March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

Salient Object Detection based on Deep Autoencoder Network with ELU Residual Block

Hoi Jun Kim[1], SangHun Lee[*2], HyunHo Han[3]
[1]Master Student, Dept. of Plasma bio display, Kwangwoon University, Korea
[*2]Associate Professor, Ingenium College of Liberal Arts, Kwangwoon University, Korea
[3]Professor, Ingenium College of Liberal Arts, Kwangwoon University, Korea
hoi97@kw.ac.kr[1], leesh58@kw.ac.kr[*2], hhhan@mail.ulsan.ac.kr[3]

Article Info
Volume 83
Page Number: 4395 - 4402
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 26 March 2020*

Abstract
In this paper, we proposed a deep autoencoder segmentation method using ELU residual block and concatenation to reduce the loss of features and improve the accuracy by salient object detection based on deep learning. The existing saliency detection and segmentation methods have an Autoencoder structure, and many features are lost in the process of extracting and compressing features, and the process of expanding and restoring the compressed features. These losses indicate that the background was segmentation, or the object was not segmentation. In the Encoder process, which was a feature extraction stage for improving such a case, detailed information was utilized through skip connection of a residual block, and loss of features is prevented by using an ELU as an activation function. After feature extraction in Encoder process, feature loss occurs because feature was simply expanded in process of Decoder. In order to prevent these losses, the features generated in the process of Encoder were connected to concatenate to utilize in Decoder. The proposed method reduced the loss of features and improved salient object detection in the Autoencoder structure. The proposed method showed improved results compared to the existing method.

Keywords: *Segmentation, Convolutional Neural Network, Autoencoder, Residual Block, Saliency map.*

1. Introduction

Recently, in video processing technology, object segmentation and detection has played an important role in fields such as video surveillance, traffic control, and motion recognition[1]. In addition, recognition by segmentation in various fields such as characters and faces is possible. In these computer vision fields, salient object detection and segmentation are very important[2-5]. By rapidly scanning the entire image, a target area where the eyes are concentrated can be secured. By focusing on the region of interest, it is possible to obtain detailed information on a necessary portion, and pay attention while suppressing information on other regions. The contents of these saliency areas are generally called salient objects.

Salient object detection includes a hand-crafted based detection method[6-9] and a method using CNN (Convolutional Neural Network), one of the deep learning technologies using end-to-end learning[10-12]. The deep learning-based algorithm mainly has an Autoencoder structure, and the Autoencoder is composed of an Encoder that compresses and extracts features and a Decoder that restores the compressed features. These algorithms show higher detection accuracy than existing hand-crafted based models. However,

since the features are compressed into the Encoder process and the image size is reduced, the feature map at the connection with the Decoder network is very small, and the loss of feature information can be increased in the restoration process. In the proposed method, in order to reduce the loss of feature information when proceeding with the Autoencoder network, features were extracted by using residual blocks and various activation functions in each feature extraction and reduction stage of the encoder. In addition, the features of each stage extracted from the Encoder process were used in the Decoder process of concatenation. The experimental results compared and analyzed U-Net[13], FCN(Fully Convolutional Networks)[14] and the proposed method.

2. Related Works

2.1 Autoencoder

Autoencoder[15] is a studied ANNs(Artificial Neural Networks) for compressing image data. Autoencoder has the same structure as FNNs(Feedforward Neural Networks), and is an unsupervised learning model. Unlike FNNs, the size of the input and output layers is always the same. Autoencoder consists of Encoder and Decoder as a neural network that simply copies input to output as shown in [Figure 1].

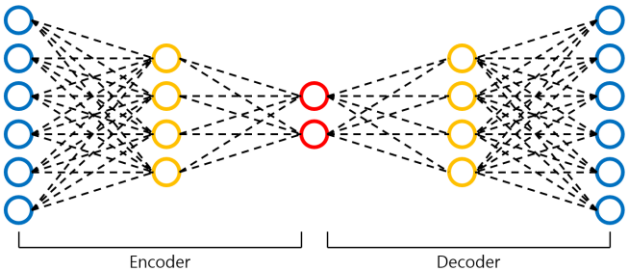


Figure 1. Autoencdoer architecture.

