# A Simple Methodology for 2D Reconstruction Using a CNN Model

Armando Levid Rodríguez-Santiago, José Anibal Arias-Aguilar, Alberto Elías Petrilli-Barceló(✉), and Rosebet Miranda-Luna

Graduate Studies Division, Universidad Tecnológica de la Mixteca, Km. 2.5 Carretera a Acatlima, 69000 Huajuapan de León, Oaxaca, Mexico
levid.rodriguez@gmail.com, {anibal,petrilli,rmiranda}@mixteco.utm.mx
http://www.utm.mx

**Abstract.** In recent years, Deep Learning research have demonstrated their effectiveness in digital image processing, mainly in areas with heavy computational load. Such is the case of aerial photogrammetry, where the principal objective is to generate a 2D map or a 3D model from a specific terrain. In these topics, high-efficiency in visual information processing is demanded. In this work we present a simple methodology to build an orthomosaic, our proposal is focused in replacing traditional digital imagen processing using instead a Convolutional Neuronal Network (CNN) model. The dataset of aerial images is generated from drone photographs of our university campus. The method described in this article uses a CNN model to detect matching points and RANSAC algorithm to correct feature's correlation. Experimental results show that feature maps and matching points obtained between pair of images through a CNN are comparable with those obtained in traditional artificial vision algorithms.

**Keywords:** Deep Learning · CNN · 2D reconstruction · Aerial images

## 1 Introduction

Image stitching produces a mosaic that corresponds to a set of images taken from one or several cameras which overlap and are joined in a single image [6]. In the generation of this mosaic several computer vision techniques are used. We worked with aerial images and computer vision strategies combined with photogrammetry techniques.

The stitching process is usually made with traditional computer vision methods as shown in Fig. 1a. It begins with a drone flight plan to image acquisition of a selected area. Then placeholders with georeferenced points are added over a map as well as flight height and overlapping percentage between each pair of acquired images. Usually a mobile application is configured with these specifications to acquire the information autonomously. Some popular free apps to help in this stage are Pix4D and DroneDeploy.


K. M. Figueroa Mora et al. (Eds.): MCPR 2020, LNCS 12088, pp. 98–107, 2020.
https://doi.org/10.1007/978-3-030-49076-8_10

Then, an image processing stage is performed. It begins with feature extraction and continues with the identification and relationship of similar features between images in overlapping areas [8,17]. Key points operators [18] are mainly used as feature extraction algorithm. They use radiometric features such as points, edges, corners, etc. that can be detected in adjacent images under normal capture conditions. They are not robust to inclination, rotation, scale or lighting changes, however, in aerial images these conditions does not occur very often. To deal with these conditions, computer vision techniques are a good option, being one the most popular Scale Invariant Feature Transform (SIFT) algorithm. SIFT [19,20] processing has four steps:

1. Scale Space Extrema Detection: identify a location and scales key points using scale space extrema in the DoG (Difference-of-Gaussian) functions with different values of standard deviation.
2. Key point Localization: key point candidates are localized and refined by eliminating low contrast points.
3. Orientation Assignment: orientation of key point is obtained based on local image gradient.
4. Description Generation: compute the local image descriptor for each key point based on image gradient magnitude and orientation at each image sample point in a region centered at key point.

These steps generate a 128-dimension key point descriptor.

Once an interesting group of features have been extracted, the next step to do is features correlation or features correspondence. It consists of vector descriptor comparison. Several methods can be used: quadratic search, kd-tree data structure, etc. Erroneous correspondences (outliers) presented in the correlation are eliminated from estimation through fundamental matrix or essential matrix (if the internal parameters of the camera are known) [2,4]. This is difficult because internal parameters of the camera very often are unknow. Therefore, other strategies are found in the literature such as LMS (Least-Median-Square) and MAPSAC, however, one the most used strategy is RANSAC (Random Sample Consensus) [9,16], which is an iterative algorithm to determine a fundamental matrix. RANSAC is essentially composed of two steps that are iteratively repeated [27]:

– Hypothesize. First minimal sample sets (MSSs) are randomly selected from the input dataset and model parameters are computed using only elements of the MSS. Cardinality of MSS is the smallest sufficient to determine the model parameters (as opposed to other approaches, such as least squares, where parameters are estimated using all data available, possibly with appropriate weights).

