Interframe Association of YOLO Bounding Boxes in the Presence of Camera Panning and Zooming

Zijiao Tian, Yaakov Bar-Shalom, Rong Yang, Hong'An Jack Huang and Gee Wah Ng

May 13, 2023

Interframe Association of YOLO Bounding Boxes in the Presence of Camera Panning and Zooming

Zijiao Tian,Yaakov Bar-Shalom
Department of ECE,
University of Connecticut,
Storrs, CT 06269, USA
Emails: zijiao.tian@uconn.edu
yaakov.bar-shalom@uconn.edu

Rong Yang, Hong'An Jack Huang
DSO National Laboratories,
12 Science Park Drive,
Singapore 118225
Emails: yrong@dso.org.sg
hhongan@dso.org.sg

Gee Wah Ng
Home Team Sci. and Tech. Agency(HTX),
1 Stars Ave, #12-01
Singapore 138507
Email: ng_gee_wah@htx.gov.sg

***Abstract*—In this paper, we develop an approach for measurement-to-track association (M2TA) in the presence of (unknown) camera panning and zooming from drone-captured video. Standard M2TA methods assume that the target motion can be used to predict the "measurement association regions" for the bounding boxes. However, if there is a sudden state change due to camera shift (panning) and zooming, it will lead to incorrect associations and poor tracking results. To solve this, the zoom ratio and panning in 2D coordinates are used to describe the camera motion parameters in each frame. The estimated parameters are obtained by a grid search combined with global assignment or directly solved using the linear least squares method, which is also combined iteratively with assignment. The goal is to achieve correct M2TA by adjusting the predicted measurements using the estimated camera parameters. These "improved" predictions can also be used to update the target state with filtering algorithms. Frames with panning or/and zooming from real data are used to illustrate the effectiveness of the proposed methods and compared with the validation gate method based on inflated covariances.**

***Index Terms*—multitarget tracking, camera panning and zooming, YOLO, validation gate, measurement-to-track association, surveillance with UAVs.**

I. Introduction

In unmanned aerial vehicles (UAVs) based tracking systems, it is common for the camera to change its field of view (FoV) to track moving targets [6], [9], [10], [14]. Validation gates are used in measurement-to-track associations (M2TA), such as Nearest Neighbor as well as Global assignment [2], [3]. However, when the camera changes its viewing direction and field of view by panning and/or zooming, there is a sudden position change for each target image between two consecutive frames. Thus, the measurements may not be correctly associated with existing tracks for M2TA, resulting in degraded tracking performance.

To solve the abrupt camera motion problem, Ref. [7] modeled the camera motion by geometric transformation based on background feature points. Ref. [8] compensated for camera movement for joint tracking and video registration based on a factorial Hidden Markov Model. Then, the maximum-a-posteri (MAP) and the reversible jump Markov chain Monte Carlo methods were employed for tracking and camera motion estimation in [4]. Ref. [5] used the Enhanced Correlation Coefficient (ECC) for aligning two image profiles. However, it is a pixel-based approach, which requires operating on every pixel on both the input and template image. This leads to a very high computation complexity, therefore making it less suitable for real-time video applications. Alternatively, the Covariance Inflation (CI) approach [2] can be used. The camera parameter changes creates bias in the estimation. Instead of estimating the bias precisely, this approach increases the measurement noise variances to embed the bias implicitly. For the association, the association gate is enlarged by an inflated measurement error variance so that shifted measurements can be associated with their tracks. Some simple logic is applied to decide if the gates need to be changed, e.g., the tracks without measurements in their gate will use inflated measurement covariance for the association. However, this is a heuristic and (possibly very) imperfect approach. The camera parameter changes causes not only the measurements to shift but also the predicted states. Without taking the state shift due to panning and zooming (PZ) into consideration, can lead to errors in both association and state update. Therefore, a better and systematic approach is desired.

In this paper, the UAV is flying at a fixed altitude and there is one camera mounted on it. Moving targets on the ground are detected by YOLO (You Only Look Once) — a real-time detection deep learning algorithm [11], [12]. It provides the bounding boxes (BB, position information) of the detected moving targets and can distinguish different objects, such as cars vs. people. We only focus on people in the video in our study and track the top left corner of the BB.

If the camera mounted on the UAV has the capability to adjust its view, such as zooming in and panning, then the tracking process can be more accurate and effective. The camera parameter is described by the zoom ratio and 2D shift due to panning in the horizontal and vertical directions. The camera parameters are constant for each person in one frame.

Proc. 26nd Intn'l Conf. on Information Fusion, Charleston, SC, USA, June 2023.

