BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 67 (71), Numărul 2, 2021
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ
DOI:10.2478/bipie-2021-0010

EYE IMAGE SEGMENTATION METHOD BASED ON THE MODIFIED U-NET CNN ARCHITECTURE

BY

CASIAN MIRON[1,*], LAURA IOANA GRIGORAS[2], RADU CIUCU[3] and VASILE MANTA[1]

[1] "Gheorghe Asachi" Technical University of Iaşi,
Faculty of Automatic Control and Computer Engineering
[2] "Gheorghe Asachi" Technical University of Iaşi,
Faculty of Electronics, Telecommunications and Information Technologies
[3]"Politehnica" University of Bucharest, Faculty of Electrical Engineering

Received: March 29, 2021
Accepted for publication: May 16, 2021

Abstract. The paper presents a new eye image segmentation method used to extract the pupil contour based on the modified U-Net CNN architecture. The analysis was performed using two databases which contain IR images with a spatial resolution of 640x480 pixels. The first database was acquired in our laboratory and contains 400 eye images and the second database is a selection of 400 images from the publicly available CASIA Iris Lamp database. The results obtained by applying the segmentation based on the CNN architecture were compared to manually-annotated ground truth data. The results obtained are comparable to the state of the art.

The purpose of the paper is to present the implementation of a robust segmentation algorithm based on the U-Net convolutional neural network that can be used in eye tracking applications such as human computer interface, communication devices for people with disabilities, marketing research or clinical studies.

The proposed method improves uppon existing U-Net CNN architectures in terms of efficiency, by reducing the total number of parameters used from 31

[*]Corresponding author; *e-mail*: casian_miron@yahoo.com

millions to 38k. The advantages of using a number of parameters approximatly 815 times lower than the original U-Net CNN architecture are reduced computing resources consumption and a lower inference time.

Keywords: deep learning; segmentation; eye dataset; U-Net; convolutional neural networks.

1. Introduction

Eye tracking has become a useful tool with multiple application in fields such as marketing (Wedel and Pieters, 2008), design, human-computer interface (Li *et al*., 2006; Zhang *et al*., 2017), driver safety (Singh *et al*., 2011), assistive technology (Bozomitu *et al*., 2019b), psychological studies on human behaviour (Rahal and Fiedler, 2019), teaching (Ujbanyi *et al*., 2019), gaming (Alhargan *et al*., 2017) and research (Hooge *et al*., 2019).

Eye image segmentation for eye tracking applications is usually performed by using feature or model based methods that can be influenced by outside factors such as image artefacts caused by variations in illumination, reflections from eyeglasses, cornea or contact lenses, the eyelashes and eyebrows, makeup, partially closed eye (Bozomitu *et al*., 2019a). Considering these factors, the need for a more robust method has risen, and such a method is segmentation based on efficient convolutional neural networks.

Traditionally, eye image segmentation has been performed using various methods adapted from other segmentation applications such as text analysis. These methods require the identification of a threshold value, either fixed or adaptive, in order to separate the pupil contour from the background image. An in-depth analysis of eye image segmentation methods is presented in (Pasarica *et al*., 2017). This study presents a comparison between three subclasses of segmentation methods: fix threshold segmentation (fixed threshold, quantitative threshold (Zhang and Gerbrands, 1994) and cumulative distribution function (Lee, 2001) method), global threshold segmentation (minimum error Kittler (Kittler and Illingworth, 1986) and characteristic separation), and local adaptive threshold (Bradley (Bradley and Roth, 2007), Bernsen (Bernsen, 1986) and Niblack (Niblack, 1985) methods). The results presented in this study show the highest segmentation accuracy for the Bradley method at approximately 85%.

This paper presents the implementation of an efficient U-Net convolutional neural network in order to perform eye image segmentation for eye tracking applications (Ronneberger *et al*., 2015).

