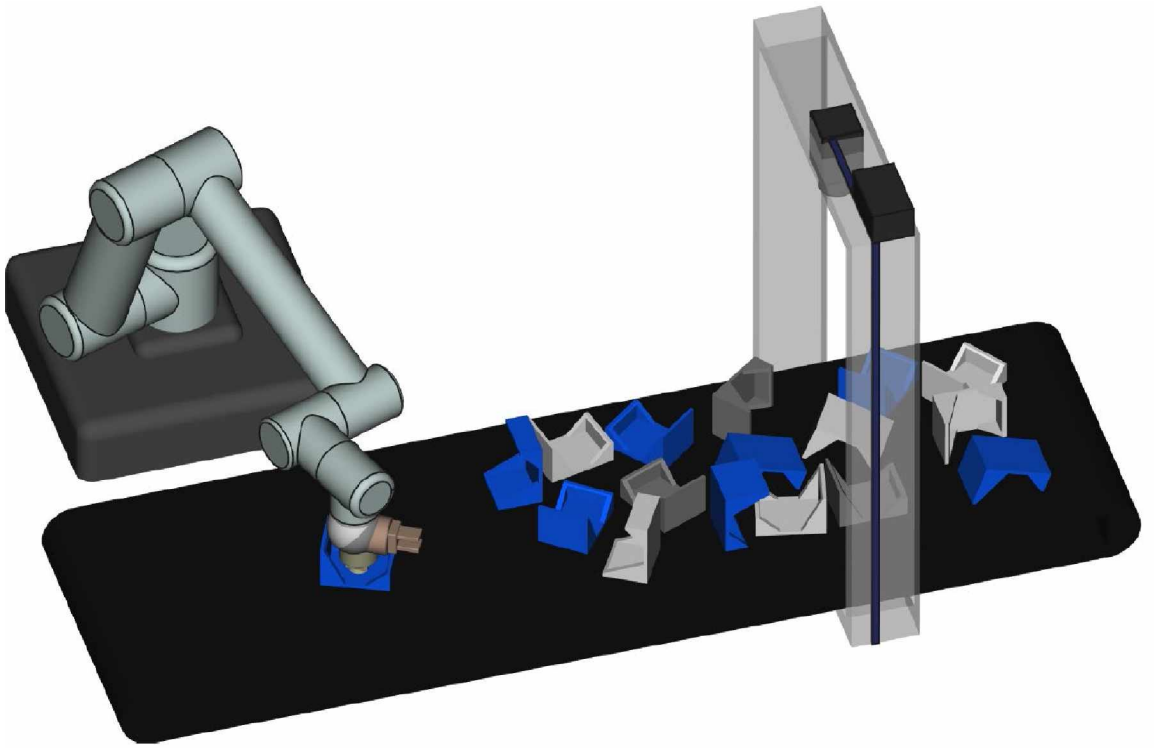


Figure 20 – Example of a robotic workplace with a perception system. Taken from (Stursa, 2021a).

Due to standard end-effector scenarios and technologies, the experiments were limited to two types of end-effectors: parallel gripper and vacuum cup. It should be noted that the method focused on objects spread out in a single-layer manner. An obvious advantage of this arrangement is the absence of the need for depth information, since all the grasping points occur at a similar vertical distance from the sensor. Therefore, the RGB data can be considered as a sufficient source of information. The non-trivial object to be manipulated has been defined and so has its possible positions on the conveyor belt. The object provided two kinds of grasping points - an edge and a plane - for manipulation by a parallel gripper or a vacuum cup depending on the position and orientation of the object. According to the possible stable positions of the

objects on the conveyor belt, five different positions with different accessibility of the grasping points of the two end effectors were considered. The following requirements were set for the grasping points to be considered feasible. A free circular area of at least 12 mm in diameter had to be available for the vacuum cup (planes). For the parallel gripper, an edge of at least 12 mm in length had to be available. In addition, for the parallel gripper, a free rectangular area of at least 10 mm on each side of the edge was required to allow the object to be grasped securely (Stursa, 2021a). The object positions considered, together with a representation of the grasping points, are shown in Figure 21.

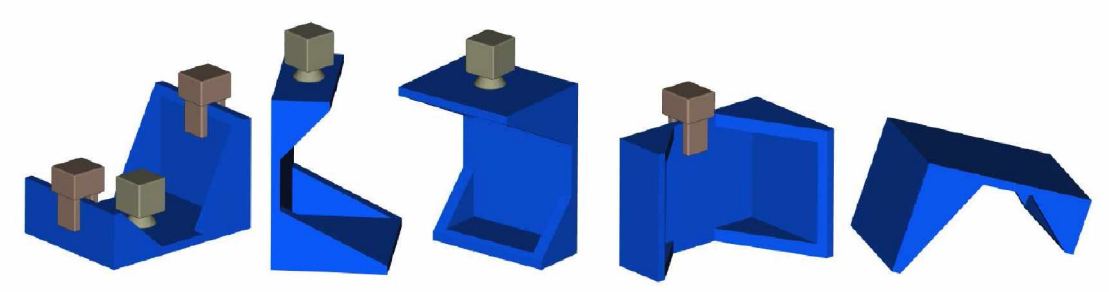


Figure 21 – Grasping points of the object at considered poses. Taken from (Stursa, 2021a).

Given these well-defined positions and the availability of grasping points, the different arrangements that may occur for objects randomly placed in a single layer on the conveyor belt were considered subsequently. These arrangements included irregular contacts between objects, including overlapping. In some cases, the space available for gripping points was reduced, while in extreme cases the gripping points were obscured and thus eliminated completely (Stursa, 2021a). Selected examples of object arrangements are shown in Figure 22.

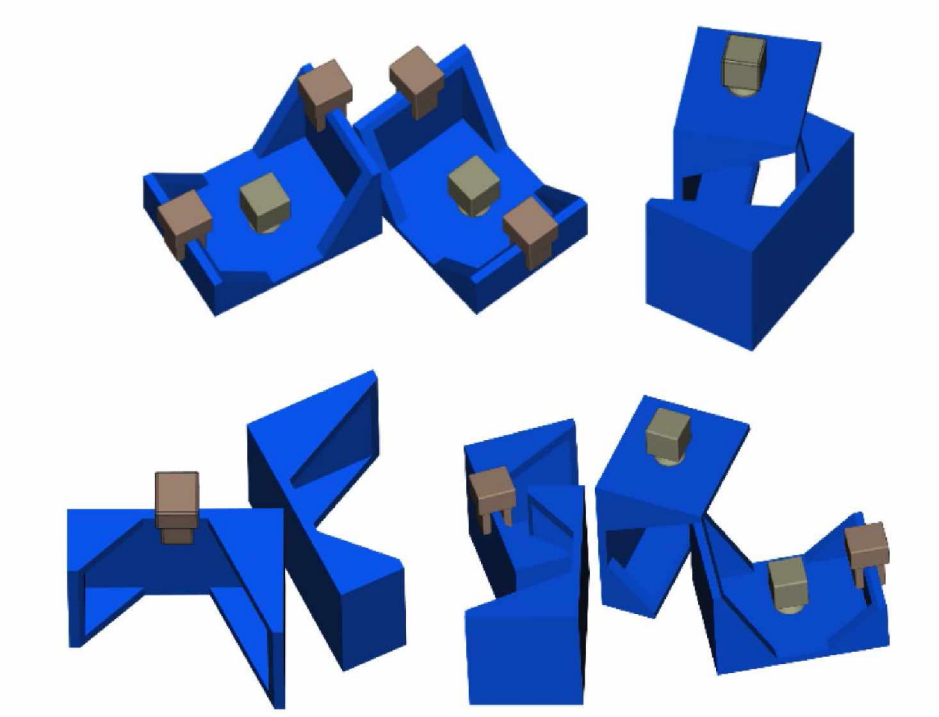


Figure 22 – Examples of object arrangements. Taken from (Stursa, 2021a).

As a result of the mentioned arrangements, the proposed perception system was expected to successfully detect only the available grasping points.

CNN for Grasping Point Detection

Based on previous research, the U-Net architecture and custom modifications were chosen as the most suitable option. Given the requirement to localize two types of grasping points, it was necessary to determine how these grasping points would be transformed for image processing purposes, in particular for segmentation.

The first transformation, specifically for the gripper points of the vacuum dish, was identical to the previous detailed localization of the human heads as described in Centroid-Based Person Detection and specifically in equation (28).

The second transformation considered was for the parallel gripper, which had to consider the angle of the planar orientation and was thus included in the labels of the surrounding pixels. Therefore, the label of each pixel was determined by considering the parallel gripper as a function of the distance of the pixel from the two endpoints of the abscissa. This approach defined both the grasping point and the angle of the parallel gripper. Specifically, for pixel x, the label h_x was defined by equation (18)

$$h_x = \left(\frac{d_{12}}{d_{1x} + d_{2x}}\right)^a \cdot \frac{1}{1 + b{d_{0x}}^c}, \tag{18}$$

where d_{12} is the length of the abscissa, i.e. the distance between the points that define the gripping point; d_{1x} is the distance between the current pixel and the first endpoint of the abscissa; d_{2x} is the distance between the current pixel and the second endpoint of the abscissa; d_{0x} is the distance between the current pixel and the midpoint of the abscissa. The parameters *a, b,* and *c* affect the size of the shape and need to be set according to the features of the scene. Using the above transformation, a spatial representation of the grasping point for the parallel gripper was obtained, where the optimal position was labelled with value 1 and the value of the label decreased with distance from it. The steepest decrease occurred perpendicular to the abscissa, while the decrease in label value was more gradual parallel to the abscissa (Stursa, 2021a). An example of the marking of the grip points for both types of end effectors is shown in Figure 23.

