Mutual superimposing of SAR and ground-level shooting images mediated by intermediate multi-altitude images

Array 12 (2021) 100102

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Mutual superimposing of SAR and ground-level shooting images mediated by intermediate multi-altitude images

Hisatoshi Toriya [a,b,*], Narihiro Owada [a], Mahdi Saadat [a], Fumiaki Inagaki [a], Ashraf Dewan [c], Youhei Kawamura [d], Itaru Kitahara [b]

[a] Faculty of International Resource Sciences, Akita University, Japan
[b] Center for Computational Sciences, University of Tsukuba, Japan
[c] School of Earth and Planetary Sciences, Curtin University, Australia
[d] Faculty of Engineering, Hokkaido University, Japan

ARTICLE INFO

Keywords:
Image registration
Keypoint matching
Synthetic aperture radar (SAR)
Remote sensing

ABSTRACT

When satellite-based SAR (Synthetic Aperture Radar) images and images acquired from the ground are registered, they offer a wealth of information such as topographic, vegetation or water surface to be extracted from the ground-level shooting images. Simultaneously, high temporal-resolution and high spatial-resolution information obtained by the ground-level shooting images can be superimposed on satellite images. However, due to the differences in imaging modality, spatial resolutions, and observation angle, it was not easy to directly extract the corresponding points between them. This paper proposes an image registration method to estimate the correspondence between SAR images and ground-level shooting images through a set of multi-altitude images taken at different heights.

1. Introduction

SAR (Synthetic Aperture Radar) mounted on satellites or aircraft is an imaging radar that uses microwaves to image the earth's surface. It is possible to obtain terrain information such as vegetation and immersion water from phase and polarization information [1]. In addition, it is also possible to superimpose terrain information acquired via ground-level shooting images and also possible to superimpose high temporal and spatial visual information on SAR satellite images, if SAR images and ground-level shooting images such as mobile cameras, car-mounted cameras, and surveillance cameras are co-registered (i.e., projection transformation between the two images is obtained). However, due to the differences in their imaging modality, spatial resolutions, and observation angle, it is difficult to directly estimate geometrical correspondence between SAR satellite images and ground-level shooting images. There have been no reports to solve this problem. This paper proposes an image-to-image registration method to find the correspondence between satellite-based SAR and ground-level shooting images through a set of multi-altitude images, taken at different heights (Fig. 1).

We achieve an image registration method between SAR satellite images and ground-captured images, which is previously unrealized. It is a novel method that allows us to integrate large-scale event recognition of objects from satellite images and detailed information such as reconstructed 3D scenes from ground-level shooting images.

In the proposed method, we classify imaging types into four categories: "SAR," "optical," "low-altitude," and "ground-level." The proposed method estimates projection transformation between "SAR image and optical image," "optical image and low-altitude image," and "low-altitude image and ground-level shooting image," respectively, by detecting the corresponding points using local features. The method integrates them to achieve an effective registration between SAR image and ground-level shooting image. Table 1 shows characteristics of the above-mentioned imaging styles and factors that made the detection of corresponding points difficult. Two-way arrows in Table 1 correspond to the following three problems. We attempted to solve them in this work by detecting corresponding points between:

I. SAR and optical images: to detect corresponding points between images that have differences in appearance due to the variations in imaging mode (modal differences) caused by the differences in observing wavelengths of the imaging sensors, we use SAR images as input and generate images that simulate the appearance of optical images (generated optical images) to achieve corresponding points

* Corresponding author. Faculty of International Resource Sciences, Akita University, Japan.
E-mail address: toriya@gipc.akita-u.ac.jp (H. Toriya).

https://doi.org/10.1016/j.array.2021.100102
Received 21 July 2021; Received in revised form 13 September 2021; Accepted 8 October 2021
Available online 15 October 2021

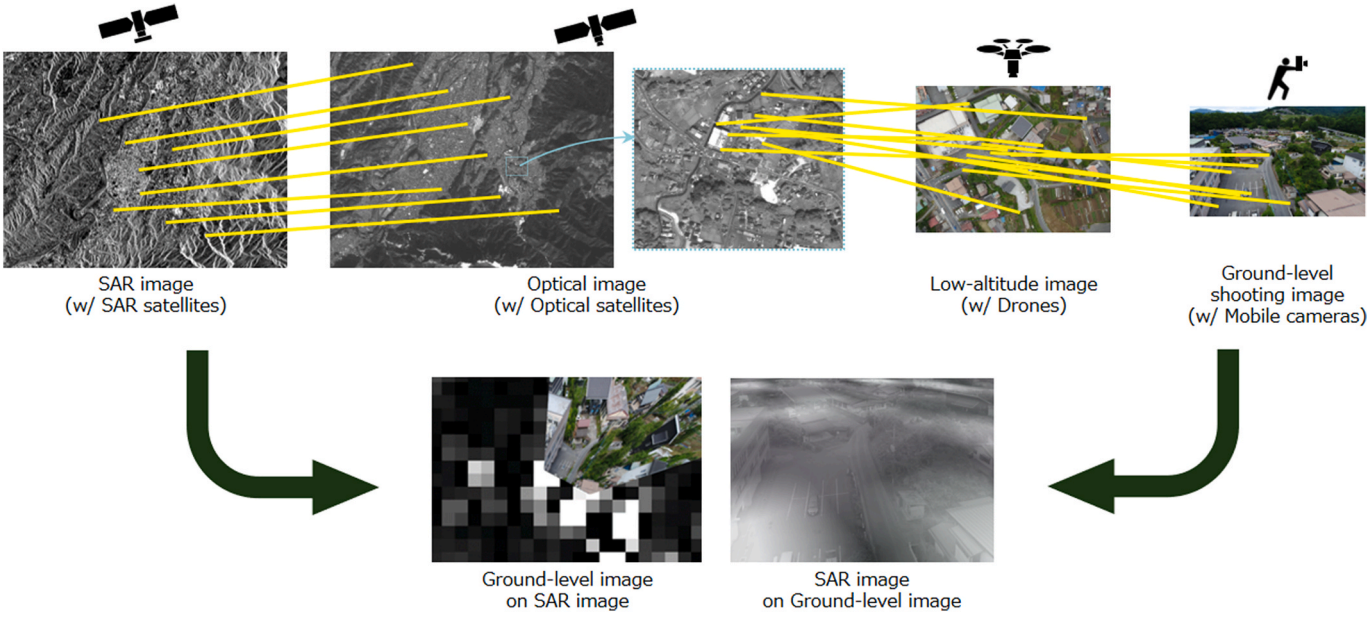
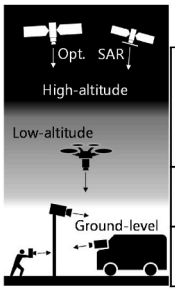


Fig. 1. Proposed method, estimating correspondence between SAR images, optical images, low-altitude images, and ground-level shooting images. By detecting the corresponding points among multi-altitude images, superimposing between SAR satellite images and ground-level shooting images is realized as shown at the bottom.

Table 1
Characteristics of image type. Two-way arrows indicate the differences in image heights..



		Capturing area	Capturing angle	Spatial resolution	Temporal resolution	Sensor
High-altitude	SAR	Wide	Vertical (downward)	Low (30+ cm/pix)	Between low and middle	SAR
	Optical	Wide	Vertical (downward)	Low (30+ cm/pix)	Low	Visible
Low-altitude		Middle	Vertical (downward)	High (10- cm/pix)	Between middle and high	Visible
Ground-level		Narrow	Horizontal	High (10- cm/pix)	High	Visible

detection by aligning image modals.

II. Optical and low-altitude images: to detect corresponding points between images with large differences in spatial resolution due to the differences in imaging altitude, optical images are upscaled using image super-resolution, and low-altitude images are downscaled to align their spatial resolutions.

III. Low-altitude and ground-level shooting images: to detect corresponding points between images with projection distortion caused by the difference in imaging directions, we generate virtual-top-view image, which is looking down from top-viewpoint, by correcting projection distortion of the ground surface area according to internal parameters of the camera and value of the accelerometer at the time of imaging. In this way, the correspondence point search between the two images is achieved.

2. Related work

2.1. Estimation of correspondence between SAR and optical images

Image registration between remote sensing images is commonly performed by using a DEM (Digital Elevation Model) and with orthorectification [2,3]. However, the spatial resolution of a DEM may be insufficient for accurate registration of SAR and optical images with high spatial resolution. Also, if we assume that we will be assessing the situation immediately after a disaster, topography of an area may change significantly after the disaster. As a result, DEM may become unsuitable to utilize. Therefore, an image feature-based registration method, which does not depend on DEM, is needed. However, wavelength of microwaves used in SAR is longer than those of visible lights. Due to the differences in their reflection and scattering characteristics, appearance of SAR and optical images differ greatly even then they represent the same area. Therefore, it is difficult to obtain common feature in both images.