Gee Wah Ng is a staff from DSO National Laboratories who is on secondment to HTX.

During the 2d assignment[1], the new predicted measurements are generated based on the estimated camera parameters and are compared with the current measurements. The camera parameters are estimated using two methods. The first is using a global optimum grid search with a prior search grid. The search grid is narrowed down for each iteration until all M2TA pairings have the minimal total cost. The second method is using the linear least squares (LLS) method. The parameter estimates are directly obtained by taking the partial derivatives of the global cost function and equating them to zero, which is simpler and more precise than the grid search method. Based on the estimated camera parameters, one calculates corrected predictions. During the filtering step, these are used with a Kalman filter and the tracking results are improved significantly. The simulations consider the panning and/or zooming cases and show that the proposed method — (assignment with camera parameter estimation) — yields better associations than the simple gating method.

The paper is organized as follows: Section II introduces the measurements obtained by YOLO and the baseline method using the inflated validation gates. Section III describes the first proposed method when the camera parameters are estimated by the grid search. The analytical minimization method based on the LLS is presented in Section IV. Section V shows the experiments on real data and discusses the results. Finally, Section VI provides concluding remarks.

II. Problem Formulation

For video tracking, a camera mounted on a UAV with a fixed altitude is pointed toward the targets. The bounding boxes of each target are acquired using a deep-learning technique. To track accurately, the camera has the capability to adjust its view (i.e., move with the targets), including zooming in/out and panning. However, the unknown camera motion (zooming and/or panning) can result in incorrect assignments using the single data association approach, such as gating.

A. YOLO Measurements

The measurements are the target's positions (bounding boxes), which are obtained by a state-of-the-art object detection algorithm — You Only Look Once (YOLO) [11]. YOLO can detect multiple objects in an image or video frame and label them with corresponding class probabilities in real time. It uses a single neural network to make predictions for multiple objects, which makes it faster and more efficient than traditional object detection algorithms. The network divides the input image into a grid and predicts bounding boxes and class probabilities for each grid cell. YOLOv3 [12], which is used here, has some minor improvements in detection and bounding boxes accuracy for smaller targets.

As shown in Fig. 1, YOLOv3 detects a car and four people from a video. It also has the ability to distinguish people and cars with different labels. Thus, we can only focus on the targets we are interested in, i.e., people.

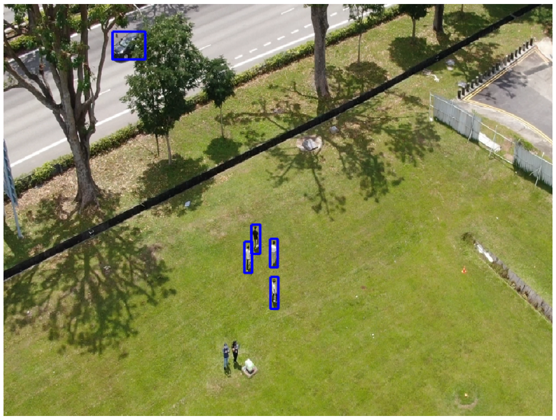


Fig. 1. The moving objects detected by YOLOv3 [12]. The blue 2D bounding boxes contain the position information of each target.

B. Validation Region and Auction-based Assignment

To reduce computation, a gating (validation) procedure is commonly used in measurement to track association [13], [15]. Assume there is a set of measurements $z_i(k)$ at time k, i=$1, 2, \ldots, N_m$ and a set of predicted measurements $\hat{z}_j(k|k-1)$ at time k, j=$1, 2, \ldots, N_t$. A filter has been initialized and each target has its track. The probability density function (pdf) of the measurement prediction from the target, designated as j, is normally distributed given by

$$p[z_j(k)] = \mathcal{N}[z_j(k); \hat{z}_j(k|k-1), S_j(k)] \quad (1)$$

where $S_j(k)$ is the associated innovation covariance [1].

The difference between each measurement and each predicted measurement is evaluated by the normalized distance d_{ji}^2 from track j to measurement i and is given by

$$d_{ji}^2 = [z_i(k) - \hat{z}_j(k|k-1)]' S_j(k)^{-1} [z_i(k) - \hat{z}_j(k|k-1)] \quad (2)$$

This is also referred to as the squared Mahalanobis distance. If $d_{ji}^2 \leq \gamma$, we consider the measurement z_i falls within the track $\hat{z}_j$'s gate (γ is the gate threshold, typically based on the chi-square distribution).

