VIDEO IMAGE TARGET RECOGNITION AND GEOLOCATION METHOD FOR UAV BASED ON LANDMARKS

Y. Zhang, C. Lan[*], Q. Shi, Z. Cui, W. Sun

Information Engineering University, Zhengzhou 450000, China - (zhangyx6656, lan_cz, hills1, zzch0908, wsun9)@163.com

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: Landmarks, Target Detection, Target Geolocation, GPS-denied Environment, UAV Video

ABSTRACT:

Relying on landmarks for robust geolocation of drone and targets is one of the most important ways in GPS-denied environments. For small drones，there is no direct orientation capability without high-precision IMU. This paper presents an automated real-time matching and geolocation algorithm between video keyframes and landmark database based on the integration of visual SALM and YOLOv3 deep learning network method. The algorithm mainly extracts the landmarks from the drone video keyframe images to improve target geolocation accuracy, and designs different processing scheme of the keyframes which contains rich and spare landmarks. For feature extraction matching, we improved ORB feature extraction strategy, and obtained a more uniformly distributed feature points than original ORB feature extraction. In the three groups of top-down drone video images experiments, the 100m, 200m, and 300m of the case were carried out to verify the robustness of the algorithm and being compared with GPS surveying data. The results show that the features of keyframe landmarks in the top-down video images within 300m are stable to match the landmark database, the geolocation accuracy is controlled within 0.8m, and it has good accuracy.

1. INTRODUCTION

The ability to accurately detection and orientation ground targets is very important in intelligence collection, surveillance, and reconnaissance (ISR) missions using UAVs (Kwon, 2012a). Global Navigation Satellite System (GNSS) such as GPS and BeiDou enable the high accuracy positioning of UAVs (Tahar, 2016a), which serves as a fundamental task for general UAV missions. However, due to the fact that GNSS signals are broadcasted from satellites, they are easily affected by weather conditions, and are even more easily to be jammed or spoofed, making GNSS based positioning schemes far from reliable for UAV missions in harsh environments and hostile terrains (Lee, 2015a, Liao, 2015a). On the other hand, for inertial positioning, the large drones with high-precision Inertial Measurement Unit (IMU) equipment have low visual positioning accuracy, and small reconnaissance drones do not have high-precision IMU equipment, so it does not have direct geolocation capability. Not all drones can be equipped with expensive high-precision equipment such as POS, IMU and other devices for acquiring their own pose and position data in real-time, and the flight environment is often complex and variable. Strong background noise and radio signal interference would limit the drone's GPS signals, which brings great difficulties in acquiring itself and the target geo-location. The task is especially challenging when GPS signals are not available in GPS-denied environments where GPS signal is not available, jammed or too weak to be used reliably. In order to realize the timely and effective geolocation of the target and self of the drone in an unknown environment, one must use other cues to geo-locate objects through registration of those objects in that environment (Shih-Ming, 2012a).

Numerous researches have been proposed to address the image registration problem that could deal with failing in GPS-denied environments. The ability to correlate two images of the same location but acquired from different sources is challenging. In this typical image registration problem, the challenges that arise when dealing with UAV image and reference image registration can be attributed to: (1) different camera position, rotation, resolution, scale, translating, different sensors and illumination conditions during the image acquisition phase resulting in different object appearance as well as occlusion problems that confuse feature-based image registration, (2) dissimilarity in camera intrinsic parameters introduces photogrammetric differences between the images pair, and (3) difference in image acquisition history may result in mismatch between the image pair due to objects appearing/disappearing making registration more difficult. (Nassr, 2018a).

Although there are many works on image landmark recognition and matching for target geolocation, it is a complex system and most works on this area are in a more heuristic and less practical approach (Filho, 2015a). The image landmark matching aims at finding scale invariant feature consistenting with the reference image in the aerial images that are captured during the drone flight by an onboard camera. After feature matching, the drone location is estimated in real time in order to accomplish targets geolocation (Deangelo, 2016a). This paper presents a visual localization method that enable geolocation the targets in GPS-denied environment (Fig 1). The contribution of this paper is a novel method that optimally combining landmarks matching with deep learning network model between reference image and UAV video image for targets geolocation when GPS signals are not available. The method includes three parts. The first part uses convolutional neural network to recognize and detect the moving targets of the UAV video images, the second part matches the landmarks of the UAV video keyframe and reference image, and the third part computes the geolocation of the ground detected targets. We implement the method using YOLOv3 network model and ORB

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-2-W16-285-2019 |

feature matching. In section 2, we describe the algorithms and implementation details of the entirely framework flow. In section 3, we present the experimental details and results. Finally, we conclude the paper with summary remarks in section 4.