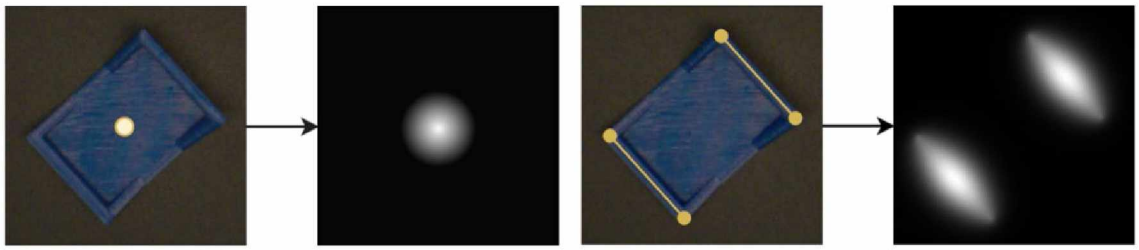


Figure 23 – Grasping point representations. Taken from (Stursa, 2021a).

With respect to the presented transformations, the neural network was required to transform the original RGB image of the scene into two grayscale schematic images where the grasping points are highlighted as previously described gradient shapes. These shapes effectively provided all the necessary information for the robotic manipulator.

ASP U-Net Architecture

The introduced ASP U-Net neural network was based on a CNN topology capable of performing the transformation shown in Figure 24. Moreover, since the perception system was designed for real-time industrial applications using single-board computer architectures, the size of the neural network and the resulting computational power requirements were also taken into account. To achieve this, the previously introduced U-Net was used as the initial architecture. The U-Net has been very successful in semantic image segmentation, especially when only small training datasets are available (Bardis, 2020). This feature has been very beneficial for embedded applications (Stursa, 2021a).

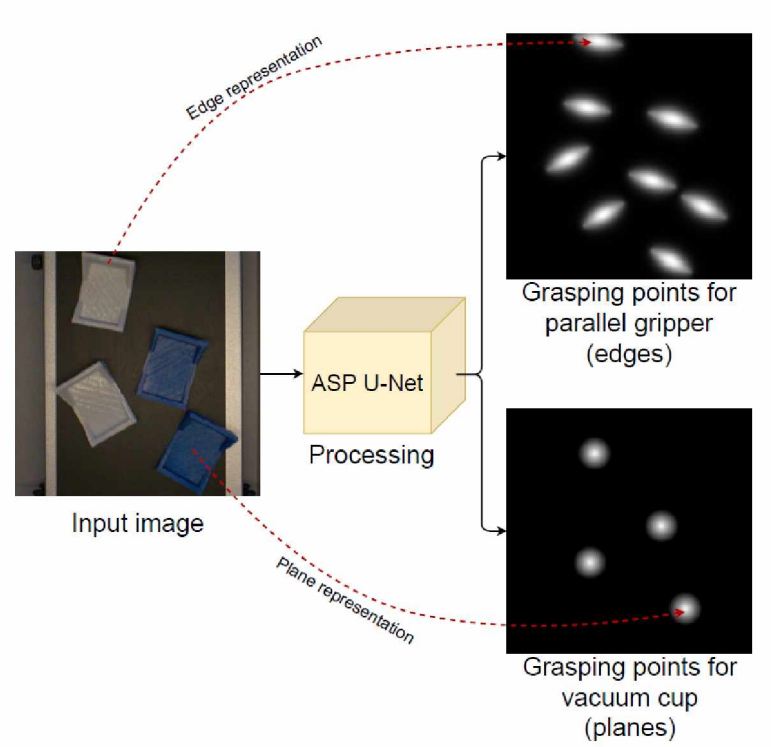


Figure 24 – Schematic representation of network outputs. Taken from (Stursa, 2021a).

However, the U-Net was defined by more than 30 million parameters with a memory size of 364 MB. Therefore, inspired by techniques like those used in SqueezeNet (Iandola, 2016) and SqueezeSegNet (Nanfack, 2018), the classic convolutional and transposed convolutional layers were replaced by layers similar to the Fire and DeFire modules, respectively. The Fire and DeFire modules were implemented in the Down sampling and Up sampling modules in the encoder and decoder parts of the U-Net. These replacements significantly reduced the number of parameters while maintaining accuracy. A detailed explanation of these modules is given in the original article (Stursa, 2021a).

Further improvements were based on the requirement of the network to label grasping points located on objects of different colors under different lighting conditions. In the decoding part of the architecture, an attention mechanism based on attention gates was implemented to filter features propagated through the skip connections. Finally, the network was designed to generate two images, each providing information about one type of grasping point. To achieve this goal, it was advantageous to parallelize the propagation of data through the network at a

point, allowing separate parts of the architecture to handle the separate processing of edges and planes. Therefore, the U part of the original U-Net architecture was replicated and mirrored, with the two parallel parts connected in the final section of the network. With this enhancement, the network was expected to adjust its parameters to handle edge-relevant elements in one parallel branch and plane-relevant elements in the other branch (Stursa, 2021a).

The resulting name ASP U-Net (short for Attention Squeeze Parallel U-Net) refers to the integration of attention mechanism, parameter reduction, parallel detection of different types of grasping points, and origin - the U-Net. All the details and specific implementations of these enhancements are thoroughly described in the original paper (Stursa, 2021a). The overall proposed ASP U-Net architecture using the previously mentioned modules is shown in Figure 25.

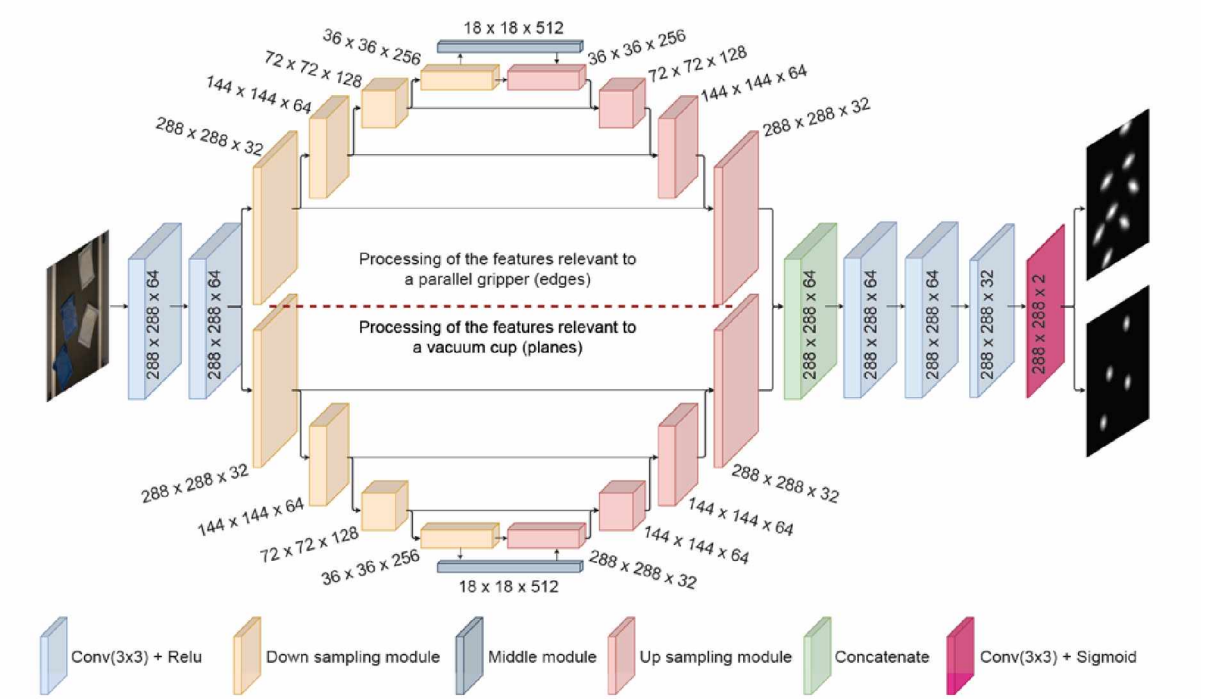


Figure 25 – ASP U-Net architecture. Taken from (Stursa, 2021a).

Architectures to Compare

To validate the ASP U-Net architecture, its performance was compared with several state-of-the-art architectures. Specifically, SegNet (Badrinarayanan, 2017), BiSeNet (Yu, 2018), U-Net, and FCN-VGG16 (Shelhamer, 2017) were implemented according to their original configurations. In addition, Squeeze U-Net (Beheshti, 2020) and Attention U-Net (Oktay, 2018) were included for comparison to match the latest trends in edge computing and attention mechanisms. In addition, some semantic segmentation neural networks based on classical convolutional neural networks were also considered as backbone networks. Specifically, ResNet 101 (He, 2016), DenseNet 121 (Huang, 2017) and MobileNet (Howard, 2017), were combined with the FCN architecture. All architectures were modified to work with the same data. The input and output layers for each architecture were replaced with the same layers used in the ASP U-Net architecture (Stursa, 2021a).

3.3.2 Experimental Procedure

Dataset Acquisition

A demonstration robotic stand was prepared to obtain training and test data. A Basler acA2500-14uc industrial RGB camera was used as the RGB sensor. This sensor could provide up to 14 5MPx RGB images per second. The camera was equipped with a Computar M3514-MP lens to view the 300 × 420 mm scanned area from a distance of 500 mm. Initially, a total of 716 images were taken for the training set. The resulting image collection contained 0 to 9 objects of three colors (blue, white, gray) in different positions and poses. Many images contained only parts of the objects. The resolution of the images was 288 × 288 RGB pixels with 8-bit depth. Each image in the training set had to be labeled to be usable for training the neural network. Specifically, the positions of the grasping points were identified, and grayscale schematic images were prepared for each image. To simplify this process, a custom labeling application (GraspLabeller) was developed to manually prepare the necessary data. The GraspLabeller application allowed the grasp points to be marked using a computer mouse and provided the resulting pair of greyscale images (Stursa, 2021a), as shown in Figure 26.

