Optik 127 (2016) 76–80

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Detecting dense crowds of microbes from microscope images in a global optimization framework

Ye Liu[a,*], Shuohong Wang[b], Hao Gao[a], Baoyun Wang[a]

[a] School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] School of Computer Science, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 13 November 2014
Accepted 25 September 2015

Keywords:
Object detection
Microscope imaging
Simulated annealing

ABSTRACT

This paper addresses the problem of detecting a large number of densely aggregated and arbitrary oriented targets from microscope images. Scientists have long been interested in the collective behavior exhibited by aggregated organisms, and analyzing the images of hundreds of single-celled organisms may reveal the underlying biological mechanisms and biomechanics; however automatically retrieving the position and orientation of each individual in such high-density crowd is challenging and difficult for existing object detection methods. We propose in this paper a method that is able to solve this problem effectively. Contrary to conventional sliding window based detection methods that make decisions based on only local image features, the proposed method seeks to make decisions from a global perspective. Detection problem is formulated as a global optimization with all the interactions and local cues integrated in one objective function. And maximizing the objective function yields much more reasonable results. Systematical experiments have been carried out and the results demonstrate the high performance of the proposed method.

1. Introduction

The collective behaviors exhibited by the aggregation of large number organisms of the same species have attracted the research attention from scientists for a long time. Recently, researchers study this kind of behavior from simple one-celled organisms such as paramecium. They captured videos of hundreds of paramecia under microscope, trying to reveal how they interact with each other and how the interactions finally contribute to their collective behaviors. However, the lack of effective technique to automatically detect these organisms has been a stumbling block for their research, as finding these targets by manual labeling is time-consuming.

From technical perspective, automatic detection is a non-trivial task, the reason is: (1) In high density crowd, the space between pair of individuals is limited and interactions commonly happen among targets which corrupts the appearances of targets on image. (2) The resolution of the captured image is relatively low, so each targets may take up only several hundreds of pixels, thus many of visual details of target's appearance is lost. (3) The texture and shape of targets vary significantly, so it is difficult to train a discriminant model that is able to account for all the variations (see Fig. 1).

Although object detection is a well-researched topic during the past two decades, most of the work on face detection [1] or pedestrian detection [2–5] or generic objects [6–8] focuses on the detection of one dominant object or a few isolated objects. Several previous work have addressed the problem of detecting dense crowd of targets. For example, in [9] human heads were detected in crowded scene which may contain tens of humans in a density-aware optimization framework. In [10], hundreds of cells are detected and tracked. The configuration of multiple occluded human was optimized in [11] making the detector robust against occlusion. Small number of interacting cells are segmented in [12]. Stochastic optimization have been studied for detecting crowds of dot-like targets [13] and partially occluded human [14]. Multiple targets with minor interactions are tracked with a MCMC based particle filter in [15]. And in [16], hundreds of flying animals moving in 3D space are detected and tracked. Multiple small targets are detected and tracked in [17]. Isolated moving object can be detected by effectively modeling the background [18,19]. However, these methods are not applicable to our problem due to the distinctive appearance characteristics and interaction mode among targets.

We propose in this paper a novel method that is able to detect hundreds of arbitrary oriented and frequently interacting targets effectively. We first extend the sliding window based detection

* Corresponding author. Tel.: +86 15062272038.
E-mail addresses: yeliu@njupt.edu.cn (Y. Liu), sh_wang@fudan.edu.cn (S. Wang), gaohao@njupt.edu.cn (H. Gao), wangby@njupt.edu.cn (B. Wang).

http://dx.doi.org/10.1016/j.ijleo.2015.09.218
0030-4026/

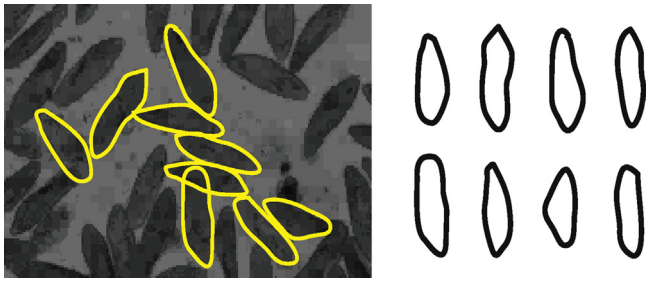

Fig. 1. The appearances (shape and texture) of the targets vary significantly among different individuals.