– Test. In the second step, RANSAC checks which elements of the entire dataset are consistent with the model instantiated using parameters estimated in the first step. The set of such elements is called consensus set (CS).

RANSAC terminates when the probability of finding a better ranked CS drops below a certain threshold. In their original formulation the ranking of CS was its cardinality (i.e. CSs that contain more elements are ranked better than CSs that contain fewer elements).

This is the best option to adjust the correspondences and eliminate features that do not meet a reference value. The final stage is to build an orthomosaic with all previously performed procedures. In this step, computer vision techniques are used to join all photographs into one.

It should be noted that the most complex task is orthomosaic generation. It is extremely complex, however, recent research has demonstrated great efficiency of convolutional neural networks (CNN) in digital image processing [1,11,22], that is why this investigation uses a CNN to built an orthomosaic from Technological University of the Mixteca (UTM) campus with aerial images obtained from an Unmanned Aerial Vehicle (UAV).

## 2 Related Work

Aerial photogrammetry is a procedure to obtain plans for large land areas by means of aerial photographs [3]. The result is a 2D map or a 3D terrain model. To do this we need to apply computer vision techniques and algorithms.

Research has been carried out with the purpose of perform improvements such as the work of [13] where SIFT algorithm is used to feature extraction and digital surface models (DSM) were generated from UAV images in high resolution. Similarly, in [15], the author proposes to use new algorithms for surface reconstruction. These approaches demand still high computational complexity.

Nevertheless, recent research has included studies in Deep Learning approaches such as presented in [5,10,24,26] where they perform image pairing and 3D reconstructions using deep neuronal network techniques. Obtained results are quite acceptable, however, proposed models are very complex and often require additional information from external sensors [14].

## 3 Methodology

Our approach for orthomosaic reconstruction consists in replacing traditional digital image processing techniques with a Convolutional Neuronal Network. We propose two stages: feature extraction with a neuronal convolutional network and correspondences correction. Methodology is shown in Fig. 1. We can see the main change between both approaches for obtaining an orthomosaic: procedure shown in (Fig. 1a) involves the use of classical computer vision techniques for digital image processing, and we propose to change almost all of these complex processes with a single convolutional neural network as shown in (Fig. 1b).

The process mentioned above for obtaining and correlating features between images is a complex stage, with the extraction of features being one of the most difficult. However, Noh et al., and Teichmann et al. presents a proposal for feature extraction, DEep Local Feature (DELF). This model is particularly useful for

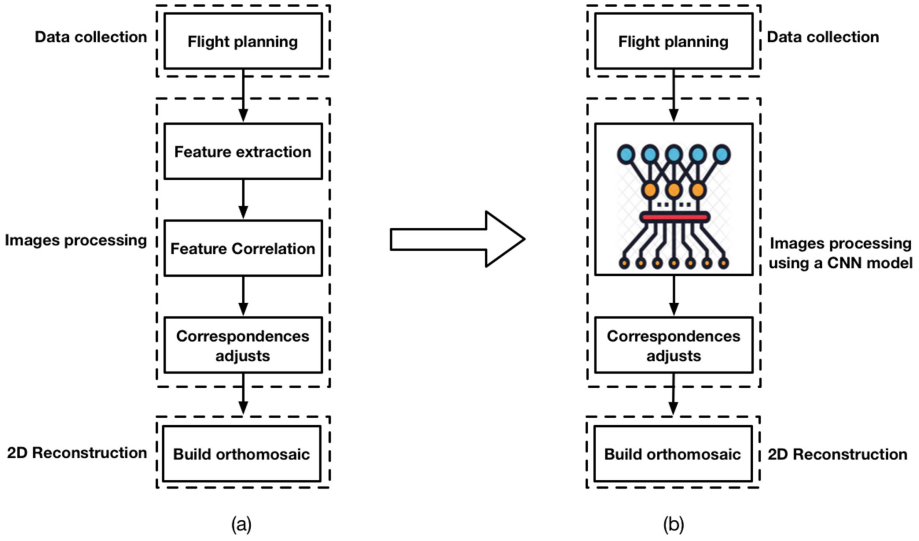



**Fig. 1.** Traditional methodology is shown on the left (a). We can see that it consists of five stages to obtain an orthomosaic. The most complex steps are those of digital image processing. On the right (b), our proposed methodology replaces the most complex stages of digital image processing with a CNN model