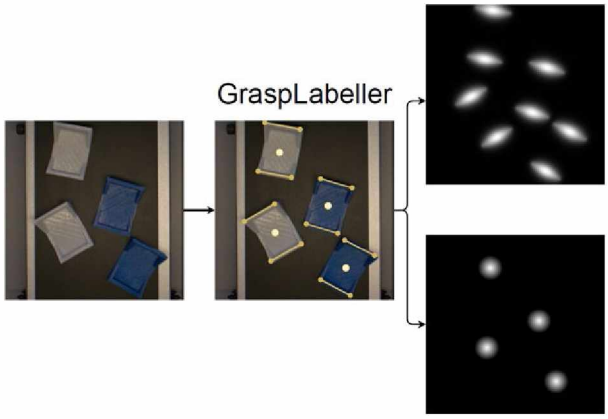


Figure 26 – Dataset labelling using GraspLabeller. Taken from (Stursa, 2021a).

The labelling procedure was carried out in accordance with the minimum dimensional requirements of the parallel gripper and vacuum cup. Considering the image resolution and the overall scene layout, the parameters in equations (7) and (18) were set as follows: $a = 10$, $b = 0.002$, $c = 2$, $R = 20$ (Stursa, 2021a).

Various data augmentation techniques have been used to augment the data sets. Geometric transformations were applied to the original RGB images and the corresponding grayscale target images to ensure label preservation. Specifically, each sample was randomly rotated by an angle between -10° and 10°. In addition, each sample was shifted by up to ±10 pixels in both horizontal and vertical directions. This procedure expanded the original dataset from 716 images to 2,148 images. A separate data set had to be created to evaluate the perceptual system. Instead of extracting a portion of the training data, a new dataset was manually created. Additional object arrangements were designed to include all possible settings. Emphasis was placed on challenging configurations such as difficult contact positions and hard-to-reach grasping points. In total, 54 unique images were collected for testing. These images contained 148 objects with 236 grasping points, of which 143 were suitable (Stursa, 2021a).

Networks Training

The ASP U-Net, along with competing state-of-the-art architectures, has been trained using the Adam optimizer, which is widely known to provide satisfactory performance. The initial weights were set randomly using a Gaussian distribution. The loss function used was the binary cross entropy, defined earlier. Thirty percent of the data set was set aside as a validation set. Training experiments were performed five times for each architecture to reduce the stochastic nature of the training and to prevent the loss function from getting stuck at a local minimum. Then, the best instances were evaluated based on the loss function in the validation set. (Stursa, 2021a).

Evaluation Metrics

After training all considered architectures, each network was evaluated. The well-known intersection over union (IoU) metric was used for accuracy. However, this metric usually evaluates detectors based on the ground truth bounding boxes and predicted bounding boxes as previously described (see 3.1.2). Since the proposed perception system did not provide bounding boxes but instead produced a grayscale image where the predicted position depended on the pixel intensity, a generalized IoU (gIoU) metric was defined. This metric was modified to evaluate overlap and union of pixel values. Since the proposed perception system provided two grayscale images as output, the metric was evaluated separately for each output, referred to as $gIoU_e$ for edges and $gIoU_p$ for planes. An example of this generalized IoU method is shown in Figure 27.

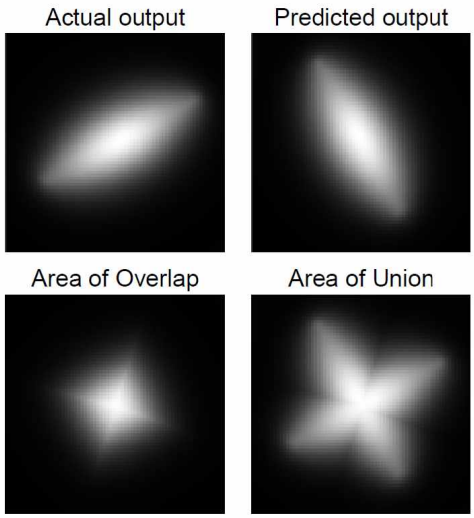


Figure 27 – Demonstration of the generalized IoU metric. Taken from (Stursa, 2021a).

In addition, a classical mean absolute error (MAE) metric was used to interpret the results. This metric measured the difference between the predicted and target outputs. Like gIoU, MAE was evaluated for each output separately, referred to as MAE_e for edges and MAE_p for planes. Equally important was the evaluation of the memory size and response time of the perception system, as it was designed for real-time industrial applications using single board computer architectures. Both the size of the neural network and its response time were evaluated on the NVIDIA Jetson NANO. All the details are thoroughly described in the original paper (Stursa, 2021a).

3.3.3 Results and Discussion

The proposed ASP U-Net was trained and evaluated five times with competing architectures using the procedure and dataset described earlier. To assess the performance of ASP U-Net against competing architectures, the $gIoU_e$, $gIoU_p$, MAE_e and MAE_p metrics were evaluated using the best performing training session for each architecture determined by the lowest value of the binary cross-entropy loss function. Next, memory size and response time were evaluated for each selected architecture using the Jetson NANO. For evaluation procedure, each original input image from the test set was first processed by a neural network. Then, these actual outputs were compared with their target outputs and the gIoU and MAE metrics were determined. After processing all samples from the test set, the overall metrics were calculated as the average of the intermediate values. The resulting values were summarized in Table 10.

Table 10 – Metrics over testing set. Taken from (Stursa, 2021a).

Architecture	$gIoU_e$	$gIoU_p$	MAE_e	MAE_p	Size, MB	T_{NANO}, s
ASP U-Net	0,8675	0,9016	0,003192	0,001103	77	0,81
Attention U-Net	0,8425	0,873	0,003806	0,001449	374	1,18
BiSeNet	0,8204	0,8539	0,003407	0,001455	463	7,39
FCN-DenseNet121	0,8285	0,8685	0,003147	0,001451	110	2,27
FCN-VGG16	0,7885	0,8031	0,003894	0,002102	933	1,83
FCN-MobileNet	0,1806	0,4938	0,019114	0,006177	93	0,33
FCN-ResNet101	0,834	0,8573	0,003605	0,001586	513	1,11
SegNet	0,8598	0,8897	0,003416	0,001243	303	3,15
Squeeze U-Net	0,8124	0,7751	0,004277	0,002323	30	0,19
U-Net	0,8494	0,8582	0,003243	0,001535	364	1,02

In addition to the results table itself, cases where the grasping locations overlap to different degrees were also selected to verify the ability of the system to reflect the grasping possibilities. The individual detection cases with metric values are shown in Figure 28.

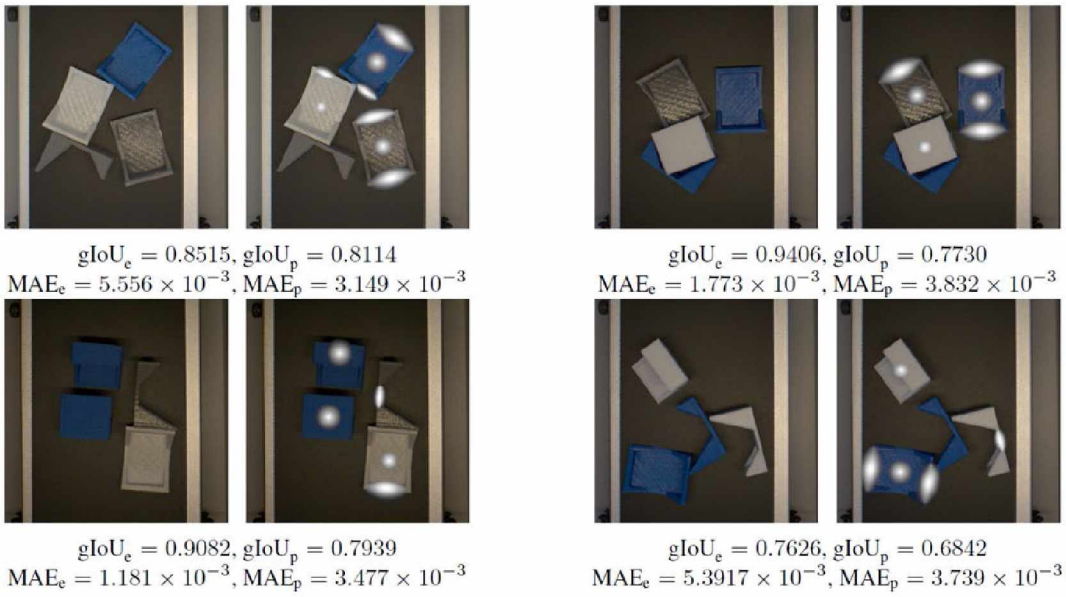


Figure 28 – Response of ASP U-Net to a scene with a group of randomly situated objects. Taken from (Stursa, 2021a).

The proposed perception system, based on the ASP U-Net architecture and a pixel-based scene transformation approach, dealt with the problem of detecting grasping points for both parallel grippers and vacuum cups. While all possible positions and interactions of objects were considered, object occlusion was simplified to a single layer. The system did not directly address 3D end-effector orientation, making it more suitable for end-effectors and fixed-position industrial robots such as SCARA.

On the test set, the results varied significantly. Considering metrics such as $gIoU_e$, $gIoU_p$, MAE_e, and MAE_p, ASP U-Net provided the best results in three of these categories. Although the BiSeNet and FCN-DenseNet121 networks performed better during the training stage in terms of the resulting error function; ASP U-Net outperformed them in the testing stage. There was a general trend where the grasping points detection metrics for the parallel gripper were lower than the metrics for the vacuum cup. U-Net and Squeeze U-Net showed similar performance for both types of grasping points. Significant differences were also observed in memory usage and response time between the architectures. Memory-efficient architectures such as FCN-MobileNet and Squeeze U-Net required much less memory than full-featured architectures. The proposed ASP U-Net with a memory requirement of less than 80 MB was also considered a memory-efficient architecture. The response time of ASP U-Net on the Jetson NANO was favorable and among the best, where it was only surpassed by very lightweight architectures. Detailed results are described in detail in the original paper (Stursa, 2021a).

