The Visual Computer
https://doi.org/10.1007/s00371-021-02118-1

ORIGINAL ARTICLE

Localizing and tracking dense crowd of microbes by joint association and detection refinement

Ye Liu[1] · Shuohong Wang[2] · Jianhui Nie[1] · Hao Gao[1]

Accepted: 22 March 2021

Abstract
This paper presents a method for detecting and tracking large number of arbitrary-oriented and densely aggregated microbes from image sequences captured under microscope. We first propose an integral channel feature (ICF)-based detector which is able to localize the dense and arbitrarily oriented targets with low false positive rate. Then instead of treating target detection and tracking as two separated problems as many previous works did, we propose to refine the detection results in the data association process. The kinematic pattern of microbes is well modeled with the proposed integral sliding energy (ISE), which is combined with detection response in a hybrid cost function. Minimizing the cost function allows us to simultaneously select the true targets from the detections and to match the targets across two consecutive frames. Systematical experiments have been conducted to demonstrate the effectiveness of proposed method.

Keywords Object detection · Multi-object tracking · Microscopy image

1 Introduction

The collective behaviors of biological systems with large number of individuals have been attracting the research attention of scientists for many years. With a video camera, we can obtain the videos which record the behaviors of the systems. However, analyzing the data by manually labeling the organisms is time-consuming, labor-intensive and error-prone, since there are usually numerous targets with similar appearance on a single image. Recently, scientists are interested in studying the collective behavior in large number of microbes; they placed hundreds of paramecium in a culture dish and recorded videos under microscope trying to studying the way they interact with each other in moving. However, obtaining the position and motion of all the microbes by manual labeling is labor-costing.

In this paper, we address the problem of detecting and tracking hundreds of microbes in the hope of providing a convenient tool for scientists to study such behaviors. However, detecting and tracking targets in such dense crowd is a non-trivial tack. This is because the targets are in a dense crowd; the targets frequently interact with each other, making the appearance of the target apt to be corrupted by near-by targets and motion pattern of targets abrupt and irregular. We propose an automatic tracking system which is able to detect and track hundreds of microbes (Paramecium) in videos captured under microscope. Our tracking system consists of two modules which resolve two tasks, respectively: detection and tracking as shown in Fig. 1. Both of the two tasks are non-trivial; next we will describe the two problems in details respectively.

Detection problem The aim of the detection task is to precisely determine the position and orientation of every microbe. Since the microbes are densely aggregated, it is a challenging problem for conventional object detection methods for two issues: (1) since the microbes are arbitrarily oriented which is different from conventional object detection problems such as human detection [5] and face detection [26,30]in which the targets are almost up-right and (2) the population density is extremely high; thus, the microbes may

✉ Ye Liu
yeliu@njupt.edu.cn

Jianhui Nie
njh19@163.com

Hao Gao
gaohao@njupt.edu.cn

[1] School of Automation and Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China

[2] Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA

Published online: 27 April 2021

squeeze one another, making the microbes exhibit diversity and complexity in their appearance.

Tracking problem The goal of tracking task is to establish correspondences among the detections across time which is a multi-object tracking (MOT) problem. However, the extremely high population density makes the problem much more challenging than conventional MOT problems due to: (1) the appearance may change dramatically between two consecutive frames making appearance clue not so stable and (2) the mutual interaction among microbes makes the motion pattern complex which is difficult to model and predict.

The tracking system presented in this paper has overcome the above challenging problems to a considerable extent. The contribution of the paper is summarized as follows:

1. We propose a rotational ICF detector which is suitable for arbitrarily oriented object detection. Also, we integrate information from multiple heterogeneous sources to boot the discriminative ability of the features for detection.
2. We propose a novel tracking framework which jointly associates the detections across time and identifies the false positives in detection. This is achieved by a specific design in the cost matrix in the association stage.
3. In order to model the complex motion pattern of the microbes, we propose the integral sliding energy (ISE) which is based on the reasonable assumption that microbe moves in an energy-saving manner. The ISE models the motion pattern of the microbes accurately and enhances the successful rate of data association.

Systematical experiments have been carried out which demonstrates the effectiveness of both the proposed detection and tracking method.

2 Related work

Computer vision and machine learning methods are playing an increasingly important role in biology and medicine[3, 12,14,35]; this is partly due to the detection and tracking algorithms can provide accurate position and motion data of the subjects automatically which facilitates quantitative analysis[2,11,19]. For example, the three-dimensional trajectories of hundreds of fruit flies were reconstructed in[18]. Groups of bats were captured with multiple cameras and the 3D motion trajectories are recovered[33]. In [31], multiple zebra fish were detected and tracked automatically in a water tank. Cell population were detected and tracked in [3,11] with non-overlapping constraints.

From technical point of view, object detection and multiple object tracking are two important problems in computer vision. Ever since the Viola–Jones face detector [30] and HOG-SVM human detector [5] were proposed, various methods have been proposed which improve the feature extraction and the classification scheme. Among them, the integral channel feature framework[6] was proposed for human detection which is able to leverage the advantages of multiple sources of information while maintaining the computational efficiency. In order to overcome the shortcomings of non-maximum suppression in detecting groups of targets, global optimization was introduced to make decision more reasonably [19](our previous work). However, the output of stochastic optimization is not stable. Recently, deep learning-based methods have been successful in many problems [13,15]. In object detection, deep learning-based detector such as faster R-CNN [28], SSD[17], YOLO[24] *etc.,* have shown superior performance in object detection. However, these methods are not suitable for detecting arbitrary-oriented objects, besides, training a deep neural network detector requires huge amount of labeled data, which is difficult in our problem. Recently, the deep learning-based detectors have been modified to regress the orientation angle for arbitrary-oriented object detection [23,34] which perform well in object detection in remote sensing and scene text detection. However, they are not suitable for our problem due to the high-density of the microbes which is difficult for both object classification and bounding box (with orientation) regression tasks.

