IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2017, accepted March 13, 2017, date of publication March 23, 2017, date of current version April 24, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2686482

Learning Spatio-Temporal Information for Multi-Object Tracking

JIAN WEI, MEI YANG, AND FENG LIU

Jiangsu Province Key Laboratory on Image Processing and Image Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Corresponding author: Jian Wei (tdweijian@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501260 and Grant 61471201, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20130867, in part by the Jiangsu Province Higher Education Institutions Natural Science Research Key Grant under Project 13KJA510004, in part by the Peak Of Six Talents in Jiangsu Province under Grant RLD201402, in part by the Natural Science Foundation of NJUPT under Grant NY214031, in part by the 1311 Talent Program of NJUPT, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

ABSTRACT The robust multi-object tracking problem is a challenging issue in the field of computer vision. In this paper, we propose a multi-object tracking algorithm with temporal-spatial information and trajectory of confidence. The whole process is divided into local and global association. Trajectories with high confidence are associated with the detection result of the current frame during local association, whereas trajectories with low confidence are associated with the detection results of the current frame are not matched during global association. We determine the association results using a combined model. By utilizing the information of spatial-temporal correlation, the model is more robust and can deal with missed detection. In addition, we measure the reliability of the spatial information by the confidence map smoothing constraint and the peak sidelobe ratio criterion. We conduct experiments using a challenging public data set, and the results show that our proposed algorithm is superior to many other popular algorithms when dealing with problems, such as missed detection and poor tracker robustness.

INDEX TERMS Multi-object tracking, trajectory of confidence, spatio-temporal information.

I. INTRODUCTION

Multi-object tracking has always been a challenging problem in the field of computer vision. Its aim is to estimate the state of multiple objects while preserve their changing appearance and motion characteristics over time. Multi-object tracking has a wide range of applications, such as video surveillance, motion analysis, anomaly detection, robot navigation, and automatic driving. However, in complex scenarios, there are particularly significant challenges as the objects may frequently occlude with each other or have similar appearances which affects the tracking process. Owing to the rapid development of target detection [1], tracking-by-detection methods has achieved very reliable test results and exhibited good performance even in complex environments. This method associates all detection results obtained by detection to generate the trajectories of objects and can be roughly divided into steps, namely batch processing and online processing. Recently, these technologies [2]–[5] have significantly improved, eliminating the problem of blurs in association and detection failure. Initially, batch-tracking methods [6], [7] usually acquire all detection results of video sequences, and connect them together to generate a short trace of target objects. Then, these short traces can be associated globally to form complete target trajectories. Thus, the global association is critical during the progress. Many global association methods [8], [9] have been proposed, but they still cannot treat cases in which target objects are blocked for a long time. Furthermore, these methods usually need to detect the entire video sequence in advance and iterate constantly to achieve the global optimal solution with a high calculation complexity. As a result, the use of batch process methods does not easily solve the problem of real-time applications in multi-object tracking. Online tracking methods [10], [11] can perform real-time processing to generate traces according to information about the current and previous frames without the need to pre-detect the entire video sequence. Nonetheless, it is more sensitive to errors and missed detections, which generates many more tracking fragments and object exchanges between each other. Online multi-object tracking methods based on the confidence of the trajectory [9] typically involve two steps. First, tracking results for the first few frames that are close to each other are connected

to a reliable short trace. Then, these trajectories with high confidence can grow gradually with local association. The confidence of the trajectories may decrease when objects are blocked, and detection is lost during detection; hence, global association is still needed to connect these fragments of low confidence to form complete trajectories. The connection process during the above two steps can be achieved by establishing the combined model using an appearance model, motion model, and constraints of shapes. Track learning [12]–[14] is also a popular research direction in the field of multi-object tracking and has attracted much attention. According to one study, it can be divided into offline track learning and online track learning. Even though offline learning cannot utilize the dynamic and historical information associated with objects when associating, it is a good solution for fuzzy problems, especially when moving objects are repeated in cases of tracking failure. While online learning can achieve real-time tracking by taking advantage of target features and priori information, there may be drifts if incorrect training samples are learned. In this article, we apply spatial and temporal information to the combined model. For the problem of tracking, temporal information refers to all of the objective information of prior frames, while spatial information refers to local target and surrounding regions in the background. Most local spatial information remains unchanged between two adjacent frames owing to the small neighboring time intervals. As a consequence, there is a strong spatial and temporal relationship between continuous frames. The proposed combined model is more robust because of the strong correlation of temporal information, and can address issues such as incorrect tracking more easily. In addition, in this paper, we propose to measure the reliability of spatio-temporal information using the confidence map smoothing constraint and the peak sidelobe ratio criterion. The proposed multi-object tracking algorithm can effectively deal with the problems of missed detection, and improve the robustness of the target tracker.