2. Eye Image Datasets

The eye image segmentation using efficient CNN was performed on two datasets. The first dataset referred to as DB1 was obtained in our laboratory using a USB camera with the spatial resolution of 640x480 pixels. The camera was modified by replacing the light source with IR leds and by applying an infrared glass filter over the camera lens. The camera was mounted on a pair of glasses in order to position it slightly beneath the eye. The dataset contains 400 images with the resolution 640x480 acquired with the eye in different positions due to the eye gaze direction (up/down or left/right). This is due to the fact that the pupil shape transitions from circular to elliptical based on the filming angle and the eye gaze (Bozomitu *et al*., 2019b).

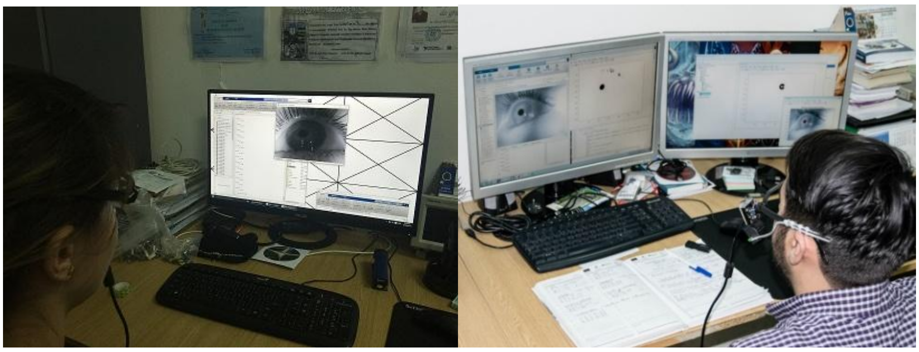

Fig. 1 – Image acquisition in our laboratory.

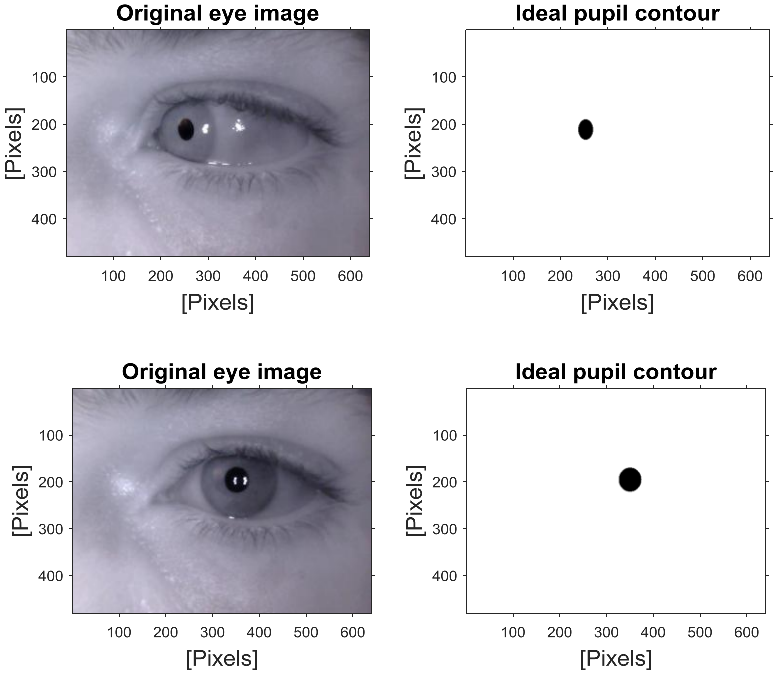

Fig. 2 – Examples of images and the manually annotated pupil contour for DB1.

The second dataset, referred to as DB2, is CASIA-Iris-Lamp which is a larger public dataset of approximately 16000 eye images ("CASIA-Iris-Lamp dataset," 2020). For this analysis we selected 10 images from each of the first 40 subjects from this dataset, resulting in a total of 400 images. The spatial resolution is 640x480 pixels and the images were acquired in IR. The position of the pupil in these images is central, the difference being the fluctuation in lighting conditions.