For single object tracking, numerous methods have been proposed in recent years to improve the learning model for appearance, for example online PCA [29], structured output SVM [8], sparse coding[21,22], correlation filter [9,10]. And recently, deep learning methods such as siamese network [1,7,16] have shown promising performance in single object tracking. Despite their success in single object tracking, they are not applicable to our problem since: (1) the targets are arbitrary oriented which is not considered in these methods, (2) the targets are similar in appearance and the appearance may change dramatically making the target may be more similar to nearby target than itself in previous frame.

For multiple object tracking, various data association methods have been proposed [20] to match the detections across time, such as MCMC[12], network flow[4], Hungarian algorithm[27], dynamic programming [36], *etc*. But these methods are not suitable for our problem due to the high density in population and the specific motion pattern of the targets. Recently, deep metric learning has been introduced into MOT to better exploit appearance coherency such as [32]; however, appearance coherency does not hold in our data since the microbes of our data may with a high probability change dramatically in appearance between two frames. Our data association strategy is still based on Hungarian algorithm, but by properly designing the cost during motion and allowing empty matching in both sides, the algorithm is able to simultaneously reject false positives and find correct correspondences across time.

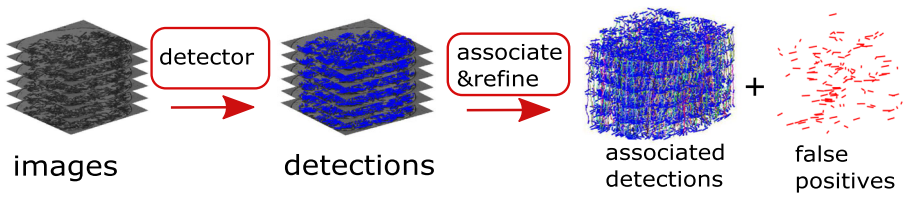

Fig. 1 The overall flowchart of proposed tracking system

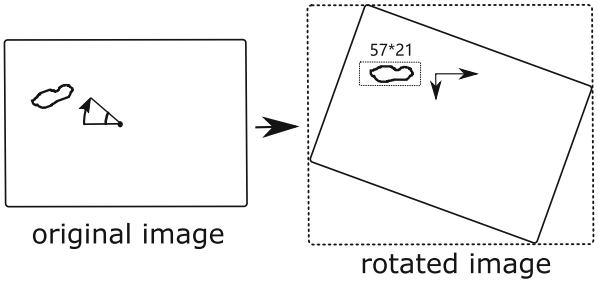

Fig. 2 Sliding windows framework for detecting arbitrary-oriented objects

3 Rotational integral channel feature detector

3.1 Rotational sliding window framework

In order to detect arbitrarily oriented targets, we modify the sliding window scheme by adding rotations to original image. As shown in Fig.2, the original image I is rotated around the image center by θ, a point (x, y) in original image I with image center (cx, cy) is transformed to (x', y') in transformed image I' with center (cx', cy') after rotation; this process can be written as:

$$\begin{pmatrix} x' - cx' \\ y' - cy' \\ 1 \end{pmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x - cx \\ y - cy \\ 1 \end{pmatrix} \quad (1)$$

Rotation angle θ is evenly spaced within a range from 0 to π. Then sliding window detection for horizontally oriented microbes is performed on each transformed image; the four corners of a detected are then rotated back to original image as shown in Fig.2. In our implementation, the sliding window size is 21×57. In this way, arbitrarily oriented target can be detected with detector trained with samples that in one orientation (horizontal). An alternative way to achieve arbitrarily oriented object detection is to rotate the sliding window; however, this will result in rotated rectangles; the sum of pixel values in them cannot be efficiently via integral image which will be introduced later.

3.2 Integral channel feature detector

We follow the integral channel feature (ICF) detection framework which is able to select the most discriminative features from heterogeneous sources of information. Each channel is a transformed image from original image with a kind of linear or nonlinear transform. The features are extracted as the sum of pixel values in rectangles with random locations and sizes in randomly chosen channels. Then the features are selected with Adaboost. By transforming a channel image into an integral image, rectangle features can be efficiently computed with several add/sub operations.

As shown in Fig. 3, we adopt the 4 types of channels for detection:

1. *Original image* Original image is used to keep the original gray scale information.
2. *Binary image* Since the microbes are darker than the background, we make use of this information by adding thresholded binary image as one of the channels.
3. *Gradient magnitude and histogram* We compute the gradient of image and compute the gradient magnitude as a channel. And for each pixel location, we put the pixels in a 6×6 window into a 6-bin histogram according to the gradient orientation. By taking out the values in the 6 bins, respectively, we can obtain the 6 gradient histogram channels.
4. *Difference of Gaussian (DOG)* Edges at different scale can be identified with DOGs, which are computed as the difference of images filtered by Gaussian kernels:.

$$DoG = G_{\sigma_1} * I - G_{\sigma_2} * I \quad (2)$$

where G_{σ_1} and G_{σ_2} are the two Gaussian kernels of different standard deviations σ_1 and σ_2, and $*$ is the convolution operator.

