CNN-Based Salient Object Detection on Hyperspectral Images Using Extended Morphology

Koushikey Chhapariya, Graduate Student Member, IEEE, Krishna Mohan Buddhiraju, and Anil Kumar

Abstract—Salient object detection in hyperspectral images (HSIs) is of interest in various image processing and computer vision applications. Many studies considering spectral information have been reported, extracting only low-level features from a HSI. This letter proposes a convolutional neural network (CNN) based salient object detection method using hyperspectral imagery to utilize spatial and spectral information simultaneously. The proposed methodology incorporates an extended morphological profile (EMP) followed by a CNN to utilize the information from nearby pixels and high-level features simultaneously. We have evaluated the performance of the proposed approach on two independent datasets to verify the generalization ability, viz.: 1) hyperspectral salient object detection dataset (HS-SOD) and 2) Pavia University (PU) dataset. An extensive quantitative analysis of the results revealed that the proposed method significantly outperforms other state-of-the-art methods by approximately $\geq 2\%$ of the area under receiver operating characteristic (ROC) curve (AUC) and F-measure and lower mean absolute error for both datasets.

Index Terms—Convolutional neural network (CNN), extended morphological profile (EMP), extended morphology, hyperspectral image (HSI), salient object detection, spectral–spatial classification.

I. INTRODUCTION

Salient object detection aims at identifying the objects or regions in an image that are visibly distinctive from its vicinity (i.e., background). Considering this, salient object detection has been used in many computer vision tasks such as object recognition [1], visual tracking [2], image retrieval [3], military target detection [4], and so on. Initially, the salient object detection methods were based on the cognitive ability of humans limited only to RGB colors. The features were extracted using contrast difference between objects and image background [5]. On the contrary, with the advancement in technology, many applications such as military intelligence, weather forecasting, monitoring crop photosynthesis, and so on, require a broad-spectrum wavelength with more than three channels (e.g., RGB or YCbCr).

Hyperspectral image (HSI) data provide spectral information from both the visible and infrared portions of the electromagnetic spectrum. However, it is difficult to implement computer vision algorithms on hyperspectral data considering high feature dimensionality and large data volumes. Methods like principal component analysis (PCA) can be employed at the preprocessing stage to reduce the dimensionality. Koch and Ullman [6] first proposed the theoretical foundation for saliency maps on images. Followed by this, Borji and Itti [5] developed a salient object detection method constrained to RGB images using distance measures and local characteristics, such as color, intensity, and so on, for feature extraction. Subsequently, many methods were introduced and modified to improve salient object detection and performance accuracy. Cheng et al. [7] proposed a region-based contrast (RC) method to segment the images based on spatial relations, thereafter extracting the features.

Earlier proposed salient object detection methods used local features such as texture, color, intensity, orientation, and so on, to generate saliency maps. These methods were able to detect and extract only low-level and hand-crafted features based on prior knowledge of the existing datasets [8]. Thus, a learning-based process such as [9] was proposed to overcome this limitation, efficiently integrating different features and improving the detection accuracy for salient objects. With the emergence of advanced learning-based algorithms, convolutional neural network (CNN) has been widely considered for learning automated feature extraction and filters for multiple objects. Attributed to its end-to-end learning tool ability, CNN has been employed for saliency detection techniques [10], [11], [12]. A multiscale deep feature (MDF) [13] using CNN was proposed for salient object detection, in which a multilevel image segmentation method was adopted. This leads to a large number of parameters resulting in heavy computation that is not ideal for HSI data having a large number of features. Wang et al. [14] proposed a recurrent fully convolutional neural network (RFCN) for salient object detection, which iteratively sends the output having any errors back to the network to rectify them. This gives a refined saliency map as an output with more accurate detection. Similarly, a deeply supervised network (DSN) [15] was introduced in which skip-layer connections were used between the loss layer and the last layer of each CNN. With its multiscale and multilevel features, CNN enhances salient regions and preserves object boundaries.

