Novel Approach for Person Detection Based on Image Segmentation Neural Network*

Dominik Stursa[1] [0000−0002−2324−162X], Bruno Baruque Zanon[2] [0000−0002−4993−204X], and Petr Dolezel[1] [0000−0002−7359−0764]

[1] University of Pardubice, Studentska 95, 53210 Pardubice, Czech Republic
[2] University of Burgos, Calle Don Juan de Austria 1, 09001 Burgos, Spain

Abstract. With the rise of the modern possibilities in computer science and device engineering, as well as with growing population in big cities among the world, a lot of new approaches for person detection have become a very interesting topic. In this paper, two different approaches for person detection are tested and compared. As the first and standard approach, the YOLO architectures, which are very effective for image classification, are adapted to the detection problem. The second and novel approach is based on the encoder-decoder scheme causing the image segmentations, in combination with the locator. The locator part is supposed to find local maxima in segmented image and should return the specific coordinates representing the head centers in the original image. Results clearly report this approach with U-Net used as encoder-decoder scheme with the locator based on local peaks as the more accurately performing detection technique, in comparison to YOLO architectures.

Keywords: Person detection · Convolutional neural network · YOLO.

1 Introduction and related work

With the rise of modern possibilities in computer science and device engineering as well as with growing population in big cities among the world, a lot of new approaches for person detection have become a very interesting topic. Person detection has an indispensable importance and is still increasingly needed for purposes of surveillance systems, safety in public transport, optimization in transport planning etc. As an initial part of person tracking, more accurate detection methods need to be developed.

These days, problems about person localization, detection and tracing are more in focus from academic and also corporate experts. Various researched approaches to person detection are based on radar sensors [1], 3D scanners [2] or infra-red sensors [3]. However, these approaches often fail to detect every human

* Supported from ERDF/ESF. "Cooperation in Applied Research between the University of Pardubice and companies, in the Field of Positioning, Detection and Simulation Technology for Transport Systems (PosiTrans)" (No. CZ.02.1.01/0.0/0.0/17_049/0008394).

passing through and are not able to track people precisely. For these difficulties, person tracking systems are still more often implemented using video processing algorithms and computer vision techniques [4].

For person detection, not only technologies and methods interfere with implementation possibilities, but also laws of the country where the detection system will be applied. As such, methods, where identification of the person is not possible, are more attractive for the corporate environment. Thus, a monitoring system placed above passing humans should naturally solve the mentioned difficulty as shown in Fig. 1.

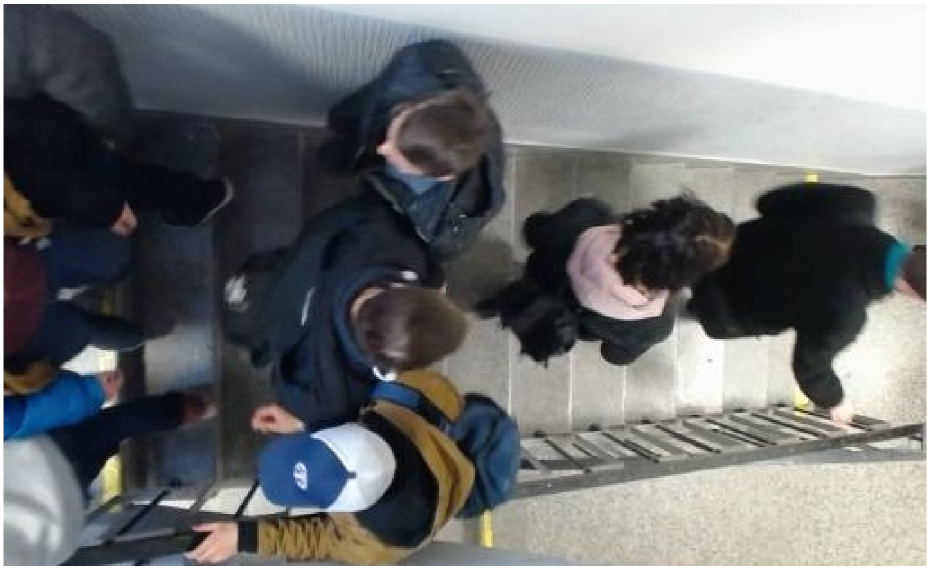

Fig. 1. Image captured from above heads (high angle - people cannot be identified).

