Development of A Registration Technique of Time-series Satellite Image Based on Improved Geometric-matching CNN

Futa Morishima
Department of Mechanical and Control Engineering
Kyushu Institute of Technology
Kitakyushu City, Japan

Tohru Kamiya
Department of Mechanical and Control Engineering
Kyushu Institute of Technology
Kitakyushu City, Japan
Email: kamiya@cntl.kyutech.ac.jp

***Abstract*— One major application of remote sensing data is detecting environmental changes. Image registration, which aligns the coordinates of multiple satellite images acquired over time by different sensors, is an important preprocessing step in this detection process. Conventional registration methods use feature matching to identify features designed by humans, such as Scale-Invariant Feature Transform (SIFT). However, deep learning–based methods for image registration are now being actively studied. While the latest satellite images have a resolution of less than 1 m, processing such high-resolution images can be computationally expensive. Recently, operational systems with Artificial Intelligence (AI) inside satellites have attracted interest. However, the lack of computational resources for onboard computers can also be an issue. In this paper, we propose a lightweight, accurate image registration model with a small number of parameters. Our model is based on geometric matching Convolutional Neural Networks (CNNs) and attempts to reduce the number of parameters by incorporating MobileNetV3, lightweight estimation (LE), and Convolutional Mixers (ConvMixer) to improve registration accuracy. Experimental results on time-series satellite images demonstrate that our method achieves a Grid Mean Squared Error (MSE) of 0.00614, representing the error in image registration. The number of parameters is small. Therefore, the proposed method achieves higher registration accuracy with fewer parameters than the base model.**

Keywords; ConvMixer; Convolutional Neural Network; Geometric-matching CNN; Image Registration, Light-weight Estimation; MobileNetV3; Remote Sensing.

I. INTRODUCTION

In recent decades, the rapid development of satellite and sensor technology has significantly increased the amount of data acquired by remote sensing, resulting in big data [1]. Earth observation satellites are one of the main sources of this data. Figure 1 shows the estimated annual volume of data acquired by three types of Earth observation satellites: Landsat, MODIS, and Sentinel [2]. The volume of data was approximately 0.25 petabytes (PB) in 2013 and exceeded 4.25 PB in 2019. Remote sensing data is used for a wide range of applications. One such application is environmental change detection, a research field that has attracted much attention. This technique uses multiple satellite images taken at the same location to recognize changes in the ground surface. When using multiple satellite images for environmental change detection, it is necessary to compensate for the displacement of the captured area on the image caused by the satellites' attitude. Therefore, image registration, the process of deforming images to overlap the captured areas, is applied as a preprocessing step.

Conventional image registration methods are based on artificial features, such as Scale-Invariant Feature Transform (SIFT) [3] and Accelerated KAZE (AKAZE) [4]. These methods are called “feature-based” and perform registration by detecting and matching feature points to estimate image deformation. Recently, research has focused on satellite image registration methods based on deep learning models. Since these models learn feature representations independently, they are expected to enable more precise image registration than traditional methods, even for satellite images from different time periods and sensors.

However, registering high-resolution satellite images presents a computational cost issue. Generally, the higher the resolution of an image, the higher its registration accuracy [5]. The latest satellites can capture images of the Earth’s surface with a resolution of less than 1 meter (m). The use of high-resolution satellite images is increasing. However, processing such images, including registration, is computationally costly due to the large amount of information they contain. Additionally, when considering applications in which an Artificial Intelligence (AI) system is mounted on a satellite's onboard computer, the computer's performance can be problematic. Onboard computers are generally small and lack computing power and memory capacity. This hinders the analysis of large deep learning models.

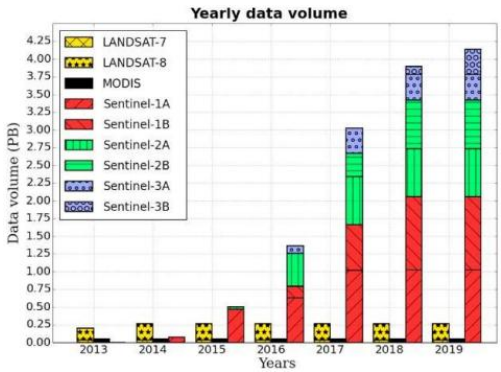


Figure 1. Estimated data volume of satellite images [2].