Imamoglu et al. [16] propose a salient object detection technique using CNN-based unsupervised segmentation for the HSIs. Since CNN-based HSI does not need labels to extract high-level features, this process is justified. However, this
technique lacks spatial resolution, which plays a crucial role in detecting objects in HSI. Object detection methods lacking spatial resolution and focusing only on spectral information are inadequate with the object complexity with respect to its background. Appice et al. [17] proposed an autoencoder-based data reconstruction method called autoencoding of hyperspectral imagery for saliency analysis (AISA). It incorporates spectral–spatial information using a set of structuring are erosion and dilation. These operators are used to extract geometrical information. The two fundamental operators are opening and closing.

In this letter, we focus on extracting high-level features by incorporating both spectral and spatial information for salient object detection on the HSI datasets. Instead of individually stacking spectral and spatial features, which leads to information loss, they are assimilated simultaneously. However, automation and extraction of high-level features comprising spectral–spatial information are complex. This motivates us to propose a method for salient object detection by integrating CNN with EMP (as discussed in Section II). By considering nearby pixel information using an EMP, our method highlights two advantages: 1) both spectral and spatial information can be combined simultaneously for extracting high-level features and 2) it detects objects of multispatial scales (size and shape) and aspect ratios in HSIs. Further, a comparative performance analysis of the proposed approach with existing conventional salient object detection methods is carried out using standard evaluation metrics.

II. METHODOLOGY

A. EMP Feature Extraction

The mathematical morphology [19] is considered one of the most used techniques for analyzing inter-pixel dependency. There are many morphological operators available for extracting geometrical information. The two fundamental operators are erosion and dilation. These operators are used to extract structural information from the image using a set of structuring elements (SEs) of known shapes and sizes. The SE detects the shape and size of geometrical structures and controls the smoothness of the output image. The morphological profile (MP) consists of the opening profile (OP) and the closing profile (CP), which are obtained after applying erosion and dilation in two different orders. OP segregates bright features while CP separates dark features from the image.

The OP and the CP at pixel $p$ of the image $I$ can be defined as an $n$-dimensional vector and are given as

$$\text{OP}_i(p) = \gamma_k(p), \quad \forall i \in [0, n],$$

$$\text{CP}_i(p) = \phi_k(p), \quad \forall i \in [0, n]$$

(1)

where $\gamma_k$ is the opening operator of the image with SE of size $i$, $\phi_k$ is the closing operator of the image with a SE of size $i$, and $n$ is the total number of opening and closing operations. The MP operator is a combination of both opening and closing operators and is given as

$$\text{MP}(p) = \{\text{OP}_n(p), \ldots, \text{CP}_1(p), p, \text{OP}_1(p), \ldots, \text{CP}_n(p)\}.$$  

(2)

An MP is usually formulated based on a single wavelength band image. Thus, in hyperspectral data, spectral information is extracted from multiple bands, and an MP for each of the individual bands is computed to present the spatial information at different spectral regions. This approach is called extended MP (EMP), an extended version of the MP [20]. Since HSIs have a large number of feature channels, it is challenging to apply MP directly to the HSI data. PCA was used initially to reduce the dimensionality, and then MP features were extracted from each principal component image. The EMP is represented as

$$\text{EMP}(p) = \{\text{MP}(x^1), \ldots, \text{MP}(x^m)\}$$

(3)

where, $x^i$ is the $i$th channel of the image $x$ after PCA processing and $m$ is the total number of principal components.

B. EMP With CNN

In order to avoid data redundancy and reduce the number of parameters used for the feature map, EMP feature extraction and CNN are combined. For the proposed methodology (Fig. 1), the principal components that account for $\geq 96\%$ of the total explained variance in the data have been used as the reduced set of features. Extended morphology using the reconstruction approach was computed to incorporate spatial information with extracted PCA bands providing spectral information.

A disk-shaped SE has been used with increasing size at each iteration to capture the spatial information. OP eliminates small objects with bright pixels from the foreground, keeping them in the background, while CP discards dark pixels. The reconstruction method restores the objects performing consecutive opening and closing operations. Each MP enlarges or reduces dark and bright parts, resulting in a more homogeneous version of the original image. The EMP output obtained is binarily masked to compare with the provided binary ground-truth data. Precision, recall, and F-measure is used as performance metrics for evaluating the proposed approach. A three-layer CNN model is integrated with the EMP to have rich semantic information with high-level features from hyperspectral imagery for salient object detection. The three convolution layers consist of convolution kernels of size $3 \times 3$ with stride 1.