For an image I, a total number of 11 channels are adopted: original image, thresholded binary image, gradient magnitude, gradient histogram of 6 channels and 2 DOG channels; these channel images are termed as a image set $\{I_c | c = 1, 2, ...11\}$. Since the detection is performed on rotated images, these channels are generated from each rotated image. For each sliding window, a number of rectangles are extracted. A rectangle R_i can be determined by 5 parameters

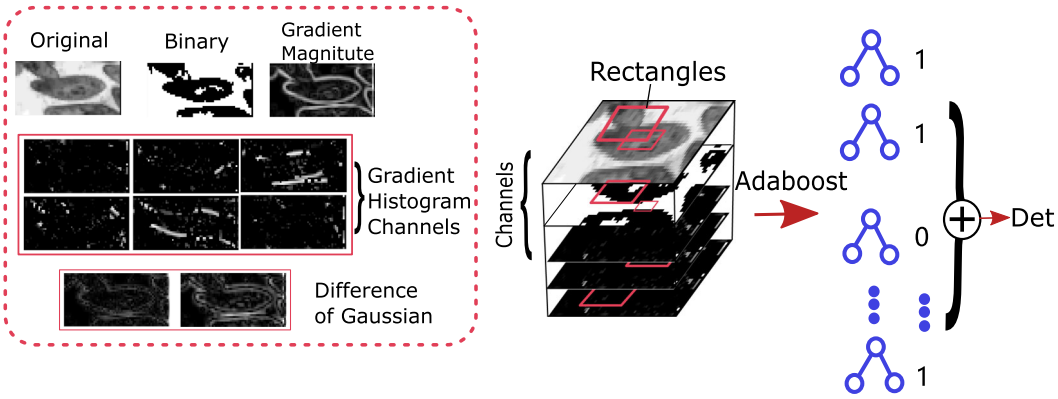

Fig. 3 The ICF detector with the 11 channels. After the features are selected, a set of depth-two decision trees are obtained. And the final detection response(Det) is the sum of the classification outputs of the decision trees

$(c_i, x_i, y_i, w_i, h_i)$, where c_i is the channel index, (x_i, y_i) is the coordinate of top-left corner, w_i and h_i are the width and height of the rectangle. And the feature value f_i is computed as the sum of pixel values in rectangle R_i of I_{c_i}. By randomly choosing the five parameters for Nf times, we can obtain a set of rectangle features for each sliding window.

We adopt the Adaboost framework [6,30] to select weak classifiers and classify the detection candidates. Each weak classifier is a depth-two decision tree which uses three rectangle features; as in [6], the three features are selected from the feature pool as the ones which achieve the lowest classification error. After one tree is constructed, the training samples are re-weighted and the next decision tree is trained. We collect a training dataset which consists of 982 positive samples and 5000 negative samples. Each positive sample is a 21×57 image, which are obtained by manually labeling the heads and tails of the microbes and then rotating the cropped image region to horizontal orientation. And the negative samples are obtained by randomly cropping image patches in labeled images which cannot be matched to any labeled target according to the Intersection over Union (IoU) score ($IoU > 0.5$); for two rectangles A and B, IoU is defined as $\frac{A \cap B}{A \cup B}$ which is the area of the intersection divided by the area of the union.

We generate a feature pool consists of $Nf = 10000$ rectangle features, and 200 depth-two decision trees are selected which uses 600 out of Nf rectangle features. A decision tree is trained as follows: (1) select one feature as the root node from the Nf features which gives the lowest classification error rate. Once selected, the optimal feature index and the corresponding threshold are kept. (2) Given the root feature index according its corresponding threshold, the dataset can be divided into two parts. For each part of the dataset, one feature is selected as the child node in the same way as in determining the root node in (1). Once two child nodes are determined, a depth-two decision tree has been build. After all the 200 decision trees have been trained, the training process ends.

3.3 Hard negative mining and NMS

After training for the first round, we further test the classifier on a set of 2000 negative samples which are not used in the first round. This is to find out hard negatives (samples that are negative but are classified as positive by the classifier). Then the hard negatives are added to the dataset to re-train the model. The final decision is made by voting the outputs of the decision trees. The detection score Det of a window is computed as the number of trees that make positive prediction as shown in Fig. 3.

After collecting all the candidate detections from rotated images of all the angles which are transformed back to the position and orientation in original image, a standard non-maximum suppression procedure is performed to remove duplicate detections. IoU is computed between two detections which are close to each other for a detected target A and a ground truth target B. If the IoU score is greater than 0.7, the detection with smaller detection score is discarded.

4 Joint association and detection refinement

4.1 The matching problem

The detector introduced in previous section is able to extract the locations and orientations of the targets in each image. In order to obtain the motion of the targets across frames, data association is performed to match the detections across time. Let $\mathcal{D}_t = \{d_t^i | i = 1, 2, ...N_t\}$ denote the detection set at t which is output by the detector, and d_t^i is the ith detection of all N_t detections. Given $\mathcal{D}_t$ and $\mathcal{D}_{t+1}$, the goal is to find a most reasonable matching $\mathcal{M} = \{(i, j) | d_t^i \in \mathcal{D}_t, d_{t+1}^j \in \mathcal{D}_{t+1}\}$, in which d_t^i in $\mathcal{D}_t$ corresponds to d_{t+1}^j in $\mathcal{D}_{t+1}$. By

defining certain cost function $C(i, j)$ and minimizing the total cost of $\mathcal{M}$, the optimum matching can be found using the Hungarian Algorithm. In our problem, two major problems must be addressed:

1. Since the targets are densely aggregated and similar in appearance, the first problem is how to define a cost function which is able to characterize the subtle differences between correct and incorrect correspondences. Two cues are considered for matching the targets: the appearance and the motion continuity. However, we observe that the appearance cue is not so stable that may lead to matching error, so it is not considered in our framework.
2. Both $\mathcal{D}_t$ and $\mathcal{D}_{t+1}$ inevitably contain false positives, so the second problem is how to identify the false positives in order to boost the matching accuracy.