The Encoder is also called a cognitive network, and extracts features of input data, compresses them, and converts them into internal representations. The Decoder is also called a generation network, and converts extracted features and a compressed internal representation to an output. Deep learning network structure often used in the field of salient object detection and segmentation.

2.2 Activation function

The activation function[16] is a function that converts the sum of the input signals to an output signal, determines how to output the input signal, and stacks layers on the network so that nonlinearity can be expressed . The activation function is roughly divided into a unipolar activation function and a bipolar activation function. However, the polarity indicates an activation function that can output only a positive output value, and the bipolar indicate an activation function that can output even if it is negative. Typical activation functions used in CNN include Sigmoid, Tanh, and ReLU (Rectified Linear Unit). The Sigmoid function normalizes the input to a value of (0, 1) and is expressed by equation (1).

$$Sigmoid(x) = 1/(1 + e^{-x}) \quad (1)$$

The Tanh function is a function that comes out to complement the Sigmoid function. The input is normalized to a value of (-1, 1) and expressed by equation (2).

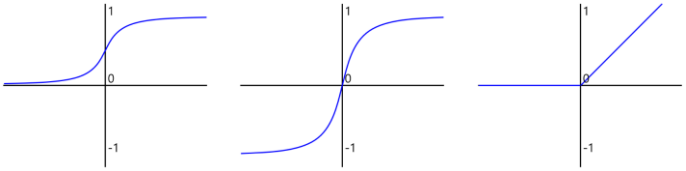
$$Tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \quad (2)$$

However, the Sigmoid and Tanh functions have small differential values, and in the process of optimizing the energy function to be learned, gradient vanishing occurs in which the slope disappears every time the layer passes. ReLU is the activation function that solves these problems and is most frequently used in CNN. The ReLU function is the same as the positive Linear function, and negative values are output as 0. In addition, it is learned 6 times faster than Sigmoid and Tanh that do not execute the $\exp(\cdot)$ function, and is expressed by equation (3).

$$ReLU(x) = max(0, x) \quad (3)$$

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

A graph of each activation function is shown in [Figure 2].



(a) Sigmoid (b) Tanh (c) ReLU

Figure 2. Activation function graph.

3. Proposed Method

In the proposed method, the Autoencoder structure used the residual block to extract features from the input image, and the Decoder used concatenate and deconvolution to restore the feature map to the input size. In the existing CNN method, learning was performed by using ReLU as an activation function and utilizing only features in the positive region without using features in the negative region. The proposed method uses the ELU function as an activation function to utilize the features of all area. A saliency map was extracted through a structure containing a lot of these information. The structure of the proposed method is [Figure 3].

3.1 Encoder for feature extraction

In this paper, as with existing segmentation methods, it was proceeded with the Autoencoder structure. The size of the input image was adjusted to 224 × 224. First, the input image was advanced by a 3 × 3 convolution layer using the Encoder of the Autoencoder to extract features, and the features were compressed using max pooling. Next, unnecessary features were removed through a residual block using the ELU activation function, and features around and inside the salient object were extracted.

If Convolution layer is deeply stacked, the feature was lost in the process of extracting and compressing the feature, and the deep structure causes overcharging. In the proposed method, loss of features was prevented by using a residual block, and overcharging was prevented by using a short skip connection inside the residual block. [Figure 4] is a residual block.