II. RELATED WORKS

In recent years, the existed methods [15]–[18] in the field of multi-object tracking are mostly based on Kalman filters and particle filters. These methods are effective for forecasting states that have a short duration rather than in complex scenes. These data association methods, such as the joint probabilistic data association filter (JPDAF) [19], multiple hypothesis tracking (MHT) [20], and Markov chain Monte Carlo sampling techniques (MCMCDA) [21] can solve the tracking problem in complex scenarios.

In this paper, we apply the Spatio-Temporal Context (STC) algorithm [22] to extract spatial-temporal information, the contribution of the paper related to [22], and we briefly introduce the algorithm below. Assuming that x^* is the center of the objects position, then the local spatial information is defined as

$$X^c = \{c(z) = (I(z), z) | z \in \Omega_c(x^*)\} \tag{1}$$

where $I(z)$ is the gray value at position z, and $\Omega_c(x^*)$ represents the neighborhood area with twice the size of the object area at position x^*.

The algorithm estimates the likelihood of the targets position by calculating a confidence map, and it solves the tracking problem as follows:

$$m(x) = P(x|o) \tag{2}$$

where $x \in R^2$ is the location of the object and o indicates the presence of an object in the scene. Eq. (2) can also be expressed as

$$m(x) = \sum_{c(z) \in X^c} P(x|c(z), o) P(c(z)|o) \tag{3}$$

where the conditional probability $P(x|c(z), o)$ is modeled as the spatial relationship between the objects center and its local spatial information, which is modeled as a priori probability $P(c(z), o)$.

This algorithms core is to learn the spatial relationship between the objects center and its local position, as previously mentioned before. Consequently, the conditional probability is defined as

$$P(x|c(z), o) = h^{sc}(x - z) \tag{4}$$

where $h^{sc}(x - z)$ represents the relative distance and direction of the objects center x and its local spatial position z.

The priori probability is related to the apparent of local spatial information, and can be expressed as

$$P(c(z)|o) = I(z)\omega_\sigma(z - x^*) \tag{5}$$

where $\omega(\cdot)$ is a Gaussian weighting function.

The objects confidence map is defined as

$$m(x) = P(x|o) = be^{-|\frac{x-x^*}{\alpha}|^\beta} \tag{6}$$

where b is the normalized constant, and α and β are the scale parameter and shape parameter, respectively. Combining Eq. (4), (5), and (6), Eq. (3) can also be expressed as

$$m(x) = h^{sc}(x) \otimes (I(x)\omega_\sigma(x - x^*)) \tag{7}$$

where $\otimes$ represents the operator of convolution. Eq. (7) can be calculated in the frequency domain using the fast Fourier transform (FFT) to learn the spatial information $h^{sc}(x)$.

The temporal information is updated as $H^{stc}_{t+1} = (1 - \rho)H^{stc}_t + \rho h^{sc}_t$, where ρ is the learning rate. Using the spatial and temporal information learned, the objects position in the next frame can be determined as

$$x^*_{t+1} = \arg\max_{x \in \Omega_c(x^*_t)} m_{t+1}(x) \tag{8}$$

where $m_{t+1}(x)$ is defined as

$$m_{t+1}(x) = H^{stc}_{t+1} \otimes (I_{t+1}(x)\omega_{\sigma_t}(x - x^*_t)) \tag{9}$$