2. RELATED WORKS

Even though landmark recognition and matching is not a new subject on the literature for ground target positioning (Farag, 2004a), the approach for UAV image registration is not well explored yet, mostly because of its complexity and high real-time requirements on precision and computer processing (Silva Filho, 2016a). However, automatic image registration servers as a fundamental part for high accuracy geolocation resolving (Liu, 2018). To extend the matching ability of handling large scale variations, abundant works on landmark recognition and matching for target geolocation takes on the results of already developed object-recognition algorithms and adapt them to the different aerial circumstances.

Feature based algorithms, such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004a), Oriented FAST and Rotation BRIEF (ORB) (Rublee, 2011a), AKAZE (Alcantarilla, 2011a), have changed the object recognition field of study (Li, 2015). In (Lee, 2010a) the method first extracts feature points from the image data taken by a monocular camera using the SIFT algorithm. The system selects landmark feature points that have distinct descriptor vectors among the feature points, calculate those points location and store them in a database. Based on the landmark information, the current position of the UAV is estimated. It considers as a landmark just the exact feature point instead of an object. This method has been used for indoor applications, which is a controlled environment. In outdoors flights, this application could not be used properly because the amount of similar features would result in a high rate of false positive encounters.

Some methods proposed recognizing image descriptors of local intensity patterns to register successive images such as the Kanade-Lucas-Tomasi feature tracker (KLT) (Vivet, 2011a) which is one of the optical flow techniques (Baker, 2004). Here, both KLT and optical flow have been used in many image registration applications under similar quality (Rebiere, 2008a). In (Kwon, 2012a) proposed a new method to compute the UAV attitude and locate mobile ground targets using ground landmarks obtained from SIFT features. In (Lin, 2007a) have proposed an UAV-based image registration system, SIFT features are used for consecutive UAV image registration. However, for a realistic application, the quality mismatch would appear between the UAV image and the reference image, this will greatly degrade the performance of image registration.

Besides, many works have focused on aggregation methods of local features, which include popular techniques such as VLAD (Jegou, 2011a) and Fisher Vector (FV) (Jegou, 2012a). The main advantage of such global descriptors is the ability to provide high-performance image retrieval with a compact index. For similarity measurement, Bag-of-Features (BOF) model is widely used in image retrieval context (Nister, 2006a, Philbin, 2007a, Sivic, 2003a). However, it is generally very hard for flat BoF model to distinguish images with large scale differences from the negative ones due to lack of overlap.
In the past few years, several global descriptors based on CNNs have been proposed to use pretrained (Babenko, 2014a, Tolias, 2015a) or learned networks (Gordo, 2016a, Radenovic, 2016a). CNNs have also been used to detect, represent and compare local image features. In (verdie, 2014a) learned a regressor for repeatable keypoint detection. In (Yi, 2016a) proposed a generic CNN-based technique to estimate the canonical orientation of a local feature and successfully deployed it to several different descriptors. In (Yuting, 2017a) proposes a deep learning method to jointly learn the feature representations and similarity metric over the training samples obtained from various imaging conditions. It turned out that CNN feature network obtained significantly better results than the local-feature-based methods and holistic-representation-based methods.

3. GEOLOCATION FRAMEWORK

The core of the proposed framework for real-time drone video image registration and target geolocation can be built as shown in Figure 1. The framework mainly consists of four modules, including a data input module, a landmark matching module, a target detection and recognition module, and a real-time target geolocation module. Each of the following sections presents details of the framework modules.

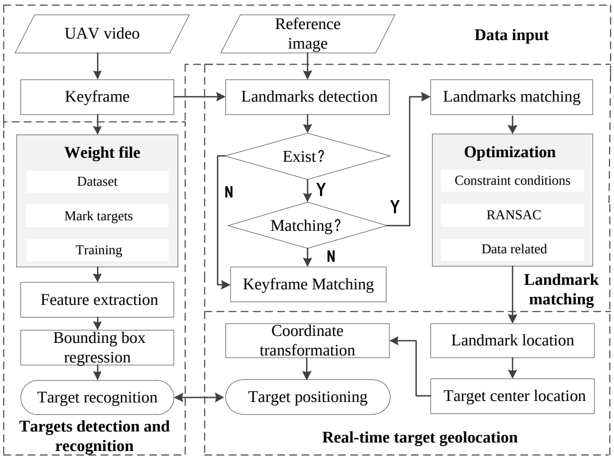


Figure 1. The major modules of proposed framework

3.1 Data Input

The framework processes image from two different sources: UAV video image and reference image. To improve the registration efficiency, we selected keyframe image from the real-time video image for registration.

