AKADÉMIAI KIADÓ

Pollack Periodica • An International Journal for Engineering and Information Sciences

17 (2022) 1, 1–6

DOI: 10.1556/606.2021.00425

ORIGINAL RESEARCH PAPER

Detection of feature-poor, small sized objects on noisy images

Tamas Storcz* and Zsolt Ercsey

Department of Systems and Software Technologies, Faculty of Engineering and Information Technology, University of Pécs, Boszorkány u. 2, H-7624 Pécs, Hungary

Received: April 9, 2021 • Revised manuscript received: June 27, 2021 • Accepted: June 30, 2021
Published online: November 3, 2021

ABSTRACT

Visual identification of objects is an important challenge today. Main target of frequently applied methods is to identify or classify complex objects. These methods are far less effective when objects are small and less complex, and thus less descriptor features are on hand. The main reason for this is that these features can significantly change on object occlusion or appearance of noise.

The presented solution performs identification of simple, small (size is 17 × 13 pixels) objects with elliptical shape. High pass filtered normalized cross correlation is used for region of interest detection and a simple deep neural network is used for classification of selected regions. The proposed method detected objects on a noisy image with accuracy of 96.2%.

KEYWORDS

image processing, deep neural network, feature poor object detection

1. THE CHALLENGE

In a biological effect experiment, experts investigate effect of chemicals with different dose and dilution. Chemical treatments are executed under fully controlled conditions, and their effect is measured based on the number of occurrences of small objects of predefined shape within a predefined area. For scientific evaluation, the number of objects must be calculated accurately. The calculation of the objects is done by visual observation of offal, collected under the treated surface. The main aim of this visual observation is to determine the number of objects as accurately as possible. Because of the specific circumstances of field sampling, the experiment and the result observation are done separated in time and space. Visual data acquisition tried to be standard, but field circumstances are restrained. When sampling, a perpendicular photo was taken by a camera of a mobile phone over a constant area (page A4), from the same distance, but natural illumination changes were detected. Quality of this camera system (optics and sensor) is medium category. Compared to photo cameras, it is rather low end. Thus, the observation was done on normal/high resolution (3456 × 4608 pixels), but often on under illuminated images.

The primary challenge is that next to the objects that has to be counted, other offal appears in the observation area, of which amount is orders of magnitude larger, while its size and shape is similar to the object sought, moreover also occlusion could happen. The identification of the object to be counted is difficult because of its homogenous dark, almost black color and relatively small size. An object in the picture has almost a perfect ellipse shape with approximately 17 × 13 pixel size. Fig. 1a shows a case, when the dark objects on white background are fully visible, the identification, therefore counting can be done accurately. Nevertheless, in case presented in Fig. 1b, when the environment is charged by a noise very similar to the objects, or in the case when the original object (presented in Fig. 1c) has been modified like it is illustrated in Fig. 1d, the object identification becomes difficult, and as a consequence the counting becomes inaccurate.

*Corresponding author.
E-mail: storcz.tamas@mik.pte.hu

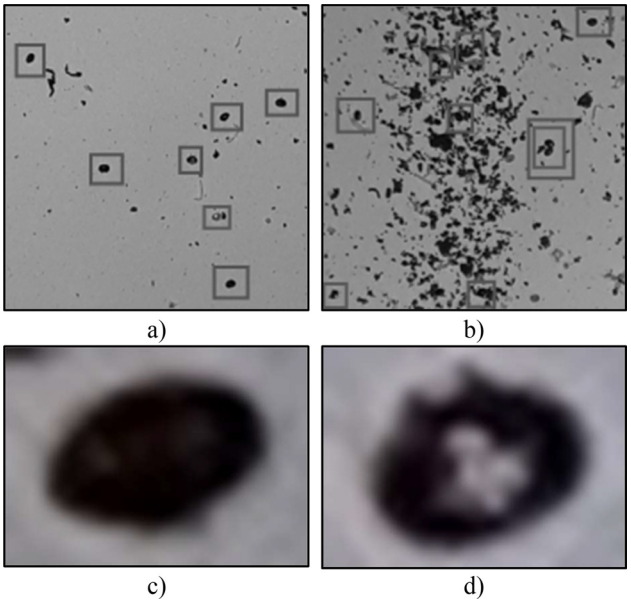

Fig. 1. Object identification tasks, a) clear environment, b) noisy environment, c) homogenous surface and shape, d) inhomogeneous surface and shape
(*Source*: Author)

2. RELATED WORKS

2.1. Blob detection

In scale space theory [1] interference of differentiate of spatially close, oriented gradients pairs, highlights those gradient pairs, if their distance fits scale space parameter. Based on these highlights, blobs [2] according to the scale space level can be identified. The advantage of this method is that when Laplace of Gaussian is replaced by difference of Gaussian, the scale space level transformation calculation becomes simpler, thus faster, and the scale space level of observed images can be predetermined and constant for this experiment.