III. OUR TRACKING FRAMEWORK

If object i appears at frame t, the binary function is set as $v^i(t) = 1$; otherwise, $v^i(t) = 0$. If $v^i(t) = 1$, the state of object i is represented as $x_t^i = (p_t^i, s_t^i, v_t^i)$, where p_t^i,s_t^i, and v_t^i indicate the position, size, and speed, respectively. The state of object i until frame t is expressed as $T^i = \{x_k^i | v^i(k) = 1, 1 \le t_s^i \le k \le t_e^i \le t\}$, where t_s^i and t_e^i represent the start and end frame numbers of object i during the trajectory. In addition, $\mathbb{T}_{1:t}$ represents all traces of the video sequence until frame t. Similarly, z_t^i stands for the detection results of object i at frame t, and $\mathbb{Z}_{1:t}$ represents all test results of the video sequence at frame t. Online multi-object tracking can be considered as to involve seeking the optimal solution of $\mathbb{T}_{1:t}$, and maximizing the posterior probability with given $\mathbb{Z}_{1:t}$.

$$\hat{\mathbb{T}}_{1:t}^{MAP} = \arg\max_{\mathbb{T}_{1:t}} p(\mathbb{T}_{1:t} | \mathbb{Z}_{1:t}) \tag{10}$$

It is impossible to solve Eq.(10) directly because of the numerous possible combinations of $\mathbb{Z}_{1:t}$. However, it can be resolved into an equation of the trajectory of confidence.

A. TRAJECTORY OF CONFIDENCE

The confidence of the trajectory can be intuitively understood as the similarity between the candidate trace and the real one. Trajectories with high confidence should meet the following requirements: (1) Length of trajectory: a short trace is more likely to be unreliable, while a long one is more likely to be the correct object trace; (2) Occlusion: a trajectory that is severely blocked by other traces should not be considered a reliable result; (3) Combined model: a trajectory that matches the test results with a high score is a more reliable result.

Based on the above requirements, the trajectory of confidence can be modeled as

$$conf(T^i) = \left(\frac{1}{L} \sum_{\substack{k \in [t_s^i, t_e^i], \\ v^i(k)=1}} \Lambda(T^i, z_k^i) \right) \times \max((1 + \beta \cdot \log((L - \omega)/L)), 0) \tag{11}$$

where L is the base of T^i, namely the length of the trajectory, with the formula expressed as $L = |T^i|$. ω represents the missing frames due to blockages by other objects or unreliable test results and is expressed as $\omega = t_e^i - t_s^i + 1 - L$. The first item in the Eq.(11) is the score of the combined model between trajectories and detection results; a high value implies a high degree of trajectory confidence. We define the combined model below. The second item in the Eq.(11) implies that a short trajectory length or congested scenario can reduce the confidence with a control parameter β, which depends on the detection performance. β is set to a large value when the detector has a high accuracy.

B. TRAJECTORY OF ASSOCIATION

In order to solve the problem of multi-object tracking effectively, Eq.(10) is combined with the trajectory of confidence and is redefined as follows

$$\hat{\mathbb{T}}_{1:t}^{MAP} = \arg\max_{\mathbb{T}_{1:t}} \iint p(\mathbb{T}_{1:t} | \mathbb{T}_{1:t}^{(hi)}, \mathbb{T}_{1:t}^{(lo)}) \times p(\mathbb{T}_{1:t}^{(lo)} | \mathbb{T}_{1:t}^{(hi)}, \mathbb{Z}_{1:t}) p(\mathbb{T}_{1:t}^{(hi)} | \mathbb{Z}_{1:t}) d\mathbb{T}_{1:t}^{(hi)} d\mathbb{T}_{1:t}^{(lo)} \tag{12}$$

where $\mathbb{T}_{1:t}^{(lo)}$ and $\mathbb{T}_{1:t}^{(hi)}$ represent trajectories with low confidence and high confidence, respectively.

As shown in Eq.(12), multi-object tracking problems are divided into two parts. First, trajectories with high confidence are associated with detection results that are provided online, namely local association. Then, trajectories with low confidence are associated with other trajectories and detection results that are unmatched in the current frame, namely global association.

During local association, trajectories with high confidence $\mathbb{T}^{i(hi)}$ are allied with detection results that are provided online $\mathbb{Z}_t$. If the number of trajectories with high confidence is h and the detection results are n in frame t, then the score matrix $S_{h \times n}$ is defined as

$$S = [s_{ij}]_{h \times n}, s_{ij} = -\log\left(\Lambda\left(\mathbb{T}^{i(hi)}, z_t^j\right)\right), z_t^j \in \mathbb{Z}_t \tag{13}$$

where $\Lambda(\mathbb{T}^{i(hi)}, z_t^j)$ is shown in Eq. (15). Matrix $S_{h \times n}$ can be calculated using the Hungarian algorithm to obtain the matching results between the trajectories and detection. When the association costs of the matching fall below a preset threshold $-\log(\theta)$, then z_t^j is associated with $\mathbb{T}^{i(hi)}$. Consequently, we perform the following steps: (1) Update the position and velocity of the trajectories with an associated z_t^j, and the size of the target track with the average of an objects size in the previous and current frames; (2) Update $conf(\mathbb{T}^{i(hi)})$ using Eq.(11) and z_t^j.