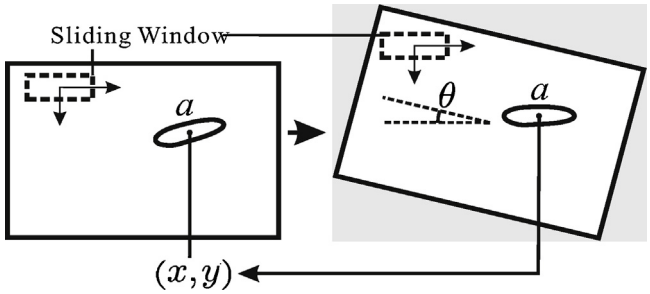

Fig. 3. The extended HOG-SVM detector for arbitrary oriented objects.

framework for detecting arbitrary oriented objects. But instead of making decisions with the detector response which reflects only local information, we propose a global optimization framework in which all the mutual interactions as well as the local information involving all the targets are modeled in one objective function, maximizing which yields more reasonable results than local non-maxima suppression methods. An overview of the method is shown in Fig. 2. Systematical experiments have been carried out to evaluate the performance of the proposed method.

2. Extended HOG-SVM detector

The HOG-SVM detector was first proposed in [2]. Despite its great success in human detection, it has problems if directly applied without modification to the problem here. In order to detect arbitrary oriented objects, we add an additional dimension to the space to be searched: (x, y, θ), which means the original image is rotated by θ around the image center and then conventional sliding window procedure is carried out on every rotated image (see Fig. 3). We tried original HOG descriptor in [2] but it did not perform well, we think this is probably because the targets are too small for conventional HOG which was designed for much bigger objects. So we made some modifications to it: we use a 5 by 6 cell and a 9 direction bins, the vertical and horizontal cell stride are both 3 pixels (no blocks and normalization here). We use linear SVM as the classifier. Each output of the detector can be represent as (x, y, θ, r), where r is the detection confidence output by SVM, (x, y) is the location in original image computed from an inverse transform (Fig. 3). After these modifications, the detector with revised HOG feature performs better than the one with original HOG, and also the detector is able to detect arbitrary oriented targets, but due to the challenges mentioned previously, the detection candidates still contain lots of false positives. Here we also considered template matching as the sliding detector, but we find it less effective as HOG-SVM detector, this is elaborated in detail in the experiment section.

Non-maxima suppression methods can be used to remove the redundant detections and some of the false positives, but in such crowded scene, such methods suffer from over-suppression or under-suppression, because such methods make decision based on the information from a local neighborhood which is insufficient to account for the complex interactions among targets.

If we consider the problem from a global perspective with all the interactions as well as all the local information of all the targets taken into consideration, we shall make more reasonable decisions. So the detection problem can be formulated as an optimization problem: Let p denote a n dimensional binary vector whose ith entry $p(i)$ equals 1 if the ith detection truly exists and 0 otherwise, for example $p = (10010)$ denotes there are 5 candidates and the 1st and 4th exist. Then a global objective function $E(p)$ which models multiple sources of information from all the candidate targets is maximized with respect to p:

$$p = arg\max E(p). \tag{1}$$

The optimization can also be viewed as filtering out false positives and duplicated detections from the candidate targets to make the objective function maximized.

3. Multi-cue objective function

3.1. Detector response

The first cue that is considered in the objective function is the detector response output by the sliding window detector in the previous section. Although the detection results from the detector are far from perfect, they are still important local cues that indicate possible target locations. Generally, it reflects to some extend the probability of the existence of a detected candidate from a local perspective. So the first term of the objective function is summation of the responses of all the candidate targets that are assumed to be real. Let $\{S_i | i = 0, 1, 2, \ldots, N\}$ denote a set of N candidates from

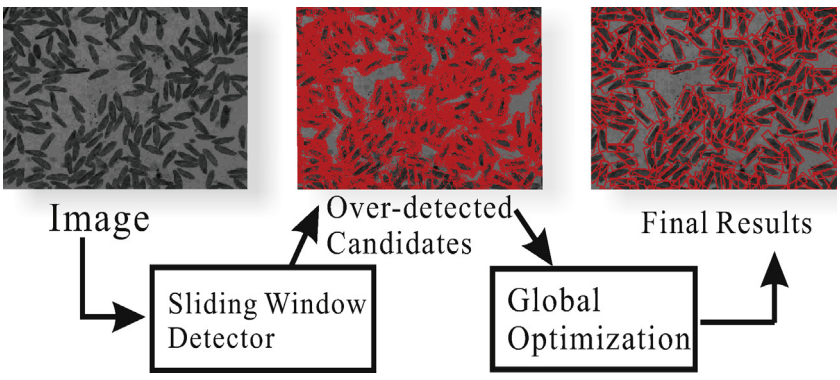

Fig. 2. Overview of proposed method.