As the view at each person is significantly limited, the main features to detect are heads and shoulders. Only a few approaches for person detection, tracking and counting with the video acquisition system placed above heads of people, have been proposed. Gao et al. [5] provide a technique combining convolutional neural networks and cascade Adaboost methods. The method based on combination of classical RGB and depth camera have been used in [6]. Both mentioned articles do not consider a strict vertical downward frame acquisition.

Detection in the image captured from above the head with image feature extraction using a histogram of oriented gradients in combination with pattern recognition network or SVM as a classifier, is shown in previous authors publication [7].

Sun et al. [8] proposed a method that utilizes the depth video stream and computes a normalized height image of the scene after removing the background.

The height image is a projection of the scene depth below the camera, which helps for better segmentation of the scene. Therefore, based on the results [8], the scene segmentation seems to be a possible approach for object detection.

The paper is structured as follows. Firstly, the problem is properly formulated. In the following section, the used methods are described and the dataset acquisition is illustrated. Then, the experiments along with the results are presented and discussed. The article is finished with the conclusions.

2 Problem formulation and methodology

Vertical downward frames captured with a monocular camera create a special scene. If the camera sensor is placed indoors and statically at one place, the scene is composed by several different types of objects, which could be divided into a few categories like person, bags, floors, railings etc.

Hence, the idea of this work is to propose a novel method for human detection and positioning based on image segmentation neural networks and compare it to a well-established approach in order to develop an efficient person detector in real-life RGB images. As such, the provided images are supposed to be derived from a video, captured from above the person heads. Inputs of the detector, in the case considered in this contribution, are size normalized RGB images cropped from a video.

The process of person detection is realized and tested by two different methods based on neural networks. The comparative method uses the neural network architectures for object detection called the YOLO, where the object and its position are represented by a class and its bounding box defined by the specific coordinates. The proposed method is based on the encoder-decoder scheme, providing the image segmentation, supported with the locator. The locator part finds local maxima in segmented image and returns the specific coordinates representing the head centers.

2.1 Novel approach for person detection

As mentioned above, our approach is composed of three parts: encoder, decoder and locator. The encoder part is based on a convolutional neural network (CNN) for image classification. Several topologies have been tested for selecting the one with best results to compare with YOLO architectures. The decoder part is created as the combination of feedforward neural network (FFNN) and CNN. The main reason for this structure is to create a segmented image where only heads are highlighted as radial gradients. Finally, the function returning coordinates of local peaks in image is used as a locator. For finding the local maxima, a maximum filter is used. This dilates the original image and merges neighboring local maxima closer than the size of the dilatation. Coordinates, where the dilated image is equal to the original image, are returned as a local maxima. The complete system is shown in Fig. 2. In this paper, the encoder-decoder part is examined in detail.

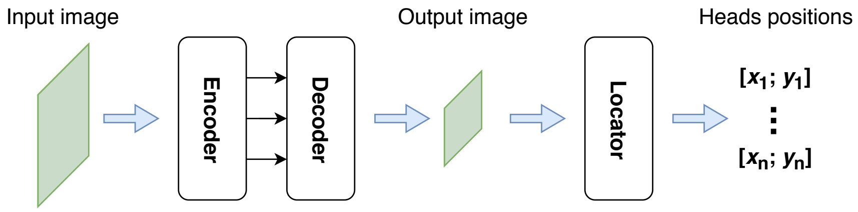

Fig. 2. Scheme of novel approach for person detection.

2.2 YOLO architectures

In order to compare the proposed approach with a state-of-the-art model, we have selected the YOLO model [9]. The reasons for this decision are derived from the problem to tackle in the long term: the detection of person on a stream of video images. This task is on the one hand straightforward, as we only want to distinguish if there is a person present in the image or not, and on the other hand demanding in the sense that it must be completed in real time. YOLO is one of the fastest architectures for object recognition in images, while at the same time has a very similar correct recognition rates as more complex neural models, even on multi-class problems.

The main idea of the architecture is to frame the recognition problem as a regression one, instead of defining a convolutional window to swipe the image looking for recognizable shapes. This allows calculation of both the bounding box of the object to be recognized and the confidence percentage of that object which belongs to a certain class. The initial YOLO model is a deep neural network architecture which includes originally 24 convolutional layers that are used to extract the main features from the image and serve as inputs to 2 fully connected layers that are utilized to predict the final coordinates for objects and their probabilities of belonging to a class.