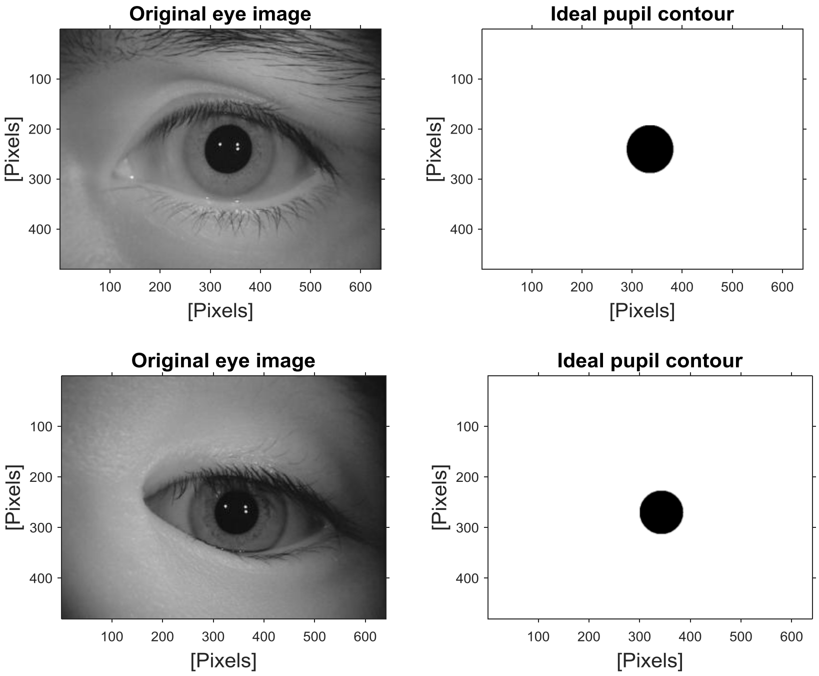

Fig. 3 – Examples of images and the manually annotated pupil contour for DB2.

In order to determine the image segmentation accuracy we compared the contour determined using the proposed efficient CNN method to the manually annotate pupil contour for each image.

3. U-Net CNN Architecture

The U-Net CNN architecture is characterised by the use of a contracting branch (the left side) and an expansive branch (the right side) (Ronneberger *et al.*, 2015), as presented in Fig. 4. The layers of the architecture are presented in detail in Table 1.

The left branch of the CNN architecture has the input image redimensioned to a lower spatial resolution, in our case from 640x480 pixels to

128x96 pixels, and is processed by applying two 3x3 separable convolutions with ReLU, the first with dropout, the second without, followed by a 2x2 max pooling operation with stride 1 for downsampling. The downsampling reduces the input image dimensions by half and we double the number of feature channels in convolutional layers on each step from 8 to 128 (more precisely 8, 16, 32, 64, 128) until the image resolution is 6x8 with 128 feature channels. Separable convolutions are used to reduce the number of parameters necessary. These are depthwise convolutions which use a single 3x3 kernel on a single input channel, for example if the input is 10x10x3 with a 3x3 kernel then the output is 8x8x3 (27 parameters without bias), but the desired final features channel is 8. In order to obtain this value, pointwise convolutions which perform a classic 1x1 convolution are used in order to have an output of 8x8x8 (32 parameters pointwise convolutions) (Chollet, 2017). The total number of parameters for depthwise convolutions and pointwise convolutions is 59. The dropout is used to simulate a larger number of different architectures by dropping nodes from the CNN during training in order to reduce overfitting and to improve generalization error.

The right branch consists of upsampling of the feature map using a 2x2 transpose convolution with stride 2x2, except for the last layers where we use 5x5 stride to upsample the output mask resolution from 96x128 to 480x640. Separable convolutions also start to decrease the number of feature channels from 128 to 8 with the output layer 480x640x1. The number of parameters of the input feature map is reduced by applying for the next layers 1x1 convolutions with a large number of filters.

The U-Net also requires for corresponding cropped feature maps from each branch of the network to be concatenate. This combines the location information from the downsampling branch with the contextual information in the upsampling branch to finally obtain a general information combining localization and context, which is necessary to predict a good segmentation map.

Overall, the U-Net architecture consists of 20 separable convolutional layers, 9 1x1 convolutional layers, 5 transpose convolutional layers, 4 maxpooling and 4 concatenated layers which uses a total of 38356 parameters, as presented in Table 1.

The model training and validation is performed on a dataset that contains images from both DB1 and DB2. For the training subset 428 eye images were randomly selected from the dataset. The validation was performed on 108 eye images and the remaining 264 images were used for testing.