In the global association, trajectories with low confidence are likely to be short fragments. Therefore, they should be associated with the other trajectories and detection results that are unmatched in the current frame. Suppose there are h and l trajectories with high and low confidence, respectively, as well as n detection results that are unmatched. Consider the following related events: (1) Event A: $\mathbb{T}^{i(lo)}$ is associated with $\mathbb{T}^{j(hi)}$; (2) Event B: $\mathbb{T}^{i(lo)}$ is terminated; (3) Event C: $\mathbb{T}^{i(lo)}$ is associated with y_t^j.

The total cost matrix during association is:

$$G_{(l+n)\times(h+l)} = \begin{bmatrix} A_{l \times h} & B_{l \times l} \\ -\log(\theta)_{n \times h} & C_{n \times l} \end{bmatrix} \tag{14}$$

where $A = [a_{ij}]$ represents Event A, $a_{ij} = -\log\left(\Lambda\left(\mathbb{T}^{i(lo)}, \mathbb{T}^{j(hi)}\right)\right)$ indicates the association costs, and is calculated using Eq.(15). $B = diag[b_1, ..., b_l]$ represents Event B, and $b_i = -\log\left(1 - conf\left(\mathbb{T}^{i(lo)}\right)\right)$ represents termination costs. $C = [c_{ij}]$ represents Event C, and $c_{ij} = -\log\left(\Lambda\left(\mathbb{T}^{i(lo)}, y_t^j\right)\right)$ can be calculated using Eq.(15). In addition, threshold θ is still used during local association. The minimum optimal cost of global association is calculated using the Hungarian algorithm, and is applied to the updated trace and confidence value.

C. COMBINED MODEL

Our paper describes $\mathbb{T}^i$ from $\{A^i, S^i, M^i\}$, which shows the model of the appearance, shape, and motion respectively. The combined model, which determines the matching results between two trajectories (or trajectory and detection), is defined as follows:

$$\Lambda(X, Y) = \Lambda^A(X, Y)\Lambda^S(X, Y)\Lambda^M(X, Y) \tag{15}$$

where X and Y are the trajectories or detection results.

Owing to the space distance that exists between each pair of trajectories, the value of correlation regarding the spatial-temporal information may be small. Based on the above consideration, the appearance model is divided into two separate models, one of which is for trajectory and detection, whereas the other is for the two trajectories in this article.

1) APPEARANCE MODEL BETWEEN TRAJECTORY AND DETECTION

First, in this paper, we describe the smooth constraint of confidence maps (SCCM), which is derived from temporal-spatial information and the peak-to-sidelobe ratio (PSR), which can better determine whether a target object is blocked or is being incorrectly tracked. The idea and formula (16) (17) (18) originated from [23]. In [23], all parts of the target are independently tracked by using Kernelized Correlation Filter (KCF) [24] trackers. When a new frame arrives, the confidence maps of these tracked parts are first computed. By assigning adaptive weights to these maps, a joint map can be constructed to predict the new state using the particle filter method.

$$PSR_t^i = \frac{\max(m_t^i) - \mu_t^i}{\sigma_t^i} \tag{16}$$

$$SCCM_t^i = \left\| m_t^i - m_{t-1}^i \oplus \Delta \right\|_2^2 \tag{17}$$

$$\rho_t^i = PSR_t^i + \eta \cdot \frac{1}{SCCM_t^i} \tag{18}$$

where m_t^i is the confidence map of object i at frame t with a mean value μ_t^i and variance σ_t^i. In addition, $\oplus$ represents the translational operation of the confidence map. Further, Δ represents the position of the maximum value in the confidence map at frame $t-1$ relative to frame t.

The values of ρ will change when the target object is tracked wrongly. As shown in Fig.1, an error occurs during the tracking near frame 62, which leads to the low value of the graph from position ②.