UAV Video: The framework accepts a video from the UAV denoted as S. From S, we extract S(i) (video keyframe) which we compare to a reference map (M). It is important to note that the initial starting GPS coordinate of the UAV is assumed to be known and can be defined as the center pixel of S(1). This assumption is made based on notion that a UAV cannot be deployed without knowing its location.

Reference Image: The framework uses a reference map M with known GPS bounds. Mainly the process is finding out where S(i) resides in M and subsequently estimating the position of S(i). Using equations (1) and (2), it is possible to calculate a certain GPS coordinate. The opposite is also possible to estimate the latitude and longitude of a pixel using equations (3) and (4) (Nassar, 2018a).

$$pix_x = \frac{(width_{\max} - width_{\min})\,(lon - lon_w)}{(lon_e - lon_w)} \tag{1}$$

$$pix_y = \frac{(height_{\max} - height_{\min})\,(lat - lat_n)}{(lat_s - lat_n)} \tag{2}$$

$$lat = \frac{lat_s + (lat_n - lat_s)\,(pix_x - height_{\min})}{(height_{\max} - height_{\min})} \tag{3}$$

$$lon = \frac{lon_w + (lon_e - lon_w)\,(pix_y - width_{\min})}{(width_{\max} - width_{\min})} \tag{4}$$

Where pix_x, pix_y = latitude and longitude of a pixel
lon_w, lat_n, lat_s, lon_e = bounds of M.

3.2 Landmark Matching

The reference image area used to extract landmarks is much larger than the UAV video frame image, and there are feature rich areas and feature sparse areas on the image. The feature rich areas matching effect is better, but the time consumption is serious, it is difficult to meet the real-time matching geolocation requirements. The feature sparse areas are difficult to match with the reference image landmarks database, resulting in low precision of the transformation matrix of the video frame image to the reference image, which does not meet the accuracy requirement of the moving target geolocation. The algorithm designs the processing scheme of the keyframe which contains rich and sparse landmarks.

For scenario one: keyframes contain rich landmarks. In order to ensure the distribution uniformity of landmarks feature extraction and matching as much as possible, and to improve the accuracy of landmark feature matching, we use the image feature pyramid model to divide the reference image into different levels, and then divide the reference image of each level into $L_i \times L_i$ sub-regions for completing the gridding of different levels of images in the image pyramid model. The number of subregions corresponding to different levels is equal to the square of the level L of the image feature pyramid, the top layer of the image pyramid is the Lth layer, and the bottom layer is the first layer, and then we extract a fixed number ORB feature from each subregion in every level. Next we calculate the ORB feature descriptors of the keyframe and based on the Euclidean distance match these descriptors with the reference image landmark feature descriptors to achieve the absolute position measurement between the UAV and the target (Filho, 2017a).

Figure 2 briefly shows the flow chart of registration from keyframe to reference image. We could provide a sequence of UAV video keyframe images $I_0, I_{1,\cdots,} I_n$, and a reference image M in practical. Here, we assume $H_{i,j}$ denote the homography from Ii to Ij , $H_{i,m}$ denotes the homography from Ii to M (Huang, 2013a, Kwon, 2012a, Lin, 2007a).

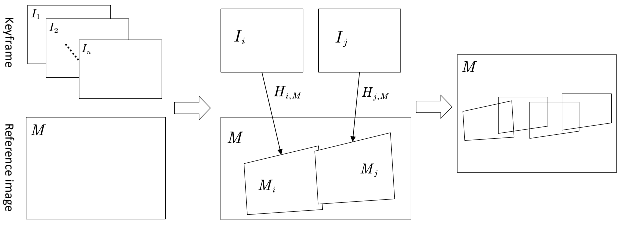


Figure 2. Registration from keyframe to reference image

For scenario two: keyframes contain sparse or even no landmarks. At this time, the keyframe and the reference image are mismatched, and the matching of the real-time video frame with the reference image is abandoned, and the matching of the dynamic keyframe in the video image is utilized to achieve the relative position estimation of the drone and the target. Assume the video image acquired by the drone is represented as *I*, there are I_i, I_j as the adjacent two keyframe images, where $j = i + f$, f represents the frame rate, and its adjustment can adapt to the matching rate of the platform while ensuring real-time performance. Usually, the continuous two keyframe images are an approximate linear smooth transformation process, there is no significant change, so the real-time performance of the ORB feature can be used for feature points extraction and matching to obtain homography matrix between video keyframes. The flow of registering as illustrated in Figure 3.