The training was performed using the eye images and the corresponding manually annotated masks in order to determine the model over a number of 50 epochs.

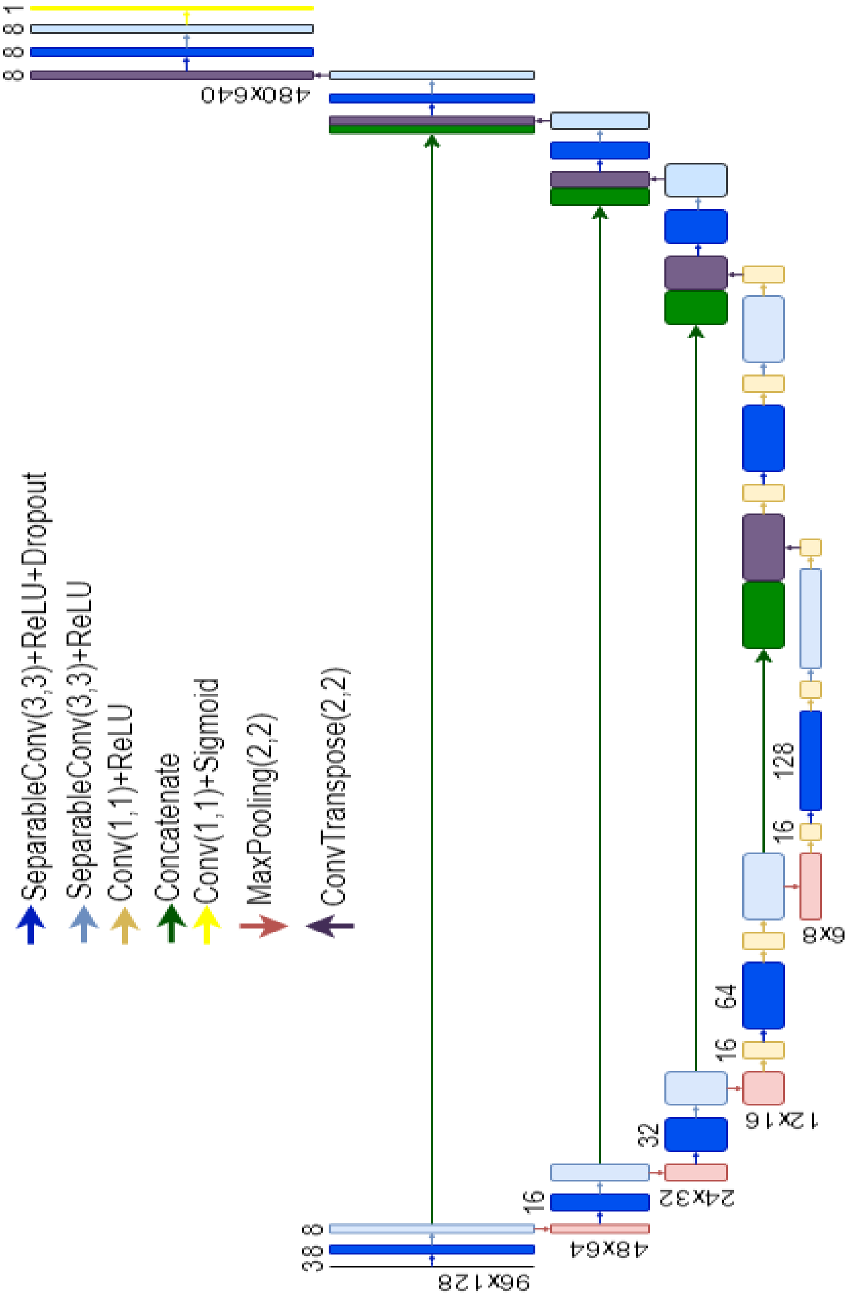

Fig. 4 – The modified U-Net CNN architecture.

