HYPERSPECTRAL AND MULTISPECTRAL IMAGE FUSION USING A MULTI-LEVEL PROPAGATION LEARNING NETWORK

Carlos A. Theran[a,b,d], *Michael A. Alvarez*[a,c,d] *Emmanuel Arzuaga*[a,b,c,d], *Heidy Sierra*[a,b,d*]

Laboratory for Applied Remote Sensing,
Imaging and Photonics (LARSIP)[a]
Department of Computer Science and Engineering[b]

Department of Electrical and Computer Engineering[c]
University of Puerto Rico Mayaguez,
Call Box 9000, Mayaguez, PR, 00681.[d]

ABSTRACT

Data fusion techniques have gained special interest in remote sensing due to the available capabilities to obtain measurements from the same scene using different instruments with varied resolution domains. In particular, multispectral (MS) and hyperspectral (HS) imaging fusion is used to generate high spatial and spectral images (HSEI). Deep learning data fusion models based on Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) have been developed to achieve such task.

In this work, we present a Multi-Level Propagation Learning Network (MLPLN) based on a LSTM model but that can be trained with variable data sizes in order achieve the fusion process. Moreover, the MLPLN provides an intrinsic data augmentation feature that reduces the required number of training samples. The proposed model generates a HSEI by fusing a high-spatial resolution MS image and a low spatial resolution HS image. The performance of the model is studied and compared to existing CNN and LSTM approaches by evaluating the quality of the fused image using the structural similarity metric (SSIM). The results show that an increase in the SSIM is still obtained while reducing of the number of training samples to train the MLPLN model.

Index Terms— Hyperspectral image, Multispectral image, Long Short Term Memory, Data fusion, Deep learning

1. INTRODUCTION

Remote sensing is experiencing rapid growth of different sensor technologies available at different scales fueled in part by new satellite system deployments that contain different optical sensors designs. As a result, these systems are capable of collecting spectral and spatial measurements. Moreover, a sensing system is characterizing by three main resolution domains: spectral, spatial, and temporal. It follows, then the quality of information gathered by a sensor is associated with its resolution domain. Design choices such as application type, size, weight, and costs affect the resolution that each sensor has across the different domains. Trade-offs associated with these choices limit high-quality information to one or two resolution domains in sensor systems. For example, different sensor technologies are capable of providing high spatial and high spectral information, but such designs tend to increase costs significantly.

Different techniques have been proposed in the literature to overcome the absence of several resolution domains on the collected data by combining different sensor modalities, each contributing a high-quality resolution domain [1, 2, 3, 4]. These techniques use the data collected by different sensors with different resolution domains over the same scene. In this manner, we are capable of improving the resolution by merging this data. The process of integrating data from multiple sources is known as data fusion.

Data fusion is becoming the preferred option to improve the data collected by multi-sources. As a result, we can achieve inferences that are not obtaining from a single source. For example, perform a better classification, and description of different terrestrial and atmospheric phenomenons are tasks that use data fusion techniques to improve performance and analysis results [5, 6]. In this manner, due to the high spectral content of HS image, it has gained relevant attention in remote sensing for image enhancement using data fusion techniques [7, 8, 9]. However, typical hyperspectral sensors generate images with a high spectral resolution while sacrificing spatial resolution. Then, to overcome the lack of spatial resolution from the HS image, it is fuse with MS image due to its high spatial resolution. This technique has gained relevant attention to generate images with high-spatial and high-spectral resolution (HSaHS) [10].

In literature, Deep Learning (DL) algorithms have gained relevant attention for its capacity to solve problems that involve data with high dimension, extract more useful information, and auto-feed or adjust their parameters during the training process [11]. Particularly, DL algorithms for single image super-resolution have been proposed [12], where a trained convolutional neural network (CNN) model is used as an end-to-end mapping between low and high-resolution images. The model takes the low-resolution image as the input

*heidy.sierra1@upr.edu

and outputs a high-resolution image. Also, Palsson proposed to train a convolutional neural network model (3D-CNN) for learning filters used to fuse the MS and HS images [8].

In this work, we propose a new model called Multi-Level Propagation Learning Network (MLPLN) that fuses HS and MS images. The architecture of this model allows the use of different sized patch as input for training, which provide an intrinsic data augmentation. It also minimizes the loss of spatial properties of the patch in the fusion process. These patch are defined by two parameters, the scalability number η and padding factor κ. In this manner, the MLPLN learns from the smallest patch and propagates the learned information through the network until it reaches the biggest patch. The performance of our model is compared with recent state of the art networks. Moreover, a thorough study of the spatial-parameters, size of patch $\zeta_{p\times p}$, scalability number η, and padding factor κ is presented in this work.