The quantitative results shown in Table 10 and the examples in Figure 24 show that ASP U-Net performs well under a variety of spatial object orientations. This architecture effectively distinguishes between feasible and not feasible grasping points based on object orientation, which is one of the key features of the proposed approach. In addition, the ASP U-Net is suitable for real-time applications due to its response time. Overall, the ASP U-Net network provided competitive accuracy compared to state-of-the-art full-weight architectures while remaining usable for edge computing due to its low memory consumption and excellent response time.

3.3.4 Conclusion

In Article 5 (Annex 5), an innovative fully convolutional neural network architecture, called ASP U-Net, was proposed for use in a robotic grasping system. ASP U-Net is widely applicable for simultaneous detection of grasping points using different types of end effectors of robotic arms. The architecture was thoroughly tested for grasping points detection for a parallel gripper and a vacuum cup using a pixel-based scene transformation. Specifically, the positions and orientations of the gripper points were encoded into gradient geometric shapes that effectively capture all the necessary information for the robotic arm to manipulate objects. The performance of ASP U-Net was compared with nine competing architectures and the results proved sufficient accuracy with effective memory requirements and fast response time. The ASP U-Net output provided information about the position of all feasible grasping points in the scene. However, some objects may have parts that are more suitable for grasping than others. For example, as with a mug, it is better to grasp it by the handle than by a hot surface, so choosing the correct point is essential for safe and effective grasping. Thus, the solution

showed that this problem can be addressed in data labeling by excluding undesirable areas from being labeled as grasping points in the dataset (Stursa, 2021a).

In addition to the scientific contribution, the proposed method has been adapted and applied in an industrial environment. This application is briefly summarized in the following section.

3.3.5 Practical Implementation

The application dealt with the implementation of a computer system for brick detection in an industrial robotic application. In the manufacturing process, interlacing bricks were used to prevent fusing of the main bricks during furnace firing. The task was to determine the exact position and orientation of these interlacing bricks, which were arranged in several layers on pallets to be handled by the robot. An image of the production process in which the system was implemented is shown in Figure 29.

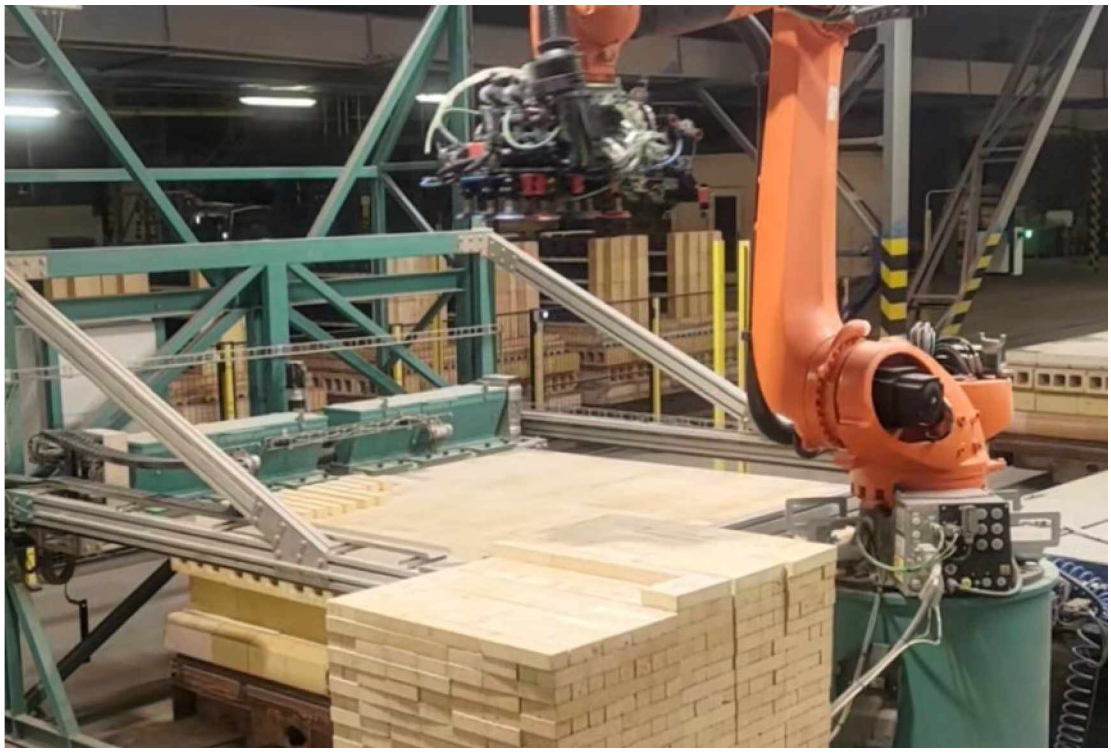


Figure 29 – Production process in which the system was implemented.

A method based on the transformation of image data into gradient ellipses, which were then identified using a standard ellipse detector in the image, was chosen as a suitable tool for position and rotation detection. This provided information on the position, rotation, and size of the objects. An example of the task is shown in Figure 30.

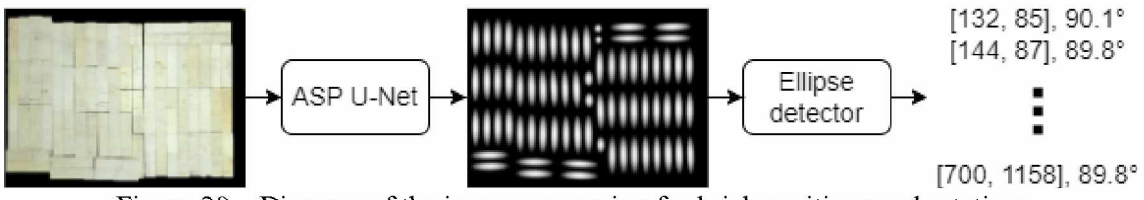


Figure 30 – Diagram of the image processing for brick positions and rotations.

For accurate detection, the ASP U-Net architecture was adapted to provide reliable brick detection in laboratory and industrial environments. The vision system used a Stereolabs ZED 2i stereo camera to capture the depth and RGB images themselves in combination with a Stereolabs ZED Box edge computing device. This setup allowed the robotic system to detect and remove bricks layer by layer with high accuracy. The brick detection system was successfully implemented in a real industrial environment where it ensured the efficiency of the robotic manipulation process. Despite the challenges of industrial environment variability, the system demonstrated the potential of machine vision in industrial automation and opened the way for further improvement and adaptation to different manufacturing environments.

Award-Winning Application

Moreover, the application was presented at a Technology Day event hosted by the University of Pardubice Technology Transfer and Knowledge Center (CTTZ). This developed technology won the first place and was awarded the best innovation at the event. This meeting bridged research with industrial partners. The competition evaluated innovative technologies on the basis of their potential for real-world application, clarity of presentation, and the marketability of the solution.

3.4 Summary

This chapter discussed advanced object detection techniques, focusing on methods that have outperformed conventional bounding box detection approaches. These methods included segmentation neural networks and one- and two-stage object detectors. The research is built on previous work presented in several papers (Appendices 3, 4, and 5), each of which contributed to the development of more accurate and computationally efficient object detection methods. One significant innovation involved transforming objects into centroids using a segmentation neural network employing an encoder-decoder architecture. This method proved particularly effective for head detection in top-down views, where input images were converted into probability maps that predicted the likelihood of an object's occurrence. Subsequent studies have extended this approach to detect other objects, including people and robot grasping points, with pixel-level localization accuracy.

4 Research Impact and Collaborations

The methods and technologies developed in this work have made significant contributions to both academia and practical applications. These methods have been successfully applied in a variety of projects, highlighting their universality and real-world impact. A notable element of success has been the collaboration with an international research group, namely the GICAP Group, and participation in cutting-edge projects. The following sections summarize the main contributions and describe how these methods have been incorporated into collaborations and industrial applications. The methods developed in this work have produced several major innovations that have expanded the possibilities in image processing.

4.1 Major Contributions of the Developed Methods

4.1.1 Transforming Image Data into Localization Maps

One of the main contributions is the approach and method of transforming image data into localization maps. This approach provides a more accurate and differentiated representation of objects, which improves detection and localization capabilities. This innovative method has proven particularly useful in tasks such as head detection from top-down views and robotic manipulation where accurate localization is critical.

4.1.2 Method for Localization Error Evaluation

A new method for localization error assessment has been developed that offers a more accurate evaluation of the performance of the detection system. This method focuses on the distances between predicted and actual object centroids and provides a fine-grained evaluation suitable for tasks requiring pixel-level accuracy. The development of this metric has improved the ability to consistently compare different detection algorithms and facilitated further improvements to proposed methods.

4.1.3 Centroid Counterpoint Module for Suppressing False Detections

Another new feature is the centroid counterpoint module, a method designed to effectively suppress false detections in complex image processing tasks. By filtering out noise and irrelevant detections, this module improves the reliability and accuracy of the detection process. This method is particularly advantageous in real-time applications where minimizing false detections is critical to maintaining system performance.

4.1.4 Development of a Custom ASP U-Net Architecture

The development of the ASP U-Net architecture is another key contribution of this work. ASP U-Net is a fully convolutional neural network designed for efficient and accurate object detection with an emphasis on resource efficiency. This architecture, which incorporates attention mechanisms and parallel processing, has proven highly effective in detecting grasping points for robotic manipulators and has been successfully deployed in several industrial projects.

4.1.5 Ability to Detect Only Relevant Features

The ASP U-Net architecture also excels in its ability to efficiently distinguish between feasible and not feasible grasping points based on object orientation. This capability is critical in applications such as robotic manipulation, where selecting the most appropriate grasping point can significantly impact performance. The architecture's ability to focus on relevant features enables it to provide accurate and actionable results, even in complex scenarios.

4.2 Industrial Applications and Real-World Impact

The methods and technologies developed in this work have been applied in many industrial contexts, demonstrating their practical value and versatility. These applications are included in a variety of sectors including surveillance, public transport, manufacturing, and robotics. This demonstrates the broad impact of the research. The following subsections provide a detailed overview of these industrial applications and their importance.

