Convolutional Neural Network for Sound Processing - Study of Deployed Application

Petr Dolezel, Dominik Stursa, Daniel Honc
Faculty of Electrical Engineering
and Informatics
University of Pardubice
Pardubice, Czech Republic
petr.dolezel@upce.cz

***Abstract*—Pest birds are considered as a special kind of vermin, since, in most of countries, their legal position does not enable their direct extermination. Therefore, in order to protect the agricultural areas indirectly from pest birds, the robust and highly selective pest bird sensor is necessary to design. In this contribution, the pest bird detection unit, based on a convolutional neural network, is presented. The convolutional neural network itself is used for the decision making about the pest bird occurrence, while sound recordings are used as input data. The testings, presented at the end of the contribution, proved a very high accuracy of the detection unit, with the results indispensably improved in comparison to previously presented approaches.**

***Index Terms*—Pest birds, spectrogram, convolutional neural network**

I. Introduction

People deal with various sorts of vermin since the beginning of agriculture. Pest birds are considered as a special kind of vermin, since, in most of countries, their legal position does not enable their direct extermination. However, the amount of damages caused by pest birds is indispensable, especially in orchards, plantations and in vineyards. Thus, a number of approaches have been proposed so far to protect agricultural areas against pest birds. As examples, propane cannons, sound alarms, mist nets, kites or falconry can be mentioned. A complex list of various possibilities is summarized in [1].

The mentioned protection systems are unfortunately in constant operation, regardless of the occurrence of pest birds. Therefore, the running is often costs-ineffective and some systems can even produce constant noise pollution. Therefore, the aim of this paper is to deal with a robust and highly selective pest bird sensor in order to provide a trigger for the mentioned protection systems. Clearly, if some detection system is able to provide a highly reliable information about the occurrence of the pest birds in the area, the protection system can operate on demand, not all the time.

Dealing with possible approaches for pest bird detection, non-contact family of sensors comes into consideration. Acoustic sensors (microphones) have been identified as very suitable ones [2]. Other approaches successfully implement radar sensors [3], [4]. Image sensors [5] and IR sensors [6] are applied only in very specific cases, where a direct visual contact is available between emitter and receiver.

Nevertheless, it is necessary to decide about the flock of pest birds presence in the monitored area. Then, if true, the position of the flock should be estimated as exactly as possible.

II. Problem Formulation

The idea of pest bird detection in monitored area, in order to control the protection system, is proposed as follows: a set of detection devices is spread over the monitored area. The consecutive devices are supposed to be located with overlapping perimeter. Then, the position of the flock, as well as its speed and direction, can be estimated using statistical approach - see [7] for particular example.

Apparently, it is necessary to develop a robust, selective and cheap enough detection device in order to implement the statistical approach proposed above. In [8] and [11], the authors proposed a detection unit, which used real-time sound recordings as an input source. The sound samples were preprocessed using common filtration and normalization techniques and then, the relevant features were extracted using Linear Prediction Coding approach (LPC) [9]. With the extracted features, a pattern recognition neural network was implemented to decide about the pest bird occurrence.

With a rapid development of special hardware for parallel processing, the products based on architectures like nVIDIA Jetson or Intel Movidius become affordable enough for commercial production - see Fig. 1. The implementation for the detection unit is naturally suggested, since the mentioned parallel processing architectures are able to process much more information than just LPC.

Therefore, the aim of this paper is to propose and test a detection unit based on convolutional neural network framework, which can be advantageously implemented using e.g. Intel Movidius. Convolutional neural networks are well known for the highly successful rate for classification problems. In addition, they do not use a separate preprocessing and feature extraction technique due to the usage of convolutional and pooling layers, implemented in the anterior part of the network. The comprehensive information about the convolutional neural networks can be found in [10].

Fig. 1. nVIDIA Jetson and Intel Movidius.

Thus, the the pest bird detection device, based on convolutional neural networks, is proposed and designed in the following sections. The performance of the proposed device is compared to the results presented in [8] and in [11], where LPC approach was implemented.

III. Proposed Solution

Convolutional neural networks are especially capable to work with an image as a source of information. However, an image sensor is generally not a good choice due to the requirement of a direct visual contact with the target. Therefore, acoustic recording is selected as a suitable source of information. Acoustic sensors are affordable and energetically effective, they are able to provide complex enough information, though [12]. However, sound sample as a source of information needs to be transformed to a data structure suitable for a convolutional neural network. Then, the network is expected to decide about the presence of the bird in the surrounding of the detection unit. Hence, the detection should work as shown in Fig. 2. Each block of the referred flow chart is discussed in the following sections.