2. MULTI-LEVEL PROPAGATION LEARNING NETWORK

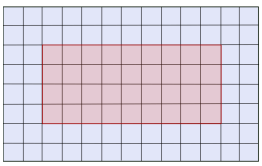
The potential of the proposed model came from the propagation learning technique provided by LSTM, combined with a fully connected layer, contributes to a new optimal architecture for super-resolution images. The idea behind a LSTM network is the inclusion of a self-loop that avoids the exploiting gradient and vanishes gradient problems. As a result, this network has the capability to retain information for long periods of time [**?**, 11]. Also, new concepts have been adopted for our proposed approach. Thus, in order to introduce the Multi-Level Propagation Learning Network (MLPLN), let us define a sequence of images Λ, padding factor, and scalability number, as follows.

Let A_0 be an image of spatial size $p \times l$, where $p, l \in \mathbf{R}$. A sequence of images Λ is defined as the set of images $A_j \subseteq A_0$, $\forall j \in \mathbf{N}$, such that the following properties are satisfied:

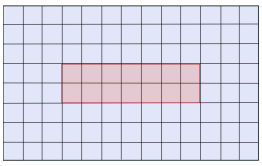
- The size of A_j is $p - 2j \times l - 2j$.
- A_j is concentric with A_0.

Each set Λ_i, for $1 \leq i \leq m$ is defined by the following parameters:

- **Padding factor**: This parameter allows cropping an image of size $p \times l$ by a factor of $\kappa \in \mathbf{N}$, the cropped image losses $2(\kappa \times p) + 2(\kappa \times l) - \kappa^2 \times 4$ pixels. Where its minimum spatial size must be 2×2, and the upper boundary for κ is defined as $\kappa_{max} \leq \dfrac{\min\{p, l\}}{2} - 1$. Figure 1 exemplifies the effect of this parameter.
- **Scalability number**: This parameter sets the number of cropped images $\eta \in \mathbf{N}$ ($|\Lambda_i| = \eta$), taken from the given image A_0. Due to that the minimum spatial size must be 2×2, the upper boundary is defined as, $\eta_{max} \leq \left\lceil \log_2 \left\{ \dfrac{\min\{p, l\} - 2}{\kappa} \right\} \right\rceil$. As a result, the parameter κ is inversely proportional to the cardinality of Λ_i.



(a) $\kappa = 2$



(b) $\kappa = 3$

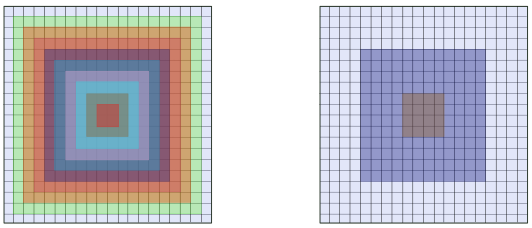
Fig. 1: Two samples of padding parameter for a matrix of size $p \times l = 8 \times 13$.

In Figure 2, we provide an example of the parameter scalability number. As mentioned before, this parameter contains the information about how deep we move into the center of the images. Also, this parameter defines the length of the sequences $[\Lambda_1, \Lambda_2, \ldots, \Lambda_m]$, where $m \in \mathbf{N}$. And each Λ_i for $1 \leq i \leq m$ is a patch with different size concentric to the biggest patch A_0.

The new proposed model to fuse MS and HS images is configured using an input layer, a LSTM layer, a fully connected layer, and a regression layer as an output layer.

The input layer receives sets Λ of patch (sequence of images), with $\Lambda_i \subset \Lambda$ for $i = 1, \ldots, \eta$. Each Λ_i is of a different size, and the spatial relationship between each patch in the set is defined by the scalability number and padding factor parameters.

Our approach can be seen as sequences of images, where each sequence has its own long-term spatial dependencies. For this reason, our model uses a LSTM layer, due to its capability to allow gradient flow for a long duration. As well as, its ability to avoid the vanishing or exploding gradient [**?**]. A fully connected layer is used to provide mapping features to a more separable space. This layer sends the output of the LSTM layer to a space that is more discriminative [13]. As a result, the MLPLN learns the weights to predict the target data linearly. In the last layer of the proposed model, a linear regression technique is used, which is the adequate method for prediction problems [14]. This layer measures how well the model fits the training data, by using a minimization problem (RMSE) to find the values of the weights generated by the fully connected layer.