In this paper, we propose a lightweight, deep learning–based image registration model with a small number of parameters to address the issues. Our method is based on the geometric-matching Convolutional Neural Network (CNN) [6], a CNN-based image registration model that incorporates mechanisms to reduce parameters and improve registration accuracy. We evaluate the accuracy of our proposed method using a time-series satellite image dataset and describe the results.

The rest of this paper is organized as follows: In Section II, we describe our proposed method in detail. Section III presents the experimental results and discussion. Section IV presents the conclusions.

II. METHODS

In this paper, we propose a new registration method for time-series satellite image based on improved geometric-matching CNN approach. The flow of the method is described below.

A. Geometric-matching CNN

This paper uses a CNN based on geometric matching as the base model. The model performs registration in three steps: feature extraction, feature matching, and estimation of geometric deformation parameters. During feature extraction, features are extracted from input images I_A and I_B for registration. The input images are normalized. Two backbones with shared weights that are pre-trained by ImageNet extract features from each image. The base model is used up to the middle layer of ResNet101. These features are L2-normalized and used as feature maps, f_A and f_B, in the next step. In the feature matching step, the intermediate output, M_{AB}, is obtained by calculating the inner product of the vectors, f_A and f_B, from the channel directions of each coordinate in the feature maps, f_A and f_B. Then, ReLU (rectified linear unit) and L2 normalization are performed on M_{AB} to produce a similarity map, f_{AB}. Next, the geometric transformation parameter estimation step produces the estimated image transformation parameters, θ_{ES} which are then applied to input images I_A and I_B via the CNN architecture. The network consists of repetitive blocks made up of convolutional layers, Batch Normalization (BN) functions, and Rectified Linear Unit (ReLU) functions. In the base model of this paper, the block containing convolutional layers with 7×7 and 5×5 kernels are repeated twice, respectively. The isotropic affine transformation matrix is output in the final fully connected layer.

B. MobileNetV3

The proposed method uses MobileNetV3 [7] as the backbone for feature extraction to reduce the number of model parameters and enable fine-tuning. MobileNetV3 has a structure with multiple bottlenecks in series. Normal convolutional layer operations are divided into pointwise and depthwise convolutions to reduce computational cost. Pointwise convolution uses a 1 × 1 kernel and operates only in the channel direction of the feature map. On the other hand, depthwise convolution uses a single kernel that operates only in the spatial direction of each channel of the feature map. Additionally, a Squeeze-and-Excitation (SE) Layer [8], a type of attention mechanism, is introduced at the bottleneck. The backbone of the proposed model is based on the larger architecture of MobileNetV3, which is more accurate.

C. Light-weight Estimation

Lightweight Estimation (LE) has been incorporated into the geometric transformation parameter estimation stage of the network structure of the proposed model. The base model's structure has a significantly higher number of parameters due to convolution on a large, 900-channel feature map. Therefore, LE reduces the number of channels in the similarity map f_{AB} via pointwise convolution. Additionally, an SE layer has been added to improve the balance between parameters and registration accuracy by emphasizing important features.

D. ConvMixer

During the geometric transformation parameter estimation step, the ConvMixer model [9] is employed to enhance the precision of image registration. This is accomplished by replacing the base model's block structure with the ConvMixer's. Convolutional layers in CNNs generally have difficulty considering the relevance of information at distant locations in the feature map. ConvMixer overcomes this limitation by using depthwise convolution with a large kernel size to extract information from distant locations. However, depthwise convolution alone cannot extract relevant information in the channel direction. Thus, pointwise convolution is added to learn from the entire feature map. The geometric transformation parameter, θ_{ES} is obtained by applying Global Average Pooling (GAP) and a Fully Connected Layer (FCL) to the output of several consecutive blocks.

E. Proposed model

Table I illustrates the structure of the geometric deformation parameter estimation step. Figure 2 illustrates the architecture of the proposed model. As shown in Figure 2, we use MobileNetV3 as the feature extraction network and perform fine-tuning in the feature extraction step to reduce the number of model parameters. In the feature matching step, we incorporate a Cosine Similarity Attention (CSA) module to enhance alignment accuracy. For the following experiments, three comparative models were prepared: the +MNV3 model with MobileNetV3 only, the +LE model with LE applied to the +MNV3 model, and the +CM model with ConvMixer applied to the +MNV3 model. In the +LE model, the number of channels in the geometric deformation parameter estimation step is reduced to 64 while maintaining the same number of parameters as in the proposed model.