Based on (2), a score matrix A with elements a_{ji} is generated and the assignment can be solved via the auction method. The auction procedure is composed of the bidding phase and the assignment phase. The outline of steps in the auction algorithm is as follows [3]:

First, select an unassigned measurement i until there is no unassigned measurement.

Second, find the best track j_i for each measurement. j_i should satisfy

$$a_{j_i i} - P_{j_i} = \max_{j=1,\ldots,n} (a_{ji} - P_j) \quad (3)$$

where $a_{j_i i}$ is the gain from assigning measurement i to track j_i and P_{j_i} is the "price" of track j_i.

[1] 2d refers to "two dimensional assignment", i.e., between two lists (in our case latest measurements and existing tracks); 2D refers to two dimensional space.

Third, unassign the measurement previously assigned to j_i and then assign track j_i to measurement i.

Fourth, update the price of track j_i to the level at which observation i is "almost happy"

$$P_{j_i} \leftarrow P_{j_i} + y_i + \epsilon \tag{4}$$

where y_i is the difference between the best and second best assignment values for measurement i. Finally, return to the second step until all measurements are assigned.

In the following update step, the filter (such as a Kalman filter) will incorporate the assigned measurements into the predicted state to obtain an updated state estimate.

C. Inaccurate Association Problem

We assume there are multiple targets in one video frame, and they move with a nearly constant velocity (NCV) [1]. The camera mounted on a UAV can change its view for better tracking. The sudden change happens[2] due to abrupt camera motion [10], [16], and thus leads to a large position change for each target between frames. While validation gates and auction can provide good M2TA pairs in target tracking, they assume that the target motion follows a predictable trajectory (NCV) with certain uncertainty.

There are two special cases for camera motion. First, when the camera is panning (horizontal or vertical or arbitrary direction), validation gates that ignore this may lead to incorrect M2TA, particularly when people are closely spaced and small. Second, when the camera is zooming in, although a larger gate size can be set in gating, one still has no information about the zoom ratio (ratio of the focal length at the current frame vs. the previous frame), which will lead to incorrect estimates of the target state in the filtering step. Fig. 2 and Fig. 3 show the above two cases at frame 233 and frame 1165 of the real database considered, respectively.

III. Association with Camera Parameter Estimation

In this section, we introduce the proposed M2TA algorithm to address the camera motion problem. The fundamental idea is to estimate a camera parameter vector that includes PZ. This results in better predicted measurements and filter updates.

A. Association Method

Assuming there is a set of measurements $z_i(k)$ at time k, i=$1, 2, \ldots, N_m$, and a set of predicted measurements $\hat{z}_j(k|k-1)$ at time k, j=$1, 2, \ldots, N_t$. The κorrected prediction conditioned on the (unknown) camera pointing shift (panning) is denoted as $\mathbf{z}_c(k)$=$[x_c(k), y_c(k)]'$ and the zoom ratio denoted as $\phi(k)$, is given by

$$\hat{z}_j^{\kappa}[k|k-1, \boldsymbol{\zeta}(k)] = \begin{bmatrix} \hat{x}_j(k|k-1)\phi(k) + x_c(k) \\ \hat{y}_j(k|k-1)\phi(k) + y_c(k) \end{bmatrix} \tag{5}$$

$$\triangleq \hat{z}_j(k|k-1)\phi(k) + \mathbf{z}_c(k) \tag{6}$$

and the zoom ratio ϕ (which multiplies the positions relative to the center of the FPA) is given by

$$\phi(k) = \frac{f(k)}{f(k-1)} \tag{7}$$

with $f(k)$ is the focal length at time k.[3] The camera parameter vector is

$$\boldsymbol{\zeta}(k) = \ [\phi(k) \ \ \mathbf{z}_c(k)']' \tag{8}$$

$$= \begin{bmatrix} \phi(k) \\ x_c(k) \\ y_c(k) \end{bmatrix} \tag{9}$$

and $\mathbf{z}_c(k)$ is the camera pointing displacement mapped to the Focal Plane Array (FPA).

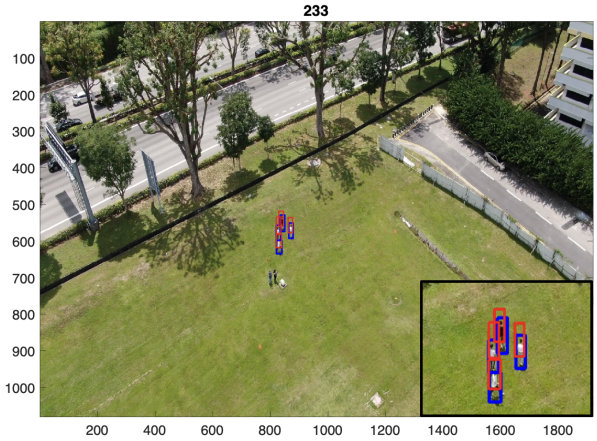


Fig. 2. Case 1: there is a large deviation between the tracking results based on gating (red bounding boxes) and the actual positions (blue bounding boxes) due to panning.