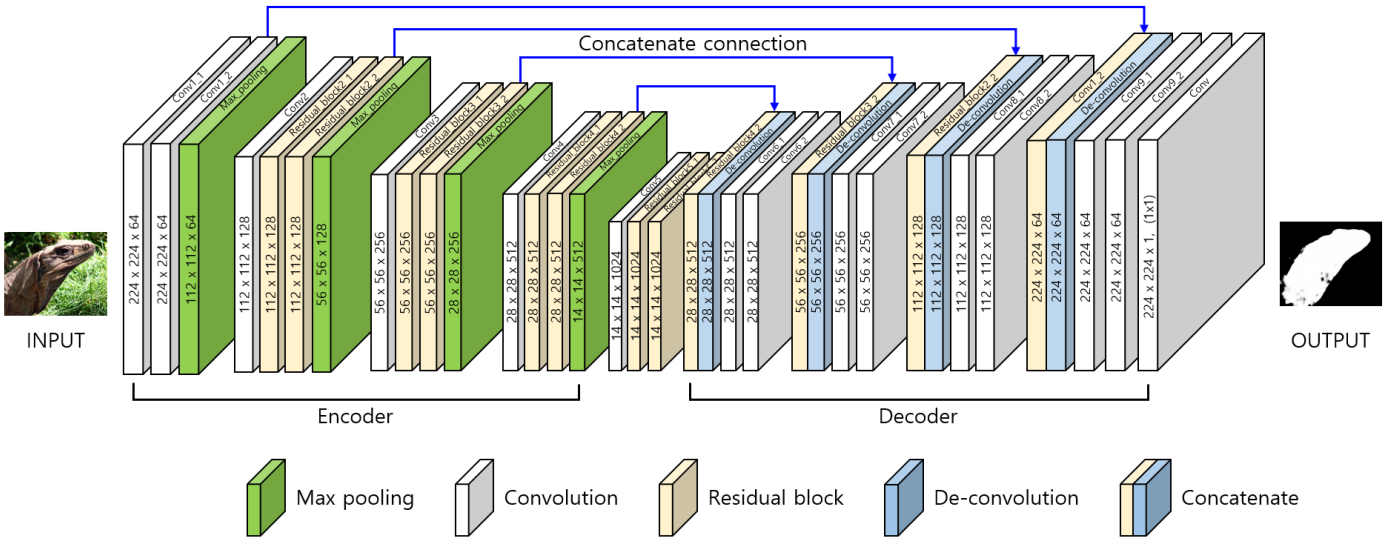
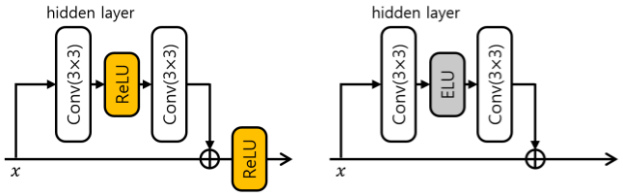


Figure 3. Proposed method architecture.

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402



(a) existing residual block (b) proposed residual block

Figure 4. Residual block architecture.

(A) of [Figure 4] is a conventional residual block composed of two 3 × 3 convolution layers and two ReLU activation functions. In the existing residual block, the result of performing a hidden layer (convolution + ReLU + convolution) and the input x were connected through a short skip connection. After that, the ReLU activation function is applied to output only the features in the positive region from which the features in the negative region have been removed. In this process, feature loss occurs.

The proposed method combines the result of the hidden layer (convolution + ELU + convolution) using the ELU activation function and the input x through a short skip connection. After that, unlike the conventional method, the concatenate was connected to the Decoder for use in the Decoder stage, which is a procedure for restoring features without applying an activation function. This process reduced feature loss and prevented overcharging. equation (4) is an expression of the provided residual block.

$$f_{res}(x) = f_{conv}\left(ELU\left(f_{conv}(x)\right)\right) + x \quad (4)$$

where, the input is x, and $f_{res}(\cdot)$, $f_{conv}(\cdot)$, $ELU(\cdot)$ mean the provided residual function, convolution operation, and ELU activation function, respectively. The ELU activation function is the same as ReLU when the input x is positive and converges to -1 when the input x is negative as shown in [Figure 5]. Therefore, at the time of ReLU as an activation function, the negative node was output to 0, and the problem that was not learned was solved, and the influence of the negative node was reduced, so the ELU activation function was used. Equation (5) is an expression of the ELU activation function.