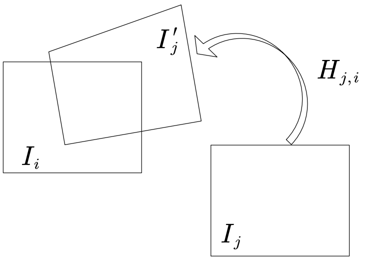


Figure 3. video image dynamic keyframes registration

Let $H_{i,j}$ denote the keyframe image I_i to I_j transform the homography matrix, then

$$H_{i.,j} I_i = I_j \tag{5}$$

Where $H_{i.,j} = K_i R_i R_j^T K_j^{-1}$, then

$$K_i = \begin{bmatrix} f_i & 0 & 0 \\ 0 & f_i & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_i = e^{[\theta_i]_x}, \quad [\theta_i]_x = \begin{bmatrix} 0 & -\theta_{i3} & \theta_{i2} \\ \theta_{i3} & 0 & -\theta_{i1} \\ -\theta_{i2} & \theta_{i1} & 0 \end{bmatrix} \tag{6}$$

Where f = focal length
$\boldsymbol{\theta} = [\theta_1 \;\; \theta_2 \;\; \theta_3]$ is camera rotation angle.

3.3 Targets Detection and Recognition

We choose a deep learning detection model based on region regression to meet the high real-time requirements for target detection. Here we use the YOLOv3 deep learning network model for target recognition and detection on drone video images (Redmon, 2018a), the MS-COCO, ImageNet and CIFAR-10 datasets published on the internet have fewer top-down images for the drone, and directly use the public dataset to train the network, which is difficult to obtain a better target recognition effect. It is not easily found datasets of aerial images with corresponding flight data in the literature results and in order to test the method and analyse the results that this particular drone image dataset was produced. Therefore, we have established a new dataset of drone top-down images with its ground-truth (Yoon, 2009a). To further enhance the performance of the proposed method, we adjusted the network model parameters according to actual situation and then trained. The target detection threshold is set to 0.1. The dataset contains 10 videos with a total of over 1000 images with an average of 40 vehicle per image. We annotate the location and class name of the object in each image by using the Yolo_mark tool. The dataset has 3 classes (car, bus, truck) and over 40000 bounding boxes. Furthermore, the shot scene and shot time of the videos are various. Therefore, the dataset has real data distribution and high diversity. Table 1 shows several attributes of the proposed UAV image dataset. The dataset contains most real world challenges including occlusion, size change, camera motion, motion blur, and dynamic change.

Table1: some attributes of video training dataset

Attribute	Value
Video sequences	10
Class numbers	3
Total Images	Over 1000
Total bounding boxes	Over 40000
Image depth	24 bits
Image resolution	4000 ×3000
Experiment platform	DJI MAVIC 2 drone

3.4 Real-time Target Geolocation

YOLOv3 deep learning network model predicts bounding boxes using dimension clusters as anchor boxes. The network predicts 4 coordinates for each bounding box, t_x, t_y, t_w, t_h. The cell is offset from the top left corner of the image by (c_x, c_y) and the bounding box prior has width and height p_w, p_h, then the predictions correspond to b_x, b_y, b_w, b_h (Redmon, 2018a). we can solve the geolocation of vehicle target by calculating the central point of the detected target according to bounding boxes coordination. In this step, coordinate system transformation is very important. we convert the image pixel coordinates of the vehicle target to the UAV camera coordinates, and then convert to the world coordinates to solve target geolocation. Here, we think the UAV coordinate system and the camera coordinate system are equal and the initial UAV geolocation is known. Considering $F{:}T(pix_x, pix_y) \mapsto G(lat, lon)$ the georeferencing relation from the video image T with the Object Reference Space Image G, and $K{:}T(x,y) \mapsto Q(X,Y)$ the Geometric Transformation that maps the video image T in the query image Q, it is possible to build the geo-referencing transform H, from the query image Q, in which:

$$H{:}Q(X,Y) \xrightarrow{K^{-1}} T(pix_x, pix_y) \xrightarrow{F} G(lat, long) \tag{7}$$