Generally, a CNN model with a $3 \times 3$ convolution kernel fails to preserve spectral information. Thus, in this research, the output of EMP is used as an input for the CNN model having both spectral and spatial information consecutively. Rectified linear unit (ReLU) is used as a nonlinear activation function followed by batch normalization (BN) and a 40% dropout layer to increase the speed of data training while avoiding overfitting (Table I). Further, the feature maps were input to fully connected layers, and the softmax cross-entropy loss function was applied to achieve convergence in the model while training. The parameters of the CNN models were updated by back-propagation using a stochastic gradient with momentum on the loss.
III. EXPERIMENTAL RESULTS AND DISCUSSION

This section has compared the proposed methodology with prominent salient object detection methods. Two different hyperspectral datasets, namely hyperspectral salient object detection (HS-SOD) and Pavia University (PU), have been considered. The experimental parameters set up for this research are described in Section III-A. The obtained results are described in Sections III-B and III-C, further discussed in Section III-D.

A. Parameter Setting

The different hyper-parameters used for this research include a learning rate of 0.001 set for stochastic gradient descent with a momentum of 0.7 and weight decay of 0.0005. Full-resolution images have been used to train the network with a mini-batch of size 10. A total number of 50 iterations was selected with 89,798 trainable parameters (Table I). The labeled dataset has been split into 25% for training and 75% for testing. The weights of the kernels were randomly initialized. The proposed framework is implemented using the TensorFlow repository on an online platform called Google Colab [21]. Google Colab is based on a Jupyter notebook that can run in any environment with a graphical processing unit (GPU), 12 GB of random access memory (RAM), and 358.27 GB of cold storage for data computation.

B. Experimental Results

1) Dataset Details: The performance of this research work was evaluated on two independent datasets to verify the generalization ability, viz., HS-SOD and PU datasets.

The HS-SOD [22] dataset consists of 60 HSIIs collected from different scenes at the public parks of Japan, between August and September 2017. The images were collected in 151 spectral bands with an image size of 1024 × 768 pixels, among which 81 spectral bands were considered to have a visible spectral range from 380 to 780 nm. The ground truth and the rendered color image (sRGB) were provided with the dataset. This dataset can be downloaded from GitHub https://github.com/gistairc/HS-SOD.

The PU dataset acquired in 2001 by the reflective optics system imaging spectrometer (ROSIS) covers a site in Northern Italy. It has a spatial resolution of 1.3 m and 115 spectral bands, each of size 610 × 340 pixels. Among 115 spectral bands, 12 noisy bands were removed. The ground truth data comprising nine land use/landcover classes were split into two halves, one for training and the other for testing.

2) Results: Among 60 HSIIs in the HS-SOD dataset, 15 images were used for training and the remaining images for performance evaluation. This section analyzes the resulting output for the proposed methodology and compares it with other developed methods. For the comparison, two conventional methods, including RC [7], MDF [13] and four recent CNN-based methods, including RFCN [14], DSN [15], AISA [17] and HSISO [18], have been considered. A total of 20 PCA bands were extracted, having ≥96% of the total explained variance in the data. In Fig. 2, the proposed methodology at different iterations for the HS-SOD dataset is illustrated. With the increase in the number of iterations, the boundary edges are reformed, thus reducing blurriness. Different images of the HS-SOD dataset considering various conditions such as large objects, small objects, simple center bias scenes, and complex and multiple object scenes were categorized to show visual comparison. From Fig. 3, it can be observed that the salient object detection results generated by the proposed method are much closer to the ground truth than other methods. This is because incorporating spatial information results in coherent boundaries and higher contrast between objects and their background. The saliency map computed by CNN-based methods yields better results than the conventional method, as shown in Fig. 3.