In our experiments we used the YOLOv2 model, which is a modification for faster performance of the original model [10]. The YOLOv2 includes many improvements to the initial architecture, in order to improve both its performance and its computational complexity. The main modification is the inclusion of pre-calculated anchor boxes, used to simplify calculations. Instead of predicting the coordinates of the bounding boxes of an object, the task is now to calculate the offset for one of the given boxes to match the detected object. This allows large simplification of the fully connected layers of the second phase of the model. In order to adapt the pre-calculated anchor boxes for a given recognition problem, a simple algorithm is used as a previous phase for the training of the model. This involves calculating a standard k-means on the sizes of the labeled objects in the images, to determine the most common sizes of the objects to detect.

In the original work, the base model for YOLOv2 includes 19 convolutional layers and 5 maxpooling layers alternated between them. This architecture uses batch normalization to stabilize training, speed up convergence, and regularize the model.

3 Dataset creation

For the purposes of human detection with mentioned methods, specific datasets were created. As both methods are based on neural networks, each dataset was composed of an input-output pair series for use in supervised learning.

Two-dimensional matrices with three layers, representing RGB picture, were used as the input for both methods. The video sequence with person walking on the staircase was captured with a monocular camera. Then, the frames with significant shift between person head positions in two consecutive frames were selected. Selected frames were cropped and resized for the purposes of tested neural network architectures. Due to the difference between the tested methods, two types of outputs were prepared.

A labeled picture is supposed to be the output of the YOLO architectures. Therefore, the picture labeling was performed using the MATLAB tool called Image Labeler. The output from the Image Labeler was then modified to a proper structure necessary for the training of the YOLO.

For the proposed method, a special training set was prepared. In particular, output images were created, where every supposed center of a head was labeled by the value 1 and the surrounding values within the defined radius were gradually decreased to zero. The input and enlarged picture of the output for encoder-decoder training is shown in Fig. 3.

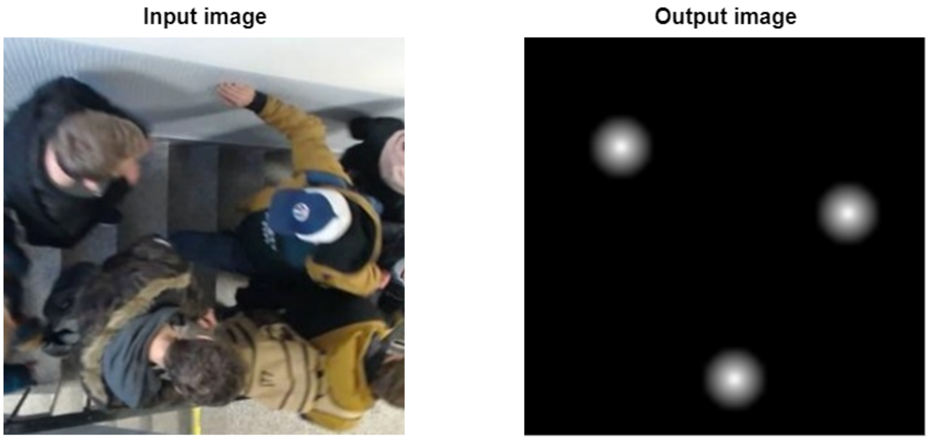

Fig. 3. Input-output pair for training of the encoder-decoder part of a novel approach.

4 Experiment procedure

For both methods, specific datasets were created. Eventually, 1173 images from the captured video were selected. Images were size normalized, which made them ready as inputs for both methods. Then, for the every image, a corresponding expected output was created.

Datasets were split into 2 groups with the ratio of 3 to 1. The first group with a total of 881 input-output pairs was randomly selected from the dataset for the purposes of neural network training. The second group with a remaining 292 pairs was left for testing.

The YOLO architecture is well known and tested by its authors in [9], [10]. Therefore, the training was realized for several of these structures with the specific data.

On the other hand, the topology considered in the case of the novel approach had to be tested first. Thus, totally 5 possible topologies were selected. Every topology was tested and evaluated 10 times. The total mean square error, defined as follows, was used as the metric.

$$E_{val} = \frac{1}{n \cdot N} \sum_{i=1}^{N} \sum_{j=1}^{n} [y_i(j) - \hat{y}_i(j)]^2, \tag{1}$$

where N is the number of the output samples in the testing set, n is the number of pixels in output, $y_i(j)$ is the desired value of pixel in the i^{th} output, and $\hat{y}_i(j)$ is the predicted value of pixel in the i^{th} actual output from the net.

All of the best tested topologies were then selected for comparison with the YOLO architecture.