4. EXPERIMENTS AND RESULTS

In this section we present the quantitative and qualitative results of applying this method on several real video images. The experiments developed intended to validate the proposed method to estimate the geolocation of ground detected target. The experimental performed focused on recognizing moving vehicle target geolocation and on how accurate those geographical coordinates were, compared with previously known DOM data. The DOM image was produced from 62 drone images with an image resolution of 5cm. The experiments were performed in a Win 10 PC with a 2.5GHz Intel Core i7, 4th generation, 8GB RAM and NVIDIA GeForce GTX 1050Ti. The programming environment is Visual Studio 2015 and Qt 5.9.3.

4.1 Landmark Extraction

We improve the existing ORB feature detection algorithm by gridding image pyramid in order to extract uniformly distributed landmark points. The main consideration is setting the scale factor (actually 1.2) and the number of pyramid layers (actually 8) when ORB feature is extracted. The original image is reduced by scale factor and scaled by 1/1.2 times, then we obtained an image pyramid with 8 layers by downsampling. we expressed the process of scaled image as $\mathrm{I}' = \mathrm{I}/\mathrm{scaleFactor}$ $(\mathrm{k} = 1, 2, \cdots, \mathrm{nlevels})$. Next, the obtained image is extracted ORB features according to the gridded image block, as shown by the yellow area in the L_1 layer, and recorded, as the figure 4 shows.

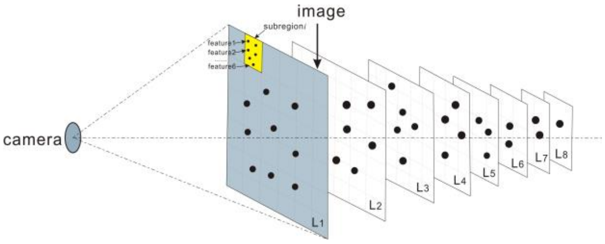


Figure 4. Image feature pyramid model

We select a small area from the reference image and extract original ORB features and the ORB features based on the image feature pyramid model. Figure 5 specifically shows the feature extraction results in two ways. It can be seen that the feature distribution region extracted by the gridding ORB algorithm is more uniform, and the matching experiment is more robust.

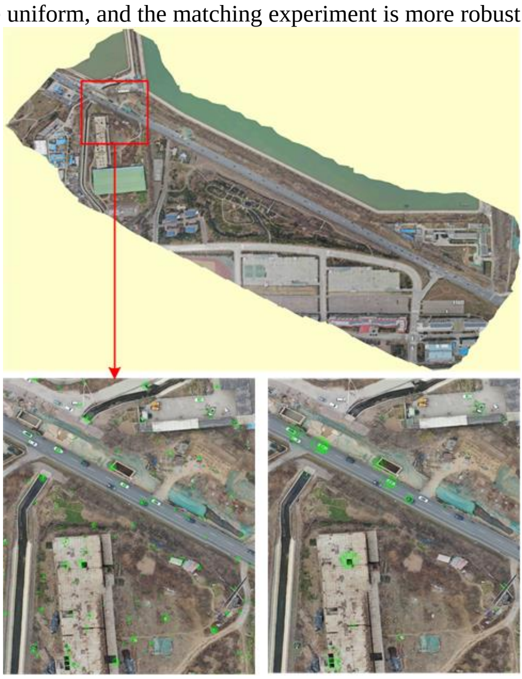


Figure 5. Feature detection comparison. The upper figure is the reference image, the lower left image is the feature extraction result of the ORB algorithm based on the image feature pyramid model, and the lower right image is the feature extraction result of the ORB algorithm.

4.2 Landmark Matching

After computing ORB-descriptors for all landmark points, feature matching is performed by computing the Euclidean distance of these feature vectors for all pairs of landmark points. We expect very similar matching distances for landmark points describing comparable scene objects in the images. If only the nearest neighbour is considered as a match, most of the possible matches will be missed. To solve this problem, a one-to-many matching scheme is applied, by taking the Bag of Words (BoW) model as putative matches (Fig. 6).

