The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

RFBfeat Deep Neural Network for Crossview Remote Sensing Image Matching

Mohamamd Erfan Messbah* Ali Jafari†

Abstract

Accurate image matching is a fundamental task in computer vision with diverse applications, but it faces several major challenges. These challenges include differences in perspective and the complexity of managing images with different resolutions. Differences in perspective, attributed to different angles of image registration, have historically confounded conventional matching methods. Another challenge is the different resolution of input images. In this article, a deep neural network is proposed to match satellite images with UAV images, which solves these two major challenges. In this network, RFB blocks are used to enhance feature maps. RFBs expand the field of influence while taking into account central information, thereby strengthening the robustness of the network and ultimately improving the matching results. The simulation results show that the proposed method has a higher efficiency than the existing methods and has been able to match two images with two different viewing angles well.

Keywords: Image Matching, Deep Neural Network, Remote Sensing Images, Drone Images

1 Introduction

The process of finding the corresponding points in two images of the same scene that were captured under different conditions is called image matching. Image matching applications include 3D reconstruction[2, 16], object recognition[5, 1] a motion tracking[13].

Image matching faces many obstacles that hinder its efficiency. Among these challenges, the most important challenges are caused by the distortions caused by the changes in the camera perspective of the images. Camera angles and perspective have a profound effect on how objects appear in images. A slight displacement of the viewing angle can lead to significant distortions such as stretching, compression and even complete visual transformations. This makes it difficult for algorithms and CNN methods to identify relevant features and leads to false matches. For example, changing the perspective in the images of a building taken from different angles can drastically change its appearance, and complex algorithms are needed to overcome these differences in matching images.

Figure 1: The difference between the two perspectives of satellite and drone

This challenge is more visible in satellite images and UAV images. Figure 1 shows the difference between the two views. Satellites take pictures with a vertical perspective and from a very long distance. These images create a broad view and allow the observation of broad patterns and trends in large areas. On the other hand, UAVs fly much lower than satellites and are only a few hundred meters high at best. As a result, they can record accurate images of small areas with high resolution. Drone images, taken from a more horizontal perspective, can suffer from lens distortion, which can make straight lines crooked or buildings appear tilted.

Another challenge of matching is the different resolutions of the images. UAVs with their lower altitude can record images with a much higher resolution than satellites, and in some cases up to a few centimeters. This allows for more detailed analysis of the Earth's surface, making UAV imagery ideal for applications that require precise measurements or identification of specific objects. On the other hand, the resolution of satellite images is limited by the size of the sensors in the satellites and the distance between the satellites and the earth. As a result, satellite images usually have a much higher resolution than drones.

To solve these problems, a new network is presented by combining RFB in the neural network in order to build a strong descriptor and multi-layer detection of key points.

*Faculty of Electrical and Computer, Malek Ashtar University of Technology, Iran, messbah.m.e@gmail.com

†Faculty of Electrical and Computer, Malek Ashtar University of Technology, Iran , iustuser@mut.ac.ir

2 Related Works

2.1 Matching

The primary goal of the multi-step computer vision process known as "image matching" is to identify a match between two or more images. Key Points or Interest Point identification is the first step in this procedure, when unique spots within photos are found. The following stage, feature description, involves calculating a descriptor for each interest point that has been identified. Finding a match between the two images Interest points is the final and most important step. Here, creating a relationship or link between the Interest points in one image and their equivalents in another image is the aim. The descriptors of interest points in the two images must be compared for this process to work.

In general, image matching can be divided into two categories:

detect, then describe: These techniques involve a step-by-step procedure. The first step is to locate important details in images. Feature descriptors are used to describe the highlighted points. Rich details about the surrounding local image context of each key point are contained in these descriptors[7, 15, 10].

detect and describe: Another method for finding matches in deep learning techniques is known as "simultaneous detection and description". This method uses a neural network architecture in which the processes of interest point detection and description run concurrently. Simultaneous detection and description enable the generation of key points and their corresponding descriptors simultaneously, as opposed to a sequential workflow. The effectiveness of this method and its smooth integration with deep learning-based image matching techniques have drawn attention to it in recent years[9, 4, 12].

2.2 Interest Points

Unique locations in an image that are easily recognized in a variety of settings, such as brightness, scale, and rotation, are known as key points or interest points. A interest point's attributes include its ability to stand out from the background, its immutability across a range of scales, its resilience to noise, and its evenly distributed density throughout the image. Because it forces us to select unique points, the proper selection of these points also contributes to the descriptors improvement.