Table 1
U-Net CNN Architecture Detailed Layers

No.	Layer	Kernel	No. param.	Dimensions
0	Input Image	-	-	96x128x3
1	SeparableConv+ReLU+Drop(0.1)	3x3	59	96x128x8
2	SeparableConv+ReLU	3x3	144	96x128x8
3	MaxPooling	2x2	-	48x64x8
4	SeparableConv+ReLU+Drop(0.1)	3x3	216	48x64x16
5	SeparableConv+ReLU	3x3	416	48x64x16
6	MaxPooling	2x2	-	24x32x16
7	SeparableConv+ReLU+Drop(0.2)	3x3	688	24x32x32
8	SeparableConv+ReLU	3x3	1344	24x32x32
9	MaxPooling	2x2	-	12x16x32
10	Conv+ReLU	1x1	528	12x16x16
11	SeparableConv+ReLU+Drop(0.2)	3x3	1232	12x16x64
12	Conv+ReLU	1x1	1040	12x16x16
13	SeparableConv+ReLU	3x3	1232	12x16x64
14	MaxPooling	2x2	-	6x8x64
15	Conv+ReLU	1x1	1040	6x8x16
16	SeparableConv+ReLU+Drop(0.3)	3x3	2320	6x8x128
17	Conv+ReLU	1x1	2064	6x8x16
18	SeparableConv+ReLU	3x3	2320	6x8x128
19	Conv+ReLU	1x1	2064	6x8x16
20	ConvTranspose	2x2	4160	12x16x64
21	Concat(13, 20)	-	-	12x16x128
22	Conv+ReLU	1x1	2064	12x16x16
23	SeparableConv+ReLU+Drop(0.2)	3x3	1232	12x16x64
24	Conv+ReLU	1x1	1040	12x16x16
25	SeparableConv+ReLU	3x3	1232	12x16x64
26	Conv+ReLU	1x1	1040	12x16x16
27	ConvTranspose	2x2	2080	24x32x32
28	Concat(8, 27)	-	-	24x32x64
29	SeparableConv+ReLU+Drop(0.1)	3x3	2656	24x32x32
30	SeparableConv+ReLU	3x3	1344	24x32x32
31	ConvTranspose	2x2	2064	48x64x16
32	Concat(5, 31)	-	-	48x64x32
33	SeparableConv+ReLU+Drop(0.1)	3x3	816	48x64x16
34	SeparableConv+ReLU	3x3	416	48x64x16
35	ConvTranspose	2x2	520	96x128x8
36	Concat(3, 35)	-	-	96x128x16
37	SeparableConv+ReLU+Drop(0.1)	3x3	280	96x128x8
38	SeparableConv+ReLU	3x3	144	96x128x8
39	ConvTranspose	2x2	265	480x640x8
40	SeparableConv+ReLU+Drop(0.1)	3x3	144	480x640x8
41	SeparableConv+ReLU	3x3	144	480x640x8
42	Conv+Sigmoid	1x1	9	480x640x1
-	Output	-	38356	480x640x1

4. Experimental Results

The model obtained after the training stage of the U-Net architecture is validated on 108 eye images, after which we tested the model segmentation accuracy on 264 test images. The output image mask for the test images is obtained by applying an additional step to the U-Net architecture which consists of a 2x2 transposed convolutional layer with a stride of 5 that upscales the output mask from the spatial resolution of 96x128 to 480x640 pixels. The segmentation similarity between the manually annotated mask and the output image mask obtained from the trained model applied for the test images is determined by computing the dice similarity coefficient (DSC) (Dice, 1945), (Sorensen *et al*., 1948). The DSC consists of the ratio between the double value of total number of correctly identified pupil and background pixels and the total number of pixels in both masks, as follows:

$$DSC = \frac{2\left|M_1 \cap M_2\right|}{PixM_1 + PixM_2} \tag{1}$$

where M_1 is the manually annotated mask, M_2 is the output image mask, $PixM_1$ and $PixM_2$ are the total number of pixels for each mask, which in our case is equal to the spatial resolution of 640x480.

Fig. 5 and 6 present the results of the testing stage where the original eye image, the manually annotated mask and the model output mask are shown for images given as examples.