(a) $\eta = 9, \kappa = 1$ (c) $\eta = 2, \kappa = 4$

(b) $\eta = 4, \kappa = 2$ (d) $\eta = 1, \kappa = 6$

Fig. 2: Samples of scalability parameter for a matrix of size 20×20.

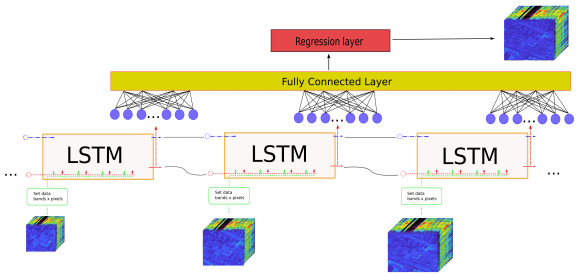
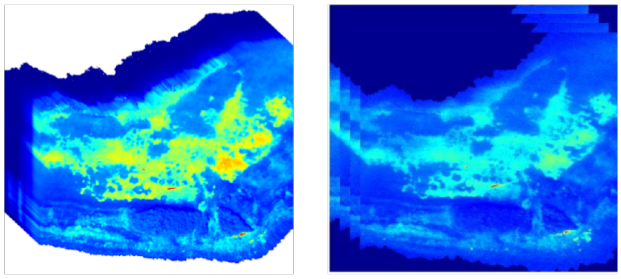


Fig. 3: MLPNL Model



(a) (b)

Fig. 4: Figure (a) is the HSI image of Enrique and figure (b) the MS image.

3. EXPERIMENTAL DATA

The experimental data used in this work are 2 different hyperspectral images: Indian Pines, and Enrique Reef. These images are described as follows: The Indian Pines hyperspectral image was gathered by the AVIRIS sensor and consists of 145×145 pixels and 224 spectral bands in the wavelength range 400 to 2500 nm. The number of bands were reduced to 200 by removing high water absorption bands. This scene has 16. The Enrique Reef hyperspectral image consists of 128 bands and it was captured using AISA Eagle sensor, this image was acquired in 2007. The spatial resolution of this data is 1m. There are 6 classes: Mangrove, Deep water, Coral, Sand, Sea grass, and Flat reef.

All of these images are used to generate simulated data following the procedure presented in [8, 9], a low spatial resolution hyperspectral image is simulated by applying image decimation by a factor of 4 to the original dataset. Likewise, a high spatial resolution multispectral image is simulated from the original data set by averaging bands from different spectral ranges: 1) Blue: 445-516nm, 2) Green: 506-595nm, 3) Red: 632-698nm and 4) NIR = 757-853nm.

The performances of the proposed network is tested using the following metric; the structural similarity index (SSIM)[15], power signal noise ratio (PSNR)[16], and relative dimensionless global error (ERGAS) [17].

4. EXPERIMENTAL RESULTS

In our approach, we reduced the spectral dimensionality of the hyperspectral images. We also removed redundant spectral bands and kept only the bands that contained the majority of information of the scene. For this purpose, we used SVD as a technique that allows separating the spatial from the spectral information. In this manner, given a hyperspectral image $A \in \mathbf{R}^{p \times b}$ where b is the number of bands and p is the number of pixels, the SVD applied to A gives the following factorization: $A = USV^T$, where $U \in \mathbf{R}^{p \times p}$ is an orthogonal matrix, whose columns are eigenvectors of pixels, $S \in \mathbf{R}^{p \times b}$ is a diagonal matrix of which elements are the eigenvalues that represent the energy of each pixel by bands, and $V \in \mathbf{R}^{b \times b}$ contains the spectral information. In this decomposition the matrix $\Gamma = US$ represents the spatial information. Figure 3 provides the description of the proposed model used for the experimental results.

On the other hands, the configuration of our model is as follow; the first layer (Input layer) receives sets of images (patch) with different size, which are defined by the parameters padding factor and scalability number. For experimental propose we are defined four different initial patch size (20×20), (16×16), (12×12) and (8×8). The images on a set are transformed as bands per pixel. As a result, the feature

of our network is the number of spectral bands. These patch are the input of the second layer LSTM, where the number of hidden units is the number of features. This layer is defined as sequences to sequences, where the output of the LSTM layer is received by a fully connected layer, and finally, a regression layer is used. To provide information about the performance of MLPLN with a different number of samples for training, a 40% and 80% of the patch are selected randomly from Γ_r for each data set describe in table 1.

