2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506625

A HYPERSPECTRAL APPROACH FOR UNSUPERVISED SPOOF DETECTION WITH INTRA-SAMPLE DISTRIBUTION

Tomoya Kaichi[†] *and Yuko Ozasa*[‡]

[†]Faculty of Science and Technology, Keio University, Japan
[‡]Department of Information System Engineering, Tokyo Denki University, Japan

ABSTRACT

Despite the high recognition accuracy of recent deep neural networks, they can be easily deceived by spoofing. Spoofs (e.g., a printed photograph) visually resemble the actual objects quite closely. Thus, we propose a method for spoof detection with a hyperspectral image (HSI) that can effectively detect differences in surface materials. In contrast to existing anti-spoofing approaches, the proposed method learns the feature representation for spoof detection without spoof supervision. The informative pixels on an HSI are embedded onto the feature space, and we identify the spoof from their distribution. As this is the first attempt at unsupervised spoof detection with an HSI, a new dataset that includes spoofs, named Hyperspectral Spoof Dataset (HSSD), has been developed. The experimental results indicate that the proposed method performs significantly better than the baselines. The source code and the dataset are available on Github.[1]

***Index Terms*—** Unsupervised spoof detection, hyperspectral image, out-of-distribution detection, intra-sample distribution, single-pixel classification

1. INTRODUCTION

Recent advances in deep learning have dramatically improved recognition accuracy. However, they are vulnerable to spoofing attacks [1, 2]. A spoof could be a printed photo of the actual object (print attack), a replayed digital video (replay attack), a fake object, such as an artificial flower (fake attack), etc. Since these spoofs can visually be similar to the actual objects (lives), RGB image classification approaches tend to classify spoofs into the learned category as lives with a high confidence score. It can be difficult to distinguish an effective spoofs from lives, even with the human eye.

Hence, we suggest using hyperspectral images (HSIs) to detect spoofs rather than RGB images that attempt to mimic the process of human vision. An HSI has about 100 or more spectral bands for each pixel. Due to its informative pixel spectra, an HSI can be utilized to distinguish between objects in image scenes in biomedical, environmental, and land survey domains [3, 4]. We believe that HSI captures the differences between the surface materials of lives and spoofs, and its spectral information enables spoof detection without the supervision of spoof samples. The unsupervised approach is desirable because an adequate number of spoof samples is not always available a priori due to the unlimited types of spoof [5, 6].

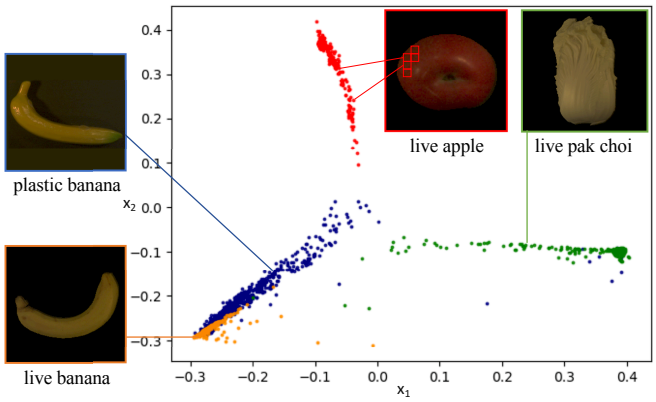

Fig. 1. A toy experiment of feature extraction by the proposed loss function on three live objects and one spoofing (plastic banana). The randomly selected 1000 pixel spectra on each object are embedded in a 2D feature space. The larger variance of the spoof's features can be observed.

In this paper, we address unsupervised spoof detection in contrast to existing RGB-based anti-spoofing approaches that rely on spoof supervision [1, 2, 7, 8]. Unsupervised spoof detection can be a task to detect a spoof as a sample outside the training distribution of live samples. This problem setting is regarded as out-of-distribution (OOD) detection, in which a spoof is treated as OOD. OOD detection is a task of identifying whether a test input is drawn far from the training distribution (in-distribution) or not [9, 10, 11, 12]. We follow the problem setting of multi-class OOD detection, i.e., our model is trained using labeled multi-class live examples, and only then can it detect spoofs in the test phase. However, learning the feature representation to identify spoofs from training data that include only live objects is a challenging problem because of the similarity in appearance of lives and spoofs.