large-scale instance-level image recognition and to index image regions. This model detects and describes semantic local features which can be geometrically verified between images showing the same object instance [21,25]. DELF use a ResNet50 [12] model trained on ImageNet Dataset [23] as a baseline to feature extracting layers trained with a classification loss. Features are localized based on their receptive fields, which are computed by means of convolutional and pooling layers of a Fully Convolutional Network (FCN). Code is provide in Tensorflow for building a model which could be used to train models for other applications.

Then, based on the DELF model and Noh's work, we used our new dataset including 880 aerial images rescaled to 250 × 250 pixels (Table 1). This dataset was created by capturing multiple aerial images of the entire university campus. Due to the terrain conditions of the campus, a minimum safe flight-height of 100 m and a maximum of 150 m were selected. Overlapping percentages among captured images were considered with two configurations, the first set with 30% both longitudinally and transversely, and the second 50% in both directions.

**Table 1.** UTM campus image dataset. This is the way images have been organized, so that they can be used to adjust the CNN model.

| Height\Overlaping | 30% × 30% | 50% × 50% |
|---|---|---|
| 100 mts. | 200 | 400 |
| 150 mts. | 100 | 180 |
| Total | 300 | 580 |

As in Noh's work, we used the original pre-trained ResNet50 model with ImageNet as a base, and we performed a fine-tuning procedure to improve our local descriptors. We employed a FCN at the output of $conv4\_x$ convolutional block of ResNet50. This output was adjusted in a way that it can be considered like a feature extraction and key points matching machine on aerial images and also this adjusted model could be a replacement for other key point detectors and descriptors. Neural model is shown with detail in Fig. 2. We can see a pair of images supplied to the CNN Pipeline. Internally, DEFL model is truncated at the output of the feature map to be connected to a FCN for finding vector descriptors of the input images and with it a geometric correction applied with the RANSAC algorithm for finally create the orthomosaic.

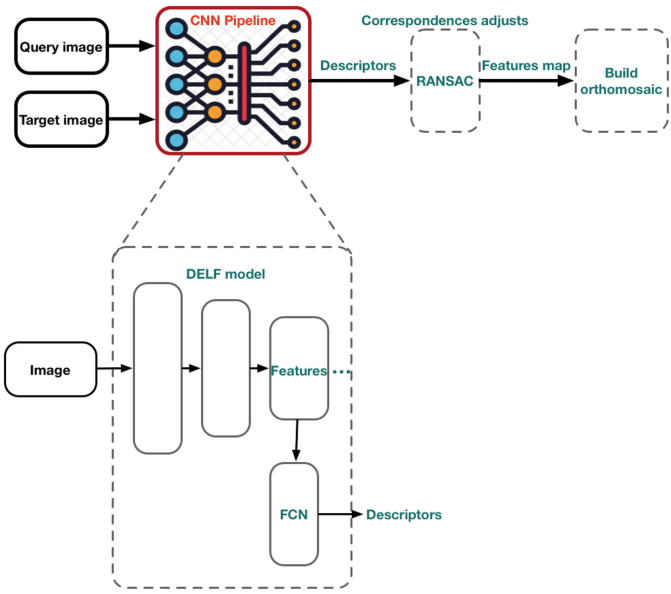



**Fig. 2.** CNN Pipeline. It uses the DELF model with ResNet50 trained on ImageNet.

After finding correspondences, outliers must be eliminated from estimation through the fundamental matrix since internal parameters of the camera are unknown. However, many of correspondences are faulty and estimating the parameter set with all coordinates is not enough. Therefore, RANSAC algorithm is used on top of the normal model to robustly estimate the parameter set by detecting outliers. The main objective is to determine geometric transformation between both images, that is, to define the fundamental matrix that relates two views of planar target. RANSAC algorithm can help computing the homography matrix [7,16] starting with acquired correspondences. Then, we use RANSAC with the feature vectors extracted from images as a set of observed

data points. Moreover, as the model that can be fitted to data points we used an affine transform model. We end up having a set of source and destination coordinates which can be used to estimate the geometric transformation between both images and building an orthomosaic with all previously performed procedures.