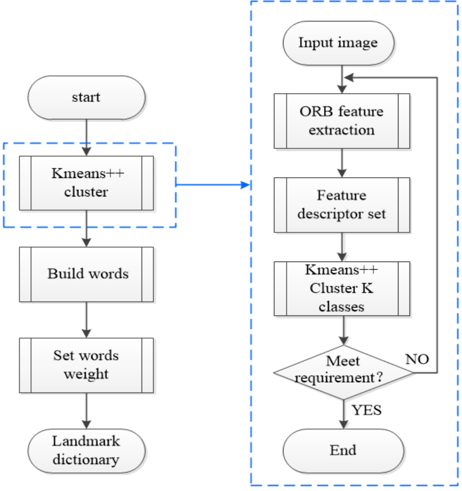


Figure 6. Landmark dictionary build flow

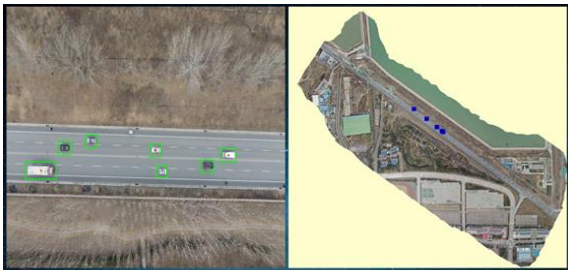
Figure 6 shows the flow of the landmark points dictionary generation. Using standard methods, the correct match would most probably be missed, because the descriptors of the putative matches are very close to each other. BoW-based matching of real-time video image and reference image landmark points are conducted with direct index method. The BoW model classifies all extracted landmarks on the image, then using the classified features instead of the original landmark descriptors could improve the efficiency of landmark search and matching.

To reduce the number of mismatching, a threshold according to the feature matching distance is applied to discard clear mismatches, as this is also proposed in the original ORB-matching (Rublee, Ethan, 2012). Our experiments show that 0.2 is a good trade-off between rejecting strong outliers and retaining enough correct matches.

4.3 Landmark Matching

In the experiment, we set up 100m, 200m, 300m drone video images of different altitudes to detect vehicle targets, and then based on the improved ORB algorithm to match real-time video images and reference images to obtain geolocation of detected moving vehicle targets. For evaluation, the matches created with our method are used to estimate a homography H and fundamental matrix F together with RANSAC. The following shows the experimental results of the target detection and geolocation of moving vehicle under different altitudes.

(1) flight altitude 100m vehicle detection and geolocation



(2) flight altitude 200m vehicle detection and geolocation

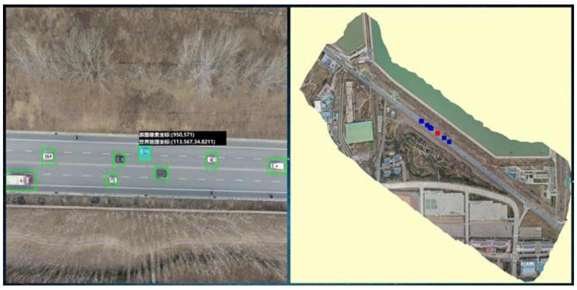
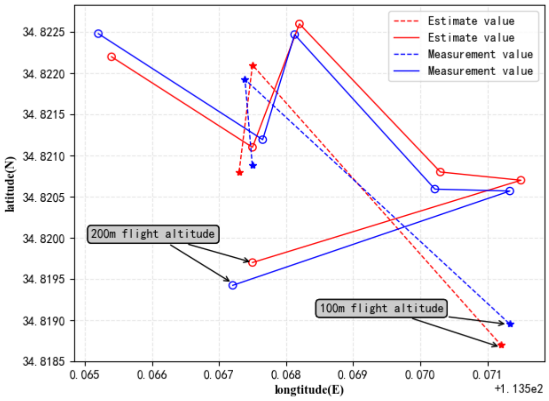


Figure 7. Moving vehicle targets detection and geolocation for different flight altitude. (In the left figure, the green box is used to express different classes of vehicle targets detected by the YOLOv3 network model. The pixel coordinates of the vehicle detected center point are indicated by red crosses, and the blue box shows the geo-location of the target，the red box shows the selected target in the right figure on the reference map.)