A. Data Preprocessing

Clearly, the acoustic data need to be preprocessed in order to provide a consistent supply of information.

The quality of original sound source is determined by the recording device. For the aims of this contribution, a microphone, which provides single-channel sound stream with sample rate of 44100 Hz and double precision, is used.

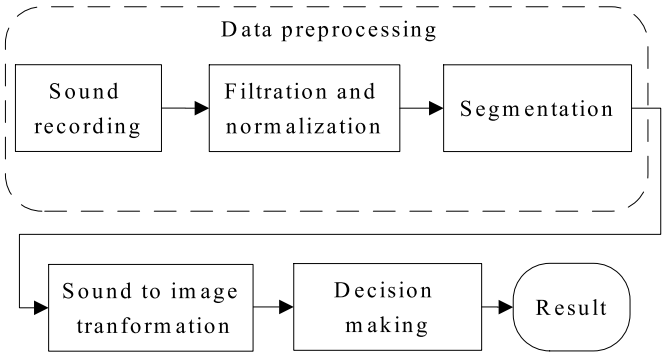

Fig. 2. Flow chart of the detection unit.

The sound stream is then continuously filtered and statistically processed in order to remove noise, silent segments and to balance the energies across the recent history of recording - see Fig. 3 for the example. Afterwards, the signal is divided into the segments of constant length. In order to keep the consistency with the reference work mentioned above [11], the the length of the segments is set to 3 s and each segment is passed forward for the further processing.

B. Sound to Image Transformation

As mentioned above, a visual representation of the sound sample is advantageous to get in order to prepare a suitable input to a convolutional neural network. Although some other possibilities are available, a spectrogram is selected for the purposes of this paper, as suggested in [13]. The spectrogram represents visually the frequencies in a sound signal as it varies with time. The vertical and horizontal axes represent time and frequency, while the third dimension indicates the amplitude of a particular frequency value at a particular time instant. Instead of 3D figure, the third dimension is mostly represented by either the intensity or color. In Fig. 4, the example of spectrogram gained by the detection unit is shown.

As seen in Fig. 4, the spectrogram is, for the purposes of this contribution, created by dividing the segments into 8 windows. Furthermore, a Hamming window is implemented to window the signal, 50 % overlap between consecutive sections is specified and a fast Fourier transform algorithm [14] using all the samples in the window is used for frequency analysis.

C. Decision Making

While a pattern recognition neural network is implemented for decision making in [11], an approach based on convolutional neural networks is used here, as described the introduction of this paper. During last decade or two, many successful architectures of convolutional networks were introduced. A brief list of the most popular ones is available in [15].

The procedure of any neural network design involves training, validation and testing set acquisition, neural network training and, eventually, network testing. The procedure is described in the following section.

