Multi-UAV joint target recognizing based on binocular vision theory

Zhang Yuan, Wang Jun

AVIC Fai, Avonic Department, 710089 Xi' an, China

Abstract. Target recognizing of unmanned aerial vehicle (UAV) based on image processing take the advantage of 2D information containing in the image for identifying the target. Compare to single UAV with electrical optical tracking system (EOTS), multi-UAV with EOTS is able to take a group of image focused on the suspected target from multiple view point. Benefit from matching each couple of image in this group, points set constituted by matched feature points implicates the depth of each point. Coordinate of target feature points could be computing from depth of feature points. This depth information makes up a cloud of points and reconstructed an exclusive 3D model to recognizing system. Considering the target recognizing do not require precise target model, the cloud of feature points was regrouped into n subsets and reconstructed to a semi-3D model. Casting these subsets in a Cartesian coordinate and applying these projections in convolutional neural networks (CNN) respectively, the integrated output of networks is the improved result of recognizing..

1 Introduction

A common method of target recognizing for UAV is to input a single image containing suspected target to the recognizing system and compare to template in the library. Affect by lighting, view point position of EOTS or noise etc., the recognizing accuracy is lower than ideal. Despite of these distractions, the suspected target also has the exclusive 3D model (Considering there are no identical targets in the image). Reconstructing the 3D model of suspect target by cloud of feature points provides extra information could be more accurate than 2D recognizing method. Based on the binocular vision theory, the 3D model also could be reconstructed by UAVs with single EOTS. Assuming the image from each EOTS was grouped into an image set. As long as the UAV group does share overlapping field of view, which require enough matched points (the number of points is agile and determined by complexity of suspected target) in the image group, the 3D model was able to be reconstructed. Figure 1 shows a scenario of 3 EOTS joint recognizing.

Although these matched feature points is able to reconstruct a coarse 3D model, the recognizing system is not in need. In the recognizing procedure, the semi-3D model is satisfied and lower resource consuming. The semi-3D model is a layer group which depth value on the light axis of EOTS is compute from matched points in couple image. This layer group is a degenerating of 3D model. Casting this layer group on Z plane and X plane of Cartesian coordinate, which the light axis of image sensor is X axis, these 2 projections contained all 3 dimension features. After identifying these 2 projections with trained convolutional neural networks (CNN), fusing outputs of CNN is indispensable and final procedure in the recognizing system. The system architecture is shown in Figure 2. The first step is data assembly. When recognizing, the center node UAV, called the image processing node assemble all sensor images (at least there are 2 images in the assembly set). The second step is the pre-processing of reconstructing. By matching feature points and computing the depth information, the space relationship of layer group was get. In this procedure, mismatching of feature points which leads to inaccurate semi-3D model should be avoiding. Matching points should be grouped into n (n is re-set and decided by complexity of targets) sets. Each set represents one feature zone which the depth of zone is represented by the center point depth itself. In the last step, all these n feature point sets reconstruct the semi-3D model of suspected target and recognize with 2 projections.

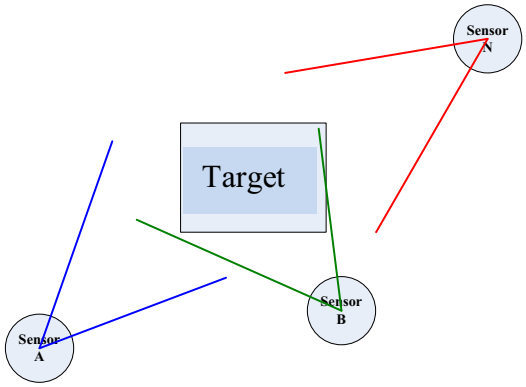

Figure 1. Scenario of multi-sensor joint recognizing.