4.2 Integral sliding energy

The key issue of our association scheme is the integral sliding energy, which models the dynamics of the target between two consecutive frames accurately. We assume the microbes move in an energy-saving manner; the total energy of the system is the sum of the energies of all individuals in the group. So data association aims at finding the matching that consumes the minimum total energy cost.

We assume a simplified mechanical model: the microbe moves and consumes energy against a resistance force. Let $fr(x, y)$ denote the resistance force at location (x, y) and $dis(x, y)$ denote the displacement at location (x, y), the total energy consumed by an individual target between two frames can be computed as the integration of the energy at every location inside the target's shape region D:

$$\begin{aligned} E &= \int\int_D fr(x, y)dis(x, y)dxdy \\ &= fr(x, y)\int\int_D dis(x, y)dxdy \end{aligned} \tag{3}$$

And we assume a constant resistance force at every location, so the energy is proportional to the integration of the displacement at every location in D.

Since it is difficult to obtain the precise shape region of each detection, a simplified model is adopted to represent the shape as a line segment with fixed length and two end points. As shown in Fig. 4, detection d_t^i is represented as line segment $l_t^i = \{e_t^{i,1}, e_t^{i,2}\}$. And its candidate matching correspondence d_{t+1}^j is defined likewise: $l_{t+1}^j = \{e_{t+1}^{j,1}, e_{t+1}^{j,2}\}$. Consider a point on the line segment l_t^i: $e_t^i(\mu) = e_t^{i,1} + \mu(e_t^{i,2} - e_t^{i,1})$ which corresponds to the point $e_{t+1}^j(\mu) = e_{t+1}^{j,1} + \mu(e_{t+1}^{j,2} - e_{t+1}^{j,1})$ on l_{t+1}^j, the energy cost at this point is proportional to the Euclidean distance it slides. And the energy cost of moving the line segment is the integral of energy on the whole line segment which we name as integral sliding energy (ISE):

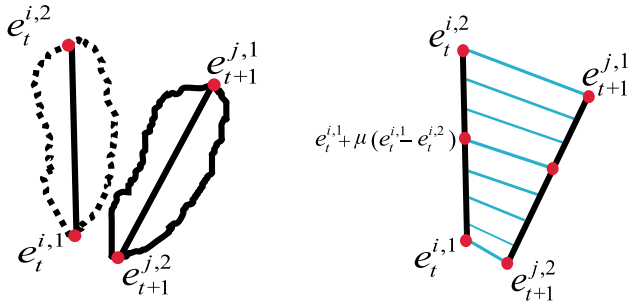

Fig. 4 Figure that illustrates the integral sliding energy. Left: two detections at t and $t+1$ are represented as two line segments. Right: ISE is computed as the sum of the length of blue lines

$$ISE(i, j) = \int_0^1 \left\| e_{t+1}^j(\mu) - e_t^i(\mu) \right\|_2 d\mu \tag{4}$$

For computational convenience, a line segment is discretized into 10 evenly spaced sample points which turns the integral into summation as shown in Fig. 4. Now the ISE energy can be written as:

$$ISE(i, j) = \sum_{n=0}^{10} \left\| e_{t+1}^j(n\Delta\mu) - e_t^i(n\Delta\mu) \right\|_2. \tag{5}$$

where $\|.\|_2$ computes the Euclidean distance (L2 norm) between two points.

4.3 Cost minimization

The ISE accurately models the energy cost by a microbe as it moves. Beyond that, the detection confidence is also incorporated into the cost function. This is because we expect the matching algorithm to be able to automatically select the detections with high confidence for matching. So the final cost function can be formulated as:

$$C(i, j) = \begin{cases} ISE(i, j) - \lambda(Det_t^i + Det_{t+1}^j) & if dc(i, j) < \sigma, \\ +\infty & otherwise. \end{cases} \tag{6}$$

where Det_t^i and Det_{t+1}^j are the detection response scores of detection d_t^i and d_{t+1}^j. λ is a weighting factor, and $dc(i, j)$ is the distance between the center points of two detections. If $dc(i, j)$ is greater than a threshold σ, the cost is set to $+\infty$. Detection pairs with higher confidence scores will have lower cost; thus, in the data association stage, the detections with high detection scores are more likely to be selected.

Note that in our cost function the appearance coherency across time is abandoned. This is because in our problem the

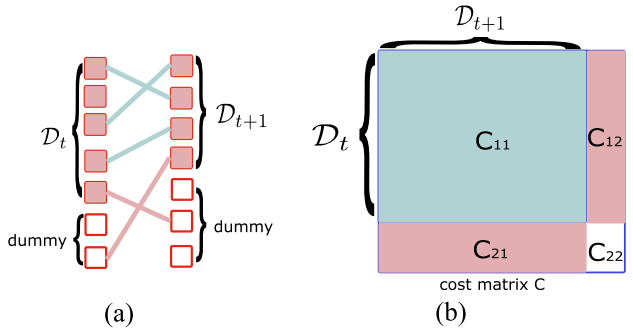

Fig. 5 Figure that illustrates the cost matrix. **a** Dummy objects are added to both sides to be matched, different colors indicate the cases that two matched items are detections (red) and one of the items is dummy (green). **b** The resulted cost matrix whose four block matrices corresponds to different cases of matching in (**a**)(see the text for details)

population density is extremely high, the microbes may be squeezed by the surrounding individuals, and its appearance may change dramatically. Thus appearance coherency is not reliable in tracking.