4.1 Tested encoder-decoder topologies

The encoder-decoder part of the approach was tested in two ways. In the first structure, the decoder part remained the same and the encoder part was progressively replaced by 4 tested topologies in total. The decoder part was composed of a flatten layer, two dense layers reshaped to a rectangle for possible connections to convolutional part, two convolution layers, max-pooling layer, and a convolution layer providing the output picture.

Encoder topologies were selected based on authors previous experience. Net1 consists of two convolutional and one max-pooling layers. Net2 is similar, but it contains a more complex sequence of anterior layer. Both networks were adapted from [11]. The third is one of the pioneering architectures - LeNet-5 [12], [13], while the fourth is probably the most cited topology - AlexNet [14].

The second structure was based on the U-Net, which is a symmetric fully convolutional network originally used for image segmentation in the medical sector and was defined in [15].

The best results of each tested structure with relative sizes of the used networks are shown for comparison in Table 1.

Table 1. Resulting values of metric (1) for every tested structure

Network	E_{val}	Depth	Size	Parameters (Millions)
AlexNet	6.93E-03	12	227 MB	61.0
Net1	9.49E-03	8	285 MB	24.9
Net2	8.58E-03	10	535 MB	46.8
LeNet	9.39E-03	8	153 MB	13.4
U-Net	3.82E-03	24	355 MB	31.0

4.2 Tested YOLO architectures

As discussed in sub-section 2.2 the main architecture of the YOLOv2 approach includes several CNN initial layers for feature extraction and some final layers to perform the detection, as a particular type of regression task. With this architecture in mind, we have completed different tests using several pre-trained CNNs used for feature extraction present in literature for the first stage, while keeping the layers of the second stage unchanged. By choosing among the fastest performing CNNs and given that the recognition task is much simpler than those that initially designed for (multi-class detection), we expect to obtain a reasonable trade off between accuracy and low complexity for the recognition task. Relative sizes of the networks used are detailed for comparison in Table 2 including AlexNet, which was not used, but is included as a reference. All have been pre-trained on a subset of the ImageNet database [16] and then were performed a transfer learning operation using the same training set used in their encoder-decoder counterparts.

5 Results and discussion

The aim of this section is to evaluate both tested approaches represented by the YOLO and the novel approach.

5.1 Metrics definition

At first, the overlap between two bounding boxes is defined and called the intersection over union (IOU). The ground truth bounding box and predicted

Table 2. Relative sizes of pre-trained models used as the part of the YOLO models tested.

Network	Depth	Size	Parameters (Millions)
squeezenet	18	4.6 MB	1.24
shufflenet	50	6.3 MB	1.4
mobilenetv2	53	13 MB	3.5
AlexNet	8	227 MB	61.0

bounded box are necessary to know in order to evaluate this metric. The intersection is given by dividing the overlapping area of these bounding boxes with the area of union between them, as shown in Fig. 4.

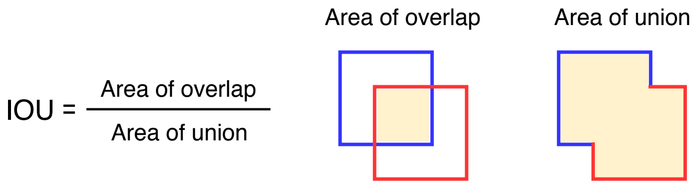

Fig. 4. Illustrative description of the IOU.

In addition, precision and recall are considered for further evaluation. The precision represents ability of a model to identify only the relevant objects. Hence, the percentage of correct positive predictions is given by following equation

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

The ability of a model to find all the relevant cases is called recall. It represents the percentage of true positive detection among all relevant ground truths given by following equation

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

In the equations above, TP means true positive, FP means false positive and FN means false negative.

5.2 Results

The best topology of every structure was tested over testing dataset. Then, the IOU (accuracy), precision and recall was calculated with defined threshold of 0.75. The resulting values of all the selected metrics, evaluated over the testing set, are summarized in Table 3.

Table 3. Resulting values of all the selected metrics

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	0.780
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	0.960	0.949	0.833	0.902

5.3 Discusion

Results obtained in the previous section clearly report U-Net, with the locator based on local peaks, as the most accurately performing detection technique in terms of IOU, precision and recall. However, other architectures (LeNet, AlexNet, Net1, Net2) used as encoders, fail to over-perform the YOLOv2 architecture, which is a generally accepted standard for object detection using deep learning. Furthermore, Table 1 and 2 obviously indicate, that the number of parameters for learning, as well as the memory necessary to store the detector, is unnecessarily big in the case of U-Net. Hence, the detectors used by the YOLOv2 approach are simpler, and probably, more computationally efficient.

