DOI: 10.61782/fa.2023.0469

forum acusticum 2023

OBJECT CLASSIFICATION IN AUTOMOTIVE ULTRASONIC SENSING USING A CONVOLUTIONAL NEURAL NETWORK

Jona Eisele[1,2,3*] **André Gerlach**[1] **Marcus Maeder**[2]
Andreas Koch[3] **Steffen Marburg**[2]
[1] Robert Bosch GmbH, Corporate Research, Robert-Bosch-Campus 1, 71272 Renningen, Germany
[2] Chair of Vibroacoustics of Vehicles and Machines, Technical University of Munich, Germany
[3] Institute for Applied Artificial Intelligence, Stuttgart Media University, Germany

ABSTRACT

The challenges of automated driving and driver assist systems increasingly require enhanced sensing of the vehicle environment. Ultrasonic sensors are used in parking and maneuvering situations to calculate the distance to obstacles using the pulse-echo method. Because of their robustness, low production costs and widespread use, increasing the performance of ultrasonic sensors is of great interest. A signal processing pipeline and deep learning methods for classifying obstacles using a single ultrasonic sensor are presented. Time-frequency images that are forwarded to a convolutional neural network are extracted using the continuous wavelet transform. The classification of seven object classes and the classification of traversability is performed in a semi-anechoic chamber and on an asphalt parking space. Promising results are achieved in classifying the traversability of obstacles. However, the discrimination of small objects can be challenging, especially on asphalt ground, which leads to interfering clutter reflections.

Keywords: *ultrasound, signal processing, deep learning*

1. INTRODUCTION

Automotive ultrasonic sensors are used for sensing the near field of vehicles, especially in parking and maneuvering situations. Currently, the distance to obstacles is calculated based on the pulse-echo method. However, further enhancement of surround sensing is crucial for automated driving applications and driver assist systems. Because of their low production costs, robustness and widespread use, increasing the performance of ultrasonic sensors is of particular interest [1]. Notably, a classification of obstacles is desirable.
In general, classification tasks using automotive ultrasonic sensors have been poorly addressed. There has been some work on classifying ground types [2, 3] or obstacle height [4]. A more comprehensive range of studies on classifying acoustic echoes is available in the fields of non-destructive testing (NDT) [5, 6], ultrasonography [7, 8], and underwater sonar [9, 10]. However, in these applications, usually transducer arrays are used, allowing beamforming and imaging methods. For automotive sensing, low-cost ultrasonic sensors consisting of a single piezo-electric transducer are employed [1]. Therefore, different approaches are required to extract relevant features from the acquired signals.
In this article, a summary of the published work in [11] about object classification in automotive ultrasonic sensing and a discussion about related and future work are given. In Sec. 2, relevant features in acoustic echoes for target discrimination are considered. In Sec. 3 – 6, the processing pipeline and classification results from [11] are summarized and discussed. In Sec. 7, a conclusion, and an outlook on future investigations in the field of automotive ultrasonic sensing are given.

2. TARGET PROPERTIES IN ACOUSTIC ECHOES

Bats, dolphins, and blind people use echolocation to navigate their surroundings. Moreover, they can extract certain features about scatterers such as distance, height, orientation, size, shape, and surface texture [12–14]. The distance to an object can be determined by echo delay, based on the

[*] ***Corresponding author**: jona.eisele@de.bosch.com*

time of flight from the transmitter to the object and vice versa. With increasing echo delay, the echo amplitude decreases, as described by the transmission loss. The transmission loss comprises geometric spreading loss and atmospheric sound absorption [15]. Sound absorption in air increases drastically with higher frequencies. Further, higher frequencies are attenuated due to the transmitter and receiver's directivity index, indicating the scatterer's direction [16]. The energy of an echo also depends on the target strength TS of the scatterer, which is defined as

$$TS = 10 \lg \left(\frac{I_{\text{echo}}}{I_{\text{imp}}} \right) \text{ dB}, \quad (1)$$

with the intensity of the echo I_{echo} and the intensity of the impinging sound I_{imp} at a defined distance from the target. Since the impedance difference between most relevant objects and air is large, TS mainly depends on the target's acoustic cross-section, i.e. the size of the reflecting surface, and can vary from different directions [17]. Further, the target strength can be decreased if the surface of the target consists of absorbing material.

