A NEW SPATIO-TEMPORAL FUSION METHOD FOR REMOTELY SENSED DATA BASED ON CONVOLUTIONAL NEURAL NETWORKS

Yunfei Li[1]*, Chenying Liu*[1]*, Lin Yan*[1]*, Jun Li*[1]*, Antonio Plaza*[2] *and Bo Li*[3]

[1]Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation,
School of Geography and Planning, Sun Yat-sen University, Guangzhou, 510275, China
[2]Hyperspectral Computing Laboratory, Avenida de la Universidad s/n, E-10003 Caceres, Spain
[3]School of Computer Science and Engineering, Beihang University, Beijing 100191, China

ABSTRACT

In some remote sensing applications such as change detection, satellite images with both high spatial and high temporal resolution are required. However, no single satellite sensor can currently provide such images due to technical specifications. To solve this problem, spatio-temporal fusion provides a cost-effective solution. In this paper, we propose a new spatio-temporal fusion approach, based on convolutional neural networks (CNNs), for Landsat and MODIS image fusion. Specifically, the proposed approach utilizes CNNs to model the heterogeneity of fine pixels from the coarse MODIS images. Here, the heterogeneity of fine pixels is defined as the difference between the reflectance changes obtained from the two types of images. After that, two transition-predicted images can be obtained using the trained CNNs, which are then fused in order to obtain a final prediction. In our newly proposed approach, CNNs are only used to learn the heterogeneity of fine pixels rather than the whole images, thus providing a more stable and less time-consuming strategy as compared to other available approaches. We evaluated the proposed approach on a public spatio-temporal fusion dataset and the obtained results suggest that our newly developed method achieves state-of-the-art performance.

Index Terms— Spatio-temporal fusion, convolutional neural networks (CNNs), heterogeneity.

1. INTRODUCTION

Temporally dense remote sensing images are necessary and important for change detection applications, such as the characterization of crop yields [1], vegetation monitoring [2] and the detailed investigation of human-nature interactions [3], where the changes need to be accounted for at a very fine scale in heterogeneous regions. In this context, such temporally dense images should also exhibit high spatial resolution. However, no available satellite instrument can currently provide such images due to technical and budget limitations. In other words, the images with high temporal resolution usually exhibit low spatial resolution, while the ones with high spatial resolution are often sparse in frequency. Hereinafter, we refer to such images as *coarse* and *fine*, respectively, in terms of their spatial resolution. Accordingly, we also refer to their pixels as coarse and fine. To tackle the aforementioned problems, spatio-temporal fusion provides a feasible and effective strategy that can generate images with high spatial and high temporal resolution by combining the two types of images above.

A key requirement of spatio-temporal fusion is to be able to model the reflectance changes of land surface, including two aspects, *i.e.*, the phenology changes (e.g., seasonal changes of crops) and land cover type changes during a certain temporal period [4]. Up to now, many spatio-temporal fusion methods have been adopted, which can be categorized into three main groups [5]: weighted function-based ones [6, 7] , unmixing-based ones [5, 8], and learning-based ones [4, 9, 10, 11]. The former two are generally based on the linear mixture model, in which a pixel in the coarse image can be regarded as a linear combination of some corresponding pixels in the fine image. Specifically, weighted function-based methods model the pixel values in the predicted fine image using the weighted sum of spectrally similar pixels. However, such methods implicitly assume that no land cover type change happens during the prediction period [5], thus performing well only for phenology changes. On the other hand, unmixing-based methods utilize the linear unmixing model to extract the temporal reflectance changes from coarse images. Operating under a similar assumption as weighted function-based ones, these approaches are also unable to precisely predict the land cover type change. On the contrary, learning-based methods (most of which are sparse representation-based ones) establish a relationship between the coarse and fine image pairs on the basis of their structural similarity, leading to a good ability to deal with land cover type changes [4, 9]. However, the image features are manually designed, making them complex and unstable. To solve this problem, a recent trend is to use CNNs rather than the sparse representation for learning purposes, due to the fact that CNNs are capable to generate the features in a data-driven context. Song et al. [11] proposed a spatio-temporal Satellite Image Fusion Using Deep Convolutional Neural Networks (STIFCNN) method which reconstructs the fine images from the coarse ones by a nonlinear mapping-based CNN and a super resolution-based CNN, and then fuses the reconstructed images with the observed ones to get the final predictions. STIFCNN is more effective than sparse representation- based approaches when extracting features from large-scale images. However, the combination of two CNNs and a fusion model makes the method complex and time-consuming. Furthermore, performing the reconstruction directly from coarse images brings instability due to the large existing gap in spatial resolution between Landsat and MODIS images.

