On possibilities of human head detection for person flow monitoring system*

Petr Dolezel[1][0000−0002−7359−0764], Dominik Stursa[1][0000−0002−2324−162X], and Pavel Skrabanek[2][0000−0001−6194−0467]

[1] Faculty of Electrical Engineering and Informatics,
University of Pardubice, Czech Republic
petr.dolezel@upce.cz
[2] Institute of Automation and Computer Science,
Brno University of Technology, Czech Republic
skrabanek@fme.vutbr.cz

Abstract. Along with the development of human society, economy, industry and engineering, as well as with growing population in the world's biggest cities, various approaches to person detection have become the subject of great interest. One approach to developing a person detection system is proposed in this paper. A high-angle video sequence is considered as the input to the system. Then, three classification algorithms are considered: support vector machines, pattern recognition neural networks and convolutional neural networks. The results showed very little difference between the classifiers, with the overall accuracy more than 95 % over a testing set.

Keywords: Person flow monitoring · Support vector machines · Pattern recognition neural network · Convolutional neural network · Histograms of oriented gradients.

1 Introduction

Along with development of human society, economy, industry and engineering, as well as with growing population in the world's biggest cities, various approaches to person detection have become the subject of great interest. Monitoring of person flow, as a branch of the person detection problem, has an indispensable importance in the public transport system safety surveillance, passenger flow prediction, transport planning, or transport vehicle load monitoring. Apparently, the possibility of a precise person flow detection has a great positive effect on public transport systems, station control and management, and cost optimization.

* The work was 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).

The monitoring of the person flow problem is constantly getting more and more focus from both academic and corporate experts. Various approaches to person flow detection are based on infra-red sensors [1], radar sensors [7], lasers [6] or 3D laser scanners [2]. However, these approaches often suffer problems of counting every object that passes through, and also are not able to track the objects precisely. So, in present days, person flow monitoring systems are often implemented using video processing algorithms and computer vision techniques. Generally, several ways of image and video processing can be considered as tools for person flow detection. In particular, statistical methods based on model learning, shape feature, skin color feature or area estimation are widely used [27].

Focusing on shape feature evaluation, human appearance, pose, orientation, and movement are typically considered as inputs for further processing [19]. However, if the monitoring is going to be used in public areas, it is appropriate to avoid collecting data that will enable a specific identification of the person (especially faces). Thus, a high-angle video acquisition tends to be a natural solution of the mentioned difficulty - see Fig. 1.

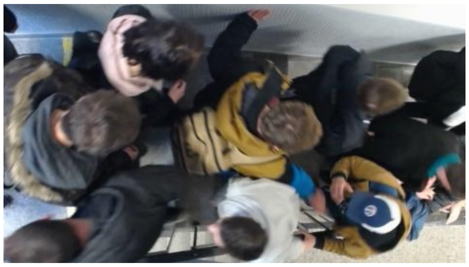

Fig. 1. High-angle shot, persons cannot be identified.

When dealing with a person flow problem using a high-angle video acquisition, only few approaches have been proposed. Gao et al. [10] provide a technique combining convolutional neural networks and cascade Adaboost methods. In [9], the authors use a depth camera along with a classical RGB camera. Both articles provide a method for head and shoulder detection, they do not consider a strict high-angle video acquisition, though. Still, a head itself can provide a strong feature due to its almost circular shape. Then, the Hough transform can be applied to human head detection for getting the flow monitoring result [21]. However, authors of this contribution propose another approach for feature extraction - histograms of oriented gradients. In the previous research, we dealt with a very specific problem of white wine grape detection and counting using visual data. Although a totally different problem, the grape shape is similar to a head. And in the research summarized in [23] and [24], histograms of oriented gradients have proven to be an optimal tool for feature extraction.

Therefore, an approach for a person's head detection is derived and comprehensively tested in this paper. This approach is intended to be used as a key part of a person flow monitoring system, which is going to be implemented in various means of transport for passenger counting. The paper is structured as follows. The problem is properly formulated in next section. Then, the used methods are described and the dataset acquisition is illustrated. The experiments are presented as the subsequent section and the paper is finished with the conclusions.