Therefore, future work needs to include several elements in order to provide satisfactory grounds for the introduced approach. Firstly, the U-net encoder-decoder architecture should be optimized to reduce the memory size and computational complexity. Then, the time consumption of the performance needs to be evaluated. And consequently, the approach has to be tested and analyzed under operating conditions with proprietary hardware.

6 Conclusion

A deep convolutional neural network based method for person detection is proposed in this paper. The proposed method is intended to be used for person a flow monitoring system in public transport. Contrary to other approaches, the proposed method uses a convolutional neural network for image segmentation. The segmented image is then processed using the local peaks approach in order to provide the positions of the people in the image. The experiments using a custom dataset provided a precision rate of more than 98 % and recall rate of 96 %. The YOLOv2 approach with various detectors was used as a competitive approach. When using the same dataset and considering all the metrics, the best performing of the new approach versions (the U-Net version) clearly outperforms an established model as the YOLOv2.

However, the work presented in this contribution is only one step in the development of the complex and robust person flow monitoring system. The future work includes neural network architecture optimizing, computational complexity testing and, obviously, testing under operational conditions.

References

1. Choi, J.W.; Quan, X.; Cho, S.H. Bi-Directional Passing person Counting System Based on IR-UWB Radar Sensors. IEEE Internet of Things Journal 2018, 5, 512522. https://doi.org/10.1109/JIOT.2017.2714181
2. Akamatsu, S.; Shimaji, N.; Tomizawa, T. Development of a person counting systemusing a 3D laser scanner. 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014), 2014, pp. 19831988. https://doi.org/10.1109/ROBIO.2014.7090627

3. Ahmed, A.; Siddiqui, N.A. Design and Implementation of Infrared Based Computer Controlled Monitoring System. 2005 Student Conference on Engineering Sciences and Technology, 2005, pp. 15. https://doi.org/10.1109/SCONEST.2005.4382890
4. He, M.; Luo, H.; Hui, B.; Chang, Z. Pedestrian flow tracking and statistics of monocular camera based on convolutional neural network and Kalman filter. Applied Sciences (Switzerland) 2019, 9. https://doi.org/10.3390/app9081624
5. Gao, C.; Li, P.; Zhang, Y.; Liu, J.; Wang, L. person counting based on head detection combining Adaboost and CNN in crowded surveillance environment. Neurocomputing 2016, 208, 108 116. SI: BridgingSemantic. https://doi.org/10.1016/j.neucom.2016.01.097
6. Fu, H.; Ma, H.; Xiao, H. Real-time accurate crowd counting based on RGB-D information. 2012 19th IEEE 370 International Conference on Image Processing, 2012, pp. 26852688. https://doi.org/:10.1109/ICIP.2012.6467452
7. Dolezel, P.; Stursa, D.; Skrabanek, P. On Possibilities of Human Head Detection for Person FlowMonitoring System. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2019, 11507 LNCS, 402413.https://doi.org/10.1007/978-3-030-20518-8_34
8. S. Sun, N. Akhtar, H. Song, C. Zhang, J. Li and A. Mian, "Benchmark Data and Method for Real-Time person Counting in Cluttered Scenes Using Depth Sensors," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 10, pp. 3599-3612, Oct. 2019. https://doi.org/10.1109/TITS.2019.2911128
9. Redmon J.; Divvala S. K.; Girshick R. B.; Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. Computing Research Repository (CoRR), 2015. http://arxiv.org/abs/1506.02640,
10. Joseph Redmon J.; Farhadi A. YOLO9000: Better, Faster, Stronger. Computing Research Repository (CoRR), 2016. http://arxiv.org/abs/1612.08242
11. Millstein, F. Deep Learning with Keras; CreateSpace Independent Publishing Platform, 2018
12. Bottou, L.; Cortes, C.; Denker, J.S.; Drucker, H.; Guyon, I.; Jackel, L.D.; LeCun, Y.; Muller, U.A.; Sackinger,E.; Simard, P.; Vapnik, V. Comparison of classifier methods: a case study in handwritten digit recognition. Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5), 1994, Vol. 2, pp. 7782 vol.2. https://doi.org/10.1109/ICPR.1994.576879
13. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998, 86, 22782324. https://doi.org/10.1109/5.726791
14. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. 421 2012, Vol. 2, pp. 10971105. https://doi.org/10.1109/TITS.2019.2911128
15. ToDo U-Net art https://doi.org/2
16. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C. & Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 2015, 115, 211-252 https://doi.org/10.1007/s11263-015-0816-y