In this paper, we develop a new CNN-based spatio-temporal fusion approach to fuse fine Landsat images with coarse MODIS images. Our newly proposed approach utilizes CNNs to model the heterogeneity of fine pixels from the coarse pixels. Specifically, we define the heterogeneity of fine pixels as the reflectance change dif-

This work was supported by National Natural Science Foundation of China under Grant 61771496, Guangdong Provincial Natural Science Foundation under Grant 2016A030313254, National Key Research and Development Program of China under Grant 2017YFB0502900.

ference on each fine pixel in comparison to the coarse one. After that, two transition-predicted images can be obtained using well-trained CNNs. Then, a fusion model is implemented on the two transition images to obtain the final prediction. In our newly proposed method, the CNNs only learn the heterogeneity of fine pixels rather than the whole images, thus making it more stable in comparison to STIFCNN. Furthermore, the proposed method is simpler and less time-consuming since only one type of CNN needs to be trained. The rest of the paper is organized as follows. In section 2, the proposed method is introduced in detail. The obtained results are presented and discussed in section 3. Finally, our conclusions and some future lines are given in section 4.

2. METHODOLOGY

In the following, we denote the dates of the obtained images as t_i, and the corresponding Landsat and MODIS images as L_i and M_i, respectively, for $i = 1, 2, 3$. A general flowchart of our method is given in Fig. 1. Specifically, we utilize the image pairs at t_1 and t_3 to train the CNN and predict L_2 from L_1, L_3, M_1, M_2, and M_3 with the trained CNN and a fusion model.

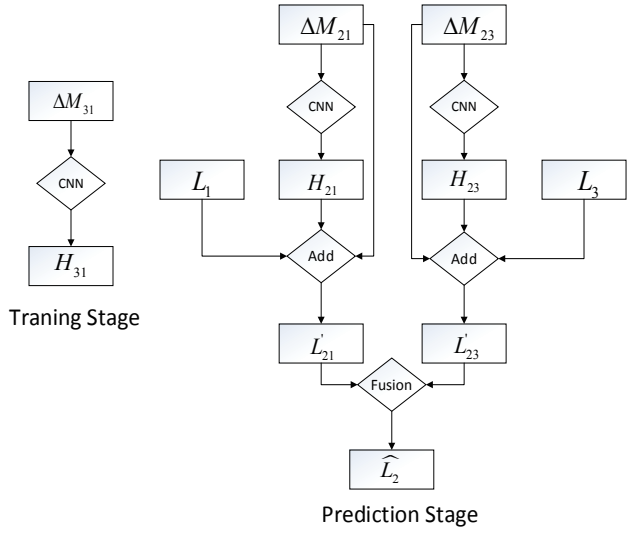

Fig. 1. General flowchart of the proposed method, in which rectangles and diamonds represent images and operations, respectively.

Let x, y and b be the coordinates along x-, y- and the spectral domain, respectively; let ΔM and ΔL be the coarse and fine reflectance changes, *i.e.*, those between the coarse and the fine images, respectively. Given a fine pixel (x, y, b), the reflectance change between dates t_i and t_j can be formulated as:

$$\Delta L_{ij}(x, y, b) = L_i(x, y, b) - L_j(x, y, b). \tag{1}$$

The corresponding coarse reflectance change is given by:

$$\Delta M_{ij}(x, y, b) = R(M_i(x, y, b) - M_j(x, y, b)), \tag{2}$$

where $R(\cdot)$ is the resampling operation of ΔM in accordance to ΔL. Then, we define the heterogeneity of the fine pixel as:

$$H_{ij}(x, y, b) = \Delta L_{ij}(x, y, b) - \Delta M_{ij}(x, y, b). \tag{3}$$

2.1. Training stage

In the training stage, we feed the CNN with the coarse reflectance change between the dates t_1 and t_3 to obtain the corresponding heterogeneity of fine pixels for each band, that is:

$$\widehat{H}_{13}(x, y, b) = G(\Delta M_{13}(x, y, b); \Theta), \tag{4}$$

where $G(\cdot)$ is a nonlinear mapping function, Θ is a set of parameters, and $\widehat{H}_{13}$ is the predicted heterogeneity of fine pixels.