As shown in Fig.2, the target object is occluded near frames 20, 74, and 140, and is finally tracked wrongly from frame 140.

Fig.3 shows that the value of the graph is relatively low when the target is occluded, but it is still tracked well at frame 62. Conversely, the value of the graph is higher at frames 20 and 80 when the object is tracked correctly.

As shown in Fig.4, the value of the graph is low when the object is occluded at frames 20, 74, and 140, but it is high at any other time when the tracking is correct.

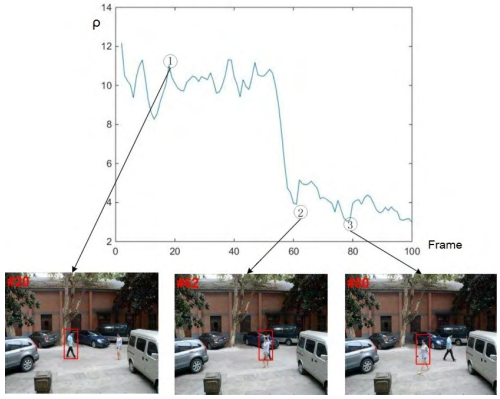


FIGURE 1. Graph of incorrect tracking for the first sequence.

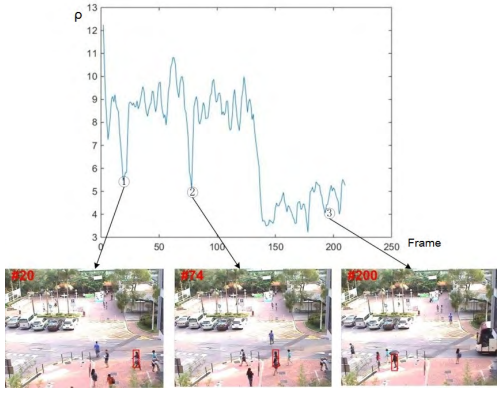


FIGURE 2. Graph of incorrect tracking for the second sequence.

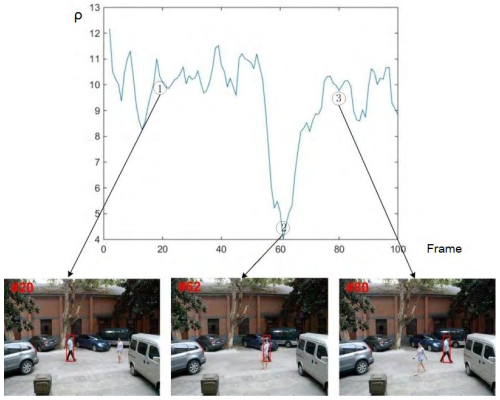


FIGURE 3. Graph of correct tracking for the first sequence under occlusion.

As a consequence, the metrics of SCCM and PSR, which were derived from the confidence map related to spatial-temporal information, give a fair assessment of whether the target is occluded or incorrectly tracked. Below, we describe the SCCM and PSR to determine the reliability of the spatial information.

Suppose the current frame is t and the target result set at frame $t-1$ is $\{l_1, ..., l_n\}$, n is the number of objects at frame $t-1$, and l_i represents the assumed position of the object i.

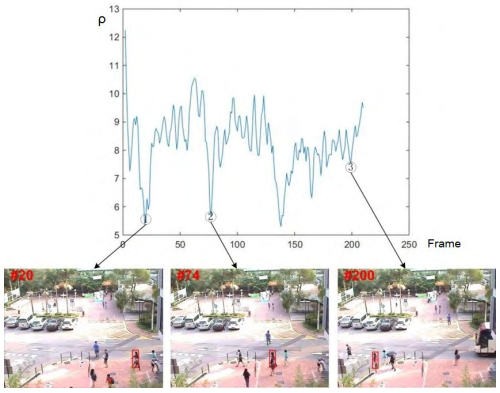


FIGURE 4. Graph of correct tracking for the second sequence under occlusion.