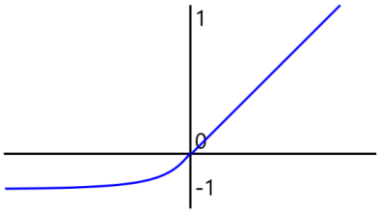


Figure 5. ELU activation function graph.

$$ELU(x) = \begin{cases} x & x \geq 0 \\ a(e^x - 1) & x < 0 \end{cases} \quad (5)$$

3.2 Decoder for feature expansion

In the Decoder process, the feature map was expanded and restored through the deconvolution and convolution layers without using residual blocks. When a feature is extended simply by using only deconvolution, loss occurs even if the relationship with peripheral features is learned. The features extracted in the process of Encoder were used for concatenate to reduce the loss of features. After that, we combined the features extracted from the Encoder process with the features extended to deconvolution through convolution operations. Stride has used deconvolution to double the size of the feature map. Equation (6) is an equation for calculating the size of the feature map via Deconvolution.

$$D(w', h') = D(s(w-1) + k - 2p, \\ s(h-1) + k - 2p) \quad (6)$$

where, w and h represent the height and width of the input, respectively, and s, k and p mean stride, kernel size and padding, respectively.

Concatenate connected the same size feature map for Encoder and Decoder. Concatenate is represented as in [Figure 6] and simply means to continue with layer. Equation (7) is concatenate.

$$f_{concat}(w', h') = \{f_{res}(w', h'), D(w', h')\} \quad (7)$$

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

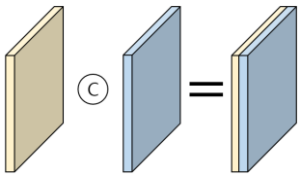


Figure 6. Concatenate-Connection.

3.3 Loss function

As the loss function, a L2 loss function was used. This is a loss function that compares the squared error between the saliency map created using the provided network and the ground truth. We learned while reducing the error between the ground truth and the predicted saliency map via equation (8).

$$f_{Loss} = \sum_{i=1}^{n}(y - \hat{y})^2 \quad (8)$$

where y is the ground truth and $\hat{y}$ is the predicted saliency map.

4. Experimental and Results

The proposed deep autoencoder network used 4447 HKU-IS datasets as training datasets, and used 1,000 ECSSD datasets and 10,000 MSRA10K datasets as experimental datasets. The proposed method compared the FCN, U-Net of deep learning based salient object detection method with the experimental results. To compare and analyze the experimental results, MAE (Mean Absolute Error), Precision, Recall, and F-Measure were used as evaluation indexes. Equation (9) is an evaluation index MAE.

$$MAE = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y) - G(x,y)| \quad (9)$$

$S(x,y)$ indicates a predicted saliency map, and $G(x,y)$ means ground truth. $W \times H$ indicates the size of image. Since MAE is an error rate indicating predicted result, the ground truth, and absolute error value, smaller value, better the performance. Equation (10) is an equation of Precision and Recall, and equation (11) is F-Measure.

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \quad (11)$$

Precision, Recall, and F-Measure are numerical values that represent accuracy, and the closer to 1, the better the performance. In the F-Measure, the value of β^2 was set to 0.3 for evaluation. Precision and Recall are calculated based on whether the pixel values at the same position in the saliency map are equal to ground truth. The predicted saliency map was binarized using the threshold as a threshold.

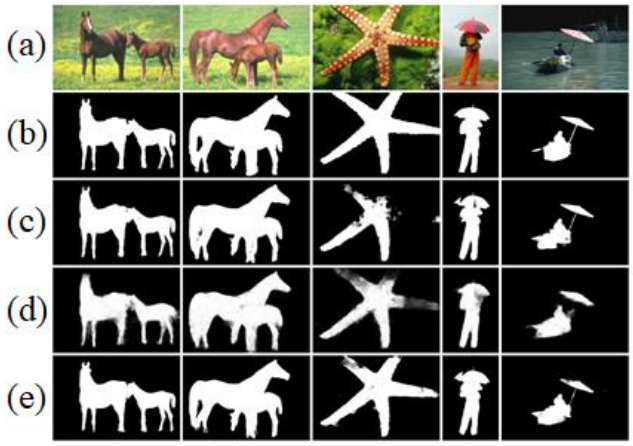


Figure 7. Qualitative comparison of predicted results 1 (ECSSD dataset).