It is necessary to detect common features from two images with different wavelengths to perform image-based registration (not DEM-based). Methods using template matching have been proposed, including NCC (Normalized Cross-Correlation) and Mutual Information (MI) [4,5]. These methods require a relatively wide range of image features. A small template size reduces accuracy of template matching, while a large template size reduces robustness to partial differences (e.g., pre- and post-disaster) and occlusions. Therefore, it is necessary to use local image features for registration in situations where the terrain changes partially, such as pre- and post-disaster.

Local image features are used to achieve image-based geometric registration [6–8]. In particular, keypoint-based methods have the advantage of using local correspondences to estimate total correspondence between two images, unlike template-based matching methods. The advantage is that even if a part of observed region changes,

correspondences can be obtained from other partial locations since local image features are used.

When dealing with SAR image registration, geometric distortions such as layover and foreshortening are core problems [1]. Since it is complex to correct for a layover, a method to detect regions where no geometric distortion occurs (such as ground, coastline, etc.) and to find corresponding points within the region is a solution. Foreshortening is also difficult to correct; for small foreshortening, it is possible to find the corresponding points by using an algorithm that is robust to viewpoint change distortion. In this case, image registration should be performed with non-linear projection. For low resolution images, the geometric distortion can be ignored in many cases.

2.2. Estimation of correspondence between optical and low-altitude images

In conventional image registration methods, involving low-altitude drone images and optical images, methods of generating a 3D model by applying Structure from Motion (SfM) [9] has been proposed [10,11]. However, these methods require many drone images. It is difficult to apply them to a single drone image. Applying them to a small number of images has problems such as generation error of the 3D model and the effect of satellite positioning error.

To reduce regression error by iterative processing using mutual information after initial registration using SIFT keypoints [12] and a method using NCC [13] have been proposed. However, the former method is known to have difficulty in image registration when the initial registration by SIFT keypoints fails, and the latter method presents difficulty in dealing with projective distortions between two images.

2.3. Estimation of correspondence between low-altitude and ground-level shooting images

A method has been proposed to detect corresponding points in the road area captured in both ground-level shooting images taken by a car-mounted camera and an aerial image taken by an aircraft [14]. In that method, ground area captured by a car-mounted camera is projected as if it would have been captured from a bird's eye view, based on the camera parameters so that the perspective distortion between the two images can be corrected. The projection is calculated based on camera's tilt angle. This method estimates the correspondence between two images by detecting corresponding points using SURF [15], and then estimates position and orientation of the car. In Ref. [14], it is assumed that the parameters for correcting projection distortion are fixed because a car-mounted camera with a fixed pose is used, but there are still problems for improvement to apply the method to mobile cameras, whose pose is not fixed.

3. Corresponding points detection between SAR and optical images

3.1. SAR to optical image translation using deep neural network for corresponding points detection

In SAR, active imaging is performed using microwave, and reflection and scattering characteristics are different from those of visible lights. As a result, images with a different appearance from optical images are captured even if both capture the same area. Therefore, it is difficult to detect corresponding points based on similarity of local area, such as SIFT [16]. In the proposed method, a DNN (Deep Neural Network)-based image translator (as "image translation DNN") trained by GAN (Generative Adversarial Networks) [17,18], especially cGAN (Conditional GAN) [19,20], is used to translate SAR image appearance into an optically translated image (as "generated optical image") before the process of corresponding points detection is conducted. In this section, we propose a method to find local feature correspondences between multimodal (SAR and optical) images using image feature-based keypoint detection, description, and matching algorithms, as shown in Fig. 2. Authors have proposed the core technology [21].

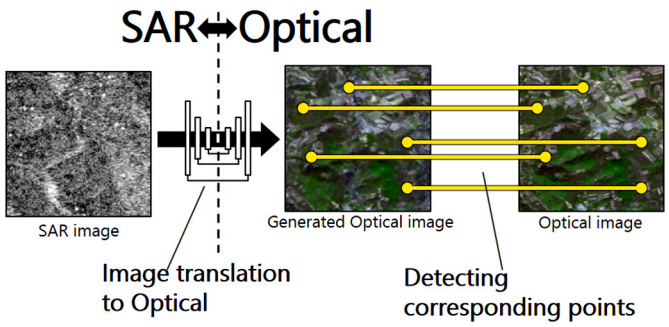


Fig. 2. Image translation and corresponding points detection using image transformation DNN, trained by GAN.

However, a previous study [21] using Pix2pix [22] to train image translation DNN by cGAN had problem of blurring details and losing local features. To obtain more accurate corresponding points, we need to use a training method that emphasizes local information. To overcome shortcoming of the previous study [21], we propose a method to obtain an image transformation DNN that does not lose local features, necessary for detecting corresponding points by applying an edge-enhancement filter to the answer image data during GAN training.

3.2. Training deep neural networks with conditional adversarial generative networks

Using SAR images as input and optical images as the answer, we train image translation DNN with cGAN and generate optical images by prediction of the image translation DNN that learned SAR to optical image translation. The GAN makes the generator and discriminator competing to produce high quality generated optical images. Accuracy of the corresponding point detection is degraded due to blur in the generated image with conventional cGAN, as shown in Fig. 3. In the proposed method, quality of the generated image by the generator G is improved by applying an edge-enhancement filter to the answer image during the training of the discriminator D.

Fig. 4 shows the cGAN model, in which DNN is trained to perform image translation. The SAR image is set as input $\boldsymbol{x}$ to the cGAN, and $F_{\mathrm{edge}}(\boldsymbol{y})$, which is the result of applying an edge enhancement filter to an optical image $\boldsymbol{y}$, is set as the answer for training the generator G (image translation DNN) and discriminator D. The process of translating SAR image to a generated optical image using this image translation DNN allows generated optical image to have the same modality as original optical image, and conventional corresponding detection process, such as SIFT, can be applied.

In this case, loss function $\mathscr{L}_{\mathrm{Proposed}}$ of the proposed method can be expressed as follows with reference to the loss function of Pix2pix [22].

$$\mathscr{L}_{\mathbf{Proposed}}(\boldsymbol{G},\boldsymbol{D}) = \mathbb{E}_{y}\left[log\boldsymbol{D}\left(\boldsymbol{F}_{\boldsymbol{edge}}(\boldsymbol{y})\right)\right] + \mathbb{E}_{x,z}[log(1-\boldsymbol{D}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})))] + \lambda\mathbb{E}_{x,y,z}\left[\left\|\mathbf{F}_{\mathbf{edge}}(\mathbf{y}) - \mathbf{G}(\mathbf{x},\mathbf{z})\right\|_{1}\right],$$

where constant λ is set to 0.1 in this work.

3.3. Corresponding points detection using keypoint detector and descriptor

An overview of corresponding points detection process of the proposed method is shown in Fig. 2. An image translation DNN is obtained using the learning model described in the previous section, and SAR image is translated into a generated optical image by the image translation DNN. SIFT [16] is used for corresponding points detection process, and it is performed between generated and actual optical image.

Fig. 3. Generated image by conventional cGAN (left) and the answer image (right). Left image, the blur smooths out the luminance gradient, which loses accuracy of the corresponding point detection.

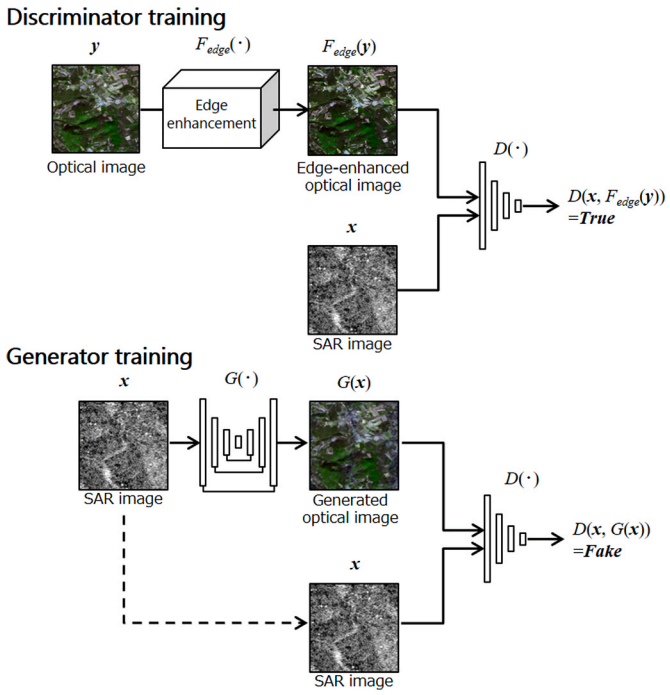


Fig. 4. A GAN that preserves local features needed for correspondence point search by adding edge enhancement filtering.