the detector where $S_i = (x_i, y_i, \theta_i, r_i)$, the first term of the objective function can be written as:

$$E_{\text{det}} = \sum_{i=1}^{N} p(i)r(i), \tag{2}$$

where $p(i)$ is 0 or 1. Maximizing this term alone is unreasonable because we can get the maximum value by simply setting all the entries of p to 1. This can be interpreted as that we encourage more candidates with high confidence to be true. In fact, if the other two cues are considered, the number of true targets can be controlled and a balance can be reached.

3.2. Mutual interaction

The targets are constrained in a flat dish, so they may collide and squeeze with each other, but they seldom cross over each other. The objective function should be able to model such kind of interactions among targets.

A weighted template T is first learned from positive training samples. It is called weighted template because for each position of the template there is a weight value indicating how confident of that position belonging to the foreground area. Thus for a candidate detection S_i, we can predict a rectangle by applying an rigid transform R_i to the bounding rectangle of the weighted template T, the transform R_i is determined by the location and orientation (x_i, y_i, θ_i). Let $O_{i,j}$ denote the overlap region of the predicted rectangular regions of target i and target j, then the pair-wise interactions cost is defined as:

$$E_{\text{int}} = \sum_{i,j} \sum_{X \in O_{i,j}} T(R_i^{-1}(X)) + T(R_j^{-1}(X)), \tag{3}$$

where $X = (x, y)$ is a point in overlap region of target i and j, and $R_i^{-1}(\cdot)$, $R_j^{-1}(\cdot)$ are the inverse transform of $R_i(\cdot)$ and $R_j(\cdot)$ respectively. This cost in a single overlap position is the sum of the weights in its two back-projected positions in template T. If the overlap position is more likely to be a foreground position, a higher cost will be added to E_p. This is illustrated in Fig. 4(a), the overlap region is bounded with yellow line and corresponding areas on the weighted template is also plotted. The case on the right side has a larger interaction cost.

The weighted template is obtained by analyzing the thresholded binary images of the positive training samples $\{B_1, B_2, \ldots, B_{N_1}\}$, the probability of a pixel location (x, y) being foreground is the number of samples that has 1 value at (x, y) divided by the total number N_1:

$$P_f(x, y) = \sum_i \frac{B_i(x, y)}{N_1}, \tag{4}$$

and the corresponding weight at (x, y) is:

$$T(x, y) = \exp[\mu P_f(x, y)]. \tag{5}$$

where μ is set to 5 in our experiments.

3.3. Foreground–background Segmentation

The background can be subtracted using simple image thresholding, and our objective function also takes this important information into account. The basic motivation is to let the targets cover as much foreground area as possible. Recalled the weighted template T, we can get a binary target template by setting a weight threshold (for example 0). So given a binary vector p, we can predict a binary image I_p by sequentially placing a target template at (x_i, y_i, θ_i) on an all-zero image, Fig. 4(b) shows a simple example of this process. The third term of the objective function is defined as:

$$E_{seg} = \frac{\left|\{(x, y) | I_b(x, y) = I_p(x, y)\}\right|}{\left|\Omega\right|}. \tag{6}$$

where $|\cdot|$ measures the cardinal number of a point set, and Ω is all the pixels of thresholded binary image I_b. This term encodes the similarity between synthetic image and segmented image.

The final objective function is the weighted sum of the three terms:

$$E(p) = E_{\text{det}} + \alpha E_{\text{int}} + \beta E_{seg}, \tag{7}$$

where α is negative number because the interaction cost should be as low as possible.

4. Optimization

Maximizing the above function with respect to p is highly complex, as the configuration space of p is 2^n (if there are 5000 detection candidates then the space to be searched is 2^{5000}). And the gradient of the function is difficult to obtain, so stochastic optimization is considered. Among the stochastic optimization methods, simulated annealing (SA) is simple and typically powerful for finding a good solution of complicated objective functions in an acceptable amount of time, and it is suitable for discrete optimization problem (the variables in this problem is a vector of binary numbers). New solution p_{i+1} is generated by randomly flipping an entry (each entry with equal probability) in p_i, if the new solution results a larger $E(p)$, it is accepted, otherwise it is accepted in an acceptance probability which decreases with time [20].