Step 1: Information about the spatial-temporal correlation is applied to predict the result set $\{l_1', l_2', \cdots, l_n'\}$ of object at frame t, where $l_i' = (lx_i', ly_i')$ is the center position of the object i. Calculate the corresponding $SCCM_t^i$ and PSR_t^i. If $SCCM_t^i < 1 \times 10^{-7}$ and $PSR_t^i > 6$, the spatial information at frame t is then considered reliable in that the object is correctly tracked. Go to Step2 to continue. Otherwise, the object is assumed occluded but tracked properly, or wrongly tracked. Then go to Step3;

Step 2: The prediction from Step1 is added to the combined model. At the same time, the value of weight ρ_t^i is assigned using the SCCM and PSR values. Therefore, the combined model is represented as follows:

$$\Lambda^A(X, Y) = STI(X, Y) \cdot BC(H_X, H_Y) \tag{19}$$

where H_X is the color histogram template of trajectory X, and H_Y is the color histogram of detection result Y at frame t. $BC(.)$ represents the Bhattacharyya distance between the histograms, $STI(.)$ is the probability of temporal and spatial correlation and is defined as follows, under the assumption that X is the trajectory of object i:

$$STI(X, Y) = \rho_t^i \times \frac{l_i' \cap Y}{l_i' \cup Y} \times \lambda \sqrt{(lx_i' - Y_x)^2 + (ly_i' - Y_y)^2} \tag{20}$$

where Y_x and Y_y are components in the x axis and y axis of the center coordinates of the detection result Y. λ is the normalization factor and ρ_t^i is the controlling parameter. We achieved the control progress by judging the reliability of the spatial information as in the following definition.

If the detection results at frame t are matched, then the spatial and temporal model, confidence map template, and the objects color histogram template are subsequently updated. Otherwise, it is assumed likely that there is a missed detection. Then, use the prediction result from Step1 as the state of frame t.

Step 3: Considering the situation where the target object is occluded but still correctly tracked, the prediction result from Step 1 is still applied in the combined model, where the weight ρ_t^i is updated using the SCCM and PSR values. The combined model is shown in Eq. (18) at Step2. If any detection results are not matched, then the prediction from Step1 is assumed as the objects state at frame t. However, regardless of whether the detection results are matched, not all models are updated because the correctness is not firmly asserted even when it is matched.

In addition, the cumulative value C records the number of situations for which the SCCM or PSR value is lower than the threshold. If $C > 15$, the object is assumed to be incorrectly tracked. Thus, the first time at which the objects SCCM or PSR value is lower than the threshold is returned, and the tracking is resumed at that point. Then, the combined model neglects the spatial and temporal information, which is formed as follows:

$$\Lambda^A(X, Y) = BC(H_X, H_Y) \tag{21}$$

Otherwise, the target is assumed occluded but still tracked correctly. Then, we continue the tracking and no time node is returned.

2) APPEARANCE MODEL BETWEEN TRAJECTORIES

Assuming that the current frame is t, the tracking results are $\mathbb{T}_{1:t}^{(lo)}$ and $\mathbb{T}_{1:t}^{(hi)}$. The combined model is used to associate trajectories that have high confidence and low confidence, and it is defined as follows:

$$\Lambda^A(\mathbb{T}^{i(lo)}, \mathbb{T}^{j(hi)}) = BC(H_{i(lo)}, H_{j(hi)}) \tag{22}$$

where $H_{i(lo)}$ and $H_{j(hi)}$ are the color histogram templates of trajectory i and j with low confidence and high confidence, respectively.

The shape model is defined as

$$\Lambda^S(X, Y) = \exp\left(-\left\{\frac{h_X - h_Y}{h_X + h_Y} + \frac{w_X - w_Y}{w_X + w_Y}\right\}\right) \tag{23}$$

where the shape model is constructed using the width and height of the object.

The motion model is defined as

$$\begin{aligned}\Lambda^M(X, Y) = {} & N(p_X^{tail} + v_X^F \Theta; p_Y^{head}, O^F) \\ & \times N(p_Y^{head} + v_Y^B \Theta; p_X^{tail}, O^B)\end{aligned} \tag{24}$$

where the forward rate v_X^F is the speed of motion, which is measured from the head of X to its tail. Meanwhile, the backward rate v_Y^B is the speed of motion, which is measured from the tail of Y to its head. In addition, we utilize only the forward movement model when calculating the allied scores between the trajectory and detection results.