Experimental results for 80% and 40% of training data for Indian Pines and Enrique Reef datasets are presented in table 2 and table 3, respectively. The highlighted results in the table 2 and table 3 show that parameters patch size, κ and η are consistent to maximize the performance model.

Percent	Patch size	Indian Pines samples	Enrique Reef samples
80%	20×20	115	1130
	16×16	179	1772
	12×12	272	3021
	8×8	562	6750
40%	20×20	29	332
	16×16	45	519
	12×12	84	933
	8×8	169	2114

Table 1: Representation of patch number used for training per percent, patch size, and data set.

Parameters					Metrics							
					SSIM		RMSE		PSNR		ERGAS	
%	Patch	κ	η	Band	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
80	20	1	4	20	0.902	3.6E-03	0.025	3.5E-04	29.87	0.10	6.14	1.1E-01
	16	**1**	**4**	**20**	**0.907**	**1.1E-03**	**0.024**	**8.8E-05**	**29.99**	**0.06**	**6.10**	**8.8E-02**
	12	1	4	20	0.906	1.1E-03	0.024	1.7E-04	29.98	0.06	6.13	2.7E-02
	8	1	3	6	0.906	1.5E-03	0.024	1.5E-04	29.97	0.04	6.05	7.1E-02
40	20	1	4	20	0.792	1.0E-02	0.042	4.8E-03	24.99	0.55	8.00	5.7E-01
	16	1	4	20	0.854	1.5E-02	0.030	8.3E-04	28.07	0.41	6.60	1.3E-01
	12	1	4	20	0.895	2.0E-03	0.025	2.5E-04	29.64	0.17	6.20	1.8E-01
	8	**1**	**3**	**20**	**0.905**	**4.6E-03**	**0.024**	**1.3E-04**	**29.94**	**0.06**	**6.10**	**5.9E-02**

Table 2: Summary of MLPLN using Indian Pines dataset. The 40% and 80% training samples were used. The presented values were the best performance of the proposed model combining patch size, scalability, and padding factor.

Parameters					Metrics							
					SSIM		RMSE		PSNR		ERGAS	
%	Patch	κ	η	Band	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
80	20	1	4	20	0.986	1.8E-04	0.010	9.0E-05	39.86	0.05	2.98	2.9E-02
	16	1	4	20	0.987	1.6E-04	0.010	4.0E-05	39.88	0.03	2.97	7.7E-03
	12	**1**	**4**	**15**	**0.987**	**2.4E-05**	**0.010**	**1.6E-05**	**39.91**	**0.02**	**2.96**	**6.5E-03**
	8	1	3	20	0.987	9.5E-05	0.010	3.8E-05	39.92	0.06	2.95	8.2E-03
40	20	1	4	20	0.973	1.1E-03	0.013	2.4E-04	37.74	0.06	3.95	8.4E-02
	16	1	4	15	0.975	1.3E-03	0.012	2.0E-04	38.04	0.14	3.77	7.7E-02
	12	1	4	20	0.982	3.6E-04	0.011	8.3E-05	38.93	0.11	3.38	4.2E-02
	8	**1**	**3**	**15**	**0.984**	**5.6E-04**	**0.010**	**1.3E-04**	**39.43**	**0.19**	**3.16**	**5.3E-02**

Table 3: Summary of MLPLN using the Enrique Reef dataset. The 40% and 80% training samples were used. The presented values were the best performance of the proposed model combining patch size, scalability, and padding factor.

In addition, we compare the MLPLN with two methods in state of the art. Figure 5 shows the comparison for the Enrique Reef dataset, in this case, the MLPLN is better than the others technique for 80% of training samples, and it is also better when the bands selected are greater or equal to 6 taken only 40% of training samples.