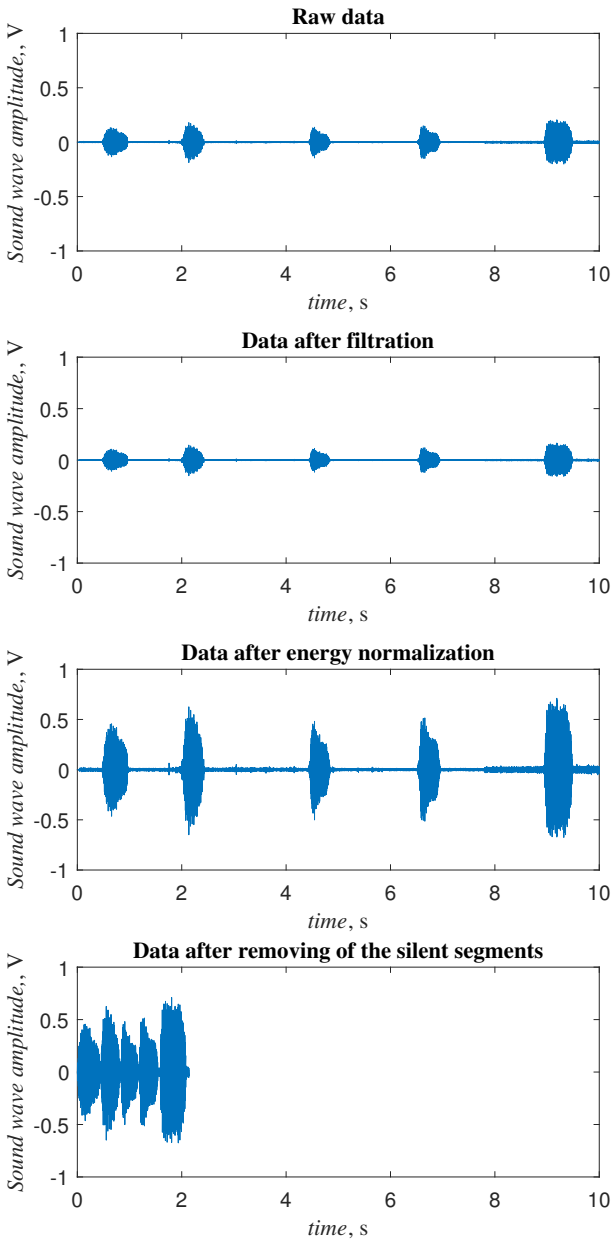

Fig. 3. Sound filtration and normalization.

IV. Convolutional Neural Network for Decision Making

A. Data Acquisition for Training Set

For a training and testing sets, songs, calls and other sounds produced by more than 30 bird species are used. About 50 minutes of total time is recorded, where 6 minutes were produced by the European starling - the specie considered as a pest bird (referred as 'positive' result). The mentioned dataset was also used in [11].

The sound samples are processed using the approach described in Section III-A and III-B.

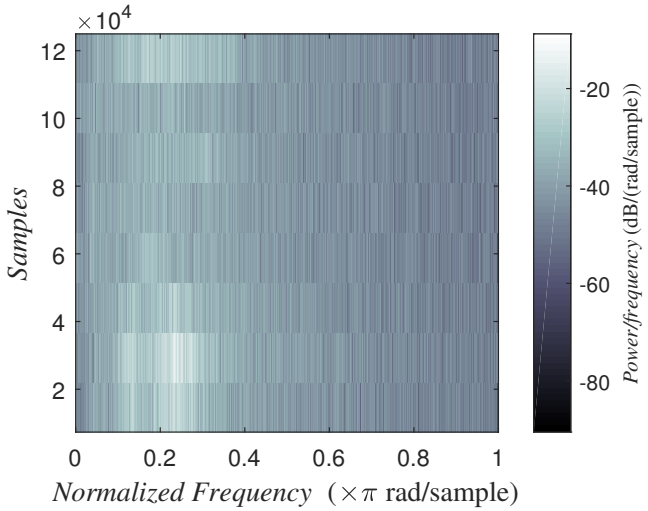

Fig. 4. Spectrogram (3 seconds of sound provide 132 300 samples).

B. Convolutional Neural Network Architectures

A number of architectures of convolutional networks is available for implementation, these days. According to the literature research as well as previous authors' experience, five specific architectures are selected for using as decision making structures. 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 [17], while the fourth one is probably the most cited topology - AlexNet [18]. 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 [19].

C. Convolutional Neural Network Training and Validation

The selected architectures are trained in order to classify correctly the pest bird from the dataset described in section IV-A (70 % used as training set, 15 % used for validation and 15 % as testing set). ADAM algorithm is implemented as optimizer [20]. Since convolutional neural networks are sensitive to the sizes of the input image, two sizes are considered for the image of spectrogram - 50×50 px and 100×100 px. The training experiments are performed a hundred times due to a stochastic character of training and a binary cross entropy function over the validation set is used as a loss function - see (1).

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

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

The resulting values are shown in Fig. 5 for the input size 50×50 px and Fig. 6 for 100×100 px.

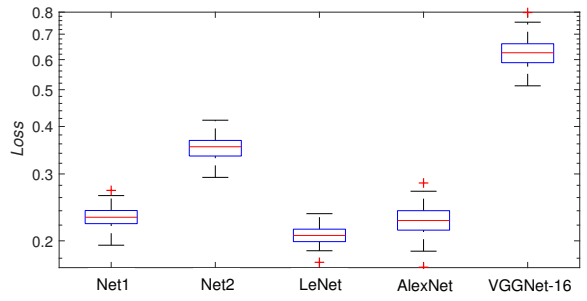

Fig. 5. Training using various architectures and 50×50 px input image.