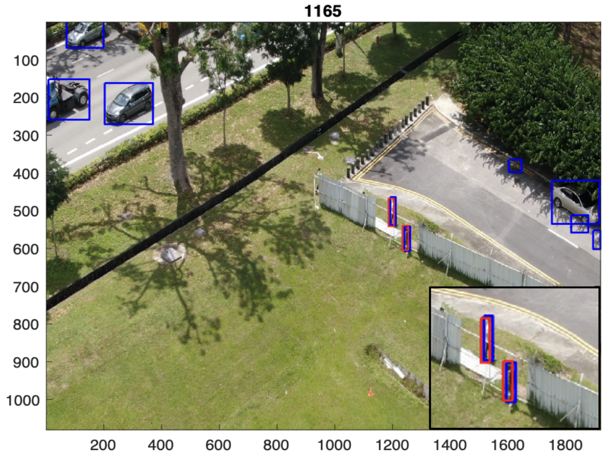


Fig. 3. Case 2: the camera is zooming in and the size of the bounding box is increased as well as its location relation to center of the frame. The gating method needs a larger gate size.

[2] In addition, wind fluctuations might cause sudden shakes in video recording.

[3] The focal lengths are unknown and not observable (unless the sizes of the length are known). However, the ratio (7) can be estimated.

The association should be done between the following:

1. Tracks represented by corrected (denoted by superscript “κ”) predictions $\hat{z}_j^{\kappa}[k|k-1, \boldsymbol{\zeta}(k)]$, j=$0, 1, \ldots, N_t$ where the index j=0 represents the “dummy tracks” to which the unassociated measurements belong (they will be used to start new tracks). Note the corrections to the predicted measurements are done with the yet to be determined $\boldsymbol{\zeta}(k)$.

2. Measurements $z_i(k)$, i=$0, 1, \ldots, N_m$ where the index i=0 represents the “dummy measurements” to which the unassociated tracks belong.

The cost of assigning $z_i(k)$ to $\hat{z}_j^{\kappa}[k|k-1, \boldsymbol{\zeta}(k)]$ is the scalar normalized squared distance[4]

$$c[i, j, k, \boldsymbol{\zeta}(k)] = ||z_i(k) - \hat{z}_j^{\kappa}(k|k-1)||^2 \tag{10}$$

Finally,

$$\hat{\boldsymbol{\zeta}}(k) = \underset{\boldsymbol{\zeta}(k)}{\operatorname{argmin}}\, C[k, \boldsymbol{\zeta}(k)] \tag{11}$$

$$= [\hat{\phi} \quad \hat{x}_c \quad \hat{y}_c] \tag{12}$$

where

$$C[k, \boldsymbol{\zeta}(k)] = \sum_{i, j_A(i)} c[i, j_A(i), k, \boldsymbol{\zeta}(k)] \tag{13}$$

for a given assignment A.

B. Grid Search

The procedure of the 3D grid (for the 3D vector (9)) with 27 candidate $\boldsymbol{\zeta}$ vectors (we set several values for each component of $[\phi, x, y]'$ to be discussed in more details) is presented below:

During each iteration, for each grid point, indexed (l, u, v) each pair in the 2d assignment (tracks and measurements) between the two lists is evaluated for[5] $[\phi_l, x_u, y_v]'$. The result is the best Global Optimal Assignment (GOP) for grid point (l, u, v)

$$C^*_{l,u,v} = C^{\text{GOP}}_{l,u,v} \tag{14}$$

The Global Optimum for the Grid (GOG) is

$$\min_{l,u,v} C^{\text{GOP}}_{l,u,v} = C^{\text{GOG}} \tag{15}$$

Due to the camera motion being the same for each object at time k, $[\phi(k), x(k), y(k)]'$ are the same for each pair in the corrected assignment. Based on (15), the pairs and the camera parameters vector estimate $\hat{\boldsymbol{\zeta}}$ are obtained. The search range is narrowed down in the next iteration until the minimal cost value C^{GOG} is within a tolerance difference from the previous one. Note that when there is no measurement associated with the current track, the cost value is very large. A threshold can be set to avoid associating a “disappeared” track.