In our problem, the number of objects do not change; however, due to detection errors, the number of elements in $\mathcal{D}_t$ and $\mathcal{D}_{t+1}$ are not equal. A conventional approach to handle this is to add dummy elements to the smaller set, and the some of the elements in the larger set will be matched to the dummy elements. In our problem, false positives exist in both $\mathcal{D}_t$ and $\mathcal{D}_{t+1}$, and we expect the false positives in both sets to match dummy elements. So dummy elements are added to both $\mathcal{D}_t$ and $\mathcal{D}_{t+1}$ as shown in Fig. 5. And we set the cost of a detection matching a dummy element a constant ξ. After dummy elements are added, the number of elements at t and $t+1$ becomes equivalent as Nd_t. Let ρ denote the false positive rate of the detector, Nd_t is computed as:

$$Nd_t = \max(|\mathcal{D}_t| + \rho|\mathcal{D}_{t+1}|, \rho|\mathcal{D}_t| + |\mathcal{D}_{t+1}|) \tag{7}$$

where $|.|$ denotes the cardinal number of a set.

The cost of a detection matching a dummy element is defined set to a constant ξ. The cost matrix is a square matrix and can be partitioned into four block matrices: C_{11}, C_{12}, C_{21} and C_{22} as shown in Fig. 5. The elements in C_{11} are computed with Eq. 6. And the elements in C_{12} and C_{21} are set to ξ. The elements in C_{22} are set to $+\infty$. The cost matrix serves as input to the Hungarian algorithm. The detections in $\mathcal{D}_t$ or $\mathcal{D}_{t+1}$ that are matched to dummy elements are labeled as false positives. The algorithm is described in detail in Algorithm 1.

The hyper parameter λ is selected using the ground truth of two frames; we iterate a range of values to obtain an optimum value. The distance threshold is set empirically to be two times the average target length. The value of ξ is also set empirically as the one slighter greater than mean cost of a pair of matched detections since we expect the algorithm to punish those detections that are matched to dummy detection.

Input: Detection sets: $\mathcal{D}_t = \{d_t^1, d_t^2, ...d_t^{N_t}\}$ and $\mathcal{D}_{t+1} = \{d_{t+1}^1, d_{t+1}^2, ...d_{t+1}^{N_{t+1}}\}$
Output: Correspondances: $\mathcal{M} = \{(i,j)|d_t^i \in \mathcal{D}_t, d_{t+1}^j \in \mathcal{D}_{t+1}\}$, false positive sets: $\mathcal{O}_t \subset \mathcal{D}_t$ and $\mathcal{O}_{t+1} \subset \mathcal{D}_{t+1}$

```
Compute Nd_t using Eq. 7
for i=1:Nd_t do
    for j=1:Nd_t do
        if i ≤ N_t and j ≤ N_{t+1} then
            Compute C(i, j) using Eq. 6
            //elements in C_11
        else
            if i > N_t and j > N_{t+1} then
                C(i, j) = ∞
            else
                C(i, j) = ξ
                //elements in C_12 and C_21
            end
        end
    end
end
//cost matrix C has been computed
Generate matching vector: m ∈ R^{Nd_t} using Hungarian algorithm with C
for i=1:Nd_t do
    if i ≤ N_t and m_i ≤ N_{t+1} then
        add (i, m_i) to M
        //Matched detection pairs are kept in M
    else
        //Detections matched to dummy are identified as false positives
        if i > N_t and m_i ≤ N_{t+1} then
            add d_{t+1}^{m_i} to O_{t+1}
        end
        if m_i > N_t and i ≤ N_t then
            add d_t^i to O_t
        end
    end
end
```

Algorithm 1: Joint data association and detection refinement

5 Experimental results

5.1 Detection performance

The data were obtained by placing hundreds of paramecium under microscopy and videos were captured with a camera. The video resolution is 712 × 536. In order to train and evaluate the performance of the proposed detector, we manually label 9 images. A total of 1082 targets on 3 of the labeled images are cropped for training. The rest 6 images are also fully labeled which contains a total of 2214 targets. Each labeled or detected target is represented as a rectangle. And

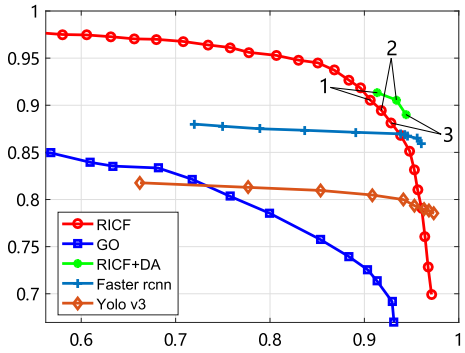

Fig. 6 The precision-recall curves of three methods: RICF (proposed detector), RICF+DA (proposed method with data association refinement) and GO (HOG-SVM detector with global optimization [19])

Table 1 Improvement on the Precision, Recall and f-score after data association(DA)

	Precision	Precision	Recall	Recall	f-score	f-score
no.	wo DA	DA	wo DA	DA	wo DA	DA
1	0.9066	0.914	0.905	0.913	0.906	0.914
2	0.918	0.934	0.895	0.906	0.906	0.92
3	0.9285	0.94	0.881	0.89	0.904	0.916

The rows correspond to the point pairs marked with 1, 2 and 3 in Fig. 6

we adopt the metric IoU for measuring the correctness of a detection. If the IoU between a detected target and a ground truth target is greater than 0.5, the detection is considered as correct. By choosing different threshold in detector, we can obtain different precision-recall (PR) values. We expect a method to be with both a high precision and a high recall.