At the beginning, n_0 targets out of N candidates are selected randomly as initialization, where n_0 is approximated as the area of segmented foreground divided by the average area of one target. Then in each iteration, an entry of p is flipped, which means a candidate is eliminated or added to the current true target set.

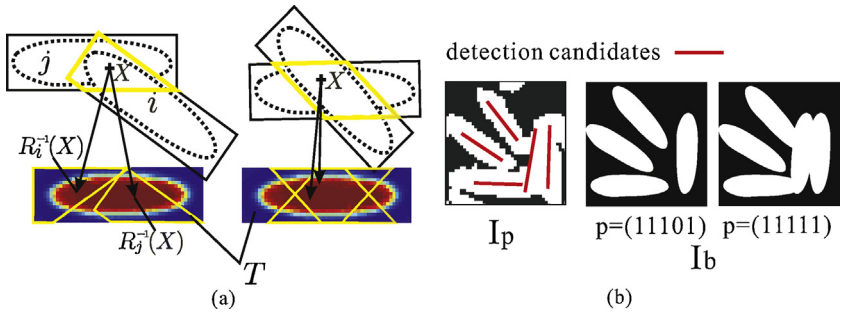

Fig. 4. Two different cues. (a) Different interaction costs. The overlap area (bounded by yellow line) is back-projected onto the weighted template. The cost on the right is larger than the left one. (b) Segmentation cue. Left: segmented image and detections. Right: synthetic binary images with different p. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

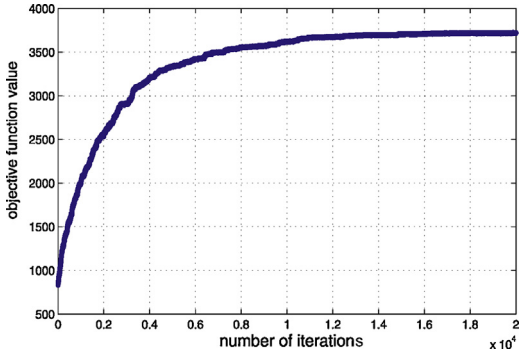

Fig. 5. The value of $E(p)$ changes with iteration number.

Fig. 5 shows the change of $E(p)$ with the number of iterations. After a dramatic increase at the beginning, the objective function gradually converges to a maximum number which is much larger than the one at the beginning. Before the optimization, a threshold is set to control the number of candidates that are passed into the optimization stage, tuning the threshold will result a trade-off between precision and recall.

5. Experimental results

5.1. Data and methods

The data used in the experiments were collected by scientists using a video camera. About 300 paramecia were put into a flat dish and observed under a microscope. The resolution of captured image is 708 × 532, we select 9 images out of a total of 700 for evaluation, the targets in all these 9 images are manually labeled (more than 2000 targets in all). The positive training samples are also obtained by manual labeling another set of images, from which 240 samples are selected. And the negative samples are selected randomly from the labeled images, and the positive samples in the selection are manually picked out and eliminated. The negative samples are determined by bounding rectangle comparison: if the rectangle of a sample is not covered by a labeled rectangle (the overlap area is less than 0.7) it is labeled as a negative sample.

Like in [2], after training for the first time, the detector is retrained with false positive samples added to the positive samples. The final number of negative samples is 3263.

We can see in Fig. 5 that the optimization converges after about 15k iterations. The maximum iteration number is set to 20k to ensure a convergence.

As most of existing object detection methods can not be applied in this problem, nor can we find a method that specially solve this same problem, so we implemented a conventional non-maxima suppression (termed NMS) method for comparison. If the overlap area between two rectangles is above 0.7 of the area of one rectangle, they are considered to be adjacent. In this way, a graph can be constructed with vertices representing the detections and edges representing the adjacency. Then for each connected component of the graph, the detection with the highest response r is kept and other detections of the connected component are discarded. Before the graph is constructed, a threshold is set to filter out the detection with low response, tuning this parameter will result in a trade-off between precision and recall. We follow the criterion of the PASCAL VOC Challenge: a successful detection is counted if the overlap area between a detection bounding box and a ground truth bounding box is more than 50% of one bounding rectangle.