As shown in Fig. 1, the used CNN for each band is composed of three convolutional layers in total, where the first two layers are followed by a batch normalization layer and a rectified linear unit (ReLU). The loss function is chosen as the mean squared error (MSE) function:

$$\ell(\Theta) = \frac{1}{N} \sum_{x,y} \|\widehat{H}_{13} - H_{13}\|^2, \tag{5}$$

where N is the number of training samples.

2.2. Prediction stage

As illustrated in Fig. 1, the prediction stage mainly contains three parts. First, we use the CNN to model $\widehat{H}_{21}$ and $\widehat{H}_{23}$, *i.e.*, the heterogeneities of fine pixels on t_2 in comparison to M_1 and M_3, from ΔM_{21} and ΔM_{23}, *i.e.*, the corresponding coarse reflectance changes, respectively. After that, we can obtain two transitional predictions as follows:

$$L'_{21}(x, y, b) = \Delta M_{21}(x, y, b) + \widehat{H}_{21}(x, y, b) + L_1(x, y, b), \tag{6}$$

$$L'_{23}(x, y, b) = \Delta M_{23}(x, y, b) + \widehat{H}_{23}(x, y, b) + L_3(x, y, b). \tag{7}$$

As shown in Eqs. (4) and (5), the CNN is trained only using M_1, M_3, H_1 and H_3, but still misses the information contained in M_2. Therefore, L'_{21} and L'_{31} are likely not precise enough, mainly due to the deviation existing in $\widehat{H}_{21}$ and $\widehat{H}_{21}$. To optimize the prediction, a fusion model F is finally implemented as:

$$\widehat{L}_2 = F\left\{L'_{23}, L'_{21}\right\}. \tag{8}$$

Specifically, F is a weighted model, where the weights of two transitional predicted images in each band are determined by the overall spectral similarity of M_2 to M_1 and M_3, respectively, using the MSE function, that is:

$$w_{23}^b = \frac{MSE_{21}^b}{MSE_{21}^b + MSE_{23}^b}, \tag{9}$$

and

$$w_{21}^b = 1 - w_{23}^b, \tag{10}$$

where

$$MSE_{2j}^b = \frac{\sum_{x,y}(M_2(x, y, b) - M_j(x, y, b))^2}{N}, \tag{11}$$

for $j = \{1, 3\}$. Using (9) and (10), the final prediction is formulated as:

$$\widehat{L}_2(x, y, b) = w_{21}^b L'_{21}(x, y, b) + w_{23}^b L'_{23}(x, y, b). \tag{12}$$

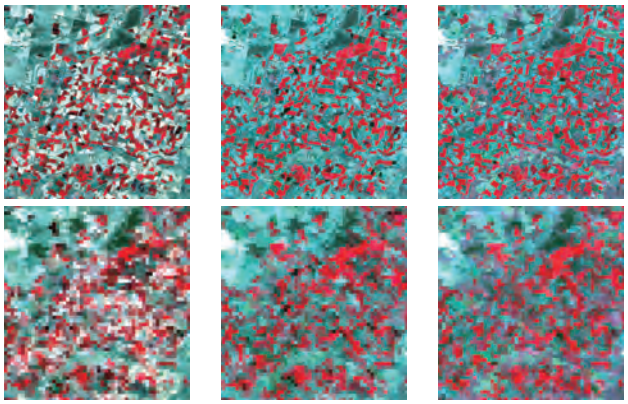

Fig. 2. Satellite image pairs used in our experiments, where the upper row displays the Landsat images and the bottom row displays the corresponding MODIS images.

Fig. 3. Original (a) and predicted Landsat images obtained by STIFCNN (b) and the proposed method (c).

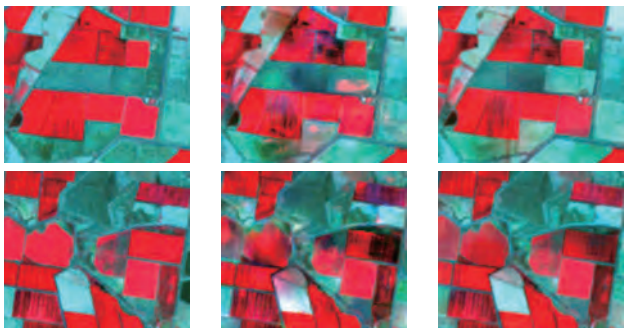

Fig. 4. Zoomed details from the results in Fig. 3, where the first and the second row show the results obtained for the subscenes marked by the black and yellow rectangles in Fig. 3(a), respectively.