Most objects consist of multiple sub-reflectors, producing multiple echoes (highlights). The number and distance of the single highlights depend on the object's geometrical features. In the overall backscatter, the single echoes are combined into a delay spread echo waveform that exceeds the length of the transmitted signal. Furthermore, movement of the scatterer or transmitter/receiver causes Doppler spreading [18]. Resulting interferences of the highlights appear as ripples and notches that can be seen in the temporal and spectral structure of the echo signal [16].

3. DATA ACQUISITION

The data set being described in [11] contains measurements of 30 relevant objects in parking and maneuvering at a sample rate of 215 kSa/s. The measurements are performed in a semi-anechoic chamber (lab data) and an asphalt parking space (field data). Each object is measured in 151 iterations at 55 positions, producing 249,150 labeled measurements per environment. A piezo-electric ultrasonic sensor is used to transmit frequency-modulated pulses (chirps) ranging from 42.5 to 52.5 kHz. At higher frequencies, the signals would be increasingly attenuated by sound absorption, resulting in a lower detection range. At lower frequencies, more extraneous sound sources would be included [1].

For automotive ultrasonic ranging, usually correlation-based thresholding methods are applied to detect single echo points [19]. These echo points, containing time of flight and amplitude information, can be seen as a very compressed representation of the full transducer signal. However, richer features based on the raw time signals should be considered for classification of obstacles. Therefore, an interface for capturing the transducer signal should be provided in future sensors. In [11], a condenser microphone as receiver is used as a practical solution for research purposes. The ultrasonic sensor's transfer function is applied to the captured signals to imply the sensor's frequency response. The transmitter and receiver are mounted in a wooden plate (Figure 1) on a linear rail allowing stationary and dynamic measurements.

Figure 1: Front view (left) and rear view (right) of the ultrasonic sensor and condenser microphone

4. FEATURE EXTRACTION

To suppress unwanted noise and to limit the signals to the frequency range of interest, a finite impulse response (FIR) bandpass filter with a lower passband frequency of 40 kHz and a higher passband frequency of 55 kHz is applied. The object-related backscatter is then cut out from the time signal based on the known object distances. The window length is 768 samples, covering the total backscatter of even broad scatterers. In practice, finding the origin of the windows may be performed by a sliding window approach or based on the conventionally used pulse-echo method for echo detection.

As discussed in Sec. 2, relevant features for object discrimination are in the temporal as well as in the spectral structure of an echo. Therefore, we apply the continuous wavelet transform (CWT), which is seen as a linear time-frequency transform, for feature extraction. The CWT of a time signal $x(t)$ is defined as

$$CWT(\tau, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\ \psi^* \left(\frac{t-\tau}{a} \right) \text{d}t, \quad (2)$$

with the complex conjugated wavelet function ψ^*, the location of the wavelet τ, and the wavelet scaling factor a. The wavelet is slid over the time signal with different scaling factors compressing or dilating the wavelet. The scaling can then be related to frequencies in the analyzed signal, based on the center frequency of the wavelet. In contrast to the more conventional short-time Fourier transformation, this allows an improved time resolution for higher frequencies and an improved frequency resolution for lower frequencies [20, 21]. The scalogram is then

derived as $S = |CWT(\tau, a)|^2$, representing frequency-related energies over time. Scalograms of the backscatter of a pedestrian and a tube are shown in Figure 2.

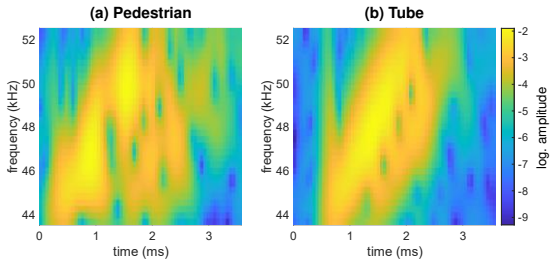

Figure 2: Scalogram features [11]

5. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks (CNNs) are artificial neural networks that are very popular in computer vision for processing image data. In contrast to conventional neural networks, CNNs use convolutions with learned kernel weights resulting in a significantly smaller number of trainable parameters due to the shared weights of the kernels. The kernels are slid over an input image, extracting feature maps by element-wise multiplications. To reduce the feature map dimensions, pooling layers are applied giving invariance to small translations [22]. Recently, CNNs are also successfully applied for tasks in the field of acoustics such as direction-of-arrival (DOA) estimation [23], acoustic scene classification [24], or sonar target recognition [9].