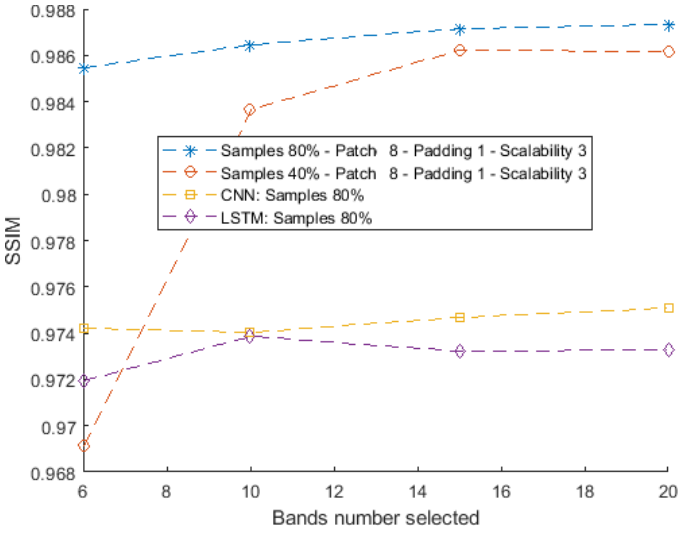


Fig. 5: Comparison best results of MLPLN with CNN and LSTM networks, for Enrique Reef Dataset

5. CONCLUSIONS AND FUTURE WORK

In this article we present a new approach to fuse HS and MS images. As a result, images with high spatial resolution and a high spectral resolution were obtained. The Scalability κ and Padding η parameters were introduced, that allow extracting and learning features from the images, and to propagate the learned feature through the network using the MLPLN proposed. With this new approach, filtering the images is not necessary, and thus the loss of spatial content is avoided. Also, the numerical results show that our model performs better in term of SIIM, PSRN, ERGAS, and RMSE. Consequently, we can deliver the following analysis: SSIM value increase w.r.t. Scalability for any patch size and padding, in all datasets, and, the smallest RMSE value and the best PSNR were obtained with a patch size of 8, $\kappa = 1$ and $\eta = 3$, comparable in the cases with 80% and 40% of training samples.

With the previous numerical results we have shown the performance of our proposed model. Which have proved the flexibility of obtained good result using 40% of data training combined with the news parameter κ and η.

The previous analysis shows that the performance of our proposed model was superior to other available deep learning models. It improves the construction of the hyperspectral cube, conserves the image's hyperspectral (LSHS) information, and improves the spatial resolution. This method will

allow us to obtain classification results with a higher accuracy with will help to classify objects within the LSHS.

6. REFERENCES

[1] R. Zurita-Milla, J. G. P. W. Clevers, and M. E. Schaepman, "Unmixing-based landsat tm and meris fr data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 453–457, July 2008.

[2] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1212–1226, May 1999.

[3] Thomas Hilker, Michael A. Wulder, Nicholas C. Coops, Julia Linke, Greg McDermid, Jeffrey G. Masek, Feng Gao, and Joanne C. White, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and modis," *Remote Sensing of Environment*, vol. 113, no. 8, pp. 1613 – 1627, 2009.

[4] Quanlong Feng, Dehai Zhu, Jianyu Yang, and Baoguo Li, "Multisource hyperspectral and lidar data fusion for urban land-use mapping based on a modified two-branch convolutional neural network," *ISPRS International Journal of Geo-Information*, vol. 8, pp. 28, 01 2019.

[5] Y. Qu, H. Qi, B. Ayhan, C. Kwan, and R. Kidd, "Does multispectral / hyperspectral pansharpening improve the performance of anomaly detection?," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2017, pp. 6130–6133.

[6] Michael Alvarez, Carlos A Theran, Emmanuel Arzuaga, and Heidy Sierra, "Analyzing the effects of pixel-scale data fusion in hyperspectral image classification performance (conference presentation)," in *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imagery XXVI*. International Society for Optics and Photonics, 2020, vol. 11392, p. 1139205.

[7] O. Eches, N. Dobigeon, and J. Tourneret, "Enhancing hyperspectral image unmixing with spatial correlations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4239–4247, Nov 2011.

[8] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, May 2017.

[9] Carlos A Theran, Michael A Álvarez, Emmanuel Arzuaga, and Heidy Sierra, "A pixel level scaled fusion model to provide high spatial-spectral resolution for satellite images using lstm networks," in *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2019, pp. 1–5.

[10] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, March 2019.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4580–4584.

[14] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal, *Locally Weighted Learning*, pp. 11–73, Springer Netherlands, Dordrecht, 1997.

[15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] M Fallah and A Azizi, "Quality assessment of image fusion techniques for multisensor high resolution satellite images (case study: Irs-p5 and irs-p6 satellite images)," vol. 38, 01 2010.

[17] P. Jagalingam and Arkal Vittal Hegde, "A review of quality metrics for fused image," *Aquatic Procedia*, vol. 4, pp. 133 – 142, 2015.