2 Problem formulation

The aim of this work is to develop a person detector in real-life RGB images. The images are supposed to be derived from a video sequence acquired orthogonally - from above. In the computer vision, the detection process is usually compounded of four steps. During the first step, an object image is acquired from a large real-life image; the second step performs image preprocessing; the third one provides extraction of features; and the final step represents the classification of the object image using the feature vector. In this particular approach, the inputs of the detector are size normalized RGB object images cropped from a real-life video. The outputs are classes of the object images - see Fig. 2 for a basic illustration of the functionality.

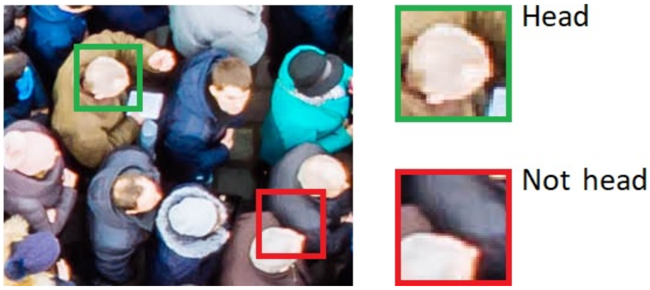

Fig. 2. Person detector functionality.

The structure of the detector is based on our previous work [25], [24] and [23]. Nevertheless, each part of the detector is redesigned in order to fit to the new purpose. The necessary details about all the parts are summarized in the following subsections.

2.1 Image preprocessing

The image preprocessing consists of two steps. The first step deals with the conversion of an input RGB object image from RGB to the grayscale format according to the ITU-R recommendation BT.601 [12]. The resulting grayscale image is obtained by eliminating the hue and saturation information, while retaining the luminance.

The second step of the preprocessing normalizes the contrast of the grayscale image. Each pixel of the resulting image can acquire values from $[0, 1]$. The output of the image preprocessing is the contrast normalized grayscale image.

2.2 Feature extraction

As mentioned above, histograms of oriented gradients (HOGs) [8] are considered as a suitable tool for feature extraction. In simple words, HOG feature descriptor provides distribution (histograms) of directions of gradients (i.e. oriented gradients) of the image. Thus, HOGs encode local shape information from regions within an image.

In order to get beneficial information, HOG cell size and a number of bins need to be set properly. HOG cell size represents the sub-frames of the image under examination. The number of bins affects the sensitivity of gradient directions - all the gradient directions in the sub-frame are divided into the particular number of bins. The volume of each bin consequently provides the information about the dominant gradient directions. It is widely recommended to use 9 bins, but there is no explicit recommendation for the HOG cell size - see Fig. 3, where HOG features are extracted using various cell sizes.

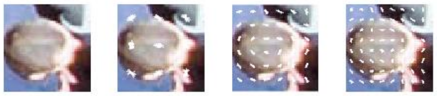

Fig. 3. HOG features for 9 bins and cell size [16, 16]px, [8, 8]px and [6, 6]px. Original size of the object image is [51, 51]px. The length of white abscissae is related to the gradients in the image.

2.3 Classification techniques

The aim of a classification technique is to decide a category of an object captured in an object image. In this contribution, two categories of objects, ’head’ and ’not head’, are assumed. The class ’head’ is called ’positive’ and the class ’not head’ is called ’negative’.

Based on previous authors’ experience, several approaches are considered as possible classification algorithms. Support vector machines (SVMs), which traditionally provide good results with HOGs, are suggested as a first possibility [18].

Except SVMs, feedforward multilayer neural networks seem to be a decent choice for classification in combination with HOGs [26]. For pattern recognition in input data, hyperbolic tangent activation functions are recommended to use in hidden layers and softmax activation functions in output layer. See [20] for detailed information. Such a topology of feedforward network is then called the pattern recognition network (PRN).

Convolutional neural networks (CNNs) are selected as the third approach to be tested. With a rapidly growing possibilities of parallel computing, CNNs became a leading methodology for image processing and analyzing [14]. Compared to other approaches, CNNs use relatively little preprocessing due to the usage of convolutional and pooling layers, traditionally implemented as anterior layers. Therefore, convolutional neural networks leverage spatial information of the object images and a separate feature extraction technique is not necessary to be employed. During the last decade or two, many successful architectures of CNNs were introduced. A brief list of the most popular ones is available in [3].

Therefore, a schematic representation of the detector is shown in Fig. 4.