The quantitative comparison was performed for the selected datasets with different methods, as shown in Table II. The proposed method yields higher accuracy for all the metrics calculated. It increases the F-measure and AUC score by 2% and gives the minimum absolute error (MAE). Thus, it is inferred that the number of false predictions is significantly

### Table I

Layer-wise Parameter Summary of the Proposed Architecture

| Layer (Type)       | Filter Size | Output Shape     | Parameters |
|--------------------|-------------|------------------|------------|
| Conv_1+ReLU+BN     | 3 x 3 x 32  | (128,128,81)    | 23,360     |
| Dropout_1          | 0.4         | (128)            | 0          |
| Conv_2+ReLU+BN     | 3 x 3 x 64  | (64,64,32)      | 18,496     |
| Dropout_2          | 0.4         | (64)             | 0          |
| Conv_3+ReLU+BN     | 3 x 3 x 64  | (32,32,64)      | 36,928     |
| Fully Connected    | 120         | (7680)           | 0          |
| Flatten            |             |                  |            |
| Output Layer       | -           | (84)             | 10,164     |
| (Dense_1)          |             |                  | 850        |
| (Dense_2)          |             |                  | 0          |

Total Parameters: 89,798
Trainable Parameters: 89,798
Non-Trainable Parameters: 0
Fig. 2. Generated saliency map for different iterations with respect to ground truth.

Fig. 3. Visual comparison of salient object detection method on sample images of HS-SOD dataset. The images are split into different categories to show the efficiency of the proposed methodology for different scenes.

Fig. 4. Comparison of various salient object detection methods on HS-SOD dataset by (a) PR curve and (b) ROC curve.

Fig. 5. Visual results of salient object detection on PU dataset.

C. Discussion

In this section, the behavior of the proposed method has been analyzed and interpreted to provide insight for future work. The quantitative evaluation and the saliency map were generated for comparison purposes. The saliency map generated in Fig. 2 shows reformed boundary edges with an increase in the number of iterations, thus yielding better optimization of the model. Figs. 3 and 5 show the visual comparison of the salient object detection method for the two different datasets to study the efficacy of the proposed approach. This demonstrates that combining spatial information and spectral information using a combination of CNN and EMP performed better in terms of accuracy and generating a saliency map. The performance of the proposed method was also compared with different published methods by computing standard evaluation metrics results on both datasets as shown in Table II. EMP considering nearby pixel information along with CNN could be the primary reason for the proposed method to yield improved accuracy with refined boundaries. However, there were a few images from the HS-SOD dataset for which the proposed methodology failed to generate a saliency map (Fig. 6). These failures are because of objects having extremely low contrast with respect to their background [Fig. 6(a) and (b)]. The other reason may be the inability to identify the depth between the ground and the object [Fig. 6(c)]. The presence of multiple objects having complex shapes, which are very close to each

| Method | HS-SOD | Pavia University |
|--------|--------|------------------|
| RC [7] | 0.741  | 0.128 0.681  0.713 | 0.310 0.887 |
| MDF [13] | 0.809 | 0.142 0.742  0.802 | 0.213 0.892 |
| RFCN [14] | 0.914 | 0.109 0.801  0.815 | 0.081 0.922 |
| DSN [15] | 0.927 | 0.017 0.871  0.824 | 0.102 0.962 |
| AISA [17] | 0.885 | 0.114 0.723  0.793 | 0.159 0.902 |
| HSISO [18] | 0.921 | 0.121 0.907  0.894 | 0.112 0.973 |
| Proposed | 0.942 | 0.013 0.921  0.903 | 0.032 0.991 |
other [Fig. 6(d)] may be responsible for less than optimal performance of the proposed methodology. The possible solution for the above-mentioned problems could be to provide more prior information to the model based on object feature similarity. This can help in updating the learnable weights of the CNN model by providing segment-level information. Apart from this, training the model with more complex scenes can eventually benefit different levels of scenes.