[1] https://github.com/t-kaichi/hyperspoof

To overcome this issue, we propose a method to discriminate spoofs by aggregating the shape-agnostic features. The proposed method takes an HSI as an input and embeds its pixel spectra onto the feature space using a single-pixel classifier. The multiple shape-independent features are obtained from a sample (HSI). As shown in Figure 1, the proposed method identifies a spoof by the distribution of its features, which we call intra-sample distribution, even if the live objects and spoofs have similar surface spectra. Our approach is inspired by the recent success on OOD detection that enables the identification of OOD examples using the OOD score calculated from the intermediate feature representations of a discriminatively trained classifier [9, 10, 11].

We show that our method accurately and robustly detects spoofs without any spoof supervision or fine-tuning with spoof examples. Following the previous studies on hyperspectral image analysis, HSIs of vegetables and fruits were used for the evaluation [13, 14, 15]. The spoofs in the test images include print, replay, and fake attacks. As this study is the first attempt at using an HSI for unsupervised spoof detection, we established a hyperspectral spoofing dataset (HSSD) comprising 15 categories of live vegetables and fruits and their print and replay attacks as spoofs. Additional experiments on the public dataset that included fake attacks [15] validate the versatility of our method. The code and the HSSD are available on our project page.[1] Our contributions are summarized as follows:

- We introduce the use of HSIs for spoof detection and define the novel problem of unsupervised spoof detection.
- We propose a new approach for unsupervised spoof detection that analyzes the intra-sample distribution of an HSI. This approach significantly outperforms the OOD-detection-based baselines on the two datasets, which include various spoofing attacks.
- We create a publicly available dataset for hyperspectral spoof detection.

2. INTRA-SAMPLE DISTRIBUTION

In the proposed approach, a model is trained to classify labeled live objects, and the intermediate features of the model are utilized for discriminating spoofs as per the modern OOD detection approaches [16, 9, 10, 11]. These RGB-based OOD detection approaches map a sample image onto the feature space. Some of them demonstrate that the large distance between the input and the in-distribution samples in the feature space suggests that the input is an OOD sample [9, 10]. However, in terms of spoof detection, training a feature extractor that maps spoof features far away from the feature vectors of live samples is difficult due to the similarity of the lives and spoofs in shape and color.

To identify the difference between lives and spoofs, we train a single-pixel classifier to extract not spatial but spectral features. It achieved high classification accuracy with HSIs due to their high spectral resolution. However, unfortunately, we found that the single-pixel classifier trained with only live samples could not embed spoof features away from the live feature distributions when the spoof surface was spectrally similar to that of the lives or the spectral resolution of the input HSI was low. Figure 1 illustrates the results of toy experiments that visualize the 2D activations from the penultimate layer of a single-pixel classifier that was trained with three live vegetables and fruits. The red, green, and yellow points denote the intra-sample feature embeddings of the live apple, pak choi, and banana, respectively. The blue points represent the embeddings of the pixel spectra of the artificial banana. The spectra of the live and spoof bananas are mapped at a short distance in the feature space, and the spoof banana is classified into a *banana class* with a high confidence score (the mean maximum softmax probability for all pixels was 0.95). This suggests that the spoof detection strategies based on the individual feature embeddings, such as distance-based [9, 10] and classification-confidence-based approaches [17], may suffer from the spoof features projected near the live features.

Therefore, we present to identify spoofs using the distribution of the embeddings within a sample (intra-sample distribution) rather than individual embeddings. Intra-sample distribution reflects the variety of the input spectra in the entire surface of a sample. The pixel spectra on HSIs are determined by the object's 3D shape and albedo in fixed viewpoint and light source setting [18]. Since spoofs differ from lives in shape (e.g., print and replay attacks) and/or albedo (e.g., print, replay, and fake attacks), the embedded feature vectors form different distributions between lives and spoofs.

In the proposed method, we incorporate the difference in intra-sample distribution into the training and spoof decision rules. Specifically, the feature extractor is trained to reduce the intra-sample variance in the feature space only with live images. As shown in Figure 1, this leads to a small intra-sample variance in the test live samples, while a spoof whose spectral distribution differs from that of the training samples has a large variance. The trained model identifies whether a test sample is a spoof from its intra-sample variance. The detailed algorithm is described below.