The proposed CNN architecture is shown in Table 1. Zero-padding is applied for each convolutional layer, followed by batch normalization and the ReLU activation function. The 64×32-pixel scalogram images are given into a 2D convolution layer of shape 5×7 to extract low-level feature maps. After an average pooling layer, two convolution layers of shape 1×5 and 5×1, respectively, are applied to separately extract temporal and frequency features. Two convolution layers of shape 3×3 are then used to extract high-level features. The feature maps are flattened and concatenated with the distance feature which is defined by the origin of the considered window. Finally, two fully connected layers are used for classification. After the second fully connected layer, the softmax function is applied to map the output to class-specific probabilities. Dropout and early stopping are used for regularization [22]. To perform a binary classification regarding traversability, the softmax function can be replaced by the sigmoid function with a single neuron in the last fully connected layer.

Stochastic gradient descent (SGD) is used as an optimizer during training with the cross-entropy loss function

$$\ell(p, q) = -\sum_{c=1}^{C} p(y)\ \log\big(q(y)\big)\,, \qquad (3)$$

where C is the number of classes, $p(y)$ the target distribution of the labels y, and $q(y)$ the estimated distribution. To enhance the model's robustness, domain-specific data augmentation is applied. A combination of methods such as noise injection and time shifting produces the best results. [11]

Table 1: CNN architecture [11]

Layer	Output dimension
Scalogram input	64×32
Convolution (5×7)	16×64×32
Avg. pooling (2×2)	16×32×16
Convolution (1×5)	32×32×16
Convolution (5×1)	32×32×16
Avg. pooling (2×2)	32×16×8
Convolution (3×3)	64×16×8
Avg. pooling (2×2)	64×8×4
Convolution (3×3)	64×8×4
Avg. pooling (2×2)	64×4×2
Flatten	512
Concatenate (+ distance)	513
Fully connected	256
Fully connected	7

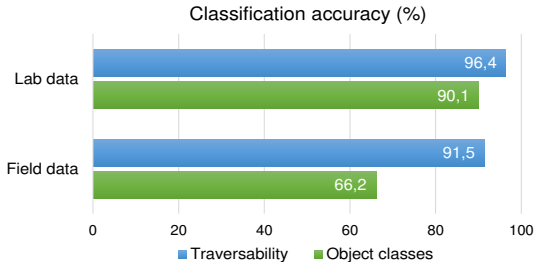

Figure 3: Classification accuracies

6. CLASSIFICATION RESULTS

The classification accuracies for seven object classes (no object, small object, curb, bag, tree, tube/pole, pedestrian) and traversability using lab and field data are shown in Figure 3. In both environments, over 90 % accuracy is achieved for traversability. Classifying the object classes is more challenging than traversability. Using lab data, 90.1 % accuracy is achieved for the object classes, while only 66,2 % is achieved for the field data. This can be ascribed to more disturbances in the field, mainly consisting of ground clutter. Especially the accurate detection of small objects is hindered by interfering clutter. The classification results are discussed in more detail in [11].

7. CONCLUSIONS AND FUTURE WORK

A processing pipeline for classifying obstacles in automotive ultrasonic sensing has been proposed using extracted scalogram images as an input to a CNN. Promising classification results have been achieved, especially regarding traversability. For future studies, it is planned to include multiple measurement cycles in the classification process to increase the model's robustness. A simple majority vote or more sophisticated architectures, such as convolutional recurrent neural networks, should be examined. To enhance the current scalogram representations, where phase information is discarded, it is also planned to investigate the value of adding phase information to the feature inputs. Further, making use of multiple sensors emitting in a round-robin fashion and receiving cross-echoes should be considered. Scanning the object from different directions adds relevant spatial context, potentially increasing classification accuracy.