C. Filtering

After association, the assigned measurements are incorporated into the updated track state estimates during the filtering stage. The “improved” prediction $\hat{z}_j^{\kappa}$ (6) can be used for the target state update in the Kalman filter. Thus the updated state (only position) estimate for target j with the filter gain K is given by

$$\hat{x}_j(k|k-1) = \hat{z}_j^{\kappa}[k|k-1, \hat{\boldsymbol{\zeta}}(k)] + K(k)\nu(k) \tag{16}$$

where

$$\nu(k) = z_j(k) - \hat{z}_j^{\kappa}[k|k-1, \hat{\boldsymbol{\zeta}}(k)] \tag{17}$$

is the measurement residual. Since the proposed method (assignment with camera parameter estimation) provides a more accurate prediction, the tracking results will be significantly better than the validation gating method when the camera is PZ.

IV. Analytical Minimization Method

Although using the grid search can yield an estimated camera parameter that is close to the optimal, it requires a prior interval for each parameter and may result in a high computing complexity due to several iterations with a narrowing grid.

The linear least squares (LLS) method is used next to solve for $\hat{\boldsymbol{\zeta}}$ directly. For an initial assignment with $\{i \leftrightarrow j_{0(i)}\}_{i=1}^{N_m}$, the cost from (10) is expressed as

$$c[i, j_0(i), k, \boldsymbol{\zeta}(k)] = ||z_i(k) - \hat{z}_{j_0}^{\kappa}(k|k-1)||^2 \tag{18}$$

$$= [x_i(k) - \hat{x}_{j_0}(k|k-1)\phi(k) - x_c(k)]^2$$
$$+ [y_i(k) - \hat{y}_{j_0}(k|k-1)\phi(k) - y_c(k)]^2 \tag{19}$$

We can then find $\hat{\boldsymbol{\zeta}}(k)$ by setting the partial derivatives of $c[i, j_0(i), k, \boldsymbol{\zeta}(k)]$ with respect to ϕ, x_c and y_c are zero, as follows

$$\nabla_{\boldsymbol{\zeta}} c[i, j_0(i), k, \boldsymbol{\zeta}(k)] = 0 \tag{20}$$

which yields

$$\frac{1}{2}\frac{\partial c}{\partial \phi} = (\hat{x}_j^2 + \hat{y}_j^2)\phi + \hat{x}_j x_c + \hat{y}_j y_c - \hat{x}_j x_i - \hat{y}_j y_i \tag{21}$$

$$\frac{1}{2}\frac{\partial c}{\partial x_c} = \hat{x}_j \phi + x_c - x_i \tag{22}$$

$$\frac{1}{2}\frac{\partial c}{\partial y_c} = \hat{y}_j \phi + y_c - y_i \tag{23}$$

Summing up (21)–(23) for all association pairs and equating to zero (omitting k for simplicity), $\hat{\boldsymbol{\zeta}}(k)$ is obtained from the matrix equation given by

$$\mathbf{A}\boldsymbol{\zeta} = \mathbf{b} \tag{24}$$

where

$$\mathbf{A} = \begin{bmatrix} \sum_{j=1}^{N}(\hat{x}_j^2 + \hat{y}_j^2) & \sum_{j=1}^{N}\hat{x}_j & \sum_{j=1}^{N}\hat{y}_j \\ \sum_{j=1}^{N}\hat{x}_j & N & 0 \\ \sum_{j=1}^{N}\hat{y}_j & 0 & N \end{bmatrix} \tag{25}$$

$$\mathbf{b} = \begin{bmatrix} \sum_{j=1}^{N}(\hat{x}_j x_{i(j)} + \hat{y}_j y_{i(j)}) \\ \sum_{j=1}^{N} x_{i(j)} \\ \sum_{j=1}^{N} y_{i(j)} \end{bmatrix} \tag{26}$$

[4] We assume the innovation covariances are all diagonal and equal, thus we can omit them.

[5] The subscript c for camera is omitted here for simplicity.

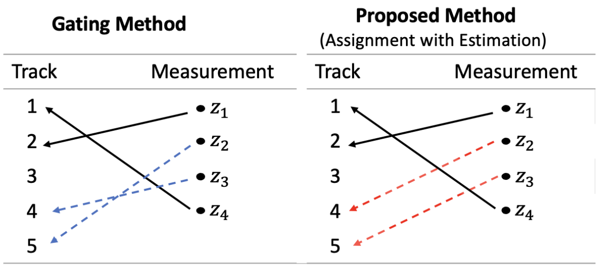


Fig. 4. 2d assignment for case 1.

where $N = \min\{N_m, N_t\}$ is the number of M2TA pairings; $i(j)$ is the index of measurement associated with track j. Note that the sum of the derivatives yields a non-singular $\mathbf{A}$; at least two pairings can determine the three camera parameters and there is no need for iterations. The predictions from (6) are then corrected based on the estimated parameters.