3. EXPERIMENTAL RESULTS

3.1. Study area and data

The dataset used for validation purposes is obtained from [12], which is widely used for spatio-temporal fusion purposes. The study area, the Coleambally irrigation district located in southern New South Wales, Australia, is typically heterogenous. There are 17 cloud-free Landsat-MODIS pairs available in 2001-2002. All the images were atmospherically and geometrically corrected, resampled to a spatial resolution of 25m, and cropped to a size of 1720×2040 pixels [12]. The bands of the MODIS images have been rearranged to match those of the Landsat images. As shown in Fig. 3, we selected three image pairs collected on January 12, 2002, February 13, 2002, and February 22, 2002, corresponding to t_1-t_3, respectively, and cropped them to a size of 1000×1000 pixels. Furthermore, we also implemented another CNN-based method, STIFCNN, for comparison purposes. STIFCNN was trained using the image pairs of October 8, 2001, December 4, 2001, and April 11, 2002, following [11].

3.2. Evaluation metrics

We used five evaluation metrics to quantitatively compare our proposed method with STIFCNN. The first one is the root mean square error (RMSE), which gauges the reflectance difference between the predicted image $\widehat{L}$ and the real image L:

$$RMSE = \sqrt{\frac{\sum_{x,y}\sum_b (L(x,y,b) - \widehat{L}(x,y,b))^2}{N}}. \quad (13)$$

The second is the correlation coefficient (r), which shows the linear relationship between the predicted and real reflectance:

$$r = \frac{\sigma_{L\widehat{L}}}{\sqrt{\sigma_L \sigma_{\widehat{L}}}}, \quad (14)$$

where $\sigma_{L\widehat{L}}$ is the covariance of real and predicted images, while σ_L and $\sigma_{\widehat{L}}$ are their variances.

The third is the structure similarity (SSIM), which evaluates the similarity of the overall structure between the predicted and real images:

$$SSIM = \frac{(2\mu_L\mu_{\widehat{L}} + C_1)(2\sigma_{L\widehat{L}} + C_2)}{(\mu_L^2 + \mu_{\widehat{L}}^2 + C_1)(\sigma_L + \sigma_{\widehat{L}} + C_2)}, \quad (15)$$

where μ_L and $\mu_{\widehat{L}}$ are the average values of the real and predicted images, respectively, and C_1 and C_2 are small constants included in order to avoid SSIM being zero. In our experiments, both C_1 and C_2 are set to 0.001.

The forth one is the spectral angle mapper (SAM), which measures the spectral distortion of the predicted image:

$$SAM = \frac{1}{N}\sum_{x,y} arccos \frac{\sum_b L(x,y,b)\widehat{L}(x,y,b)}{\sqrt{\sum_b L^2(x,y,b)\sum_b \widehat{L}^2(x,y,b)}}. \quad (16)$$

The last metric is the erreur relative global adimensionnelle de synthese (ERGAS), which indicates the overall spectral similarity of two images:

$$ERGAS = 100\frac{f}{c}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\frac{RMSE_b^2}{\mu_{L_b}^2}}, \quad (17)$$

where f is the resolution of the fine image, c is the resolution of the coarse image, and μ_{L_b} and $RMSE_b$ are the average value and RMSE of the bth band of the real image, respectively.

As it can be observed from the equations above, a smaller RMSE, SAM and ERGAS (and a bigger r and SSIM) indicate better performance of the evaluated methods.

3.3. Results and analysis

Fig. 3 shows the original (a) and predicted images obtained by STIFCNN (b) and the proposed method (c), respectively, for the Landsat image collected on February 13, 2002. For further comparison, we extracted two subscenes from this scene and zoomed the obtained results for these subscenes in Fig. 4, where the first and second row respectively correspond to the subscenes marked by a black and yellow rectangle in Fig. 3(a). As it can be observed, the proposed method exhibits a good ability to capture the surface feature changes. Furthermore, it can generate a smoother image than the one generated by the STIFCNN method, in which many rough patches can be appreciated. This was expected since, in our proposed method, CNNs are designed to carry out a "carving" process and only need to learn the heterogeneity of fine pixels, while STIFCNN directly predicts the whole image from the coarse one (with low spatial resolution). The considered fusion methods are statistically evaluated in Table 1. As it can be seen, the proposed method achieves better results in terms of all the considered quality metrics, suggesting that it can generate higher-quality predictions from both the spectral and structural viewpoints.