III. EXPERIMENTAL RESULTS AND DICUSSIONS

In our experiments, we applied the proposed method to a time series of satellite images. Some of the results and discussions are provided below.

A. Detail of the dataset

In our experiment, we apply an isotropic affine transformation to the target image and then register it with the source image. This transformation is randomly selected with a rotation angle between -π/12 and π/12, a scale factor between 0.75 and 1.25, and a translation factor between -0.125 and 0.125. We obtained 2,880 pairs of time-series satellite images from a dataset using Google Earth Pro. The image size is 480X480 pixels. The range of values for the deformation parameters was determined by implementing the geometric matching CNN. Due to the limited number of time-series satellite image pairs, an additional 13,500 images from the PASCAL VOC 2011 dataset [10] are used for training only. This experiment is based on three-fold cross-validation. In each split, the training, validation, and test data consist of 10,500, 420, and 960 image pairs, respectively.

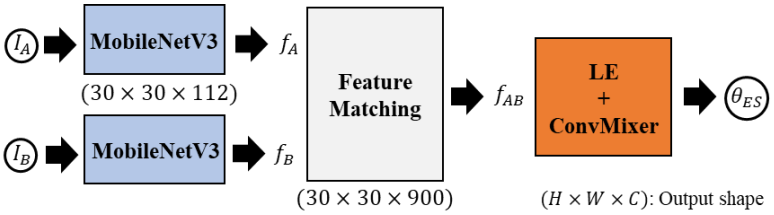


Figure 2 Architecture of proposed model.

TABLE I. STRUCTURE OF PARAMETER ESTIMATION STEOP

(a) BASE MODEL

Layer	Size	Output	Activation
Input f_{AB}	-	30×30×900	-
Conv.1	7×7	24×24×512	BN, ReLU
Conv.2	7×7	18×18×256	BN, ReLU
Conv.3	5×5	14×14×128	BN, ReLU
Conv.4	5×5	10×10×64	BN, ReLU
FCL	-	6	-

(b) PROPOSED MODEL

Layer	Size	Output	Activation
Input f_{AB}	-	30×30×900	-
PwConv.1	1×1	30×30×256	GELU, BN
DwConv.1	9×9	30×30×256	GELU, BN, Res
PwConv.2	1×1	30×30×256	GELU, BN
DwConv.1 ~ PwConv.2 ×2			
SE Layer	-	30×30×256	GELU
DwConv.1 ~ SE Layer ×4			
GAP	-	1×1×256	-
FCL	-	6	-

Pw: Pointwise, Dw: Depthwise, Res: Residual

TABLE II. EXPERIMENTAL RESULTS (EPOCHS=100)

Model	Grid MSE	Params[M]	Throughput [it/s]
Base	0.01579	57.602	15.48
+MNV3	0.01050	30.860	**29.41**
+LE	0.00858	**1.687**	28.46
+CM	0.00645	1.728	27.65
Proposed	**0.00614**	1.760	27.07

B. Evaluation metrics

Grid Mean Squared Error (Grid MSE) is used to evaluate the accuracy of image registration. It is also used as a loss function, where a smaller error results in a closer value to zero.

$$MSE_{Grid} = \frac{1}{N}\sum_{i=0}^{N} d((\tau\theta_{ES}(G_i), \tau\theta_{GT}(G_i))^2 \quad (1)$$

In this experiment, N is set to 400 and represents the number of grid points. G_i is the coordinate of the i-th grid point, and $\tau\theta_{ES}$ and $\tau\theta_{GT}$ represent the change in the coordinate of G_i due to the estimated and ground truth geometric transformation parameters, respectively. We also compare the number of parameters and throughput of the model, as well as registration accuracy.

C. Results and discussion

Table II shows the experimental results. The values in the table are the averages from three-fold cross-validation. Compared to the base model, the proposed model has fewer parameters and improves registration accuracy. Figure 3 shows examples of output images from the base and proposed models. The top row shows a target image (the registration target) and a source image (the image used for registration). The second and subsequent rows, from left to right, show the model's output image and the different image. The difference image shows registration errors; red indicates larger deformation than in the target image, and green indicates smaller deformation.