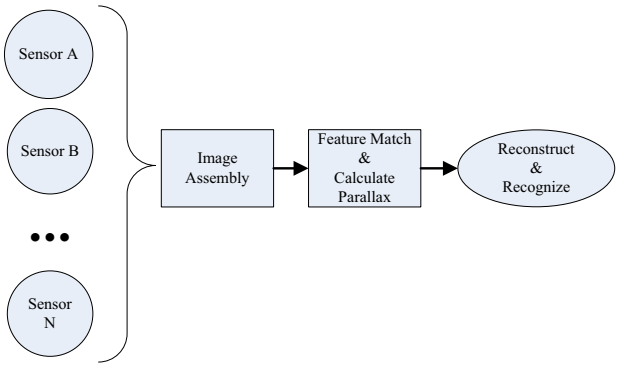

Figure 2. Recognizing System Architecture.

2 Target processing and recognizing

2.1 Matching feature points

In recognizing, the first important prerequisite is the field of view of each couple of EOTS is overlap and the suspected target should be in this field. Otherwise, there are no matched points in the field of view to identify the suspected target or even reconstruct the 3D model. In figure 1, both couples (A-B, B-N and A-N) and the UAV group (including A, B and N) could be joint recognizing. But the couple of A and N, which has a few overlap fields of view, could not reconstruct the semi-3D model for lacking or missing feature points.

In case of matching 2 images whose size is M*N, the image from UAV A was called image A and the image from UAV B was called image B. After extracting and matching feature points with SIFT, the mismatching points in result should be eliminate for badly reconstructing. Assuming U and V are column and row coordinates respectively. Based on the hypothesis that the layout relationship of matched feature points is almost coincident (electric-optic turret of UAV is able to keep the image plane from view rotating, shown in figure 3 a) and b)), set up a vector as follows: $f = [l, k]$ where:

$$l = (U_A + N - U_B)^2 + (V_A + M - V_B)^2 \quad (1)$$

$$k = (U_A + N - U_B) / (V_A + M - V_B) \quad (2)$$

For the layout relationship of feature points is stable, the vector f has a narrow variable range. Computing the density of sub-space constructed by each couple points with follows:

$$D_{k,l} = \frac{\Sigma i}{\pi r^2} \quad (3)$$

Where: The subscript k and l is the center point and edge point index respectively, this point couple construct a circle area as the selected space. πr^2 = the number of matched feature points in selected space, Σi = the area of selected space and the radius r is the distance between point k and l. Computing each space density and selecting the most density space as the room which the matched feature points in. The feature points scattered before and after eliminating were shown in figure 3 c) and d).

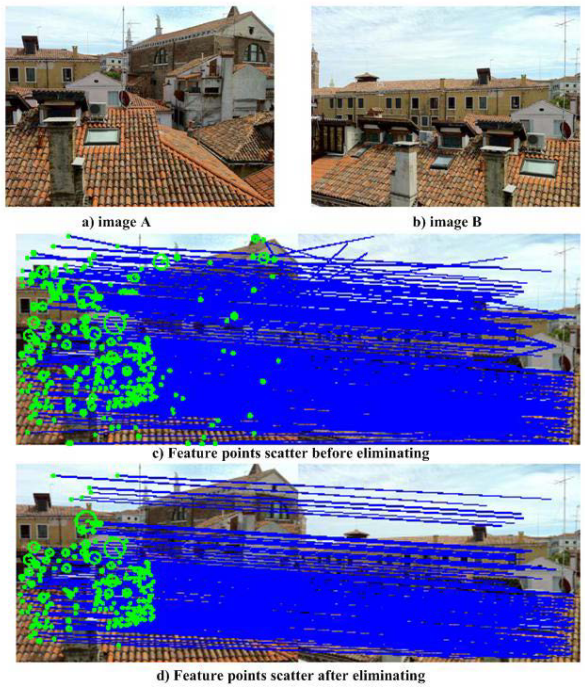

Figure 3. Feature points scatter: a) image A; b)image B; c) before eliminating; 4)after eliminating

2.2 Reconstructing semi-3D model and Casting

According the binocular vision theory, the depth information of feature points which is the foundation to reconstruct 3D model of suspected target is exclusive and certain. The abundant 3D information benefits the accurate rate of recognizing system but consuming amount of memory and computing resource. Therefore 3D model is not required and even harmful to system in the target recognizing task. For avoiding useless depth information overload the computer resource, the feature points will be divided into n subsets which each subset just has single depth value. K-means cluster algorithm which took the distance between each point into account is used to decide the subset. Figure 4 shows the clustering result for the image in sensor A. In this case, n was set to 4.