The selection of key points in SIFT[7] method is applied to different scales constructed from the image using Gaussian differential filter and the localization process is applied to accurately estimate the location and scale. In superpoint[3], the shared encoder processes the image and classifies the sub-network. In this part, different techniques such as gradient analysis, corner detection and image bubble analysis are analyzed and feature points are extracted. In Loftr[12], using the attention mechanism, the regions that are probably similar to each other are selected, then matching is done in that part of the list.

Figure 2: example of Intrest Points

2.3 Feature Descriptor

Feature description involves the creation of compact and distinct representations of points of interest detected in an image. These descriptors contain essential information about local visual features around points of interest. Advanced deep learning techniques and neural network architectures aim to create descriptors that remain invariant to changes such as scale, rotation, and lighting changes. These descriptors serve as representations of strong and distinct characteristics. By capturing essential details while reducing noise and variation, feature description techniques in these methods increase the accuracy and efficiency of subsequent computer vision applications. In [7], the description for each key point is

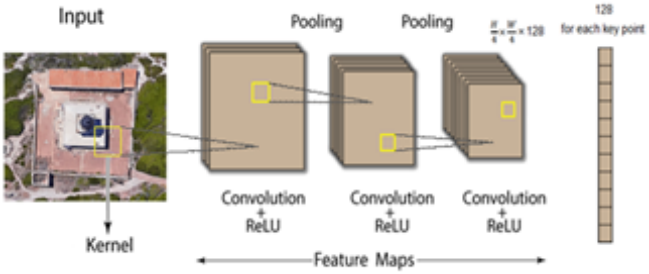

Figure 3: How to obtain descriptors in deep neural networks

obtained separately, the descriptor divides the regions around the key point into sub-regions and calculates the histogram of the gradient directions for it, and then normalizes the obtained histogram. In [4, 9], the extracted

features are transferred from the common encoder and this sub-network calculates descriptions with a fixed length. In [12], the output four-dimensional space of the attention mechanisms of two images are multiplied together and form a probability space, in this part, the regional slices of two images are multiplied again and form a new probability space, and the correlation of each the point in the state space is multiplied and the output is obtained.

3 Method

3.1 RFB

Receptive Field Block[6], which originated from detection techniques, offers a compelling strategy to enhance a neural network's capacity to describe images. Generally speaking, adding more filters to a network enhances its accuracy; however, this also adds weight to the network and As a result, the network needs more data for training, the amount of time needed to test the network. This network's design includes a particular block called RFB, which is essential to the creation of feature maps. A "field of influence" in computer vision and deep learning is the area of the input image that influences a specific neuron's output. This sphere of effect in RFB is purposefully expanded to enable it to catch a greater portion of the input image. Figure 4 depicts

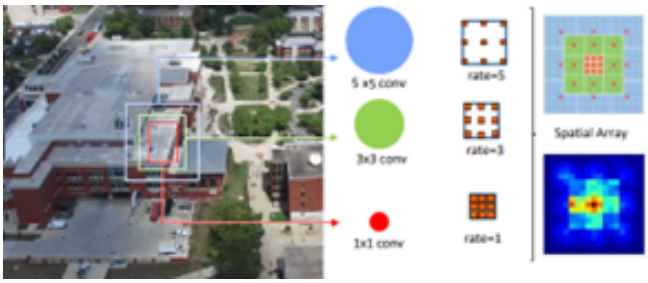

Figure 4: RFB is a multi-layer convolution block that uses various kernel sizes to combine features with variable field-of-effect sizes, improving feature extraction. This enhances the network's capacity to precisely recognize features of various sizes and aids in the capture of local and global characteristics

the rfb block's architecture. This design applies a series of parallel operations, starting with a 1*1 convolution layer, followed by 3*3 convolutions in rows 1, in row 2, 3*3 convolution layer, followed by 3*3 convolutions, and 3, as well as 5*5 convolutions before 3*3 convolutions in rows 3. The final output is joined together. A 1*1 convolutional filter's output stabilization is used. This increase in the field of vision is highly helpful since it enables the network to take into account the larger context in addition to the surrounds of a pixel or feature. Consider By taking into account more information from the image, this enlarged view strengthens the traits that RFB extracts.

The way RFB can strike a balance between emphasizing the center region surrounding a point and maintaining a broad field of vision is what makes it so successful. RFB prioritizes the core portion while keeping an eye on the larger context because the central region surrounding a point frequently contains crucial information and distinguishing characteristics. The capacity of the network to comprehend and identify elements in both photos is crucial in the area of image matching, particularly when working with images taken from different perspectives. Every view offers a different viewpoint on the scene, and different views may make some elements or objects harder to see or distinguish.