3. PROPOSED METHOD

The single-pixel classification model f is trained with labeled multi-class HSIs of live objects. Given N number of HSIs $\{x_1, x_2, ..., x_N\}$ in a batch and their corresponding labels $\{y_1, y_2, ..., y_N\}$, where $\forall y_i \in \{1, 2, ..., c\}$ and c is the number of classes, the classifier f takes a hyperspectral pixel $\mathbf{p}_j$ $(1 < j \leq M)$ within x_i as an input and extracts the feature $\mathbf{z}_j$; $\mathbf{z}_j$ is then linearly projected to the classification score $f(\mathbf{p}_j)$ through the softmax activations. Here, M is the number of HSI pixels on the object to be classified, and $\mathbf{z}_j$

is the activations from the penultimate layer of f. The classifier f, constructed using a simple four-layer Multi-Layer Perceptron, is trained using the following loss function,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\{\mathcal{L}_{\mathrm{CE}}(y_i, \frac{1}{M}\sum_{j=1}^{M} f(\mathbf{p}_j)) + \lambda\, \mathsf{tr}(\hat{\mathbf{\Sigma}}_i)\}, \quad (1)$$

$$\text{where } \hat{\mathbf{\Sigma}}_i = \frac{1}{M}\sum_{j=1}^{M}(\mathbf{z}_j - \bar{\mathbf{z}})(\mathbf{z}_j - \bar{\mathbf{z}})^{\mathrm{T}},\ \bar{\mathbf{z}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{z}_j\ . \quad (2)$$

Here, $\mathcal{L}_{\mathrm{CE}}(\cdot)$ and $\mathsf{tr}(\cdot)$ represent the cross-entropy loss and trace of a matrix, respectively. λ is a hyperparameter to balance the contribution of each term. We have explicitly included the variance loss term $\mathsf{tr}(\hat{\mathbf{\Sigma}}_i)$, punishing the large intra-sample variance.

Upon obtaining the feature representations through the proposed training objective, we define a score function that segregates the spoofs based on the variance of the intra-sample distribution. Given a test HSI x and its pixels $\mathbf{p}_k$ $(1 < k \leq W \times H)$, where W and H represent the width and height of x, respectively, the spoof score $s(x)$ is defined by

$$s(x) = \mathsf{tr}(\frac{1}{\|P\|}\sum_{k:\mathbf{p}_k \in P}(\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})^{\mathrm{T}}), \quad (3)$$

$$\text{where } \bar{\mathbf{z}} = \frac{1}{\|P\|}\sum_{k:\mathbf{p}_k \in P}\mathbf{z}_k\ . \quad (4)$$

Here, P denotes the set of the foreground pixels of the input x. $\mathbf{z}_k$ denotes the extracted feature from $\mathbf{p}_k$. A large value of $s(x)$ suggests a high likelihood of being a spoof. We utilize the activations from the penultimate layer $\mathbf{z}_k$ without ensembling the features from multiple layers, unlike [10], because it would require access to the spoof samples during the training.

In the experiments, we extracted the foreground pixels P from x using U-Net [19] trained with live object images. It segmented the foreground of the spoof images as well as the live images due to the similarity in their appearance. Throughout all the experiments on the two datasets against three types of spoofing, the parameter λ in the loss function was fixed at 2.0×10^{-4}.

4. EXPERIMENTAL SETUP

4.1. Dataset

To validate our approach, we created a Hyperspectral Spoofing Dataset (HSSD) that includes live and spoof images of vegetables and fruits. Many studies on hyperspectral analysis or classification focus on vegetables and fruits due to their inter-class similarity and large intra-class variation in shape and color [13, 14, 15]. Hence, the HSSD should contribute to this area.

The HSSD is a collection of 150 live and 30 spoofing HSIs that were captured with a hyperspectral camera (GS, EBA JAPAN Co. Ltd.) covering the visible wavelength from 400 nm to 800 nm with a spectral resolution of 5 nm. For the live images, we captured two sides of five instances each for 15 categories of vegetables and fruits. In this experiment, we used four instances for training and the rest for the test. For the spoof images, we created print and replay attacks that captured the test instance images printed by the laser printer and displayed on the screen. The RGB images corresponding to the HSIs were created by the toolbox [20]. The images were manually masked to allow discrimination based solely on the appearance of the objects without relying on the background.

4.2. Baseline methods and evaluation metrics

As this work is the first attempt at hyperspectral unsupervised spoof detection, we started by comparing our approach to three RGB-based approaches: state-of-the-art anomaly detection [22], multi-class OOD detection [12], and ResNet-OOD: an RGB image classifier with a ResNet-18 [21] backbone that identifies spoofs through generally used OOD detection techniques, i.e., the maximum probability of the softmax [17] and dropout sampling [23].