The experimental results in Figure 7 show that the flight altitude 100m and 200m drone video recognition results for three classes of moving vehicle targets are quite good, without missed recognition and false recognition. Vehicle targets geolocation are stable, the recognized vehicle targets can be mapped to the reference image robustly. We calculated the geographical coordinates estimate of the detected vehicle target is (113.5670E, 34.8211N) when flight altitude is 200m, and then we obtained the actually measure value is (113.56695E, 34.82119N), the error control at 0.0001 degrees within the range of valid numbers reserved.

For evaluation moving vehicle target positioning accuracy robustly, we randomly selected 9 pairs of detected geographical coordinates of the vehicle and calculated the difference between the target geographical coordinate estimate value and measure value (Fig.8). The experimental data shows that there are differences between the estimate value and measurement value of the target geographical coordinates of the flight altitude of 100m and 200m. However, it can be seen from the figure that the difference between the estimate value and measurement value is not significant, and the differences of the longitude and latitude are less than 0.0002 degrees when the flight altitude is 100m, the geolocation accuracy is relatively stable, and the longitude and latitude differences fluctuates sharply when the flight altitude is 200m, the geolocation accuracy is unstable, the maximum and minimum difference reaches 0.0045 degrees. Besides, the average error of these 9 pairs of data can meet the requirement of less than 0.001degrees from the average error histogram figure.



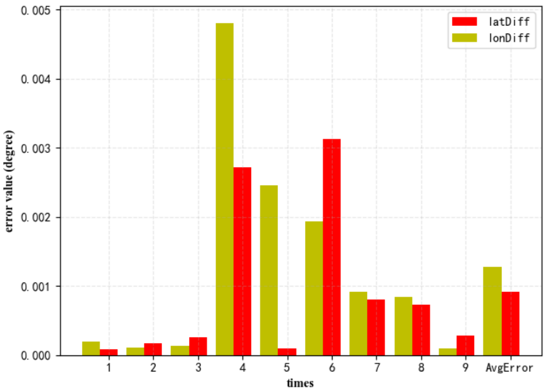


Figure 8. Geolocation difference comparison (The upper figure shows the scatter difference of target geolocation data of 100m and 200m flight altitude, the under figure shows the difference of the estimate value and measurement value)

The first two sets experiment show that target recognition and geolocation are robust, but vehicle targets geolocation are unstable with increasing mismatch between keyframes and landmark database when the flight altitude exceeds 200m. The following figure shows the result of vehicle detection and geolocation of 300m flight altitude. It can be found that different classes of vehicle targets could be detected, but the target geolocation effect on the reference image is poor, which seriously deviates from the original position of the target and could not meet the geolocation requirements.

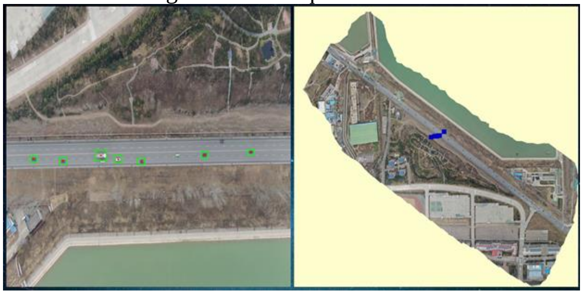


Figure 9. flight altitude 300m vehicle detection and geolocation

In summary, the experimental shows that the target real-time detection and recognition have strong stability and high robustness, and the target geolocation accuracy is controlled within 0.8m when flight altitude are below 200m by comparing target estimated geolocation with measurement geolocation, and the target geolocation accuracy is very unstable and poorly robust when the flight altitude exceeds 200m.

5. EXPERIMENTS AND RESULTS

In this paper, low-attitude drone was applied to acquire high resolution reference image and real-time video images, we proposed a visual localization method based on YOLOv3 deep learning network model and ORB feature extraction matching. For moving vehicle target detection, we obtained the weight file of the YOLOv3 model by training over 1000 images of the DJI MAVIC 2 drone containing more than 40000 targets. For feature extraction matching, we improved ORB feature extraction strategy, and obtained a more uniformly distributed feature points than raw ORB feature extraction. Experiments on the benchmark visual localization dataset shows that our method performs well in visual geolocation within 300m of flight altitude. Besides, we compared geographical coordinates of the detected vehicle target estimate value and its measure value, the proposed method did not require any prior information, which makes our method more practical. In the future, we would like to improve the image matching algorithm and attempt to replace the YOLOv3 model for target detection.