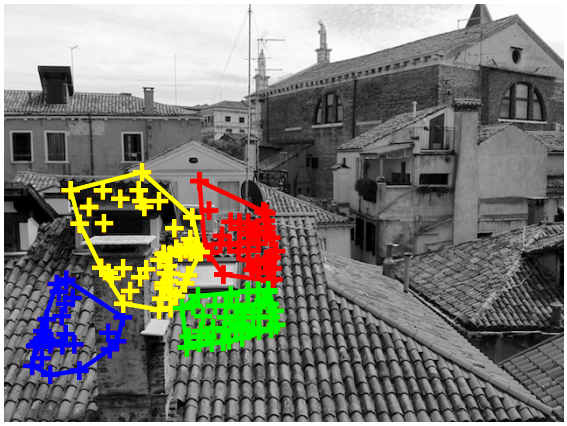

Figure 4. The clustering result in image from sensor A.

After clustering, each subset is confirmed and its depth value is represented by the point nearby the geometry center point of subset. The depth information in EOTS A coordinate system computing follows （Origin of this coordinate is the center point of sensor A imaging plane, x axis is the light axis of sensor A, y and z axis are parallel to the imaging plane row and column.）:

$$\frac{y_{OA}}{U_{OA}} = \frac{x_{OA}}{f_{OA}} \tag{4}$$

$$\frac{z_{OA}}{V_{OA}} = \frac{x_{OA}}{f_{OA}} \tag{5}$$

$$tan\theta = \frac{x_{BA}y_{OA} - y_{BA}x_{OA}}{y_{BA}y_{OA} + x_{OA}x_{BA}} \tag{6}$$

Where: x_{OA}, y_{OA}, z_{OA}, x_{BA} and y_{BA} is the target and UAV B position in A coordinate, θ is the angle between line of sensor A-target to sensor B-target. The position of UAV B in geographic coordinate is come from navigation system.

For target position in sensor B coordinate, the position is computing as follows:

$$\frac{y_{OB}}{U_{OB}} = \frac{x_{OB}}{f_{OB}} \tag{7}$$

$$\frac{z_{OB}}{V_{OB}} = \frac{x_{OB}}{f_{OB}} \tag{8}$$

$$tan\theta = \frac{x_{OB}y_{AB} - y_{OB}x_{AB}}{y_{OB}y_{AB} + x_{OB}x_{AB}} \tag{9}$$

Where: x_{OB}, y_{OB}, z_{OB}, x_{AB} and y_{AB} is the target and UAV A position in B coordinate. The position of UAV A in geographic coordinate is come from navigation system.

Transforming the point in B to A coordinate with equation (10), the target point coordinate should be similar with its position coordinate in A coordinate.

$$\begin{bmatrix} x_A \\ y_A \end{bmatrix} = \begin{bmatrix} cos\gamma & sin\gamma \\ -sin\gamma & cos\gamma \end{bmatrix} \begin{bmatrix} x_B \\ y_B \end{bmatrix} + \begin{bmatrix} x_{BA} \\ y_{BA} \end{bmatrix} \tag{10}$$

Where: γ is the angle from light axis of EOTS B to EOTS A.

As the position of each UAV is given (the navigation system of each UAV broadcast their self-positions in the group), after transforming the geographic coordinate system to Cartesian coordinate system, depth of each feature point could be compute for combining the equation (4) to equation (10).

2.3 Reconstructing semi-3D model and casting

After reconstructing, analyzing the semi-3D model of suspected target with CNN require the model 2D features. Casting the semi-3D model on planes parallel to light axis and vertical to light axis, figure 5 shows the projection on z plane and projection on x plane. The projection on z plane, which represents the grey value accumulation of points with same z position, shows the geometry relationship of each layer and the scatter of feature points on z plane. The grey scale value of each pixel of the projection on the x plane is multiply the original grey scale value by the depth value after normalizing.