We considered the baseline methods of hyperspectral spoof detection. Hyper-OOD is an intuitive approach that trains an HSI classifier and uses it as a feature extractor to discriminate between spoofs in the same manner as ResNet-OOD. We implemented this method by incorporating the classifier proposed for vegetable and fruit classification [14]. In addition, the OOD score, defined by the Mahalanobis distance from the input to the intra-class distribution of the training samples in the feature space [10], was applied to the features extracted from the classifier [14] (referred to as MD-HSI). All the baselines and the proposed method trained the models without access to any spoof samples.

We report the true negative rate at 95% true positive rate (TNR #95), the detection accuracy (DA), and the area under the receiver operating characteristic curve (AUC). These are widely used evaluation metrics in OOD detection [10, 11, 24], which summarize the performance of a binary classifier discriminating with a score (e.g., an OOD score or spoof score).

The baselines and our method were evaluated on the two test datasets: the HSSD, including print and replay attacks, and the C2H, including fake attacks [15]. We used all four categories of the objects found in the C2H, which provided pairs of lives and spoofs. The test set of the HSSD includes 30 live HSIs and 15 HSIs for each spoof attacks. In order to match the number of lives and spoofs and to increase the diversity of the test set, the test images were augmented by applying random (but realistic) transformations, i.e., image rotation, vertical and horizontal flips, and random changes to the images' contrast and brightness. The baselines and our approach were evaluated on the augmented HSSD and C2H, both of which include 120 lives and 120 spoofs in the test set.

Table 1. TNR #95, DA, and AUC for spoof detection on the HSSD and C2H, including three types of spoofing, i.e., print, replay, and fake attacks. All values are percentages. The best scores are described in bold.

Input	Method	Print attack (HSSD)			Replay attack (HSSD)			Fake attack (C2H)		
		TNR #95	**DA**	**AUC**	**TNR #95**	**DA**	**AUC**	**TNR #95**	**DA**	**AUC**
RGB	ResNet-OOD [21]	34.2	76.2	81.7	60.0	80.0	81.3	2.50	50.8	32.6
RGB	skipGANomaly [22]	40.9	69.2	72.9	10.8	62.1	62.1	36.7	65.8	57.8
RGB	CSI [12]	19.2	76.7	82.7	16.7	69.6	73.9	9.17	54.6	54.6
HSI	Hyper-OOD [14]	78.3	88.3	91.9	70.0	86.2	89.6	50.0	72.5	71.8
HSI	MD-HSI [14, 10]	70.8	83.7	89.2	25.8	81.7	86.5	4.17	58.7	54.9
HSI	Ours	**85.0**	**91.3**	**97.2**	**99.2**	**99.6**	**99.9**	**67.5**	**83.7**	**81.0**

5. RESULTS

The experimental results obtained using the setup described in Section 4 are shown in Table 1. As seen from this table, the proposed method outperformed the other methods on the two datasets when applied to three types of spoofing, showing that spoof detection training in the proposed approach yields better scores with regard to identifying the spoofs.

From the results on the HSSD whose data have 81 bands of spectral information, the methods trained by the HSI dataset achieved higher performance than the RGB-based methods. The proposed approach achieved high score especially against the replay attack because its light spectra from the screen are significantly different from the lives' spectra. This suggests that the spectral information played an important role in spoof detection. Although the available HSIs in the C2H were compressed to 30 bands, the results of the Hyper-OOD and our method showed a significant improvement over those of the RGB-based methods. The results revealed that the proposed framework performed better than the baselines even with 30 bands of spectral information.

Although Hyper-OOD and MD-HSI have common intermediate features, MD-HSI degrades spoof detection accuracy. This could be because MD-HSI was unable to reconstruct the intra-class distribution from the training samples due to the small number of training samples in each class. Actually, the HSSD has only eight HSIs per class, and C2H has many HSIs, but the same objects are captured from different views. In contrast, the proposed approach takes a pixel as an input. Therefore, it can learn the effective feature representation from a smaller number of images.