REFERENCES

Alcantarilla, P. F., Solutions, T., 2017. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7): 1281-1298.

Babenko, A., Slesarev, A., Chigorin, A., 2014. Neural codes for image retrieval. *European conference on computer vision. Springer, Cham*. doi.org/10.1007/978-3-319-10590-1_38.

Baker, S., Matthews, I., 2004. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Visio*n, 56(3):221-255.

Chen, Y., Liu, L., Gong, Z., 2017. Learning CNN to Pair UAV Video Image Patches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP(99):1-17. doi.org/10.1109/JSTARS.2017.2740898.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).* IEEE, 2: 2161-2168. doi.org/10.1109/CVPR.2006.264

Filho, P. S., Shiguemori, E. H., 2017. Uav Visual Autolocalizaton Based on Automatic Landmark Recognition. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*. doi.org/10.5194/isprs-annals-IV-2-W3-89-2017

Gordo, A., Almazán, J., Revaud, J., 2016. Deep image retrieval: Learning global representations for image search. *European conference on computer vision. Springer, Cham*.

Huang, S., M., Huang, C. C., 2012. Image registration among UAV image sequence and Google satellite image under quality mismatch. *International Conference on ITSTelecommunications. doi.org/10.1109/ITST.2012.6425189.*

Jégou, H., Douze, M., 2010. Aggregating local descriptors into a compact image representation. *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 3304-3311.

Jegou, H., Perronnin, F., Douze, M., 2011. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9): 1704-1716.

Kwon, H., Sharma, R., 2012. Robust mobile ground target localization using ground image features with UAV position compensation techniques. *2012 12th International Conference on Control, Automation and Systems.* IEEE, 454-458.

Lee, J., H., Kwon, K., C., 2015. GPS spoofing detection using accelerometers and performance analysis with probability of detection. *International Journal of Control, Automation and Systems*, 13(4):951-959.

Liao, X., Zhang, W., 2014. A novel method for gps anti-jamming based on blind source separation. *2014 Seventh International Symposium on Computational Intelligence and Design*. doi.org/10.1109/ISCID.2014.130.

Liu, X., J., Tao, X., M., Duan, Y., P., 2017. Visual information assisted UAV positioning using priori remote-sensing information. *Multimedia Tools & Applications*, (6):1-20.

Lin, Y., Yu, Q., 2007. Medioni G. Map-enhanced UAV image sequence registration. *2007 IEEE Workshop on Applications of Computer Vision (WACV'07)*. doi.org/10.1109/WACV.2007.42.

Lowe, D., G.., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91-110.

Nassar, A., Amer, K., ElHakim, R., 2018. A Deep CNN-Based Framework For Enhanced Aerial Imagery Registration with Applications to UAV Geolocalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* doi.org/10.1109/CVPRW.2018.00201.

Philbin, J., Chum, O., Isard, M., 2007. Object retrieval with large vocabularies and fast spatial matching. *2007 IEEE Conference on Computer Vision and Pattern Recognition.* doi.org/10.1109/CVPR.2007.383172.

Radenović, F., Tolias, G., Chum, O., 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. *European conference on computer vision. Springer, Cham,* 3-20.

Rebiere, N., Auclair-Fortier, M., F., 2008. Image mosaicing using local optical flow registration. *2008 19th International Conference on Pattern Recognition.* IEEE, 1-5.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv*:1804.02767.

Rublee, Ethan , 2012. ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision.* doi.org/*10.1109/ICCV.2011.6126544*.

Tahar, K., N., Kamarudin, S,. S,. 2016. UAV Onboard GPS in Positioning Determination. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B1:1037-1042.

Verdie, Y., Yi, K., Fua, P., 2015. TILDE: a temporally invariant learned detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tolias, G., Sicre, R., Jégou, H., 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv*:1511.05879.

Vivet, M., Brais, Martínez, Binefa, X., 2011. DLIG: Direct Local Indirect Global Alignment for Video Mosaicing. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(12):1869-1878.

Moo, Yi, K., Verdie, Y., Fua, P,. 2016. Learning to assign orientations to feature points. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi.org/10.1109/CVPR.2016.19.

Yoon, Y., Gruber, S., Krakow, L., 2009. Autonomous Target Detection and Localization Using Cooperative Unmanned Aerial Vehicles. *Lecture Notes in Control & Information Sciences*, 381:195-205. doi.org/10.1007/978-3-540-88063-9_12.