4.2.1 Person Tracking Algorithm and Intelligent Image Sensor

One of the key industrial applications of the developed people detection technology was its integration into the *Research and development of the next generation of FareOn NextGen intelligent system* project. This project was focused on the development of public transport systems through intelligent monitoring solutions. Person detection technology was used to create a sophisticated algorithm for tracking people's movements – *Person Tracking Algorithm*. In addition to the software algorithm, the technology has also been adapted into an *Intelligent Image Sensor Prototype*. This sensor was designed to extend the real-time monitoring capabilities of the system by providing highly accurate detection of people in different environments.

4.2.2 Robotic Grasping and Automated Production Line

Object detection methods developed for the identification of grasping points in robotic systems have also found practical applications in industrial automation. These methods have been used in contract research described in 3.3.5 . Another application was the project *Research and development of a modular automated production line based on innovative robotic modules and its application for the production of medical catheters*. The aim of this project was to make the production of medical catheters more efficient by use in a modular automated production line that uses robotic systems for the precise handling of medical products.

4.2.3 Smart Fencing: Airspace Object Detection

Another important application of the developed detection methods was in the field of security, specifically in monitoring the airspace around sensitive areas. The methods have been adapted for use in the *Smart contactless technology development for smart fencing* project. This project focused on the development of advanced systems for monitoring and securing airspace over critical or restricted zones using non-contact detection technologies to identify unauthorized objects or intrusions.

4.3 Collaboration with GICAP Group

A certain success factor in the research was the collaboration with the *GICAP group*, a leading research team in the field of computer vision and artificial intelligence. This collaboration provided valuable expertise and support in developing and improving the proposed methods. Through this collaboration, the research has profited from expertise in machine learning and computer vision, leading to the integration of advanced techniques such as segmentation neural networks and object detection algorithms. This collaboration not only contributed to the technical success of the research, but also extended its reach and fostered continued innovation.

4.4 New Projects and Future Directions

The author's most recent research involves two projects. Current research has focused on several areas in surveillance systems and image processing. A major project, being implemented in collaboration with Quantasoft, is the *Innovative Intelligent Video Analysis System for 3D visualization and monitoring of perimeter protection status*. This project is dedicated to improving image processing solutions for security and surveillance and uses deep learning techniques to improve real-time detection, tracking, and analysis of objects.

In addition, the author's current research is further performed in a project *Multi-sector and Interdisciplinary Cooperation in Research and Development of Communication, Information and Detection Technologies for Control and Signalling Systems (CIDET)*. This project addresses problems related to the optimization of detection methods and the processing of multidimensional and multispectral data.

Conclusions

The main goal of this dissertation was to present the developed methods based on deep neural networks for specific image processing tasks, in particular for detection, localization, and classification of people and objects in image data. The research initially focused on traditional image processing methods and gradually incorporated advanced techniques such as convolutional neural networks (CNNs). Various methodologies were investigated throughout the study, with considerable emphasis on balancing accuracy, computational efficiency, and real-time usability.

One of the key findings was the identification of histogram of oriented gradients (HOG) feature extraction combined with support vector machine (SVM) classifiers as an efficient traditional approach for person detection. Despite the simplicity of these methods compared to CNNs, they achieved high accuracy and maintained low computational complexity, making them suitable for real-time applications. Research has shown that traditional methods can compete with more complex CNN-based solutions in terms of performance after optimization, especially in controlled environments with normalized image size.

The research was further developed to explore the potential of CNNs, particularly focusing on the use of fully convolutional networks such as U-Net for person detection and localization by transforming image data into probabilistic maps. This approach has proven to be highly effective, with the U-Net based centroid detection method outperforming conventional bounding box-based detectors such as YOLOv2 in both accuracy and computational efficiency. In particular, the reduced U-Net version offered a significant reduction in inference time without compromising detection accuracy, making it suitable for edge computing applications. The ability of the centroid-based method to accurately generalize in different environments, including those with different elevation profiles, further highlighted its robustness and practical applicability.

Research has further advanced by introducing a pixel-precise person detection technique that uses centroid-based localization. This method transformed scene images into localization maps and encoded the position and size of the head as gradient ellipses for accurate centroid determination. By focusing on centroids instead of bounding boxes, the proposed approach, especially when using the reduced U-Net architecture, outperformed traditional bounding box-based methods such as YOLO and CenterNet in both accuracy and computational efficiency. In addition, this approach has demonstrated strong generalization capability in different environments with different elevation profiles. The reduced U-Net variant not only improved detection performance, but also achieved a 40% reduction in inference time – enhancing its potential for real-time person detection in edge computing scenarios. These discoveries further validated the benefits of centroid-based detection methods in practical applications, which is in line with the overall focus of the dissertation on developing robust, efficient, and accurate deep learning techniques for image processing tasks.

The final phase of this research was focused on the development of an innovative method for detecting the grasping points of complex objects. This is crucial for industrial robotic manipulation. Based on the success of previous methods, a new fully convolutional network called ASP U-Net was introduced. This architecture efficiently transformed RGB

scene images into grayscale maps, where grasping points were highlighted as gradient geometric shapes. The ASP U-Net was particularly effective in identifying grasping points on objects with multiple edges and planes, suitable for a variety of end-effectors, including parallel grippers and vacuum cups. The ASP U-Net was optimized for computational efficiency and was designed to run on single board computers such as the NVIDIA Jetson NANO, providing fast response and low memory consumption. Compared to several state-of-the-art architectures, ASP U-Net has demonstrated superior performance in terms of both accuracy and efficiency, making it a strong candidate for real-time industrial applications. This final phase of research has successfully demonstrated the potential of deep learning techniques in advanced robotic manipulation and laid the foundation for future practical implementations in industrial environments.

The contribution of this dissertation goes beyond theoretical advances and has significant implications for both academia and practical applications. One of the key achievements has been the development of methods for transforming image data into localization maps that have improved detection and localization tasks such as head detection and robotic manipulation. In addition, a new method for evaluating localization errors was introduced to enable more accurate evaluations for applications requiring pixel-level accuracy. The introduction of the centroid counterpoint module has further improved detection reliability by effectively suppressing false positives, a critical feature for real-time systems. Another significant contribution was the design of the ASP U-Net architecture, which combined attention mechanisms and parallel processing to achieve efficient and accurate object detection, especially in the identification of grasping points for robotic systems. The ability of this architecture to detect only relevant features has significantly increased its effectiveness in industrial applications.

The methods developed in this dissertation have been successfully applied in various industries, from surveillance systems to robotic automation and airspace monitoring, demonstrating their versatility and real-world impact. Collaboration with the GICAP group has also played a significant role, enhancing the research with expertise in machine learning and computer vision. In addition, the research covered in the dissertation has set the foundations for new projects, including advanced video analysis systems and optimized detection methods, which promise further innovations in image processing and detection technologies.

References

ALCANTARILLA, Pablo Fernández; BARTOLI, Adrien a DAVISON, Andrew J. KAZE Features. Online. In: FITZGIBBON, Andrew; LAZEBNIK, Svetlana; PERONA, Pietro; SATO, Yoichi a SCHMID, Cordelia (ed.). *Computer Vision – ECCV 2012. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, s. 214-227. ISBN 978-3-642-33782-6. At: https://doi.org/10.1007/978-3-642-33783-3_16.

ARRORA, Rhohan, 2020. Convolutional implementation of the sliding window algorithm. In: *Medium* [online]. USA [cit. 2021-8-8]. At: https://medium.com/ai-quest/convolutional-implementation-of-the-sliding-window-algorithm-db93a49f99a0

BADRINARAYANAN, Vijay; KENDALL, Alex and CIPOLLA, Roberto, 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. Online. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017-12-1, vol. 39, no. 12, pp. 2481-2495. ISSN 0162-8828. Available at: https://doi.org/10.1109/TPAMI.2016.2644615.

BASLER, "Basler ace," 2020, At: https://www.baslerweb.com/en/products/cameras/area-scan-cameras/ace/aca2500-60uc/.

BEHESHTI, Nazanin and JOHNSSON, Lennart, 2020. Squeeze U-Net: A Memory and Energy Efficient Image Segmentation Network. Online. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, p. 1495-1504. ISBN 978-1-7281-9360-1. Available at: https://doi.org/10.1109/CVPRW50498.2020.00190.

BOHREN, Craig F. a Eugene E. CLOTHIAUX, 2006. Fundamentals of Atmospheric Radiation. 1. Weinheim, Německo: Wiley-VCH. ISBN 978-3-527-40503-9.

BOSCH, Anna, Andrew ZISSERMAN, Xavier MUNOZ a C. SCHMID, 2007. Representing shape with a spatial pyramid kernel. *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*. New York, New York, USA: ACM Press, 2008, **30**(1), 401-408. ISBN 9781595937339. ISSN 0162-8828. At: https://doi.org/10.1145/1282280.1282340

BREIMAN, Leo; FRIEDMAN, Jerome H.; OLSHEN, Richard A. a STONE, Charles J. *Classification And Regression Trees*. Online. Routledge, 2017. ISBN 9781315139470. At: https://doi.org/10.1201/9781315139470.

BARDIS, Michelle; HOUSHYAR, Roozbeh; CHANTADULY, Chanon; USHINSKY, Alexander; GLAVIS-BLOOM, Justin et al., 2020. Deep Learning with Limited Data: Organ Segmentation Performance by U-Net. Online. *Electronics*. vol. 9, no. 8. ISSN 2079-9292. Available at: https://doi.org/10.3390/electronics9081199.

CARVALHO, Vitor H, 2012. *Image Processing: Methods, Applications & Challenges*. Nova Science. ISBN 9781620818442

CHAKRAVORTY, Pragnan, 2018. What is a Signal? [Lecture Notes]. IEEE Signal Processing Magazine. 35. 175-177. At: https://doi.org/10.1109/MSP.2018.2832195.

COMPUTAR, "Computar lenses," 2020, At: https://computar.com/product/705/M3514-MP.