IV. CONCLUSION
In this letter, we proposed a CNN-based salient object detection method on HSI data. The proposed technique simultaneously utilizes spectral and spatial information using EMP and nearby pixel information. Further, a three-layer CNN is integrated with EMP to extract high-level features and preserve the edges of the boundary. The proposed methodology can detect salient objects of multiple spatial scales and aspect ratios. The experimental results on HSIs prove that the proposed method increases accuracy by ≥2% in terms of F-measure and AUC score and low MAE, thus outperforming other methods. The generated saliency map for visual comparison shows a better result with enhanced boundary edges when compared with other methods. On the other hand, we observed that detecting the edges for objects with very low contrast and complex features was comparatively less effective. This ineffectiveness in object detection could be because of fixed convolutional kernel size, insufficient parameter tuning, and optimization. Notably, the proposed CNN-based EMP method for saliency detection is significant in comparison with state-of-the-art techniques and can be further used in studying different saliency objects. Further work is being taken up wherein the spectral–spatial characteristics, and a CNN model with more tunable parameters can be integrated to detect complex salient objects from the image.

REFERENCES
[1] U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom-up attention useful for object recognition?” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2, 2004, pp. 37–44.
[2] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, “Adaptive object tracking by learning background context,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2012, pp. 23–30.
[3] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, “3-D object retrieval and recognition with hypergraph analysis,” IEEE Trans. Image Process., vol. 21, no. 9, pp. 4290–4303, Sep. 2012.
[4] I. Zafar, U. Zakir, I. Romanenko, R. M. Jiang, and E. Edirisighhe, “Human silhouette extraction on FPGAs for infrared night vision military surveillance,” in Proc. 2nd Pacific-Asia Conf. Circuits, Commun. Syst., Aug. 2010, pp. 63–66.
[5] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 185–207, Jan. 2012.
[6] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” in Matters of Intelligence. Berlin, Germany: Springer, 1987, pp. 115–141.
[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, “Global contrast based salient region detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, Mar. 2015.
[8] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient object detection: A benchmark,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, “Salient object detection: A discriminative regional feature integration approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2083–2090.
[10] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, “PoolNet+: Exploring the potential of pooling for salient object detection,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jan. 4, 2022, doi: 10.1109/TPAMI.2021.3140168.
[11] Y. Liu, S. Zhang, Z. Wang, B. Zhao, and L. Zou, “Global perception network for salient object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–12, 2022.
[12] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, “Bi-directional progressive guidance network for RGB-D salient object detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 8, pp. 5346–5360, Aug. 2022.
[13] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in Proc. IEEE Comput. Conf. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 5455–5463.
[14] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Salient object detection with recurrent fully convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1734–1746, Jul. 2019.
[15] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, “Deeply supervised salient object detection with short connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 815–828, Apr. 2016.
[16] N. Imamoglu, G. Ding, Y. Fang, A. Kanazaki, T. Koyama, and R. Nakamura, “Salient object detection on hyperspectral images using features learned from unsupervised segmentation task,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 2192–2196.
[17] A. Appice, F. Lomuscio, A. Falini, C. Tamborrino, F. Mazzia, and D. Malerba, “Saliency detection in hyperspectral images using autoencoder-based data reconstruction,” in Proc. Int. Symp. Methodologies Intell. Syst. Cham, Switzerland: Springer, 2020, pp. 161–170.
[18] C. Huang, T. Xu, Y. Zhang, C. Pan, J. Hao, and X. Li, “Salient object detection on hyperspectral images in wireless network using CNN and saliency optimization,” Ad Hoc Netw., vol. 112, Mar. 2021, Art. no. 10369.
[19] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. A. Chanussot, and J. C. Tilton, “Advances in spectral-spatial classification of hyperspectral images,” Proc. IEEE, vol. 101, no. 3, pp. 652–675, Mar. 2012.
[20] Y. Gu, T. Liu, X. Jia, J. A. Benediktsson, and J. Chanussot, “Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 6, pp. 3235–3247, Jun. 2016.
[21] T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. R. Filho, “Performance analysis of Google colaboratory as a tool for accelerating deep learning applications,” IEEE Access, vol. 6, pp. 61677–61685, 2018.
[22] N. Imamoglu et al., “Hyperspectral image dataset for benchmarking on salient object detection,” in Proc. 10th Int. Conf. Quality Multimedia Exp. (qoMEX), 2018, pp. 1–3.