Disadvantage is that the method reacts poorly to appearance of noise, especially when the noise properties are similar to the object properties. Furthermore, when white content appears on the surface of a dark object, as it is shown in Fig. 1d, the recognized patch size will be smaller. Here patches have an elliptical shape, and thus the identification is orientation dependent.

2.2. Cross-correlation

The cross-correlation [3] is a measure that represents similarity of two series of data. General area of utilization is identifying a short series (function) in a longer series of data (function). This in fact almost matches the considered problem when looking for a small object in a big image, but unfortunately this method is not rotation independent, i.e., the method is only capable of identify objects, when they have the same orientation as the sample has. The utilization possibility is further reduced by that the correlation coefficient is reduced when shape similarity is not perfect. Thus, high pass filtering of correlation coefficient is required. But high threshold increases the number of false negatives, while low threshold increases the number of false positives. Threshold selection becomes even harder when correlation coefficient gets lower by insufficient illumination. That is because the sample background will be lighter than the object background. That effect can be reduced by smaller sample size, thus smaller background rate, but thresholding in this way further increases the number of false positive identifications.

2.3. Convolution

Convolution's [4] operation and meaning is very similar to cross-correlation. Both are shift invariant and linear, but convolution is commutative, associative, and distributive for addition, therefore more generally applicable and in practice more often applied. Due to its simplicity and local data requirements, it is well parallelizable. When running on GPU, its computation performance can be significantly improved. Unfortunately, it does not provide solution for orientation dependency and for the threshold selection problems mentioned above.

2.4. Normalized cross-correlation

Normalized Cross-Correlation (NCC) [5] is a modified version of generic cross-correlation, which is used more often in machine vision. The essence of the amendment is that the result is a value in the $[-1, \ldots, 1]$ interval, that helps input scale independent threshold selection, but in practice, less generally applicable because of its bigger computational resource needs. As described in [6], the NCC operation can be accelerated when based on convolution and executed on GPU, therefore the execution time may get shorter. The appropriate incrementation of execution speed makes it possible to apply a well-designed (rotated) kernel set, to make convolution and NCC orientation independent. That widens application opportunities but does not provide a solution for the threshold selection, thus the expected rate of false positive identifications in the output will be high.

2.5. Relative position of simple geometric features

Scale Invariant Feature Transform (SIFT) [7] and Speed Up Robust Features (SURF) [8] methods use different levels of the previously mentioned scale space theory [1] to determine scale space related features and their relative positions to extract then find complex featured objects. These methods approximate the Laplace of Gaussian operator of scale space theory differently. The SIFT uses difference of convolutional steps made by the Gaussian kernel, and SURF applies a box filter. Another difference is the method of feature extraction. While SIFT uses interference peaks of local gradient differences in different orientations, SURF identifies them using the Hessian matrix.

Drawback of both methods is that, for identification of objects, a certain number of unique descriptors and their relative positions are required. When identifying feature-poor objects, this condition is hardly met. When only a few

features exist, number of false positive identifications will grow with the similarity of noise and object properties and with the growth of spatial density. Because of this, these methods cannot be directly applied in the current task.

2.6. Deep neural nets

Analog processor units (artificial neurons) can be organized into layers then layers can be organized into multilayer networks, based on universal approximation theorem [9]. Deep neural networks [10, 11] usually consist of a multilayer feature extractor and a multilayer classifier. Feature extractor contains convolutional layers for feature extraction, max pooling layers for subsampling and other layer types to support model optimization. Classifier is a multilayer, fully forward connected mainly homogeneous network. These types of models are good for complex classification tasks, but when input images contain objects of unknown number, single classification operation is meaningless.

When scale space parameter is well-known and constant, classification can be done in an object size window moved on the image, by a convolution like operation. This can be executed parallel, but still would be computational resource demanding operation.