(a) Input image (b) ground truth (c) U-Net
(d) FCN (e) Proposed method

[Figure 7] is an image comparing the proposed method's saliency map with other algorithms using the ECSSD dataset as experimental data. In the comparison algorithm, we used FCN and U-Net to learn and experiment under the same conditions. U-Net lacked information on the boundaries of salient objects, and failed to detect starfish images due to loss of features at salient parts. The FCN well detected the boundary information of the salient object, but lost the feature information inside the object. The proposed method showed better detection results than the two algorithms, and also reduced the loss of features.

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

Figure 8. Qualitative comparison of predicted results 2 (ECSSD dataset).

(a) Input image (b) ground truth (c) U-Net
(d) FCN (e) Proposed method

[Figure 8] shows the experimental results using the ECSSD dataset. Similar to the previous results, U-Net and FCN showed a loss of feature information inside the salient object and lacked boundary information. The proposed method reduced the loss of boundary information and internal feature information of salient objects and showed improved detection results.

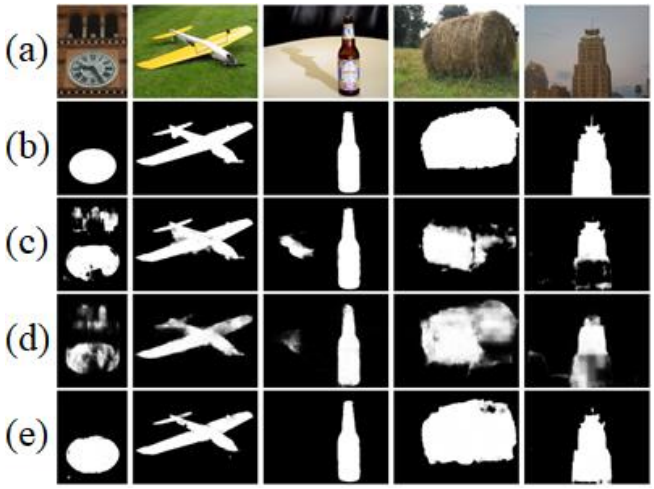


Figure 9. Qualitative comparison of predicted results 3 (MSRA10K dataset).

(a) Input image (b) ground truth (c) U-Net
(d) FCN (e) Proposed method

[Figure 9] shows the experimental results using the MSRA10K dataset. U-Net and FCN showed that loss of feature information for salient objects was large, and the background area was detected. In addition, the FCN occurred a large loss of features that progress through the Deconvolution process, and the input data is expanded to a large size at a time, causing a checkerboard phenomenon. The proposed method suppressed the information of the background area, and also showed that the loss of the feature of the salient object was reduced and improved.

Figure 10. Qualitative comparison of predicted results 4 (MSRA10K dataset).

(a) Input image (b) ground truth (c) U-Net
(d) FCN (e) Proposed method

[Figure 10] shows the experimental results using the MSRA10K dataset. U-Net and FCN showed the result that feature information inside the salient object was lost and the detection failed as in previous result. The proposed method reduces the loss of feature information of salient objects and shows the detection result closest to ground truth.

[Figure 11] shows the proposed method of MAE of ECSSD and MSRA10K dataset and the performance of U-Net and FCN. The proposed method showed lower error rates at 0.0773 and 0.0451 for the two datasets, respectively, and performed better than the other two algorithms.