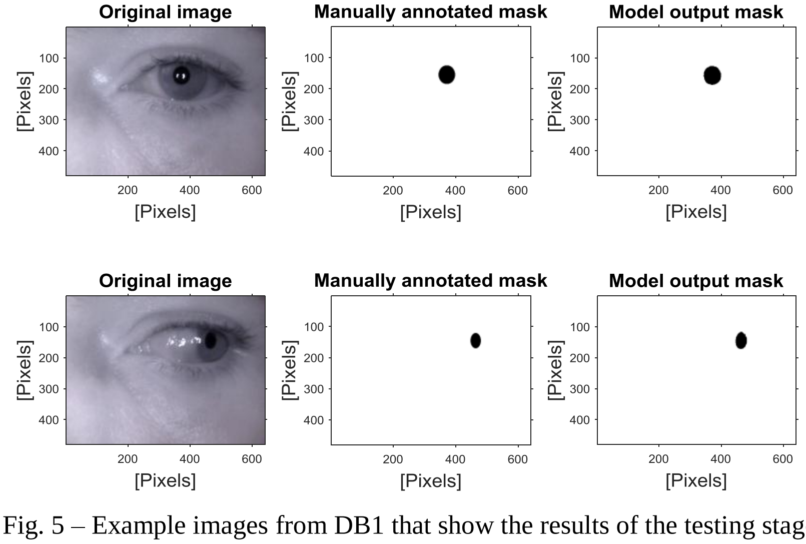

Fig. 5 – Example images from DB1 that show the results of the testing stage.

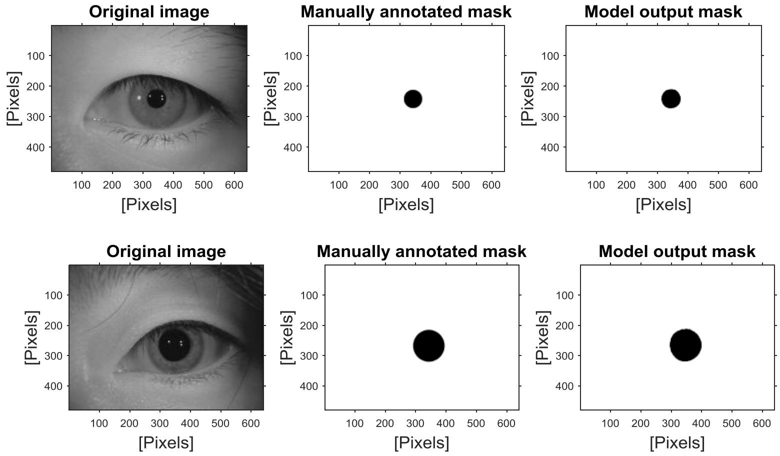

Fig. 6 – Example images from DB2 that show the results of the testing stage.

The overall DSC value obtained after analysis the test subset is 99.81% with the time of processing 1 image is less than 1ms. Table 2 presents similar values obtained for different threshold selection methods, as presented in (Păsărică *et al*., 2017). This study presents a comparison between three subclasses of segmentation methods: fix threshold segmentation (fixed threshold, quantitative threshold and cumulative distribution function method), global threshold segmentation (minimum error Kittler and characteristic separation), and local adaptive threshold (Bradley, Bernsen and Niblack methods).

Table 2
Segmentation Accuracy Using Different Methods from the State of the Art

Segmentation method	Accuracy DB1 (%)	Accuracy DB2 (%)
U-Net architecture (proposed)	99.81	99.81
Fixed threshold	83.41	79.17
Quantitative t = 0.5 % (Zhang and Gerbrands, 1994)	76.04	88.59
Cumulative distribution function (Lee, 2001)	75.87	56.54
Characteristic separation (Păsărică *et al*., 2017)	80.94	80.86
Kittler (Kittler and Illingworth, 1986)	84.77	83.87
Bradley (Bradley and Roth, 2007)	82.42	88.96
Bernsen (Bernsen, 1986)	81.92	85.94
Niblack (Niblack, 1985)	64.87	29.86

5. Conclusions

The eye image segmentation based on the U-Net system architecture is an efficient and robust method. The outside factors that can influence image segmentation which are more prevalent in model or feature based threshold selection do not influence our proposed method. The two datasets analysed present these types of factors, such as different illumination, corneal reflection, different physiological characteristics or different pupil shapes based on gaze direction and filming angle.