Compared to the base ResNet101 model, MobileNetV3 improves registration accuracy while reducing parameters. This shows that it's possible to improve accuracy while

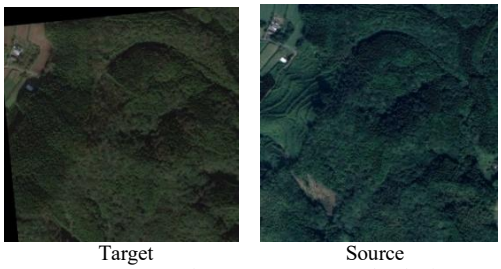
reducing parameters by fine-tuning a dataset suited for the purpose — such as time-series satellite images — rather than just pre-training with ImageNet. Using such a lightweight backbone could be beneficial in situations where computational resources for model training are limited.

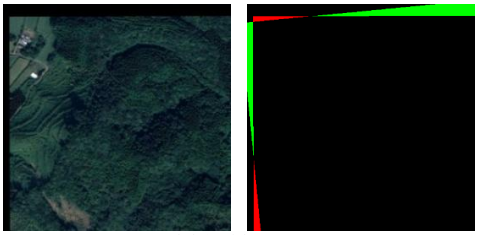
In the case of the LE, it is possible to reduce the number of parameters while improving registration accuracy. This is the result of eliminating redundancy in the convolution layer. The SE layer in LE selects important features and reduces redundant parameters through pointwise convolution. This improves accuracy and maintains it. This demonstrates that accuracy can be improved while reducing the number of parameters through pre-training using ImageNet and by fine-tuning the model on purpose-specific datasets, such as time-series satellite imagery. Using such a lightweight backbone is beneficial in situations where computational resources for model training are limited.

The ConvMixer improves registration accuracy without significantly increasing the number of parameters. This improvement stems from depthwise convolution with a nine-kernel size, which has a wide receptive field. This allows the relationship between a pixel and its surrounding regions in the similarity map f_{AB} to be captured. In the ConvMixer block structure, features are extracted while maintaining the feature map's height and width (30×30). Lower-resolution feature maps tend to lose high-frequency feature components, so maintaining the resolution enables registration with more detailed features than the base model.

Due to the modifications described above, the proposed method achieves lower parameters and higher registration accuracy than the base model. However, comparing the throughput of each model shows that reducing parameters does not necessarily lead to faster processing in the execution environment. The +MNV3 model has the fastest throughput; adding LE and ConvMixer decreases processing speed, though it is still faster than the base model. This may be because the implementation in this paper is based on PyTorch, a deep learning framework. Pointwise/depthwise convolutions are computationally less expensive than normal convolutions. Therefore, theoretically, the proposed method should be faster than the +MNV3 model. However, PyTorch optimizes GPU computation for normal convolutions, resulting in faster processing.

Figure 4 shows example images of the alignment results for the proposed method, as well as for image alignment methods based on feature point matching SIFT and AKAZE features. The RANSAC algorithm [12] is applied to the feature point matching method. As can be seen, the proposed method aligns the target image more accurately than the methods using SIFT and AKAZE features, which both produce large errors in image alignment. These errors occur because the correspondence between the detected keypoints is inaccurate, resulting in an incorrect estimated affine transformation matrix. As Figure 4 shows, SIFT and AKAZE features are not robust to time series and color variations among satellite images. However, the proposed method enables the model to learn these differences and perform correct alignment.





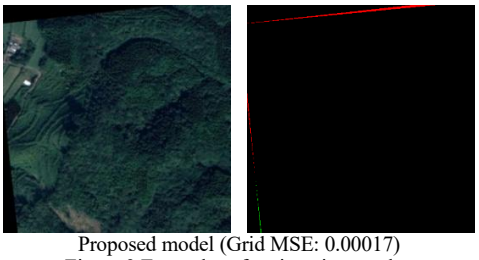
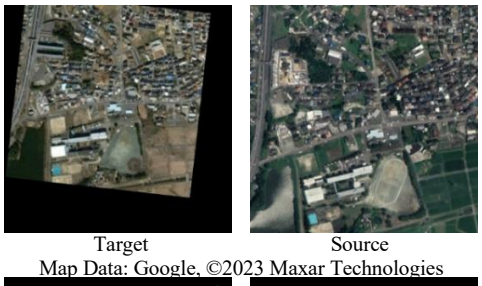


Figure 3 Examples of registration results.