DALAL, N. a B. TRIGGS. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, 886-893. ISBN 0-7695-2372-2. At: https://doi.org/10.1109/CVPR.2005.177

DUAN, Kaiwen; BAI, Song; XIE, Lingxi; QI, Honggang; HUANG, Qingming et al. CenterNet: Keypoint Triplets for Object Detection. Online. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, s. 6568-6577. ISBN 978-1-7281-4803-8. At: https://doi.org/10.1109/ICCV.2019.00667.

DOLEZEL, Petr; **STURSA, Dominik**; KOPECKY, Dusan a JECHA, Jiri. Memory Efficient Grasping Point Detection of Nontrivial Objects. Online. *IEEE Access*. 2021, roč. 9, s. 82130-82145. ISSN 2169-3536. At: https://doi.org/10.1109/ACCESS.2021.3086417.

DOLEZEL, Petr; SKRABANEK, Pavel; **STURSA, Dominik**; BARUQUE ZANON, Bruno; COGOLLOS ADRIAN, Hector et al. Centroid based person detection using pixelwise prediction of the position. Online. *Journal of Computational Science*. 2022, roč. 63. ISSN 18777503. At: https://doi.org/10.1016/j.jocs.2022.101760.

FERRARI, V., L. FEVRIER, F. JURIE a C. SCHMID, 2008. Groups of Adjacent Contour Segments for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **30**(1), 36-51. ISSN 0162-8828. At: https://doi.org/10.1109/TPAMI.2007.1144

FU, Huiyuan; MA, Huadong a XIAO, Hongtian. Real-time accurate crowd counting based on RGB-D information. Online. In: *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, s. 2685-2688. ISBN 978-1-4673-2533-2. At: https://doi.org/10.1109/ICIP.2012.6467452.

GIRSHICK, Ross; DONAHUE, Jeff; DARRELL, Trevor a MALIK, Jitendra. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, s. 580-587. ISBN 978-1-4799-5118-5. At: https://doi.org/10.1109/CVPR.2014.81.

GONZALEZ, Rafael C. a Richard E. WOODS, 2002. Digital Image Processing. 3. Londýn: Addison-Wesley Pub (Sd). ISBN 978-0-2011-8075-6.

GUILLAUME Dave, Xing, CHAO a Kishore, SRIADIBHATLA, 2010. Face Recognition in Mobile Phones. Department of Electrical Engineering, Stanford University.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing and SUN, Jian, 2016. Deep Residual Learning for Image Recognition. Online. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 770-778. ISBN 978-1-4673-8851-1. Available at: https://doi.org/10.1109/CVPR.2016.90.

HOULT, D. I. a B. BHAKAR. NMR signal reception: Virtual photons and coherent spontaneous emission. Concepts in Magnetic Resonance. 1997, str. 277-297. https://doi.org/10.1002/(SICI)1099-0534(1997)9.

HOWARD Andrew, ZHU, Menglong, CHEN, Bo, KALENICHENKO, Dmitry, WANG, Weijun, WEYAND, Tobias, ANDREETTO, Marco, HARTWIG, Adam. *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017.

HUANG, Gao; LIU, Zhuang; VAN DER MAATEN, Laurens and WEINBERGER, Kilian Q., 2017. Densely Connected Convolutional Networks. Online. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 2261-2269. ISBN 978-1-5386-0457-1. Available at: https://doi.org/10.1109/CVPR.2017.243.

IANDOLA, F.N., MOSKEWICZ, M.W., ASHRAF, K., HAN, S., DALLY, W.J., KEUTZER, K., *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size*, 2016, CoRR abs/1602.07360, At: https://doi.org/10.48550/arXiv.1602.07360

KRIZHEVSKY, Alex; SUTSKEVER, Ilya a HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. Online. *Communications of the ACM*. 2017, roč. 60, č. 6, s. 84-90. ISSN 0001-0782. At: https://doi.org/10.1145/3065386.

LAW, Hei a DENG, Jia. CornerNet: Detecting Objects as Paired Keypoints. Online. *International Journal of Computer Vision*. 2020, roč. 128, č. 3, s. 642-656. ISSN 0920-5691. At: https://doi.org/10.1007/s11263-019-01204-1.

LECUN, Y.; BOTTOU, L.; BENGIO, Y. a HAFFNER, P. Gradient-based learning applied to document recognition. Online. *Proceedings of the IEEE*. Roč. 86, č. 11, s. 2278-2324. ISSN 00189219. At: https://doi.org/10.1109/5.726791.

LIM, J.S. Two-Dimensional Signal and Image Processing. Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 478-488.

LIU, Wei; ANGUELOV, Dragomir; ERHAN, Dumitru; SZEGEDY, Christian; REED, Scott et al. SSD: Single Shot MultiBox Detector. Online. In: LEIBE, Bastian; MATAS, Jiri; SEBE, Nicu a WELLING, Max (ed.). *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016, s. 21-37. ISBN 978-3-319-46447-3. Dostupné z: https://doi.org/10.1007/978-3-319-46448-0_2.

NANFACK, Geraldin; ELHASSOUNY, Azeddine; OULAD HAJ THAMI, Rachid; ZHOU, Jianhong; RADEVA, Petia et al., 2018. Squeeze-SegNet: a new fast deep convolutional neural network for semantic segmentation. Online. In: *Tenth International Conference on Machine Vision (ICMV 2017)*. SPIE, 2018-4-13, 24-. ISBN 9781510619418. Available at: https://doi.org/10.1117/12.2309497.

NGUYEN, Ethan H.; YANG, Haichun; DENG, Ruining; LU, Yuzhe; ZHU, Zheyu et al. Circle Representation for Medical Object Detection. Online. *IEEE Transactions on Medical Imaging*. 2022, roč. 41, č. 3, s. 746-754. ISSN 0278-0062. At: https://doi.org/10.1109/TMI.2021.3122835.

NISTÉR, David a STEWÉNIUS, Henrik. Linear Time Maximally Stable Extremal Regions. Online. In: FORSYTH, David; TORR, Philip a ZISSERMAN, Andrew (ed.). *Computer Vision – ECCV 2008. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, s. 183-196. ISBN 978-3-540-88685-3. At: https://doi.org/10.1007/978-3-540-88688-4_14.

OJALA, T.; PIETIKAINEN, M. a MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Online. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002, roč. 24, č. 7, s. 971-987. ISSN 0162-8828. At: https://doi.org/10.1109/TPAMI.2002.1017623.

OKTAY, Ozan et al. Attention U-Net: Learning Where to Look for the Pancreas. ArXiv abs/1804.03999, 2018. At: https://doi.org/10.48550/arXiv.1804.03999

OLIVO, Alessandro; MATERNINI, Giulio a BARABINO, Benedetto. Empirical Study on the Accuracy and Precision of Automatic Passenger Counting in European Bus Services. Online. *The Open Transportation Journal*. 2019, roč. 13, č. 1, s. 250-260. ISSN 1874-4478. At: https://doi.org/10.2174/1874447801913010250.

PARKER, J. R. Algorithms for Image Processing and Computer Vision. John Wiley, 2010. ISBN 978-0470643853.

REDMON, Joseph; DIVVALA, Santosh; GIRSHICK, Ross a FARHADI, Ali. You Only Look Once: Unified, Real-Time Object Detection. Online. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, s. 779-788. ISBN 978-1-4673-8851-1. At: https://doi.org/10.1109/CVPR.2016.91.

REDMON, Joseph a FARHADI, Ali. YOLO9000: Better, Faster, Stronger. Online. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, s. 6517-6525. ISBN 978-1-5386-0457-1. At: https://doi.org/10.1109/CVPR.2017.690.

RONNEBERGER, Olaf; FISCHER, Philipp a BROX, Thomas. U-Net: Convolutional Networks for Biomedical Image Segmentation. Online. In: NAVAB, Nassir; HORNEGGER, Joachim; WELLS, William M. a FRANGI, Alejandro F. (ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2015, s. 234-241. ISBN 978-3-319-24573-7. At: https://doi.org/10.1007/978-3-319-24574-4_28.

ŠAJDÍKOVÁ, Martina, Patrik MAĎA a Josef FONTANA, 2018. Zrakový systém: Fyziologie a metabolické pochody zrakového vnímání. In: Funkce buněk a lidského těla: Multimediální skripta [online]. Praha: Univerzita Karlova. At: http://fblt.cz/skripta/xiii-smysly/1-zrakovy-system/.

SANDLER, Mark; HOWARD, Andrew; ZHU, Menglong; ZHMOGINOV, Andrey a CHEN, Liang-Chieh. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Online. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, s. 4510-4520. ISBN 978-1-5386-6420-9. At: https://doi.org/10.1109/CVPR.2018.00474.

SHELHAMER, Evan; LONG, Jonathan and DARRELL, Trevor, 2017. Fully Convolutional Networks for Semantic Segmentation. Online. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017-4-1, vol. 39, no. 4, pp. 640-651. ISSN 0162-8828. Available at: https://doi.org/10.1109/TPAMI.2016.2572683.

SILVA, Eduardo A.B. a Gelson V. MENDONÇA, 2005. The Electrical Engineering Handbook: Digital Image Processing. 1. Amsterdam, Nizozemsko: Elsevier. ISBN 978-0-12-170960-0.

SIEBERT, Michael a ELLENBERGER, David. Validation of automatic passenger counting: introducing the t-test-induced equivalence test. Online. *Transportation*. 2020, roč. 47, č. 6, s. 3031-3045. ISSN 0049-4488. At: https://doi.org/10.1007/s11116-019-09991-9.

SIMONYAN, K., and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, 2015, pp. 1–14.

ŠKRABÁNEK, Pavel a MAJERÍK, Filip. Detection of grapes in natural environment using HOG features in low resolution images. Online. *Journal of Physics: Conference Series*. 2017, roč. 870. ISSN 1742-6588. At: https://doi.org/10.1088/1742-6596/870/1/012004.