The system architecture is designed efficiently by reducing the number of parameters used, with a total of 38356 with the time of inference <1ms. This is performed by using downsampling and upsampling branches, by concatenating the corresponding feature maps from each branch and by using separable convolutions combined with 1x1 convolutional layers and dropout, the latter having the extra benefit of preventing overfitting of the model.

The training and validation stages provide a working model that was used to perform the eye image segmentation for the test subset. The results were quantified by determining the similarity coefficient between the manually annotated eye image mask and the model output mask. The results show a similarity coefficient of 99.81%, which is significantly improved when compared to the reported value of 85% for the same datasets analysed using model and feature based threshold selection methods.

Acknowledgements. The authors would like to thank PhD. Alexandru Păsărică of the “Gheorghe Asachi” Technical University of Iași, Faculty of Electronics, Telecommunications and Information Technologies for the input in writing this paper and for providing the datasets and manually annotated masks.

REFERENCES

Alhargan A., Cooke N., Binjammaz T., *Affect Recognition in an Interactive Gaming Environment Using Eye Tracking*, In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). Presented at the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 285-291, https://doi.org/10.1109/ACII.2017.8273614.

Bernsen J., Dynamic thresholding of grey-level images fcV, In: Proceeding of the 8 International Conference O11 Pattern Rec-Gn Ition, 1986, pp. 125l-1255.

Bozomitu R.G., Niţă L., Cehan V., Alexa I.D., Ilie A.C., Păsărică A., Rotariu C., *A New Integrated System for Assistance in Communicating with and Telemonitoring Severely Disabled Patients*, Sensors 19, 2026, 2019a, https://doi.org/10.3390/s19092026.

Bozomitu R.G., Păsărică A., Tărniceriu D., Rotariu C., *Development of an Eye Tracking-Based Human-Computer Interface for Real-Time Applications*, Sensors 19, 3630, 2019b, https://doi.org/10.3390/s19163630.

Bradley D., Roth G., *Adaptive Thresholding Using the Integral Image*, J. Graph. Tools **12**, 13-21 (2007).

CASIA-Iris-Lamp dataset [WWW Document], 2020. URL http://biometrics.idealtest.org/, Casia-Iris-Lamp.

Chollet F., *Xception: Deep Learning with Depthwise Separable Convolutions*, Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1251-1258.

Dice L.R., *Measures of the Amount of Ecologic Association Between Species*, Ecology **26**, 1945, 297-302, https://doi.org/10.2307/1932409.

Hooge I.T.C., Hessels R.S., Niehorster C.D., Diaz G.J., Duchowski A.T., Pelz J.B., *From Lab-Based Studies to Eye-Tracking in Virtual and Real Worlds: Conceptual and Methodological Problems and Solutions*, J. Eye Mov. Res. **12** (2019), https://doi.org/10.16910/jemr.12.7.8

Kittler J., Illingworth J., *Minimum Error Thresholding*, Pattern Recognit. **19**, 41-47 (1986), https://doi.org/10.1016/0031-3203(86)90030-0.

Lee H., *Method and Circuit for Extracting Histogram and Cumulative Distribution Function for Image Enhancement Apparatus*, Google Patents, 2001.

Li D., Babcock J., Parkhurst D.J., *OpenEyes: A Low-Cost Head-Mounted Eye-Tracking Solution*, In: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, ETRA ’06. Association for Computing Machinery, San Diego, California, 2006, 95-100, https://doi.org/10.1145/1117309.1117350.

Niblack W., *An Introduction to Digital Image Processing*, Strandberg Publishing Company, 1985.

Păsărică A., Bozomitu R.G., Tărniceriu D., Andruseac G., Costin H., Rotariu C., *Analysis of Eye Image Segmentation Used in Eye Tracking Applications*, Rev Roum Sci Techn – Électrotechn Énerg, **62**, 215-222, 2017.

Rahal R.-M., Fiedler S., *Understanding Cognitive and Affective Mechanisms in Social Psychology Through Eye-Tracking*, J. Exp. Soc. Psychol. 85, 103842 (2019). https://doi.org/10.1016/j.jesp.2019.103842.