Comparison with our previous work Figure 6 shows the precision-recall curves of the proposed detector (RICF) and the method in [19] using global optimization (GO); the proposed method has a remarkable improvement compared with [19].

Table 2 The mean successful rate of matching (SRM) of the four methods on every ten frames: 1–10, 11–20, 21–30

Frames	NC	ISE	DeepSort	Ours(NC)	Ours(ISE)
1–10	0.903	0.925	0.931	0.952	0.964
11–20	0.886	0.909	0.926	0.959	0.961
21–30	0.889	0.912	0.918	0.943	0.950
31–40	0.891	0.907	0.914	0.945	0.96
41–50	0.886	0.91	0.922	0.942	0.963

Table 3 Mean interventions per trajectory (MIPT) of four methods for trajectories with different lengths

Traj. len	NC	ISE	DeepSort	Ours(NC)	Ours(ISE)
20	2.64	2.57	2.37	2.22	2.15
30	3.35	3.24	3.15	2.75	2.67
40	4.04	3.92	3.62	3.27	3.15
50	4.64	4.51	4.11	3.76	3.60

Comparison with deep learning methods Conventional detectors including deep learning detectors are not suitable for arbitrarily oriented objects. And recently they were modified for detecting arbitrarily oriented scene texts and remote sensing targets [23,34]. We tried to test these method on our dataset, and however, they were not able to successfully generate normal detection results. We guess the cause for this is that the density of microbes is too high which makes it difficult to make neither classification nor regression predictions in the grids of feature map with downsampled resolution output by a convolutional neural network. We then considered replacing the ICF detector of our method with the deep learning detectors, which means that deep learning detectors are trained for horizontal microbes, and then, the image is rotated to detect arbitrarily oriented microbes, and finally NMS is

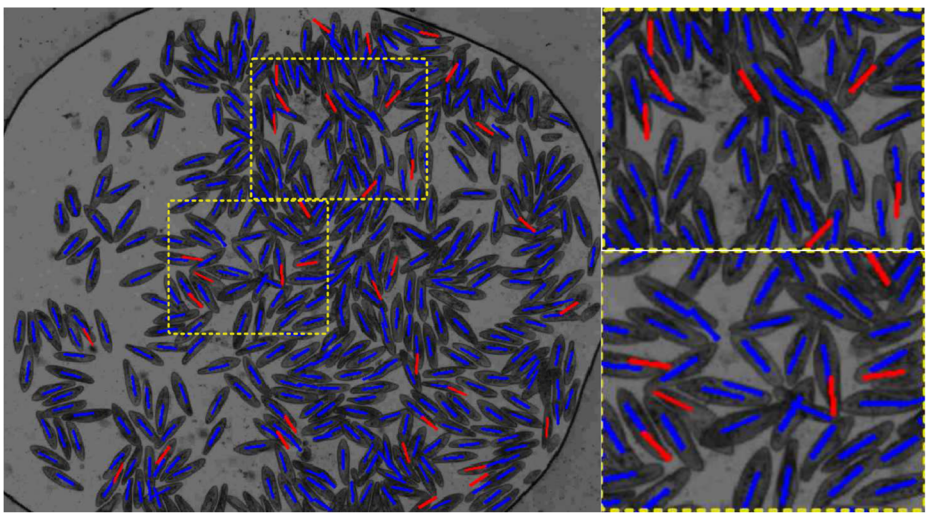

Fig. 7 The detection results of proposed detector. A total of 378 detections have been extracted. Red markers indicate the detections identified as false positives by data association. Among the 35 red markers 27 are true false positives

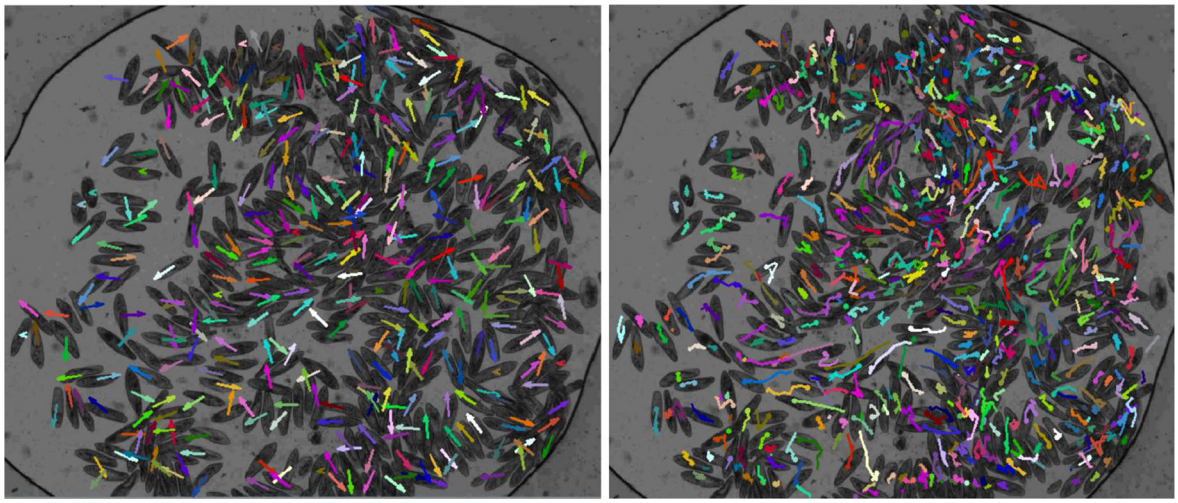

Fig. 8 Tracking result: 361 trajectories have been extracted. Left: the moving direction of the center point of each target is marked with a distinctively colored arrow, 343 (94.8%)of the targets are correctly matched. Right: the corrected 10 time-step trajectories of the targets