2.7. YOLO

You Only Look Once (YOLO) method, as a new generation of object identification system, instead of classifying contents of images, puts emphasis on approximation of spatially separated object bounding boxes. Novelty of this process is that a deep neural network estimates the bounding box properties beside the probability of bounded object class with a single scan of the input image. As the first step, the neural net creates many bounding boxes for an object candidate. From these, with a special method, the final bounding box and class probability can be calculated. First YOLO implementation contained 24 convolutional and 2 dense layers. In newer implementations, extension of DarkNet [12] contained 53 convolutional layers and ResNet [13] worked with more than 100 layers. YOLO type algorithms perform well on identification of small number of objects in observed areas, even if objects are partially occluded, but the number of objects in an area have to be predefined and can only be a few. Since this last condition in the current experiment cannot be guaranteed, only should the observation area size be reduced to too small which significantly extends execution time. Another problem, as detailed previously, that the feature-poor object identification in a noisy area results unclear identifications.

3. THE PROPOSED METHOD

As explained above, the main task is to count relatively small (17×13 pixels), feature-poor objects on a relatively big (3456×4608 pixels) image without limitation of occurrences. The object identification must be as exact, as possible. Experts determined a 5% accuracy as definitely acceptable for biological experiments. This criterion led us to draw up two minor criteria. Identification must minimize the number of false positive and false negative identifications in the final output.

3.1. Structure

Based on the reviewed experiences of related works, the proposed system consists of the following two main parts. The first component determines square regions, which potentially could contain objects (Region of Interest (RoI)). This method is expected to minimize the false negative selections. The second component is a classifier performing detailed examination of the selected region proposals and determining the object representation probability of each RoI. The classifier is expected to maximize F_1-score. During the design of the main components, next to the expected accuracy, execution time is also an important aspect.

When applying a generic pixel statistics-based shadow removing method [14, 15] or color correction [16] as an image preprocessing step, the system efficiency was not improved. The reasons were that the objects and also the noise have almost 2-dimensional shape, thus do not drop significant shadow. Furthermore, the main contents of the images are black (objects and noise) and white (background) and the low color information were further weakened by insufficient illumination and under exposition.

3.1.1. RoI extraction. For RoI detection, normalized cross-correlation is applied as described in chapter 2.4, because the method does not require large amount of teaching samples to be collected or created and the execution time can be improved by applying convolution and parallelization by applying GPU. As described before, NCC is orientation dependent, so the correlation coefficient depends on the relative orientation of the sample and image. To get rid of the orientation dependency, NCC kernel collection is augmented by variants of the original kernel. Central and axial symmetry of the objects allowed making this kernel augmentation by 8 rotation steps. Increasing the number of rotation steps increases accuracy but in parallel degrades execution speed. To get RoI centers, a non-max suppression on each kernel layer and a max selection overall kernel layers is performed. The result then is filtered with a high-pass filter. A low threshold is selected upon validation tests to eliminate the false negative identifications and to minimize the false positive identifications.

In the applied NCC implementations, the base kernel of size 31×31, shown in Fig. 2a, was rotated with 8 rotation steps between -90 and $+90$ degrees, then 19×19 pixels were copied from its center. That resulted 9 (1 original plus 8 augmented) kernels. Bigger original kernel size and center clipping was designed to eliminate the effect of image rotation on background corners, as shown in Fig. 2b.

The execution time of NCC can be reduced, when instead of the 3 channels (Red, Green, Blue) of the true color input image, the method works on its grayscale representation. Loosing additional color information represented by RGB channels could further increase the number of false

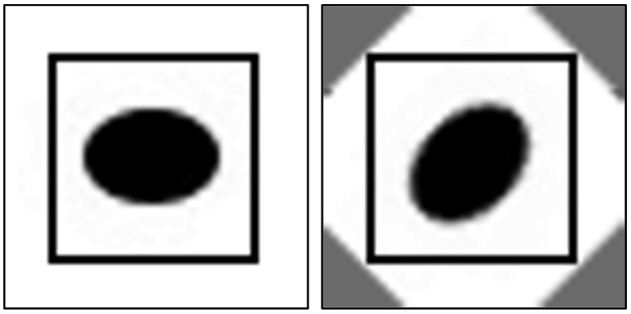

Fig. 2. RoI extraction kernels augmentation and clipping, a) default kernel, b) rotation effect on background (*Source*: Author)

positive identifications, but the second main component uses RGB input, therefore using grayscale input in the first phase does not affect the final accuracy. Because of the small size of sought objects, down sampling cannot be applied to improve execution speed.

After NCC pixel-wise maximums of calculated correlation coefficients are selected, then the suppression of high values spatially close to maximums (non-max suppression) is performed. RoI proposal centers are calculated by high pass filtering of previously combined and suppressed result.