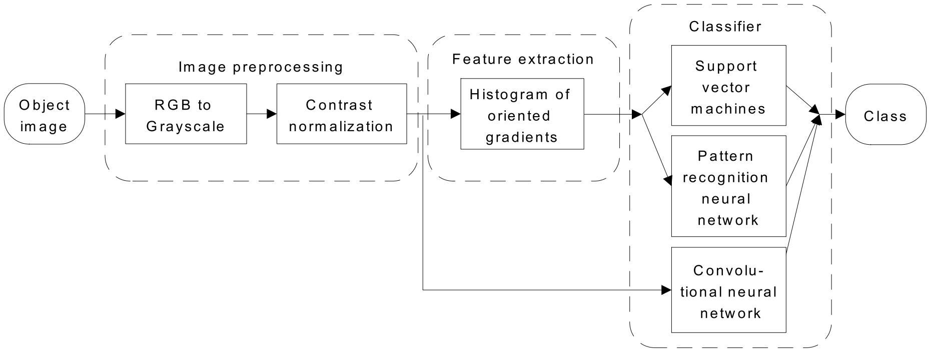

Fig. 4. Flow chart of the person detector.

3 Dataset creation

The important step for a person detector design is a preparation of appropriate training and evaluation sets. The source data should be apparently acquired within the conditions as close to the real situation as possible. Therefore, in this case, the initial video sequences were acquired in the public places both inside and outside under various light conditions. Then, a number of object images was cropped from those sequences. Eventually, 1562 original object images were acquired with the size normalized to 51px × 51 px. The data were divided into four subsets according to Table 1. Some examples are illustrated in Fig. 5.

In addition, in order to support the generalization of the detector, the training set was artificially enhanced - each object image was transformed to provide three more descendants using 90, 180 and 270 degree rotation.

Table 1. Dataset

Training set		Testing set	
Positive	Negative	Positive	Negative
375	406	379	402

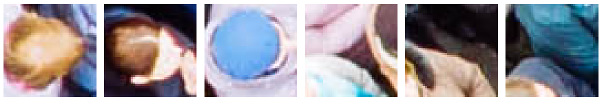

Fig. 5. Three positive (left) and three negative examples from training set.

4 Experiments with classifiers

As mentioned in section 2.3, three approaches are supposed to be tested in this contribution. In addition, each approach provides a number of variants. Thus, the particular conditions of the testing experiments are defined in the following subsections. Note that the conditions are set after a huge set of blind experiments, which are not described here in detail.

4.1 Extraction of histograms of oriented gradients

The following setting of the descriptor has demonstrated to be sufficient. Specifically, a linear gradient direction dividing into 9 bins in $0^o - 180^o$; cells of size 8×8 px; blocks of 2×2 cells; and 1 overlapping cell between adjacent blocks in both directions. Therefore, each object image, which consists of 2601 px, provides 900 elements in the feature vector.

4.2 Support vector machine classifier

SVM classifiers provide various results depending especially on the selected kernel. Linear, polynomial or radial basis function (RBF) kernels are implemented in the most of the cases. Beside the applied kernel function, the performance of a SVM classifier is also influenced by a regularization constant C. Performance of a classifier with the RBF kernel is further influenced by a kernel width σ. In order to tune these parameters, a grid search algorithm [4] combined with the 10-fold cross-validation is used.

Therefore, a set of experiments is performed in order to design an optimal SVM classifier. SVM classifiers with a linear kernel, RBF kernel and polynomial kernel with order equal to 2, 3 and 4 are considered. The training set described in section 3 is used. Since the grid search algorithm belongs to a family of stochastic algorithms, the SVM classifier optimization is performed a hundred times for each classifier and the resulting values of a loss function obtained by cross-validated SVM classifier are observed. The loss function is defined as follows.

$$E_{SVM} = \sum_{j=1}^{n} w_j I \left\{ \hat{y_j} \neq y_j \right\}. \tag{1}$$

Loss function represents the weighted fraction of misclassified observations, where y_j is the class label, $\hat{y_j}$ is the class label corresponding to the class with the maximal posterior probability, w_j is the weight for the observation j, $I\{.\}$ is the indicator function and n is the sample size. In our case, all the weights are equal and $\sum_j w_j = 1$.