This is when having a large grid view scope-like RFB's becomes advantageous. A greater portion of the input images may be covered by expanding the network's field of vision. This implies that information or characteristics that are not visible directly in an image may still be caught in the range of vision that has been enlarged, enabling the network to identify them.

in producing a feature map that is more thorough. In addition to the main features, this map also includes the surrounding context, which is particularly crucial for matching images taken from various viewpoints. In theory, RFB's capacity to widen the network's sphere of influence aids in overcoming the constraints imposed by disparate viewpoints. By doing this, it is made sure that the network can identify elements and objects that might be unclear or dissimilar between the two images. This method helps the network obtain a more thorough understanding of the scene from various angles, which in turn leads to a more accurate and reliable image matching process.

3.2 RFB feat

The detection and description method is applied to key parts and the creation of descriptive maps in this article, as shown in Figure 5. It combines the identification of discrete interest points and the creation of pertinent feature descriptors, two crucial image processing steps for matching. The objective is to obtain important visual information while preserving an efficient computational process by first identifying pertinent keys and simultaneously describing them with corresponding descriptors. Using this technique, specific features in photos can be automatically recognized and displayed. This provides a strong basis for attaining the precision needed for computation, which is essential for the success of research.

There are various benefits to integrating interest point detector and descriptor into the network.

- Allows for end-to-end training and concurrent

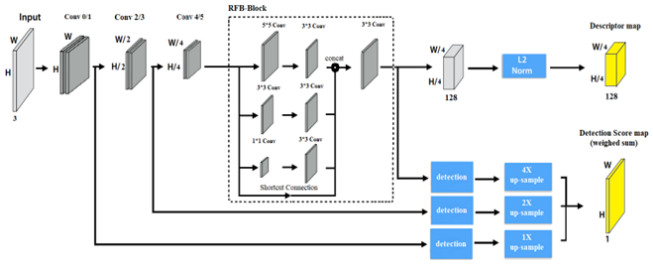

Figure 5: RFB feat

learning of the best representations for both tasks by the network.

- Guarantees intrinsic alignment between the detector and the descriptor, improving matching precision.
- Removes the requirement for distinct steps for detection and description, hence reducing computational complexity.

3.3 Interest Points Detection

In computer vision, interest detection can be achieved through a variety of techniques, one popular approach being the multiscale method. U-Net-like networks, comprising of an encoder and a decoder connected at a bottleneck point, are used in keypoint selection techniques. These networks make it easier to connect encoder and decoder elements and extract features.

Figure 6: multi level detection

This paper employs a unique multi-level approach, just like Figure 6. Three different parts of the network are identified as key points, and these key points are located using the feature maps generated by the chosen layers. This multi-level approach adds to the network's accuracy and robustness by offering a useful tool for locating important points in an image at various scales and levels of detail. This method enables the consideration of both high-level details pertaining to each major point and low-level details like corners and edges. Each interest point is given a final score, and points that are higher than a predefined threshold are then classified as interest points. The robustness and efficacy of interest point identification are increased by this multifaceted scoring and selection process.

The input images are used by the rfb_feat deep neural network to extract features. The network is trained to predict the location, scale, and orientation of feature points, among other fundamental properties. This feature improves the geometric invariance in the extracted features predictability.

A local score is produced for each location (i, j) in the descriptor map, much like formula 1. The feature response value is first exponentiated, and it is then normalized using the total of the exponential responses of the adjacent pixels.

$$a_{ij}^c = \frac{\exp(y_{ij}^c)}{\sum_{(i',j')\in N(i,j)} \exp(y_{i'j'}^c)} \tag{1}$$

In Formula 2, a point is determined for each location in terms of channel. This score is obtained by dividing the feature response value by the maximum feature response across all channels at that particular location.

$$B_{ij}^c = \frac{y_{ij}^c}{\max_t y_{ij}^t} \tag{2}$$

Finally, in formula 3, the final detection score is calculated by choosing the highest score in terms of channel. This score is further influenced by the local score on all channels. This integration combines information from both feature and peak response measurements, ultimately resulting in a single score for each potential interest point.

$$S_{ij} = \max_t (a_{ij}^c \beta_{ij}^c) \tag{3}$$

A subset of features is selected as interest points. This selection process considers factors such as response strength and spatial distribution. Using a non-maximal suppression algorithm ensures the selection of the most salient interest points while distributing good coverage throughout the image. As a result, the score map is generated through a weighted calculation, formula 4.

$$\hat{s} = \frac{1}{\sum_l w_l} \sum_l w_l s^l \tag{4}$$

An important step involves modifying the characteristics of the selected key points, especially modifying their location and scale. This fine-tuning is done through a sub-pixel correction algorithm. This fine tuning helps to increase the accuracy of interest point localization, which is important.