V. Real Data Results

The camera is at a relatively high altitude from a UAV and the targets (people) are small and closely spaced. The sampling frequency is 30 Hz (for 1 second there are 30 frames/scans). The frame has a size of 1920×1080 (pixels). Three cases are considered to illustrate the effectiveness of the proposed method.

The first case with panning is shown in Fig. 2. At frame 233, there are four measurements detected by YOLO and there are five predicted measurements (tracks). Due to panning, there are large tracking errors (the red bounding boxes have large deviations from the blue bounding boxes) based on the gating method.

Table I shows the estimated parameters by using the grid search. We set 3 search values of each component of $[\phi \;\; x_c \;\; y_c]$ and, as such, there are 27 candidates ζ vectors for each iteration. After one iteration, the global optimal estimated parameters $[\hat{\phi} \;\; \hat{x}_c \;\; \hat{y}_c]$ are obtained. Then the search range of parameters is narrowed down based on the previously estimated parameters. The final best camera parameter estimate is [1.005 -2.5 17.5] for this case. In addition, we can estimate the camera parameters using the LLS method, which yields [0.999 2.9 21.0]. The difference between these can be attributed to noise[6]. The assignment results compared with the gating method are shown in Fig. 4. The gating method assigns z_2 and z_3 to tracks 5 and 4, respectively, which is not correct. The proposed method using both parameters estimation (grid and LLS) approaches yields correct M2TA pairs[7].

The global cost for the M2TA pairings based on the Gating Method (Assignment w/o correction since it does not estimate $\boldsymbol{\zeta}$) and the grid search, the LLS are 1730, 7.8, 5.3, respectively. It shows the predictions are more precise from both proposed

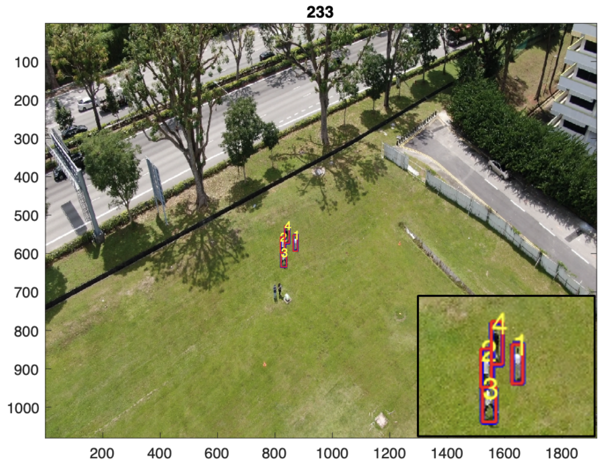


Fig. 5. Case 1: the proposed method shows better tracking results (the red bounding boxes are close to the blue bounding boxes) compared to the tracking results based on the gating method in Fig. 2.

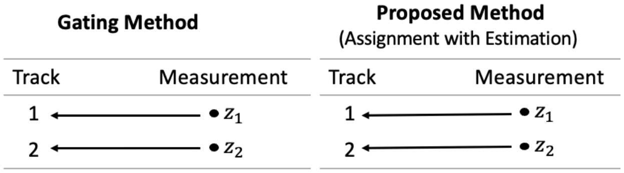


Fig. 6. 2d assignment for case 2.

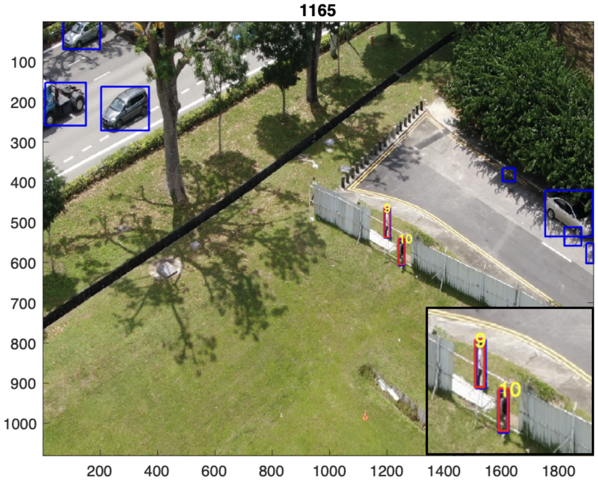


Fig. 7. Case 2: the proposed method shows better tracking results compared to the tracking results based on the gating method in Fig. 3.