The observed values are shown in Fig. 6 for all the selected kernel functions. The central lines in the box graphs, shown in the figure, are medians of loss function resulting values; the edges of the boxes are 25^{th} and 75^{th} percentiles; and the whiskers extend to the most extreme data points (except outliers).

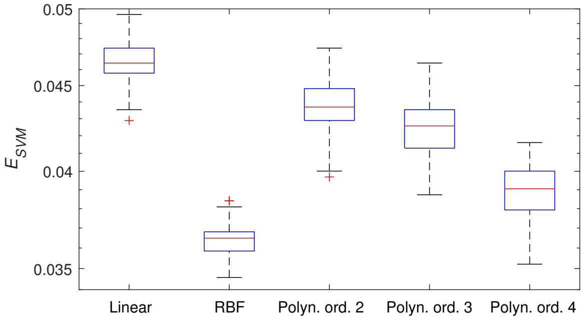

Fig. 6. Resulting values of loss function for SVM classifiers with various kernel functions.

The values pictured in Fig. 6 indicate, that the SVM classifier with the RBF kernel function provides the most suitable behavior. However, the more important quantity would be obtained by the evaluation of the classifiers using the testing set (see section 3). This evaluation is provided in the next section.

4.3 Pattern recognition neural network classifier

Beside training set acquisition, the procedure of PRNs design also involves training, pruning and testing. The essential information about this procedure is described here. More information about the process can be found e.g. in [11].

While a training of PRNs means to find suitable weights and biases of the network, the pruning converts the net into a simpler one while the performance is kept close to the original one. In our approach, a topology search is performed in the following way: PRNs of various topologies are trained using a scaled conjugate gradient algorithm [17] (random 85 % of the training data set - see Section 3 - is used for training, 15 % for validation) and the performance is observed. Similarly to the previous experiment, PRN training is a stochastic process. Therefore, the experiments are performed a hundred times and the results are statistically evaluated. A loss function for the evaluation is defined using a binary cross entropy function.

$$E_{PRN} = -\frac{1}{n}\sum_{j=1}^{n}[o_j \ln(y_j) + (1-o_j)\ln(1-y_j)], \tag{2}$$

where o_j is the desired output, y_j is the actual output of the neural network and n is the number of samples.

The observed resulting values of E_{PRN} are shown in Fig. 7 for various topologies of PRN beginning with two neurons in one hidden layer and ending with two hidden layers, each witch five neurons.

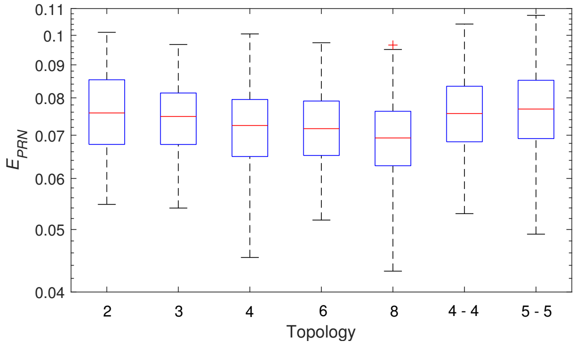

Fig. 7. Resulting values of loss function for PRN classifiers. Labels along the X-axis represent number of neurons in hidden layers.

Looking at Fig. 7, all the topologies provide similar results. Again, the more important tests using testing set are provided in the next section.

4.4 Convolutional neural network classifier

As mentioned above, a number of architectures of CNNs is available for testing, these days. According to some literature research as well as previous authors' experience, five specific architectures are selected for implementation. The first two architectures are relatively simple. Net1 consists of one convolutional layer, one max-pooling and one fully connected dense layer with 512 neurons. The layer with 2 neurons and softmax activation function is implemented as the output layer. Net2 is similar, but it contains a more complex sequence of anterior layers. In particular, it is convolutional layer + convolutional layer + max-pooling layer + convolutional layer + max-pooling layer. Both networks are adapted from [16]. The third one is one of pioneering architectures - LeNet [5], while the fourth one is probably the most cited topology - AlexNet [15]. This architecture was originally designed to win the ImageNet Large-Scale Visual Recognition Challenge, but it spread to a huge number of industrial and engineering applications. The last selected architecture is called VGG-16 net, based on the a large number of simple and repetitive layers, which is, in some cases, effective to implement [22].