4 Training

In the context of detection, we aim for interest points to be repeatable under varying viewpoints or illumination. For description, we seek descriptors to be unique, preventing mismatches. Given a pair of images (I,I') and their corresponding point matches C, the loss function minimizes the distance between matching descriptors while maximizing the distance to other irrelevant descriptors.

The loss function is defined as follows:

$$Loss(l, l') = \frac{1}{|C|} \sum_{c \in C} \frac{\hat{s}_c \hat{s}_{c'}}{\sum_{q \in C} \hat{s}_q \hat{s}_{q'}} M(f_c, f_{c'}) \quad (5)$$

where $\hat{s}$ and $\hat{s'}$ are the soft detection scores at points in I and I', respectively, C is the set of all correspondences between I and I', and q and q' are the corresponding descriptors.

To calculate $M(f_c, f_{c'})$, we identify four distinct descriptors: and from the same location, (the closest descriptor to in I), and (the closest descriptor to in I').

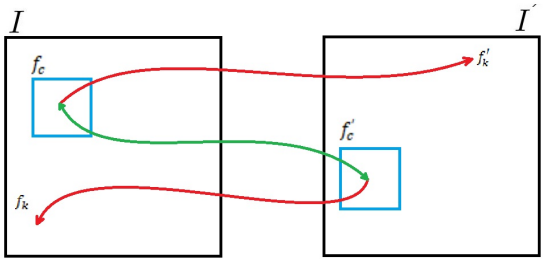

Figure 7: multi level detection

$M(f_c, f_{c'})$ is computed as:

$$M(f_c, f_{c'}) = [D(f_c, f_{c'}) - m_p] + [m_n - \min_{k \neq c}(\min_{k \neq c} D(f_c, f_k), \min_{k \neq c} D(f_k, f_{c'}))] \quad (6)$$

The Euclidean distance between two descriptors is calculated using D. The parameters m_p *and* m_n are set to 0.2 and 1.0, respectively, for positive and negative cases.

5 Experimental Evaluation

5.1 Dataset

The GL3D dataset[11] is a valuable resource for training image matching networks, providing essential information such as camera positions and depth maps[14]. Camera poses describe the camera's location and orientation, while depth maps represent the distance to objects in the scene. This data is crucial for training networks to learn how to predict camera positions and depth maps from a pair of input images. By learning from this data, the networks can develop the ability to understand the spatial relationships between objects in the scene and accurately estimate the camera's perspective.

The GL3D dataset[11] is particularly valuable because it provides a large and diverse collection of image pairs with corresponding camera poses and depth maps[8]. This allows networks to be trained on a wide range of scenarios, improving their generalization ability. Additionally, the high quality of the data in the GL3D dataset[11] ensures that the networks can learn from accurate and reliable information, leading to more robust and accurate models.

We created an extensive test dataset in order to effectively evaluate how well various networks and techniques handled a range of drone and satellite viewing angles. This dataset is made up of multiple image pairs that cover a variety of situations and environments. This article's comparisons and results are based on how well different networks and techniques worked with this carefully chosen test dataset.

We were able to determine the most efficient techniques for image matching and analysis in the context of drone and satellite photography by using this test dataset, which gave us important insights into the advantages and disadvantages of each strategy.

5.2 Evaluation

Repeatability and correct rate are two critical parameters for image matching. The correct rate calculates the proportion of exact matches among all matches generated by the algorithm. It is determined by dividing the total number of matches by the number of correct matches. Conversely, repeatability measures the accuracy of matches by examining the minimum critical parts in both images. It is calculated by multiplying the minimum key points of both photos in the number of exact matches.