3.4. Removal process of false corresponding points

The corresponding points detection process is performed between generated and actual optical images. In the process of detecting corresponding points, some false corresponding points are identified, and these points reduce accuracy of the image registration process. In the case of registration between high-altitude images, false corresponding points can be removed according to the scale and orientation of the keypoints since they can be regarded as the differences in scale, rotation, and translation [23].

4. Corresponding points detection between optical and low-altitude images

In this section, we describe a method for detecting corresponding points between images with significantly different spatial resolutions,

Fig. 5. Difference in spatial resolution.

such as low-altitude images (drone images) and optical images (optical satellite images), as authors have proposed the core method in Ref. [24]. The spatial resolution is increased by applying a super-resolution processing to low resolution (LR) image, and at the same time, the spatial resolution of the high resolution (HR) image is decreased to align two images for improving accuracy.

As shown in Fig. 5, spatial resolutions of low-altitude and optical images differ greatly (a scale ratio of 12 times is assumed in this paper). On the other hand, a scale ratio of about 5 times is assumed in the corresponding points detection process using image features [25].

In this case, an approach to increase accuracy of the corresponding points detection process can be considered by applying super-resolution processing to an LR image (optical satellite image) to increase spatial resolution and reduce the difference between two images.

4.1. Corresponding points detection between two images with significantly different spatial resolutions by image super-resolution of characteristic regions

It is a fact that man-made structures in top-view images appear as relatively simple shapes and apply super-resolution processing to LR images to reduce the difference in spatial resolution between the two images and improve the accuracy of corresponding points detection. As an example, Fig. 6 shows a comparison between Bicubic interpolation and super-resolution. It can be seen that the proposed method restores local regions such as the white lines of roads and contours of structures compared to Bicubic interpolation.

We use a DNN-based method (as “super-resolution DNN”) for super-

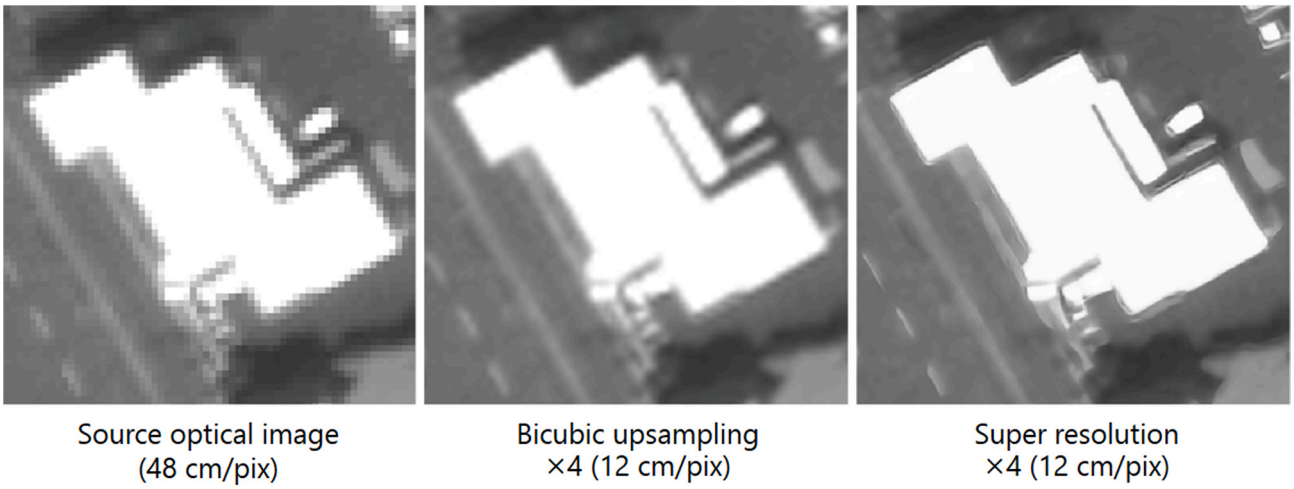


Fig. 6. Comparison of corner and edge feature estimation.

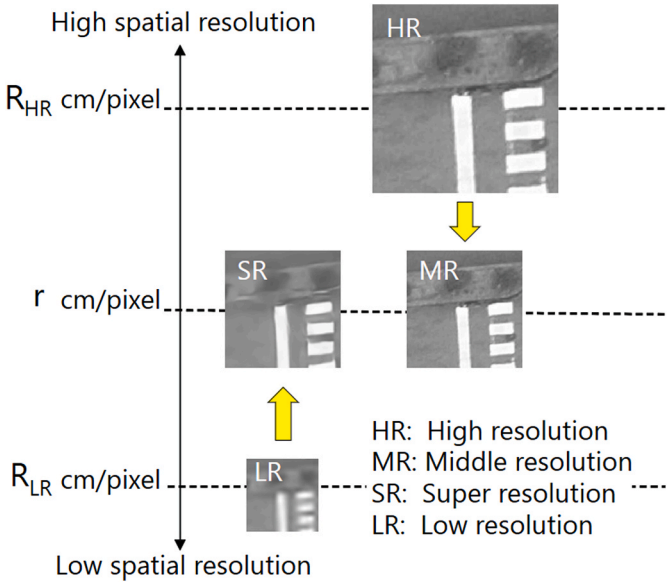


Fig. 7. Accuracy of the detection of corresponding point, spatial resolutions of two images are aligned to an appropriate value by combining super-resolution and downscaling processing.

resolution. The higher the super-resolution factor, the better the image is because it improves accuracy of corresponding points; however, if the factor is too high, artifacts are generated in the image, and many false corresponding points occur. In our investigation of the balance between super-resolution factor and false corresponding points [24], we confirmed that it is appropriate to limit the factor to 4–6 times when the difference in spatial resolution is 12 times. In our method, as shown in Fig. 7, super-resolution processing is applied to LR images (optical images), and at the same time, the downscaling process is applied to HR images (low-altitude images) to achieve a good balance between the spatial resolution and correspondence point search accuracy.

4.2. Super-resolution processing of local area

We adopt a DNN-based method for super-resolution processing. It is necessary to decide an effective region as training data to train a DNN that performs super-resolution of local regions for detecting corresponding points. Fig. 8 shows the process of training data generation. The SIFT keypoint detection is applied to a set of HR images taken by a drone in an urban area, and the regions around the derived keypoints are extracted as patches. Also, HR image is scaled down to generate the LR image. In LR image, patches are extracted from the same area as a patch in the HR image. If size of the keypoints (the diameter of the region around the keypoints used for feature detection) is too small, spatial resolution of the LR image is not adequate. Therefore, sizes of the keypoints are important consideration during the keypoint detection in the HR image, and the patch extraction is performed only around the feature points with a certain size using an appropriate threshold. The threshold value is determined according to the difference in the spatial resolution of the HR image and the LR image.

4.3. Downscaling process considering spatial resolution and correspondence points detection accuracy

In this section, we describe the process of aligning spatial resolutions of two images for corresponding point detection. To achieve accurate keypoint detection, description, and matching, spatial resolution of the two images should be the same [25]. However, as the factor of super-resolution increases, artifacts appear in the estimation by DNN, and the accuracy of the corresponding point detection decreases. Therefore, we apply high-factor super-resolution processing to the LR image so that the accuracy of the corresponding point detection does not decrease, and then downscale HR to align with LR image to generate a middle resolution (MR) image with a spatial resolution between the HR and LR images, and perform the correspondence point search with the same spatial resolution. A smoothing filter corresponding to the downscaling factor is applied in advance before the downscaling process to prevent moiré (interference fringes).

4.4. Detecting corresponding points between super-resolution and middle-resolution images and calculating the homography matrix

Since the purpose of super-resolution processing in the proposed method is to estimate features of the edges and corners of structures, the keypoint detector should distinguish those features efficiently. Also, since the orientation information is not always retained in the low-altitude images taken by drones, the keypoint detector should be able to detect corresponding points in a way that is robust to changes in rotation. In the proposed method, ORB [26] is used as keypoint detector, and the SIFT is used as the keypoint descriptor because of such conditions. The ORB keypoint detector is based on FAST [27], which can detect edges and corner features quickly, and the ORB also calculates the direction of the brightness gradient when detecting keypoints, thus, meeting the above conditions. Robust estimation of RANSAC [28] is