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

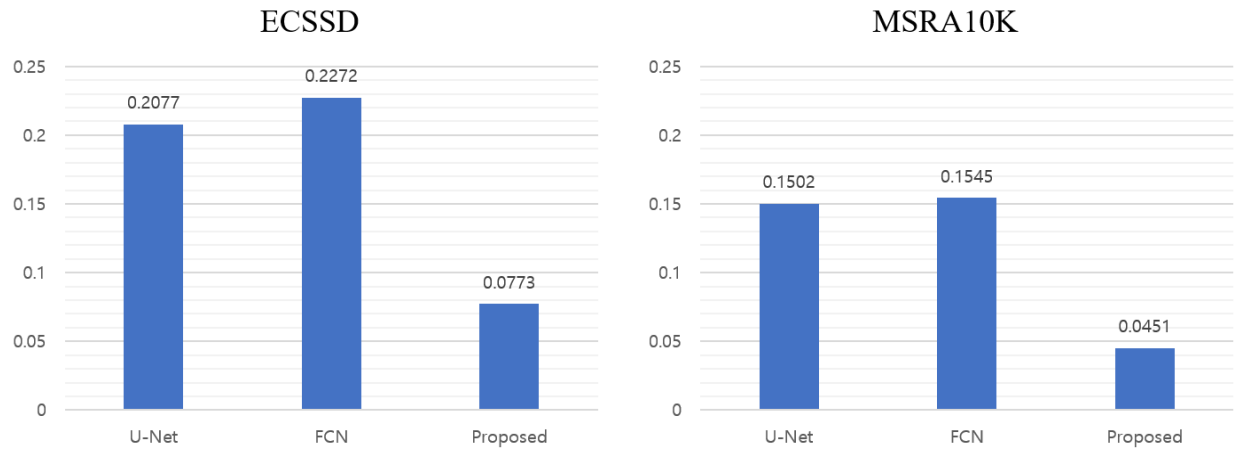


Figure 11. The Mean Absolute Error of which our proposed method and other methods on benchmark dataset.

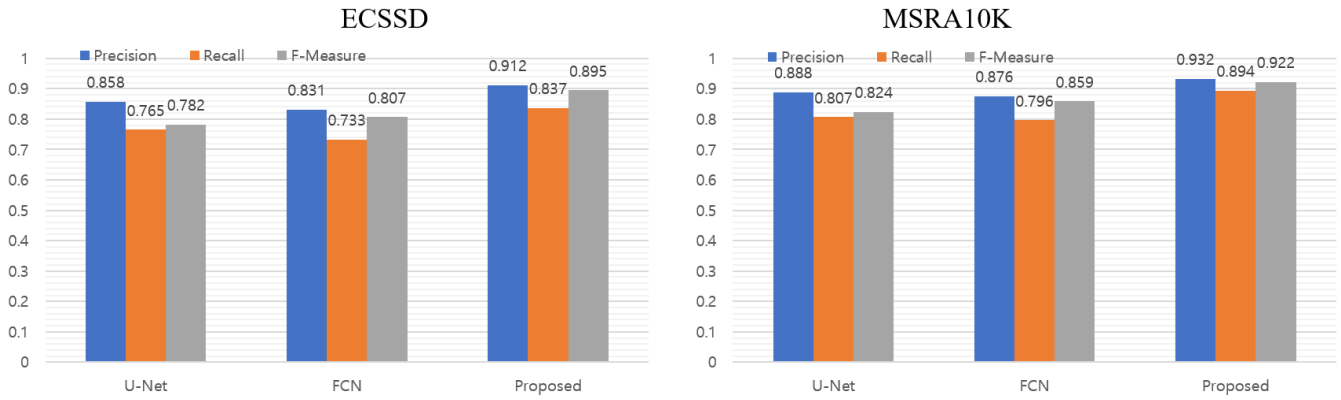


Figure 12. The Precision, Recall and F-Measure of which our proposed method and other methods on benchmark dataset.

[Figure 12] shows the proposed method of ECSSD, Precision, Recall and F-Measure of MSRA10K dataset and the performance of U-Net and FCN. The proposed method showed higher Precision and Recall in ECSSD dataset with Precision and Recall of 0.912 and 0.837, respectively, than the other two algorithms, and the MSRA10K dataset also showed higher performance in 0.932 and 0.894, respectively. Also, F-Measure also showed higher performance than FCN and U-Net respectively 0.895,0.922 of two datasets.

5. Conclusion

The proposed method used a residual block for the deep structure of the Autoencoder structure, and proposed salient object detection to reduce the loss of features and improve accuracy. Existing salient object detection showed the result that the object was not segmented or the object, but the background, was segmented. The proposed method used a residual block to reduce the loss of features during the Encoder process and to construct a deep structure. In addition, using ELU as an activation function, features were extracted using both negative and positive regions, and objects and backgrounds were distinguished. In the Decoder process, concatenate was used to minimize the loss of features in the process of restoring features to the input size. The proposed method showed improved results over existing methods. In future research needs to improve the loss of internal features of objects.