[6] The statistical significance of the estimates will be discussed in the expanded version of this paper, which will also include a tilting parameter.

[7] Since there is no formal ground truth, this was ascertained by visual inspection of the frames.

TABLE I. Camera parameter search and the best parameters for case 1

Search Iteration	**Grid Points**			**Best Parameters**
	ϕ	x_c	y_c	$[\hat{\phi}\ \ \hat{x}_c\ \ \hat{y}_c]$
1	[0.96 1 1.04]	[-20 0 20]	[-20 0 20]	[1 0 20]
2	[0.99 1 1.01]	[-5 0 5]	[15 20 25]	[1.01 -5 15]
3	[1.005 1.01 1.015]	[-7.5 -5 -2.5]	[12.5 15 17.5]	**[1.005 -2.5 17.5]**

TABLE II. Camera parameter search and the best parameters for case 2

Search Iteration	**Grid Points**			**Best Parameters**
	ϕ	x_c	y_c	$[\hat{\phi}\ \ \hat{x}_c\ \ \hat{y}_c]$
1	[0.96 1 1.04]	[-20 0 20]	[-20 0 20]	[1 0 0]
2	[0.99 1 1.01]	[-5 0 5]	[-5 0 5]	**[1.01 -5 -5]**

TABLE III. Camera parameter search and the best parameters for case 3

Search Iteration	**Grid Points**			**Best Parameters**
	ϕ	x_c	y_c	$[\hat{\phi}\ \ \hat{x}_c\ \ \hat{y}_c]$
1	[0.96 1 1.04]	[-20 0 20]	[-20 0 20]	[1 0 20]
2	[0.99 1 1.01]	[-5 0 5]	[15 20 25]	[1.01 -5 25]
3	[1.005 1 1.015]	[-7.5 -5 -2.5]	[20 25 30]	[1.01 -7.5 30]
4	[1.0025 1.01 1.0175]	[-10 -7.5 -5]	[27.5 30 32.5]	**[1.01 -10 32.5]**

methods. The tracking results based on the proposed method[8] are shown in Fig. 5. Thanks to the estimated camera parameters, the tracking positions of each target are much accurate than the results in Fig. 2.

The second case with zooming in is shown in Fig. 3. There are two detected measurements. The estimated parameters from the grid search are $[1.01\ \ -5\ \ -5]$ in Table II, and parameters from the LLS are $[1.022\ \ -20\ \ -12]$. The final cost for the Gating Method is 97.9 while for the proposed method using the grid search and the LLS are only 4.6 and 2.3, respectivley. Fig. 6 shows the assignment results and Fig. 7 shows the tracking results. Although the gating method did find the correct pairs, the gate size was changed to a larger one than the normal size. However, if the gate size is too large, then false measurements may be associated with targets, leading to incorrect target state estimates. The proposed method can provide the zoom ratio without the need for a heuristic gate size, and the tracking results are better than the gating method.

The third case with both zooming and panning is considered in Fig. 8. Based on 4 iterations, the best parameters from the grid search are $[1.01\ \ -10\ \ 32.5]$ from Table III. It is noted that since there is only one detected target, the estimated camera cannot be obtained from the LLS method (only one pair results in a singular matrix $\mathbf{A}$ (25)), which inidicates the benefits of the grid search. The tracking results are shown in Fig. 8. Due to camera motion, the red bounding box has a large deviation from the blue (actual) bounding box from the gating method as shown in Fig. 8(a). Based on the estimated camera parameters, the accuracy of the estimated target position is largely improved in Fig. 8(b).

TABLE IV. Average tracking errors (RMSE in pixels of the estimated bounding box position) for the three cases