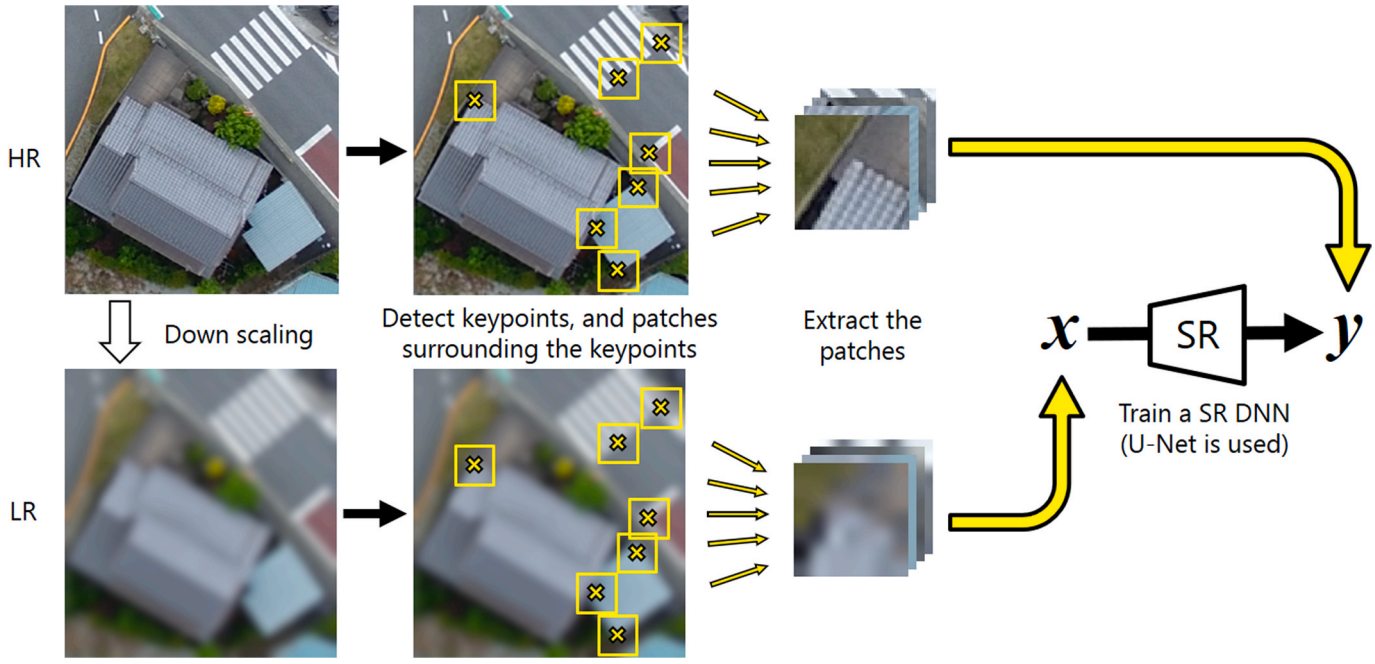


Fig. 8. The process of training dataset generation to train a super-resolution DNN to estimate local regions. The process involves detecting keypoints on the HR image and extracting patches from surrounding region, and then extracting patches from the same region of LR image.

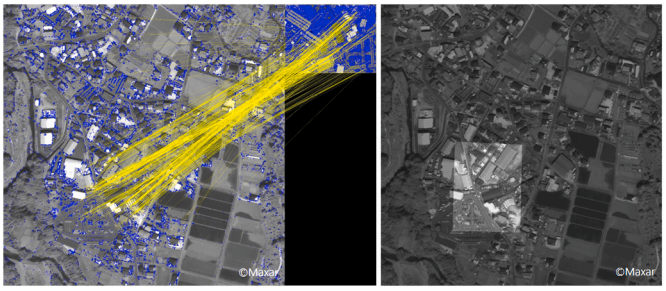


Fig. 9. Corresponding points detection by ORB + SIFT (left) and example of image registration (right). In the left, blue points indicate keypoints for which no corresponding points were found, and yellow points indicate keypoints for which corresponding points were found. The corresponding points are connected by yellow lines. The right figure shows overlaying of low-altitude image onto optical image by calculating homography transformation, based on the corresponding points information.

Fig. 10. Perspective distortion of ground area, caused by the difference in imaging direction. Left figure shows a ground-level, and right figure shows a low-altitude image. When ground surface is taken horizontally, spatial resolution is higher in the foreground area and lower in the background (perspective distortion).

performed to estimate the correspondence using the obtained correspondence information. Fig. 9 shows an example of the corresponding points detection and image registration.

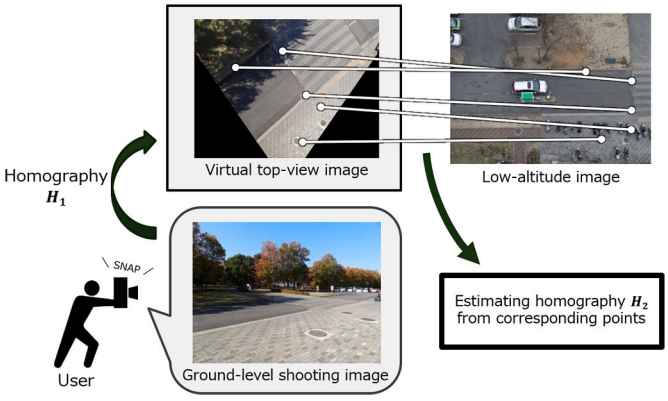


Fig. 11. An overview of corresponding points detection between low-altitude and ground-level shooting images.

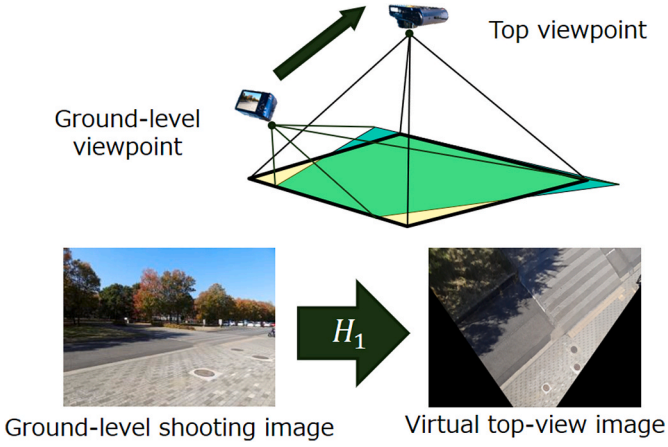


Fig. 12. Virtual top-view image generation.

Fig. 13. Detecting corresponding points between virtual top-view and low-altitude images.

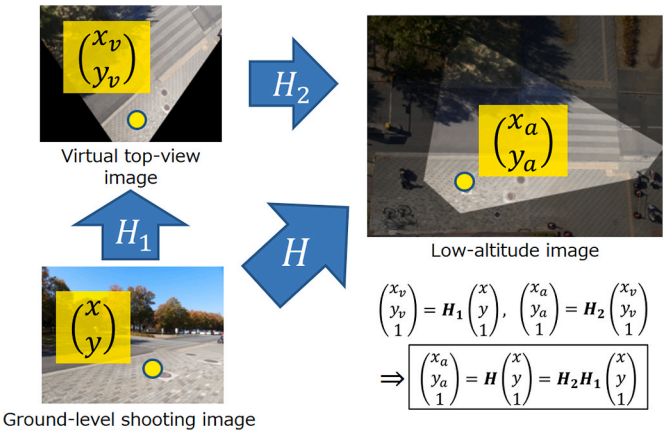


Fig. 14. Calculation of homography matrix from ground-level to a low-altitude image.

5. Corresponding points detection between low-altitude and ground-level shooting images

We propose a method to calculate geometric correspondences between low-altitude images looking down vertically and ground-level shooting images taken with mobile, car-mounted, or surveillance cameras by detecting corresponding points using keypoint matching. The direction of imaging differs greatly between low-altitude images and ground images. As a result, visibility of the object (especially ground area) changes due to perspective distortion, as shown in Fig. 10, making it difficult to detect corresponding points. Authors have proposed the core technology for mobile camera localization [23,29].

As shown in Fig. 11, this method corrects perspective distortion and achieves corresponding point detection by projecting ground area in the ground-level shooting image into a virtual top-view image, as if looking down from top-view [23].

Fig. 12 shows procedure for generating a virtual top-view image. Assuming that internal parameters of camera used for taking ground-level shooting images are known, the rotation matrix is calculated according to the camera's tilt angle at the time of taking the image, and the homography matrix $\boldsymbol{H}_1$ is estimated to transform the image into one that looks down from the vertical top-viewpoint (virtual top-view image).

As shown in Fig. 13, corresponding points between generated virtual top-view and low-altitude images are detected by the SIFT. The homography matrix $\boldsymbol{H}_2$ from virtual top-view to the low-altitude image is estimated from corresponding points.