Similarly to the previous cases, the mentioned architectures are trained in order to classify correctly the dataset described in section 3. ADAM algorithm is implemented as optimizer [13]. Again, the experiments are performed a hundred times due to a stochastic character of training and a binary cross entropy function is used as loss function - see (2). The resulting values are shown in Fig. 8.

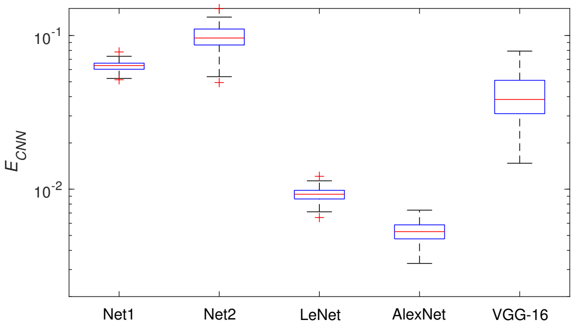

Fig. 8. Resulting values of loss function for CNN classifiers.

After the training process, AlexNet seems to be the most suitable CNN to be implemented in the person detector.

In the next section, all the variants designed here are tested using the testing set. The testing procedure should denote the best possibility among of all.

5 Results and discussion

The aim of this section is to evaluate all the proposed classifiers. A good practice for the evaluation is to determine the accuracy of the classifiers over the testing set. However, two additional metrics, precision and recall, are proposed to evaluate the classifiers comprehensively. The metrics are described by following equations.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

where TP (true positive) is the number of correctly classified positive images, FN (false negative) is the number of misclassified positive images, FP (false positive) is the number of misclassified negative images, and TN (true negative) is the number of correctly classified negative images.

The resulting values of the metrics for all the classifiers are summarized in Table 2.

Table 2. Testing results

Classifier	**Accuracy**	**Precision**	**Recall**
SVM with linear kernel function	0.9539	0.9561	0.9485
SVM with RBF kernel function	0.9645	0.9718	0.9545
SVM, polynomial order 2	0.9539	0.9525	0.9525
SVM, polynomial order 3	0.9545	0.9544	0.9518
SVM, polynomial order 4	0.9581	0.9589	0.9545
PRN, [2]	0.9539	0.9567	0.9479
PRN, [3]	0.9529	0.9536	0.9492
PRN, [4]	0.9549	0.9532	0.9538
PRN, [6]	0.9529	0.9653	0.9367
PRN, [8]	0.9542	0.9654	0.9393
PRN, [4 4]	0.9504	0.9474	0.9505
PRN, [5 5]	0.9542	0.9531	0.9525
Net1	0.9129	0.9356	0.8813
Net2	0.9206	0.9249	0.9103
LeNet	0.9501	0.9359	0.9631
AlexNet	0.9040	0.8671	0.9472
VGG_16	0.8886	0.8544	0.9288

The testing results, shown in Table 2, indicate several interesting outcomes. First of all, best accuracy and precision along all the classifiers is provided by the SVM model with a RBF kernel function, which may be a surprising fact, considering the list of classifiers. Then, the tested CNNs provide generally the worst performance. And thirdly, all the performances are very similar, accuracy between 95% and 96%. This feature could indicate, that although the classifiers are generally trained sufficiently, some samples in the testing set can be outside the regular position. Comprehensive check of the results shows, that if a testing sample is misclassified by one classifier, it is misclassified by at least 8 other classifiers in more than 60 % of cases.

6 Conclusion

In this contribution, a set of classifiers for person detection from a high-angle image is introduced, designed and tested. According to the results presented above, the image feature extraction using a histogram of oriented gradients in combination with pattern recognition network or support vector machine as a classifier looks like an effective solution for such issues. Apparently, not only the accuracy, but also computation time is necessary to be tuned in order to provide a suitable tool for the monitoring of person flow in real-life conditions. Hence, computational efficiency and classifier implementation using special hardware for parallel processing will be the next subject of interest.