Case	**Gating method**	**Proposed method**
1	29.5	1.03
2	19.3	1.31
3	37.6	1.30

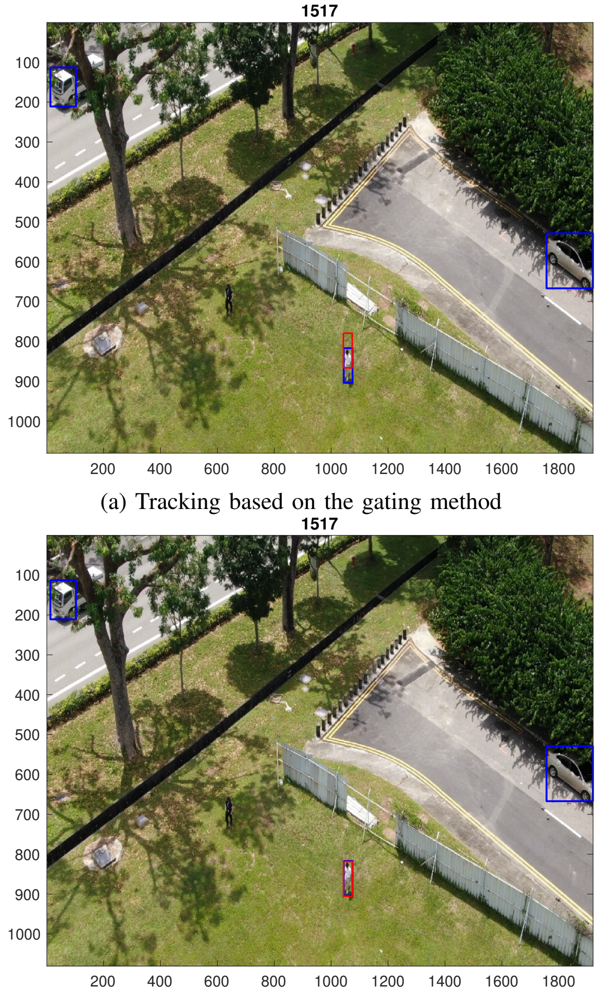
Finally, Table IV shows the average position Root Mean Square Error (RMSE) of the estimated positions of the bounding boxes for the above three cases (3 frames). With camera parameters estimate, the proposed method yields significantly smaller tracking errors vs. the gating method.

VI. Conclusions and Future Work

In this paper, we solved the inaccurate measurement-to-track association problem caused by camera motion. The camera mounted on an UAV and its parameters are determined by the zoom ratio and panning in 2D coordinates, which are estimated by the grid search or the least squares method. At each frame, the measurements are associated with the current tracks based on an improved prediction from the estimated camera parameters. The proposed approach (assignment with camera parameter estimation) not only yields correct M2TA pairings but also can provide better state estimation in the update step. Simulations show that the proposed method has better performance than the validation gating method when the camera changes its field of view by panning and zooming.

Further comparisons of the two algorithms proposed here with the baseline YOLO tracker on the real data will be carried out to obtain the tracking accuracy (innovation norms) and reliability (possible breakages and false tracks) over the entire

[8] Here we show only the results of the method with the grid search to save space.



(a) Tracking based on the gating method

(b) Tracking based on the proposed method

Fig. 8. Case 3: the camera is both zooming in and panning. Tracking results based on (a) the gating method and (b) the proposed method.

length of the data and, if available, additional real scenarios will be considered. Also, improving the basic tracking filter design parameters [1] will be done.

References

[1] Y. Bar-Shalom, X. R. Li and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation*, Hoboken, NJ, USA: Wiley, 2001.

[2] Y. Bar-Shalom, P. Willett and X. Tian *Tracking and Data Fusion: A Handbook of Algorithms*, YBS Publishing, 2011.

[3] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*, Dedham, MA, USA:Artech House, 1999.

[4] W. Choi, C. Pantofaru and S. Savarese, "A General Framework for Tracking Multiple People from a Moving Camera", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1577-1591, July 2013.

[5] G. D. Evangelidis and E. Z. Psarakis, "Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858-1865, Oct. 2008.

[6] B. Kiefer, *et al.*, "1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 2023.

[7] S. Li and D.-Y. Yeung, "Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models", *AAAI*, vol. 31, no. 1, Feb. 2017.

[8] X. Mei and F. Porikli, "Joint tracking and video registration by factorial Hidden Markov models", *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 973-976, Las Vegas, NV, USA, 2008.

[9] M. Mueller, G. Sharma, N. Smith and B. Ghanem, "Persistent Aerial Tracking system for UAVs", *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Korea (South), 2016, pp. 1562-1569.

[10] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking", *in ECCV 2016*, pp. 445–461.

[11] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[12] J. Redmon, and A. Farhadi, "Yolov3: An incremental improvement". arXiv preprint arXiv:1804.02767, 2018.

[13] D. Reid, "An algorithm for tracking multiple targets", *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843-854, December 1979.

[14] J. Thomas, J. Welde, G. Loianno, K. Daniilidis and V. Kumar, "Autonomous Flight for Detection, Localization, and Tracking of Moving Targets With a Small Quadrotor", *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1762-1769, Jul. 2017.

[15] B.-N. Vo *et al.*, "Multitarget tracking", *in Wiley Encyclopedia of Electrical and Electronics Engineering.* New York, NY, USA: Wiley, 2015.

[16] P. Zhu et al., "Detection and Tracking Meet Drones Challenge", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380-7399, 1 Nov. 2022.