As Fig. 14 shows, after $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are computed, homography matrix from the ground-level to low-altitude image can be represented as their product. When $(x,\ y)$, $(x_v,\ y_v)$, $(x_a,\ y_a)$, which represent points on a ground-level shooting image, a virtual top-view image and a low-altitude image, have corresponded; they are described as the following:

$$\begin{pmatrix} \boldsymbol{x_v} \\ \boldsymbol{y_v} \\ 1 \end{pmatrix} = \boldsymbol{H}_1 \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ 1 \end{pmatrix}, \begin{pmatrix} \boldsymbol{x_a} \\ \boldsymbol{y_a} \\ 1 \end{pmatrix} = \boldsymbol{H}_2 \begin{pmatrix} \boldsymbol{x_v} \\ \boldsymbol{y_v} \\ 1 \end{pmatrix} .\#(2)$$

Therefore, the homography matrix $\boldsymbol{H}$ from a ground-level shooting image to low-altitude image is described as the following:

$$H = \boldsymbol{H}_2 \boldsymbol{H}_1 \#(3)$$

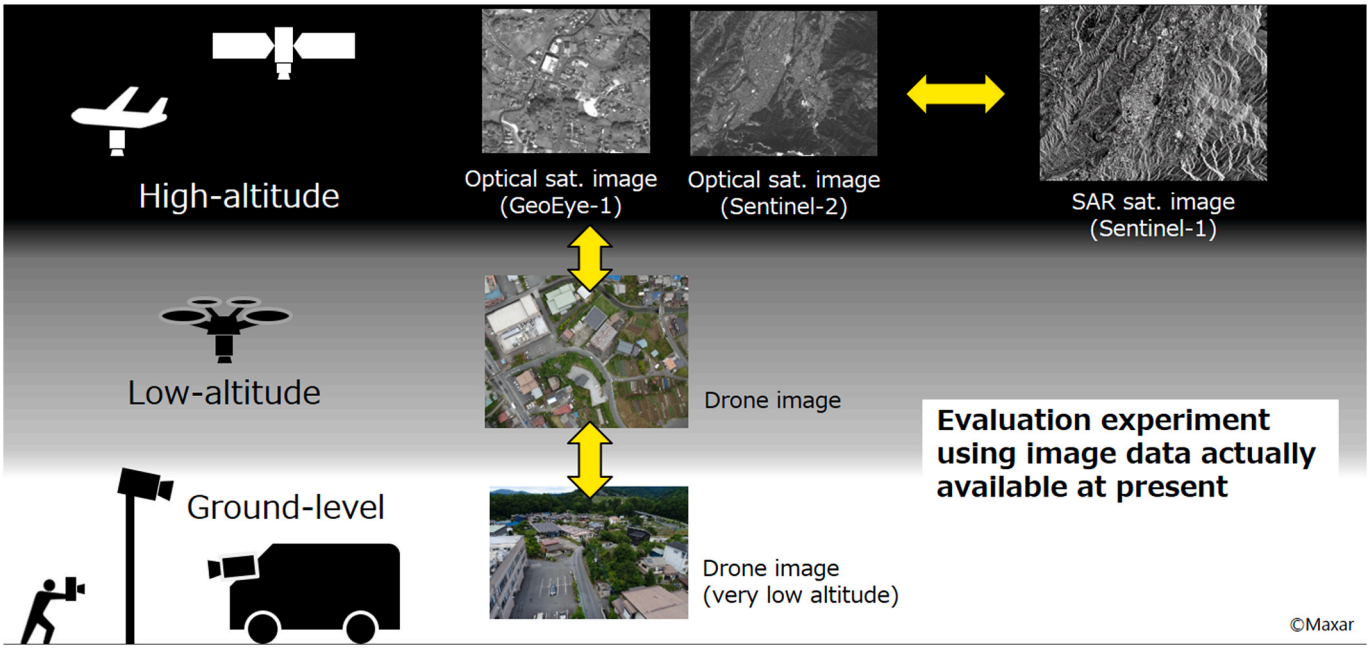


Fig. 15. Overview of the validation in this section.

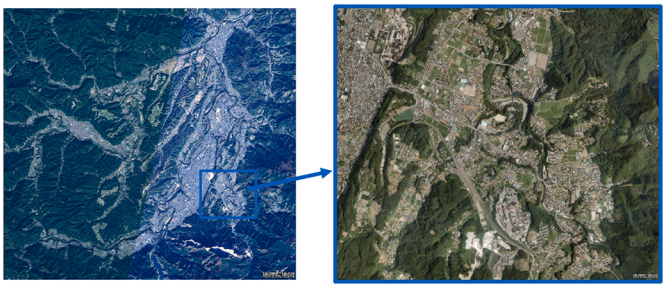


Fig. 16. Aerial view of the experimental site (Yokoze-city, Chichibu-county, Saitama-prefecture, Japan).

6. Validation of image registration process

We validate image registration between SAR and ground-level shooting images by integrating methods described in Section 3, 4, and 5, as shown in Fig. 15.

6.1. Experimental site and dataset

6.1.1. Experimental site

Fig. 16 shows an aerial image of experiment site. The experimental location was Yokoze-city, Chichibu-county, Saitama-prefecture, Japan, and the experimental dataset consisted of Sentinel-1 satellite images (SAR images), GeoEye-1 and Sentinel-2 satellite images (optical images), low-altitude images, and ground-level shooting images of a common area. The reason for using two types of optical satellite images with different spatial resolutions is to prepare images with intermediate spatial resolutions and interpolate the differences, because the spatial resolution of SAR images, which are generally available at present, is several m/pixel at best, while the spatial resolution of optical satellite images is several tens of cm/pixel. The GeoEye-1 and Sentinel-2 satellite images have geographic information assigned to each pixel (i.e., they are already registered), so it is possible to obtain accurate correspondences.

6.1.2. Training dataset for image translation DNN

Sentinel-1 SAR satellite images and Sentinel-2 optical satellite images (optical images) of the experiment site are used to train the image translation DNN. 16 images, which are not covered with cloud, were selected of 67 images taken by Sentinel-2A and Sentinel-2B satellites during the period from January 1 to December 31, 2019, and Sentinel-1 satellite images with close taking dates were selected as a pair, respectively. Table 2 shows selected image datasets. One of the 16 images was used as validation dataset.

Each image pair was registered using a DEM (Digital Elevation Model), obtained from Fundamental Geospatial Data Download Service of the Geospatial Information Authority of Japan [30], and spatial resolution of each pair was set to 12.5 m/pixel so that the images overlapped pixel-by-pixel. Patches of 256x256 pixels were cut out every 128 pixels in height and width from the images, and 896 patches were prepared as the training and 56 patches as validation datasets. The brightness values in the range of $[\mu-2\sigma,\ \mu+2\sigma]$ were linearly normalized to the range of $[-1,\ 1]$ and used for training and validation.

6.1.3. Training dataset for super-resolution DNN

As HR image, we used aerial images of urban area taken by a drone in Oshima-town, Tokyo-prefecture, Japan, on September 19, 2019. The drone was a DJI Phantom 4, and spatial resolution of the HR images was

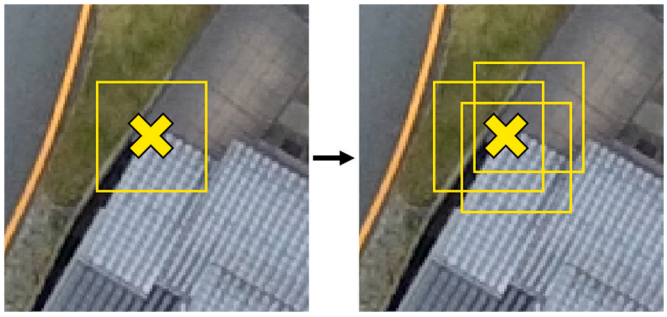


Fig. 17. Random shift during patch extraction.

Table 2
Date and time of the image dataset used to train image translation DNN. Highlighted rows refer to validation dataset.

SAR satellite images Satellite: Sentinel-1 Relative orbit number: 46		Optical satellite images Satellite: Sentinel-2 Relative orbit number: 74	
Unit	Time (UTC)	Unit	Time (UTC)
A	2019/01/03 - 20:43	B	2019/01/03 - 01:30
A	2019/01/15 - 20:43	B	2019/01/13 - 01:30
A	2019/01/27 - 20:43	B	2019/01/23 - 01:29
A	2019/02/08 - 20:43	B	2019/02/12 - 01:27
A	2019/02/20 - 20:43	A	2019/02/17 - 01:27
A	2019/03/04 - 20:43	A	2019/03/09 - 01:26
A	2019/03/16 - 20:43	B	2019/03/14 - 01:26
A	2019/03/28 - 20:43	B	2019/03/24 - 01:26
A	2019/04/09 - 20:43	B	2019/04/13 - 01:26
A	2019/04/21 - 20:43	A	2019/04/18 - 01:27
A	2019/05/03 - 20:43	A	2019/04/28 - 01:27
A	2019/05/15 - 20:43	B	2019/05/13 - 01:27
A	2019/10/06 - 20:43	B	2019/10/10 - 01:26
A	2019/10/30 - 20:43	B	2019/10/30 - 01:27
A	2019/11/11 - 20:43	B	2019/11/09 - 01:28
A	2019/11/23 - 20:43	A	2019/12/04 - 01:30