During the implementation, the proper selection of a high-pass filter threshold is an important question. This selection was based on utilization of labeled data that was generated previously by experts. Therefore, a very threshold could be set, which selected all RoI-s of the expert generated label database. Based on the details presented earlier, the assumption was that the rate of false positive identifications will be high. And indeed, the method identified 25,136 regions, from which 5,297 was excluded because of low average intensity and variance; these regions were almost fully black. The RoI extraction false positive rate exceeded (25,136–5,297)/724 = 2,740% of the true positive identifications, thus its accuracy was around 3.5%.

3.1.2. RoI classification. The output of the RoI selection phase is obviously unsuitable for direct application as output, because of its low precision, caused by the high rate of false positive identifications. But if data requirements of machine learning methods are considered, that output can be considered as very useful.

The next step must be to create and teach a machine learning based classifier system to classify contents of the previously selected 19 × 19 pixel sized RoI proposals. Because the color information is not intended to be lost, classification is done on a 19 × 19 × 3 sized matrix, clipped from 3 color channels of the input RGB image. For the classification, a simple deep neural network classifier was designed and thought. Color correction described in [17] would only change the rate of the color channels, thus human perception of pixels, however it did generate new information for the classification, therefore its application was dropped.

3.1.3. Deep neural network. The proposed deep neural network, as detailed in chapter 2.6, consists of a convolution based automatic feature extractor module, and a fully forward connected, dense classifier module.

The proposed feature extractor module is much simpler than YOLO structures. Because of fix scale of the input images and small size of the sought objects, it contains 3 convolutional layers each of which applies 3 × 3 pixel sized kernels and creates 128 features for all 3 color channels. These features are processed by a dense net with 1 hidden layer, containing 1,000 neurons with rectified Linear Unit (ReLU) activation and a dense output layer with sigmoid activation. The classifier network output is the probability of the RoI representing an object. The final object representation results of a RoI to be true when the probability is above 0.5, and false otherwise.

3.2. Training

For training an Adam optimizer [17] and a binary cross-entropy error function were used to fit the binary classification task.

3.2.1. Data acquisition. To teach the system for detecting objects, training samples had to be defined from field images. Experts involved in the experiment labeled 22 randomly selected RGB input images, on which 728 positive samples were marked. NCC RoI extraction executed on that, 22 labeled images resulted 724 positive, previously labeled samples and 25,136 negative samples (false RoI positive identification). By an intensity high-pass filter, 5,297 negative samples were excluded, thus 19,839 negative samples were acquired. Available samples are divided into training and testing datasets by a 60–40% split in a stratified manner, which means positive and negative samples are split separately. After splitting, each positive training sample was planned to be augmented by 4 random rotations, so the augmented positive sample space would be extended to 2,170 elements. The main target of augmentation is to decrease rotation dependency of the classifier. But it also increases the repetition of main features of samples in training data therefore that would turn the classifier into a memory of positive samples.

Exact and detailed structure of sample space is presented in Table 1.

3.2.2. Batch creation. Because the rate of the positive samples was only 3.5% before the test augmentation and just 11% after it, it is clearly visible, that the sample space is asymmetric. Teaching batch generation in an asymmetric sample space has to be done carefully, because if only one class is represented in a batch, the classifier instead of classification will become a single memory, which stores the represented class as output. Therefore, based on the criteria above, batches containing 200 items of which 30% positive samples are generated. That means, each training batch will contain 60 positive and 140 negative samples. Within the current experiment, 12,337/200 = 62 batches could be created from training samples. On stratified batch

Table 1. Structure of sample space before and after augmentation

	Train		Test	Total	
	Original	Augmented		Original	Augmented
Positive	434	2,170	290	724	2,460
	2.1%	9.7%	1.4%	3.5%	11%
Negative	11,903	11,903	7,936	19,839	19,839
	57.9%	53.4%	38.6%	96.5%	89%
Total	12,337	14,073	8,226	20,563	22,299
	60%	63.1%	40%	100%	100%

Table 2. False positive and false negative identification statistics

#	1	2	3	4	5	6	7	8	9	10	Sum	*Avg*
FP	42	3	36	38	37	26	12	28	21	34	287	*28.7*
FN	1	3	0	2	0	0	3	2	0	0	11	*1.1*

generation, number of negative batches is 11,903/140 = 86. When sharing 343 positive samples between these 86 batches, each would contain only 4.9 positive samples. To match 60 required items, the missing items are taken from other batches, which results repetition of positive samples in the full training process. No further geometrical augmentation was applied, not to increase positive item repetition rate. And classification was rotation independent, as it was proven by final accuracy.