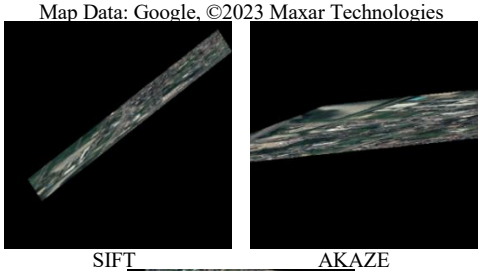


Figure 4 A comparison of the proposed method and the feature point matching methods.

IV. CONCLUSION AND FUTURE WORK

This paper was built upon the geometric-matching CNN model, reducing parameters and improving alignment accuracy. In this paper, we propose a method of aligning time series satellite images using deep learning models. The method uses CNN architecture to estimate the geometric transformation parameters necessary for image alignment, taking different time-series satellite images of the present and past as input. We designed the proposed method with a geometrically matching CNN as the base model. We incorporated mechanisms to create a low-parameter model with high positioning accuracy. For instance, we employed the MobileNetV3 backbone structure and LEs to minimize the number of parameters. We also employed the SE Layer and CSA attachment mechanisms. Additionally, we adopted the ConvMixer CNN architecture to improve accuracy while reducing the number of parameters.

Experiments using time-series satellite images were conducted to evaluate the accuracy of the proposed method. Grid MSE, which measures the error in the coordinates of the grid points in the image, and the number of model parameters were used as evaluation indices. The results showed that the proposed method (plus LE and the CM model) performed best with all parameters except CSA. This method reduced Grid MSE by 0.00965 compared to the basic model. Additionally, the number of parameters was smaller than that of the basic model, indicating an improved trade-off between the number of parameters and positioning accuracy in the proposed method.

This experiment evaluates performance based solely on Grid MSE. It is limited in its ability to assess robustness to lighting, seasonal changes, parallax, and scene-specific evaluations. We plan to address these limitations in the future. It will be necessary to further reduce the computational load and improve image alignment accuracy. This will require considering not only the theoretical computational load but also the characteristics of real-world environments, such as those of the PyTorch framework. Additionally, we should consider the practical application of the proposed method to various tasks, such as recognizing environmental changes using time-series satellite images.

ACKNOWLEDGMENT

This paper uses satellite images according to the guidelines of Google Earth [13].

REFERENCES

[1] X. Zhang et al., “Deep Learning for Processing and Analysis of Remote Sensing Big Data: A Technical Review”, Big Earth Data, Vol. 6, No. 4, pp. 527-560, 2022.

[2] P. Soille et al., “A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data”, Future Generation Computer Systems, Vol. 81, pp. 30-40, 2018.

[3] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.

[4] P. F. Alcantarilla et al., “Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces”, Proc. of British Machine Vision Conference, 2013.

[5] C. Zhao et al., “Effects of Spatial Resolution on Image Registration”, Proc. of Society of Photo-optical Instrumentation Engineers-the International Society for Optical Engineering, Vol. 9784, pp. 1-16, 2016.

[6] I. Rocco et al., “Convolutional Neural Network Architecture for Geometric Matching”, Proc. of Computer Vision and Pattern Recognition, pp. 6148-6157, 2017.

[7] A. G. Howard et al., “Searching for MobileNetV3”, Proc. of IEEE/CVF International Conference on Computer Vision, pp. 1314-1324, 2019.

[8] J. Hu et al., “Squeeze-and-Excitation Networks”, Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, 2018.

[9] A. Trockman et al., “Patches Are All You Need?”, arXiv preprint arXiv: 2201.09792, 2022.

[10] M. Everingham et al., “The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results”, http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html (Accessed:March/2024)

[11] M. A. Fischler et al., “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”, Communication of the ACM, Vol.24, No.6, pp.381-395, 1981.

[12] R.Raguram et al., “USAC: A universal framework for random sample consensus”, IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI), Vol.35, No.8, pp.2022-2038(2013).

[13] Google, “Brand Resource Center”, https://about.google/brand-resource-center/products-and-services/geo-guidelines/ (Accessed:March/2024).