Apart from the HOG-SVM detector, we also implemented a template matching (TM) method for generating detection candidates. The image is also rotated, and a template is matched

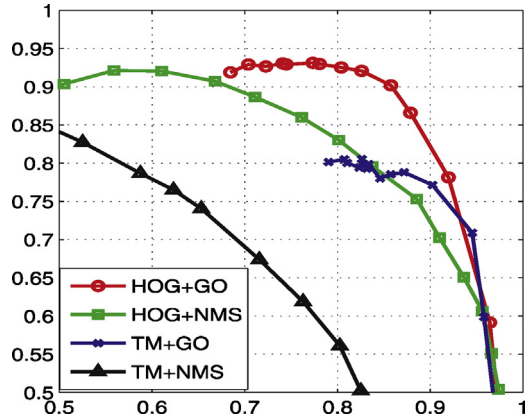

Fig. 6. PR curves of the four methods.

on the rotated images using a fast template matching algorithm [21]. A set of detections can be obtained by setting a threshold to the matching score (normalized cross-correlation). And these detection candidates are further filtered respectively using two methods: non-maxima suppression (TM+NMS) and the previously introduced global optimization (TM+GO). The difference between TM+GO and HOG+GO is the detector that generates the initial detection candidates.

5.2. Performance

The precision-recall curves of the four methods are shown in Fig. 6. The method with HOG detector and global optimization (HOG+GO) performs best. We notice that template matching perform poorly with non-maxima suppression but if the global optimization procedure is applied the performance is significantly enhanced, which demonstrates the effectiveness of the proposed global optimization framework.

In order to better explore the preference of the methods with different target density, for each image, we randomly select 10 sub-images of size 150 × 150 covering nearly the whole image (except those regions that are too crowded even for human to distinguish), the evaluation are carried out on these sub-images to investigate the performance of the methods on different density data.

The performances under different target densities are evaluated, as the target numbers on the sub-images are different. We divide the range of target numbers into 3 intervals: 14–19, 20–25,. . ., 26–30, which are labeled as 1, 2, 3 respectively, and for each interval, the average miss rate (1-recall) and f-measure (the harmonic mean of precision and recall) of four methods on sub-images versus the target density are computed (see Fig. 7). The parameters of the four methods are selected as the ones with maximum f-measure. The miss rates (Fig. 8) of all the four

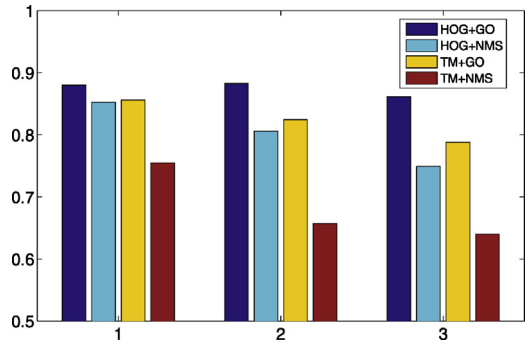

Fig. 7. F-measures of the four methods on data with different densities.

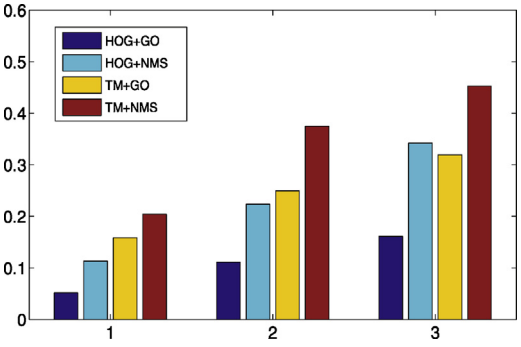

Fig. 8. Miss rates of the four methods on data with different densities.

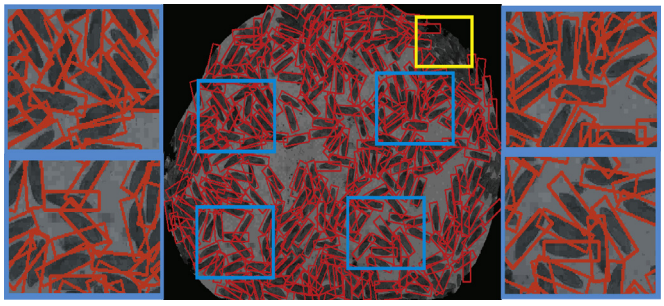

Fig. 9. Qualitative detection results. The images on the bottom and right are enlarged version of ROIs of the bigger image.

methods increase with the target density. HOG+GO has a higher and relatively stable f-measure than the rest methods, which to some extend reflects the robustness of the proposed method.