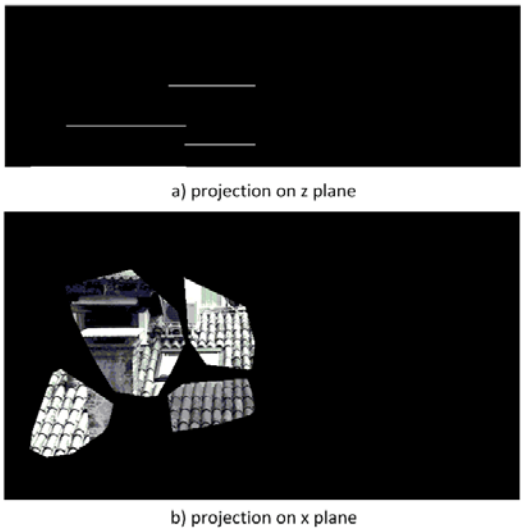

Figure 5. Projection of semi-3D model: a) projecting on z plane; b) projection on x plane.

Setting up CNN (in this case, LeNet-5 was apply to recognizing system) networks to analyzing these 2 projections and fusing each output together, the final outputs is the recognizing outputs of suspected target. To fuse the result, Establish the fuse principle follows:

$$minJ = (y - a_1y_1 - a_2y_2)^2 \tag{11}$$

Where: y is the standard output of target model in the library, a1 is the weight of CNN output for z plane projection, y1 is the output of CNN output for z plane projection, a2 is the weight of CNN output for x plane projection, y2 is the output of CNN output for x plane projection. a1 and a2 should be training with gradient descent method before recognizing suspected target.

3 Simulation and analyzing

Set three kinds of building as targets for testing this recognizing system: department, mall and bridge. Stp1, establishing the target library by these VR models. Stp2, arranging 2 or 3 virtual moving cameras from different view point. Stp3, processing these three models follows the procedure shown in figure 1and training CNN networks. Stp5, set up a test set included 3 models above and castle and tank. The recognizing result is shown in table I and table II. The recognizing rate is improving with the increasing of feature point number. The complicated surface structure (complicated surface texture, the shortage of valid depth feature point information, the onefold spatial target structure, etc.)

requires more feature points in overlap field for archiving similar recognizing rate which means the larger overlap field and smaller angle between target to sensors. Table II shows the increasing of recognizing rate with the number of feature points. After appending one EOTS to recognizing system, the overlap field is larger and the feature points in each 2 field of view could be reconstructing a model in larger scale. Feature points in the public field of 3 EOTS, after reconstructing by its depth information, could be compensate the depth value which provide a more precision semi-3D model. The recognizing rate difference is shown in Table 1 and Table 2.

Table 1. Recognizing results of 2 EOTS joint recognize.

	Result		
	department	mall	bridge
False alarm rate(%)	12	10	15
Recognizing rate(%)	67	70	66
Feature points in total	35	42	37

Table 2. Recognizing results of 3 EOTS joint recognize.

	Result		
	department	mall	bridge
False alarm rate(%)	12	10	15
Recognizing rate(%)	71	72	70
Feature points in total	52	60	49

4 Conclusions

This paper provides a method of recognizing suspected target with multi-image sensor. By reconstructing a semi-3D model of suspected target, the exclusive semi-3D geometry features was set up and represented by layer. The recognizing system not only takes layers constructed by feature points as input but also the projection of these layers which is represent the position relationship of these layers. Benefited with the abandoned geometry feature provided by semi-3D model, the multi-image sensor joint recognizing can improve the estimation. Though the joint recognizing, a more accurate and reliable result can be achieved.

References

1. 1.Hartley Zisserman, “Multiple View Geometry in Computer Vision” Phil. Cambridge University Press, New York, pp. 237 – 360, 2000.
2. 2.Ke Y, Su kthankar R, “PCA- SIFT: a more distinctive representation for local image descriptors”, Proceedings of IEEE Conference on Computer Vision and Pattern, Washington D.C.506-513, 2004.
3. Hinton G E, Salakhutdinov R R, “Reducing the dimensionality of data with neural networks”, Science, **313**(5786):504-507, 2006.
4. Rumelhart D E, Hinton G E, Williams R J, “Learning representations by back-propagating errors”, Nature, 323:533-538,1986.
5. Krizhevsky A, “Learning multiple layers of features from tiny images”, University of Toronto, Toronto, Canada, 2009.