SKRABANEK, Pavel; DOLEZEL, Petr; NEMEC, Zdenek; **STURSA, Dominik** a XIE, Kun. Person Detection for an Orthogonally Placed Monocular Camera. Online. *Journal of Advanced Transportation*. 2020, roč. 2020, s. 1-13. ISSN 2042-3195. At: https://doi.org/10.1155/2020/8843113.

STURSA, Dominik; HONC, Daniel a DOLEZEL, Petr, 2020. *DEVELOPMENT OF IMAGE PROCESING SYSTEM FOR PERSON DETECTION*. Online. MM Science Journal. 2020-10-7, roč. 2020, č. 3, s. 4000-4006. ISSN 18031269. At: https://doi.org/10.17973/MMSJ.2020_10_2020032.

STURSA, Dominik; ZANON, Bruno Baruque a DOLEZEL, Petr, 2021. Novel Approach for Person Detection Based on Image Segmentation Neural Network. In: *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)*. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, s. 166-175. ISBN 978-3-030-57801-5. At: https://doi.org/10.1007/978-3-030-57802-2_16.

SU, Lihong a HUANG, Yuxia. Support Vector Machine (SVM) Classification: Comparison of Linkage Techniques Using a Clustering-Based Method for Training Data Selection. Online. *GIScience & Remote Sensing*. 2013, roč. 46, č. 4, s. 411-423. ISSN 1548-1603. At: https://doi.org/10.2747/1548-1603.46.4.411.

SZEGEDY, Christian; VANHOUCKE, Vincent; IOFFE, Sergey; SHLENS, Jon a WOJNA, Zbigniew. Rethinking the Inception Architecture for Computer Vision. Online. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, s. 2818-2826. ISBN 978-1-4673-8851-1. At: https://doi.org/10.1109/CVPR.2016.308.

TAN, Mingxing; PANG, Ruoming a LE, Quoc V. EfficientDet: Scalable and Efficient Object Detection. Online. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, s. 10778-10787. ISBN 978-1-7281-7168-5. At: https://doi.org/10.1109/CVPR42600.2020.01079.

THABET, Rafika, Ramzi MAHMOUDI a M.H., BEDOUI, 2015. Image processing on mobile devices: An overview. International Image Processing, Applications and Systems Conference, IPAS 2014. https://doi.org/10.1109/IPAS.2014.7043267

WEINBERGER, K.Q. and Saul, L.K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. Journal of Machine Learning Research 10(9), 2009, pp. 207−244.

WU, Xia. Design of Person Flow Counting and Monitoring System Based on Feature Point Extraction of Optical Flow. Online. In: *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*. IEEE, 2014, s. 376-380. At: https://doi.org/10.1109/ISDEA.2014.92.

YU, Changqian; WANG, Jingbo; PENG, Chao; GAO, Changxin; YU, Gang et al., 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. Online. In: FERRARI, Vittorio; HEBERT, Martial; SMINCHISESCU, Cristian and WEISS, Yair (eds.). *Computer Vision – ECCV 2018*. Lecture Notes in Computer Science. Cham: Springer International Publishing, p. 334-349. ISBN 978-3-030-01260-1. Available at: https://doi.org/10.1007/978-3-030-01261-8_20.

ZHANG, Xiangyu; ZHOU, Xinyu; LIN, Mengxiao and SUN, Jian, 2018. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. Online. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, p. 6848-6856. ISBN 978-1-5386-6420-9. Available at: https://doi.org/10.1109/CVPR.2018.00716.

ZHANG, Qi. A novel ResNet101 model based on dense dilated convolution for image classification. Online. *SN Applied Sciences*. 2022, vol. 4, no. 1. ISSN 2523-3963. At: https://doi.org/10.1007/s42452-021-04897-7.

ZHOU, Xingyi; ZHUO, Jiacheng a KRAHENBUHL, Philipp. Bottom-Up Object Detection by Grouping Extreme and Center Points. Online. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, s. 850-859. ISBN 978-1-7281-3293-8. At: https://doi.org/10.1109/CVPR.2019.00094.

Student's Publications and Research Activities

Articles Used in the Dissertation

Article 1 – Section 2.1

STURSA, Dominik; HONC, Daniel a DOLEZEL, Peter. DEVELOPMENT OF IMAGE PROCESING SYSTEM FOR PERSON DETECTION. Online. *MM Science Journal*. 2020, vol. 2020, no. 3, pp. 4000-4006. ISSN: 18031269.
At: https://doi.org/10.17973/MMSJ.2020_10_2020032.

Article 2 – Section 2.2

SKRABANEK, Pavel; DOLEZEL, Petr; NEMEC, Zdenek; **STURSA, Dominik** a XIE, Kun. Person Detection for an Orthogonally Placed Monocular Camera. Online. *Journal of Advanced Transportation*. 2020, vol. 2020, pp. 1-13. ISSN: 2042-3195. At: https://doi.org/10.1155/2020/8843113.

Article 3 – Section 3.1

STURSA, Dominik; ZANON, Bruno Baruque a DOLEZEL, Petr. Novel Approach for Person Detection Based on Image Segmentation Neural Network. Online. In: HERRERO, Álvaro; CAMBRA, Carlos; URDA, Daniel; SEDANO, Javier; QUINTIÁN, Héctor et al. (ed.). *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020). Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2021, pp. 166-175. ISBN: 978-3-030-57801-5. At: https://doi.org/10.1007/978-3-030-57802-2_16.

Article 4 – Section 3.2

DOLEZEL, Petr; SKRABANEK, Pavel; **STURSA, Dominik**; BARUQUE ZANON, Bruno; COGOLLOS ADRIAN, Hector et al. Centroid based person detection using pixelwise prediction of the position. Online. *Journal of Computational Science*. 2022, vol. 63. ISSN 18777503. At: https://doi.org/10.1016/j.jocs.2022.101760.

Article 5 – Section 3.3

DOLEZEL, Petr; **STURSA, Dominik**; KOPECKY, Dusan a JECHA, Jiri. Memory Efficient Grasping Point Detection of Nontrivial Objects. Online. *IEEE Access*. 2021, vol. 9, pp. 82130-82145. ISSN 2169-3536. At: https://doi.org/10.1109/ACCESS.2021.3086417.

Dissertation Topic Related Publications

CHOUAI, Mohamed; DOLEZEL, Petr; STURSA, Dominik a NEMEC, Zdenek. New End-to-End Strategy Based on DeepLabv3+ Semantic Segmentation for Human Head Detection. Online. *Sensors*. 2021, roč. 21, č. 17. ISSN 1424-8220. At: https://doi.org/10.3390/s21175848.

DOLEZEL, Petr; STURSA, Dominik a SKRABANEK, Pavel. On Possibilities of Human Head Detection for Person Flow Monitoring System. Online. In: ROJAS, Ignacio; JOYA, Gonzalo a CATALA, Andreu (ed.). *Advances in Computational Intelligence. Lecture Notes in Computer*

Science. Cham: Springer International Publishing, 2019, s. 402-413. ISBN 978-3-030-20517-1. At: https://doi.org/10.1007/978-3-030-20518-8_34.

DOLEZEL, Petr; STURSA, Dominik a HONC, Daniel. Rapid 2D Positioning of Multiple Complex Objects for Pick and Place Application Using Convolutional Neural Network. Online. In: *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*. IEEE, 2020, s. 213-217. ISBN 978-1-7281-9809-5. At: https://doi.org/10.1109/ICSTCC50638.2020.9259696.

DOLEZEL, Petr; STURSA, Dominik; HONC, Daniel; MERTA, Jan; ROZSIVALOVA, Veronika et al. Counting Livestock with Image Segmentation Neural Network. Online. In: HERRERO, Álvaro; CAMBRA, Carlos; URDA, Daniel; SEDANO, Javier; QUINTIÁN, Héctor et al. (ed.). *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020). Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2021, s. 237-244. ISBN 978-3-030-57801-5. At: https://doi.org/10.1007/978-3-030-57802-2_23.

DOLEZEL, Petr; STURSA, Dominik a KOPECKY, Dusan. Suitable ASP U-Net training algorithms for grasping point detection of nontrivial objects. *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*. 2022, s. 1586-1591. ISBN 978-1-6654-9607-0. At: https://doi.org/10.1109/CoDIT55151.2022.9803900.

DOLEZEL, Petr; STURSA, Dominik a KOPECKY, Dusan. Memory Efficient Deep Learning-Based Grasping Point Detection of Nontrivial Objects for Robotic Bin Picking. Online. *Journal of Intelligent & Robotic Systems*. 2024, roč. 110, č. 3. ISSN 1573-0409. At: https://doi.org/10.1007/s10846-024-02153-9.

STURSA, Dominik; DOLEZEL, Petr a HONC, Daniel. Multiple Objects Localization Using Image Segmentation with U-Net. Online. In: *2021 23rd International Conference on Process Control (PC)*. IEEE, 2021, s. 180-185. ISBN 978-1-6654-0330-6. At: https://doi.org/10.1109/PC52310.2021.9447488.

STURSA, Dominik; DOLEZEL, Petr a HONC, Daniel. Grasping Point Detection Using Monocular Camera Image Processing and Knowledge of Center of Gravity. Online. In: SILHAVY, Radek (ed.). *Artificial Intelligence Trends in Systems. Lecture Notes in Networks and Systems*. Cham: Springer International Publishing, 2022, s. 531-541. ISBN 978-3-031-09075-2. At: https://doi.org/10.1007/978-3-031-09076-9_48.

STURSA, Dominik; KOPECKY, Dusan; ROLECEK, Jiri; DOLEZEL, Petr a BARUQUE ZANON, Bruno. Classification of Polymers Based on the Degree of Their Transparency in SWIR Spectrum. Online. In: GARCÍA BRINGAS, Pablo; PÉREZ GARCÍA, Hilde; MARTINEZ-DE-PISON, Francisco Javier; VILLAR FLECHA, José Ramón; TRONCOSO LORA, Alicia et al. (ed.). *17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022). Lecture Notes in Networks and Systems*. Cham: Springer Nature Switzerland, 2023, s. 371-382. ISBN 978-3-031-18049-1. At: https://doi.org/10.1007/978-3-031-18050-7_36.