## 4 Experimental Results

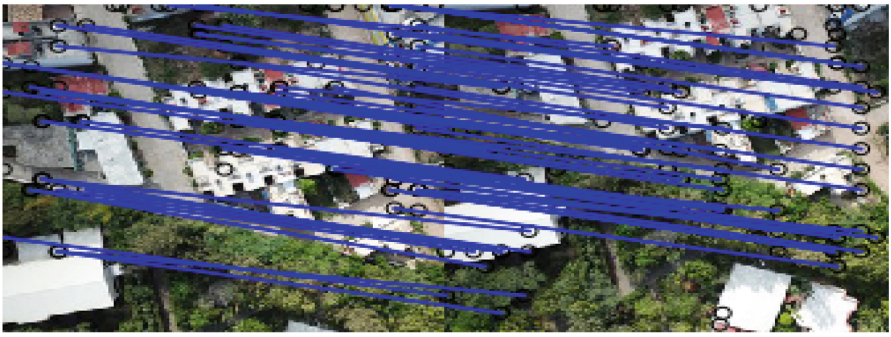
In order to evaluate our proposal we analyze qualitative results in two stages. In the first one, we determine the efficiency of our process for feature extraction and matching features in the dataset. In the second experiment, we check results for orthomosaic generation.

Our methodology achieves the goals to obtain a feature map by training a CNN model that encodes learning to select features for the matching task. Figure 3 shows feature correspondences between a pair of images from our database. It successfully matches them in a challenging environment as the UTM campus. It could include changes in contrast, sharpness, brightness and rotations in the images. Moreover, results shown that RANSAC algorithm improves correction of correspondences obtained in (Fig. 3a and 3b). Furthermore, matching points are acceptable and can be compared to those of SIFT algorithm, showing equivalent results Fig. 3c. It is a good benchmark by the SIFT algorithm robustness.

The described process permit to obtain acceptable feature maps to pair aerial images. In Fig. 4, it is shown an example of 2D reconstruction with high-resolution aerial images. In this experiment we used 100 images to perform an orthomosaic reconstruction. This images cover approximately an area of 100 $km^2$ from UTM campus (the campus has around 104 hectares). Some areas do not have constructions (Fig. 4a) and other have buildings (Fig. 4b). Resulting orthomosaics present high-definition details that are acceptable and suitable to be employed for several purposes.

On the other hand, we analyze the similarity of the resulting orthomosaics versus a manual reconstruction, an aerial image that covers the same area and an orthomosaic obtained from Pix4DMapper. We use Euclidean distance to determine the similarity between each one (smaller distance, greater similarity). The results are shown in Table 2. It shows that the resulting orthomosaics with our methodology are similar to those obtained by a traditional or manual process, but with high-definition details and less processing time.

(a)



(b)

(c)

**Fig. 3.** Figures show feature maps obtained with real images. (a) Matching points without geometric correction. (b) Geometric correction with RANSAC and (c) Results obtained with SIFT.

(a) (b)

**Fig. 4.** 2D reconstructions of UTM campus. We present two examples of an orthomosaic reconstruction using 100 high definition images at an altitude of 150 m and 50% of overlapping (a) without buildings and (b) with buildings.

**Table 2.** Comparison between our resulting orthomosaics and other reconstructions. This table shows the Euclidean distance as a measure of similarity between orthomosaics. Manual reconstruction was performed with images at 50% of their original resolution. Aerial image was taken at twice the reference height. Pix4DMapper's orthomosaic only shows 75% of total established area.

| Resulting orthomosaic vs | Euclidean distance |
|---|---|
| Manual reconstruction | 11.564167 |
| Image at twice of the reference height | 16.99647 |
| Orthomosaic from Pix4DMapper | 20.794645 |

## 5 Conclusions

In this work a simple methodology to built orthomosaics using aerial images is presented. This study focuses on verify the methodology that uses a deep neuronal network model. Preliminary results generating orthomosaics have been verified qualitatively obtaining feature maps and matching points between images pairs.