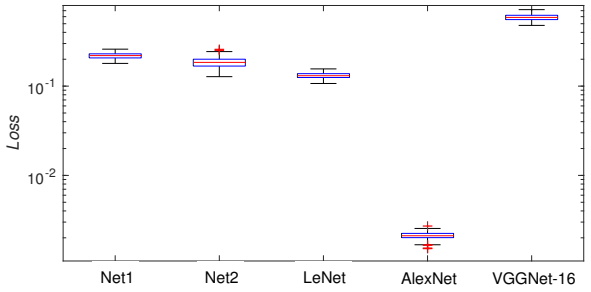

Fig. 6. Training using various architectures and 100×100 px input image.

In the mentioned figures, the central lines in the box graphs 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).

Considering the values of the loss function using the validation set, the AlexNet network with the 100×100 px input image seems to provide the best results. However, the testing procedure should be performed in order to get the most relevant results. Thus, The metrics, described by following equations, are evaluated over the testing set using the most successful networks from training.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4)$$

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

The resulting values are summarized in the following table.

TABLE I
TESTING RESULTS I

Classifier	Accuracy	Precision	Recall
Net1 50×50 px	0.9114	0.8888	0.9952
Net2 50×50 px	0.8885	0.9353	0.8995
LeNet 50×50 px	0.9278	0.9560	0.9377
AlexNet 50×50 px	0.7540	0.7359	1.0000
VGG_16 50×50 px	0.6852	0.6852	1.0000
Net1 100×100 px	0.9311	0.9196	0.9856
Net2 100×100 px	0.9180	0.9842	0.8947
LeNet 100×100 px	0.9770	0.9809	0.9856
AlexNet 100×100 px	0.7081	1.0000	0.5741
VGG_16 100×100 px	0.6852	0.6852	1.0000

The resulting metric values evaluated over the testing set show a bit different trends than presented in Fig. 5 and Fig. 6. Some architectures are apparently overfitted, while the LeNet network provided the most consistent results. Therefore, the LeNet architecture is selected for the further testing.

Another set of training experiments is performed as a next step. Now, the LeNet architecture is considered and the input size of the image is tested - the sizes go from 25×25 px to 200×200 px. Again, the training experiments are performed a hundred times, and a binary cross entropy function over validation set is used as loss function. The resulting values are shown in Fig. 7. In addition, the testing procedure is performed using the same testing set, as in the previous step - see Table II.

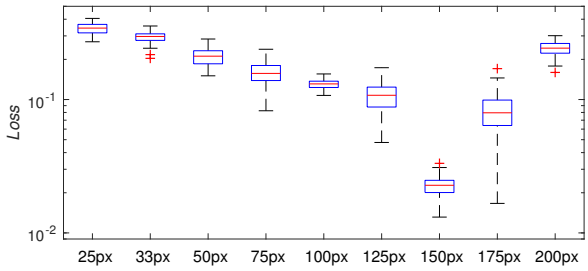

Fig. 7. Training using LeNet architecture and various sizes of input image.

V. Results and Discussion

The pest bird detection unit, based on the convolutional neural network architecture, is designed in the previous sections. The presented training results as well as the testing experiments indicate, that a very selective convolutional neural network with high accuracy, precision and recall can be designed. However, the procedure of design is heavily affected by many tunable parameters; some architectures provide very poor results (esp. VGG_16) and the size of the input image

TABLE II
TESTING RESULTS II

Classifier	Accuracy	Precision	Recall
LeNet 25×25 px	0.8786	0.9215	0.8995
LeNet 33×33 px	8819	8649	0.9808
LeNet 50×50 px	0.9278	0.9560	0.9377
LeNet 75×75 px	0.9409	0.9847	0.9282
LeNet 100×100 px	0.9770	0.9809	0.9856
LeNet 125×125 px	0.9901	0.9858	1.0000
LeNet 150×150 px	1.0000	1.0000	1.0000
LeNet 175×175 px	0.9639	0.9500	1.0000
LeNet 200×200 px	0.8950	1.0000	0.8468

produces decent differences in performance, too; as seen in Table II.