8. REFERENCES

[1] Noll M, Rapps P. Ultrasonic Sensors for a K44DAS. In: Winner H, Hakuli S, Lotz F, Singer C (eds) *Handbook of Driver Assistance Systems*. Cham: Springer International Publishing, pp. 303–323.

[2] Riopelle N, Caspers P, Sofge D. Terrain Classification for Autonomous Vehicles Using Bat-Inspired Echolocation. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. Rio de Janeiro: IEEE, pp. 1–6.

[3] Bystrov A, Hoare E, Tran T-Y, Clarke N, Gashinova M, Cherniakov M. Road Surface Classification Using Automotive Ultrasonic Sensor. *Procedia Engineering* 2016; 168: 19–22.

[4] Pöpperl M, Gulagundi R, Yogamani S, Milz S. Capsule Neural Network based Height Classification using Low-Cost Automotive Ultrasonic Sensors. In: *2019 IEEE Intelligent Vehicles Symposium (IV)*. Paris, France: IEEE, pp. 661–666.

[5] Sambath S, Nagaraj P, Selvakumar N. Automatic Defect Classification in Ultrasonic NDT Using Artificial Intelligence. *J Nondestruct Eval* 2011; 30: 20–28.

[6] Masnata A, Sunseri M. Neural network classification of flaws detected by ultrasonic means. *NDT & E International* 1996; 29: 87–93.

[7] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis* 2017; 42: 60–88.

[8] Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, Ni D, Wang T. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering* 2019; 5: 261–275.

[9] Neupane D, Seok J. A Review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition. *Electronics* 2020; 9: 1972.

[10] Castro-Correa JA, Badiey M, Neilsen TB, Knobles DP, Hodgkiss WS. Impact of data augmentation on supervised learning for a moving mid-frequency source. *The Journal of the Acoustical Society of America* 2021; 150: 3914–3928.

[11] Eisele J, Gerlach A, Maeder M, Marburg S. Convolutional neural network with data augmentation for object classification in automotive ultrasonic sensing. *The Journal of the Acoustical Society of America* 2023; 153: 2447–2459.

[12] Surlykke A, Nachtigall PE, Fay RR, Popper AN (eds). *Biosonar*. New York: Springer New York. Epub ahead of print 2014. DOI: 10.1007/978-1-4614-9146-0.

[13] Thaler L, Goodale MA. Echolocation in humans: an overview. *WIREs Cogn Sci* 2016; 7: 382–393.

[14] Stroffregen TA, Pittenger JB. Human Echolocation as a Basic Form of Perception and Action. *Ecological Psychology* 1995; 7: 181–216.

[15] Madsen PT, Surlykke A. Echolocation in Air and Water. In: Surlykke A, Nachtigall PE, Fay RR, Popper AN *Biosonar*. New York: Springer New York, pp. 257–304.

[16] Simmons JA, Houser D, Kloepper L. Localization and Classification of Targets by Echolocating Bats and Dolphins. In: Surlykke A, Nachtigall PE, Fay RR, Popper AN (eds) *Biosonar*. New York: Springer New York, pp. 169–193.

[17] Le Chevalier F. Target and Background Signatures. In: *Principles of radar and sonar signal processing*. Boston: Artech House, 2002, pp. 207–281.

[18] Ricker DW. Spread Scattering and Propagation. In: *Echo Signal Processing*. Boston: Springer US, pp. 319–405.

[19] Qiu Z, Lu Y, Qiu Z. Review of Ultrasonic Ranging Methods and Their Current Challenges. *Micromachines* 2022; 13: 520.

[20] Chen VC, Ling H. Time-Frequency Transforms. In: *Time-frequency transforms for radar imaging and signal analysis*. Boston: Artech House, 2002, pp. 25–46.

[21] Stark H-G. Continuous Analysis. In: *Wavelets and signal processing: an application-based introduction*. Berlin: Springer, 2005, pp. 13–40.

[22] Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press, 2016.

[23] Chakrabarty S, Habets EAP. Broadband doa estimation using convolutional neural networks trained with noise signals. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, USA: IEEE, pp. 136–140.

[24] Abeßer J. A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Applied Sciences* 2020; 10: 2020.