Resulting orthomosaics were evaluated using Euclidean distance as a similarity measure. Orthomosaic obtained was compared with: a manual reconstruction, an image captured at a higher height and a reconstruction obtained with commercial software. It is showed that our methodology provides similar results to those obtained as described before but with a high-definition details. Our results are as well comparable with those obtained with traditional computer vision algorithms.

On the other hand, reconstruction of larger areas such as the entire campus of the university with a high-resolution orthomosaic map is being considered for future work.

## References


1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)
2. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J. ACM (JACM) **45**(6), 891–923 (1998)
3. Asale, R., Rae: fotogrametría: Diccionario de la lengua española, November 2019. https://dle.rae.es/fotogrametria
4. Barazzetti, L., Remondino, F., Scaioni, M.: Extraction of accurate tie points for automated pose estimation of close-range blocks. In: ISPRS Technical Commission III Symposium on Photogrammetric Computer Vision and Image Analysis (2010)
5. Chen, Y., Liu, L., Gong, Z., Zhong, P.: Learning CNN to pair UAV video image patches. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(12), 5752–5768 (2017)
6. Cheng, Y., Xue, D., Li, Y.: A fast mosaic approach for remote sensing images. In: 2007 International Conference on Mechatronics and Automation, pp. 2009–2013. IEEE (2007)
7. Dung, L.R., Huang, C.M., Wu, Y.Y., et al.: Implementation of RANSAC algorithm for feature-based image registration. J. Comput. Commun. **1**(6), 46–50 (2013)
8. Escalante Torrado, J.O., Porras Díaz, H., et al.: Ortomosaicos y modelos digitales de elevación generados a partir de imágenes tomadas con sistemas uav. Tecnura **20**(50), 119–140 (2016)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
10. Ghamisi, P., Yokoya, N.: Img2dsm: height simulation from single imagery using conditional generative adversarial net. IEEE Geosci. Remote Sens. Lett. **15**(5), 794–798 (2018)
11. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: learning global representations for image search. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VI. LNCS, vol. 9910, pp. 241–257. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_15
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Li, J., Ai, M., Hu, Q., Fu, D.: A novel approach to generating DSM from high-resolution UAV images. In: 2014 22nd International Conference on Geoinformatics (GeoInformatics), pp. 1–5. IEEE (2014)
14. Li, S., Zhu, Z., Wang, H., Xu, F.: 3D virtual urban scene reconstruction from a single optical remote sensing image. IEEE Access **7**, 68305–68315 (2019)
15. Li, T., Hailes, S., Julier, S., Liu, M.: UAV-based SLAM and 3D reconstruction system. In: 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 2496–2501. IEEE (2017)

16. Li, X., Liu, Y., Wang, Y., Yan, D.: Computing homography with RANSAC algorithm: a novel method of registration. In: Electronic Imaging and Multimedia Technology IV, vol. 5637, pp. 109–112. International Society for Optics and Photonics (2005)
17. Lingua, A., Marenchino, D., Nex, F.: Automatic digital surface model (DSM) generation procedure from images acquired by unmanned aerial systems (UASS). RevCAD J. Geodesy Cadastre **9**, 53–64 (2009)
18. Lingua, A., Marenchino, D., Nex, F.: A comparison between "old and new" feature extraction and matching techniques in photogrammetry. RevCAD J. Geodesy Cadastre **9**, 43–52 (2009)
19. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
21. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3456–3465 (2017)
22. Radenović, F., Tolias, G., Chum, O.: CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 3–20. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_1
23. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Tang, J., Folkesson, J., Jensfelt, P.: Geometric correspondence network for camera motion estimation. IEEE Robot. Autom. Lett. **3**(2), 1010–1017 (2018)
25. Teichmann, M., Araujo, A., Zhu, M., Sim, J.: Detect-to-retrieve: efficient regional aggregation for image search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5109–5118 (2019)
26. Weerasekera, C.S., Latif, Y., Garg, R., Reid, I.: Dense monocular reconstruction using surface normals. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 2524–2531. IEEE (2017)
27. Zuliani, M.: Ransac for dummies. Vision Research Lab, University of California, Santa Barbara (2009)