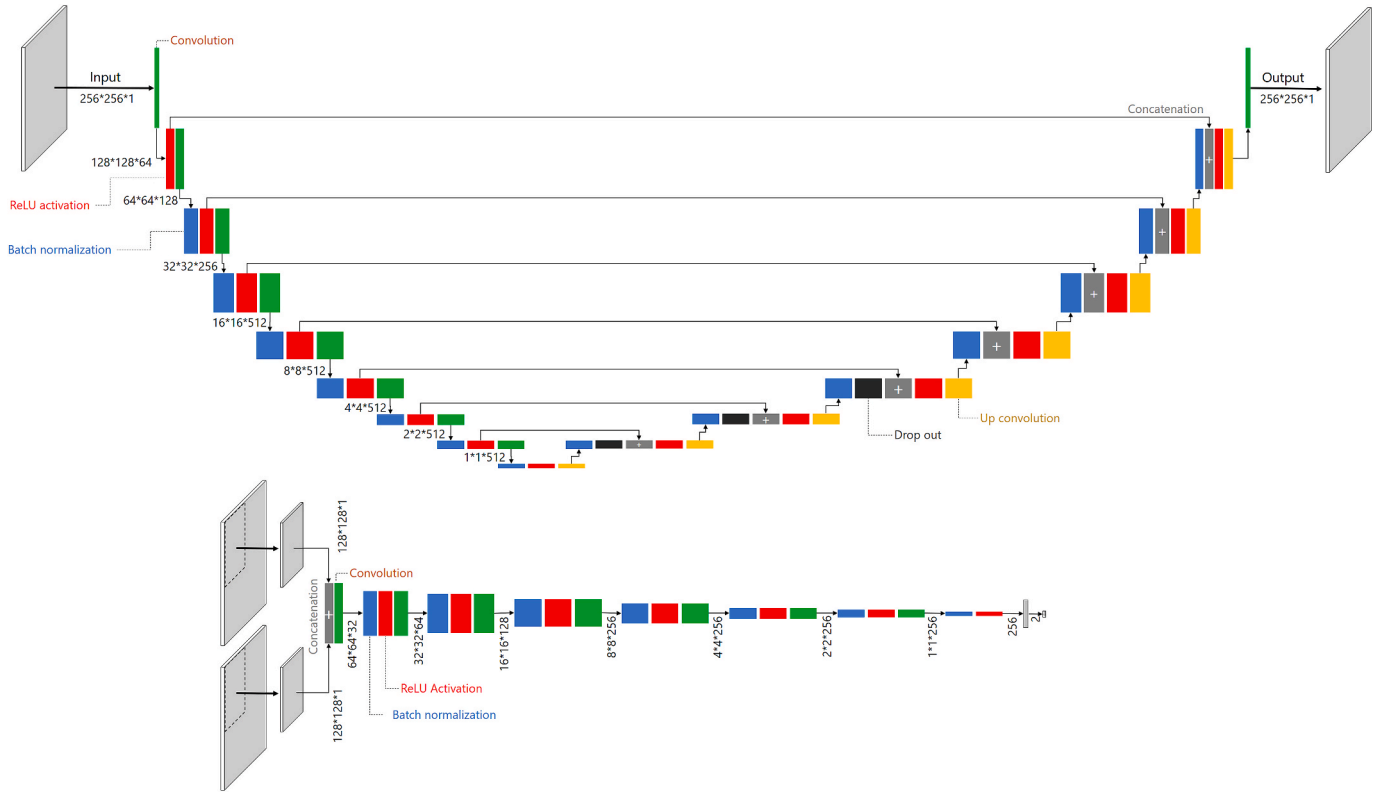


Fig. 18. Structure of generator (up) and discriminator (bottom) for image translation.

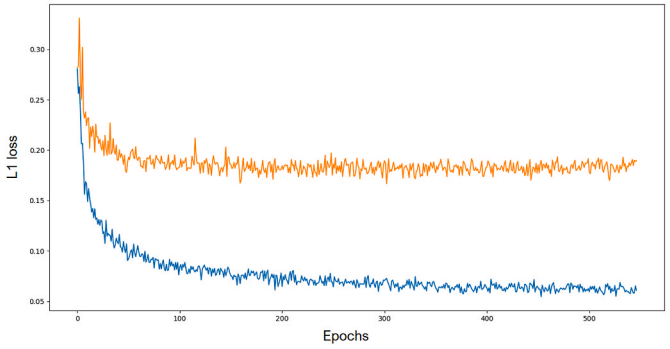


Fig. 19. Graph of loss function of the generator in training an image translation DNN. The blue and orange lines show loss function for training and validation data, respectively.

4.0 cm/pixel. The HR image was reduced to 1/4 of its original size to obtain the LR image for training. The threshold of the keypoint size was set to 12.0 pixels. As shown in Fig. 17, a random shift was added to the patch extraction process. This is because prior experiments showed that the keypoints are always located in the center of the patch so that point-like artifacts are observed in the super resolution (SR) image without random shift. The patch size was 128x128 pixels, and the number of patches was about 36,000. Each pixel depth of the images in the training dataset was unsigned 8-bit RGB, which was first converted to grayscale to make one channel, and then the brightness values in the range of $[0, 255]$ were linearly normalized to the range of $[-1, 1]$ as the input.

6.2. DNN training details

6.2.1. Image translation DNN

U-Net [31] was used as image translation DNN to translate SAR images to generated optical images and was trained by cGAN, as shown in Fig. 18. The image patch size used for training was 256x256 pixels, the input batch size was 16, and the training time was 12 h. The learning rate of the generator and discriminator was set to 1.0×10^{-3}. The loss functions were the L1 norm and binary cross-entropy, respectively. The patch size of PatchGAN [22] was set to 128x128 pixels. We used Cygnus [32], a high-performance computer at the Center for Computational Science, University of Tsukuba. As for the training time, as shown in Fig. 19, it was confirmed that the value of the loss function in the validation data of the generator (image translation DNN) did not change for a long time, so the training was terminated after 12 h, and the image translation DNN was constructed using the parameters whose loss function in the validation data had the minimum value of 1.668×10^{-1}. The image translation DNN was constructed using the parameters whose loss function was 1.668×10^{-1}.

The edge-enhancement filter for training the discriminator was implemented as:

$$F_{edge}(v, \boldsymbol{I}) = \begin{bmatrix} -v & -v & -v \\ -v & 1+8v & -v \\ -v & -v & -v \end{bmatrix} \boldsymbol{I},$$

where $\boldsymbol{I}$ is input image and v is a parameter for setting the strength of the edge enhancement. In this experiment, $v = 0.1$ was used because it revealed promising results in preliminary experiments.

6.2.2. Super-resolution DNN

U-Net was used as super-resolution DNN for optical images, and it was trained by GAN, as shown in Fig. 20. The image size used for training was 128x128 pixels, the input batch size was 128, and the training time was 48 h. The initial learning rate of the generator and discriminator was set to 1.0×10^{-3}, which was halved after 24 h. The loss functions were the L1 norm and binary cross-entropy, respectively. The patch size of PatchGAN was set to 64x64 pixels. As for the training time, as shown in Fig. 21, it was confirmed that there was little change in the value of the loss function in the verification data of the generator

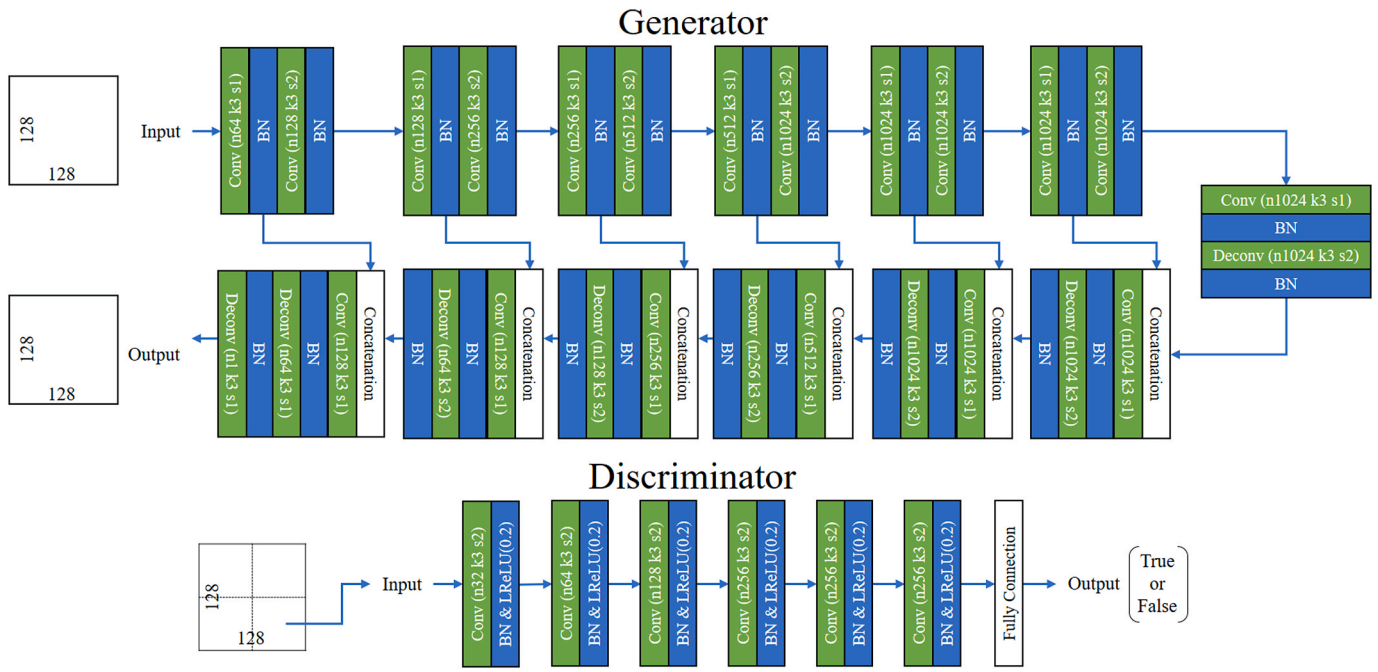


Fig. 20. The structure of the generator and discriminator for super-resolution.