References

1. Ahmed, A., Siddiqui, N.A.: Design and implementation of infra-red based computer controlled monitoring system. In: 2005 Student Conference on Engineering Sciences and Technology. pp. 1–5 (Aug 2005). https://doi.org/10.1109/SCONEST.2005.4382890
2. Akamatsu, S., Shimaji, N., Tomizawa, T.: Development of a person counting system using a 3d laser scanner. In: 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014). pp. 1983–1988 (Dec 2014). https://doi.org/10.1109/ROBIO.2014.7090627
3. Aloysius, N., Geetha, M.: A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP). pp. 0588–0592 (April 2017). https://doi.org/10.1109/ICCSP.2017.8286426
4. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. The Journal of Machine Learning Research **13**, 281 – 305 (Feb 2012)
5. 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. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5). vol. 2, pp. 77–82 vol.2 (Oct 1994). https://doi.org/10.1109/ICPR.1994.576879
6. Chen, Z., Yuan, W., Yang, M., Wang, C., Wang, B.: Svm based people counting method in the corridor scene using a single-layer laser scanner. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). pp. 2632–2637 (Nov 2016)
7. Choi, J.W., Quan, X., Cho, S.H.: Bi-directional passing people counting system based on ir-uwb radar sensors. IEEE Internet of Things Journal **5**(2), 512–522 (April 2018). https://doi.org/10.1109/JIOT.2017.2714181
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 vol. 1 (June 2005). https://doi.org/10.1109/CVPR.2005.177
9. Fu, H., Ma, H., Xiao, H.: Real-time accurate crowd counting based on rgb-d information. In: 2012 19th IEEE International Conference on Image Processing. pp. 2685–2688 (Sep 2012). https://doi.org/10.1109/ICIP.2012.6467452
10. Gao, C., Li, P., Zhang, Y., Liu, J., Wang, L.: People counting based on head detection combining adaboost and cnn in crowded surveillance environment. Neurocomputing **208**, 108 – 116 (2016). https://doi.org/https://doi.org/10.1016/j.neucom.2016.01.097, http://www.sciencedirect.com/science/article/pii/S0925231216304660, sI: BridgingSemantic
11. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1999)

12. ITU-R Recommendation BT.601: Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios (Mar 2011)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
14. Kizuna, H., Sato, H.: The entering and exiting management system by person specification using deep-cnn. In: 2017 Fifth International Symposium on Computing and Networking (CANDAR). pp. 542–545 (Nov 2017). https://doi.org/10.1109/CANDAR.2017.40
15. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. vol. 2, pp. 1097–1105 (2012)
16. Millstein, F.: Deep Learning with Keras. CreateSpace Independent Publishing Platform (2018)
17. Moller, M.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks **6**(4), 525 – 533 (1993)
18. Paisitkriangkrai, S., Shen, C., Zhang, J.: Performance evaluation of local features in human classification and detection. IET Computer Vision **2**(4), 236–246 (December 2008). https://doi.org/10.1049/iet-cvi:20080026
19. Pore, S.D., Momin, B.F.: Bidirectional people counting system in video surveillance. In: 2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT). pp. 724–727 (May 2016). https://doi.org/10.1109/RTEICT.2016.7807919
20. Resch, C., Pineda, F., Wang, J.J.: Automatic recognition and assignment of missile pieces in clutter. In: Neural Networks, 1999. IJCNN '99. International Joint Conference on. vol. 5, pp. 3177–3181 vol.5 (1999). https://doi.org/10.1109/IJCNN.1999.836162
21. Shang, H., Wang, T.: Bus passenger counting based on frame difference and improved hough transform. In: 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). pp. 3132–3135 (April 2012). https://doi.org/10.1109/CECNet.2012.6201616
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556 (09 2014)
23. Skrabanek, P., Dolezel, P.: Robust grape detector based on svms and hog features. Computational Intelligence and Neuroscience **2017** (2017). https://doi.org/10.1155/2017/3478602
24. Skrabanek, P., Majerik, F.: Evaluation of performance of grape berry detectors on real-life images. pp. 217–224 (2016)
25. Skrabanek, P., Runarsson, T.P.: Detection of grapes in natural environment using support vector machine classifier. In: Proceedings of the 21st International Conference on Soft Computing MENDEL 2015. pp. 143 – 150. Brno University of Technology, Brno, Czech Republic (23–25 June 2015)
26. Taskiran, M., Cam, Z.G.: Offline signature identification via hog features and artificial neural networks. In: 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 83–86 (Jan 2017). https://doi.org/10.1109/SAMI.2017.7880280
27. Wu, X.: Design of person flow counting and monitoring system based on feature point extraction of optical flow. In: 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications. pp. 376–380 (June 2014). https://doi.org/10.1109/ISDEA.2014.92