March - April 2020
ISSN: 0193-4120 Page No. 4395 - 4402

6. Acknowledgment

The present Research have been conducted by the Research Grant of Kwangwoon University in 2020.

References

[1] Song H, Zheng Y, Zhang K. Robust visual tracking via self-similarity learning. Electronics Letters [Internet]. 2017 Jan 5;53(1):20–2. DOI:10.1049/el.2016.3011

[2] Pyo S-K, Lee G, Park Y-S, Lee S-H. A license plate detection method based on contour extraction that adapts to environmental changes. Korea Convergence Society [Internet]. 2018 Sep 28;9(9):31–9. DOI: 10.15207/JKCS.2018.9.9.031

[3] Kim H-J, Park Y-S, Kim K-B, Lee S-H. Modified HOG Feature Extraction for Pedestrian Tracking. Korea Convergence Society [Internet]. 2019 Mar 28;10(3):39–47. DOI:10.15207/JKCS.2019.10.3.039

[4] Kim DI, Lee GS, Han GH, Lee SH. A Study on the Improvement of Skin Loss Area in Skin Color Extraction for Face Detection. Korea Convergence Society [Internet]. 2019 May 28;10(5):1–8. DOI:10.15207/JKCS.2019.10.5.001

[5] Lee DW, Lee SH, Han HH, Chae GS. Improved Skin Color Extraction Based on Flood Fill for Face Detection. Korea Convergence Society [Internet]. 2019 Jun 28;10(6):7–14. DOI: 10.15207/JKCS.2019.10.6.007

[6] Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 2012 Nov;34(11):2274–82. Available from: http://dx.doi.org/10.1109/TPAMI.2012.120

[7] Allili MS, Ziou D. Object of Interest segmentation and Tracking by Using Feature Selection and Active Contours. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. IEEE; 2007. Available from: http://dx.doi.org/10.1109/CVPR.2007.383449

[8] Byoung Chul Ko, Jae-Yeal Nam. Automatic Object-of-Interest segmentation from natural images. In: 18th International Conference on Pattern Recognition (ICPR'06) [Internet]. IEEE; 2006. Available from: http://dx.doi.org/10.1109/ICPR.2006.302

[9] Gang Hua, Zicheng Liu, Zhengyou Zhang, Ying Wu. Iterative Local-Global Energy Minimization for Automatic Extraction of Objects of Interest. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 2006 Oct;28(10):1701–6. Available from: http://dx.doi.org/10.1109/TPAMI.2006.209

[10] Han L, Li X, Dong Y. Convolutional Edge Constraint-Based U-Net for Salient Object Detection. IEEE Access [Internet]. 2019;7:48890–900. Available from: http://dx.doi.org/10.1109/ACCESS.2019.2910572

[11] Jiang X, Gao Y, Fang Z, Wang P, Huang B. An End-to-End Human Segmentation by Region Proposed Fully Convolutional Network. IEEE Access [Internet]. 2019;7:16395–405. Available from: http://dx.doi.org/10.1109/ACCESS.2019.2892973

[12] Meng F, Guo L, Wu Q, Li H. A New Deep Segmentation Quality Assessment Network for Refining Bounding Box Based Segmentation. IEEE Access [Internet]. 2019;7:59514–23. Available from: http://dx.doi.org/10.1109/ACCESS.2019.2915121

[13] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Lecture Notes in Computer Science [Internet]. Springer International Publishing; 2015. p. 234–41. Available from: http://dx.doi.org/10.1007/978-3-319-24574-4_28

[14] Long, J., Shelhamer, E., Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015;(pp. 3431-3440).

[15] Baldi, P. Autoencoders, Unsupervised Learning, and Deep Architectures. Proceedings of ICML Workshop on Unsupervised and Transfer Learning, in PMLR. 2012;27:37-49

[16] Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S. Activation functions: Comparison of trends in practice and research for deep learning. 2018; arXiv preprint arXiv:1811.03378.