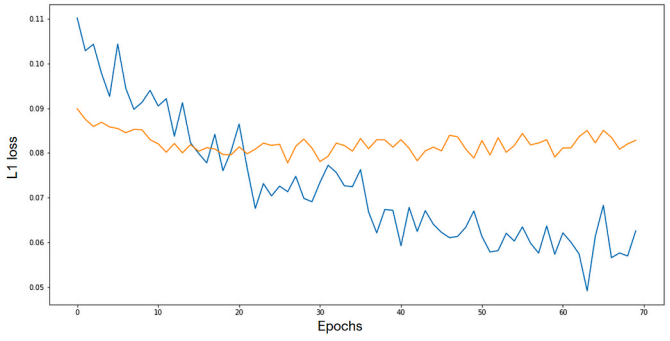


Fig. 21. Loss function of the generator in training a super-resolution DNN. The blue and orange lines show loss function of training and validation data.

(super-resolution DNN) for a long time, so the training was terminated after 48 h. The super-resolution DNN was constructed using parameters when the loss function in the validation data was a minimum value of 7.775 × 10–2.

7. Result and discussion

Fig. 22 shows resulting images of the corresponding point detection between each image. It can be confirmed that the proposed method works effectively in each processing step, and accurately detect correspondence.

Fig. 23 and Fig. 24 show results of registration of SAR and ground-level shooting images based on detected corresponding points. Information obtained by SAR can be overlaid on ground-level shooting image. Also, it can be confirmed that information with high temporal (high frequency) and high spatial resolution, which is close to the ground surface and the information on the ground surface can be understood in detail, is overlayed on SAR image.

One possible application of this research is sharing of information of damaged areas immediately after a disaster. We can share this by taking advantage of SAR characteristics, which enables observation even in poor atmospheric condition, and characteristics of ground-level shooting images, which enables us to capture localized changes that are difficult to capture with high-altitude.

Another possible application is outdoor XR (a collective term for Augmented Reality, Virtual Reality, and Mixed Reality). As shown in Fig. 23, it is possible to realize a system that overlays geographic information on ground-level viewpoint. However, as shown on the right of Fig. 23, if satellite images are directly overlaid on ground-level viewpoint, useful information cannot be presented due to a difference in spatial resolution. Therefore, it is necessary to construct a system that takes into account the difference in spatial resolution, such as converting information extracted from high-altitude images into a vector layer and overlaying it on the ground-level viewpoint.

8. Conclusion

In this work, we proposed a method to superimpose SAR images with ground-level shooting images by detecting corresponding points between multi-altitude images taken at different altitudes. The possibility and usefulness of the method were demonstrated by an experiment using actual image. This makes it possible to mutually superimpose information from SAR satellite images and ground-level shooting images, which were previously handled completely separately.

We have achieved a method of image registration between SAR satellite images and ground-captured images. It is a novel method that allows us to integrate large-scale event recognition of objects from satellite images and detailed information such as reconstructed 3D scenes from real-world information of ground-level shooting images.

As a further prospect for the future, by applying our proposed method to a group of images collected by crowdsourcing [33], it will be possible to achieve high spatial and temporal resolution at any point, and we can expect to construct a “next-generation GIS (Geographic Information System)” that is distinct from the current GIS. The “next-generation GIS” will enable instantaneous sharing of information that transcends differences in imaging devices, locations, and time. Also, it is expected to be used as an essential information source for society as an information infrastructure.

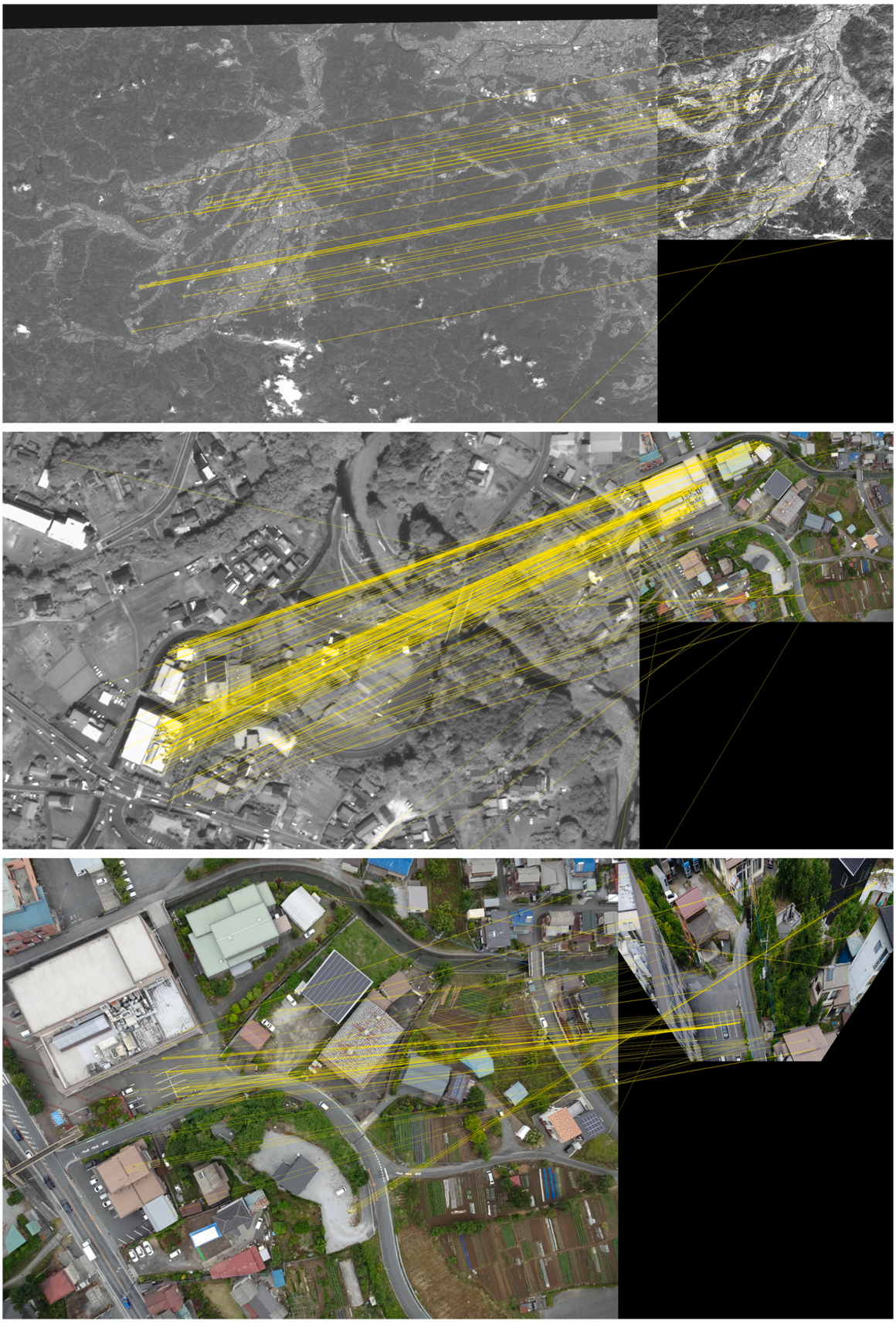


Fig. 22. Results of corresponding point detection between multi-altitude images. Results for optical and generated optical images (top), optical and low-altitude images (center), and low-altitude and ground-level shooting images (bottom) are shown.