IV. EXPERIMENTS

In this paper, we used the CLEAR evaluation [25], which is composed mainly of two parts: multiple object tracking precision (MOTP), which reflects the accuracy when determining the targets location, and multiple object tracking accuracy (MOTA), which reflects the accuracy when determining the number of goals and the related properties. Both of them jointly measure the ability of the algorithm to perform continuous tracking.

FIGURE 5. Test results for the first sequence from S2.L1 using the TC-ODAL algorithm.

FIGURE 6. Test results for the first sequence from S2.L1 using our proposed algorithm.

FIGURE 7. Test results for the second sequence from S2.L1 using the TC-ODAL algorithm.

TABLE 1. Performance comparison for multi-target tracking algorithms.

Test videos	Algorithms	MOTP	MOTA
PETS(S2.L1)	Conf.Map	56.30%	79.70%
	Energy Min	-	80.20%
	TC-ODAL	69.39%	78.19%
	Our algorithm	70.75%	82.86%
PETS(S2.L2)	Conf.Map	51.30%	50.00%
	Energy Min	-	59.40%
	TC-ODAL	54.52%	67.79%
	Our algorithm	56.28%	70.57%

A. PERFORMANCE COMPARISON

In order to evaluate our proposed algorithm, we selecte test videos S2.L1 and S2.L2 in VS-PETS 2009, whose detection results can be derived from literature [5]. In addition, we choose the three most representative multi-object tracking algorithms for comparison with our experiment results using the same detection results as in our algorithm. These three algorithms are Conf. Map [26], Energy Min [5], and TC-ODAL [14]. The performance comparison results for MOTP and MOTA are shown as Table 1.

Fig.5-12 shows specific test results for our algorithm and TC-ODAL with respect to S2.L1. The test results for the TC-ODAL algorithm are shown in Fig.5, where we can see that the trace of ID = 11 gives an error at frame 277, the trace of ID = 4 and ID = 11 have been exchanged from frame 308, and the trace of ID = 6 and ID = 5 deviated from the target position at frames 318 and 326, respectively. However, from Fig.6, we can see that our algorithm can correctly track objects.

Fig.7 shows results obtained for the TC-ODAL algorithms. After crossing with the trace of ID = 15 at frames 667, 672, and 677, the trace of ID = 18 is matched to the trace of ID = 20, which introduced overlaps in the trajectories. Finally, the trace of ID = 18 is lost at frame 679, and is truncated at that time. Meanwhile, Fig.8 shows that our algorithm can also track the target correctly.

Fig.9 shows the results for the TC-ODAL algorithm. We can see that the tracking of ID = 23 is wrong at frame 710, the trace of ID = 1 is mismatched, and is even replaced by that of ID = 23 at frame 716 and 722. At frame 725, the trace of ID = 1 is mismatched to another detection result, and the tracking of ID = 20 gives an error at the same time. Meanwhile, Fig.10 shows that although the test result obtained using our algorithm experiences interference at frame 710 near the trace of ID = 6, it can still be tracked steadily.

From Fig.11, we can see that the result using the TC-ODAL algorithm, the tracking of ID = 22 is affected by interference near frame 777. The trace of ID = 22 is occluded by that of ID = 23 at frame 788 near the instance of ID=24. Instances of ID = 24 appear beside that of ID = 16 at frame 792,

FIGURE 8. Test results for the second sequence from S2.L1 using our proposed algorithm.

FIGURE 9. Test results for the third sequence from S2.L1 using the TC-ODAL algorithm.

FIGURE 10. Test results for the third sequence from S2.L1 using our proposed algorithm.

FIGURE 11. Test results for the fourth sequence from S2.L1 using the TC-ODAL algorithm.

FIGURE 12. Test results for the fourth sequence from S2.L1 using our proposed algorithm.

and the trace of ID = 24 eventually replaces the object of ID = 16 at frame 795, which breaks the trajectory of the latter. Meanwhile, Fig.12 shows that our proposed algorithm can track objects without errors.

Test results for S2.L2 using our proposed algorithm are shown in Fig.13-14.

B. SOLVE MISSED DETECTION

Even though detection-based tracking methods can improve the tracking accuracy and precision on the bases of detection results, missed and error detection still exist because the currently used detector does not work correctly all of the time, which leads to a reduced tracking performance. Under

FIGURE 13. Test results for S2.L2 using our proposed algorithm.

FIGURE 14. Test results for S2.L2 using our proposed algorithm.