Furthermore, the results from the previous work, using LPC coefficients and a pattern recognition neural network, are summarized in Table III. It is obvious, that well-trained LeNet network with the 150×150 px input image provides much better results with the same testing data. Clearly, the expectations of 100% accuracy, as seen in Table II, would not be fulfilled under the real conditions. Nevertheless, the application of convolutional neural networks brings indispensable improvement in comparison to the previous work.

TABLE III
RESULTS OBTAINED IN [11]

	Calls	Alarms	Begging calls	Songs	Total
Accuracy	0.9043	0.9516	0.8857	0.8627	0.8963
Precision	0.9362	0.9375	0.9355	0.9400	0.9375
Recall	0.8800	0.9677	0.8286	0.8103	0.8621

VI. CONCLUSION

The pest bird detection unit design, based on convolutional neural network, is presented in this contribution. The process of design is especially focused on the selection of the most suitable network architecture. The tests performed in this paper indicate, that the selection of a suitable architecture affects heavily the overall performance of the detection unit. While the VGG_16 and the AlexNet architectures provide relatively poor results, the LeNet architecture shows no misclassified results over the whole testing set. Furthermore, the comparison to previously designed detection unit, based on linear predictive coefficients and pattern recognition neural network, proved indispensable improvements in accuracy.

ACKNOWLEDGMENT

The work has been supported by the IGA Funds of the University of Pardubice, Czech Republic. This support is very gratefully acknowledged.

REFERENCES

[1] J. Bishop, H. McKay, D. Parrot, and J. Allan, *Review of international research literature regarding the effectiveness of auditory bird scaring techniques and potential alternatives*. Central Science Laboratories, 2003.

[2] D. Stowell, E. Benetos, and L. F. Gill, "On-bird sound recordings: Automatic acoustic recognition of activities and contexts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1193–1206, June 2017.

[3] C. Wasserzier, D. Fischer, and T. Rheinhard, "Development of a radar sensor for reducing the risk of bird collisions with wind turbines," in *2017 18th International Radar Symposium (IRS)*, June 2017, pp. 1–9.

[4] Y. Su, C. Chang, J. Kuo, and J. Kuo, "Design of self-injection-locked radar system for birds wingbeat frequency detection," *IEEE Sensors Journal*, vol. 18, no. 24, pp. 10 010–10 017, Dec 2018.

[5] W. K. Poon, C. J. Wong, K. Abdullah, E. S. Lim, and C. K. Teo, "Development of migratory birds population monitoring system using digital single reflex camera," in *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, Aug 2011, pp. 136–140.

[6] L. Wei, G. Mirzaei, M. W. Majid, M. M. Jamali, J. Ross, P. V. Gorsevski, and V. P. Bingman, "Birds/bats movement tracking with ir camera for wind farm applications," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 341–344.

[7] J. Jerabek, L. Zaplatilek, and M. Pola, "A proposal of radio ultrawideband systems for precision indoor localization," 2015, pp. 355–358.

[8] P. Dolezel, M. Mariska, and I. Taufer, "Possibilities of feedforward multilayer neural network classifier as a detector of pest birds in vineyards," *International Journal of Engineering Research in Africa*, vol. 18, pp. 184–191, 2015.

[9] J. Markel and A. Gray, *Linear Prediction of Speech*. Springer Berlin Heidelberg, 1976, iSBN: 978-3-642-66288-1.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[11] P. Dolezel, P. Skrabanek, and L. Gago, "Pattern recognition neural network as a tool for pest birds detection," in *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 2017.

[12] D. Moore, "Demonstration of bird species detection using an acoustic wireless sensor network," in *Local Computer Networks, 2008. LCN 2008. 33rd IEEE Conference on*, 2008, pp. 730–731.

[13] P. Li, M. Chen, F. Hu, and Y. Xu, "A spectrogram-based voiceprint recognition using deep neural network," in *The 27th Chinese Control and Decision Conference (2015 CCDC)*, May 2015, pp. 2923–2927.

[14] J. W. Cooley, "The re-discovery of the fast fourier transform algorithm," *Mikrochimica Acta*, vol. 93, pp. 33–45, 08 1987.

[15] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 International Conference on Communication and Signal Processing (ICCSP)*, April 2017, pp. 0588–0592.

[16] F. Millstein, *Deep Learning with Keras*. CreateSpace Independent Publishing Platform, 2018.

[17] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik, "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, Oct 1994, pp. 77–82 vol.2.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 2, 2012, pp. 1097–1105.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980