Other Academic Publications

DOLEZEL, Petr; HONC, Daniel a STURSA, Dominik. Predictive Controller Based on Feedforward Neural Network with Rectified Linear Units. Online. In: SILHAVY, Radek; SILHAVY, Petr a PROKOPOVA, Zdenka (ed.). *Intelligent Systems Applications in Software Engineering. Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2019, s. 1-12. ISBN 978-3-030-30328-0. At: https://doi.org/10.1007/978-3-030-30329-7_1.

DOLEZEL, Petr; STURSA, Dominik a HONC, Daniel. Convolutional Neural Network for Sound Processing - Study of Deployed Application. Online. In: *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019, s. 1-5. ISBN 978-1-5386-9322-3. At: https://doi.org/10.1109/RADIOELEK.2019.8733479.

DOLEZEL, Petr; STURSA, Dominik a HONC, Daniel. One Step Deep Learning Approach to Grasp Detection in Robotics. Online. In: SILHAVY, Radek; SILHAVY, Petr a PROKOPOVA, Zdenka (ed.). *Data Science and Intelligent Systems. Lecture Notes in Networks and Systems*. Cham: Springer International Publishing, 2021, s. 8-17. ISBN 978-3-030-90320-6. At: https://doi.org/10.1007/978-3-030-90321-3_2.

DOLEZEL, Petr; HOLIK, Filip; MERTA, Jan a STURSA, Dominik. Optimization of a Depiction Procedure for an Artificial Intelligence-Based Network Protection System Using a Genetic Algorithm. Online. *Applied Sciences*. 2021, roč. 11, č. 5. ISSN 2076-3417. At: https://doi.org/10.3390/app11052012.

DOLEZEL, Petr; ROZSIVALOVA, Veronika; PAKOSTA, Marek a STURSA, Dominik. Automated Dataset Enhancement Using GAN for Assessment of Degree of Degradation Around Scribe. Online. In: *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2023, s. 1454-1458. ISBN 979-8-3503-1140-2. At: https://doi.org/10.1109/CoDIT58514.2023.10284338.

DVORAK, Miroslav; DOLEZEL, Petr; STURSA, Dominik a CHOUAI, Mohamed. Genetic Algorithm-Based Task Assignment for Fleet of Unmanned Surface Vehicles in Dynamically Changing Environment. Online. *Cybernetics and Systems*. S. 1-18. ISSN 0196-9722. At: https://doi.org/10.1080/01969722.2023.2240645.

HOLIK, Filip; DOLEZEL, Petr; MERTA, Jan a STURSA, Dominik. Development of Artificial Intelligence Based Module to Industrial Network Protection System. Online. In: ARAI, Kohei; KAPOOR, Supriya a BHATIA, Rahul (ed.). *Intelligent Systems and Applications. Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2021, s. 229-240. ISBN 978-3-030-55189-6. At: https://doi.org/10.1007/978-3-030-55190-2_18.

REJFEK, Lubos; PIDANIC, Jan; STURSA, Dominik; NGUYEN, Tan N.; TRAN, Phuong T. et al. Passage Detection of a Train via a Reference Point. Online. In: TRONG DAO, Tran; HOANG DUY, Vo; ZELINKA, Ivan; DONG, Chau Si Thien a TRAN, Phuong T. (ed.). *AETA 2022—Recent Advances in Electrical Engineering and Related Sciences: Theory and Application. Lecture Notes in Electrical Engineering*. Singapore: Springer Nature Singapore, 2024, s. 119-130. ISBN 978-981-99-8702-3. At: https://doi.org/10.1007/978-981-99-8703-0_10.

ROZSIVALOVA, Veronika; DOLEZEL, Petr; STURSA, Dominik a ROZSIVAL, Pavel. Sequence of U-Shaped Convolutional Networks for Assessment of Degree of Delamination Around Scribe. Online. *International Journal of Computational Intelligence Systems*. 2022, roč. 15, č. 1. ISSN 1875-6883. At: https://doi.org/10.1007/s44196-022-00141-1.

STURSA, Dominik a DOLEZEL, Petr. Comparison of ReLU and linear saturated activation functions in neural network for universal approximation. Online. In: *2019 22nd International Conference on Process Control (PC19)*. IEEE, 2019, s. 146-151. ISBN 978-1-7281-3758-2. At: https://doi.org/10.1109/PC.2019.8815057.

STURSA, Dominik; HAVLICEK, Libor; KUPKA, Libor a DOLEZEL, Petr. IMC Strategy Using Neural Networks for 3D Printer Bed Temperature Control. Online. In: SILHAVY, Radek; SILHAVY, Petr a PROKOPOVA, Zdenka (ed.). *Software Engineering Perspectives in Intelligent Systems. Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2020, s. 979-989. ISBN 978-3-030-63321-9. At: https://doi.org/10.1007/978-3-030-63322-6_84.

STURSA, Dominik; HAVLICEK, Libor a KUPKA, Libor. Robotic Sorting Line Model Using Coloured Petri Net. Online. In: SILHAVY, Radek (ed.). *Software Engineering and Algorithms. Lecture Notes in Networks and Systems*. Cham: Springer International Publishing, 2021, s. 709-717. ISBN 978-3-030-77441-7. At: https://doi.org/10.1007/978-3-030-77442-4_59.

STURSA, Dominik; DOLEZEL, Petr a HONC, Daniel. Basic Urinal Flow Curves Classification with Proposed Solutions. Online. In: ARAI, Kohei; KAPOOR, Supriya a BHATIA, Rahul (ed.). *Intelligent Systems and Applications. Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, 2021, s. 737-746. ISBN 978-3-030-55179-7. At: https://doi.org/10.1007/978-3-030-55180-3_56.

STURSA, Dominik; DOLEZEL, Petr a MERTA, Jan. Airspace Object Detection Above the Guarded Area Using Segmentation Neural Network. Online. In: ARAI, Kohei (ed.). *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2. Lecture Notes in Networks and Systems*. Cham: Springer International Publishing, 2022, s. 283-292. ISBN 978-3-030-89879-3. At: https://doi.org/10.1007/978-3-030-89880-9_22.

Utility Models

K2 Machine s.r.o., University of Pardubice. *Linka pro výrobu dvouplášťových katetrů.* Inventors: KOT, Michal; STURSA, Dominik. Czech Republic. IPC: B65G37/00, B25J9/16. Utility Model, CZ 37329 U1, application no. 2023-41237, registered on September 25, 2023. At: https://isdv.upv.gov.cz/webapp/resdb.print_detail.det?pspis=PUV/41237&plang=EN.

Research Projects

Mezisektorová a mezioborová spolupráce ve výzkumu a vývoji komunikačních, informačních a detekčních technologií pro řídicí a zabezpečovací systémy

EN: Intersectoral and Interdisciplinary Collaboration in Research and Development of Communication, Information, and Detection Technologies for Control and Security Systems

Date: April 1, 2024 - November 30, 2028

Program: Operační program Jan Amos Komenský - výzkum

Provider: Ministerstvo školství, mládeže a tělovýchovy

Position in the project: Co-researcher

Inovativní inteligentní video analytický systém pro 3D vizualizaci a monitorování stavu ochrany perimetru

EN: Innovative Intelligent Video Analytical System for 3D Visualization and Monitoring of Perimeter Protection Status

Date: January 1, 2024 - June 30, 2026

Program: TREND - 10. VS

Provider: Technologická agentura České republiky

Position in the project: Co-researcher

Výzkum a vývoj modulární automatizované výrobní linky na bázi inovovaných robotických modulů a její aplikace na výrobu lékařských katetrů

EN: Research and Development of a Modular Automated Production Line Based on Innovative Robotic Modules and Its Application in the Manufacture of Medical Catheters

Date: May 1, 2021 - May 31, 2023

Program: Aplikace

Provider: Ministerstvo průmyslu a obchodu

Position in the project: Lead researcher

Výzkum a vývoj nové generace inteligentního systému FareOn NextGen

EN: Research and Development of the Next Generation Intelligent System FareOn NextGen

Date: May 1, 2021 - May 31, 2023

Program: Aplikace

Provider: Ministerstvo průmyslu a obchodu

Position in the project: Co-researcher

Vývoj bezkontaktní technologie pro inteligentní ochranu zájmových prostor

EN: Smart contactless technology development for smart fencing

Date: January 1, 2020 - December 31, 2022

Program: INTER-ACTION

Provider: Ministerstvo školství, mládeže a tělovýchovy

Position in the project: Co-researcher

Spolupráce Univerzity Pardubice a aplikační sféry v aplikačně orientovaném výzkumu lokačních, detekčních a simulačních systémů pro dopravní a přepravní procesy (PosiTrans)

EN: Cooperation in Applied Research between the University of Pardubice and companies, in the Field of Positioning, Detection and Simulation Technology for Transport Systems (PosiTrans)

Date: October 1, 2018 - June 30, 2022

Program: Operační program Výzkum, vývoj a vzdělávání (OP VVV) - výzkum

Provider: Ministerstvo školství, mládeže a tělovýchovy

Position in the project: Co-researcher

Other Projects

Studijní program Automatizace (SPAUT)

EN: Study Program Automation (SPAUT)

Date: July 1, 2022 - June 30, 2024

Program: Národní plán obnovy - SC B

Provider: Ministerstvo školství, mládeže a tělovýchovy

Position in the project: Co-Investigator

INnoVation and ENTrepreneurship in HEIs (INVENTHEI)

Date: October 1, 2022 – December 31, 2022

Program: EIT HEI Initiative

Provider: EU – European Union

Position in the project: Co-Investigator

INnoVation and ENTrepreneurship in HEIs (Deep INVENTHEI)

Date: February 1, 2024 - July 31, 2024

Program: EIT HEI Initiative

Provider: EU – European Union (EIT Health, EIT Raw materials)

Position in the project: Co-Investigator

List of Annexes