To validate the effectiveness of the proposed loss function and the spoof score, we implemented two baselines that had the same architecture as the proposed model. The first was trained without the proposed variance loss term, i.e., only through the cross-entropy loss function ($\mathcal{L}_{\mathrm{CE}}$), and its spoof score was defined by the mean maximum softmax probability [17]. The second was trained by $\mathcal{L}_{\mathrm{CE}}$, and the spoof was discriminated by the proposed intra-sample variance-based score (Eq. (3)). Table 2 exhibits the spoof detection results of these two baselines and the proposed method. The comparison between the second baseline and the proposed method revealed that the model trained to explicitly reduce intra-sample variance achieved high spoof-detection scores.

Table 2. AUC to evaluate the contribution of the proposed loss term and spoof score to spoof detection. PA, RA, and FA represent print, replay, and fake attacks, respectively. The last row denotes the results of the proposed method.

Loss function	Spoof score	PA	RA	FA
$\mathcal{L}_{\mathrm{CE}}$	max prob.	83.8	61.9	49.4
$\mathcal{L}_{\mathrm{CE}}$	variance	63.8	88.8	59.1
$\mathcal{L}_{\mathrm{CE}} + \mathrm{var}(Z)$	variance	97.2	99.9	81.0

6. CONCLUSION

This paper introduces the use of an HSI for spoof detection and presents an unsupervised approach to detect spoofs. It has been demonstrated that spectral information is useful for identifying spoofs and that the proposed loss function, which explicitly reduces intra-sample variance in the feature space, helps discriminant feature representation to be learned for spoof detection. Thus, it is able to effectively distinguish spoofs from lives using the proposed spoof score. In addition, we developed the first hyperspectral spoof detection dataset (HSSD) that contains print and replay attacks. The experiments on the HSSD revealed that the proposed method performed well with the small number of training samples.

7. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 19K20299. We thank Toshiki Kikuchi for helpful suggestions and the students in our lab for their careful annotations.

8. REFERENCES

[1] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *IEEE Int. Conf. on Image Processing*, 2015, pp. 2636–2640.

[2] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise modeling, synthesis and classification for generic object anti-spoofing," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 7292–7301.

[3] Fuan Tsai and William D Philpot, "A derivative-aided hyperspectral image analysis system for land-cover classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 40, no. 2, pp. 416–425, 2002.

[4] Robert T Kester, Liang Gao, and Tomasz S Tkaczyk, "Development of image mappers for hyperspectral biomedical imaging applications," *Applied optics*, vol. 49, no. 10, pp. 1886–1899, 2010.

[5] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu, "Deep tree learning for zero-shot face anti-spoofing," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 4680–4689.

[6] Yunxiao Qin et al., "Learning meta model for zero-and few-shot face anti-spoofing," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2020, vol. 34, pp. 11916–11923.

[7] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*. Springer, 2015.

[8] Junying Gan, Shanlu Li, Yikui Zhai, and Chengyun Liu, "3d convolutional neural network based on face anti-spoofing," in *2nd int. conf. on multimedia and image processing*. IEEE, 2017, pp. 1–5.

[9] Jim Winkens et al., "Contrastive training for improved out-of-distribution detection," *arXiv preprint arXiv:2007.05566*, 2020.

[10] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, 2018, pp. 7167–7177.

[11] Chandramouli Shama Sastry and Sageev Oore, "Detecting out-of-distribution examples with gram matrices," in *Int. Conf. on Machine Learning*, 2020, pp. 8491–8501.

[12] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *Advances in Neural Information Processing Systems*, 2020.

[13] Robert Ennis, Florian Schiller, Matteo Toscani, and Karl R Gegenfurtner, "Hyperspectral database of fruits and vegetables," *JOSA A*, vol. 35, no. 4, pp. B256–B266, 2018.

[14] Jan Steinbrener, Konstantin Posch, and Raimund Leitner, "Hyperspectral fruit and vegetable classification using convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 162, pp. 364–372, 2019.

[15] Longbin Yana, Xiuheng Wang, Min Zhao, Maboud Farzaneh Kaloorazi, Jie Chen, and Susanto Rahardja, "Reconstruction of hyperspectral data from rgb images with prior category information," *IEEE Trans. on Computational Imaging*, 2020.

[16] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[17] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[18] Eric Veach and Leonidas J Guibas, "Optimally combining sampling techniques for monte carlo rendering," in *Proc. of the 22nd annual conf. on Computer graphics and interactive techniques*, 1995, pp. 419–428.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] "HyperSpectralToolbox," `https://github.com/davidkun/HyperSpectralToolbox`.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *Int. Joint Conf. on Neural Networks*. IEEE, 2019, pp. 1–8.

[23] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.