4. RESULTS

From the total 730 objects identified by experts, RoI selection identified 724, while the number of total false positive identifications was 19,839 after intensity filter. Although the total accuracy of RoI selection is just 3.5%, and rate of false positives and true positives is 6,849%, the recall [18] of positive samples as given in Eq. (1) is 724/730 = 99.1%, therefore RoI selection method satisfies the requirements imposed on it, namely its false negative rate is 0.9%

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{1}$$

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \tag{2}$$

$$\text{Object detection error} = \left| \frac{\text{RoI}^{+} - \text{DNN}^{\text{FN}} + \text{DNN}^{\text{FP}}}{\text{positive samples}} - 100 \right|. \tag{3}$$

The deep neural network classifier on a device equipped with Core I7, 16 GB RAM, and Nvidia GeForce 1660Ti video card with 6 GB RAM, 4,000 epochs long training process runs 6 h on average. From 10 executions, *total* number of false negative identifications was 11, *total* number of false positive identifications was 287. So *average* false negative identification of these 10 runs is 1.1 and *average* false positive identification is 28.7 as it is shown in Table 2.

Instead of accuracy, quality of method is measured by precision of positive class, as described in Eq. (2). Thus, the performance of system $724/(724 + 28.7) = 96.2\%$ and its error is 3.8%. Final object detection error of the complete system, as Eq. (3) specifies $|(724 - 1.1 + 28.7)/730 - 100| = 3\%$.

5. CONCLUSION

The proposed system has the structure of an NCC based RoI selector connected to a deep neural network-based classifier. This system after a parameter optimization based on expert labelling, RoI extraction and augmentation, performed the object identification task with better than 5% average accuracy. This means that the error of original object counting task solution will not exceed 5%. This expected error level is considered to be good by experts of the biological problem domain; therefore the method is applied as part of biological experiments.

The fine-tuning of positive teaching sample augmentation process could decrease the number of false negative and false positive identifications, thus the general accuracy of the object counting task was considered to be good.

ACKNOWLEDGEMENTS

The research project is conducted at the University of Pécs, Hungary, within the framework of the Biomedical Engineering Project of the Thematic Excellence Programme 2020 – National Excellence Sub-program (2020–4.1.1-TKP2020).

REFERENCES

[1] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Boston, MA: Springer, 1994.

[2] H. Kong, H. C. Akakin, and S. E. Sarma, "A generalized Laplacian of Gaussian filter for blob detection and its applications," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1719–1733, 2013.

[3] R. B. Fisher and P. Oliver,"Multi-variate cross-correlation and image matching," in *Proceedings of the 6th British Conference on Machine Vision*, Dept. of Electron. and Electr. Eng. Univ. of Surrey, Guildford Surrey, UK, 1995, pp. 623–632.

[4] P. Gertreuer, "A survey of Gaussian convolution algorithms," *Image Process. Line*, vol. 3, pp. 286–310, 2013.

[5] F. Zhao, Q. Huang, and W. Gao, "Image matching by normalized cross-correlation," in *IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, Toulouse, France, May 14–19, 2006, pp. II–II.

[6] K. Briechle and U. D. Hanebeck, "Template matching using fast normalized cross correlation", in *Proceedings of SPIE: Optical Pattern Recognition XII*, Orlando, FL, US, Apr. 16–20, Paper no. 4387, 2001, pp. 95–102.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, pp. 91–110, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comp. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.

[9] A. N. Kolmogorov, "On the representations of continuous functions of many variables by superpositions of continuous functions of one variable and addition," *Doklady Akademii Nauk USSR*, vol. 114, no. 5, pp. 953–956, 1957.

[10] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 23–28, 2014, pp. 2147–2154.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 27–30, 2016, pp. 770–778.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection" in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 27–30, 2016, pp. 779–788.

[13] L. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 21–26, 2017, pp. 6517–6525.

[14] T. Storcz, Z. Ercsey, and G. Várady, "Histogram based segmentation of shadowed leaf images," *Pollack Period.*, vol. 13, no. 1, pp. 21–32, 2018.

[15] T. Storcz, G. Várady, and Z. Ercsey,. "Identification of shadowed areas to improve ragweed leaf segmentation," *Tech. Gaz.*, vol. 28, no. 4, 2021, in press.

[16] G. Várady, "Color information correction of images using lightweight camera systems," *Pollack Period.*, vol. 14, no. 1, pp. 3–14, 2019.

[17] I. K. M. Jais, A. R. Ismail, and S. Q. Nisa, "Adam optimization algorithm for wide and deep neural network," *Knowl. Eng. Data Sci.*, vol. 2, no. 1, pp. 41–46, 2019.

[18] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, & correlation," *J. Machine Learn. Tech.*, vol. 2, no. 1, pp. 37–63, 2011.