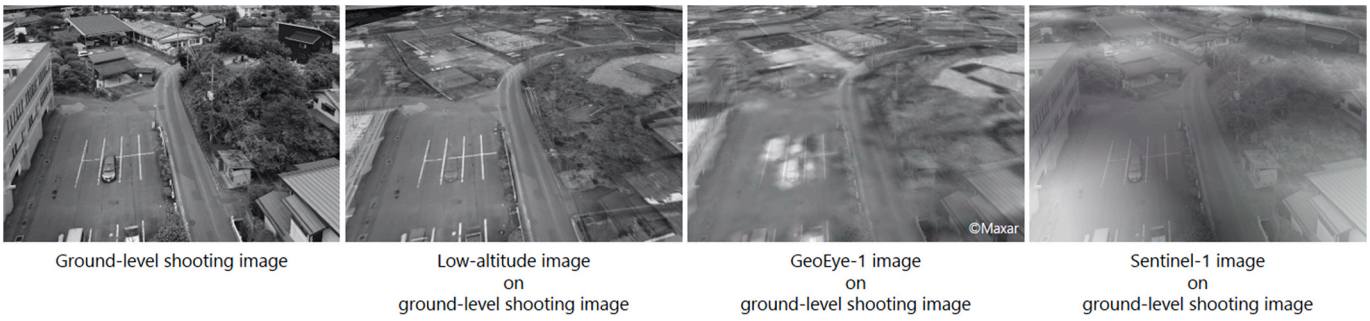


Fig. 23. The result of overlaying multi-altitude images onto a ground-level shooting image.

Fig. 24. Registration result between SAR and ground-level shooting images. It can be seen that areas where retroreflections occur, such as buildings, are represented by bright pixels.

Author contributions

Hisatoshi Toriya: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing–original draft preparation, Writing–review and editing, Investigation, Visualization. **Narihiro Owada**: Writing–original draft preparation, Validation, Visualization. **Mahdi Saadat**: Methodology. **Fumiaki Inagaki**: Validation. **Ashraf Dewan**: Conceptualization, Resources, Writing–review and editing, Supervision. **Youhei Kawamura**: Supervision. **Itaru Kitahara**: Conceptualization, Methodology, Resources, Writing–review and editing, Supervision, Project administration, Funding acquisition.

Funding

This work was supported by Grant-in-Aid for JSPS Fellows Number JP19J11514, Japan; and JST CREST Grant Number JPMJCR16E3, Japan.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Drone photography was conducted in cooperation with DRONEBIRD by Crisis Mappers Japan (NPO), a disaster drone rescue team. (http://dronebird.org/, accessed: Apr. 13, 2021.)

References

[1] Floyed M H, Anthony J L. Principles and applications of imaging radar. In: Manual of remote sensing. John Wiley & Sons, Inc.; 1998. third ed., Volume 2, third ed.
[2] Le Moigne J, Netanyahu N, Eastman R. Image registration for remote sensing. Cambridge University Press; 2011.
[3] Jensen JR. Introductory digital image processing: a remote sensing perspective. fourth ed. Prentice Hall Press; 2016.
[4] Dame A, Marchand E. Accurate real-time tracking using mutual information. 2010. https://doi.org/10.1109/ISMAR.2010.5643550.
[5] Xiong B, Li W, Zhao L, Lu J, Zhang X, Kuang G. Registration for SAR and optical images based on straight line features and mutual information. 2016. https://doi.org/10.1109/IGARSS.2016.7729667.
[6] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, "Robust optical-to-SAR image matching based on shape properties," Geosci Rem Sens Lett IEEE, vol. 14, no. 4, pp. 564–568, Apr. 2017, doi: 10.1109/LGRS.2017.2660067.
[7] Xiang Y, Tao R, Wang F, You H. Automatic registration of optical and SAR images VIA improved phase congruency. " in *international Geoscience and remote sensing symposium*. 2019. p. 931–4. https://doi.org/10.1109/IGARSS.2019.8898506.
[8] Ye Y, Bruzzone L, Shan J, Bovolo F, Zhu Q. Fast and robust matching for multimodal remote sensing image registration. IEEE Trans Geosci Rem Sens 2019; 57(11). https://doi.org/10.1109/TGRS.2019.2924684. 9059–9070, Nov.
[9] Snavely N, Seitz SM, Szeliski R. Photo tourism: exploring photo collections in 3D. " in *ACM SIGGRAPH*. 2006. p. 835–46. https://doi.org/10.1145/1179352.1141964.
[10] M. A. Fonstad, J. T. Dietrich, B. C. Courville, J. L. Jensen, and P. E. Carbonneau, "Topographic structure from motion: a new development in photogrammetric measurement," Earth Surf Process Landforms, vol. 38, no. 4, pp. 421–430, Mar. 2013, doi: 10.1002/esp.3366.
[11] Kobayashi K, Shishido H, Kameda Y, Kitahara I. A method to collect multi-view images of high importance using disaster map and crowdsourcing. " in *IEEE international Conference on big data*. Jan. 2019. p. 3510–2. https://doi.org/10.1109/BigData.2018.8622193.
[12] Huang SM, Huang CC, Chou CC. Image registration among UAV image sequence and Google satellite image under quality mismatch. " in *international Conference on ITS telecommunications*. 2012. p. 311–5. https://doi.org/10.1109/ITST.2012.6425189.
[13] Fan B, Du Y, Zhu L, Tang Y. The registration of UAV down-looking aerial images to satellite images with image entropy and edges. In: International conference on intelligent robotics and applications; Nov. 2010. p. 609–17. https://doi.org/10.1007/978-3-642-16584-9_59.
[14] Noda M, et al. Vehicle ego-localization by matching in-vehicle camera images to an aerial image. " in *asian Conference on computer vision*. 2010. p. 163–73. https://doi.org/10.1007/978-3-642-22819-3_17.
[15] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). Comput Vis Image Understand 2008. https://doi.org/10.1016/j.cviu.2007.09.014.
[16] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94. Nov.
[17] Goodfellow IJ, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014; 3:2672–80. Accessed: Sep. 19, 2020. [Online]. Available: http://www.github.com/goodfeli/adversarial.
[18] Radford A, Metz L, Chintala S. "Unsupervised representation learning with deep convolutional generative adversarial networks," Nov. 2016 [Online]. Available: http://arxiv.org/abs/1511.06434.
[19] Mirza M, Osindero S. "Conditional generative adversarial nets,". http://arxiv.org/abs/1411.1784. [Accessed 14 December 2020]. accessed.
[20] Miyato T, Koyama M. cGANs with projection discriminator. " Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.05637.
[21] Toriya H, Dewan A, Kitahara I. "SAR2OPT: image alignment between multi-modal images using generative adversarial networks. " in *international Geoscience and remote sensing symposium*. 2019. p. 923–6. https://doi.org/10.1109/IGARSS.2019.8898605.

[22] Isola P, Zhu JY, Zhou T, Efros AA. “Image-to-image translation with conditional adversarial networks. IEEE conference on computer vision and pattern recognition, vol. 2017; Nov. 2017. p. 5967–76. https://doi.org/10.1109/CVPR.2017.632.
[23] Toriya H, Kitahara I, Ohta Y. Mobile camera localization using aerial-view images. IPSJ Trans. Comput. Vis. Appl. 2014;6:111–9. https://doi.org/10.2197/ipsjtcva.6.111.
[24] Toriya H, Dewan A, Kitahara I. Adaptive image scaling for corresponding points matching between images with differing spatial resolutions. ” in *IEEE international Conference on big data*. 2021. p. 3088–95. https://doi.org/10.1109/bigdata50022.2020.9377754.
[25] Tareen SAK, Saleem Z. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK,” in. International Conference on Computing, Mathematics and Engineering Technologies 2018;January:1–10. https://doi.org/10.1109/ICOMET.2018.8346440.
[26] Rublee E, Rabaud V, Konolige K, Bradski G. “ORB: an efficient alternative to SIFT or SURF,” in IEEE conference on computer vision and pattern recognition. 2011. p. 2564–71. https://doi.org/10.1109/ICCV.2011.6126544.
[27] Rosten E, Drummond T. Machine learning for high-speed corner detection. In: European conference on computer vision, vol. 3951. LNCS; 2006. p. 430–43. https://doi.org/10.1007/11744023_34.
[28] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM Jun. 1981;24(6):381–95. https://doi.org/10.1145/358669.358692.
[29] Toriya H, Kitahara I, Ohta Y. A mobile camera localization method using aerial-view images. 2013. https://doi.org/10.1109/ACPR.2013.27.
[30] Fundamental geospatial data (FGD) - geospatial information authority of Japan. accessed Apr. 07, 2021), https://www.gsi.go.jp/kiban/index.html.
[31] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 2015;9351:234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
[32] “Supercomputers – Center for Computational Sciences. accessed Apr. 07, 2021, https://www.ccs.tsukuba.ac.jp/eng/supercomputers/.
[33] Howe J. The rise of crowdsourcing. Wired Mag. 2006. https://doi.org/10.1086/599595.