5.2.1 Correct Rate

A key parameter to evaluate the accuracy of the image matching algorithm is its correct rate. Formula 6 gives information about how well the system can find and match similar features in different photos. We evaluate the accuracy of feature identification and matching by calculating the percentage of exact matches to all matching cases. While a lower correct rate indicates that there may be an error or inconsistency in the matching process, a higher correct rate indicates a more accurate image matching algorithm.

$$Correct\ rate = \frac{Correct\ matches}{All\ matches} \quad (7)$$

Method	Correct rate	Rep * 100
SIFT[7]	0.07	0.28
D2net[4]	0.47	1.75
Aslfeat[9]	0.30	0.59
Loftr[12]	0.473	1.2
RFB-feat	**0.655**	**1.84**

Table 1: comparision between our method and existing methods

5.2.2 Repeatability

Repeatability focuses on the accuracy of the matches that the algorithm produces. Since this serves as a standard comparison point, it considers the lowest possible key points in both images. We may evaluate the algorithm's Repeatability of matching results between different photos, similar to formula 6, by dividing the number of correct matches by this minimum value. If the algorithm consistently matches features, even when there are fewer key points to compare, it has a better repeatability score.

$$Repeatability = \frac{Correct\ matches}{\min(n_1, n_2)} \tag{8}$$
$$n_i = number\ of\ keypoint\ in\ image\ i$$

Some tasks may require high correct rate but can tolerate less repeatability, while other tasks may require high repeatability but can tolerate less accuracy.

Comparison The integration of receptive field blocks (RFBs) into our image matching network has significantly enhanced performance, particularly in handling diverse viewpoints. These RFBs effectively address the challenges posed by eye-captured images, improving discriminative measures and leading to more accurate matching results. While RFBs excel in managing viewpoint changes, they exhibit limitations in dealing with image rotations. Further research is needed to address this aspect and ensure the network's robustness across a broader range of transformations.

A comprehensive comparison between our proposed method and established techniques like SIFT[7], d2Net[4], Aslfeat[9], and LOFTR[12] demonstrates the significant superiority of our approach. Our method consistently outperforms these existing methods in terms of accuracy, providing compelling evidence of its enhanced performance and efficiency.

Key Findings from Our Analysis:

- **SIFT's Limitations:** Despite its value, SIFT[7] struggles with viewpoint changes, often producing inconsistent results.
- **D2Net's Potential:** D2Net[4] shows promise and can achieve good performance in certain scenarios, but is generally outperformed by our proposed method.
- **LOFTR's False Positives:** LOFTR[12] generates a large number of matches but suffers from a high false positive rate, often producing incorrect matches even when the input images do not correspond.
- **Aslfeat's Promising Results:**Aslfeat[9] offers promising results similar to our method, but our approach excels in both the quantity and quality of matches, demonstrating superior performance across various criteria.

Figure 8: Result of RFB feat

6 Discussion

The RFB component is particularly effective in managing different viewpoints, demonstrating its superiority in scenarios involving images captured from various angles or perspectives. Its adaptability to varying viewpoints enables it to create feature maps that capture the essential characteristics of the scene, facilitating more accurate matching. This capability is especially advantageous when working with remote sensing images,

which often exhibit significant perspective changes due to different viewing angles and sources. RFB's flexibility in capturing the nuances of these diverse perspectives allows it to extract meaningful information from the images, even in challenging conditions where traditional methods might struggle.

References

[1] A robust feature matching strategy for fast and effective visual place recognition in challenging environmental conditions.

[2] Y. Bai. Enhancing image matching for 3d reconstruction via deep learning with key. net and adalam. In *2024 6th International Conference on Image, Video and Signal Processing*, pages 67–72, 2024.

[3] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[4] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-net: A trainable cnn for joint detection and description of local features.

[5] J. Lavín-Delgado, J. F. Gómez-Aguilar, D. Urueta-Hinojosa, Z. Zamudio-Beltrán, and J. Alanís-Navarro. An efficient technique for object recognition using fractional harris–stephens corner detection algorithm. *Multimedia Tools and Applications*, 83(8):23173–23199, 2024.

[6] S. Liu, D. Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 385–400, 2018.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[8] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *European Conference on Computer Vision (ECCV)*, 2018.

[9] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020.

[10] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[11] T. Shen, Z. Luo, L. Zhou, R. Zhang, S. Zhu, T. Fang, and L. Quan. Matchable image retrieval by learning from surface reconstruction. In *The Asian Conference on Computer Vision (ACCV*, 2018.

[12] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. Loftr: Detector-free local feature matching with transformers.

[13] C. Suryaraj and M. Geetha. Block based motion estimation model using cnn with representative point matching algorithm for object tracking in videos. *Expert Systems with Applications*, page 124407, 2024.

[14] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016.

[16] L. Zhao, H. Wang, Y. Zhu, and M. Song. A review of 3d reconstruction from high-resolution urban satellite images. *International Journal of Remote Sensing*, 44(2):713–748, 2023.