FIGURE 15. Real trajectory of objects.

FIGURE 16. Test trajectory with TC-ODAL.

such circumstances, it is important to ensure that objects are continue to be tracked well under missed and error detection.

In this paper, we consider the spatial information continuity between adjacent frames to solve the problem of missed detection. As shown in Fig.15, the true trajectory curve is in red, and missed the target near the street lamp. If we use a simple method to estimate the missed target position, an estimation error would be introduced, which may lead to the missed tracking of all instances in subsequent frames. While the algorithm continues to perform estimation, the object will be lost completely during tracking, as shown by the red curve in Fig.16, which deviates largely from the true trace in Fig.15. The above problem can be solved well using multi-object tracking with temporal information. As shown in Fig.17, the test result agrees well with the real trajectory.

FIGURE 17. Test trajectory with our algorithm.

V. CONCLUSIONS

In our paper, we proposed a multi-object tracking algorithm with temporal-spatial information and a trajectory of confidence based on detection and tracking. The entire process is divided into local and global association. Trajectories with high confidence are associated with the detection result of the current frame during local association, whereas trajectories with low confidence are associated with high confident trajectories and detection results that are not matched during global association. We determine the association result using

the combined model. Even though detection-based tracking methods can improve the tracking accuracy and precision based on the detection results, there are still missed and error detections because existing detectors do not work correctly all the time, leading to a reduced tracking performance. In such circumstances, it is important to ensure that objects continue to be tracked well. As a consequence, in this paper, we apply spatial and temporal information to the combined model. The improved model is more robust because of the correlation relationship of the spatial-temporal information, and it helps to solve the problem of missed detection. In addition, we measure the reliability of the spatial-temporal information using the principles of SCCM and PSR. We use a publicly available dataset to perform experiments using our proposed algorithm, and our results show that our algorithm can deal with problems such as missed detection, simultaneously improving the robustness of the track detector, hence, it is superior to many other popular algorithms.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 886–893.

[2] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[3] A. A. Butt and R. T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1846–1853.

[4] L. Leal-Taixe, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3542–3549.

[5] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.

[6] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. CVPR*, Jun. 2011, pp. 1273–1280.

[7] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Proc. CVPR*, Jun. 2011, pp. 1217–1224.

[8] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1918–1925.

[9] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2034–2041.

[10] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1815–1821.

[11] X. Song, J. Cui, X. Wang, H. Zhao, and H. Zha, "Tracking interacting targets with laser scanner via on-line supervised learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 2271–2276.

[12] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2953–2960.

[13] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 685–692.

[14] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1218–1225.

[15] W. F. Leven and A. D. Lanterman, "Unscented Kalman filters for multiple target tracking with symmetric measurement equations," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 370–375, Feb. 2009.

[16] K. Smith, D. Gatica-Perez, and J.-M. Odobez, "Using particles to track varying numbers of interacting people," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 962–969.

[17] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1805–1819, Nov. 2005.

[18] M. Yang, Y. Liu, L. Wen, Z. You, and S. Z. Li, "A probabilistic framework for multitarget tracking with mutual occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1298–1305.

[19] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, Jul. 1983.

[20] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979.

[21] S. Oh, S. Russell, and S. Sastry, "Markov chain Monte Carlo data association for multi-target tracking," *IEEE Trans. Autom. Control*, vol. 54, no. 3, pp. 481–497, Mar. 2009.

[22] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 127–141.

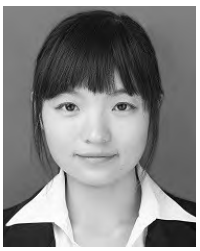
[23] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4902–4912.

[24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[25] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," in *Multimodal Technologies for Perception of Humans*. Berlin, Germany: Springer, 2008, pp. 3–34.

[26] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

JIAN WEI is currently pursuing the Ph.D. degree with the Jiangsu Province Key Laboratory on Image Processing and Image Communications, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include machine learning and visual tracking.



MEI YANG received the M.S. degree in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016. Her research interests include machine learning and visual tracking.

FENG LIU received the M.S. and Ph.D. degrees from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 1993 and 1997, respectively. He is currently a Full Professor with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing. His research interests include machine learning, visual tracking, and intelligent video analysis. He is a member of the IEEE Signal Processing Society.