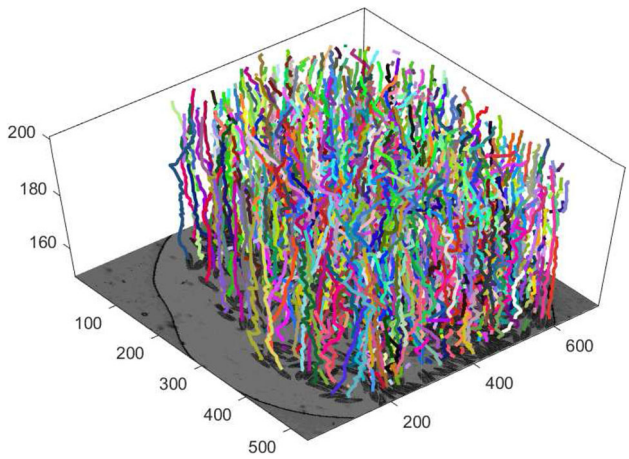

Fig. 9 The trajectories visualized in a 3D volume whose Z-axis is the frame number

applied. We evaluated two state-of-the-art deep detectors: faster RCNN [28] and Yolo V3 [25]. Both of the methods were able generate high-quality detection results, especially the faster RCNN. However, as shown in Fig. 6 a problem with both of the two methods is the recall score cannot be raised by lowering the threshold of detection confidence. This means the rate of missed objects is relatively high.

The role of refinement by data association We also show the performance of the detections refined by data association (RICF+DA) which gains further improvement in detection. We only plot three points of the PR curve of RICF+DA because the proposed data association can boost performance only when the detections at hand are already of relatively high quality with a small percentage of false positive rate. As in Table 1, the f-score of three points marked with 1, 2 and 3 in the PR curve has been boosted after data association. Figure 7 shows the detection results refined by data association. The detections with red mark are filtered out during data association. In this example, there are a total of 378 targets are detected on this image, among them 35 are identified as false positive, 27 out of 35 are true false positives, 8 are mislabeled. This demonstrates the effectiveness of our scheme of identification of false positives through data association.

5.2 Tracking performance

We first evaluate the performance of the proposed method in associating the detections between two consecutive frames. Success rate of matching (SRM) is computed as the number of correctly matched targets divided the total number of targets. Since the motion pattern of adjacent frames are similar, we divide a total of 50 images into five groups: 1–10, 11–20, 21–30, 31–40 and 41–50. The mean SRMs of four methods are calculated on these groups respectively. The performance of five data association methods is evaluated: DeepSort[32], association by finding nearest center (NC), association by finding smallest ISE (ISE), proposed data association method with distance between the center points instead of ISE (Ours(NC)) and the proposed method with ISE (Proposed). The results are listed in Table 2; the proposed joint association and detection refinement strategy can substantially boost the matching accuracy if we compare the results of NC and ISE to Ours(NC) and Ours(ISE). From the table, we can also observe that proposed integral sliding energy (ISE) can further boost the performance compared with nearest-center(NC), which demonstrates that the ISE models the motion pattern more effectively. DeepSort does not perform well in our dataset; this is possibly due to the appearance coherency does not hold in our dataset, since the microbes may squeeze each other; thus, the deep metric

learning which is the method's key component makes little effect on the tracking performance.

Since our method aims at providing accurate position and motion data for scientists with as few manual interventions as possible, we propose a new metric for evaluating the tracking performance: mean interventions per trajectory (MIPT) which counts the mean times of interventions to get a correct trajectory. Human intervention is required if the current trajectory starts to follow a wrong detection or ends at current position. We obtain ground truth trajectories of 50 time-step long by manual labeling and then evaluate the performance on trajectories truncated to different lengths: 20, 30, 40 and 50.

The results are shown in Table 3. Either for simple nearest neighbor strategy or proposed data association method, the performance boosts if the ISE is used instead of center distance, which proves the effectiveness of ISE. And the proposed data association method outperforms the simple nearest neighbor by a large margin whether with ISE or center distance. Figure 8 gives an example of the tracking results; the direction of motion in center point is visualized in the left image, and the corrected trajectories are visualized in the right image. In this example, 361 trajectories have been extracted; in the left image 343 (94.8%)of the targets are correctly matched. All the trajectories are visualized in a 3D volume as in Fig. 9.

6 Conclusion

We propose in this paper a method which is able to detect and track hundreds of densely aggregated and arbitrarily oriented microbes. A rotational ICF detector is proposed to detect the densely aggregated and arbitrary oriented microbes. And we propose a joint data association and detection refinement framework with a novel integral sliding energy incorporated which is able to match the targets across time while eliminating false positives in detections at the same time. Experimental results have demonstrated the effectiveness of proposed method in both localizing the targets and tracking the targets through time.

7 Future work

Our future work will focus on developing a software with a user-friendly graphical user interface (GUI), which is convenient for scientists to studying the behaviors of microbes in dense crowds. We also consider improving the deep learning detectors for better detecting the densely aggregated and arbitrarily oriented objects and also introducing deep learning techniques into the matching stage to further boost the detection and matching accuracy.

Acknowledgements The research work of this paper is sponsored by Natural Science Foundation of China under Grant 61602255 and 61931012. The authors would like to thank Prof. T. Vicsek and his group for providing the research data of this work.

Declaration

Conflict of interest Authors of this manuscript declare that they have no conflict of interest.