Ronneberger O., Fischer P., Brox T., *U-Net: Convolutional Networks for Biomedical Image Segmentation*, In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, Springer International Publishing, Cham, 234-241, 2015, https://doi.org/10.1007/978-3-319-24574-4_28.

Singh H., Bhatia J.S., Kaur J., *Eye Tracking Based Driver Fatigue Monitoring and Warning System*, In: India International Conference on Power Electronics 2010 (IICPE2010). Presented at the India International Conference on Power Electronics 2010 (IICPE2010), 1-6, 2011, https://doi.org/10.1109/IICPE.2011.5728062.

Sorensen T.A., Sørensen T., Sørensen T.A., Sørensen T.J., Sørensen T.J., Sorensen T., Sorensen T., Sorensen T.A., Sørensen T., Biering-Sørensen T., *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content, and its Application to Analyses of the Vegetation on Danish Commons*, 1948.

Ujbanyi T., Sziladi G., Katona J., Kovari A., *Pilot Application of Eye-Tracking to Analyze a Computer Exam Test*, In: Klempous R., Nikodem J., Baranyi P.Z.

(Eds.), Cognitive Infocommunications, Theory and Applications, Topics in Intelligent Engineering and Informatics, Springer International Publishing, Cham, 329-347, 2019, https://doi.org/10.1007/978-3-319-95996-2_15.

Wedel M., Pieters R., *Eye Tracking for Visual Marketing. Found*, Trends® Mark. 1, 231-320, 2008, https://doi.org/10.1561/1700000011.

Zhang X., Liu X., Yuan S.-M., Lin S.-F., *Eye Tracking Based Control System for Natural Human-Computer Interaction* [WWW Document], Comput. Intell. Neurosci., 2017, https://doi.org/10.1155/2017/5739301.

Zhang Y., Gerbrands J.J., *Objective and Quantitative Segmentation Evaluation and Comparison*, Signal Process. **39**, 43-54, 1994.

METODA DE SEGMENTARE A IMAGINILOR OCHIULUI PE BAZA ARHITECTURII U-NET CNN MODIFICATĂ

(Rezumat)

Lucrarea prezintă o nouă metodă de segmentare a imaginii ochilor, utilizată pentru extragerea conturului pupilei, bazată pe arhitectura CNN U-Net modificată. Analiza a fost realizată folosind două baze de date care conțin imagini IR cu o rezoluție spațială de 640x480 pixeli. Prima bază de date a fost achiziționată în laboratorul nostru și conține 400 de imagini ale ochiului, iar a doua bază de date este o selecție de 400 de imagini din baza de date CASIA-Iris-Lamp disponibilă public. Rezultatele obținute prin aplicarea segmentării bazate pe arhitectura CNN au fost comparate cu datele adnotate manual. Rezultatele obținute sunt comparabile cu cele din stadiul actual al domeniului.

Scopul lucrării este de a prezenta implementarea unui algoritm de segmentare robust bazat pe rețeaua neuronală convoluțională U-Net, care poate fi folosită în aplicații de urmărire a ochilor, cum ar fi interfața umană cu computer, dispozitive de comunicare pentru persoanele cu dizabilități, cercetări de marketing sau studii clinice.

Segmentarea imaginii oculare este de obicei realizată folosind un algoritm bazat pe un model sau bazat pe caracteristici, care poate fi influențat de artefacte ale imaginii cauzate de condiții de iluminare, cornee, ochelari sau lentile de contact, caracteristici fiziologice ale ochilor sau prezența genelor sau sprâncenelor. Acest lucru necesită un algoritm mai robust și adaptativ care poate fi utilizat pentru a determina cu exactitate conturul pupilei în imagini oculare, cum ar fi în cazul rețelei neuronale convoluționale U-Net. Segmentarea imaginii ochilor folosind diferite metode, cum ar fi pragul de segmentare cantitativ, metoda Bradley, metoda Kittler, metoda de separare a caracteristicilor a fost efectuată în Păsărică și colab., 2017. Acest articol prezintă un studiu comparativ complet al acestor metode, analizând aceleași baze de date propuse în această lucrare.