Some qualitative results are given in Fig. 9. The small images on the right and bottom show the enlarged areas in the image. The method fails to give correct detections in the yellow boxed area. We found that it is difficult even for human to distinguish the targets here, so the targets in this area are not counted for evaluation. The global optimization is effective in eliminating false positives while allowing the targets to be closely adjacent.

6. Conclusion

We propose in this paper a novel method that is able to detect dense crowd of arbitrary oriented microbes under microscope. The method determines the existence of over-detected candidates in a global optimization framework. Experimental results show satisfactory performance of the method on challenging datasets with hundreds of closely interacting targets.

Acknowledgements

The research work of this paper is sponsored by NUPTSF (Grant Nos. NY213093 and NY213167). The authors would like to thank Prof. T. Vicsek and his group for providing the research data of this work, and Dr. Haishan Wu for helpful discussions.

References

[1] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, IEEE, Vol.1, 2005, pp. 886–893.
[3] X. Wang, T.X. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in: ICCV, IEEE, 2009, pp. 32–39.
[4] J. Wu, C. Geyer, J.M. Rehg, Real-time human detection using contour cues, in: IEEE International Conference on Robotics and Automation (ICRA), 2011, pp. 860–867.
[5] S. Paisitkriangkrai, C. Shen, J. Zhang, Performance evaluation of local features in human classification and detection, IET Comput. Vis. 2 (4) (2008) 236–246.
[6] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[7] J. Gall, A. Yao, N. Razavi, L. Van Gool, V. Lempitsky, Hough forests for object detection, tracking, and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2188–2202.
[8] W. Chang, S.-Y. Lee, Description of shape patterns using circular arcs for object detection, IET Comput. Vis. 7 (2) (2013) 90–104.
[9] M. Rodriguez, I. Laptev, J. Sivic, J.-Y. Audibert, Density-aware person detection and tracking in crowds, in: ICCV, IEEE, 2011, pp. 2423–2430.
[10] K. Li, T. Kanade, Cell population tracking and lineage construction using multiple-model dynamics filters and spatiotemporal optimization, in: International Workshop on Microscopic Image Analysis with Applications in Biology, 2007.
[11] S. Rujikietgumjorn, R. Collins, Optimized pedestrian detection for multiple and occluded people, in: CVPR, IEEE, 2013.
[12] Z. Wu, D. Gurari, J.Y. Wong, M. Betke, Hierarchical partial matching and segmentation of interacting cells, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI, Springer, 2012, pp. 389–396.
[13] X. Descombes, R. Minlos, E. Zhizhina, Object extraction using a stochastic birth-and-death dynamics in continuum, J. Math. Imaging Vis. 33 (3) (2009) 347–359.
[14] T. Zhao, R. Nevatia, in: CVPR, IEEE, Vol. 2, Tracking multiple humans in crowded environment (2004), pp. II-406.
[15] Z. Khan, T. Balch, F. Dellaert, MCMC-based particle filtering for tracking a variable number of interacting targets, IEEE Trans. Pattern Anal. Mach. Intell. 27 (11) (2005) 1805–1819.
[16] Y. Liu, H. Li, Y.Q. Chen, Automatic tracking of a large number of moving targets in 3D, in: ECCV, Springer, 2012, pp. 730–742.
[17] R. Yao, Y. Zhang, Y. Zhou, S. Xia, Multiple small objects tracking based on dynamic Bayesian networks with spatial prior, Optik 125 (10) (2014) 2243–2247.
[18] P. Chiranjeevi, S. Sengupta, Moving object detection in the presence of dynamic backgrounds using intensity and textural features, J. Electron. Imaging 20 (4) (2011), 043009-043009.
[19] G. Jin, Y. Li, Y. Gu, J. Li, D. Gao, L. Liu, Moving target detection approach based on spatio-temporal salient perception, Optik (2014).
[20] P.J. Van L, E.H. Aarts, Simulated Annealing, Springer, 1987.
[21] J.P. Lewis, Fast template matching, in: Vision Interface, Vol 95, 1995, pp. 15–19.