References

1. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865. Springer (2016)
2. Betke, M., Wu, Z.: Data association for multi-object visual tracking. Synth. Lectures Computer Vis. **6**(2), 1–120 (2016)
3. Bise, R., Sato, Y.: Cell detection from redundant candidate regions under nonoverlapping constraints. IEEE Trans. Med. Imaging **34**(7), 1417–1427 (2015)
4. Butt, A.A., Collins, R.T.: Multi-target tracking by lagrangian relaxation to min-cost network flow. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2005)
6. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: British Machine Vision Conference (2009)
7. Dong, X., Shen, J.: Triplet loss in siamese network for object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 459–474 (2018)
8. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S.L., Torr, P.H.: Struck: structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intel. **38**(10), 2096–2109 (2015)
9. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2014)
10. Hu, H., Ma, B., Shen, J., Sun, H., Shao, L., Porikli, F.: Robust object tracking using manifold regularized convolutional neural networks. IEEE Trans. Multimed. **21**(2), 510–521 (2018)
11. Kanade, T., Yin, Z., Bise, R., Huh, S., Eom, S., Sandbothe, M.F., Chen, M.: Cell image analysis: algorithms, system and applications. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 374–381. IEEE (2011)
12. Khan, Z., Balch, T., Dellaert, F.: An mcmc-based particle filter for tracking multiple interacting targets. In: European Conference on Computer Vision, pp. 279–290. Springer (2004)
13. Kuanar, S., Athitsos, V., Mahapatra, D., Rao, K., Akhtar, Z., Dasgupta, D.: Low dose abdominal ct image reconstruction: an unsupervised learning based approach. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1351–1355. IEEE (2019)
14. Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K.R.: Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2576–2580. IEEE (2018)
15. Kuanar, S., Rao, K., Bilas, M., Bredow, J.: Adaptive cu mode selection in hevc intra prediction: a deep learning approach. Circuits, Syst., Signal Process. **38**(11), 5081–5102 (2019)
16. Liang, Z., Shen, J.: Local semantic siamese networks for fast tracking. IEEE Trans. Image Process. **29**, 3351–3364 (2019)

17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
18. Liu, Y., Wang, S., Chen, Y.Q.: Automatic 3d tracking system for large swarm of moving objects. Pattern Recognit. **52**, 384–396 (2016)
19. Liu, Y., Wang, S., Gao, H., Wang, B.: Detecting dense crowds of microbes from microscope images in a global optimization framework. Optik **127**(1), 76–80 (2016)
20. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X., Kim, T.K.: Multiple object tracking: A literature review. arXiv preprint arXiv:1409.7618 (2014)
21. Ma, B., Hu, H., Shen, J., Liu, Y., Shao, L.: Generalized pooling for robust object tracking. IEEE Trans. Image Process. **25**(9), 4199–4208 (2016)
22. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2259–2272 (2011)
23. Ming, Q., Zhou, Z., Miao, L., Zhang, H., Li, L.: Dynamic anchor learning for arbitrary-oriented object detection. arXiv preprint arXiv:2012.04150 (2020)
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
25. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
26. Rehman, B., Ong, W.H., Tan, A.C.H., Ngo, T.D.: Face detection and tracking using hybrid margin-based roi techniques. Vis Computer **36**(3), 633–647 (2020)
27. Reilly, V., Idrees, H., Shah, M.: Detection and tracking of large number of targets in wide area surveillance. In: European Conference on Computer Vision, pp. 186–199. Springer (2010)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
29. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Computer Vis. **77**(1–3), 125–141 (2008)
30. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Computer Vis. **57**(2), 137–154 (2004)
31. Wang, S.H., Cheng, X.E., Qian, Z.M., Liu, Y., Chen, Y.Q.: Automated planar tracking the waving bodies of multiple zebrafish swimming in shallow water. PloS One **11**(4), e0154714 (2016)
32. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649. IEEE (2017)
33. Wu, Z., Hristov, N.I., Hedrick, T.L., Kunz, T.H., Betke, M.: Tracking a large number of objects from multiple views. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1546–1553. IEEE (2009)
34. Yang, X., Yan, J.: Arbitrary-oriented object detection with circular smooth label. In: European Conference on Computer Vision, pp. 677–694. Springer (2020)
35. Zhao, J., Wang, S.H., Liu, X., Liu, Y., Chen, Y.Q.: Early diagnosis of cirrhosis via automatic location and geometric description of liver capsule. Vis. Computer **34**(12), 1677–1689 (2017)
36. Zou, D., Zhao, Q., Wu, H.S., Chen, Y.Q.: Reconstructing 3d motion trajectories of particle swarms by global correspondence selection. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1578–1585. IEEE (2009)

Ye Liu received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2013, and the B.Eng. degree in computer science from Tongji University, Shanghai, China, in 2007. Since 2013, he has been a faculty member with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include object detection and tracking, medical image analysis etc.

Shuohong Wang received her Ph.D. degree from Fudan University, China, in 2017 and her B.Eng. degree from East China University of Science and Technology, China, in 2012. She is now a postdoc at Harvard University. Her research interests include biomedical image analysis and machine learning.

Jianhui Nie received the B.Eng. degree in automation and the Ph.D. degree in control science and engineering from the Dalian Maritime University, Dalian, China, in 2007 and 2012, respectively. He is currently an Associate Professor with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include three-dimensional reconstruction and computer graphics.

Hao Gao received the M.Sc. and Ph.D. degrees in computer science from Jiangnan University, Wuxi, China, in 2006 and 2009, respectively. From 2009 to 2011 and 2012 to 2013, he was a Postdoctoral Fellow with Tsinghua University and City University of Hongkong, respectively. He is currently a Professor with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China. He has authored or coauthored more than 50 international journal and conference papers. His current research interests include evolutionary algorithms and computer vision.