Table 1. Quantitative assessment of the considered fusion methods.

	STIFCNN			Proposed		
	RMSE	r	SSIM	RMSE	r	SSIM
Band 1	0.0114	0.9077	0.9462	**0.0103**	**0.9207**	**0.9541**
Band 2	0.0120	0.8895	0.9374	**0.0109**	**0.9083**	**0.9480**
Band 3	0.0118	0.9066	0.9442	**0.0107**	**0.9220**	**0.9536**
Band 4	0.0153	0.8388	0.9038	**0.0128**	**0.8838**	**0.9297**
Band 5	0.0121	0.9061	0.9427	**0.0116**	**0.9205**	**0.9495**
Band 6	0.0104	0.9249	0.9557	**0.0100**	**0.9356**	**0.9605**
ERGAS		1.2594			**1.1132**	
SAM		0.1358			**0.1239**	

4. CONCLUSION

In this paper, we propose a new CNN-based spatio-temporal fusion approach for Landsat and MODIS image fusion. Our newly proposed approach utilizes CNNs to learn the heterogeneity of fine pixels from coarse images for prediction purposes. Here, we define the heterogeneity of fine pixels as the difference between the reflectance changes obtained from two types of images, *i.e.*, the fine ones and the coarse ones. Resulting from this process, our method obtains two transitional predicted images and then fuses them to generate the final prediction. Our experimental results, conducted on a standardized spatio-temporal fusion dataset, demonstrate that our newly developed approach achieves state-of-the-art performance. In the future, we will test our proposed approach on additional datasets.

5. REFERENCES

References

[1] Michael D. Johnson, William W. Hsieh, Alex J. Cannon, Andrew Davidson, and Frédéric Bédard, "Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods," *Agricultural and Forest Meteorology*, vol. 218-219, pp. 74–84, 2016.

[2] Miaogen Shen, Yanhong Tang, Jin Chen, Xiaolin Zhu, and Yinghua Zheng, "Influences of temperature and precipitation before the growing season on spring phenology in grasslands of the central and eastern qinghai-tibetan plateau," *Agricultural and Forest Meteorology*, vol. 151, no. 12, pp. 1711–1722, 2011.

[3] Li Xuecao, Yuyu Zhou, G Asrar, Jiafu Mao, Xiaoma Li, and Wenyu Li, "Response of vegetation phenology to urbanization in the conterminous united states," *Glob Chang Biol*, vol. 23, no. 7, pp. 2818–2830, 2017.

[4] Huang Bo, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3707–3716, 2012.

[5] Xiaolin Zhu, Eileen H. Helmer, Feng Gao, Desheng Liu, Jin Chen, and Michael A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sensing of Environment*, vol. 172, pp. 165–177, 2016.

[6] Feng Gao, Jeffrey G. Masek, Mathew R. Schwaller, and Forrest Hall, "On the blending of the landsat and modis surface reflectance: predicting daily landsat surface reflectance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[7] Wang Jing and Huang Bo, "A rigorously-weighted spatiotemporal fusion model with uncertainty analysis," *Remote Sensing*, vol. 9, no. 10, pp. 990, 2017.

[8] Mingquan Wu, Wenjiang Huang, Zheng Niu, and Changyao Wang, "Generating daily synthetic landsat imagery by combining landsat and modis data," *Sensors*, vol. 15, no. 9, pp. 24002–24025, 2015.

[9] Huihui Song and Bo Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 1883–1896, 2013.

[10] Jiang Cheng, Hongyan Zhang, Huanfeng Shen, and Liangpei Zhang, "Two-step sparse coding for the pan-sharpening of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 5, pp. 1792–1805, 2014.

[11] Huihui Song, Qingshan Liu, Guojie Wang, Renlong Hang, and Huang Bo, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018.

[12] Irina V. Emelyanova, Tim R. Mcvicar, Thomas G. Van Niel, Tao Li Ling, and Albert I. J. M. Van Dijk, "Assessing the accuracy of blending landsat-modis surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sensing of Environment*, vol. 133, no. 12, pp. 193–209, 2013.